THE UNIVERSITY *of* EDINBURGH

Modelling Cross-lingual Transfer For Semantic Parsing

Thomas Rishi Sherborne

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2024

Abstract

Semantic parsing maps natural language utterances to logical form representations of meaning (e.g., lambda calculus or SQL). A semantic parser functions as a human-computer interface by translating natural language into machine-readable logic to answer questions or respond to requests. Semantic parsing is a critical technology within language understanding systems (e.g., digital assistants) for accessing computational tools using natural language without expert knowledge or programming skills.

Cross-lingual semantic parsing adapts a parser to map more natural languages to logical form. Contemporary advances in semantic parsing generally only study parsing of English. Successful cross-lingual transfer for a semantic parser improves the utility of parsing technologies by enabling broader access to these tools. However, developing a cross-lingual semantic parser introduces additional challenges and trade-offs. High-quality data for new languages is scarce and requires complex annotation. Given available data, a parser must adapt to language variations in expressing meaning and intent. Existing multilingual models and corpora also exhibit extant biases for English, with variable cross-lingual transfer to languages with fewer speakers or resources. At present, there is no optimal strategy or modelling solution for teaching a new language to a semantic parser.

This thesis considers the efficient adaptation of a semantic parser from English to new languages. We are motivated by a case study of an engineer expanding a natural language database interface to new customers, seeking accurate parsing of new languages under a constrained budget for annotation. Overcoming the development challenges of cross-lingual semantic parsing requires innovation in model design, optimisation algorithms and strategies for sourcing and sampling data.

Our overarching hypothesis is that cross-lingual transfer is achievable through aligning representations between a high-resource language (i.e., English) and new languages unseen for the task. We propose different strategies for this alignment, exploiting existing resources such as machine translation, pre-trained models, data for adjacent tasks, or a few annotated examples in each new language. We propose different modelling solutions suited to the quantity and quality of cross-lingual data. First, we propose an ensembled model to bootstrap a parser from multiple machine-translation sources, improving robustness by exploiting lower-quality synthetic data. Second, we propose a zero-shot parser using auxiliary tasks to learn cross-lingual representation alignment without any training data in new languages. Third, we propose an efficient meta-learning algorithm optimising cross-lingual transfer during training

with a few labelled examples in new languages. Finally, we propose a latent variable model explicitly minimising divergence between representations across languages using Optimal Transport. Our results reveal that accurate cross-lingual semantic parsing is possible by composing minimal samples of target language data within models explicitly optimising for accurate parsing and cross-lingual transfer.

Lay Summary

Semantic parsing is a tool to convert human languages (e.g., English or French) into languages understandable to computers (e.g., code or logic). This is useful for computers to interpret the meaning of an instruction or question. For digital assistants such as Alexa or Siri, you would speak a command to your assistant, and the assistant must convert your command into computer logic to understand and respond to your command. Without semantic parsing, these assistant technologies are only helpful if you can communicate in complex computer logic.

However, a drawback of modern semantic parsing technologies is while they are generally adequate at understanding English, they can fail to understand any other language as successfully. This imbalance is a consequence of most data and tools for parser development existing only in English. Generating more data for this task is often complex and expensive. To teach a semantic parser more languages—we must examine how to use very little data efficiently. This goal is desirable as it will allow more demographics to access and use parsing technologies such as virtual assistants.

In this thesis, we consider different strategies to develop a semantic parser which can understand multiple languages (ranging from English and French to Chinese, Hindi or Thai) and generate logic outputs. If this semantic parser is successful, users who speak any of the available languages can use these tools just as an English speaker would. We analyse what makes this goal challenging for a semantic parser to propose new types of parsers to accurately understand questions from any language.

We propose multiple parser models using this high-quality data, focusing on a scenario where you have as little high-quality data as possible. With very little data, we can more easily create a parser requiring less expensive data creation. This thesis proposes four methods to create a parser for many languages. Our proposals for parsers can understand a command like 'How hot will it be at 3 pm?', or in Spanish '¿Qué tanto calor hará a las 3pm?', into the logical command to perform this function. We identify that building a parser requires high-quality data reflecting how real users ask questions. Overall, we build new parsing models by translating as little as 1% of the available data from English into these languages. Our models can better interpret commands and questions from any language from innovation in model and algorithm design. Our research contributes to improving semantic parsing for more different languages, making it more accessible and useful for a broader audience.

Acknowledgements

I would neither have started nor finished my PhD without the enduring support of my supervisor Mirella Lapata. I can express nothing short of infinite gratitude for Mirella's support, feedback and guidance since I started this journey in 2018. Mirella has supported me through every first draft, every paper rejection, every talk and poster, and every acceptance notification. Even when one reviewer refused to back down from the assertion that SQL is definitely just English—Mirella invested more kindness and guidance than I can ever repay into our shared work being the best it could possibly be. I sincerely thank Mirella for stopping me from taking a software engineer job in Tokyo to start a PhD.

I also express my gratitude to Mark Steedman, Alexandra Birch, Ann Copestake, and Polina Bayvel as supervising collaborators throughout my research career. I would not be where I am today without your efforts and mentorship. I thank my collaborators in Edinburgh: Nikita Moghe, Tom Hosking, Naomi Saphra, and Yumo Xu for turning our shared rants into peer reviewed contributions. I also thank my collaborators at the Allen Institute for Artificial Intelligence during my internship in 2023: Pradeep Dasigi, Hao Peng, Ananya Harsh Jha, Clara Na, Ian Magnusson and Emma Strubell. Thank you also to Valentina Pyatkin, Jesse Dodge, Hamish Ivison, Jacob Morrison, Nishant Subramani, Matt Finlayson, and the intern class of 2023 for challenging my unfounded theories on sharpness aware minimisation.

Innumerable people have contributed to my PhD journey. Thank you to my cohort in the CDT in Data Science. Thank you to Nelly Papalampidi and Seraphina Goldfarb-Tarrant for being such aspirational officemates. Thank you to my 2018 cohort of the Cambridge MPhil where I started my journey. Thank you to Heather Lent and Ruixiang Cui in Copenhagen for always listening to my first ideas. Thank you to my research group for always improving my second ideas. Thank you to all the friends I have made at conferences since ACL 2022 in Dublin.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Thomas Rishi Sherborne*)

Table of Contents

Chapter 1

Introduction

1.1 Semantic Parsing as a Human-Computer Interface

Semantic parsing is the task of realising the semantics of a natural language utterance in a computer-executable language. A semantic parser translates a natural language input into a machine-readable expression of meaning e.g., logical form, SQL, SPARQL, or code (Wilks and Fass, 1992; Mooney, 2007). When this logical form is expressed in some machine-readable language, a semantic parser functions as a human-computer interface by translating natural language into machine-readable logic for system control or interaction (Zelle and Mooney, 1996).

Semantic parsing is a long-standing natural language processing task with applications ranging from virtual environment control (Winograd, 1971), machine translation (Huang, 1990), robot interaction (Thomason et al., 2020), and formal linguistic analysis of corpora (Hershcovich et al., 2019). Early examples of semantic parsing include the SHRDLU system (Winograd, 1971), using a semantic parser to control a virtual "blocks world" using natural language. Users manipulate 3D virtual objects by expressing instructions in natural language, and the semantic parser translates instructions into computer-readable actions in the Lisp language. SHRDLU extends the pattern-matching design of earlier systems like ELIZA (Weizenbaum, 1966) for more advanced semantic understanding. Semantic parsing is the accessible interface allowing non-expert English speakers to access computational tools otherwise requiring expert programming skills.

Semantic parsing is vital in contemporary language understanding technologies, including virtual assistants such as Apple Siri or Amazon Alexa. A user speaks to a virtual assistant in natural language, and a semantic parser must translate this request into the respective logic to answer a question, create a calendar event, or respond to a

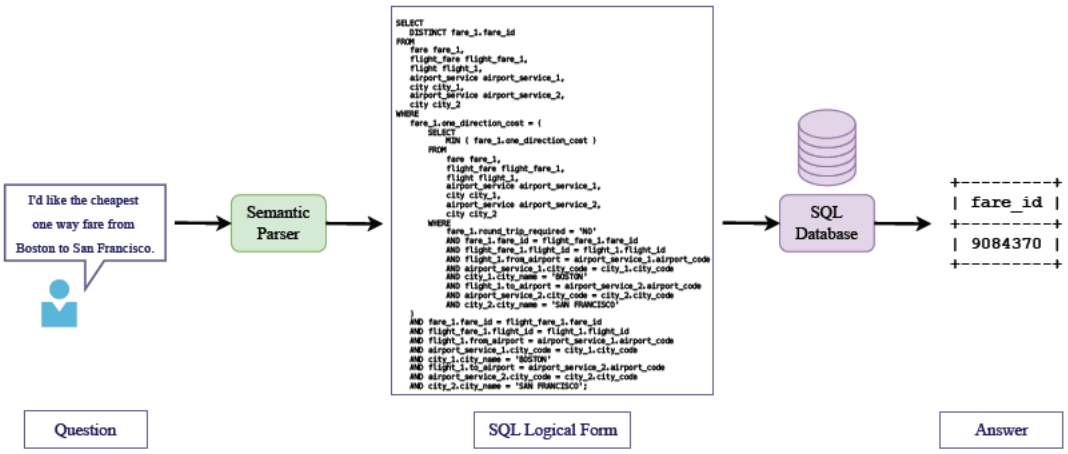

Figure 1.1: Executable text-to-SQL semantic parsing provides a natural language interface to structured data. The user can express a question in natural language and the semantic parser translates this question into the formal SQL data query language. We consider this SQL expression as the logical form of the input semantics with equivalent meaning. This logical form is executed in a database to return an answer to the user's question. Example taken from the ATIS dataset (Hemphill et al., 1990; Dahl et al., 1994).

text message (Kollar et al., 2018). Some type of semantic parsing underpins how these natural language utterances become machine-interpretable instructions. This thesis considers semantic parsing grounded in a relational database (i.e., semantic parsing with outputs executable in a database shown in Figure 1.1), and semantic parsing for spoken language understanding (i.e., SLU semantic parsing in Figure 1.2).

Semantic parsing is critical to modern natural language understanding engines in providing a capability to parse logical meaning from natural language (Kamath and Das, 2019). For semantic parsing to be accurate and useful, the system must overcome many challenges in comprehension. These challenges include:

- **Ambiguity** can arise due to underspecification, or the polysemy of words in an input. For example, a request to "give me directions to a park" must resolve the polysemic meaning of 'park' as a public garden or a space demarcated for a specific usage (e.g., 'car park').

- **Context** must be inferred to express meaning from utterances with uncommunicated detail. A speaker might assume implicit context for more efficient communication. However, this context must be made explicit in formal logic for a precise expression. For example, responding to the request "find me local Thai restaurants" must infer the current location to provide relevant results. Similarly,

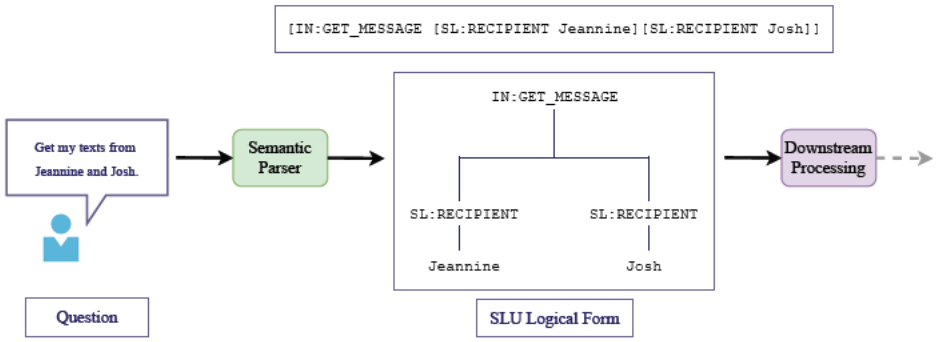

Figure 1.2: Semantic parsing for spoken language understanding (SLU) provides rich, fine-grained annotations to a question for downstream functionality. For virtual assistant systems, this parses entities, relationships and actions for task-specific processing. As a simulation of a chatbot-style interface, this is also referred to as *task-oriented dialogue* semantic parsing. Example taken from the TOP dataset (Gupta et al., 2018).

responding to "book me a flight to Dallas next week" requires knowing the current time, resolving when 'next week' occurs, and inferring the intended source airport.

- **World knowledge** is required to interpret relative relationships or meaning beyond verbatim comprehension. For example, the command "text sister good morning" must interpret the familial structure of the user. Similarly, knowledge of idioms is required to understand the phrase 'put lipstick on a pig' is not an instruction to apply cosmetics to an ungulate mammal.

Moreover, the modelling of a parser introduces additional difficulties in accurately representing logic for correct machine execution. The parser must generate a *well-formed* expression syntactically valid in the machine-readable language. This expression must be *concise* to unambiguously describe the input semantics without redundancy. The logic must also not be *spurious* in expressing inaccurate semantics with the same outcome e.g., the logical form of 'what is three squared?' is '3^2' and not '$3+3+3$', even if both expressions evaluate to '9'. The challenges of well-formedness, conciseness and spuriousness, among others, compound the difficulty of the task. This compounded complexity renders accurate and generalisable parsing as a long-standing challenge within natural language processing (Mooney, 2007).

Recent advancements in semantic parsing have demonstrated the strength of modelling the parser as a neural network (Dong and Lapata, 2016). This framework interprets semantic parsing as a case of machine translation from natural to formal languages

(Andreas et al., 2013). Modern systems, typically based on the Transformer encoder-decoder architecture (Vaswani et al., 2017), establish state-of-the-art performance across old and new benchmark datasets. Progress in semantic parsing is rapidly improving the capability to parse more complex language and generate more accurate logical forms.

However, this progress is largely limited to semantic parsing of *English* natural language. While this bias follows broader trends towards English within natural language processing, another barrier to multilingual semantic parsing is the annotation complexity requiring expert annotators and translators. Successful cross-lingual semantic parsing is desirable to improve access to parsing technologies for more users. Despite this, there is presently no obvious or easy strategy for increasing the language capabilities of a parser without a trade-off between accuracy, efficiency and cost. Without this vision, multilinguality in semantic parsing is less feasible as a realisable technology.

This thesis addresses these challenges by considering how to efficiently adapt a semantic parser from English to additional natural languages, subject to constraints on efficiency and cost. This is a cross-lingual transfer problem, defined in Section 1.2, with the objective of transferring accurate and generalisable semantic parsing from English to additional natural languages. We outline the challenges for this objective in Section 1.3. In Section 1.4, we frame our investigation as a case study of a budget-limited engineer extending an English-language interface to a database and virtual assistant. This case study motivates our hypotheses and contributions in Chapters 3 to 6. We consider the cost-performance trade-off to propose strategies for rapid, accurate cross-lingual semantic parsing with minimal resources.

1.2 Multilingual Natural Language Understanding

Multilinguality within natural language processing is the design of a system with consideration for more than a single natural language. Increasing the multilinguality of natural language processing systems is vital to allow speakers of more languages to access these systems (Joshi et al., 2020). Multilinguality methods include combining many natural languages as a practice for combined effort and resources. Examples of this include Universal Dependencies[1] (de Marneffe et al., 2021), unifying $>$ 100 languages within a single dependency parsing framework, and CommonVoice[2] (Ardila et al., 2020), a speech corpus of transcribed speech in $>$ 100 languages. This large-scale

[1] https://universaldependencies.org/
[2] https://commonvoice.mozilla.org/

multilingual modelling can mitigate a bias for any dominant language (typically English) and broadly improve performance by jointly modelling all languages simultaneously. Another perspective for multilingual natural language processing is the *cross-lingual* case of transfer learning where we assume data and models for a task in a *source* language (e.g., English), and desire to transfer task-specific capability from the *source* language to *target* language(s), where little or no data is available. We define *cross-lingual transfer* as the success or failure of a model generalising to a target language, with equivalent performance as the source language. Data-efficient cross-lingual transfer requires few, or zero, data examples in target languages. Inefficient cross-lingual transfer will demand more data examples, and associated annotation burden, for the same outcome. Cross-lingual transfer is often integrated with the aforementioned multilingual resource pooling. However, the cross-lingual perspective is the explicit goal of expanding the task from known to unknown natural languages. In this thesis, we follow this paradigm to define and investigate cross-lingual semantic parsing from English to target natural languages.

Contemporary multilingual modelling begins at model pre-training. The multilinguality of large pre-trained models can be principally improved by increasing the diversity of languages, and quantity of data, in pre-training corpora. Pre-training on multilingual corpora allows a model to learn representations or inherit knowledge from multiple languages, and combine knowledge from lexical, syntactic and semantic features across languages. Monolingual pre-trained models such as BERT (Devlin et al., 2019) or BART (Lewis et al., 2020a) generally fail at cross-lingual transfer given the lack of adequate vocabulary or exposure to multilingual data. The creation of large multilingual text corpora such as the *multilingual Colossal Common Crawl Corpus* (Xue et al., 2021, mC4) has enabled pre-training of large models with nascent multilingual capabilities. Models such as XLM-ROBERTA (Conneau et al., 2020), MBART (Liu et al., 2020), and MT5 (Xue et al., 2021) broadly improve cross-lingual transferability for natural languages included during pre-training. However, these task-agnostic models typically must be specialised for task-specificity using relevant training data i.e., a *fine-tuning* training pipeline. Modern challenges to multilingual modelling often centres around applying monolingual fine-tuning (e.g., training on semantic parsing data in English only) to languages with scarce task-specific resources.

Modern approaches to cross-lingual semantic parsing focus on leveraging large language models (LLMs) to mitigate the difficulty and expense of multilingual data generation. This includes using LLMs to generate synthetic data for training in a low-

resource language (Rosenbaum et al., 2022), or using LLMs to provide language-specific annotation of LLM-generated synthetic data (Nicosia et al., 2021). This generated data is termed "silver-standard" as it is machine-generated without human annotation. These approaches combine LLMs with lower-quality data to improve cross-lingual parsing accuracy. Similar work has identified that LLMs can be prompted for effective compositional generalisation (Drozdov et al., 2023) or fine-tuned for low-resource domain adaptation (Schucher et al., 2022) within semantic parsing. While cross-lingual prompting is an expanding domain (Shi et al., 2022; Asai et al., 2023), fine-tuning presently proves more effective for low-resource languages often poorly represented in pre-training corpora (Ruder et al., 2023). Solutions in this thesis focus on fine-tuning a multilingual pre-trained <1 billion parameter model for our task. We employ fine-tuning as a proven strategy of cross-lingual adaptation across many tasks (Zhao et al., 2021) and consider a smaller model scale to exploit efficient optimisation with fewer computational resources. In Chapter 6, we identify how smaller models with higher-quality data can actually outperform the aforementioned larger language models. We consider our methods as a proof-of-concept for efficient cross-lingual semantic parsing. In Chapter 7, we discuss how our contributions can be applied to LLM-scale models in future work.

1.3 Challenges

Building accurate and generalisable cross-lingual semantic parsing requires addressing challenges in both structured prediction and cross-lingual modelling. Table 1.1 outlines examples in English and the same meaning in different languages which we now desire to parse. While we mention difficulties inherent to semantic parsing in Section 1.1, this thesis primarily considers solutions for challenges in cross-lingual transfer. These challenges include:

Data scarcity influences the success of multilingual models by limiting the tasks and languages with available resources. At a minimum, building a task-specific model for a new language requires data to evaluate generalisation. However, translating all available training data into all >7000 global languages (Dryer and Haspelmath, 2013) is an intractable task, leading to a significant focus on avoiding generating training data for every target language. This includes using machine translation (Conneau et al., 2018a), zero-shot transfer from English (Hu et al., 2020; Liang et al., 2020), or translating as

MultiATIS++SQL (see Chapter 2)	
EN	I'd like the cheapest one way fare from Boston to San Francisco.
FR	J'aimerais trouver le tarif aller simple le moins cher de Boston à San Francisco.
ZH	请帮我查找从波士顿到旧金山最低的单程票价
MTOP (Li et al., 2021)	
EN	Can you tell me how much rain is expected tonight?
ES	¿Puedes decirme cuánta lluvia se espera para esta noche?
DE	Kannst du mir sagen, wie viel Regen heute Nacht erwartet wird?

Table 1.1: Input data examples for MultiATIS++SQL (see Chapter 2), and MTOP (Li et al., 2021) datasets in English (EN), French (FR), Simplified Chinese (ZH), Spanish (ES), and German (DE). The logical form output is semantically equivalent for all natural language inputs. The parser must recognise and interpret variations in utterance structure, lexical similarity, politeness, and tense across languages.

few examples as possible (Zhao et al., 2021). While this can improve performance, professional or machine translation can result in multilingual systems which poorly model native speakers (Koppel and Ordan, 2011), leading to poor generalisation for actual users. Data annotation for semantic parsing is also highly complex—requiring domain expertise and skill in manually writing logical forms (Perez-Beltrachini et al., 2023). To address these challenges of data scarcity, this thesis considers multiple strategies for data collection and sampling to effectively build accurate and useful target language data resources.

Cross-lingual language variation influences lexical, syntactic and semantic similarities and dissimilarities between target languages. These contribute to the difficulty in learning unified semantic and syntactic structures from data across any task. Lexical variation includes how homographs vary across languages: the German noun "Küche" has multiple meanings translating to either "kitchen" or "cuisine" in English. Similarly, the Chinese proper noun "圣保罗" is a phonetic translation for either "São Paulo" and "St. Paul". Resolving the correct entity is critical for logical forms executing a search for flights to one location. Syntactic variation is observed when queries follow different structures: the English question "Do you have X?" can be misinterpreted as the declarative statement in Chinese "你有一个X [you have one X]", but Chinese

more commonly uses a positive-negative alternation pattern to query possession ("有没有一个X?" [have not have one X?]). Semantic variation includes differences in politeness between languages: some dialects of English may drop politeness (e.g., 'please', 'thank you') when interacting with computer interlocutors. However, a Chinese speaker may use "请 [qǐng]" at the beginning of an utterance to denote a polite request to a computer. Revisiting idioms, a system must infer that "you can't teach an old dog new tricks" and "loro viejo no aprende a hablar [an old parrot never learns to talk]" are equivalent expressions in English and Spanish. Additional challenges include grammatical variation across languages including the morphology for tense or grammatical gender. Successful cross-lingual transfer requires resolving these ambiguities or dissimilarities for accurate parsing.

Existing resources are biased towards English and other languages with more available resources (i.e., high-resource languages). As a consequence, multilingual systems often are most performant on English tasks and task-specific capability attenuates for other languages (Blasi et al., 2022; Søgaard, 2022). This English-centric bias is observed in pre-trained models (Lai et al., 2023), and multilingual datasets (Kreutzer et al., 2022) widely used to develop and evaluate cross-lingual transfer. The overrepresentation of English disproportionately allocates these resources to English-centric tasks. As a consequence, a majority of natural language processing technologies are realistically only technologies for English (Ananiadou et al., 2012). Furthermore, normalising only English data for multilingual resources biases systems for the cultural norms and perspectives of English-speaking nations (Hershcovich et al., 2022). This imbalance exacerbates the challenge in semantic parser development by correlating the success of cross-lingual transfer with the similarity between the target language and English. Languages which share many features (e.g., words or syntax) with English (e.g., French, German) are more likely to benefit from transfer from English than languages which share less (e.g., Chinese, Thai) (Lauscher et al., 2020). A parser should overcome this English-centric bias for accurate cross-lingual transfer to similar and dissimilar target languages.

Representation alignment is a long-standing research objective for enabling cross-lingual transfer (Conneau et al., 2017; Lauscher et al., 2020; Wei et al., 2021). If a system can build a representation space where equivalent semantics from multiple languages are represented equivalently; then any task conditioned upon this representation

space will perform equivalently for any input language (Conneau et al., 2020). However, generalisable representation alignment is challenging for successful cross-lingual transfer without sufficient high-quality data. Multilingual pre-trained models learn some common linguistic features but often allocate specific representation subspaces to different languages without global representation alignment (Chang et al., 2022). Practically, a representation is more likely to be closer to any representation from the same language, rather than the same meaning in a different language (i.e., monolingual clustering). This limits the cross-lingual transfer capability for tasks reliant on aligning semantics from different input languages. For an encoder-decoder semantic parser, cross-lingual alignment can be studied as the similarity between latent encodings output from the encoder. If an encoder outputs the same latent representation of meaning from an input in any language, the decoder will deterministically map this representation to the same logical form. In this thesis, we interpret cross-lingual alignment between latent representations as a sufficient condition for cross-lingual transfer. Our methods focus on producing this alignment within a parser, and our results study and visualise this alignment as an interpretation of parser success. We further discuss the conditions for this alignment in Section 2.4.2.

These challenges unify into **no obvious modelling solutions** for cross-lingual semantic parsing. Prior efforts scarcely centre the cross-lingual perspective to exploit the fewest examples for the most accurate system in new languages. Jones et al. (2012) propose an adaptation of the English GeoQuery dataset (Zelle and Mooney, 1996) into six languages, but produce only parallel monolingual parsers for each language. Susanto and Lu (2017a) propose an early neural network semantic parser jointly modelling all languages in a single model. Neither explicitly models cross-lingual transfer without parallel data. Duong et al. (2017a) partially addresses this challenge using machine translation for parsing German or English. However, the evaluation does not extend to any less similar languages (e.g., Chinese) due to a lack of available data. Representation alignment is also not studied as a cause for, or effect of, cross-lingual parsing success.

In this thesis, we present solutions for these challenges in building a cross-lingual semantic parser. We propose different modelling solutions suited to the quantity and quality of cross-lingual data. When data is not available, prior work does not address the poor robustness of machine translation or the potential to exploit available data from adjacent tasks. This thesis addresses these issues through an ensembling approach in Chapter 3, and a multi-task zero-shot parser in Chapter 4. If data is available, prior

work does not address developing a multilingual Transformer parser without suffering *the curse of multilinguality* (Conneau et al., 2020): a phenomenon where expanding the supported languages in a system reduces overall performance. This thesis addresses these issues through algorithm design in Chapter 5, and a new parser architecture in Chapter 6.

We highlight a rare quality of semantic parsing well suited for evaluating cross-lingual transfer. Semantic parsing requires fine-grained semantic understanding to map the natural language input into the machine-readable language output. In a cross-lingual scenario, the same fine-grained semantic understanding must wholly transfer for successful parsing. The misinterpretation of entities, relationships, and relative or numerical expressions can all result in an incorrect parse. Evaluation for executable semantic parsing is strict by evaluating if a parser output is syntactically correct or represents equivalent precise meaning. This provides a binary signal testing if the required fine-grained semantic understanding is transferred to each target language. We argue that tasks like entailment classification or machine translation are susceptible to misrepresenting cross-lingual transfer. Semantic equivalence can often be ignored in machine translation, where sensitivities to idioms and culturally-specific expressions often fail to translate across languages without explicit modelling (Dankers et al., 2022). A similar semantic misinterpretation may not be detectable in an evaluation reliant on labels or overlap-based metrics. In contrast, failure in cross-lingual transfer for semantic parsing results in null accuracy. Therefore, our contributions evaluate fine-grained semantic transfer between languages, with robust applicability to future tasks and problems.

1.4 A Case Study for an Engineer

This thesis proposes modelling methodologies for cross-lingual adaptation of a semantic parser. For this objective, we consider the following case study:

> A company manages many databases of structured information. These databases are accessible using a Structured Query Language (SQL) server, where a user provides SQL logic to retrieve the desired information from the database. The company desires to make this information more accessible to consumers. However, the target customer is not an expert in databases or programming. The company desires a natural language interface to the databases to allow more natural customer interaction. The company also desires this interface to be conversational (i.e., a chatbot). The company hires an engineer to develop this interface.
>
> The engineer builds an initial interface using a text-to-SQL semantic parser for database interaction. The parser is a neural network: specifically, a pre-trained model fine-tuned on semantic parsing data. The engineer builds the initial system for interactions in English because the company's databases are structured in English.
>
> However, the company has a global customer base speaking multiple languages. If the interface can fluently handle requests from more languages, the system will be accessible to more customers. The engineer is now tasked with prototyping a multilingual interface. The system must be able to interpret an input from any of the languages to be supported. The ideal solution will generate a machine-readable expression from the input with equivalent accuracy to the currently supported English. The engineer considers this as a cross-lingual transfer problem: adapt a semantic parsing model from English to additional languages. Additionally, there are no multilingual datasets suitable for training this model and only a small budget to create data for the prototype. Therefore, any expense must maximise the cross-lingual transfer capability to validate the prototype as a feasible product. The engineer must now examine how to build this system efficiently and within budget.

This thesis directly addresses the goals of the case study by (i) proposing modelling and optimisation solutions to developing a cross-lingual semantic parser, and (ii) investigating data efficiency and sampling strategies to maximise the utility of a limited data budget. This case study assumes access to multilingual *evaluation* data to test

generalisation and cross-lingual transfer i.e., producing evaluation data is outside of the limited prototype budget. This data evaluates if a model will generalise to *real users*. A single data example is a paired natural language utterance input and logical form output. The budget is allocated for annotating training data to develop the parser and computational resources for model training. Extrinsic success for the prototype is measured by comparing the accuracy of the parser in each target language to the parsing accuracy for English. A successful parser will have equivalent accuracy in all languages. Intrinsic success is cross-lingual representation alignment within the neural network model. A successful parser will generate equivalent latent representations from inputs with equivalent semantics from different languages.

1.5 Thesis Overview

Our overarching hypothesis is that **cross-lingual semantic parsing is enabled by aligning representations between English and target languages**. Our investigation studies how to achieve this alignment with data-efficient methods. For this study, we investigate four strategies for cross-lingual adaptation evaluating individual hypotheses. We consider strategies for data using machine translation, borrowing target language data from similar tasks for zero-shot cross-lingual transfer, and few-shot sampling to exploit a few annotated examples. This represents the least to most expensive method for generating data for the case study. Our modelling and data strategies are analysed by studying the cross-lingual representation alignment within the model. We define this objective in Section 2.4.2, and propose alignment-forward methods in Chapter 4 and Chapter 6. The outline and hypotheses for each chapter are as follows:

- Chapter 2 describes background material for our contributions. We describe the parsing task, natural and formal languages, primary parser model, model optimisation, evaluation criteria, and analysis methods.
- Chapter 3 considers the hypothesis that **machine translation can adequately approximate natural language for cross-lingual transfer in semantic parsing**. We investigate if open-source machine translation can adequately translate test data from target languages to English, or translate training data from English to target languages. Given this cost-effective "silver-standard" data, we explore if these resources are sufficiently similar to produce systems which generalise to data from speakers of target languages. We propose FATES, an ensemble

parsing model combining multiple translations to improve cross-lingual transfer robustness by modelling and fusing multiple, parallel language models from each translation source.

- Chapter 4 considers the hypothesis that **auxiliary multilingual tasks and data can induce cross-lingual representation alignment without target-language training data**. We investigate training a semantic parser on English data combined with simultaneous training on additional loss functions for other tasks where multilingual data is readily available. Our study considers learning cross-lingual representation alignment by exploiting alternative resources, without creating more data specifically for our task. This system is 'zero-shot' as the parser does not observe labelled parsing data in any target language during training. We propose ZEUS, a multi-task model optimising a semantic parser with domain-adaptive pre-training, machine translation, and input language classification. We study which auxiliary objectives contribute to representation alignment, and analyse the role of different types of multilingual data for this effect.

- Chapter 5 considers the hypothesis that **meta-learning improves few-shot cross-lingual transfer by promoting gradient-level cross-lingual regularisation of task-specific training**. We now examine if creating a small sample of target language examples can improve upon our prior proposals. We propose DRAKON, a meta-learning approach optimised for data-efficient few-shot cross-lingual semantic parsing when translating $\leq 10\%$ of the dataset. DRAKON aligns gradients from data in English and target languages, leading to significant improvement in parsing capability above any few-shot competitor methodology. We also identify how few-shot transfer yields more accurate parsing without inheriting issues of fluency (from Chapter 3) and domain relevancy (from Chapter 4). We also observe that DRAKON improves cross-lingual representation alignment via optimising cross-lingual gradient similarity during training.

- Chapter 6 considers the hypothesis that **explicit cross-lingual representation alignment improves the transfer of task knowledge to target languages**. We consider explicitly minimising the divergence between representations across languages. This technique directly targets the representation alignment hypothesis previously studied as an outcome of other methods. We propose MINOTAUR, a methodology for minimising cross-lingual representation divergence using Optimal Transport. We define a model augmented with latent variables, which

uses a multi-level divergence penalty to directly target cross-lingual representation alignment. MINOTAUR demonstrates more accurate parsing using fewer examples and training computation than DRAKON. We identify how MINOTAUR yields the closest representation alignment in this thesis, such that equivalent semantics from different languages are now much closer in latent space.

- Chapter 7 summarises our contributions and defines topics of future work applying our contributions to larger models, and extending our conclusions to languages with fewer resources.

We highlight that the contributions in this thesis have previously been reported in the following publications:

Chapter 3 Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a Crosslingual Semantic Parser. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 499–517, Online. Association for Computational Linguistics.

Chapter 3 Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic Evaluation of Machine Translation Metrics. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Chapter 4 Tom Sherborne and Mirella Lapata. 2022. Zero-Shot Cross-lingual Semantic Parsing. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Chapter 5 Tom Sherborne and Mirella Lapata. 2023. Meta-Learning a Cross-lingual Manifold for Semantic Parsing. Transactions of the Association for Computational Linguistics, 11:49–67.

Chapter 6 Tom Sherborne, Tom Hosking, and Mirella Lapata. 2023. Optimal Transport Posterior Alignment for Cross-lingual Semantic Parsing. Transactions of the Association for Computational Linguistics, 11:1432–1450.

Chapter 2

Background

In this chapter, we discuss the format of semantic parsing in terms of natural language input (Section 2.1.1) and logical form output (Section 2.1.2). We present the primary Transformer model we build on in our contributions (Section 2.2) and introduce datasets used for evaluation (Section 2.3). Finally, Section 2.4 formally defines our cross-lingual transfer framework for sampling data distributions, learning scenarios, and representation alignment.

2.1 Natural Languages and Logical Forms

2.1.1 Natural Languages

Languages develop within communities as a tool to satisfy the need for communication (Christiansen and Kirby, 2003). Semantic parsing facilitates communication for a human-computer interface translating between human- and machine-readable languages. Semantic parsing differs from general purpose language understanding tasks, such as language modelling, in generally focusing on *utterances* rather than any form of language. An utterance can be characterised as a 'continuous piece of speech separated by silence' (Harris, 1960). We inherit this terminology from the foundations of the task as a communication interface translating natural language utterances into commands (Winograd, 1971). For our task, the utterance is one input to be parsed. This is typically a single sentence, and we do not consider document or book-level parsing in this thesis. Natural language utterances in semantic parsing are typically *imperative* or *interrogative* (see examples in Table 2.1). Instructions are often imperative, commanding the parser to execute some function, and questions are often interrogative, requesting some

Form	Example
Declarative	Rain is expected for tonight.
Exclamatory	It's going to rain tonight!
Imperative	Tell me how much rain is expected tonight.
Interrogative	How much rain is expected tonight?

Table 2.1: Examples of utterances in English illustrating variations in surface form for approximately equivalent intent (Radford, 2009). Semantic parsing typically focuses on imperative and interrogative utterances (*italicised*) typically corresponding to queries and requests within a human-computer interface.

information from the interlocutor (Radford, 2009). Building a semantic parser requires accurately interpreting this structure to infer meaning.

Moreover, a parser must be sensitive to pragmatics to infer the correct action from any sentence structure. Declarative sentences can contain questions or instructions depending on the speaker's original intent. For example, the interrogative command 'Can you tell me how much rain is expected tonight?' can be literally interpreted as asking whether it is possible to determine the expected rainfall. The parser should recognise that because action is possible, the system should execute the requested instruction. Similarly, the utterance 'I want to know the expected rainfall tonight' is declarative, but the parser should interpret this as an executable instruction similar to the imperative form. We broadly group samples of language into *questions*, encompassing any utterance directed at a parser with an intended action, and *declarative text*, encompassing any arbitrary statement assumed not to contain an actionable request. The generalisation capability of a parser is influenced by the parser's ability to apply domain- or language-specific rules and structure to new utterances (Herzig and Berant, 2018). We will revisit this contrast in Chapter 4 for adapting a parser using samples of web-sourced text.

As discussed earlier in Section 1.3, semantic parsing follows the trend of bias toward English within natural language understanding tasks (Søgaard, 2022). Recent advances in semantic parsing research primarily focus on data models for English only. To address this imbalance, this thesis considers parsing the following languages (relevant demographic details from Dryer and Haspelmath (2013) and Eberhard et al. (2019)):

- **English** (EN): Germanic Indo-European language using Latin script. Natively

used by 380 million speakers and in common usage by $>$ 1 billion second-language speakers. EN has official status in 47 countries and territories. English is the *source* language in this thesis.

- **French** (FR): Romance Indo-European language using Latin script. FR is the native language of approximately 99 million people and is an official language in 39 countries.
- **Portuguese** (PT): Romance Indo-European language using Latin script. PT is an official language in 10 countries with approximately 236 million native speakers.
- **Spanish** (ES): Romance Indo-European language using Latin script. ES is an official language in 22 countries with approximately 485 million native speakers.
- **German** (DE): Germanic Indo-European language using Latin script. DE is an official language in six European countries and is the native language of approximately 75 million people.
- **Chinese** (ZH) i.e., Simplified Mandarin Chinese: This is the official form of Chinese in the People's Republic of China with 939 million native speakers. ZH is a Sino-Tibetan language using a logographic script where each character represents a singular morpheme. We consider only Simplified Chinese characters in this work. While widely used as a writing system for Chinese, we note that Simplified Chinese is not a universal written language for the Mandarin dialect. Traditional Chinese characters are preferred in Taiwan. Simplified Chinese is also not used in Hong Kong within Cantonese, and is not borrowed in Korea or Japan for their respective writing systems.
- **Hindi** (HI): Indic Indo-European language using the Devanagari script. HI is the primary official language of India with 345 million native speakers. Devanagari script is an *abugida*, a writing system where consonant-vowel sequences are a singular unit (Matthews, 2014). Each unit (analogous to a Latin character) uses a consonant letter with a vowel noted using a diacritic. We note that HI has no concept of character case.
- **Thai** (TH): official language of Thailand in the Kam-Tai genus from the Tai-Kadai language family with 21 million native speakers. Thai is related to Lao, the official language in Laos, and Shan, a minority language in Myanmar. The writing system of TH is similar to HI as an abugida alphabet without casing.

We consider these target languages given prior data availability, outlined in Section 2.3, as a proof of concept for our case study. We note that this thesis does not consider any language genuinely considered low-resource, as all languages under consideration have millions of native speakers (Cieri et al., 2016). As our methods assume little resource parity between languages—we expect our contributions to extend to lower-resource languages in the future. While our discussion of differences in declarative or interrogative utterances uses English, we assume that differences between questions and declarative text follow similar trends in each target language (Culbertson et al., 2020). Our cross-lingual semantic parser must adapt to language-specific utterance structure, in addition to the challenges of lexical ambiguity, syntactic variation, and semantic variation outlined in Section 1.3.

2.1.2 Logical Forms

Semantic parsing translates a natural language utterance into precise meaning representation for automated reasoning (Zelle, 1995; Zettlemoyer and Collins, 2005). This representation is a logical form (LF) in some machine-readable language (MRL) representing meaning abstracted from natural language (Fodor, 1975). Some machine-readable languages, such as $\lambda-$calculus or $\lambda-$dependency compositional semantics, can be considered formal languages exactly describing computation by applying functions and abstractions over defined variables. For example, the GeoQuery dataset defines a deterministic mapping from $\lambda-$calculus LFs to Prolog programs for execution (Zelle and Mooney, 1996; Kate et al., 2005). Logical forms such as SQL are not strictly considered formal languages as they do not follow a formal grammar, and are not designed for concisely expressing input semantics. More generally, a logical form can be considered the specification of a program of execution. This program describes natural language meaning as logical processing steps for a computer to act on, regardless of the intended use of the MRL. The design and structure of an MRL influence how accurately an LF can model semantics (Mooney, 2007), and be learned by a neural parsing model (Li et al., 2022). While we generate LFs as parser outputs, this thesis does not study the design, structure and suitability of MRLs. This has been studied extensively in prior work (e.g., Zelle, 1995; Zettlemoyer, 2009; Wang, 2021), and our contributions focus on improving parsing more diverse languages for a given parser. The machine-readable languages used in this thesis are as follows:

2.1.2.1 Structured Query Language

Structured Query Language (SQL) is a programming language for interaction with relational databases (Beaulieu, 2009). A user composes an SQL query to retrieve and manipulate the data stored in the relevant database. Each database stores data in *tables*, with an element represented as a *row* with attributes indexed by *columns*. The complexity in SQL is often associated with retrieving and aligning information from multiple tables for processing into a result (Li et al., 2022). Table 2.2 shows an example of SQL as a logical form. The equivalent $\lambda-$calculus succinctly describes the input semantics, and SQL expresses the same meaning in a form suitable for database execution. While not concise, parsing into SQL (i.e., *text-to-SQL* semantic parsing) is practical by allowing direct interaction between a parser and any relational data store. This enables the deployment of a parser by ignoring any 'tidiness' in expressing meaning in a formal language.

We refer to text-to-SQL parsing as *executable* semantic parsing, given we can evaluate SQL LFs in an existing database. A user asks a question, the semantic parser translates the utterance to an SQL logical form, and the database evaluates this LF to return an answer to the user. This satisfies the case study as a pipeline for question answering over structured data. An advantage of SQL as a logical form is that SQL is *grounded* in a database. Within semantic parsing, grounding refers to the environment used to validate the truth of the logical form (Clark and Brennan, 1991; Chandu et al., 2021). A grounded logical form is one constructed with respect to some environment usable for evaluating the program logic. SQL is grounded in a database, and the database can be used to execute the logical form to look up an answer (or *denotation*). Ungrounded logical forms lack this relationship to an environment for verification e.g., dependency trees (Reddy et al., 2016) or universal dependency graphs (Reddy et al., 2017).

2.1.2.2 Task-Oriented-Parsing Logical Form Language

Task-Oriented-Parsing Logical Forms (TOP-LF) were introduced in Gupta et al. (2018) as a semantic annotation scheme for expressive spoken language understanding. The logical form encapsulates the sentence-level intent classification and token-level slot labelling tasks jointly referred to as "spoken language understanding" (SLU). Each intent and slot annotation is used in downstream applications in a dialogue system. Gupta et al. (2018) argue that prior non-hierarchical SLU formats are insufficiently

EN	What flights are there from Newark to Seattle on Saturday?
λ−calculus	`lambda $0 e (and (flight $0) (from $0 newark:ci) (to $0 seattle:ci) (day $0 saturday:da))`
SQL	SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2, days days_1, date_day date_day_1 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'NEWARK' AND(flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'SEATTLE' AND flight_1.flight_days = days_1.days_code AND days_1.day_name = date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 7 AND date_day_1.day_number = 26);

Table 2.2: An example of a English (EN) input and the associated logical forms in λ−calculus from Zettlemoyer and Collins (2005), and SQL from Dahl et al. (1994). The SQL LF is much less succint in expressing the input semantics than λ−calculus. However, designing a semantic parser for SQL allows a parser to directly interface with a relational database for question answering.

expressive to allow recursive functions or represent complex annotation. This motivates the proposal for TOP-LF for more expressive logical forms with easy annotation and simple evaluation. The TOP-LF language is defined as an annotated tree structure similar to a constituency parse tree for syntax. The TOP-LF logical form parses *intents*, actions with '`IN:`' labels, and *slots*, attributes and entities with '`SL:`' labels as annotations in the utterance. These roles map to actions, with respective inputs from slots, defined externally by the dialogue system. Table 2.3 shows an example of this structure with a nested intent ('`IN:GET_RESTAURANT_LOCATION`') within the slot '`SL:DESTINATION`'. This recursive expressivity is not possible in a simpler SLU schema. The tree structure always uses an intent label as the root node and subtrees are

EN	How far is the coffee shop?
TOP-LF	`[IN:GET_DISTANCE[SL:DESTINATION [IN:GET_RESTAURANT_LOCATION [SL:TYPE_FOOD coffee]]]]`

Table 2.3: An example of a English (EN) input and the associated logical form in TOP-LF from the TOP dataset (Gupta et al., 2018). The TOP-LF form provides a rich hierarchical annotation on an utterance useful for processing in a spoken language understanding pipeline. The parser extracts *intents*, actions with '`IN`' labels, and *slots*, attributes and entities with '`SL`' labels. An intent defines a function for the system to execute with slots defining the inputs for each function.

indicated by bracketing where paired brackets, '`[ ]`', indicate a new subtree. Similar to SQL, TOP-LF is not a strict formal language as a rich labelling schema for utterances.

We refer to generating TOP-LF logical forms from utterances as *parsing for spoken language understanding*, or SLU semantic parsing. Gupta et al. (2018) propose TOP-LF as a form of semantic parsing motivated to allow precise semantic annotation for spoken language understanding tasks. We follow this by considering TOP-LF outputs as logical forms within a sequence-to-sequence semantic parsing framework. This addresses the case study in building a semantic parser suitable for dialogue-based customer interaction i.e., a chatbot.

The format of TOP-LF is suitable for a dialogue system where downstream applications will act on the parsed intents and slots. These applications are external to the parser and excluded from our system. There is no publicly available dialogue system for evaluating TOP-LF logic. Therefore, TOP-LF logic is *ungrounded* without an environment representing the ground truth[1]. Evaluating TOP-LF logic compares to gold-standard parses provided by annotators, analogous to evaluating SQL by comparing SQL query tokens instead of the executed result from predicted queries. We further discuss evaluating TOP-LF logic in Section 2.3.2.

Either executable or SLU semantic parsing provides an interface for user interactions. The logical forms produced by a semantic parser must be syntactically well-formed in the respective MRL, semantically faithful to the input, and sufficiently expressive to unambiguously describe the utterance as a program. We now describe developing a parser to satisfy these requirements in Section 2.2, and we revisit datasets and evaluating

[1] Gupta et al. (2018) claim that existing dialogue systems could easily be adapted to evaluate TOP-LF. However, we consider TOP-LF ungrounded in the absence of an exemplar evaluation environment.

logical forms in Section 2.3.

2.2 Modelling

Task Definition Given a natural language *utterance* $x = \{x_0, x_1, \ldots, x_T\}$ comprised of T tokens in functional domain $\mathcal{X}$, a semantic parser maps input x to a *logical form*, $y = \{y_0, y_1, \ldots, y_{T'}\}$ of T' tokens in functional domain $\mathcal{Y}$. In the executable case (e.g., if y is SQL), the logical form or program can be executed inside an associated database environment, e, to retrieve a *denotation*, d, as an answer to the original natural language query x. We describe the parser as a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ parameterised by weights θ and optimised using some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$. We note that unlike prior work considering parsing for new domains (Wang, 2021), we do not consider the environment e as a model input. We consider the following notation in this thesis:

X, Y, Z	Random variables
x, y, z, $\mathbf{z}$	Observations
P_X, P_Y, P_Z	Probability distributions
$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$	Functional domains
$\mathcal{D}$, $\mathcal{S}$, $\mathcal{T}$, $\mathcal{U}$	Data distributions and samples

2.2.1 Designing a Semantic Parser

We define the semantic parser function, f, as a sequence-to-sequence Transformer model (Sutskever et al., 2014; Vaswani et al., 2017). We follow this framework given recent strengths in semantic parsing (Dong and Lapata, 2016, 2018), and cross-lingual transfer (Duong et al., 2017b; Susanto and Lu, 2017b; Liu et al., 2020; Tang et al., 2021). The model comprises an encoder, Q, and a decoder, G. The encoder is responsible for transforming the input $x \in \mathbb{R}^{T \times |V|}$, T tokens indexed in encoder vocabulary V of size $|V|$ from domain $\mathcal{X}$, into a latent representation $E \in \mathbb{R}^{T \times d}$ of T vector representations of x with dimensionality d. The decoder is responsible for predicting output $y \in \mathbb{R}^{T' \times |V'|}$, T' tokens in decoder vocabulary V' of size $|V'|$ from domain $\mathcal{Y}$, from the latent tensor E. The function parameters θ can be decomposed into $\theta = \{\phi, \psi\}$ where ϕ are the encoder parameters and ψ are the decoder parameters. The encoder is Q_ϕ: function Q with parameters ϕ, and the decoder is similarly denoted as G_ψ. Equation (2.1) is

the parsing function using these components, and the initial output token, y_0, to start decoder generation.

The parser models the likelihood of output y using function f as Equation (2.2), where each element of y is predicted sequentially from input x and prior outputs $y_{<t}$. The output distribution for token y_t is defined in Equation (2.3): the decoder, G_ψ, predicts a logits output over the vocabulary conditional on prior outputs, $y_{<t}$, and the encoded input, $Q_\phi(x)$. This logits output is reparameterised as a distribution using the softmax function to predict a probability for y_t. Unless specified otherwise, we use a singular encoder model and a singular decoder model for all inputs and outputs.

$$f(x,\ \theta) :- G_\psi\left(y_0,\ Q_\phi(x)\right) \tag{2.1}$$

$$p_f(y|x,\ \theta) = \prod_{t=1}^{T'} p_f(y_t|y_{<t},\ x,\ \theta) \tag{2.2}$$

$$p_f(y_t|y_{<t},\ x,\ \theta) = \text{softmax}\left(G_\psi\left(y_{<t},\ Q_\phi(x)\right)\right) \tag{2.3}$$

Transformer Layers The encoder and decoder are stacked Transformer layers comprising multi-head attention and feedforward sublayers. Figure 2.1 outlines the Transformer architecture as defined by Vaswani et al. (2017). An encoder layer computes a contextual representation for each input element using *multi-head self-attention*, with the final encoder layer producing the latent representation. The decoder conditions upon this latent representation during the autoregressive generation of output tokens (i.e., one step of Equation (2.3)). From Vaswani et al. (2017), the multi-head attention layer for inputs query (Q), key (K), and value (V), is defined as Equation (2.4), with each head defined in Equation (2.5) using the attention mechanism in Equation (2.6).

$$\text{MultiHeadAttention}(Q,\ K,\ V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O \tag{2.4}$$

$$\text{head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.5}$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}})V \tag{2.6}$$

$$\text{FeedForward}(X) = \text{ReLU}(XW_i + b_i)W_j + b_j \tag{2.7}$$

$$\text{ReLU}(x) = \max(x, 0) \tag{2.8}$$

Each head, $i \in h$, uses dimensionality $d_h = \frac{d}{h}$ to define learnable parameters W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d \times d_h}$. Each head specialises in some degree of language understanding with these specialisations combined during forward processing (Voita et al.,

2019). Multi-head attention combines these states and projects through learned output parameter $W^O \in \mathbb{R}^{hd_h \times d}$. The feedforward sublayer is defined in Equation (2.7), where ReLU is the rectified linear unit nonlinearity activation function (Fukushima, 1975) in Equation (2.8), d_f is the feedforward projection dimensionality, $W_i \in \mathbb{R}^{d_f \times d}$ and $b_i \in \mathbb{R}^{d_f}$ are learned input parameters, and $W_j \in \mathbb{R}^{d \times d_f}$ and $b_j \in \mathbb{R}^d$ are learned output parameters.

The encoder computes *self-attention* to compute a contextual representation of the input as a latent representation. The decoder computes *masked self-attention*, similar to the encoder but without the ability to 'attend' to future states, and *cross-attention*, to compute an interaction between the current outputs and the latent representation of the inputs. In general, we use the Transformer architecture unmodified from the original definition. However, in Chapter 3 we define an augmentation on Equation (2.4) for multi-head attention over multiple parallel encodings.

Model Definition and Pre-trained Initialisation The model is constructed as an encoder-decoder Transformer network illustrated in Figure 2.2. The encoder comprises 12 layers and the decoder comprises 6 layers. The model uses a dimensionality of $d = 1024$ and 16 attention heads per layer resulting in head dimensionality $d_h = 64$. Feedforward projection dimensionality is set to $d_f = 4096$. The embedding dimensionality is also set to $d = 1024$. We further discuss the vocabularies for each dataset in Section 2.3.

In this thesis, the model design is largely informed by a pre-trained encoder. Pre-training on large corpora allows a model to learn more general representations of language (Collobert et al., 2011). Contemporary pre-trained models are widely used to initialise the parameters of models fine-tuned for a specific task (Peters et al., 2018; Devlin et al., 2019). As discussed in Section 1.2, multilingual pre-training initialises a model with generalisable knowledge of language understanding in many languages. For our task, we use a pre-trained model to initialise the parser's multilingual language understanding capability. The encoder is initialised using the MBART50 pre-trained model (Tang et al., 2021). This is a sequence-to-sequence model pre-trained on multilingual corpora using an unsupervised language modelling objective in Liu et al. (2020), and further pre-trained using a machine translation task in Tang et al. (2021). MBART50 supports 50 languages including all target languages outlined in Section 2.1.1. Our encoder follows the twelve-layer architecture for MBART50. The encoder parameters, ϕ, are initialised using MBART50 and then further fine-tuned for semantic parsing.

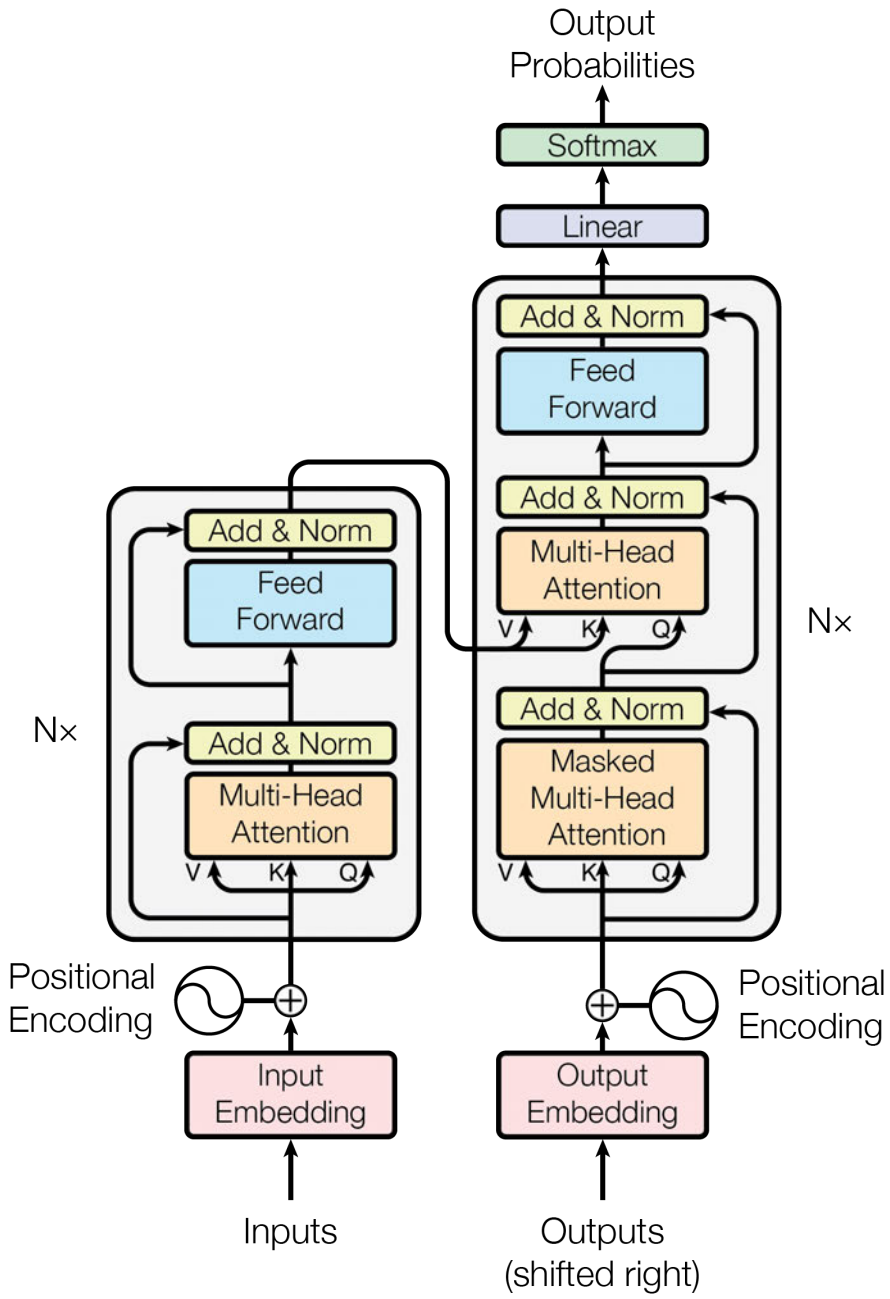

Figure 2.1: Illustration of the Transformers architecture adopted in this thesis from Vaswani et al. (2017). A single layer comprises multi-head attention and feedforward layers with additional skip connections. These layers are stacked to produce the complete model. Each multi-head attention layer receives input queries, keys, and values (Q, K, and V respectively), and outputs a contextual representation of V weighted by the dot-product interaction between Q and K. See arxiv.org/abs/1706.03762 for authors' permission for reproduction of graphics for academic purposes.

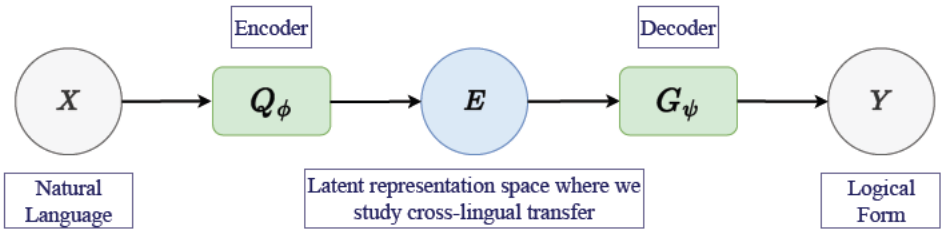

Figure 2.2: Diagram of the encoder-decoder parser mapping from natural language to logical form. All modelling contributions in Chapters 3 to 6 derive from this base model. We use the MBART50 pre-trained model (Tang et al., 2021) to initialise the encoder, Q_ϕ.

The input vocabulary also follows from MBART50 using the pre-trained 250,054 embeddings as the encoder vocabulary (i.e., $|V| = 250,054$).

In our experiments, we did not observe any benefit to also initialising the decoder with parameters from MBART50. We conjecture that this is attributed to MBART50 pre-training optimising the decoder to generate *natural* languages, whereas semantic parsing optimises a decoder to generate *machine-readable* languages. As a consequence, we train the decoder from a random initialisation in all our experiments. For similar reasons, we observed no benefit to combining encoder and decoder vocabularies into a single embedding matrix. Our final parser comprises approximately 208 million learnable parameters[2].

2.2.2 Optimising a Semantic Parser

The parser function is optimised to predict y from x by modelling the likelihood of y as Equation (2.2). The parser is trained by minimising the negative log-likelihood of true output y from x. The loss function, ℓ, for a single input-output example, (x, y), is Equation (2.9) where y_i is the index of the ground truth token in decoder vocabulary $|V'|$ and $p_f(y_i|x)$ is the probability of predicting y_i using function f with input x.

$$\ell(x,\ y) = -\sum_{i=0}^{|V'|} y_i \log p_f(y_i|x) \tag{2.9}$$

This objective is equivalent to minimising the cross-entropy between the predicted distribution, $\hat{y} = p_f(y|x)$, and true distribution, y when both variables are $\mathcal{R}^{1\times|V'|}$ shape. This loss is minimised using a minibatch gradient descent. In Chapter 4 and Chapter 5,

[2]Calculated as `total_params=sum(param.numel() for param in model.parameters())`

we define augmentations on Equation (2.9) to improve cross-lingual transfer during training.

Training Specification The parser is optimised using the loss function in Equation (2.9) oversampled data batches. We use the Adam optimiser (Kingma and Ba, 2014) for updating parameters with a learning rate of 1×10^{-4}, and $(0.99, 0.998)$ decay factors. We regularise learning using a model activation dropout factor of 10% (Srivastava et al., 2014; Gal and Ghahramani, 2016). We do not employ weight decay regularisation as this was observed to damage the meta-learning algorithm proposed in Chapter 5. We select the best model during training using the loss on a held-out validation dataset. We use only the English-language validation dataset for model selection. Multilingual validation data for model selection has been identified as useful for optimising cross-lingual transfer (Keung et al., 2020). However, we consider this scenario to invalidate 'zero-shot' cross-lingual transfer results, which should not evaluate generalisation to new languages before inference. This constraint is required for our zero-shot parser proposed in Chapter 4. The batch size for each model is 16 input-output examples unless listed otherwise. All models in this thesis can be trained on a single NVIDIA A100 80GB GPU in under 36 hours.

Inference Specification The optimised parser infers an output logical form from a test input by autoregressively generating tokens in the process outlined by Equation (2.3). We follow prior practice (Dong and Lapata, 2016) in generating LFs via beam search using a beam width of five hypotheses. The beam search decoder stores five 'hypotheses' for possible output sequences during incremental generation. The final output is the hypothesis with the highest likelihood.

2.3 Data for Cross-lingual Semantic Parsing

2.3.1 MultiATIS++SQL

The original dataset contains 5,418 utterances requesting US flight and airport information. Hemphill et al. (1990) and Dahl et al. (1994) produced the original ATIS corpus pairing English utterances with annotated SQL queries to answer respective questions using a relational database. The original SQL queries were complex with a high quantity of nested subqueries (Finegan-Dollak et al., 2018) resulting in an increased challenge

	Train	Validation	Test	Total
MultiATIS++SQL				
English (EN)	4,473	497	448	5,418
French (FR)	4,473	497	448	5,418
Portuguese (PT)	4,473	497	448	5,418
Spanish (ES)	4,473	497	448	5,418
German (DE)	4,473	497	448	5,418
Chinese (ZH)	4,473	497	448	5,418
MTOP				
English (EN)	15,602	2,229	4,457	22,288
French (FR)	11,609	1,658	3,317	16,584
Spanish (ES)	10,821	1,546	3,092	15,459
German (DE)	13,152	1,879	3,757	18,788
Hindi (HI)	11,292	1,613	3,226	16,131
Thai (TH)	11,141	1,592	3,182	15,915

Table 2.4: Dataset sizes per partition and language for MultiATIS++SQL (Section 2.3.1) and MTOP (Li et al., 2021). MultiATIS++SQL is fully parallel with all examples translated into each language. MTOP is partially parallel where all examples are sourced from English but a non-exclusive subset is translated into each target language.

for a parser. Iyer et al. (2017) proposed simplified SQL queries to improve parser accuracy by generating "easier" outputs. Xu et al. (2020) created MultiATIS++SQL by professionally translating the English queries into French, Portuguese, Spanish, German and Mandarin Simplified Chinese for the spoken language understanding version of the ATIS task.[3]

Dataset Creation We create a new version of ATIS for our objective of executable cross-lingual semantic parsing titled **MultiATIS++SQL**. This variant of ATIS pairs the simplified executable SQL queries from Iyer et al. (2017) with utterances in six languages from Xu et al. (2020). This creates a multilingual dataset with six parallel

[3]This is an intent-classification and slot-filling variant of ATIS for spoken language understanding. This form is not executable, similar to TOP-LF, and we do not consider this version of ATIS in this thesis. We also note that Xu et al. (2020) translated ATIS queries into Japanese but the mapping between utterances from English to Japanese is not publicly available. We ignore this language for this reason.

MultiATIS++SQL	
EN	I'd like the cheapest one way fare from Boston to San Francisco.
FR	J'aimerais trouver le tarif aller simple le moins cher de Boston à San Francisco.
PT	Eu gostaria da tarifa mais barata de ida de Boston para São Francisco.
ES	Quisiera la tarifa más barata de ida desde Boston hasta San Francisco.
DE	Ich suche den günstigsten Preis für einen Einzel flug von Boston nach San Francisco.
ZH	请帮我查找从波士顿到旧金山最低的单程票价

```
SQL  SELECT
       DISTINCT fare_1.fare_id
     FROM
       fare fare_1, flight_fare flight_fare_1, flight flight_1, airport_service
       airport_service_1, city city_1, airport_service airport_service_2, city city_2
     WHERE
       fare_1.one_direction_cost =(
         SELECT
           MIN( fare_1.one_direction_cost)
         FROM
           fare fare_1, flight_fare flight_fare_1, flight flight_1,
           airport_service airport_service_1, city city_1,
           airport_service airport_service_2, city city_2
         WHERE
           fare_1.round_trip_required = 'NO'
           AND fare_1.fare_id = flight_fare_1.fare_id
           AND flight_fare_1.flight_id = flight_1.flight_id
           AND flight_1.from_airport = airport_service_1.airport_code
           AND airport_service_1.city_code = city_1.city_code
           AND city_1.city_name = 'BOSTON'
           AND flight_1.to_airport = airport_service_2.airport_code
           AND airport_service_2.city_code = city_2.city_code
           AND city_2.city_name = 'SAN FRANCISCO'
       )
       AND fare_1.fare_id = flight_fare_1.fare_id
       AND flight_fare_1.flight_id = flight_1.flight_id
       AND flight_1.from_airport = airport_service_1.airport_code
       AND airport_service_1.city_code = city_1.city_code
       AND city_1.city_name = 'BOSTON'
       AND flight_1.to_airport = airport_service_2.airport_code
       AND airport_service_2.city_code = city_2.city_code
       AND city_2.city_name = 'SAN FRANCISCO';
```

Table 2.5: An example of MultiATIS++SQL data in English (EN), French (FR), Portugeuse (PT), Spanish (ES), German (DE), and Simplified Chinese (ZH). The input utterance is natural language and the output is SQL referencing an English-language relational database. All inputs map to the same SQL output for this dataset.

input utterances paired with an SQL logical form. A parser must generate syntactically and semantically correct SQL (using entities and table features in English) from any input language. Table 2.5 shows an example from MultiATIS++SQL where all utterances correspond to the same SQL logical form. The complete data split is outlined in Table 2.4 using the partitioning from Kwiatkowski et al. (2011). We note that this contribution was originally reported in Sherborne and Lapata (2022).

Data Preprocessing Input utterances are tokenised using the SentencePiece tokeniser (Kudo and Richardson, 2018). This processes inputs to match the vocabulary and pre-trained embeddings from MBART50. The vocabulary of a SQL query comprises the language keywords (e.g., `SELECT` or `INNER JOIN`) and attributes from the respective database, including table and column names (e.g., `airport_service_1.airport_code`). This vocabulary is *closed* i.e., a fixed-size lexicon entirely derived from the grounding environment. We tokenise output LFs using whitespace. We use whitespace tokenisation as the meaning of each token in an LF is already atomic. For example, the meaning of '`SELECT`' is defined for the SQL language and we consider subword tokenisation, e.g., to form tokens '`SEL`, '`ECT`", as redundant. This formatting approach may result in the model struggling to parse entities unseen during training without a pre-trained vocabulary. However, the set of possible entities in this single-domain dataset is small enough that we empirically do not observe entity unfamiliarity during inference to be a larger issue than the difficulty in adapting a pre-trained decoder to SQL syntax. We acknowledge this method could be suboptimal for multi- or open-domain datasets with more possible entities only seen during test. This difficulty is a contributor to why we use *sentinel tokens* for MTOP, explained below in Section 2.3.2. The decoder vocabulary comprises 593 tokens of SQL keywords and features from the ATIS database.

MultiATIS++SQL: Execution Accuracy As SQL logical forms are executable within an available relational database, we evaluate performance on MultiATIS++SQL using *Execution Accuracy*. This metric, also called denotation accuracy, evaluates if the predicted SQL retrieves the same answer from the database as the label logical form. For test input x with label LF y and predicted LF $\hat{y}$, we evaluate the execution accuracy using grounding environment, e, as Equation (2.10).

$$e(\hat{y}) == e(y) \qquad \text{Denotation accuracy for prediction } \hat{y} \tag{2.10}$$

MTOP	
EN	Can you tell me how much rain is expected tonight?
FR	Peux-tu me dire quelle quantité de pluie est prévue ce soir?
ES	¿Puedes decirme cuánta lluvia se espera para esta noche?
DE	Kannst du mir sagen, wie viel Regen heute Nacht erwartet wird?
HI	क्या तुम मुझे बताओगे कि आज रात कितनी बारिश होने वाली है?
TH	ช่วยบอกทีว่าคืนนี.คาดว่าฝนจะตกหนักแค่ไหน?
LF (EN)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE rain][SL:DATE_TIME tonight]]`
LF (FR)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE pluie][SL:DATE_TIME ce soir]]`
LF (ES)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE lluvia][SL:DATE_TIME para esta noche]]`
LF (DE)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE Regen][SL:DATE_TIME heute Nacht]]`
LF (HI)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE बारिश][SL:DATE_TIME आज रात]]`
LF (TH)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE ฝน][SL:DATE_TIME คืนนี้]]`

Table 2.6: An example of MTOP data in English (EN), French (FR), Spanish (ES), German (DE), Hindi (HI), and Thai (TH) from Li et al. (2021). The input utterance is natural language and the output is a logical form designed for spoken language understanding pipelines. The relevant function and action is executed dependent upon the parsed intent (`IT:` label) slots (`SL:` label). Unlike MultiATIS++SQL, the MTOP LF contains tokens from the input as slots for downstream actions. Therefore, we consider these outputs non-identical but *semantically* equivalent in our experiments. Section 2.3.2 outlines our preprocessing steps for MTOP.

Execution accuracy is useful in evaluating if the predicted LF has the same semantics as the gold LF i.e., if the prediction *means* the same concept. This metric implicitly will also evaluate for well-formed SQL, as an ill-formed query will not return the correct result. We also do not consider if the LFs are identical with this metric. This can risk introducing spurious predictions in generated LFs as we do not evaluate if the correct answer is retrieved through incorrect logic. However, spuriousness has been observed as a marginal issue unless training a weakly supervised semantic parser without logical forms (Lee et al., 2023). As our system is fully supervised, we do not explicitly model this concern.

2.3.2 MTOP

This dialogue understanding dataset from Li et al. (2021) comprises 22,288 questions in English mapping to the TOP-LF form described in Section 2.1.1. Generating the hierarchical TOP-LF output requires a sequence-to-sequence solution as the LF is tree-structured rather than a label per token. MTOP is a translation of TOP (Gupta et al., 2018) into French, Spanish, German, Hindi and Thai. Table 2.6 shows a single example from the MTOP dataset where the logical form output uses tokens from the input utterance. We consider the outputs from parallel utterances as *semantically equivalent*, but note they are typically not identical. The full dataset split is outlined in Table 2.4, highlighting that different proportions of the original English queries are translated into each language.

Data Preprocessing We use the same SentencePiece tokenisation for input utterances for MTOP to match the pre-trained encoder vocabulary. Output LF tokenisation uses whitespace with additional splitting on brackets ('`[`' or '`]`') and colons to demarcate intent and slot names i.e., '`IN:GET_WEATHER`' split to '`IN:`' and '`GET_WEATHER`'.

MTOP differs from MultiATIS++SQL due to the presence of input tokens in the output logical form (e.g., "rain" and "pluie" in Table 2.6). Implementing a parser with an *open* vocabulary requires additional engineering effort beyond the parser outlined in Section 2.2.1. For our cross-lingual case, we observe modelling solutions for an open vocabulary parser to incur weak generalisation during inference. Exploring a copy mechanism (Jia and Liang, 2016), we found that the system struggled to copy the correct tokens when a target language was not observed during training. Additionally, we found that combining the embedding spaces (i.e., V from domain $\mathcal{X}$ and V' from domain $\mathcal{Y}$) led to poor generalisation as the parser will hallucinate incorrect natural language tokens in the output LF. As an alternative, we pre-process MTOP into a closed vocabulary format to limit hallucination during decoding. To produce a closed vocabulary, we use *sentinel word* preprocessing from spoken language understanding (Raman et al., 2022). Sentinel preprocessing augments the utterance input with a label word per token, and replaces this token in the LF with the label. Table 2.7 shows an example of this relabelling schema. This pre-processing removes utterance tokens from LF outputs for a total decoder vocabulary of 206 tokens.

MTOP: Space and Case Invariant Exact Match Accuracy We cannot evaluate MTOP using execution accuracy as TOP-LF outputs are not executable. Rosenbaum

MTOP (Sentinel)	
EN	Can you tell me how much rain is expected tonight?
FR	Peux-tu me dire quelle quantité de pluie est prévue ce soir?
EN Sentinel	`word0` Can `word1` you `word2` tell `word3` me `word4` how `word5` much `word6` rain `word7` is `word8` expected `word9` tonight?
FR Sentinel	`word0` Peux-tu `word1` me `word2` dire `word3` quelle `word4` quantit'e `word5` de `word6` pluie `word7` est `word8` pr'evue `word9` ce `word10` soir?
LF (EN)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE rain][SL:DATE_TIME tonight]]`
LF (FR)	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE pluie][SL:DATE_TIME ce soir]]`
LF (EN) Sentinel	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE word6][SL:DATE_TIME word9]]`
LF (FR) Sentinel	`[IN:GET_WEATHER[SL:WEATHER_ATTRIBUTE word6][SL:DATE_TIME word9 word10]]`

Table 2.7: Example of sentinel word preprocessing (Raman et al., 2022) to produce a closed decoder vocabular for MTOP Li et al. (2021). Each utterance token is given a label '`wordi`' for token index '`i`'. The respective token is replaced in the LF by string matching.

et al. (2022) propose *Space and Case Invariant Exact-Match* (SCIEM) accuracy for evaluating TOP-LF outputs. This evaluates if the generated LF matches the gold LF by comparing output tokens. SCIEM accuracy preprocesses an output LF by normalising spacing and casing in outputs; evaluating the processed output against the preprocessed gold logical form. For SCIEM normalising function, SCIEM, this metric is computed as Equation (2.11).

$$\text{SCIEM}(\hat{y}) == \text{SCIEM}(y) \qquad \text{SCIEM accuracy for prediction } \hat{y} \tag{2.11}$$

SCIEM accuracy does not evaluate if the predicted LF has an equivalent meaning to the gold LF. Evaluating the tokens of the LF is a more literal metric testing if the logical form is 'correct', and is sensitive to spuriousness unlike Execution Accuracy. Both forms of evaluation are complementary within our experiments to verify that generated logical forms are accurate on a surface-form and semantic level.

2.4 Cross-lingual Transfer as Generalisation

2.4.1 Sampling Distributions for Data

We define some true distribution over the data for our task as $\mathcal{D}$. Sampling from this distribution in Equation (2.12) yields paired data examples of English language input utterances, x_{EN}, and output logical forms y. We describe the expected loss for the true distribution under function f, with parameters θ, as Equation (2.13) where the loss, ℓ is the negative log-likelihood function defined in Equation (2.9).

$$(x_{\text{EN}},\ y) \sim \mathcal{D} \tag{2.12}$$

$$\ell_{\mathcal{D}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x,\ y)] \tag{2.13}$$

$\ell_{\mathcal{D}}$ defines the upper-bound performance of the parser as the distribution p_f when observing the true $\mathcal{D}$. Note we consider minimum loss and maximum parsing accuracy as dual characteristics of f. However, $\mathcal{D}$ is not observable as we do not have access to infinite data to define the entire distribution. Therefore, we approximate $\ell_{\mathcal{D}}$ using an empirical estimate, $\ell_{\mathcal{S}}(\theta)$, as Equation (2.14). The empirical estimate ℓ is learned from data sample $\mathcal{S}$ comprising n examples, $\{(x_i, y_i)\}_{i=1}^{n}$ i.e., the training data.

$$\ell_{\mathcal{S}}(\theta) = \frac{1}{n}\sum_{\mathcal{S}\sim\mathcal{D}} \ell(x,\ y) \tag{2.14}$$

We aim to learn some parameter setting, θ, minimising the error between $L_{\mathcal{D}}$ and $L_{\mathcal{S}}$. This is defined as the generalisation error in Equation (2.15).

$$\Delta_g = \ell_{\mathcal{D}}(\theta) - \ell_{\mathcal{S}}(\theta) \tag{2.15}$$

In typical machine learning scenarios, minimising Δ_g encompasses generalising from training data to held-out test data sampled from the same true distribution. The error Δ_g describes the *in-distribution* generalisation goal if sample $\mathcal{S}$ adequately approximates true distribution $\mathcal{D}$.

In cross-lingual terminology, we can only sample from the true distribution for English ($\mathcal{D}$) but desire to generalise to additional languages such as French ($\mathcal{D}_{FR}$) or Hindi ($\mathcal{D}_{HI}$). We generalise these distributions to a true distribution for any target language, l, as Equation (2.16). Here we produce sample $\mathcal{S}_l$ from language l comprising

N pairs of input-output examples. We define $\mathcal{D}^*$ as sampled from the set of true distributions for L target languages in Equation (2.17). The objective in cross-lingual transfer is for a model trained on the empirical estimate $\ell_{\mathcal{S}}(\theta)$ to generalise to any true target language distribution. The cross-lingual transfer objective minimises the *cross-lingual generalisation error*, or *cross-lingual transfer gap*[4] in Equation (2.18). More generally, this is an *out-of-distribution* generalisation objective.

$$\mathcal{S}_l = \{(x_1, y)\}_{i=1}^{N} \sim \mathcal{D}_l \tag{2.16}$$

$$\mathcal{D}^* \sim \{\mathcal{D}_1, \ldots, \mathcal{D}_L\},\ l = \{1, \ldots, L\} \tag{2.17}$$

$$\Delta_x = \ell_{\mathcal{D}^*}(\theta) - \ell_{\mathcal{S}}(\theta) \tag{2.18}$$

We assume that our target language evaluation data for each dataset is an accurate held-out sample of $\mathcal{D}^*$. The contributions in this thesis consider different strategies for sampling $\mathcal{S}$ to generalise to $\mathcal{D}^*$. The case study in Section 1.4 can be considered an exercise in building $\mathcal{S}$ for this generalisation objective. We now define the three core scenarios for sampling $\mathcal{S}$ in this thesis.

Generalisation from Silver-standard Data uses an intermediate data generator, e.g., machine translation, to simulate language-specific distributions. In the case study context, this approach assumes sampling $\mathcal{D}_l$ for training data is too expensive. Chapter 3 proposes an economical strategy to maximise parsing accuracy using an intermediary machine translation system for synthetic data. We first examine a TRANSLATE TEST setup: using machine translation for test data by translating this data into English and predicting outputs using a model trained on English data. We then examine a TRANSLATE TRAIN setup: using machine translation to approximate $\mathcal{D}_l$ without sampling the true distribution. Both approaches consider building a parser to generalise to $\mathcal{D}_l$ using synthetic samples.

Zero-Shot Generalisation proposes to generalise to $\mathcal{D}_l$ without using synthetic distributions (as above), and without sampling $\mathcal{D}_l$. The training sample $\mathcal{S}$ contains no samples from the $\mathcal{D}_l$ distribution for any target language l. In Chapter 4, we propose a zero-shot parser by sampling distributions for *alternative* tasks with available data. This is permitted as $\mathcal{D}_l$ is a task-specific distribution of paired data for semantic parsing only.

[4]The corollary is the *cross-lingual transfer penalty* as the difference in performance (accuracy) between source and target language.

Few-Shot Generalisation is a more relaxed constraint that *few* samples from $\mathcal{D}_l$ are permitted for cross-lingual generalisation during training. We sample $\mathcal{D}_l$ for a small set of labelled examples, $\mathcal{S}_l \sim \mathcal{D}_l$, to augment the English-language data during training. Few-shot transfer considers how few samples are required to minimise the cross-lingual transfer gap in Equation (2.18). We propose few-shot generalisation strategies in Chapter 5 and Chapter 6 for the case study objective.

Measuring Cross-lingual Transfer The aim of this thesis is not to achieve 100.0% accuracy on all datasets. The error in Equation (2.18) outlines that the aim is performance parity *between* languages such that a system performs equivalently regardless of utterance language. For example, a system reporting 77.4% accuracy for English should achieve 77.4% accuracy for German. Ideally, a system would report greater accuracy from learning mutually useful features for multiple languages. Contributions solely to improve this 77.4% accuracy to 100.0% in English are orthogonal to our contributions. The upper bounds in our experiments are trained on each language, and a system trained on all data in all languages representing the expensive ideal scenario where 100% target language translation is possible. We propose methods to reach, or surpass, these upper bounds with minimal data. We report an average across target languages (TARGET AVG.) as the key metric reporting average improvement in cross-lingual transfer. To measure the improvement between models for this objective, we report statistical significance where possible. We use the Wilcoxon ranked sign test (Wilcoxon, 1945) to evaluate independent samples across test datasets with a significance threshold of $p < 0.01$.

2.4.2 Cross-lingual Representation Alignment

The cross-lingual generalisation error defined in Equation (2.18) provides a simplified metric for evaluating success. However, this scalar does not introspect whether the parser is encoding and organising the language understanding capability similarly for different languages. We argue that variations in a natural language's surface forms should express the same underlying semantics. We desire to analyse and inspect if this multilingual semantic similarity is being adequately encoded. To yield this insight, we study the cross-lingual transfer objective as a *cross-lingual representation alignment* problem.

Consider the semantics of different parallel inputs mapping to an equivalent output, y.

Sequence-to-sequence modelling delegates the distinct responsibilities of understanding x and generating y to the encoder and decoder respectively. Therefore, the encoder maps different x to the same latent representation to guarantee generating the same y regardless of natural language. Consider parallel encodings E_{EN} and E_l from parallel input sentences in Equation (2.19) and Equation (2.20) respectively. A decoder, G_ψ, trained only with English data is *likely* to correctly interpret E_{EN} to generate an accurate prediction $\hat{y}_{EN}$ in Equation (2.21). If E_{EN} and E_l are distant and the decoder has not observed this representation, we argue the decoder is *unlikely* to correctly map E_l to an accurate prediction $\hat{y}_l$ equivalent to $\hat{y}_{EN}$.

$$E_{EN} = Q_\phi(x_{EN}) \quad \text{Latent encoding from input in English} \tag{2.19}$$

$$E_l = Q_\phi(x_l) \quad \text{Latent encoding from input in } l \tag{2.20}$$

$$\hat{y}_{EN} = G_\psi(E_{EN}) \quad \text{Decode encoding of English to logical form} \tag{2.21}$$

$$\hat{y}_l = G_\psi(E_l) \quad \text{Decode encoding of } l \text{ to logical form} \tag{2.22}$$

$$x_{EN} \neq x_l \quad \text{Equivalent semantics in languages English and } l \tag{2.23}$$

$$Q_\phi(x_{EN}) \approx Q_\phi(x_l) \quad \text{Approximately similar latent encodings} \tag{2.24}$$

$$\hat{y}_{EN} = \hat{y}_l \quad \text{Parse to the same logical form} \tag{2.25}$$

As the decoder is simply interpreting the given latent representation, we consider this as an encoding challenge: desiring $E_l \approx E_{EN}$ by optimising Q_ϕ i.e., Equation (2.24). The ideal case is *language agnostic* representations from Q_ϕ for equivalent outputs, y, from encodings in any language. We assume this subsequently minimises the cross-lingual generalisation error, and validate this assumption through analysis in each chapter. We study the latent representation space from each system by computing a representation for each input sequence, x, by pooling the token-level encodings in Equation (2.26).

$$q_x = \frac{1}{T} \sum_{t \in T} Q_\phi(x)_t \tag{2.26}$$

We use t-SNE (van der Maaten and Hinton, 2008) to visualise the relationships between encodings from different languages. We note that the distances between clusters in t-SNE are generally not meaningful, but the separability of clusters provides valuable insight into cross-lingual semantic similarity. We complement this visualisation with quantitative analysis on the high-dimensional q_x representations before t-SNE dimensionality reduction. We consider two measurements of representation similarity for our

analysis. First, we evaluate the average **cosine similarity** between representations of parallel utterances. This provides an absolute measurement of similarity where zero represents no similarity (perpendicular in vector space) and one represents identical representations. For L target languages, we measure cosine similarity between each representation and the source English representation and average over L measurements. This is averaged over the test set for a single metric. We also report average **Top-k ranked similarity** as a measurement of relative similarity between representations in the latent space. For a representation q_x from language l, we rank all other representations from L target languages and English by cosine similarity. We then evaluate if a parallel utterance in *any* other language ($\neg l$) is within the top k most similar representations. This evaluates if the representations of similar meaning are closer in latent space than a representation of arbitrarily different natural language. As we expect paraphrases and potential duplicates in datasets, we consider Top-1, Top-5, and Top-10 (i.e., $k = \{1, 5, 10\}$) ranked similarity to allow for semantic equivalence in k sized clusters.

2.5 Summary

This chapter outlines the objectives of the thesis and details the relevant data and modelling resources we use in our contributions. Our contributions build on the Transformer model defined in Section 2.2 for the cross-lingual representation alignment objective defined in Section 2.4. Chapters 3 to 6 now details the contributions of this thesis using the measurement of success in cross-lingual transfer, as defined in Section 2.4.2, to evaluate our central hypothesis concerning representation alignment for cross-lingual semantic parsing.

Chapter 3

The Role of Machine Translation in Cross-lingual Transfer

In this chapter, we investigate if automatic machine translation (MT) is a viable cross-lingual transfer tool for our case study. In an ideal scenario, we would require no additional annotation or effort to produce a system capable of parsing all target languages. Kann (2023) propose that perfect machine translation would eliminate the need for modelling cross-lingual transfer. Exploiting MT offers multiple benefits addressing our overarching hypothesis. Translation resources are available between many language pairs (Koehn et al., 2022) and can be sourced either for free (e.g., the open source OPUS system (Tiedemann and Thottingal, 2020)), or at a low cost per translation (e.g., Google Translate (Wu et al., 2016)). Furthermore, perfect MT would offer a rapid, economical strategy to parse any target language equivalently to our upper bounds. Before we consider any novel modelling (i.e., in Chapters 4 to 6), we first consider if machine translation is the simplest and fastest route to cross-lingual semantic parsing.

We frame the translation pipeline as an intermediary tool to simulate a natural language using 'silver-standard' synthetic samples from a different natural language. If this simulation is accurate, training or inference using MT will perform comparably to the equivalent model using gold-standard translated data ('gold data'). Machine translation has proved an adequate language simulation for classification tasks (Conneau et al., 2018a), but can struggle to represent fine-grained semantics across languages (Artetxe et al., 2020). For our task, any MT system must generate fluent utterances in each target language to accurately imitate how a native speaker would compose a question. A machine translation must also be faithful to the original utterance, with equivalent underlying semantics, to map to an equivalent logical form (LF). MT

without fluent and faithful translations introduces errors leading to weaker cross-lingual generalisation.

First, we consider using translation during inference to predict an LF using the existing model for English i.e., the TRANSLATE TEST pipeline (Conneau et al., 2018a). A target language utterance is translated from the respective target language into English, and this synthetic English utterance is input to a parser trained on the source English dataset. Using TRANSLATE TEST enables parsing of target languages without developing any additional parsing models as the parser for English utterances already exists. TRANSLATE TEST is a widely successful strategy for cross-lingual classification tasks (Artetxe et al., 2023), however, semantic parsing is additionally challenging by requiring accurate representations of entities, reasoning about relationships, and inferring context. This challenge may require direct modelling of target languages for accurate parsing.

Second, we consider the inverse case where we translate during training to build a parser using pairs of machine-translated utterances and logical forms i.e., the TRANSLATE TRAIN pipeline (Conneau et al., 2018a). The source English dataset is translated into each target language, and a parser model is trained on these synthetic target language utterances paired with equivalent logical forms. This parser is used to predict logical forms from gold target language utterances during inference. This pipeline requires training more models than TRANSLATE TEST but can be further optimised to improve the approximation of each target language using data augmentation and multilingual modelling (Singh et al., 2019a; Edunov et al., 2020).

In this chapter, we consider the hypothesis that **machine translation adequately approximates natural language for cross-lingual transfer in semantic parsing**. We examine TRANSLATE TEST or TRANSLATE TRAIN for our task in the context of advancing progress in neural machine translation (NLLB Team et al., 2022). Using each pipeline, we present an analysis of how MT approximates a native speaker (including variation in question structure, formality, and tone). Owing to the fine-grained semantic transfer required of our task, we can precisely analyse and identify where MT is, and is not, suitable for cross-lingual semantic parsing. We observe that individual MT systems produce different styles of output translation varying in fluency and faithfulness. This limits the performance ceiling for TRANSLATE TEST, which cannot model the style of synthetic English utterances by design (Riley et al., 2020). For TRANSLATE TRAIN, we can mitigate the error from any one system by querying multiple machine translation systems to produce target language paraphrases as data augmentation. We

propose to further improve the TRANSLATE TRAIN pipeline by modelling multiple MT systems as parallel language models ensembled into a single parser model. This system: FATES: **F**used **At**tention **T**ransformer **E**nsembling for **S**emantic Parsing uses multiple encoders to model each MT system output independently, and then fuses these 'perspectives' on a target language to predict a logical form. Our experimental results on MULTIATIS++SQL and MTOP validate that paraphrase-based data augmentation and ensembled language modelling can maximise parsing when limited to silver-standard data.

3.1 Problem Formulation

This chapter examines the economical approach to building a semantic parser using machine translation. In the distribution sampling framework outlined in Section 2.4.1, our objective is to produce a parser capable of generalising to a target language distribution, $\mathcal{D}_l$, subject to the inability to sample from this distribution. As we cannot sample $\mathcal{D}_l$, we instead use a translation system to approximate a distribution with 'silver standard' synthetic data. Parser evaluation uses gold data from each language to evaluate generalisation to $\mathcal{D}_l$, observing samples from MT approximating $\mathcal{D}_l$. A parser with a lower generalisation error between MT and gold distributions is said to be *robust* to the transfer from synthetic to authentic samples of a language.

Machine translation is a widely used resource for generating translations of text in a desired language with MT systems ever improving and expanding to new languages (Koehn et al., 2022). Open-source MT models are often state-of-the-art on competitive benchmarks such as the Conference on Machine Translation (WMT) with successive years of WMT measuring progress in translation capability and evaluation. Within semantic parsing, Duong et al. (2017b) and Moradshahi et al. (2020) validate that machine translation forms a noisy, but reasonable, synthetic approximation to a target language.

3.1.1 Translation at Inference

We propose that the fastest and cheapest strategy for expanding supported languages in our case study is to translate utterances into English during inference. Inference time translation can be input to the existing semantic parser for English to predict a logical form i.e., the TRANSLATE TEST pipeline (Conneau et al., 2018a). This requires no

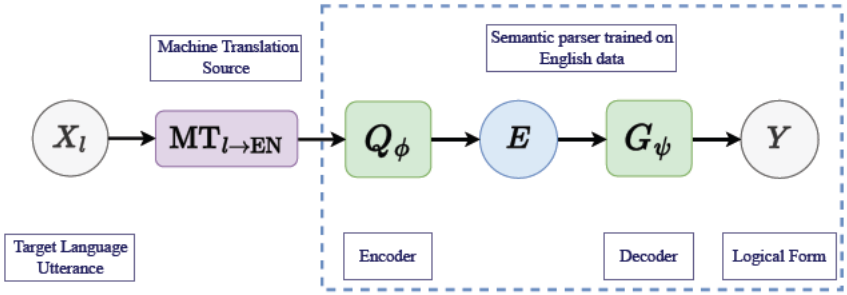

Figure 3.1: The typical TRANSLATE TEST pipeline for cross-lingual semantic parsing. A parser is trained on existing training data in English to produce a model capable of predicing logical forms from English utterances. The target language test utterance is input to a machine translation system to produce a translation in English. This translation is input to the parser to predict a logical from. While extremely cost effective, TRANSLATE TEST can be highly sensitive to translation quality.

new modelling effort and can be implemented rapidly using commercial or open-source machine translation.

Figure 3.1 describes the TRANSLATE TEST pipeline using a given MT system, $\text{MT}_{l\to\text{EN}}$, from target language l to English (EN). Equation (3.1) describes the cross-lingual generalisation challenge for TRANSLATE TEST wherein the translated utterance from target language l must approximate the equivalent English utterance. More generally, we require translated test data sample $\text{MT}_{l\to\text{EN}}(\mathcal{S}_l)$ to approximate the true distribution for English, $\mathcal{D}$. Assuming the MT output is of sufficient quality, we consider $\bar{x}$ as semantically equivalent to x, mapping to an equivalent y logical form.

$$\bar{x}_{\text{EN}} = \text{MT}_{l\to\text{EN}}(x_l) \approx x_{\text{EN}} \tag{3.1}$$

We consider this initial approach as the simplest strategy for our overarching hypothesis. We expect machine translation to mimic an upper bound of professional translation into target languages with the latent representation training with MT to follow suit. Furthermore, TRANSLATE TEST is the minimal effort baseline for all chapters in this thesis. However, in Chapter 2 raised that linguistic variation introduces challenges in representing semantics in different languages. In TRANSLATE TEST, the MT system must represent the meaning of a target language utterance equivalently for accurate inference. However, this initial parser has only observed data from native English speakers. While machine translation capability is ever increasing, the output of these systems is often considered a unique dialect of a language containing errors and idiosyncrasies

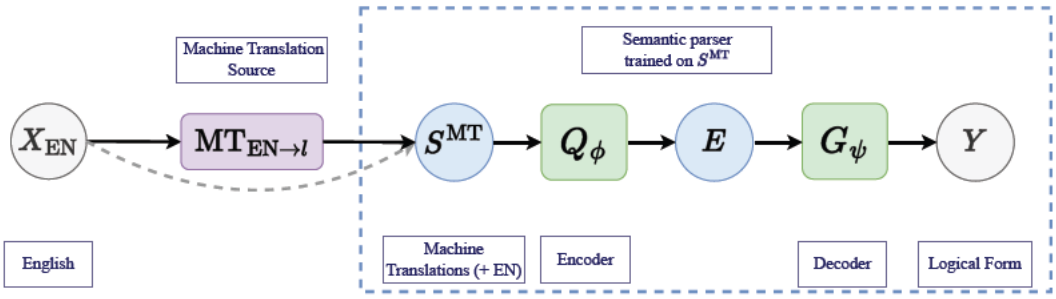

Figure 3.2: The typical TRANSLATE TRAIN pipeline for cross-lingual semantic parsing. Machine translation is used to generate synthetic utterances in a target language. A parser is trained on these synthetic utterances to generalise to gold test utterances. English data can also be added to the data sample for 'bilingual' modelling (dashed grey arrow).

unseen in data from native speakers i.e., the *translationese* dialect (Koppel and Ordan, 2011). Monolingual models can struggle to understand this dialect, and incorporating this dialect during training can be detrimental to generalisation (Riley et al., 2020). Therefore, we conjecture that our initial parser is likely to model translationese poorly when fine-grained semantic accuracy is required for accurate LF prediction. On this basis, we also explore alternatives using synthetic data during training.

3.1.2 Generating Training Data with Machine Translation

We now consider using machine translation to build training data in target languages i.e., the TRANSLATE TRAIN pipeline (Conneau et al., 2018a). TRANSLATE TEST evaluates if the parser for English can generalise to translationese from target languages. TRANSLATE TRAIN is the inverse approach where we evaluate a model trained with translationese data generalising to gold data.

Figure 3.2 describes the TRANSLATE TRAIN pipeline using a given MT system, $\text{MT}_{\text{EN}\to l}$, from English (EN) to target language l. Similar to above, Equation (3.2) describes the approximation goal for synthetic utterances to accurately represent the equivalent utterance from a native speaker. We make the same assumptions of quality in $\bar{x}$ mapping to equivalent y here as for TRANSLATE TEST above.

$$\bar{x}_l = \text{MT}_{\text{EN}\to l}(x_{\text{EN}}) \approx x_l \tag{3.2}$$

Given the MT system, $\text{MT}_{\text{EN}\to l}$, we pair n translated samples $\{\bar{x}_i\}_{i=1}^{m}$ with logical

forms from the source data, $\{y\}_{i=1}^{m}$, to create a synthetic training dataset, $\mathcal{S}_l^{\text{MT}}$, for language l in Equation (3.3). The data $\mathcal{S}_l^{\text{MT}}$ are synthetic samples from the true target language distribution via sampling from the output of the MT system. Equation (3.4) describes the ideal scenario for perfect machine translation when $\mathcal{S}_l^{\text{MT}}$ approximates an authentic sample $\mathcal{S}_l$.

$$\mathcal{S}_l^{\text{MT}} = \{\bar{x}_i,\ y\}_{i=1}^{n} \tag{3.3}$$

$$\mathcal{S}_l^{\text{MT}} \approx \mathcal{S}_l \tag{3.4}$$

First, we study if the parser trained on a singular S_l^{MT} generalises to data from the true distribution $\mathcal{D}_l$. We propose to use high-quality machine translation systems for synthetic training data in every target language. We select four machine translation systems across different sources and model sizes (discussed below in Section 3.2.2). Sherborne et al. (2020) originally considered commercial MT for this task, however, these outputs are not reproducible given the unknown implementation behind commercial models. For transparency and reproducibility, we consider only open-source neural machine translation systems in this thesis. We generally observe that open-source MT produces a comparable parser as our results in Section 3.3 are similar, or superior, to the original findings in Sherborne et al. (2020).

While TRANSLATE TRAIN requires more modelling effort to train a parser than TRANSLATE TEST, the approximation of the target language can be improved using augmentations described in Section 3.1.4 and Section 3.1.5 respectively. Additionally, directly modelling target languages allows a model to benefit from the diverse pre-training resources and shared linguistic features from multilingual modelling (Philippy et al., 2023).

3.1.3 Semantic Parsing with Machine Translation

We build a parser mapping from natural language x to logical form y as an encoder-decoder Transformer defined in Section 2.2.1. For the TRANSLATE TRAIN parser, the encoding process is expressed as Equation (3.5), and the decoding process is Equation (3.6).

$$\bar{E} = Q_\phi(\bar{x}_l) \qquad \text{Encode MT input into vector} \tag{3.5}$$

$$\hat{y} = G_\psi(\bar{E}) \qquad \text{Predict LF from vector encoding} \tag{3.6}$$

This process differs from Section 2.2.1 by explicitly labeling $\bar{x}_l$ as a translated sentence from Equation (3.2) in some language l. Optimisation uses the same cross-entropy loss for predicted $\hat{y}$ in Section 2.2.2. Following training, evaluation predicts a logical form from gold utterance x_l via Equations (3.7) and (3.8).

$$E = Q_\phi(x_l) \qquad \text{Predict encoding vector} \tag{3.7}$$

$$\hat{y} = G_\psi(E) \qquad \text{Predict logical form} \tag{3.8}$$

Successful machine translation can build an accurate and generalisable semantic parser capable of mapping utterances from native speakers to logical forms. However, even modern, state-of-the-art MT systems are prone to errors which can damage parsing performance. Errors in MT can result in unnatural phrasing (i.e., unnatural translationese) or an erroneous mapping to an inaccurate logical form (discussed in Section 3.3.7). The consequential effect is a poorer parser failing to generalise. To mitigate this issue, we now discuss a data augmentation approach combining outputs from multiple MT systems.

3.1.4 Ensembling Data with Machine Translation

In Section 3.1.2, we describe the TRANSLATE TRAIN pipeline using a single MT source to simulate target languages. This singular source will generate a translationese version of each target language specific to the MT system and respective training data, architecture, or generation algorithm. We hypothesise that we can improve our approximate of a target language by combining *multiple machine translation sources* into a single training dataset. Singh et al. (2019a) describe this combination as *paraphrasing* data augmentation useful for cross-lingual entailment classification and question answering. We expect that each MT system will produce variants of the meaning of the source utterance in different surface forms. We propose that diverse surface forms will be useful for training a parser with silver-standard data. This parser will be more capable of accurate inference on gold data, as the model has observed more linguistic variation with equivalent semantics to the logical form. Figure 3.3 outlines the paraphrase data augmentation within the TRANSLATE TRAIN pipeline.

Formally, we hypothesise that the combination of samples across K approximate distributions (Equation (3.9)) better approximates the true sample (Equation (3.10)) than any singular sample $\mathcal{S}_l^{\text{MT}}$. We discuss the empirical setting of K as part of our results in Section 3.3.

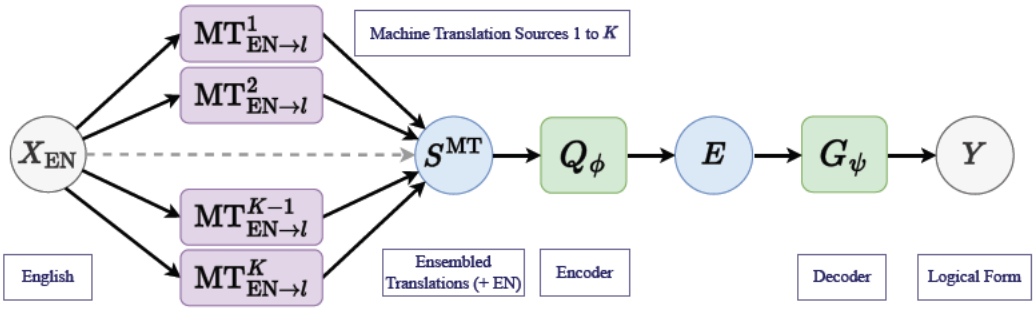

Figure 3.3: Paraphrase augmentation for more robust TRANSLATE TRAIN modelling. The original English utterance is input to K machine translation (MT) engines to produce K paired data examples in each target language. These samples from each MT source are combined into a single dataset to train an encoder-decoder parser. The source English paired data can also be added to the pooled data (dashed grey arrow). When $K = 1$, this is typical TRANSLATE TRAIN modelling, and we identify that increasing K improves cross-lingual transfer by introducing more diversity in the training data.

$$\mathcal{S}_l^{\text{MT}-K} = \{\mathcal{S}_l^{\text{MT}-1}, \mathcal{S}_l^{\text{MT}-2}, \ldots, \mathcal{S}_l^{\text{MT}-K}\} \tag{3.9}$$

$$\lim_{K\to\infty} \mathcal{S}_l^{\text{MT}-K} \approx \mathcal{S}_l \tag{3.10}$$

Table 3.1 illustrates the benefit of this *paraphrasing* augmentation for transfer from English to French. For the MultiATIS++SQL translation, no MT system exactly matches the professional translation, but all MT systems provide different approximations with different features. Systems vary in request phrasing using different politeness, 'J'aimerais [I would like]' or 'Je voudrais [I want]', and contrast between requesting a price ('le prix [the price]') or the ticket itself ('le billet [the ticket]'). For the MTOP translation, different systems choose imperative ('Obtenez [obtain]' or 'Recevez [receive]') or declarative ('Je reçois [I get]' or 'Je veux [I want]') structure. The vocabulary for 'texts' also varies between more formal 'les messages [the messages]' and informal 'mes textos [my texts]'. We argue this variation reflects natural variation in phrasing from native speakers. Paraphrasing in semantic parsing has been successful for monolingual models (Berant and Liang, 2014; Dong et al., 2017a; Iyer et al., 2017; Su and Yan, 2017) and in multilingual scenarios in similar tasks (Ganitkevitch and Callison-Burch, 2014; Mallinson et al., 2017; Dong et al., 2017b). We consider if singular or paraphrase-augmented MT systems can better approximate a target language without gold data.

EN	I'd like the cheapest one way fare from Boston to San Francisco
FR	J'aimerais trouver le tarif aller simple le moins cher de Boston à San Francisco.
OPUS	J'aimerais le prix le moins cher de Boston à San Francisco.
M2M1001.2B	J'aimerais le prix le moins cher à partir de Boston à San Francisco.
NLLB3.3B	Je voudrais le billet aller simple le moins cher de Boston à San Francisco.
NLLB1.3B	Je voudrais le billet le moins cher de Boston à San Francisco.
EN	Get my texts from Jeannine and Josh.
FR	Obtenez mes textos de Jeannine et Josh.
OPUS	Obtenez mes textos de Jeannine et Josh.
M2M1001.2B	Recevez mes textes de Jeannine et Josh.
NLLB3.3B	Je reçois les messages de Jeannine et Josh.
NLLB1.3B	Je veux mes messages de Jeannine et Josh.

Table 3.1: Examples from ATIS (upper) and MTOP (lower) from source annotators (English, EN), professional translators (French, FR), and four machine translation systems. We propose to use the variation in output from different MT systems for data augmentation to train a semantic parser. The intuition is that increasing the variation in query structure will improve the approximation of target languages and, therefore, improve parsing of languages without gold-standard training data.

3.1.5 FATES: Ensembling Parallel Encoders

The encoder, E_ϕ, within the parser, is responsible for mapping from natural language inputs to contextual vector representations in some latent space of fixed dimensionality. Considering input x_{EN} or translated input $\bar{x}_l$ map to the equivalent y, we approximately describe the encoder as responsible for the *natural language understanding* within the model. Complementing this, the decoder is responsible for the *formal language generation*. Informally, the encoder parameters ϕ control the model 'perspective' on language understanding, informed by training data S_l^{MT} for target language l and each MT system. Considering the variation in parallel MT inputs from different sources (from Section 3.1.4), we frame each paraphrase as a different interpretation of the target language. We hypothesise that the variation in surface form expression from different MT sources can be better combined by learning MT source-specific sub-encoders. Equations (3.11) to (3.13) illustrate parallel encodings $\bar{E}^k$ for K MT sources using parallel encoders $Q_{\phi-k}$ with distinct parameters $\phi - k$. Each source-specific encoder

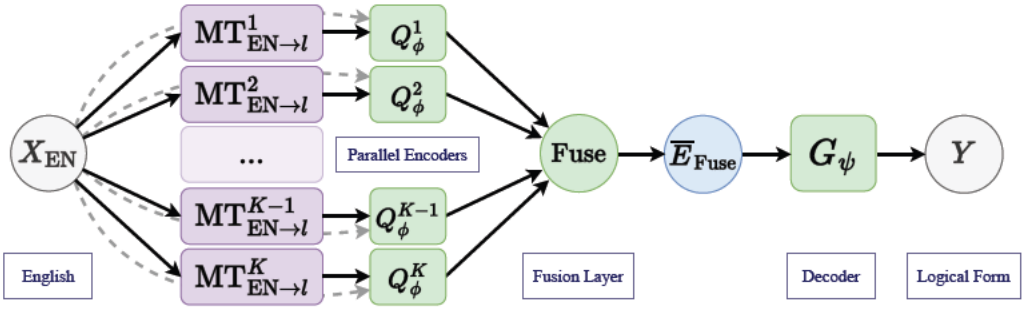

Figure 3.4: The FATES model combining encodings for parsing. FATES uses K parallel encoders for K machine translation (MT) sources to model each variant of 'translationese' independently. Each encoder is combined using a fusion layer (Fuse) for the decoder to incorporate pooled information from multiple MT sources. Each encoder can also be trained using the source English paired data (dashed grey arrow). During inference, the test utterance is input to all encoders simultaneously.

learns different parameters individually tuned to the *perspective* of each source MT system. We hypothesise that each encoder can learn improved encodings for the output distribution of an individual MT model, rather than a single encoder learning multiple distributions.

$$\bar{E}^{1} = Q_{\phi-1}\left(\bar{x}^{1}_{l}\right) \qquad \text{Encode MT input 1} \qquad (3.11)$$

$$\bar{E}^{2} = Q_{\phi-2}\left(\bar{x}^{2}_{l}\right) \qquad \text{Encode MT input 2} \qquad (3.12)$$

$$\bar{E}^{K} = Q_{\phi-K}\left(\bar{x}^{K}_{l}\right) \qquad \text{Encode MT input K} \qquad (3.13)$$

These parallel encodings can be combined for the decoder to jointly exploit encoded information from each input. For this proposal, we introduce FATES: **F**used **A**ttention **E**nsembling for **S**emantic Parsing. FATES is illustrated in Figure 3.4, which describes how parallel encoder models differ from the typical TRANSLATE TRAIN pipeline.

Section 3.1.4 defines an ensemble combining data sources into one dataset to train an encoder-decoder model. FATES differs as we learn a distinct encoder specialised for each MT system (i.e., for each system's specific translationese dialect) for fusion before the decoder. The intuition is that learning separate encoders, with different parameters, specialises in multiple respective MT output distributions during training. We expect that by learning parallel encoders, each will contribute different information to the decoder. FATES comprises K parallel Transformer encoders with parameters $\phi - k$ and a single Transformer decoder with parameters ψ. Training uses the paraphrased

synthetic data defined in Section 3.1.4 formatted as Equation (3.14) pairing output y with K parallel translations from different MT systems. Similar to Section 3.1.4, we examine the setting of K encoders from parallel sources in Section 3.3.

$$\mathcal{S}^{\text{MT}-K} = \{\bar{x}^i_{l-1},\ \bar{x}^i_{l-2}, \ldots, \bar{x}^i_{l-K}, y^i\}_{i=0}^{N} \tag{3.14}$$

The FATES encoders and decoder follow the structure given in Section 2.2.1, with an additional fusion layer to combine the parallel encoders. For K parallel inputs with lengths T_k: parallel encodings, $\bar{E}^k \in \mathbb{R}^{T_k \times d}$, are fused into a single state, $\bar{E}_{\text{Fuse}} \in \mathbb{R}^{\max_k(T_k) \times d}$, through Equation (3.15). The combined $\bar{E}_{\text{Fuse}}$ replaces the typical encoder output, $E \in \mathbb{R}^{T \times d}$ in the encoder-decoder cross-attention layer of the Transformer decoder (Equation (3.16)). Therefore enabling the decoder to simultaneously attend to the fused perspectives of each encoder.

$$\bar{E}_{\text{Fuse}} = \text{Fuse}\left(\bar{E}^1, \ldots, \bar{E}^K\right) \tag{3.15}$$

$$D' = \text{MultiHeadAttention}\left(\text{Query} = D,\ \text{Key} = \bar{E}_{\text{Fuse}},\ \text{Value} = \bar{E}_{\text{Fuse}}\right) \tag{3.16}$$

Equation (3.17) defines our options for the Fuse function. We consider no fusing by concatenating the encodings, a geometric mean over encodings (Mean), or a gating network following similar ensembling methods for machine translation (Garmash and Monz, 2016; Firat et al., 2016). The gating function computes a weighting for each of K encodings such that each encoding contributes a varying ratio of useful information to the ensemble. The intuition is that the gating function uses learnable parameters adaptively selecting important features from each input during fusion. Equation (3.18) defines the gating function g with learnable parameters $W_g \in R^{K \times d}$, $W_h \in R^{d \times Kd}$. As each gating probability g_k sums to one, the gating network in the model must decide which proportion of K encodings is most beneficial for the task objective via observing all encodings globally. We also observe the fusion layer as a necessary bottleneck to control for the magnitude variability in the non-probabilistic encodings, E_k. As these encodings lack any formal structure (e.g., the variational reparameterisation used later in chapter 6), we identify a utility in providing a combined and (approximately) normalised encoder output to attend to. Section 3.3 identifies how concatenation (i.e., no real fusion) is suboptimal compared to mean-pooling or a gating mechanism.

English	What's the name of the Denver airport?
Spanish	¿Cuál es el nombre del aeropuerto de Denver?
SQL	`SELECT DISTINCT airport_1.airport_code FROM airport airport_1, airport_service airport_service_1, city city_1 WHERE airport_1.airport_code = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = DENVER;`
English	Get my texts from Jeannine and Josh.
French	Obtenez mes textos de Jeannine et Josh.
TOP-LF	`[IN GET_MESSAGE [SL RECIPIENT Jeannine] [SL RECIPIENT Josh]]`

Table 3.2: Data examples for MultiATIS++SQL (Chapter 2), and MTOP (Li et al., 2021). We show the source English sentence, the gold-standard professional translation, and the logical form in respective LF language. Similar to Table 3.13.

$$\text{Fuse}\left(\bar{E}^1,\ldots,\bar{E}^K\right) = \begin{cases} \left[\bar{E}^1;\ldots;\bar{E}^K\right] & \text{(Concatenate)} \\ \frac{1}{K}\sum_{k=1}^{K}\bar{E}^k & \text{(Mean)} \\ \sum_{k=1}^{K} g_k \bar{E}^k & \text{(Gated)} \end{cases} \tag{3.17}$$

$$g_{1,\ldots,K} = \text{softmax}\left(W_g \tanh\left(W_h[\bar{E}^1;\ldots;\bar{E}^K]\right)\right) \tag{3.18}$$

We contrast FATES with the paraphrasing augmentation outlined in Section 3.1.4. We also contrast to single source TRANSLATE TEST and TRANSLATE TRAIN as baselines.

3.2 Experiments

3.2.1 Datasets

MultiATIS++SQL For TRANSLATE TEST, we use the gold test data in target languages: French (FR), Portuguese (PT), Spanish (ES), German (DE), and Chinese (ZH). Each utterance is translated to English (see Section 3.2.2) and used to predict a logical

form. Accurate TRANSLATE TEST inference will output an identical SQL LF from any target language. For TRANSLATE TRAIN and our augmented approaches, we use the English training utterances translated to each target language to produce training data. Section 2.3.1 details a complete description of MultiATIS++SQL.

MTOP We similarly evaluate TRANSLATE TEST using the gold test data in target languages: French (FR), Spanish (ES), German (DE), Hindi (HI), and Thai (TH). Each utterance is translated into English to predict a logical form. Similarly, TRANSLATE TRAIN uses machine translation from English to target languages to generate training data. A complete description of MTOP is given in Section 2.3.2.

MTOP differs from MultiATIS++SQL as the output LFs are *semantically equivalent* but not identical across languages. The LFs contain tokens from the input utterance in each language, and we replace these tokens with sentinel word labels (see description in Section 2.3.2). To produce translated data which matches this formatting, we must replace these tokens in the LF outputs of translated utterance inputs. For example, if the entity 'Jeannine' in the MTOP example in Table 3.2 moves from word 4 to word 2 in the translated sentence, the corresponding LF must replace '`word4`' for '`word2`' to maintain an accurate TOP-LF output. For this preprocessing step, we follow Rosenbaum et al. (2022) in using SimAlign (Jalili Sabet et al., 2020) for cross-lingual word alignment. SimAlign is an automatic multilingual BERT-based word alignment toolkit which supports all our target languages for MTOP. SimAlign receives the original and translated utterances as input and outputs an alignment matrix of lexical similarity pairs. We use this matrix to modify TOP-LF outputs for each translated dataset imitating gold standard input-output pairs for any language. We can now expect TRANSLATE TEST and TRANSLATE TRAIN data to match the structure of the gold MTOP examples with a closed decoder vocabulary. We note that we do not require this word alignment tool in later chapters which do not use synthetic training data.

3.2.2 Translation Systems

Table 3.3 outlines the key features of each MT system we consider in this thesis for TRANSLATE TEST as $\text{MT}_{l \to \text{EN}}$, and for TRANSLATE TRAIN as $\text{MT}_{\text{EN} \to l}$. We use a range of competitive >1B parameter MT Transformer models (NLLB variants, M2M100) supporting multidirectional translation (i.e., not tuned for a specific input-output language pair). We also consider the smaller OPUS Transformer models which

Name	Source	Model Type	MT Direction	Languages	Parameters
OPUS	Tiedemann and Thottingal (2020)	Transformer	EN $\rightarrow l$, $l \rightarrow$ EN	231	~78M
M2M1001.2B	Fan et al. (2021)	Transformer	Any $\leftrightarrow$Any	100	1.2B
NLLB1.3B	NLLB Team et al. (2022)	Transformer	Any $\leftrightarrow$Any	200	1.3B
NLLB3.3B	NLLB Team et al. (2022)	Transformer	Any $\leftrightarrow$Any	200	3.3B

Table 3.3: Comparison between the MT systems we evaluate in this chapter. We consider three >1 billion parameter multilingual MT systems with competitive performance in all MT directions. TRANSLATE TRAIN uses the EN $\rightarrow$ l MT direction and TRANSLATE TEST uses the l $\rightarrow$EN.

comprise many pairwise bilingual MT models. OPUS supports 233 source languages and 231 target languages for 1739 total possible MT directions. OPUS is competitive over multiple MT benchmarks (Tiedemann and Thottingal, 2020) and is more efficient during inference owing to smaller model sizes.

We verify the suitability of these systems by computing the sentence-level BLEU score (Papineni et al., 2002) between predicted translations and the gold-standard training data over all languages in MultiATIS++SQL and MTOP. We report the BLEU scores and language- and system-level averages in Table 3.4. For MultiATIS++SQL, we observe high-performance translation for NLLB and M2M100 models on Indo-European languages. All models perform similarly on MTOP with OPUS reporting better translation than larger models in some languages. These scores support the argument that different MT sources have different characteristics and suggest MT could be useful as a target language paraphrase generator in semantic parsing. We estimate that lower BLEU scores for Asian languages (Chinese, Thai) are attributable to differences in tokenisation between MT systems and gold-standard data.

3.2.3 Experimental Setting

3.2.3.1 Setting and Comparison

MONOLINGUAL **Gold** A monolingual Transformer model is trained on the gold training dataset for each target language. The model follows the encoder-decoder architecture outlined in Section 2.2.1 using the same MBART50 pre-trained encoder. This model represents the monolingual performance **upper-bound** for each language without any data constraints.

MultiATIS++SQL							
MT System		FR	PT	ES	DE	ZH	Avg. (System)
OPUS	Tiedemann and Thottingal (2020)	46.4	49.7	42.8	32.0	4.6	35.1
M2M1001.2B	Fan et al. (2021)	80.5	85.1	51.4	29.9	5.1	50.4
NLLB1.3B	NLLB Team et al. (2022)	80.5	93.1	51.4	63.6	6.9	59.1
NLLB3.3B	NLLB Team et al. (2022)	59.3	91.9	41.5	62.6	5.3	52.1
	Avg. (Language)	66.6	79.9	46.8	47.0	5.5	
MTOP							
MT System		FR	ES	DE	HI	TH	Avg. (System)
OPUS	Tiedemann and Thottingal (2020)	30.7	39.3	21.7	25.1	6.6	24.7
M2M1001.2B	Fan et al. (2021)	15.1	39.3	23.0	37.8	14.3	25.9
NLLB1.3B	NLLB Team et al. (2022)	15.1	39.3	21.7	23.1	18.6	23.6
NLLB3.3B	NLLB Team et al. (2022)	17.8	41.5	21.7	23.1	14.5	23.7
	Avg. (Language)	19.7	39.8	22.0	27.3	13.5	

Table 3.4: Sentence-level BLEU score (Papineni et al., 2002) between machine translated training split and gold training split across ATIS (Hemphill et al., 1990) and MTOP (Li et al., 2021). We report metrics for six different target languages (French, Portuguese, Spanish, Chinese, Hindi, Thai). We select high quality MT systems which can generate accurate translations in many languages. Most BLEU scores are similar across different systems identifying the ease or difficulty of the translation task. Most systems do not produce identical outputs suggesting combining from multiple MT sources may be advantageous.

MULTILINGUAL **Gold** A multilingual Transformer is trained on the union of all professionally translated data from individual MONOLINGUAL Gold models for each dataset. This follows the same Transformer architecture as MONOLINGUAL Gold. This model is the multilingual performance **upper-bound** for each language—representing the potential performance with access to many high-quality training data examples in every target language.

EN **Only** The monolingual Transformer for English from the MONOLINGUAL Gold upper-bound is used to predict logical forms directly from target language test utterances without machine translation. This is a **zero-shot lower-bound** for our models, representing the minimum performance training a model on English and predicting on target languages without engineering any specific target language parsing capability or cross-lingual representation alignment.

TRANSLATE TEST As discussed in Section 3.1.1, we predict a logical form through the parser for English by simulating the source language using machine translation. This is a **translation lower-bound** for our methods representing the current capability of the English parser to map translated target languages to logical forms without any additional target language models. In Section 3.3.2, we also analyse if any other natural languages are appropriate pivot languages for TRANSLATE TEST i.e., if any target language is a better language than English for TRANSLATE TEST inference.

TRANSLATE TRAIN Monolingual Similarly discussed in Section 3.1.2, we use translation to generate target language training data and train a parser on this data as described in Section 3.1.3. This model is a **translation lower-bound** for our methods as the minimum target language performance using a single MT source training on a singular target language.

TRANSLATE TRAIN Multilingual This is a multilingual version of TRANSLATE TRAIN Monolingual, wherein we train a single parser model for all target languages. This model may benefit from additional combined training data and combined multilingual modelling allowing feature sharing across target languages.

Silver-standard Data in Large Language models We compare to two methods for parsing MTOP using silver-standard data. First, "Translate-and-Fill" (Nicosia et al., 2021, TaF) generates training data using the mT5 pre-trained model (Xue et al., 2021), and then uses the same pre-trained model to project the word alignment labels similar to our method using SimAlign in Section 3.2.1. TaF then trains a parser using this synthetic data. The parser is based on mT5-large with 700 million parameters, or mT5-XL with 3.3 billion parameters. Second, CLASP (Rosenbaum et al., 2022) uses MT and prompting to generate multilingual training data. This data is preprocessed using sentinel labels similar to our process in Section 2.3.2 but uses string-matching-based replacement to process the translated utterances. A parser based on the 500 million parameter AlexaTM-500M (FitzGerald et al., 2022) is then trained on this synthetic data. Both prior works employ some variant of TRANSLATE TRAIN using larger pre-trained base models to compensate for synthetic data quality. In this chapter, we compare our smaller model to these techniques with similar quality data. In later chapters, we comment on how using alternative methods (including gold data) with the same smaller model can improve on these prior results.

		EN	FR	PT	ES	DE	ZH	TARGET AVG.
Lower	EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
Upper	MONOLINGUAL Gold	72.3	74.2	72.5	71.5	73.2	73.0	72.9
	MULTILINGUAL Gold	74.9	74.2	73.0	70.4	74.6	73.7	**73.2**

(a) MultiATIS++SQL

		EN	FR	ES	DE	HI	TH	TARGET AVG.
Lower	EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
Upper	MONOLINGUAL Gold	72.4	66.5	69.5	62.4	59.4	53.0	62.2
	MULTILINGUAL Gold	75.5	69.7	72.4	67.9	65.5	54.6	**66.0**

(b) MTOP

Table 3.5: Results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy for zero-shot transfer lower bound and gold data upper bounds. Lower bound performance assumes only English data and no machine translation. Upper bound performance represents an ideal scenario without resource constraints in any target language. The significant best TARGET AVG. performance is bolded.

3.2.3.2 Model Training

Our experimental setup largely follows the template of Section 2.2. The model is a Transformer (Vaswani et al., 2017) encoder-decoder. The encoder is pre-trained using the MBART50 pre-trained model and the decoder parameters are randomly initialised. For the FATES model: we select K parallel encoders receiving parallel inputs from K different MT systems. The parallel encoders in FATES are all similarly initialised using the encoder parameters from MBART50. We report analysis on both the setting of K parallel inputs and the choice of fusion function (Equation (3.17)). The fusion parameters in FATES match the decoder dimensionality ($d = 1,024$) and are randomly initialised similar to the decoder.

3.3 Results

3.3.1 What are the Upper and Lower Bounds?

We first consider the performance lower and upper bounds applicable to FATES, and the contributions in Chapters 4 to 6. The lower bound systems represent the minimum viability for our contributions, and the upper bounds represent the ideal scenario using 100% translation into target languages. Table 3.5 shows both bounds for MultiATIS++SQL and MTOP. The lower bound uses only cross-linguistic information from multilingual pre-training for cross-lingual transfer after fine-tuning on English data. Our 'EN Only' results identify that this lower bound is insufficient for accurate cross-lingual transfer. Compared to the 'MONOLINGUAL' upper bound average, the lower bound is -25.1% weaker for MultiATIS++SQL and -28.4% weaker for MTOP. This gap is largest for languages dissimilar to English, with -34.5% gap for ZH in MultiATIS++SQL, and -40.2% gap for TH in MTOP. This contrast is indicative of challenging cross-lingual transfer to languages which share few features to minimally benefit from multilingual pre-training. The overall contrast identifies there is a large gap in potential progress to improve parsing beyond the lower bound without the required effort of the upper bound.

Comparing across upper bounds, we identify that 'MULTILINGUAL Gold' is significantly superior to 'MONOLINGUAL Gold' modelling. We observe improvement for both source EN, and the average across target languages. This overall improvement suggests that at our model and dataset sizes — we are unlikely to observe *the curse of multilinguality* proposed in Conneau et al. (2018b) i.e., the system is is unlikely to perform worse overall from training on more input languages. While 'MULTILINGUAL Gold' is overall superior, the 'MONOLINGUAL Gold' is competitive in each language. In four of five target languages, the monolingual MultiATIS++SQL accuracy improves over the performance for English. Overall, our upper bound results highlight the benefit of access to high-quality training data in each target language.

3.3.2 Is TRANSLATE TEST Competitive with the Upper Bound?

Automatic translation is the fastest and most cost-effective strategy for cross-lingual semantic parsing in our case study. If MT is viable, then engineering effort and cost in producing target language parsers can be circumvented. Given the established bounds, we now consider the economical TRANSLATE TEST strategy using the 'MONOLINGUAL Gold' parser for EN to predict logical forms using machine translation (from language l

		EN	FR	PT	ES	DE	ZH	TARGET AVG.
Lower Bound	EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
Upper Bound	MONOLINGUAL Gold	72.3	74.2	72.5	71.5	73.2	73.0	72.9
	MULTILINGUAL Gold	74.9	74.2	73.0	70.4	74.6	73.7	73.2
TRANSLATE TEST	OPUS	—	57.7	58.1	58.3	58.8	50.9	56.8
	M2M1001.2B	—	58.8	58.1	59.8	59.1	56.6	58.5
	NLLB1.3B	—	60.4	58.3	58.8	59.0	58.5	**59.0**
	NLLB3.3B	—	56.6	58.8	56.0	59.0	50.9	56.3

(a) MultiATIS++SQL

		EN	FR	ES	DE	HI	TH	TARGET AVG.
Lower Bound	EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
Upper Bound	MONOLINGUAL Gold	72.4	66.5	69.5	62.4	59.4	53.0	62.2
	MULTILINGUAL Gold	75.5	69.7	72.4	67.9	65.5	54.6	66.0
TRANSLATE TEST	OPUS	—	44.9	63.1	39.1	47.1	54.2	49.7
	M2M1001.2B	—	45.3	63.7	40.3	46.8	53.5	49.9
	NLLB1.3B	—	45.0	63.4	39.5	46.8	53.9	49.7
	NLLB3.3B	—	45.5	66.0	39.3	47.0	54.1	**50.4**

(b) MTOP

Table 3.6: TRANSLATE TEST results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy compared to upper and lower bounds. TRANSLATE TEST is the fastest strategy for supporting new languages in our parsing pipeline case study. The significant best TARGET AVG. performance for TRANSLATE TEST is bolded.

to EN) as an intermediary during inference.

Table 3.6 outlines our experiments using one of four possible translation sources translating into English. 'TRANSLATE TEST' proves more capable than the lower bound across all languages and MT systems. The weakest 'TRANSLATE TEST' method improves on the upper bound by +8.5% average for MultiATIS++SQL and +16.0% average for MTOP. Practically, these methods are nearly identical except' TRANSLATE TEST' adds an MT system before the prediction. Unsurprisingly, simulating EN improves performance above inference from an utterance in any other input language. Across MT sources, different MT systems report different strengths across individual target languages, with no singular MT system as strictly superior to others. However, the gain over other models for average improvement is significant (bold in

	Test Language						
Train Language	EN	FR	PT	ES	DE	ZH	TARGET AVG.
EN	*72.3*	57.7	58.1	58.3	58.8	50.9	56.8
FR	46.5	*74.2*	54.9	39.0	68.1	13.6	44.4
PT	54.9	59.9	*72.5*	28.2	50.9	2.4	39.3
ES	51.9	50.5	42.5	*71.5*	49.3	4.7	39.8
DE	50.7	67.1	47.4	43.4	*73.2*	40.8	49.9
ZH	44.6	37.1	24.9	12.0	32.9	*73.0*	30.3
LANG AVG.	49.7	54.5	45.6	36.2	52.0	22.5	

Table 3.7: Execution accuracy for MultiATIS++SQL varying training language for TRANSLATE TEST evaluation. We use only OPUS machine translation (Tiedemann and Thottingal, 2020) to produce TRANSLATE TEST inputs. *Italics* denote the training language accuracy which other languages should match if translation is accurate. English (EN) is empirically the best pivot language for TRANSLATE TEST from the best TARGET AVG. score. We suggest this effect is a function of both translation quality and pre-training corpora size for the base MBART50 model (Tang et al., 2021).

Table 3.6). We highlight that the improvement from TRANSLATE TEST is insufficient to compete with the upper bounds, as the strongest 'TRANSLATE TEST' is -14.2% average below 'MULTILINGUAL Gold' for MultiATIS++SQL and -15.6% average below 'MULTILINGUAL Gold' for MTOP. 'TRANSLATE TEST' performance is also always below the respective performance for EN. This suggests that our hypothesis that machine translation is an adequate substitute is potentially invalid, given the lack of parity between upper bounds and 'TRANSLATE TEST'. We note that only for TH MTOP do we observe similar performance between 'TRANSLATE TEST' and the upper bound, but this accuracy is below other languages and can likely be improved overall.

Is English the Best Pivot Language?

Our results for TRANSLATE TEST in Table 3.6 implicitly assume English is the optimal setting for monolingual modelling to translate into during inference (i.e., the pivot language). This is a reasonable assumption given the original task is designed for English and later translated into target languages. English is also the common focus as the output language in MT (Koehn et al., 2022). However, Moghe et al. (2023a) recently

highlighted that other languages can perform comparably as the pivot language in low-resource scenarios. From a linguistic perspective, an ideal pivot language would be the easiest to translate into. This could manifest with minimal inflectional morphology or grammatical gender to require fewer assumptions from the MT system during inference. English has some of these features; but Chinese arguably goes further in having no inflectional morphology for gender, tense or case (Packard, 2000).

To examine if English is the best pivot in our setup, we evaluate monolingual and TRANSLATE TEST performance for target languages in MultiATIS++SQL. We test only OPUS translation as a control and consider if any other language can function as the TRANSLATE TEST pivot. Our results in Table 3.7 verify that English is empirically the best pivot language in our task. German and French are the closest competitors which could potentially be on par with English given additional tuning. Moghe et al. (2023b) reports wider analysis highlighting that MT metrics have minimal correlation with parsing performance regardless of pivot language. Consequently, selecting an appropriate pivot is not possible from MT evaluation alone. Despite an idealised framing of pivoting through Chinese, we observe the lowest TRANSLATE TEST using ZH as the pivot. The contrast between in-language performance (73.0%) and the TRANSLATE TEST accuracy (average 30.3%) highlights that this failure in TRANSLATE TEST is due to poor MT. Consequently, we argue English is currently the only sensible choice for a pivot language.

As raised in Section 3.1.1, we consider the 'TRANSLATE TEST' as a robust comparison which is inflexible to improve from multilingual modelling or improved representation alignment. 'TRANSLATE TEST' improves parsing by viewing the entire task as monolingual. However, we have already observed some benefits of multilingual parser modelling in the contrast between 'MONOLINGUAL Gold' and 'MULTILINGUAL Gold' upper bounds. We consider if this contrast is also reflected when training on MT data. Therefore, we now consider the 'TRANSLATE TRAIN' framework to contrast directly modelling a target language with translationese training data.

3.3.3 Is TRANSLATE TRAIN Competitive with the Upper Bound?

Table 3.8 compares single source 'TRANSLATE TRAIN' to lower and upper bounds, the best TRANSLATE TEST system, and recent work using LLMs with synthetic data for MTOP. In general, we find that single-source 'TRANSLATE TRAIN' is weaker than 'TRANSLATE TEST'. For MTOP, 'TRANSLATE TRAIN' actually performs poorer

		EN	FR	PT	ES	DE	ZH	TARGET AVG.
Lower Bound	EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
Upper Bound	MONOLINGUAL Gold	72.3	74.2	72.5	71.5	73.2	73.0	72.9
	MULTILINGUAL Gold	74.9	74.2	73.0	70.4	74.6	73.7	73.2
TRANSLATE TEST	NLLB1.3B (best)	—	60.4	58.3	58.8	59.0	58.5	**59.0**
TRANSLATE TRAIN	OPUS	—	56.8	39.1	51.8	60.4	59.6	53.5
	M2M1001.2B	—	59.1	46.1	40.8	63.0	42.9	50.4
	NLLB1.3B	—	55.8	28.7	43.3	59.0	51.1	47.6
	NLLB3.3B	—	62.6	50.9	41.6	58.5	50.3	52.8

(a) MultiATIS++SQL

		EN	FR	ES	DE	HI	TH	TARGET AVG.
Lower Bound	EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
Upper Bound	MONOLINGUAL Gold	72.4	66.5	69.5	62.4	59.4	53.0	62.2
	MULTILINGUAL Gold	75.5	69.7	72.4	67.9	65.5	54.6	66.0
Prior work	TaF mT5-large	83.5	71.1	69.6	70.5	58.1	57.5	65.4
	TaF mT5-XL	85.9	74.0	71.5	72.4	61.9	60.2	68.0
	CLASP	84.4	72.6	68.1	66.7	58.1	—	—
TRANSLATE TEST	NLLB3.3B (best)	—	45.5	66.0	39.3	47.0	54.1	**50.4**
TRANSLATE TRAIN	OPUS	—	24.4	23.1	32.7	22.4	9.5	22.4
	M2M1001.2B	—	22.2	20.4	28.1	24.8	10.9	21.3
	NLLB1.3B	—	23.5	23.8	28.3	27.6	8.5	22.3
	NLLB3.3B	—	23.7	23.4	28.5	27.5	10.4	22.7

(b) MTOP

Table 3.8: TRANSLATE TRAIN results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy compared to upper and lower bounds. We constrast single source TRANSLATE TRAIN to lower and upper bounds and the best TRANSLATE TEST result. For MTOP, we also compare to Translate-and-Fill (Nicosia et al., 2021, TaF), and CLASP (Rosenbaum et al., 2022). These methods use large language models with silver standard data as an alternative formulation of TRANSLATE TRAIN.

than the zero-shot lower bound. In no dataset or language is it observed that single source TRANSLATE TRAIN modelling is a sufficient proxy for professional translation. All models from any MT source lag behind the upper bounds. Compared to the monolingual model, the best outcomes for each dataset are -10.2% gap for German MultiATIS++SQL using M2M1001.2B, and -29.7% for German MTOP using OPUS. This best-case scenario is further evidence to reject this chapter's hypothesis.

Across 'TRANSLATE TRAIN' systems, German proves the "easiest" language to model using MT with the lowest gap between 'MONOLINGUAL Gold' and any singular translation system for both datasets. The most difficult languages to model using MT are MultiATIS++SQL Portuguese (PT, -21.6%) and MTOP Spanish (ES, -45.7%). This may be a manifestation of fewer pre-training tokens for these languages, compared to higher-resource German and Chinese (Tang et al., 2021). In Sections 3.3.4 to 3.3.5, we propose strategies to improve 'TRANSLATE TRAIN' improving generalisation to target languages with data augmentation and encoder ensembling.

Comparing TRANSLATE TRAIN and TRANSLATE TEST We observe that 'TRANSLATE TEST' generally performs above 'TRANSLATE TRAIN' across both datasets. For MultiATIS++SQL, 'TRANSLATE TRAIN' is only above 'TRANSLATE TEST' for high-resource language pairs in using some MT systems e.g., NLLB3.3B in FR and DE. These high-resource languages (FR, DE) are more frequently included in competitive machine translation benchmarks (Koehn et al., 2022). Therefore, it is unsurprising that the strongest 'TRANSLATE TRAIN' modelling is reported for these languages. Without additional modelling effort, 'TRANSLATE TEST' is a stronger baseline technique owing to the reduced complexity of modelling the task entirely in high-resource English.

For MTOP, our results are significantly poorer than 'TRANSLATE TEST', but also dramatically weaker than 'TaF' and 'CLASP' which report accuracies much closer to the upper bound. These systems use large models trained on more diverse corpora for semantic parsing, and our smaller model with automatically generated data fares poorly across every language. We will revisit the comparison to 'TaF' and 'CLASP' in Chapter 6, where our results using gold few-shot sampling are competitive with these methods. While not definitive, we conjecture that our failure in accurate target language parsing of MTOP is owed to poor automatic word alignment. We use SimAlign (Jalili Sabet et al., 2020) following CLASP (Rosenbaum et al., 2022) to project word labels across datasets. However, we observe that the alignment output from this model fails when entities are mistranslated. As this is a common error (see Section 3.3.7), our

synthetic data for modelling MTOP using 'TRANSLATE TRAIN' is a poor representation of any target language. 'TRANSLATE TEST' can circumvent this issue by translating from the target language into English. We observe the token-level alignments here to be much more robust to MT variation with fewer failures from alignments to rare entities. We observe word alignment for every target language than for English, leading to lower 'TRANSLATE TRAIN' performance relative to other parsers.

No singular translation system is strictly superior to others across all datasets. Notably, we do not observe a positive correlation between the BLEU score for MT outputs (Table 3.4) and translation quality. We do not observe a significant positive correlation between BLEU and 'TRANSLATE TRAIN' (Pearson $\rho = -0.19,\ p = 0.42$), or BLEU and 'TRANSLATE TEST' (Pearson $\rho = 0.48,\ p = 0.15$). A weak correlation here suggests that semantic parsing accuracy is not reflected in contemporary MT evaluation, and is poorly predicted by metrics like BLEU. As a result, it is potentially unwise to recommend an optimal translation system for our task. In later chapters, we opt to use translations from OPUS in baselines as the most computationally efficient MT prediction strategy.

'TRANSLATE TEST' is generally superior to 'TRANSLATE TRAIN' across both datasets. Neither Table 3.6 or Table 3.8 can validate our hypothesis on the adequacy of MT for cross-lingual semantic parsing. As previously discussed, 'TRANSLATE TEST' cannot benefit from multilingual modelling or data augmentation by using translation at inference. This undesirably constrains all possible improvements to 'TRANSLATE TEST' as a function of translation quality. Given this constraint, we now focus on methods for improving the 'TRANSLATE TRAIN' parsing pipeline.

3.3.4 Monolingual, Bilingual or Multilingual Modelling?

We experiment with data augmentation within 'TRANSLATE TRAIN' to evaluate if single source parsing can be improved by enhancing the training data. We consider introducing additional MT sources as 'paraphrase augmentation' (i.e., $K > 1$ in Equation (3.9) as described in Section 3.1.4), additionally training on English (+EN), and multilingual modelling combining the translationese training data from all target languages simultaneously. Table 3.9 outlines our results for MultiATIS++SQL and MTOP. For selecting MT sources for each sample, we select the K best systems from Table 3.8 for each language.

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
MONOLINGUAL $K=1$	—	62.6	50.9	51.8	63.0	59.6	57.6
MONOLINGUAL $K=2$	—	65.3	59.8	39.3	61.3	59.1	57.0
MONOLINGUAL $K=3$	—	66.3	65.5	58.3	64.0	61.0	63.0
MONOLINGUAL $K=4$	—	65.1	65.5	58.5	62.3	63.8	63.0
MONOLINGUAL $K=1$ +EN	—	66.2	66.2	53.0	40.8	61.3	57.5
MONOLINGUAL $K=2$ +EN	—	67.0	65.1	50.7	65.5	62.1	62.1
MONOLINGUAL $K=3$ +EN	—	65.9	64.6	60.0	66.5	65.5	64.5
MONOLINGUAL $K=4$ +EN	—	66.8	67.4	60.0	64.3	59.6	63.6
MULTILINGUAL $K=1$	—	63.4	61.7	56.0	63.8	59.6	60.9
MULTILINGUAL $K=2$	—	63.4	61.7	56.0	63.4	59.6	60.8
MULTILINGUAL $K=3$	—	65.3	65.1	52.4	64.9	59.6	61.5
MULTILINGUAL $K=4$	—	65.7	65.5	57.7	66.2	62.1	63.4
MULTILINGUAL $K=1$ +EN	71.1	67.0	62.6	55.4	64.0	64.0	62.6
MULTILINGUAL $K=2$ +EN	72.5	67.0	64.3	61.0	66.2	64.3	64.6
MULTILINGUAL $K=3$ +EN	71.1	67.6	65.9	61.0	66.2	62.3	64.6
MULTILINGUAL $K=4$ +EN	70.4	65.9	66.5	62.7	65.7	64.0	**65.0**

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
MONOLINGUAL $K=1$	—	24.4	23.8	32.7	27.6	10.9	23.9
MONOLINGUAL $K=2$	—	25.0	18.9	28.9	28.3	10.3	22.3
MONOLINGUAL $K=3$	—	25.1	20.7	29.8	25.9	12.7	22.8
MONOLINGUAL $K=4$	—	25.5	23.7	30.4	29.8	12.7	24.4
MONOLINGUAL $K=1+$ EN	—	32.3	30.0	38.1	31.2	13.9	29.1
MONOLINGUAL $K=2+$ EN	—	32.3	32.8	38.3	32.7	12.7	29.8
MONOLINGUAL $K=3+$ EN	—	35.7	32.9	38.6	33.9	13.9	31.0
MONOLINGUAL $K=4+$ EN	—	38.8	36.5	38.9	36.8	13.7	32.9
MULTILINGUAL $K=1$	—	27.2	26.9	32.3	30.2	14.5	26.2
MULTILINGUAL $K=2$	—	28.6	26.8	31.6	31.3	15.4	26.7
MULTILINGUAL $K=3$	—	28.5	27.1	35.2	32.6	12.8	27.2
MULTILINGUAL $K=4$	—	28.6	27.0	33.4	32.0	15.4	27.3
MULTILINGUAL $K=1+$ EN	64.9	32.0	31.4	39.0	33.1	14.2	29.9
MULTILINGUAL $K=2+$ EN	64.8	32.7	31.8	40.8	33.8	15.5	30.9
MULTILINGUAL $K=3+$ EN	76.5	39.8	40.2	45.9	35.3	15.7	**35.4**
MULTILINGUAL $K=4+$ EN	70.5	37.0	36.6	41.8	35.1	14.1	32.9

(b) MTOP

Table 3.9: Paraphrase augmentation for (a) MultiATIS++SQL and (b) MTOP. We compare between monolingual (MONOLINGUAL) and multilingual (MULTILINGUAL) training with and without English (+EN). We sample K MT sources and report results for $K = \{1,2,3,4\}$. The significant best result is bolded.

Improving Monolingual Parsing with More Translations: We observe that increasing the number of source translations improves parsing accuracy. For the monolingual case, $K = 4$ always improves on $K = 1$ in average accuracy in Table 3.9. This supports the argument that ensembling MT systems can function as paraphrases to improve the parser. However, we highlight that the improvement from increasing sources K is not strictly increasing.

For MultiATIS++SQL and MTOP: there is > 1 case where average accuracy at K sources is lower than the respective metric at $K-1$ sources. While MultiATIS++SQL FR and MTOP FR, the $K = 4$ approach always improves on $K = 1$. We note that MultiATIS++SQL DE and MTOP PT and ES show a *negative* trend for increasing K. We posit that combined MT outputs inherit the strengths *and weaknesses* of each constituent system and this pattern may be a consequence of an unfortunate combination of MT noise (i.e., negative interference). Despite this pattern, the average benefit of data ensembling with $K = 4$ suggests that sufficient MT diversity can overcome inconsistency in source selection.

Across all datasets, the major beneficiaries of more sources are languages with unique scripts. Parsing MultiATIS++SQL ZH improves by $+4.2\%$, and MTOP HI and TH improve by $+2.2\%$ and $+1.8\%$ respectively. The uniqueness of the script in each of these languages results in fewer shared features from pre-training within the model. Therefore, increasing the quantity of data in each language has a straightforward improvement in performance. This gain overcomes the performance degradation in monolingual ensembling for aforementioned languages, yielding the greatest improvement to average target language accuracy.

Can We Augment Our Data with Source English? Given the variable benefit in ensembling monolingual sources, we examine if introducing the source English data can improve the parsing of target languages (i.e., 'bilingual' modelling). Our intuition is that exposing the model to higher-quality input-output examples without translationese improves parser generalisation. As before, increasing K and training in English improves the average target language accuracy across all datasets. Introducing English to monolingual ensembles has a larger benefit than increasing target-language MT diversity: each 'MONOLINGUAL+EN' model at some K broadly performs above a respective 'MONOLINGUAL' with K' sources when $K' > K$. This improvement is prominent for MTOP, where adding high-quality data (even without ensembling at $K = 1$) outperforms the $K = 1$ monolingual comparison. We infer that English data mitigates

the aforementioned word alignment failures of TRANSLATE TRAIN. We suggest that the balance between improvements in adding English, and detriment from MT noise has an inflexion point of some K MT sources. At a certain K, the additional MT sources negate the benefit of additional English.

We identify a reversed trend for unique script languages than observed for monolingual ensembles. Here, introducing additional sources hurts performance everywhere when English is present. This is evidenced by the 'MONOLINGUAL $K = 1$ +EN' performance for MTOP HI and TH is *greater* than the 'MONOLINGUAL $K = 4$ +EN' performance. Increasing K sources reduces performance further. This suggests that English data is *more valuable* than any additional noisy target language data and increasing sources may *only increase system noise*.

Monolingual or Multilingual Modelling of Translation data? While the above experiments train a single parser *per language*, we also examine if a single parser for all target languages is feasible (i.e., 'multilingual' modelling). Multilinguality reduces training effort by producing one model per L languages over L models. However, a multilingual parser can struggle to model each relevant language together, whereas L monolingual models can adapt to each language without this concern. Across all datasets, multilingual training yields a more accurate parser for single-language source ($K = 1$) and multi-language source ($K = 4$) scenarios following the trend for the upper bound. Languages with more shared similarity benefit more from multilingual modelling. Contrasting between 'MONOLINGUAL $K = 4$' and 'MULTILINGUAL $K = 1$', we observe that FR and DE benefit the most from multilingual modelling across both datasets. These European languages share an alphabet and some lexical similarity (between Romance and Germanic families), to explain how jointly training on both sources contributes *positive interference* to benefit the multilingual parser. In contrast, languages with fewer mutual features benefit the least from multilingual modelling: MultiATIS++SQL, ZH performance decreases by -1.1% moving from a monolingual focus to a multilingual combination.

The Kitchen Sink: Should We Combine Everything? We finally examine combining English with the multilingual data ensemble from machine translation. Similar to above, introducing English improves overall target language accuracy. This combination of data is generally the best strategy across most languages. For example, the best strategy for MultiATIS++SQL is 'MULTILINGUAL $K = 4$ +EN' using all source English and

ATIS	EN	FR	PT	ES	DE	ZH	TARGET AVG.
FATES $K = 3$ (Concatenate)	—	63.6	68.7	32.8	59.3	49.3	54.7
FATES $K = 3$ (Mean)	—	67.4	64.2	54.8	61.1	66.7	62.8
FATES $K = 3$ (Gated)	—	69.2	66.7	47.5	67.5	68.5	**63.9**
MTOP	EN	FR	ES	DE	HI	TH	TARGET AVG.
FATES $K = 3$ (Concatenate)	—	15.6	15.7	29.2	7.3	5.0	14.6
FATES $K = 3$ (Mean)	—	27.4	27.5	32.2	28.8	12.1	25.6
FATES $K = 3$ (Gated)	—	27.4	26.5	33.1	30.9	11.9	**26.0**

Table 3.10: Fusion functions in FATES for MultiATIS++SQL and MTOP. We contrast between (a) no combiner function (the decoder attends to K concatenated utterance encodings); (b) mean combining to pool the encoded representations; (c) gated combination using a learned combiner function in Equation (3.16) and Equation (3.18).

every available MT example. We identify a similar English and MT noise inflexion point for MTOP, where $K = 3$ produces better outcomes than $K = 4$. Given the additional data from multilingual training, we observe that introducing English yields lesser benefit here than in 'MONOLINGUAL $K = 4$+EN'. The improvement between $K = 1$ to $K = 4$ is -3.7% lower for MultiATIS++SQL and -0.8% lower for MTOP between multilingual and monolingual data augmentation by adding English. Generally, English appears the most helpful for the *worst case* parser. We now evaluate modelling each paraphrase as a representation of equivalent semantics in different surface forms through the FATES model.

3.3.5 FATES: Ensembling Encoders

We now consider if parallel encoders for each of K MT sources can further reduce the cross-lingual transfer gap. Section 3.1.4 identifies that additional data can contribute more noise than benefit to the parser. FATES explicitly adapts to each translationese language using parallel encoders for each source. The intuition is that specialising a set of K parallel encoders to parallel sources reduces demand for the encoder to adapt toward multiple translationese samples. Each encoder models a single distribution, and the perspective of each encoder is combined before the decoder. By training on multiple paraphrases simultaneously, the model observes K surface forms of equivalent semantics

mapping to the same logical form. We expect this to improve parser robustness and the accuracy of representing semantics in the latent space.

Which fusion function? We consider three different functions for combining parallel encoder outputs for FATES on MultiATIS++SQL and MTOP in Table 3.10. We experiment with only the multilingual setting with $K = 3$ sources as an experimental control. The 'Concatenate' method concatenates all encoder outputs for the decoder to attend to all combined outputs without fusion. The 'Mean' computes the mean of all tokens at each time step. The 'Gated' computes a distribution over K inputs to dynamically compute a weighted sample over encodings. Across both MultiATIS++SQL and MTOP, the 'Gated' function yields significantly superior performance. The 'Gated' can dynamically select which of K inputs is more important for decoding using dynamic, encoder-specific scalar weightings. This strategy will allow a parser to better exploit more relevant inputs and ignore less relevant encoded information. Henceforth, we use the 'Gated' fusion for FATES.

Does Parallel Encoding Improve Cross-lingual Transfer? We train FATES on each configuration of data discussed in Section 3.1.4: monolingual ensembling (MONOLINGUAL), multilingual ensembling (MULTILINGUAL) and jointly training with English (+EN). Table 3.11 shows results for MultiATIS++SQL and MTOP. There are no $K = 1$ experiments as FATES requires > 1 parallel encoders. FATES with $K = 1$ is equivalent to TRANSLATE TRAIN.

FATES performs above paraphrase augmentation for most languages. Comparing each equivalent experiment (data source and K), FATES is a more accurate parser for $^{2}/_{3}^{\text{rd}}$ of experiments for MultiATIS++SQL or MTOP. Whereas increasing K had varied effects with paraphrase augmentation, FATES more consistently improves parsing with increasing K. We interpret this benefit as improved robustness by modelling each translationese dialect independently. The largest accuracy decrease from increasing K is reduced from -0.9% to -0.1% for MultiATIS++SQL and -1.3% to $+0.3\%$ for MTOP. Similar to Section 3.1.4, multilingual modelling for FATES improves over monolingual modelling. However, we do not observe a negative trend for non-Latin languages using FATES. Similar to FR and DE, performance for ZH, HI, or TH now improves with multilingual training and additional sources. We suggest that this effect is a consequence of the additional parameters from more encoders—each encoder is now more flexible to model each language independently without negative interference from combining

FATES	EN	FR	PT	ES	DE	ZH	TARGET AVG.
MONOLINGUAL $K=2$	—	65.4	65.4	36.4	63.4	63.6	58.8
MONOLINGUAL $K=3$	—	63.7	64.2	42.4	63.1	62.4	59.2
MONOLINGUAL $K=4$	—	66.0	64.2	47.5	64.7	66.9	61.9
MONOLINGUAL $K=2$ +EN	—	66.9	66.0	61.8	62.9	59.3	63.4
MONOLINGUAL $K=3$ +EN	—	67.5	69.8	55.6	65.4	67.8	65.2
MONOLINGUAL $K=4$ +EN	—	68.0	67.4	57.3	67.8	69.4	66.0
MULTILINGUAL $K=2$	—	68.7	58.2	53.3	68.0	63.6	62.4
MULTILINGUAL $K=3$	—	69.5	66.7	47.5	67.5	68.5	63.9
MULTILINGUAL $K=4$	—	69.4	66.5	50.2	65.6	67.4	63.8
MULTILINGUAL $K=2$ +EN	70.7	68.5	67.8	56.1	67.5	62.7	64.5
MULTILINGUAL $K=3$ +EN	75.4	69.2	69.8	54.2	67.8	69.2	66.0
MULTILINGUAL $K=4$ +EN	74.9	70.5	69.2	62.4	68.7	66.5	**67.5**

(a) MultiATIS++SQL

FATES	EN	FR	ES	DE	HI	TH	TARGET AVG.
MONOLINGUAL $K=2$	—	21.2	21.8	28.3	26.7	11.2	21.8
MONOLINGUAL $K=3$	—	23.2	19.4	28.1	28.6	9.3	21.7
MONOLINGUAL $K=4$	—	24.8	21.7	29.2	30.1	11.0	23.4
MONOLINGUAL $K=2$ +EN	—	32.5	32.7	41.8	33.2	13.9	30.8
MONOLINGUAL $K=3$ +EN	—	36.6	38.5	45.4	36.1	13.3	34.0
MONOLINGUAL $K=4$ +EN	—	41.2	39.6	45.0	38.1	13.9	35.6
MULTILINGUAL $K=2$	—	26.8	26.8	32.4	30.1	13.5	25.9
MULTILINGUAL $K=3$	—	27.4	26.5	33.1	30.9	11.9	26.0
MULTILINGUAL $K=4$	—	28.1	27.2	32.3	31.0	12.7	26.3
MULTILINGUAL $K=2$ +EN	72.6	42.5	44.3	48.0	34.2	11.5	36.1
MULTILINGUAL $K=3$ +EN	69.2	43.3	42.2	48.9	33.6	11.2	35.8
MULTILINGUAL $K=4$ +EN	69.7	44.7	45.7	49.0	32.9	13.4	**37.1**

(b) MTOP

Table 3.11: FATES: encoder ensembling for (a) MultiATIS++SQL and (b) MTOP. We compare between monolingual (MONOLINGUAL) and multilingual (MULTILINGUAL) training with and without English (+EN). We show results for K parallel encoders sampling K MT sources for $K = \{2,3,4\}$. The significant best result is bolded.

Method Type	Method	Cosine (↑)	Top-1	Top-5	Top-10
Pre-trained Model	MBART50	0.576	0.521	0.745	0.796
Lower Bound	EN Only	0.364	0.669	0.775	0.964
Upper Bound	MULTILINGUAL Gold	0.698	0.784	0.981	0.991
Machine Translation	FATES	**0.670**	**0.720**	**0.957**	**0.992**

Table 3.12: Average similarity between encodings of English and target languages for MultiATIS++SQL. Cosine similarity evaluates average distance between encodings of parallel sentences. Top-k evaluates if the parallel encoding is ranked within the k most cosine-similar vectors (higher (↑) is better). Best excluding the upper-bound is bold.

MT distributions.

Does English Data Improve Parallel Ensembling? Augmenting FATES with English yields further performance improvement. Each encoder in the model benefits from additionally observing gold-standard training data to raise the worst-case parser. For multilingual modelling with $K = 4$, the average benefit to adding English is $+3.5\%$ for MultiATIS++SQL, and $+10.9\%$ for MTOP. For MTOP, the improvement adding English is $+17.3\%$ for European languages but only 1.3% for non-European languages. At present, MT quality creates a bottleneck for TRANSLATE TRAIN and TRANSLATE TEST which is not yet surpassed in this setup. FATES also does not surpass the LLM-based models outlined in Table 3.8. We identify that FATES now performs significantly above TRANSLATE TEST, highlighting that multilingual modelling and improved robustness are possible when directly modelling target languages.

We do not achieve parity to our upper bounds using FATES or with paraphrase augmentation. Therefore, we can broadly consider our hypothesis as invalid, given that machine translation does not produce equivalent parsing outcomes as training with gold data. However, with the lowest budget for our case study—FATES offers a competitive parsing strategy with the least sensitivity to MT errors and translationese disfluency when using synthetic data.

3.3.6 Visualising Latent Representation Similarity

Our overarching hypothesis addresses cross-lingual transfer through latent representation alignment. In this chapter, we evaluate cross-lingual parsing by simulating target

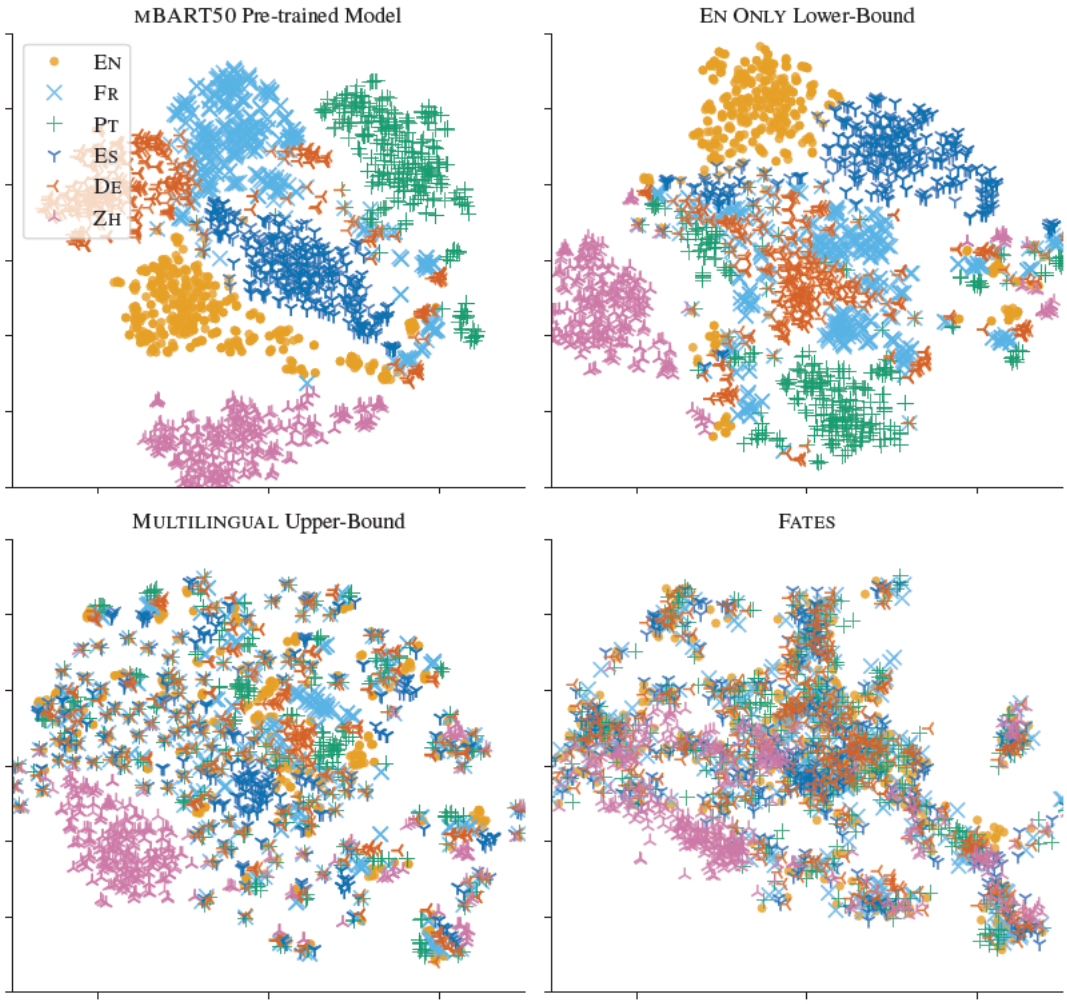

Figure 3.5: Visualisation of MultiATIS++SQL encodings (test set; 50% random parallel sample) using t-SNE. We compare the original MBART50 pre-trained model, the EN-ONLY zero-shot lower bound, MULTILINGUAL training upper bound and FATES from Chapter 3. FATES is more similar to the upper bound with improved latent cross-lingual similarity.

language data with machine translation. While not directly addressing representation alignment,[1] we expect that providing even synthetic training data will improve the latent representation structure to address the core hypothesis of this thesis. To evaluate the latent representation structure of the FATES parser, we report quantitative analysis on latent representations of MultiATIS++SQL test set in Table 3.12. We also visualise FATES compared to the initial model, lower- and upper-bounds in Figure 3.5. The methodology for this analysis and visualisation is detailed in Section 2.4.2.

Figure 3.5 identifies that the original MBART0 model and the 'EN Only' baseline exhibit *monolingual clustering* artefacts. Informally, the representations from these systems are more similar to any representation in the same language than the equivalent semantics of an utterance in a different language. This is an undesirable outcome for our objective of representation alignment as the latent space is *not semantically distributed*. This *clustering* artefact justifies how cross-lingual transfer is difficult to achieve without additional effort in aligning representations across languages. We observe this poor latent similarity in high dimensions via the low cosine-similarity between parallel encodings in Table 3.12. We also note that the lower bound model has *lower* cosine similarity to the original pre-trained encoder. We argue this is a consequence of monolingual fine-tuning contributing to catastrophic forgetting of other languages (McCloskey and Cohen, 1989; French, 1999). This further highlights the challenge of accurately representing semantics during cross-lingual transfer.

We identify that the MULTILINGUAL model upper bound largely avoids producing visually similar monolingual clustering. The exception to this is ZH, as the most dissimilar language sharing the fewest features during multilingual training. On average, MULTILINGUAL cosine and ranked similarity are much improved over the baselines or original model identifying the upper bound of parser behaviour in the ideal data scenario. We observe fewer language-level clusters with more examples clustering by parallel meaning (as supported by improved Top-*k* ranked similarity). We identify that FATES approximates the upper bound with similar latent structure and monolingual artefacts for ZH. FATES also reports similar quantitative similarity metrics with the best Top-10 ranked similarity of any model. We interpret this result as positive given that synthetic data allows a system to approximate the ideal scenario with gold data. However, we suggest that this upper bound for data is not necessarily the best latent structure. For the explicit target of improving representation alignment, subsequent chapters consider novel methods for improving representation alignment further.

[1] Chapter 4 and Chapter 6 propose methods directly addressing representation alignment.

Noun/Adjective Ambiguity ("first-class fares" is a noun object)	
EN	Show me the first class fares from Baltimore to Dallas
DE MT	Zeigen Sie mir die **erstklassigen** Tarife von Baltimore nach Dallas
DE Gold	Zeige mir die Preise in der ersten Klasse von Baltimore nach Dallas
Entity Misinterpretation (Airline names aren't preserved)	
EN	Which Northwest and United flights go through Denver before noon?
DE MT	Welche **Nordwesten** und **Vereinigten** Flüge gehen durch Denver vor Mittag
DE Gold	Welche Northwest und United Flüge gehen durch Denver vor Mittag
Question to Statement Mistranslation (rephrased as "You have a…")	
EN	Do you have an 819 flight from Denver to San Francisco?
ZH MT	你有一个从丹佛到旧金山的819 航班
ZH Gold	有没有从丹佛到旧金山的819 航班

Table 3.13: Error examples from MultiATIS++SQL. Utterances are translated into Chinese and German using both machine translation and native speakers. We highlight issues with the noisy MT data (underlined and bolded) compared to improved human translations (underlined).

3.3.7 Error Analysis

We randomly sample 50 data pairs where the gold-standard monolingual model produces a correct parser but the best TRANSLATE TRAIN model failed. As the parser can correctly parse a gold-standard utterance, but cannot parse a translated utterance in the same target language—we consider failure in these cases as entirely due to translation error (i.e., *MT breakdown* from Moghe et al. (2023b)). We discuss trends in errors here with examples shown in Table 3.13. We note that these examples are also cited as motivation for the challenges in cross-lingual parsing in Chapter 1.

Fluency and Entity Misinterpretation Proper nouns can often be ignored or mistranslated. For example, the airlines 'Northwest' or 'United' are directly translated into German equivalents when the proper noun should be preserved. While it may be entirely permissible to translate 'United' to 'Vereinigten', a native German speaker would be unlikely to produce a similar translation. Similarly, the English expression "dinner flights" can be directly translated to German as "Abendessenflug [dinner flight]",

but “Flug zur Abendszeit [evening flight]” is considered a more natural phrasing in German dialogue. Training on data with such errors limits the parser from generalising to utterances from native speakers.

Contextual Inference Ambiguity in the English input can lead to incorrect translations. For example, word “圣保罗” is a phonetic phrasing for both “São Paulo” and “‘St. Paul”. A professional translator could resolve this ambiguity by specifying country e.g., “巴西圣保罗 [Brasil São Paulo]” or “美国圣保罗 [USA St. Paul]”. However, MT can fail in resolving this ambiguity in both TRANSLATE TRAIN and TRANSLATE TEST scenarios, consequently requesting an incorrect flight destination.

Tone and Formality Another issue is the diverging politeness and tone between an MT system and a native speaker. We observe a specific issue when MT fails to model equivalent formality to a native speaker of ZH. Many utterances in MultiATIS++SQL ZH begin with “请 [qǐng]”, a politeness marker used to denote a request. No MT system we evaluate consistently also uses this politeness marker—leading to poor comprehension when present during inference.

3.4 Related Work

There is ongoing interest in machine translation as data augmentation for cross-lingual semantic parsing. Duong et al. (2017a) use TRANSLATE TRAIN to develop a joint parser for English and German notably identifying that MT enables models to better interpret *code-switched* text containing multiple languages from bilingual speakers. Sherborne et al. (2020) reports the initial version of our findings in this chapter promoting the use of MT as paraphrasing for cross-lingual parsing. Our contributions motivated **follow up work** researching localising open-domain knowledge base parsers (Moradshahi et al., 2020) and SLU parsers (Xia and Monti, 2021) to new target languages.

Recent work has highlighted that translation can be iteratively improved within the parsing pipeline, rather than queried once before training. Dou et al. (2023) create a multilingual version of Spider (Yu et al., 2018a) by iteratively translating a sentence, evaluating if the sentence can be parsed, and returning to human annotators for improvement and revised fluency. This process produces a multilingual model which can accurately parse fluent text from native speakers. Similarly, Li et al. (2023) use active learning to augment machine-translated data with gold-standard annotations for

examples where the model fails to predict an accurate LF. Iteratively improving on these 'harder' examples improves the cross-lingual generalisation capability of the model.

Adjacent work has considered whether Large Language Models (LLMs) are capable of similar data augmentation. The previously discussed TaF (Nicosia et al., 2021) and CLASP models (Rosenbaum et al., 2022) compare an LLM to specialised MT system for the TRANSLATE TRAIN scenario. This work identifies that both LLMs and MT engines are similarly competitive for silver-standard data generation. Shi et al. (2022) propose to use a vector database indexing multilingual parsing examples for retrieval augmentation in cross-lingual semantic parsing. The database is used to retrieve semantically similar examples to an utterance, and these form a single prompt for in-context learning using the Codex LLM (Chen et al., 2021). This result identifies the benefit of augmenting prompt-based inference with a semantically similar context in multiple languages.

3.5 Summary

In this chapter, we evaluate how machine translation can play a role in cross-lingual transfer for semantic parsing. We evaluate both TRANSLATE TEST and TRANSLATE TRAIN, and identify how MT systems can be used to produce multiple paraphrases in a target language. We propose methods for ensembling these paraphrases under the assumption that increasing the diversity of surface form structure improves a semantic parser trained with synthetic data. Our main contribution is FATES, a model for encoding parallel inputs from different MT systems, modelling each "translationese" dialect of a target language independently. Experimental results highlight that our hypothesis on the adequacy of machine translation for semantic parsing is rejected—building a parser with MT is not equivalent to using gold data in any language. Therefore, we argue that MT does not solve our overarching problem of building a parser for new languages with minimal resources. Our experiments suggest that data *quality* is critically important e.g., the observed improvement from introducing English data. In Chapter 4, we pivot to explore if high-quality data from *alternative* tasks, e.g., translation or language modelling can be used for cross-lingual semantic parsing exploiting existing gold data without demanding annotations in target languages.

Chapter 4

Zero-Shot Cross-lingual Semantic Parsing

In the previous chapter, we examined how machine translation can mitigate data scarcity for parsing languages beyond English. We observed that automatic translation is not a universal substitute for target language training data. Our error analysis highlights that machine translation struggles to fluently represent native-speaker utterances. We also identified that sampling data from synthetic distributions is insufficient alone for improving cross-lingual representation similarity. Additionally, this approach to semantic parsing is reliant upon translation quality, which often suffers for lower-resource languages (Ko et al., 2021). Given these reliability and quality concerns, machine translation is a suboptimal strategy for the goals of our case study.

In Chapter 2, we conjecture that latent representation alignment is a sufficient condition for cross-lingual transfer in semantic parsing. If we remove silver-standard data strategies for this goal, we now require a new source of suitable data. We maintain that sampling the target language distribution for training data is beyond feasible. However, there already exists multilingual corpora for adjacent natural language understanding tasks such as question answering, machine translation, or language modelling. This data is easy to collect and fluently represents native speakers of target languages. If we can exploit this readily available data for cross-lingual representation alignment—we can circumvent the requirement for semantic-parsing training data entirely.

In this chapter, we consider exploiting adjacent gold-standard target language corpora to learn cross-lingual representation alignment. By learning this alignment from other tasks, we anticipate cross-lingual semantic parsing without any task-specific training data. To achieve this, we introduce a multi-task objective combining monolingual

semantic parsing from English with *auxiliary* tasks (language modelling, translation, and language classification) training on existing multilingual corpora. By removing machine translation to focus on only gold data, we expect to overcome the fluency issues from the previous chapter. For our case study, this approach combines only the available English data and publicly available multilingual corpora to train the parsing model.

This chapter considers zero-shot cross-lingual parsing wherein the parser only observes target language semantic parsing inputs during inference i.e., 'zero-shot' parsing. In Chapter 3, we examined latent similarity using only pre-trained cross-linguistic information (i.e., the 'EN Only' baseline). We observe that pre-training is insufficient for generalising to all desired target languages. Analysis in Section 3.3.6 identifies that languages occupy different latent subspaces in the monolingual lower bound, the zero-shot baseline, and FATES. These representation spaces must be better aligned for cross-lingual transfer through representation alignment.

This chapter evaluates the hypothesis **auxiliary multilingual tasks and data can induce cross-lingual representation alignment without target-language training data**. We propose ZEUS: a **Ze**ro-shot **U**niveral **S**emantic parser. ZEUS uses multi-task learning to jointly learn semantic parsing, language modelling, translation, and language classification. We expect that jointly learning on these tasks optimises the latent representation space for improved cross-lingual similarity during fine-tuning. By improving latent cross-lingual similarity, a decoder trained to generate logical forms from English will better interpret an encoding from a target language utterance. The net effect will be cross-lingual logical form prediction from languages without task-specific training data. Additionally, removing machine translation from the parser removes any failures in translation or word alignment propagating to the parser. Experimental results on MULTIATIS++SQL and MTOP validate the benefit of multilingual auxiliary tasks for latent cross-lingual similarity. We observe that ZEUS significantly improves target language parsing beyond FATES or zero-shot methods without representation alignment.

4.1 Problem Formulation

4.1.1 Representation Alignment for Cross-lingual Transfer

We expect variations in any natural language's surface forms to express the same underlying semantics (Fodor, 1975). Figure 4.1 illustrates how errors in latent alignment

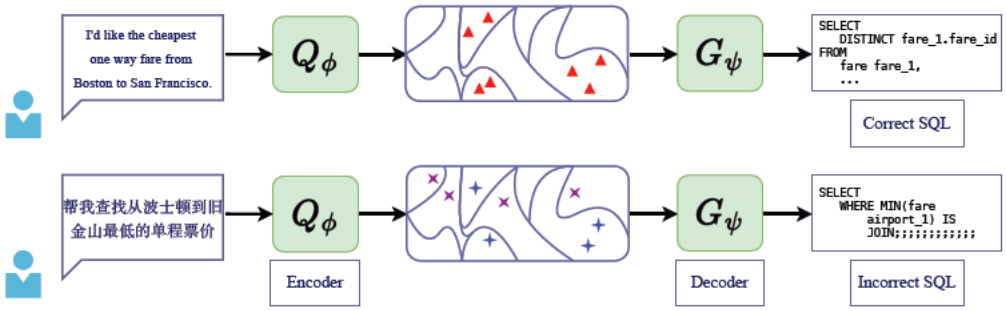

Figure 4.1: Accurate cross-lingual semantic parsing requires alignment of latent semantic representations across languages. The encoder generates a representation of the English utterance (red triangle) to condition upon during decoding. Producing the same logical form from the equivalent Chinese utterance requires a similar encoding. However, without cross-lingual alignment, the representation may partially match (blue star) or not at all (purple cross), leading the decoder to generate an inaccurate, ill-formed query.

make cross-lingual transfer difficult in semantic parsing. For this task, the equivalent semantics of different natural language inputs, x, map to an equivalent output, y, expressing equivalent underlying semantics. We frame that the encoder must map different x to the same latent representation to guarantee generating the same y from any natural language. Consider parallel encodings E_{EN} and E_l from parallel input sentences in Equation (2.19) and Equation (2.20) respectively (restated below). A decoder, G_ψ, trained only with English data is *likely* to correctly interpret E_{EN} to generate an accurate prediction $\hat{y}_{\text{EN}}$ in Equation (2.21). However, this decoder is *unlikely* to correctly interpret E_l in Equation (2.22) to generate an accurate prediction $\hat{y}_l$ equivalent to $\hat{y}_{\text{EN}}$. As the decoder is deterministically interpreting the given latent representation, we isolate alignment as an encoding challenge: desiring $E_l \approx E_{\text{EN}}$ by optimising Q_ϕ. Our aim is to produce *language agnostic* representations from Q_ϕ to produce equivalent outputs. This chapter addresses inducing this encoding behaviour without parallel semantic parsing data or machine translation.

$$E_{\text{EN}} = Q_\phi(x_{\text{EN}}) \quad \text{Latent encoding from input in English} \quad (2.19)$$

$$E_l = Q_\phi(x_l) \quad \text{Latent encoding from input in } l \quad (2.20)$$

$$\hat{y}_{\text{EN}} = G_\psi(E_{\text{EN}}) \quad \text{Decode encoding of English to logical form} \quad (2.21)$$

$$\hat{y}_l = G_\psi(E_l) \quad \text{Decode encoding of } l \text{ to logical form} \quad (2.22)$$

$$x_{\text{EN}} \neq x_l \quad \text{Equivalent semantics in languages English and } l \quad (2.23)$$

$$Q_\phi(x_{\text{EN}}) \approx Q_\phi(x_l) \quad \text{Approximately similar latent encodings} \quad (2.24)$$

$$\hat{y}_{\text{EN}} = \hat{y}_l \quad \text{Parse to the same logical form} \quad (2.25)$$

A zero-shot semantic parser is expected to produce the same logical form output from English utterance and any zero-shot target language i.e., Equation (2.25). Equation (2.24) frames the condition for this success as approximately equivalent encodings from parallel utterances in different languages. We expect parsing into inaccurate or illogical logic without satisfying these conditions. However, multiple barriers make achieving the ideal scenario non-trivial. First, a parser must overcome unfamiliarity with utterance-style data in each target language. The quantity and style of data used for pre-training are highly variable across target languages. For example, MBART50 (Tang et al., 2021) is pre-trained using 45 million sentences in German from multiple sources, but only 50,000 sentences in Portuguese (0.11% of German) sourced from transcripts of TED Talks (Qi et al., 2018). We argue that this variability provides no guarantee that the semantics of any target language are accurately represented within a pre-trained model. Second, our analysis in Section 3.3.6 raises scepticism that sufficient data encourages similarity in the latent representation space. Adapting a model's latent space to a *semantically distributed* structure (i.e., the same semantics are close in latent space) is needed to guarantee zero-shot cross-lingual transfer with language-agnostic representations.

4.1.2 Modelling Distributions for Auxiliary Tasks

The previous chapter considered using *synthetic* samples, $\mathcal{S}_l^{\text{MT}}$ to generalise to $\mathcal{D}_l$. Here we use samples from *alternative* distributions for the same generalisation goal. These alternatives are gold standard data sources for other tasks without semantic parsing annotation. We conceptualise these samples, $\mathcal{T}_l$, as Equation (4.1) from some underlying data distribution $\mathcal{U}_l$ for task distribution $\mathcal{U}$ in language l as Equation (4.2). Chapter 3 simplified the distributions to only natural language-logical form pairs. Here,

$\mathcal{T}_l$ abstractly describes the sample for any task other than semantic parsing. For language modelling, $\mathcal{T}_l$ describes unlabelled sequences, $x = (x_1, x_2, \ldots, x_T)$ of T tokens, where y is the next token in x. For translation, the (x, y) pairs in $\mathcal{T}_l$ are input and output sentences.

$$\mathcal{T}_l = \{x_i, y_i\}_{i=1}^{n} \tag{4.1}$$

$$\mathcal{T}_l \sim \mathcal{U}_l \tag{4.2}$$

4.1.3 Auxiliary Tasks

We adopt a multi-task sequence-to-sequence model (Luong et al., 2016) combining logical form generation with additional strategies for representation alignment. First, we fine-tune this encoder to encode and decode target languages through a language modelling objective (Section 4.1.3.1). We also propose further improvement using a translation objective where available (Section 4.1.3.2). Second, we directly penalise the ability to discriminate the language from the latent representation (Section 4.1.3.3). Our intuition is that our auxiliary losses minimise cross-lingual variance in latent encoding space by optimising for language-agnostic representations.

4.1.3.1 Target Language Adaptation with Unlabeled Data

The first auxiliary task is to continue the pre-training objective of language model denoising reconstruction (Lewis et al., 2020a). We introduce an additional decoder for natural language, G_ω, in parallel to the existing logical form decoder G_ψ. This forward path, encoder Q_ϕ and decoder G_ω, is trained to encode a corrupted input $\tilde{x}$ and reconstruct x during decoding. Pre-trained models are generally trained on web-sourced corpora of declarative text (e.g., Wikipedia). Our intuition is that fine-tuning a model on utterances data (i.e., 'questions' as discussed in Section 2.1.1) can improve parsing by increasing exposure to the style of target language test data. This is motivated by *domain-adaptive pre-training* (DAPT) an approach to fine-tuning continuing a pre-training objective using only data relevant to the desired end task (Gururangan et al., 2020). In our context, we hypothesise that fine-tuning a multilingual encoder on multilingual question-style data improves parsing by reducing the loss for target language utterances in this style.

Decoder G_ω reconstructs the input by autoregressively generating each token of x, similar to logical form generation described in Section 2.2.1. For each time step t,

the decoder predicts x_t using Equation (4.4), as a distribution over the vocabulary conditioned on previous outputs and the encoding, E, of noised input $\tilde{x}$. The noised input $\tilde{x}$ is sourced from x (Equation (4.3)) but intentionally corrupted (discussed below). Similar to Section 2.2.2, the loss is the cross-entropy between the prediction of x, as Equation (4.5), and the source x as Equation (4.6).

$$\tilde{x} = \text{Noise}(x) \tag{4.3}$$

$$p(x_t|x_{<t}, \tilde{x}) = \text{softmax}(G_\omega(x_{<t})) \tag{4.4}$$

$$p(x|\tilde{x}) = \prod_{t=1}^{T} p(x_t|x_{<t}, \tilde{x}) \tag{4.5}$$

$$\ell_{\text{NL}} = -\sum_{x} \log p(x|\tilde{x}) \tag{4.6}$$

The noising function (Equation (4.3)) intentionally corrupts x challenging the model to predict the original x with some tokens removed, shuffled or replaced. The denoising objective reduces the model's dependency on specific linguistic patterns by randomly removing, replacing or masking words. This improves generalisation as a regulariser, similar to the role of dropout reducing reliance on specific activations (Srivastava et al., 2014). This objective produces highly generalisable models such as BART (Lewis et al., 2020a) and MBART (Liu et al., 2020). This auxiliary task requires only unlabelled text for training as the input is also the output. Therefore, this objective improves the target language modelling in the encoder using only abundant monolingual data. As discussed in Section 2.1.1, utterances for semantic parsing often model natural speech rather than written language. A competent parser must be sensitive to the natural variation in expression for utterance data. We consider this challenge when selecting monolingual data for this task in Section 4.2.1.

4.1.3.2 Target Language Translation with Bitext data

Pre-training on multiple languages enables a model to learn some shared multilingual information (e.g., word order, syntax or grammatical gender) for language understanding (Liu et al., 2020). However, this lacks a mechanism for explicit cross-lingual semantic equivalency (Tang et al., 2021). As described in Section 2.2.1, this deficiency motivates the creation of MBART50 by continuing pre-training of MBART on a multi-language machine translation objective.

We presume that our system can similarly benefit from explicit supervision of equivalent cross-lingual semantics by introducing a translation objective to the generation

decoder, G_ω, using available bitext. We consider bitext corpora to contain *latent* labels of semantic equivalence from different utterances. We expect that introducing this latent supervision improves the similarity between latent representations of equivalent meaning. This task is introduced by modifying the denoising objective described in Section 4.1.3.1. Now, the decoder either generates the original x_l input (denoising) or the English parallel sentence (translation). To select between objectives, we uniformly sample a value r (Equation (4.7)) and use translation if r is above some threshold τ in Equation (4.8). We optimise τ as a ratio between denoising and translation during training. Equation (4.9) describes the probability of generating y' dependent on $\tilde{x}$ and r. The translation objective uses the cross-entropy between the predicted $\hat{x}_{\text{EN}}$ and gold x_{EN} similar to Equation (4.6).

$$r \sim \mathcal{U}(0,1) \tag{4.7}$$

$$y' = \begin{cases} x_l & 0 \leq r < \tau \\ x_{\text{EN}} & \tau \leq r \leq 1 \end{cases} \tag{4.8}$$

$$p\left(y'|\tilde{x},r\right) = \prod_{t=1}^{T} p\left(y'_t|y'_{<t},\tilde{x}\right) \tag{4.9}$$

The denoising or translation objective (collectively the 'generative' objectives) improve the quality of target language representations by increasing exposure to these surface forms. This task improves accuracy in semantic modelling of under-resourced languages and explicitly encourages latent semantic equivalence between parallel utterances. These goals minimise the loss for target languages to reduce any expectation that the latent encoding will be inaccurate or difficult to decode i.e., the motivating example in Figure 4.1. We now describe an additional task directly optimising for latent language similarity.

4.1.3.3 Discriminating Language from the Latent Representation

We introduce a penalty on language discriminability from the latent representation to discourage representations from forming monolingual clusters. The encoding is input into a shallow *language prediction* linear classifier, LP, optimised to predict the input natural language as a label. During training, we use *gradient reversal* (Ganin et al., 2016) to optimise the encoder, Q_ϕ, *against* this objective. This makes the encoder and the LP subnetwork adversarial. The subnetwork discriminates the language only from

the latent representation. The encoder is adversarially optimised to produce encodings where the language is less discriminable. The intuition is that the encoder now observes a gradient directly *opposing* the optimisation of the LP network for language discriminability. Gradient reversal generates more language-agnostic representations targetting our representation alignment objective. This technique was originally proposed for domain adaptation (Ganin et al., 2016), identifying that penalising the domain specificity of representations improves the domain transferability of a model. In our case, we consider different input languages as domains.

For some Transformer encoding, $E \in \mathbb{R}^{T \times d}$ over T time steps with dimensionality d, we pool the representations over time to produce a single vector $\bar{E}$ in Equation (4.10). This is input to the **L**anguage **P**redictor: a shallow linear classifier (Equation (4.11)) with parameters $W_{\text{LP}} \in \mathbb{R}^{L \times d}$ and $b_{\text{LP}} \in \mathbb{R}^{L}$. The network predicts language l, from original input x_l, of L possible target languages using the distribution defined in Equation (4.12).

$$\bar{E} = \frac{1}{T} \sum_{t} E_t \tag{4.10}$$

$$\text{LP}(x) = W_{\text{LP}} x + b_{\text{LP}} \tag{4.11}$$

$$p(l|x_l) = \text{softmax}(\text{LP}(\bar{E})) \tag{4.12}$$

We follow the approach of Ahmad et al. (2019) in using a shallow linear classifier for language prediction. We observe that adding more layers in the LP submodel introduces undesirable training instability and poorer cross-lingual parsing. The language prediction submodel, LP is trained using a categorical cross-entropy loss in Equation (4.13). Equation (4.14) describes how the gradient is reversed during the backward pass propagating from the LP model to the encoder with parameters ϕ. This directly optimises the encoder *against* language prediction, with a net effect reducing language specificity of encoded latent representations.

$$\ell_{\text{LP}} = -\sum_{x} \log\ p(l|x) \tag{4.13}$$

$$\frac{\partial \ell_{\text{LP}}}{\partial \phi} = -\frac{\partial \ell_{\text{LP}}}{\partial \text{LP}} \tag{4.14}$$

Our intuition is that language-agnostic representations improve the recognition of target language encodings for the logical form decoder to generate logical forms without sampling target language training data.

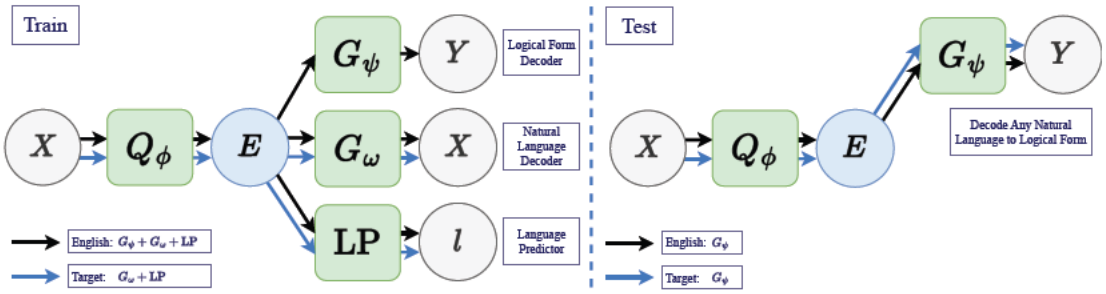

Figure 4.2: ZEUS, is a **Ze**ro-shot **U**niveral **S**emantic parser. ZEUS augments the encoder-decoder semantic parser with auxiliary objectives designed to optimise cross-lingual representation similarity. The Encoder, Q_ϕ, generates a representation, E, input to the logical form decoder, G_ψ, reconstruction decoder, G_ω, or language prediction classifier, LP. During training (left), English is input to all objectives and additional languages are incorporated using *only* the additional objectives $\{G_\omega$, LP$\}$. During test (right), logical forms are predicted using G_ψ from utterances in all languages.

4.1.4 Zero-shot Transfer from Multi-task Modelling

The model, ZEUS, combines the auxiliary objectives for a multi-task model simultaneously optimising for all tasks: semantic parsing, denoising reconstruction, translation and language prediction. Figure 4.2 outlines the complete model: a single multilingual encoder Q_ϕ; a logical form decoder, G_ψ; a natural language decoder, G_ω; and the language predictor, LP.

During training, an English query is encoded and input to all objectives for combined loss as ℓ_{LF}, ℓ_{NL}, and ℓ_{LP}. ℓ_{LF} is the loss for generating the paired logical form (Equation (2.9)), ℓ_{NL} is the loss from denoising to reconstruct the input (Equation (4.6)), and ℓ_{LF} is the loss for language prediction (Equation (4.13)). For target languages without (x, y) pairs, the encoded utterance is input only to the auxiliary tasks for a combined loss as ℓ_{NL}, and ℓ_{LP}. ℓ_{NL} is now the loss of either denoising or translation according to sampled probability r and threshold τ.

Each output loss back-propagates the gradient signal from the respective objective function during the backward pass. For the encoder with parameters ϕ, the combined gradient is Equation (4.15) where α_{NL} and α_{LP} are scalar loss weighting hyperparameters and $\lambda_{\text{LP}}(t)$ is an additional scheduled weighting on the language prediction gradient (Ganin et al., 2016). The language predictor contributes minimally to the encoder during early training—when this submodel is poor at language prediction. The gradient contribution increases during training to mitigate gradient noise during

early training. Without λ_{LP}, the noisy early predictions from ℓ_{NL} dominate the learning process to produce an overall poorer parser. Equation (4.16) controls this schedule for the weighting at training step t of T total steps. Equation (4.16) uses a hyperparameter γ to control the rate of increase in $\lambda_{LP}(t)$ overtraining.

$$\frac{\partial \ell}{\partial \phi} = \frac{\partial \ell_{LF}}{\partial \phi} + \alpha_{NL} \frac{\partial \ell_{NL}}{\partial \phi} - \lambda_{LP} \alpha_{LP} \frac{\partial \ell_{LP}}{\partial \phi} \tag{4.15}$$

$$\lambda_{LP}(t) = \frac{2}{1 + e^{-\gamma \frac{t}{T}}} - 1 \tag{4.16}$$

During inference, an utterance is encoded and *always* input to G_ψ to predict a logical form, $\hat{y}$, regardless of test language, l.

4.2 Experiments

4.2.1 Datasets

MultiATIS++SQL In this chapter, we use only the English language utterance-logical forms for training data. Each natural language utterance maps to an executable SQL logical form. We expect a parser to handle the target language inputs without observing task-specific training data in each language. Target languages are incorporated using the auxiliary data outlined in Section 4.2.2. Section 2.3.1 details a complete description of MultiATIS++SQL with input-output examples shown in Table 2.5. We use the MultiATIS++SQL multilingual test set for evaluating cross-lingual transfer from English (EN) to French (FR), Portuguese (PT), Spanish (ES), German (DE), and Chinese (ZH).

MTOP We similarly use MTOP using only the English language utterance-logical form training data. Each natural language utterance maps to a TOP-LF logical form hierarchically representing the intent and slot features of the utterance. Section 2.3.2 details a complete description of MTOP with input-output examples shown in Table 2.6. We use the MTOP multilingual test set for evaluating cross-lingual from English (EN) to French (FR), Spanish (ES), German (DE), Hindi (HI), and Thai (TH).

4.2.2 Data for Auxiliary Tasks

ZEUS is trained for semantic parsing using only the training data pairs for English. We improve latent cross-lingual alignment by sourcing parallel (bitext) and non-parallel

Data Source	Type	Relevant Languages	# Examples
MKQA (Longpre et al., 2021)	Questions (*Q*)	EN, FR, PT, ES, DE, ZH, TH	10,000
MLQA (Lewis et al., 2020b)	Questions (*Q*)	EN,ES, DE, ZH,HI	EN 12,738, DE 5029, ES 5754, ZH 5641, HI 5425
ParaCrawl 7.1 (Bañón et al., 2020)	Declarative (*D*)	EN, FR, PT, ES, DE, ZH	14 million (ZH)-296 million (ES)
CCAligned (El-Kishky et al., 2020)	Declarative (*D*)	EN, FR, PT, ES, DE, ZH, HI, TH	8 million (HI)-92 million (DE)

Table 4.1: Data and language sources for latent cross-lingual alignment with ZEUS. Data is sourced for languages: English (EN), French (FR), Portuguese (PT), Spanish (ES), German (DE), Chinese (ZH), Hindi (HI), and Thai (TH). We primarily use questions as data but also use declarative web-scraped text. For questions, we use all available data in each language. For declarative text, we match this data quantity with a random subsample of data from each language.

(unlabelled text) corpora for the auxiliary tasks described in Section 4.1.3. We use the same data sources for all auxiliary tasks simultaneously during training. All auxiliary data resources are outlined in Table 4.1.

Question Data Domain-adaptive pre-training is effective at improving language modelling when fine-tuning on curated corpora relevant to the desired end task (Gururangan et al., 2020). Therefore, we do not continue training on a large text corpus such as C4 (Raffel et al., 2020) and instead focus on available data for parsing-adjacent tasks. We primarily use input utterances from the MKQA dataset (Longpre et al., 2021) as our auxiliary alignment corpus. MKQA is a professional translation of 10,000 questions from the English NaturalQuestions dataset for question answering (Kwiatkowski et al., 2019) into 26 languages. MKQA is suitable for our objectives as: (i) the data is wholly parallel to control for the number of examples in each language; (ii) the utterances can be used as unlabelled text or bitext between each language and English; and (iii) the utterances in MKQA represent natural questions from native speakers of target languages. For languages not included in MKQA, we augment our dataset with examples from MLQA (Lewis et al., 2020b) with similar quality. MLQA is partially parallel with different utterances translated from English into different target languages. These desirable properties enable ZEUS to fine-tune towards the style of language in our test data providing fluent variation in utterance style.

Declarative Data We primarily examine if auxiliary tasks using data matching the syntactic style as our test data (i.e., questions). However, this type of data, either unlabelled or as bitext, can be challenging to source for some lower-resource languages. Therefore, we also examine if ZEUS can improve cross-lingual transfer using arbitrary scraped web text. For many natural languages, this is abundant and easy to source.

Data is sourced from ParaCrawl 7.1 (Bañón et al., 2020) and CCAligned (El-Kishky et al., 2020). These are large-scale multilingual corpora with document- and sentence-level alignment between English and >20 languages. Parallel alignments are generally noisy and can produce false pairings (see Table 4.2). We generally consider this text as 'declarative' i.e., statements not requiring a response. This contrasts with interrogative questions and imperative instructions, which typically demand a response (see Section 2.1.1). We randomly sample these corpora for the same quantity of data as the available data in MKQA and MLQA. We similarly use the aligned English equivalent for the translation objective. We also note that these samples are not parallel *between* target languages: the samples for any two languages are not parallel and there is no similarity between the respective aligned English samples.

For question data, we use 60,000 utterances from MKQA for MultiATIS++SQL, and 50,000 utterances from MKQA and 5,425 Hindi utterances from MLQA for MTOP. For declarative data, we use an equivalent quantity sampled from ParaCrawl and CCAligned. Examples from each surface form style are shown in Table 4.2. We argue that using data in the question style will benefit cross-lingual transfer for ZEUS more than using declarative text. However, as questions may not always be available, we experiment with both surface-form structures to explore more ideal or pragmatic methods of training ZEUS.

4.2.3 Experimental Setting

4.2.3.1 Setting and Comparison

Our results in Chapter 3 highlight that multilingual modelling (i.e., one model for N languages) is either comparable or superior to monolingual modelling (i.e., one model for each of N languages). Therefore, we only implement multilingual modelling henceforth for Chapters 4 to 6.

MULTILINGUAL **Gold** A multilingual Transformer is trained on the union of all professionally translated data. As in Chapters 3 to 4, this represents the **upper bound**

(a) Questions from MKQA (Longpre et al., 2021)	
EN	Who coined the phrase “let sleeping dogs lie”?
FR	Qui a inventé la phrase let sleeping dogs lie?
PT	Quem inventou a frase let sleeping dogs lie?
ES	Quién acuño la frase “let sleeping dogs lie”?
DE	Wer prägte die Phrase, lass schlafende Hunde liegen.
(b) Declarative Text from ParaCrawl 7.1 (Bañón et al., 2020)	
EN	Ready to wear for: a special evening, the office, a day at the park.
FR	Prêt à porter pour le bureau, une séance de magasinage, une activité familiale.
EN	Travel from Portici to Pozzuoli and discover Campania city of 45.6 Thousand inhabitants.
FR	Voyagez de Nice à Turin et découvrez cette ville du Piemonte de 870 Thousand habitants.
EN	All rooms feature a balcony, most of which with a sea view.
ZH	所有客房都设有一个阳台,其中大部分享有海景。
EN	It should be possible to write an interpreter for the language.
ZH	应当被设计成安全地执行远端代码。

Table 4.2: Data examples for (a) MKQA (Longpre et al., 2021) and (b) ParaCrawl 7.1 (Bañón et al., 2020). MKQA is a parallel translation from English into 26 languages—the sample from each language represents the same semantic meaning. ParaCrawl contains data scraped from the web and automatically aligned across languages. This can result in erroneous cross-lingual alignment highlighted in red. Each bitext pair is unrelated to samples from other languages.

for our model using all available data without few-shot constraints.

EN Only A monolingual Transformer is trained on only English training data. This model is evaluated on the target language test data with no translation. This is our lower-bound as the baseline model for zero-shot parsing with **no representation alignment**.

TRANSLATE TEST A monolingual Transformer is trained on source English data. Machine translation is used to translate test data from target languages into English. Logical forms are predicted from translated data using the English model. This is the **silver-standard** baseline method for Chapter 3. We report this baseline using OPUS translation (Tiedemann and Thottingal, 2020).

TRANSLATE TRAIN Machine translation is used to translate English training data into each target language as described in Chapter 3. A monolingual Transformer is trained on translated training data and logical forms are predicted using this model. This is the **silver-standard** baseline method for Chapter 3. We report this baseline using OPUS translation (Tiedemann and Thottingal, 2020).

We also compare the best FATES model from Chapter 3 to contrast between our best solution using machine translation, and this zero-shot method without translation.

4.2.3.2 Model Training

ZEUS follows the Transformer encoder-decoder setup from Section 2.2.1. The encoder, Q_ϕ, is pre-trained using the encoder parameters from the MBART50 pre-trained model (Tang et al., 2021). ZEUS uses two identical Transformer decoders for logical forms (G_ψ) and natural language G_ω with separate parameters. Both decoders are trained from scratch—we did not observe any benefit to initialising either decoder using pre-trained parameters. The language predictor submodel is also trained from scratch. One epoch of training consumes all parsing and auxiliary data where each step samples a single batch for either semantic parsing or auxiliary tasks.

Hyperparameters were chosen by training a reference model for parsing English utterances and selecting the system with minimum validation loss. The noise function for the denoising task uses token masking (i.e., a token is randomly replaced with "[mask]"). The noising function samples an integer value v from $U(0, \nu)$ and masks v tokens in an input sentence. The maximum masking factor, ν was optimised in the range $[0, 10]$ with an optimal setting of 4. Loss weightings $\alpha_{\{\text{LP, NL}\}}$ were selected from range $\{1,\ 0.5,\ 0.33,\ 0.25,\ 0.1\}$. The final setting is $\alpha_{\text{LP}} = 0.33$ and $\alpha_{\text{NL}} = 0.1$. The scaling factor for language predictor scheduling, γ in Equation (4.16), was optimised from range $\{0,\ 5,\ 10,\ 20,\ 40,\ 50,\ 100\}$ and set to 20. This setting corresponds to ℓ_{LP} reaching 95% of the maximum value after approximately 18% of training progress.

4.3 Results

4.3.1 Is Zero-Shot Parsing Competitive with the Upper Bound?

The first comparison is between zero-shot cross-lingual transfer and the upper bound using gold-standard professional translation into all target languages. This provides a contrast between the lowest-resource case (English data and a pre-trained model) and

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
MULTILINGUAL Gold	74.9	74.2	73.0	70.4	74.6	73.7	73.2
EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
ZEUS (best)	74.4	72.3	69.7	68.5	69.0	69.2	**69.7**

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
MULTILINGUAL Gold	75.5	69.7	72.4	67.9	65.5	54.6	66.0
EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
ZEUS (best)	77.5	66.2	67.4	64.2	59.4	47.7	**61.9**

(b) MTOP

Table 4.3: Results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy contrasting between: training on gold-standard translations in every target language (MULTILINGUAL Gold); training only on the English semantic parsing dataset (zero-shot 'EN Only'); and the best ZEUS model. ZEUS uses auxiliary tasks to improve cross-lingual transfer and improves on the unaugmented zero-shot parsing performance for all languages. The significant best model, exluding upper bound, is bolded.

the highest-resource case (high-quality data in all languages). If a zero-shot approach is competitive with the upper bound, this could circumvent any annotation requirements for cross-lingual transfer.

As discussed in Chapter 3, the zero-shot lower bound without representation alignment effort does not approach the upper bound without data scarcity. Table 4.3 identifies that the best ZEUS model significantly improves zero-shot parsing to approach the upper bound without target language parsing data. Using auxiliary tasks positively improves zero-shot performance for all languages. The best version of ZEUS uses interrogative and declarative text ($Q+D$) as auxiliary data. Distant languages improve the most using ZEUS parsing MultiATIS++SQL ZH improves by 41.7% and parsing MTOP HI or TH improves by 55.4% and 63.9% respectively. For similar languages (e.g., FR), the best ZEUS model is now competitive with the upper bound, offering an alternative to annotation using only English task-specific data. ZEUS is only -3.5% below the MultiATIS++SQL upper bound, and only -5.0% upper bound for MTOP. This represents an 86.2% and 84.5% reduction in the error between zero-shot and upper-bound parsing for MultiATIS++SQL and MTOP respectively. This improvement supports our hypothesis that multilingual auxiliary tasks improve the cross-lingual transferability of a monolingual task.

4.3.2 Is Zero-Shot Transfer Superior to Machine Translation?

We compare between ZEUS and the machine translation-based methods in Chapter 3 in Table 4.4. As above, we compare the unaugmented zero-shot model and the best ZEUS model to an example of TRANSLATE TEST or TRANSLATE TRAIN and the best FATES parser. Zero-shot modelling ('EN only') performs below any translation method for MultiATIS++SQL but above TRANSLATE TRAIN for MTOP. In Chapter 3, we raise that poor cross-lingual word alignment resulted in weaker parsing. Even without auxiliary tasks, zero-shot parsing is above training with translations if the translation system accumulates errors. We note that ZEUS does not surpass the silver-standard LLM-based methods we discuss in Chapter 3 for MTOP.

We observe that the best ZEUS model produces a significantly improved parser than FATES. We infer that the latent alignment objectives and training without "translationese" issues reduce the error in adaptation to natural native-speaker utterances during inference. Notably, ZEUS improves for MTOP by $+12.2\%$ over TRANSLATE TEST and $+24.8\%$ over FATES. This suggests a modelling preference for zero-shot alignment

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
TRANSLATE TRAIN (OPUS)	—	56.8	39.1	51.8	60.4	59.6	53.5
TRANSLATE TEST (OPUS)	—	57.7	58.1	58.3	58.8	50.9	56.8
FATES (best)	74.9	70.5	69.2	62.4	68.7	66.5	67.5
EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
ZEUS (best)	74.4	72.3	69.7	68.5	69.0	69.2	**69.7**

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
TRANSLATE TRAIN (OPUS)	—	24.4	23.1	32.7	22.4	9.5	22.4
TRANSLATE TEST (OPUS)	—	44.9	63.1	39.1	47.1	54.2	49.7
FATES (best)	69.7	44.7	45.7	49.0	32.9	13.4	37.1
ZEN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
ZEUS (best)	78.8	66.8	67.9	64.6	61.0	49.2	**61.9**

(b) MTOP

Table 4.4: Results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy contrasting between machine translation, zero-shot cross-lingual transfer and the best ZEUS model. We compare to single source TRANSLATE TRAIN and TRANSLATE TEST using the OPUS MT system from Chapter 3. We also compare to the best parser FATES parser from Chapter 3 (FATES from Table 3.11). We contrast MT-based systems to zero-shot cross-lingual transfer without auxiliary tasks (training on English), and the best performing ZEUS parser discussed further in Section 4.3.3. The significant best result is bolded.

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
(i) EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
(ii) $LF + NL$ ($\tau = 0.5$)	77.7	62.7	54.9	58.2	61.1	51.2	57.6
(iii) $LF + LP$	76.3	57.2	53.7	51.8	58.6	44.1	53.1
(iv) $LF + LP + NL$ ($\tau = 0.5$)	74.4	72.3	69.7	68.5	69.0	69.2	**69.7**

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
(i) EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
(ii) $LF + NL$ ($\tau = 0.5$)	77.5	60.2	64.3	61.2	60.4	48.7	59.0
(iii) $LF + LP$	76.3	63.0	65.0	60.7	57.3	44.9	58.2
(iv) $LF + LP + NL$ ($\tau = 0.5$)	78.8	66.8	67.9	64.6	61.0	49.2	**61.9**

(b) MTOP

Table 4.5: Results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy ablating auxiliary objectives. Objectives are: natural language to logical form prediction (LF); natural language generation (NL) of either reconstruction or translation according to threshold τ; and language prediction from the latent encoding LP. We show: (i) training only on the English semantic parsing dataset without auxiliary tasks; (ii) training on semantic parsing and natural language objectives with 50% probability of denoising reconstruction or translation ($\tau = 0.5$); training on semantic parsing and language prediction; and the ZEUS model using all objectives. ZEUS benefits from all auxiliary tasks for the best accuracy in cross-lingual parsing. We discuss the setting of τ further in Section 4.3.4.

over translation when translation is a noisier pipeline. For MultiATIS++SQL, where machine translation does not introduce word-alignment issues, the improvement for ZEUS over FATES is a lesser 2.2%. Both methods are competitive here, but it may be preferable to rely on gold data rather than machine translation within our case study.

4.3.3 Which Auxiliary Objective Matters?

Ablations to the model are shown in Table 4.5, identifying the contributions of different objectives. As discussed above, zero-shot parsing without auxiliary objectives (Model (i)) is the weakest approach. This is unsurprising, as this approach uses only pre-trained

cross-lingual information without additional effort to improve similarity.

Model (ii) adds the auxiliary natural language task to the model. We set the threshold to $\tau = 0.5$ for a 50%-50% split between denoising reconstruction and translation objectives for this auxiliary task. This is discussed further in Section 4.3.4. We observe that allowing the model to fine-tune towards the target languages (without parsing data) improves transfer to all target languages by an average of +9.3% for MultiATIS++SQL and 25.3% for MTOP. This agrees with prior results for other tasks identifying the benefit of domain-specific adaptation during fine-tuning (Gururangan et al., 2020).

Model (iii) evaluates the language prediction classifier (*LP*) as the only auxiliary task. We observe that this method improves over the baseline by an average of +5.3% for MultiATIS++SQL and +24.5% for MTOP. During training, the language prediction accuracy for the validation dataset peaks at 87% after 14% progress and subsequently decreases to <8% beyond 30% of training. Language prediction accuracy for the test set is 7.2%. This suggests that this classifier correctly reduces the input language discriminability from the latent representation. Comparing objectives in (ii) and (iii), the benefit of the language prediction classifier is weaker than the natural language decoder tasks. We interpret that adaptation towards specific surface form patterns is more valuable for representation alignment than latent language discriminability.

The best version of ZEUS (iv) uses all auxiliary tasks with a cumulative benefit in multi-task modelling. Compared to (i)'EN only', the full model improves by a target language average of +21.9% for MultiATIS++SQL and +28.2% for MTOP. This result supports our claim that zero-shot parsing can be optimised using auxiliary objectives. We conjecture that this combination benefits from constructive interference—the language prediction loss promotes representation invariance in cooperation with multilingual generation tasks fine-tuning the encoder toward target languages test utterances.

4.3.4 Is Translation or Reconstruction More Beneficial for Parsing?

Section 4.3.3 contrasts between auxiliary objectives identifying a positive improvement including generation objectives. As outlined in Sections 4.1.3.1 to 4.1.3.2, the natural language decoder is used for either denoising reconstruction or translation from the target language into English. The ratio between either objective is controlled by threshold hyperparameter τ. We examine the influence on parsing by varying objectives by varying τ between zero and one in Figure 4.3 for MultiATIS++SQL and Figure 4.4

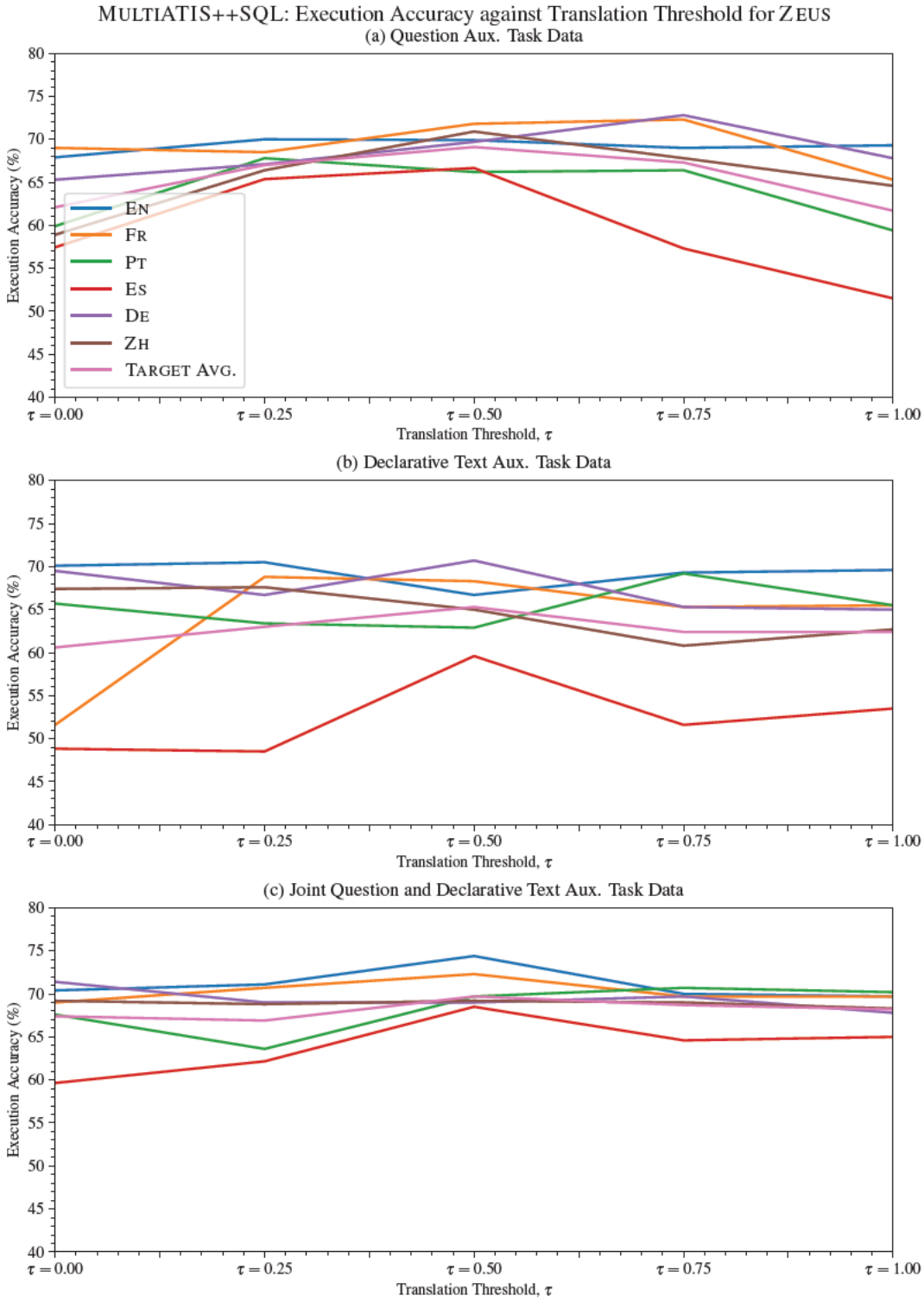

Figure 4.3: MultiATIS++SQL execution accuracy for ZEUS varying the probability threshold τ for reconstruction $r < \tau$ or translation into English $r \geq \tau$. We report results training ZEUS varying τ using (a) question data, (b) declarative text data, and (c) combined question data and declarative text. At each step, r is a uniform random sample, $r \sim \mathcal{U}(0, 1)$, to select a natural language task. We generally observe that $\tau = 0.5$ is approximately the best setting. This identifies that both translation and denoised reconstruction are useful tasks for ZEUS.

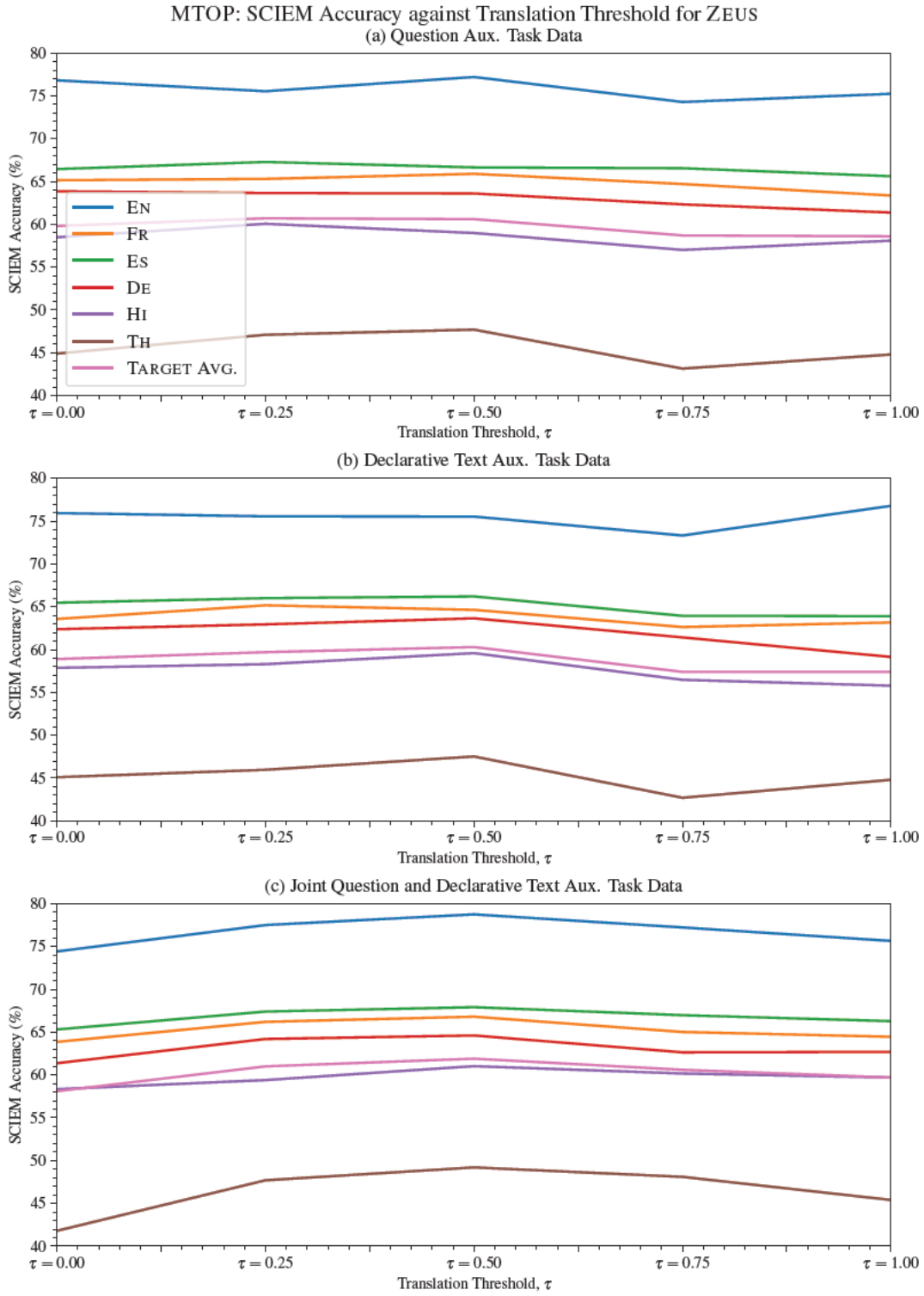

Figure 4.4: MTOP SCIEM accuracy for ZEUS varying the probability threshold τ for reconstruction $r < \tau$ or translation into English $r \geq \tau$. We report results training ZEUS varying τ using (a) question data, (b) declarative text data, and (c) combined question data and declarative text. At each step, r is a uniform random sample, $r \sim \mathcal{U}(0, 1)$, to select a natural language task. We generally observe that $\tau = 0.5$ is approximately the best setting. This identifies that both translation and denoised reconstruction are useful tasks for ZEUS.

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
Q	69.9	71.8	66.2	66.7	69.7	70.9	69.1
D	66.7	68.3	62.9	59.6	70.7	65.0	65.3
$Q+D$	74.4	72.3	69.7	68.5	69.0	69.2	69.7

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
Q	77.2	65.9	66.6	63.6	59.0	47.7	60.6
D	75.5	64.6	66.2	63.7	59.6	47.5	60.3
$Q+D$	78.8	66.8	67.9	64.6	61.0	49.2	61.9

(b) MTOP

Table 4.6: Results for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy contrasting between question (Q) or declarative text (D) style used for adaptation to target languages for ZEUS. We evaluate training ZEUS with only questions (Q), only declarative text (D), and combining both data sources $Q+D$. The combined strategy demonstrates superior accuracy, with Q only performingly competitively in most languages. The ratio between denoising and translation is fixed for all experiments ($\tau = 0.5$). Data sources for each surface form style is given in Table 4.1.

for MTOP. In general, the setting of $\tau = 0.5$ demonstrates the highest average accuracy. This setting of τ is used in all other results for ZEUS. For most languages, we do not identify a strong bias for either objective (i.e., the line is flat in the figure). We observe a benefit from combining objectives but either approach individually would yield competitive parsing. Similarly, there is no strong trend contrasting between auxiliary data styles (Q, D, or $Q+D$). In a novel setting without available bitext, the denoising reconstruction alone could be sufficient without implementing translation. We also note that separating these tasks into different decoders did not yield significant benefits but practically increased training time and GPU memory requirements.

4.3.5 Does Auxiliary Data Style Matter?

We examine the role of surface form style in adaptation towards target languages. Table 4.6 compares between training ZEUS using interrogative questions (Q), declarative text (D), and jointly on both sources ($Q+D$). We observe that adaptation towards

questions is superior to arbitrary declarative text, likely as this style is closer to the utterances in each test set. In addition, Figure 4.3 and Figure 4.4 also highlight the benefit of questions over declarative text is valid over every setting of τ. We note that ZEUS using declarative text performs above all FATES from Chapter 3—identifying that our method can robustly improve cross-lingual transfer if question data cannot be sourced for a new target language. This further supports the preference for zero-shot modelling over-reliance on synthetic data. We observe that the best ZEUS model for either dataset combines both data sources ($Q+D$). The improvement for combining data is mostly sourced from languages similar to English (FR, PT) with some less similar languages (ZH) demonstrating significantly lower performance on MultiATIS++SQL with the additional data than questions alone. This is likely because dissimilar languages share fewer features, and therefore benefit less from additional data over improved data quality.

4.3.6 How does Alignment Data Quantity Influence Transfer?

Previous findings in this chapter assume access to all data outlined in Table 4.1. We now consider if the quantity of data used for auxiliary tasks influences the ability for ZEUS to improve latent cross-lingual similarity. Figure 4.5 and Figure 4.6 demonstrates the parsing performance for MultiATIS++SQL and MTOP respectively when varying the auxiliary dataset size. For both datasets, we observe a positive trend in that increasing the corpora size improves cross-lingual parsing. For MultiATIS++SQL, the correlation between data quantity and improved parsing is more strongly correlated for questions (Pearson $\rho = 0.97$ $p < 0.01$) than for declarative text (Pearson $\rho = 0.85$ $p < 0.01$). This trend is similar for MTOP with Pearson $\rho = 0.78$ ($p < 0.01$) for questions and Pearson $\rho = 0.68$ ($p < 0.01$) for declarative text. The contrast between these correlations for Q and D reaffirms increased value in adapting towards questions over any available data. For joint data, the equivalent correlation is Pearson $\rho = 0.96$ ($p < 0.01$) for MultiATIS++SQL and Pearson $\rho = 0.82$ ($p < 0.01$) for MTOP. This trend is still positive, but we note that the performance using 100% of question data is above using 50% of joint $Q+D$ data. This suggests that including questions in auxiliary task corpora is critical for improvement and any additional gain from declarative text is a marginal benefit only from including more data during learning.

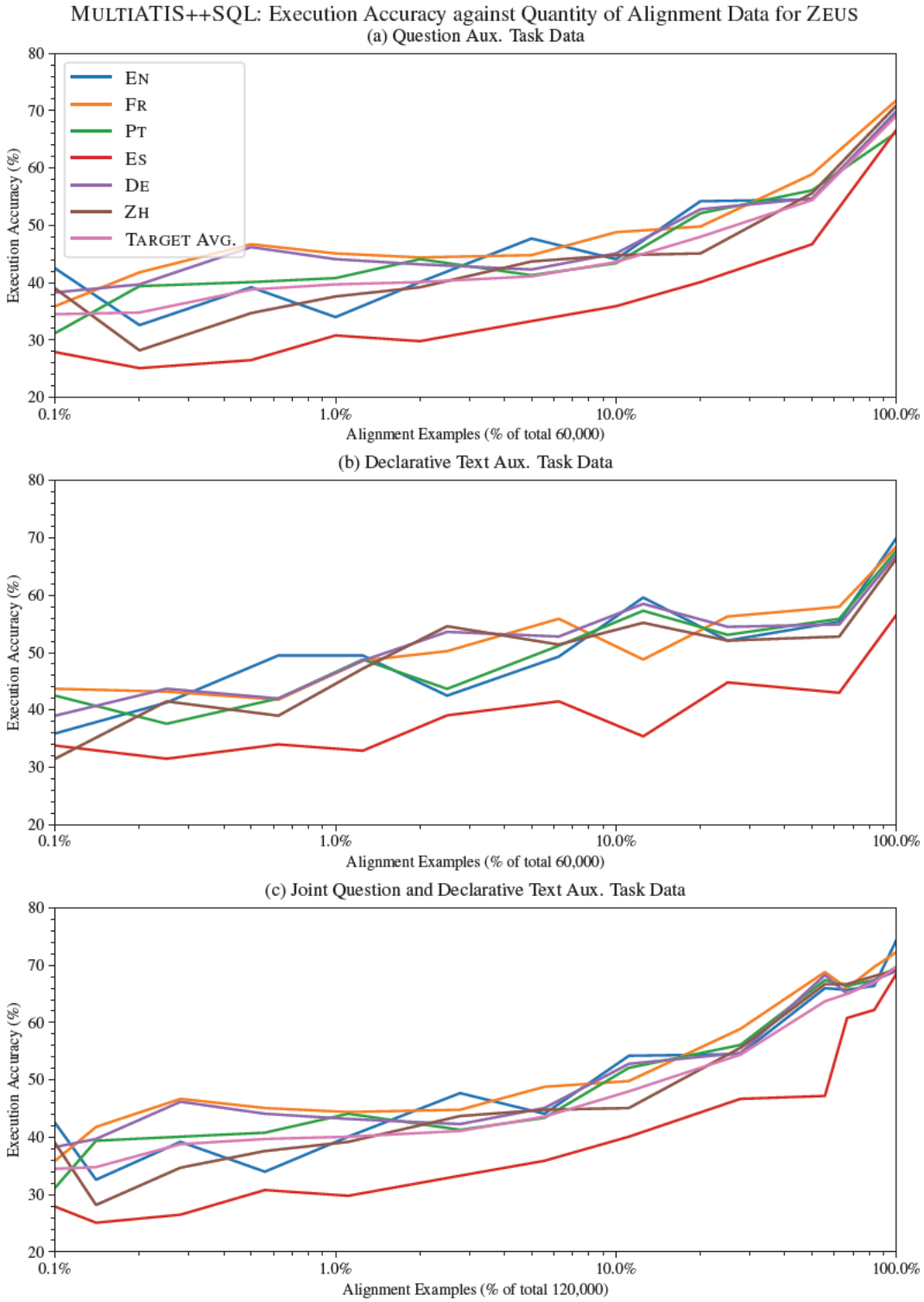

Figure 4.5: MultiATIS++SQL execution accuracy for ZEUS varying the number of examples used for auxiliary tasks as a percentage of total examples (55,425). We report results training ZEUS varying the auxiliary task dataset using (a) question data, (b) declarative text data, and (c) combined question data and declarative text. More data positively improves parsing similar to MultiATIS++SQL but we identify improved cross-lingual transfer with fewer examples for this dataset. Table 4.1 further details data sources.

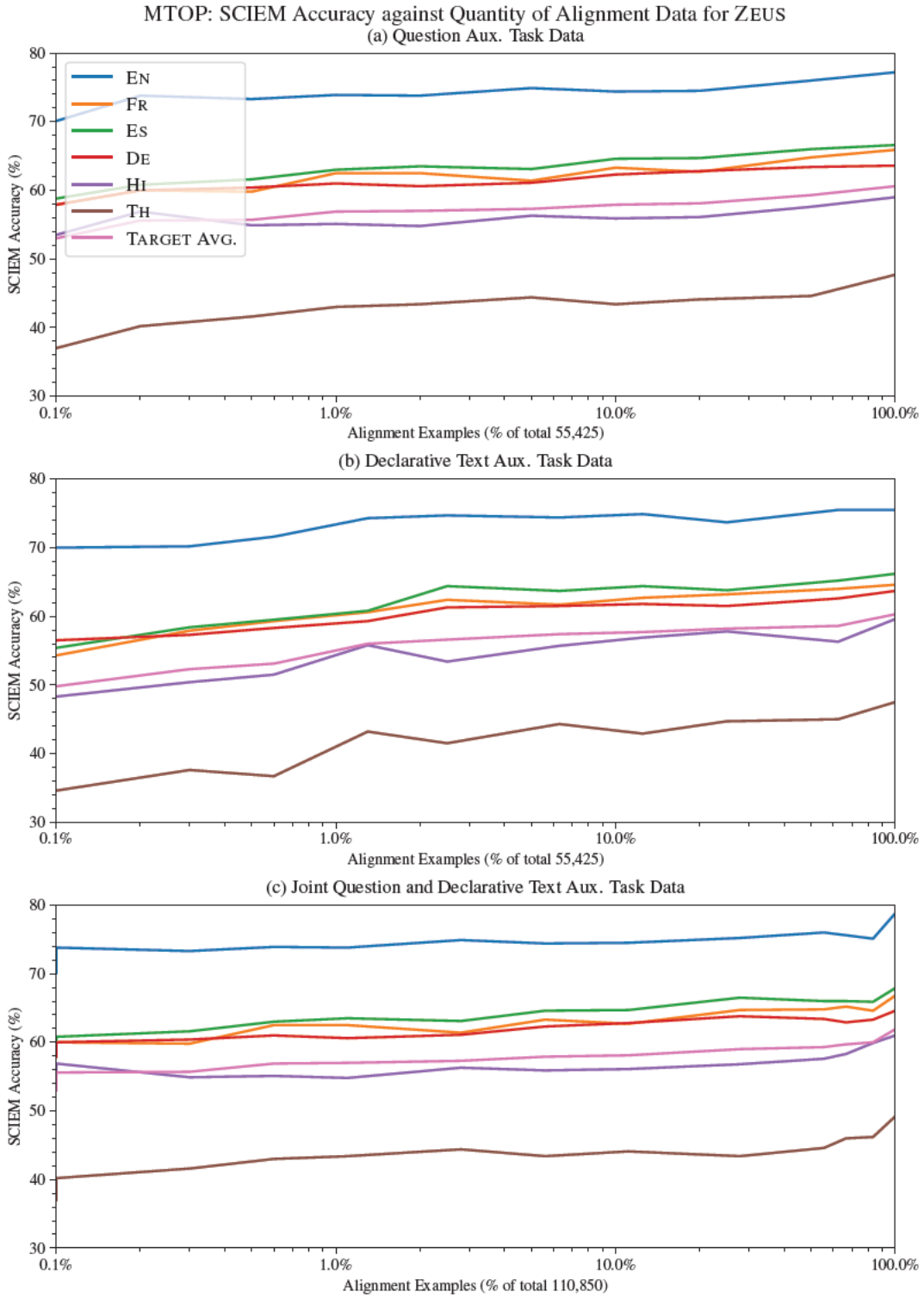

Figure 4.6: MTOP SCIEM accuracy for ZEUS varying the number of examples used for auxiliary tasks as a percentage of totaltotal examples (58,000). We report results training ZEUS varying the auxiliary task dataset using (a) question data, (b) declarative text data, and (c) combined question data and declarative text. Across all languages, we observe that more auxiliary task data benefits parsing performance despite this data containing no labels for semantic parsing. Table 4.1 further details data sources.

Method Type	Method	Cosine (↑)	Top-1	Top-5	Top-10
Pre-trained Model	MBART50	0.576	0.521	0.745	0.796
Lower Bound	Train-EN Only	0.364	0.669	0.775	0.964
Upper Bound	MULTILINGUAL Gold	0.698	0.784	0.981	0.991
Machine Translation	FATES	0.670	0.720	**0.957**	**0.992**
Zero-shot	ZEUS	**0.760**	**0.832**	0.944	0.971

Table 4.7: Average similarity between encodings of English and target languages for MultiATIS++SQL. Cosine similarity evaluates average distance between encodings of parallel sentences. Top-k evaluates if the parallel encoding is ranked within the k most cosine-similar vectors (higher (↑) is better). Best excluding the upper-bound is bold.

4.3.7 Visualising Latent Representation Similarity

We investigate our claim that ZEUS improves latent cross-lingual similarity by visualising the latent space as outlined in Section 2.4.2. Table 4.7 shows a quantitative analysis of high-dimensional representation similarity for MultiATIS++SQL. As discussed in Section 3.3.6, monolingual clustering artefacts arise without some guidance for cross-lingual representation alignment. ZEUS demonstrates reduced monolingual separability compared to the pre-trained model and the zero-shot lower bound. We observe ZEUS approximately produces a single multilingual cluster, whereas the upper-bound and FATES demonstrate structure without global multilingual similarity. We note the similarity in all methods of poorer alignment of distant languages (ZH). Quantitatively, ZEUS also produced the highest cosine-similarity and Top-1 ranked similarity above all other models thus far. Notably, we surpass the upper-bound in similarity which has access to 100% translation into each target language. Similarly, encouraging representation alignment produces more similar representations than using synthetic data without encouraging alignment. We infer that targeting representation similarity in ZEUS is how we surpass the cross-lingual similarity of the upper bound without explicit representation alignment. We interpret our results and quantitative analysis to further affirm our hypothesis that ZEUS improves cross-lingual representation alignment to improve zero-shot cross-lingual transfer. The benefit of ZEUS highlights that optimising representation alignment is valuable for the thesis beyond zero-shot parsing. We revisit this idea in Chapter 6 to explicitly optimise representation alignment as a loss function.

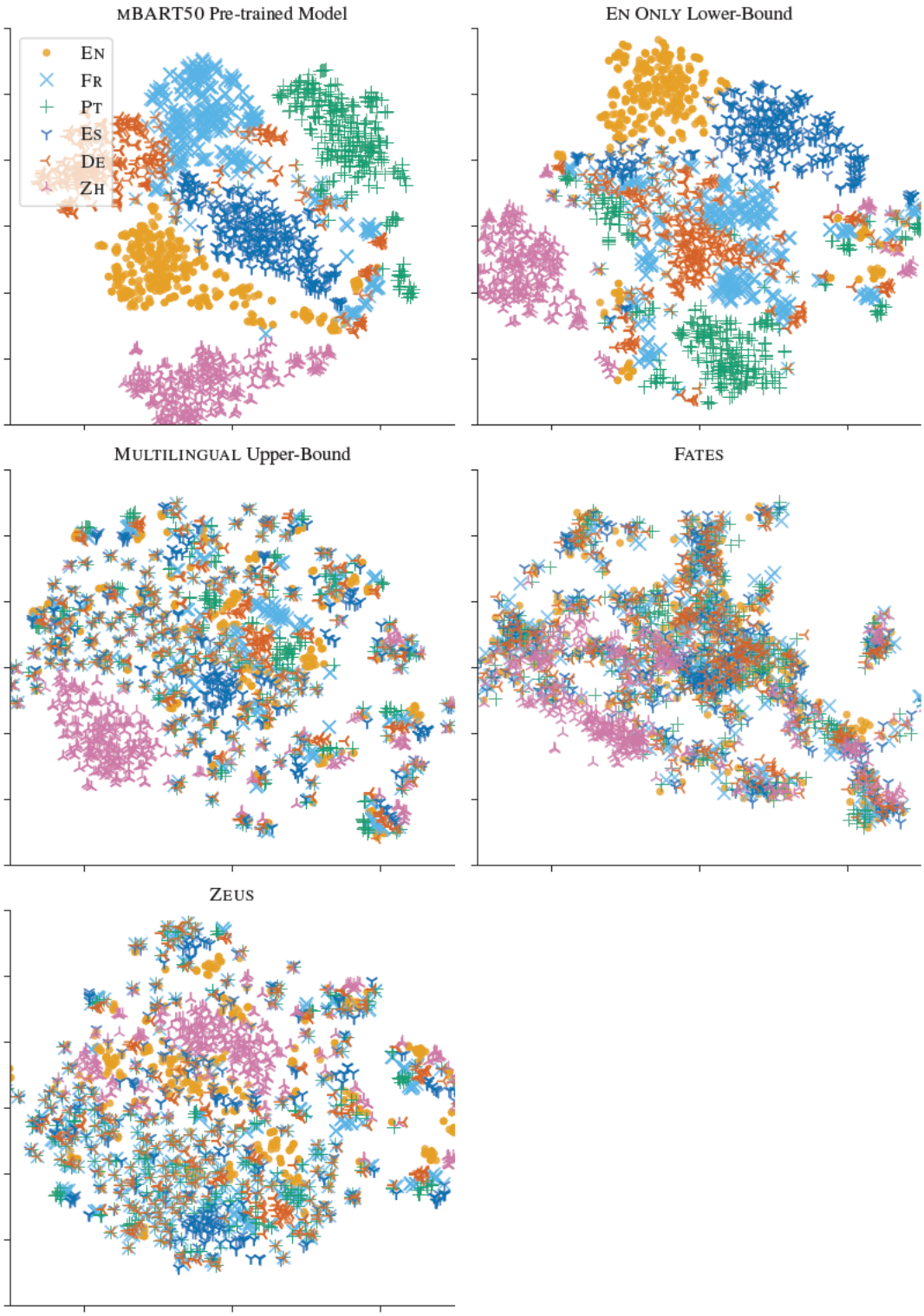

Figure 4.7: Visualisation of MultiATIS++SQL encodings (test set; 50% random parallel sample) using t-SNE. We compare the original mBART50 pre-trained model, the EN-ONLY zero-shot lower bound, MULTILINGUAL training upper bound, FATES from Chapter 3, and ZEUS. ZEUS reduces the monolingual separability in the latent space for improved cross-lingual representation alignment.

EN	What ground transportation is there in Baltimore?
ZH	巴尔的摩有什么地面交通
LF (Error)	`SELECT DISTINCT ground_service_1.transport_type FROM ground_service ground_service_1, city city_1 WHERE (( ground_service_1.city_code = city_1.city_code AND city_1.city_name = ´BALTIMORE ´ );`
LF (Correct)	`SELECT DISTINCT ground_service_1.transport_type FROM ground_service ground_service_1, city city_1 WHERE ground_service_1.city_code = city_1.city_code AND city_1.city_name = ´ BALTIMORE ´;`
EN	What type of plane is a D9S?
FR	Quel type d'aéronef est un D9S?
LF (Error)	`SELECT DISTINCT aircraft_1.aircraft_code FROM aircraft aircraft_1 WHERE aircraft_1.airport_code = ´DFW´ ;`
LF (Correct)	`SELECT DISTINCT aircraft_1.aircraft_code FROM aircraft aircraft_1 WHERE aircraft_1.aircraft_code =´D9S´;`

Table 4.8: Error examples from MultiATIS++SQL and MTOP for ZEUS. The primary improvement from ZEUS is a reduction in unexecutable or malformed logical forms. In the upper example, the TRANSLATE TRAIN model has generated incorrect brackets without closing all pairs. ZEUS can often handle generic entities but can fail when parsing rare entities not included in the auxiliary task corpora. In the lower example, the ZEUS model has not observed the 'D9S' entity in a French utterance and erroneously referenced the 'DFW' entity instead.

4.3.8 Error Analysis

We randomly sample 50 MultiATIS++SQL test examples where ZEUS correctly predicted the outputs but the best TRANSLATE TRAIN model failed. We also sample 50 examples where the gold-standard multilingual model produces a correct parser but the best ZEUS model failed. Similar to Chapter 3, we discuss trends in improvement and identify failures in the model for further study. Examples of errors are shown in Table 4.8. We note that these error patterns also extend to MTOP.

Executability Comparing to TRANSLATE TRAIN, ZEUS generates 32% fewer ill-formed SQL requests and 24% fewer extraneous queries accessing unrelated tables in the database. We observe many cases of malformed SQL in weaker models, often as the input utterance is dissimilar to the English or machine-translated training data. ZEUS makes fewer of these mistakes but can still generate incorrect logical forms. The improvement with ZEUS shifts the focal point of errors from malformed logical forms to semantically inaccurate logical forms i.e., the motivating example in Figure 4.1.

Rare Entities Translation can fail when entities are mishandled and ZEUS generates 36% fewer queries with erroneously named entities. However, many entities are still mishandled with ZEUS and weaker models. For example, the airline 'Tower Air' appears in the dataset as the original data was collected in 1990 and 1993 (Hemphill et al., 1990; Dahl et al., 1994). However, this business ceased trading in 2000. Therefore, this entity is likely to be rare, or nonexistent, in auxiliary task corpora sourced from modern web text. This causes all models to often misinterpret the entity when observed in the context of each target language. This issue is further exacerbated when the entity 'Tower Air' is transliterated into Chinese as "宝塔航空 [pagoda airlines]". This increases the difficulty for the parser to interpret a transliteration of a rare entity into a unique form unlikely to be prevalent in any auxiliary data. These types of errors are a major outstanding barrier to improving cross-lingual semantic parsing—accounting for 58% of remaining errors with the most in Chinese (60%) and least in French (36%).

4.4 Related Work

Auxiliary tasks for cross-lingual transfer have proved beneficial for multiple tasks. Mallinson et al. (2020) propose to jointly learn sentence simplification for English and language modelling for German for zero-shot English-to-German sentence simplification. Similarly, Zhao et al. (2020) propose a similar method combining language modelling and iterative back-translation tasks for sentence simplification demonstrating a denoising objective (similar to Section 4.1.3.1) can transfer between 'simple' and 'complex' language in a semi-supervised objective.

Gradient reversal techniques for language-agnostic representations have also been explored across multiple tasks. Ahmad et al. (2019) propose an auxiliary classifier, similar to our LP submodel in Section 4.1.3.3, for language agnostic representations to improve cross-lingual dependency parsing. Arivazhagan et al. (2019) propose the

same auxiliary loss for improving zero-shot machine translation. However, neither work introduces this auxiliary loss to adapt a monolingual task: either this technique complements multilingual learning with access to target language data or it is used to connect pairs of MT languages using Bitext.

Auxiliary tasks have demonstrated improvement in cross-lingual semantic parsing. Shao et al. (2020) propose learning a similar language predictor network using adversarial losses (Goodfellow et al., 2014) for spoken language understanding (SLU). This model is similar to our gradient reversal approach but negates the loss before the backward pass rather than reversing the gradient within the network. Yang et al. (2021) use linguistic analysis-oriented tools to improve zero-shot cross-lingual parsing to Discourse Representation Structure (DRS) logical forms. They show that language-independent features such as Universal Dependency labels (de Marneffe et al., 2014, UD), and Universal Part-of-speech tagging (Petrov et al., 2012) benefit a similar zero-shot setting to ours. van der Goot et al. (2021) also report a similar auxiliary task framework using translation, masked language modelling and UD prediction to improve zero-shot SLU parsing.

This chapter initially appeared as Sherborne and Lapata (2022); where we propose a novel combination of domain-adaptive pre-training and adversarial learning for language-agnostic representations. Unlike adjacent work, ZEUS requires no additional annotation (e.g., UD labels) and can scale to languages where only unlabelled data is available. Our examination of the role of surface form *style* and *quantity* also separates our work from others. Prior work leverages data for auxiliary tasks without introspection on the influence of corpora size and data quality. In contrast, we highlight that the style of additional data is strongly influential on the parser performance. In **recent follow up work**, Wang and Hershcovich (2023) extend Sherborne and Lapata (2022) for parsing Chinese into SPARQL logical forms. Their results highlight that the compositional generalisation capability of a parser, similar to ZEUS, scales linearly with increasing auxiliary data.

4.5 Summary

In this chapter, we examine how auxiliary tasks and loss functions can improve latent cross-lingual representation alignment for zero-shot cross-lingual semantic parsing. We validate a hypothesis suggesting we can jointly learn monolingual parsing and extra multilingual tasks (with existing data) to zero-shot improve. We propose ZEUS, a

parser using natural language denoising, machine translation and language prediction to improve zero-shot cross-lingual transfer. ZEUS demonstrates that improving latent representation similarity enables a parser to accurately predict a logical form for a language without training data. Ablations identify how many examples are sufficient to benefit latent alignment and the contribution of each auxiliary task. We also observe an optimal balance between reconstruction and translation tasks—highlighting the benefit of both unlabelled data and bitext for aligning representations. Analysing the latent representations learned from ZEUS demonstrates a broad improvement in latent similarity. Our error analysis further identifies that the core improvement from ZEUS is a reduction in generating unexecutable or syntactically invalid logical forms from target language inputs. Whereas Chapter 3 identifies outstanding errors in modelling fluency, we find that ZEUS can struggle to model domain-specific entities relevant only to the test dataset. After having examined silver-standard and zero-shot approaches, we now consider if few-shot data sampling can overcome these issues to further improve cross-lingual parsing in Chapter 5 and Chapter 6.

Chapter 5

Meta-Learning a Cross-lingual Manifold for Semantic Parsing

We previously considered the merit of machine translation in Chapter 3, and auxiliary data from adjacent tasks in Chapter 4. Either method can partially mitigate the performance gap between English and target languages. However, data quality issues lower the generalisation ceiling for a parser relying on alternatives to gold-standard data. Our findings identify that synthetic data can struggle to fluently represent native speakers. Similarly, 'gold-standard' data for auxiliary tasks improves parsing but yields minimal benefit when the *domain* and *style* of such data poorly resembles the test data. Neither method provides a reliable mechanism for improving cross-lingual representation alignment of utterances we desire to parse.

Given the absence of an exemplar alternative, we now consider relaxing the constraint on sampling the target language data distribution. Using gold data could improve upon FATES using fluent translations, and upon ZEUS using relevant language. We consider a 'few-shot' approach sampling a small set of labelled utterance-logical form pairs for each target language. Few-shot cross-lingual transfer demonstrates improvement over zero-shot transfer or silver-standard data methods in a range of tasks such as entailment detection and machine translation (Zhao et al., 2021). The benefit of gold data motivates our transition to few-shot sampling. We contrast all data strategies in Table 5.1. While few-shot sampling demands expert labelling, the benefit to parsing accuracy of fluent and relevant target language data may outweigh this cost. Furthermore, access to target language data permits novel modelling contributions optimising data efficiency within cross-lingual transfer (here and Chapter 6.) Practically, it is reasonable to assume that data sampled from the target distribution will support generalisation to

Chapter	Modelling	Data Strategy	Advantage	Disadvantage
Chapter 3	FATES: multi-encoder ensemble	'Silver-standard' machine translation	Low cost to generate	Low fluency
Chapter 4	ZEUS: alignment with auxiliary tasks	Zero-shot using auxiliary corpora	Typically easy to source	Low relevance
Chapter 5	DRAKON: meta-learning for transfer	Few-shot data sampling	Natural and relevant to task	Expensive

Table 5.1: Comparing the data strategies from current and prior chapters. In Chapter 5 and Chapter 6, we consider the possible merits of few-shot data sampling. We evaluate if few-shot sampling, despite the expense of annotation, is more valuable for the task than the alternative methods proposed in Chapter 3 and Chapter 4.

the target distribution. However, the expense of large-scale multilingual translation mandates responsibility for 'few-shot' sampling to minimise this quantity of 'few' examples. Therefore, this chapter examines how sampling gold-standard data improves cross-lingual semantic parsing, and the scale of required sampling to supersede methods from previous chapters.

This chapter investigates a few-shot approach to cross-lingual semantic parsing derived from meta-learning (Finn et al., 2017). The goal of meta-learning is to optimise both the performance and the learning algorithm for any task. Typical meta-learning optimises the *learning capability* of a model, such that a system can rapidly learn a new task using few examples (i.e., improving learning to learn). This is often applied in a multi-task learning framework episodically optimising the model to adapt towards a sampled task with minimal training (Wang et al., 2020; Hospedales et al., 2022). This is suitable for our case study for data-efficient transfer across languages. Therefore, we follow cross-lingual adaptations of meta-learning by modelling parsing each target language as a task (Nooralahzadeh et al., 2020). As outlined in Sections 5.1.2 to 5.1.3, this approach offers a sample-efficient algorithmic solution to few-shot cross-lingual transfer by integrating gradients from tasks with plentiful data (i.e., the source language) and tasks with minimal data (i.e., target languages).

This chapter considers the hypothesis that **meta-learning improves few-shot cross-lingual transfer by promoting gradient-level cross-lingual regularisation**. We propose DRAKON,[1] a method of efficiently integrating task fine-tuning for semantic parsing and sample-efficient few-shot cross-lingual transfer. DRAKON uses first-order meta-learning to approximate an ideal training step towards the optimal parameters for semantic parsing (defined in Section 5.1). DRAKON augments this training step with gradients from target language samples. These target language steps are now closely

[1] DRAKON is not an acronym and instead invokes an evolution of the Reptile algorithm (Nichol et al., 2018) which we build on.

aligned with the ideal training trajectory to improve the similarity in solutions for source and target languages (see Section 5.1.6). We hypothesise that DRAKON improves representation alignment by enforcing cross-lingual similarity in the parameter solution during optimisation.

Our experimental results highlight that DRAKON generally improves on FATES or ZEUS sampling gold-standard data while also requiring fewer dependencies with potential failures (e.g., machine-translation engines or domain-irrelevant auxiliary corpora). Meta-learning using DRAKON demonstrates data-efficient parsing and improves representation similarity beyond our previous methods.

5.1 Problem Formulation

5.1.1 Observing the Target Language Distribution

From Chapter 2, we recall that for each target language l, there exists some true underlying distribution, $\mathcal{D}_l$ over the data for the task. Previous chapters assume we cannot sample $\mathcal{D}_l$ to produce an empirical estimate of the loss over this distribution. This previously motivated sampling from *alternative* distributions we proposed could generalise to test data from $\mathcal{D}_l$. We now relax the constraint on sampling the true distribution for training data, allowing a small quantity of natural language-logical form paired training data examples in each target language. We previously discussed sampling the target language distribution as uneconomical, however, the alternative proposals risk additional costs to verify the viability of each alternative. These costs may be disproportionate to the minimal effort of data collection and could accumulate over time to cost more overall. As a counter, we conjecture that few-shot sampling requires less quality auditing by virtue of sampling from the same distribution as the test set. This sunk cost may ultimately prove more useful for our task if few-shot annotation requirements can be minimised (Garrette and Baldridge, 2013).

We interpret the meaning of *few* in the *few-shot* cross-lingual transfer scenario as a ratio of *gold-standard translation from source to target language*. Given a set of n training examples, we partition this set into n_{EN} samples in English ($\mathcal{S}_{\text{EN}}$ i.e., the source language) as Equation (5.1), and n_l examples translated into all target languages, $\mathcal{S}_l$, as Equation (5.2). The total unique examples remains $n = n_{\text{EN}} + n_l$ and n_l is presumed to be small (i.e., $n_l \ll n_{\text{EN}}$). The size of n_l required for accurate cross-lingual parsing is an empirical question which we address in Section 5.3.

$$\mathcal{S}_{\text{EN}} = \{x_{\text{EN}},\, y\}_{i=0}^{n_{\text{EN}}} \tag{5.1}$$

$$\mathcal{S}_{l} = \{x_{l},\, y\}_{i=0}^{n_{l}} \tag{5.2}$$

5.1.2 Model Agnostic Meta-Learning (MAML)

The objective of meta-learning is to prime a model to quickly adapt to new tasks using few training examples or steps (Wang et al., 2020, 2021b; Hospedales et al., 2022). A common algorithmic methodology for meta-learning is the **model-agnostic meta-learning** (MAML) framework (Finn et al., 2017) which we define in this section.

MAML introduces a "meta-training" optimisation phase wherein the objective is to *rapidly* adapt to a set of known tasks by iteratively simulating few-shot learning. The intention is to improve *the speed of learning tasks* rather than conventionally optimising an objective function. Learning the capability to learn faster gives rise to the "learning to learn" interpretation of how MAML works. At each episode, MAML meta-training randomly samples a task τ with associated k batches of training data. A model is optimised for τ with only k batches. MAML is typically employed in multi-task training e.g., each 'task' is a specific classification scenario such as 'classify dogs from 10 sampled animal categories' or 'classify the sentiment for this language with four examples'.

Equation (5.3) defines the MAML optimisation problem for each episode: optimise model parameters θ such that loss for task τ, using loss function ℓ_τ, is lowest after k update steps (represented by operator U_τ^k updating θ k times on task τ). The update, U_τ^k, is the conventional optimisation step to update θ based on loss, ℓ_τ, at each k step. The episode's final "meta-testing" update computes the loss after k updates and backpropagates this loss through the computation graph of all k updates. The effect of this is that θ improves at task τ within the budget of k updates.

$$\min_{\theta} \mathbb{E}_\tau \left[\ell_\tau \left(U_\tau^k (\theta) \right) \right] = \mathbb{E}_\tau \left[\ell_\tau \left(U_\tau^k \left(U_\tau^{k-1} \left(U_\tau^{k-2} \left(\ldots (U_\tau^1 (\theta_1)) \right) \right) \right) \right) \right] \tag{5.3}$$

Typical MAML training is a multi-task episodic training loop repeatedly sampling multiple tasks to improve model capability and speed for few-shot (k-shot) learning. MAML training typically produces parameter initialisation only. Further conventional fine-tuning is required to adapt to the desired task. The advantage of MAML is that this final training requires fewer steps and data than fine-tuning without MAML. MAML

training is ultimately useful for adaptation to out-of-domain or low-resource scenarios where data efficiency is increasingly important. However, a major criticism of MAML is the computational complexity, as the final gradient requires backpropagating through previous gradient updates. This higher-order gradient computation (i.e., "gradient through a gradient") calculates gradients over a complex and memory-intensive computation graph (see the unrolled Equation (5.3) for updating θ). This is often prohibitive when training large models with high GPU memory demands. This has motivated study into *first-order* meta-learning methods offering similar outcomes without higher-order gradient computation.

5.1.3 First Order Approximations of MAML

Reptile (Nichol et al., 2018) offers a similar meta-learning algorithm with reduced computational demand by transforming the meta-train gradient into a heuristic step using the distance between the initial and final parameters. Nichol et al. (2018) demonstrate that Reptile is convergent on a similar solution to MAML minimising the expected loss over meta-training tasks. Reptile replaces the meta-test step, with respective higher-order gradients, using the difference $U_\tau^k(\theta) - \theta$ as a gradient. Equation (5.4) describes this objective. When $k = 1$, this is equivalent to gradient descent over expected loss $\nabla_\theta \ell_\tau(\theta)$, but if $k > 1$ then Reptile training includes additional second-order derivatives promoting cross-batch generalisation (discussed below).

$$\min_\theta \mathbb{E}_\tau \left[U_\tau^k(\theta_1) - \theta_1 \right] \tag{5.4}$$

Nichol et al. (2018) use a Taylor Series expansion on each gradient step in a Reptile episode, g_i, to demonstrate how Reptile approximates MAML. The components of this analysis are shown in Equations (5.5) to (5.9), where ℓ is the loss function for the task, θ are the model parameters, and α is the step size between steps in a single Reptile episode.

$$g_i = \ell_i'(\theta_i) \quad \text{Gradient at step } i \tag{5.5}$$

$$\bar{g}_i = \ell_i'(\theta_1) \quad \text{Gradient at initial point} \tag{5.6}$$

$$\bar{H}_i = \ell_i''(\theta_1) \quad \text{Hessian at initial point} \tag{5.7}$$

$$\theta_i = \theta_{i-1} - \alpha g_{i-1} \quad \text{Parameter vector at } i \text{ from } i-1 \tag{5.8}$$

$$\theta_k - \theta_1 = -\alpha \sum_{j=1}^{k-1} g_j \quad \text{Total parameter vector at } k \tag{5.9}$$

The gradient at some step i can be expressed as a Taylor Series expansion in Equation (5.10). The final term here describes non-critical effects proportional to step size α omitted for clarity. Rearranging the components of this analysis yields a gradient update at i with respect to the original parameters, θ_1, as Equation (5.11).

$$g_i = \ell_i'(\theta_i) = \ell_i'(\theta_1) + \ell_i''(\theta_1)(\theta_i - \theta_1) + O\left(\alpha^2\right) \tag{5.10}$$

$$g_i = \bar{g}_i + \bar{H}_i(\theta_i - \theta_1) + O\left(\alpha^2\right) \tag{5.11}$$

Using a substitution of Equation (5.9), we can express the gradient at i as Equation (5.12). The first term in Equation (5.12) is equivalent to the conventional training loss minimisation. The second term optimises the inner product between the gradient Hessian at i and the gradient from the previous steps $< i$. This secondary term promotes similarity between gradients across multiple batches. This additional gradient component is how Reptile improves over conventional gradient descent using only $\bar{g}_i$.

$$g_i = \bar{g}_i - \alpha \bar{H}_i \sum_{j=1}^{i-1} \bar{g}_j + O\left(\alpha^2\right) \tag{5.12}$$

Reptile can be interpreted as optimising towards the *solution manifold* for each task τ (i.e., the subspace in θ of an optimal solution). Nichol et al. (2018) highlight that Reptile training is an iterative solution minimising the distance between current parameters and the manifold of optimal parameters for each task. Kedia et al. (2021) further raise that Reptile training for a *single task* converges on this solution manifold.

The final Reptile update can be geometrically interpreted as the trivial vector addition of each step's gradient. The direction of the final update is the shared vector component across all steps (i.e., $g_i \cos(\varphi)$ for some angle φ). As k increases, the cross-batch inner product term promotes generalisation across more batches from the training data

distribution. The learned solution now improves loss on all batches (or all data as k increases). The inner product in Equation (5.12) promotes learning in the direction of the shared vector component to maximise the utility of the meta-training stage. Reptile is similar to MAML in requiring a secondary fine-tuning stage for utility for an end task.

5.1.4 Domain Generalisation MAML (DG-MAML)

The fine-tuning requirements of models trained with meta-learning create additional effort beyond the existing expense of meta-training. A desire to remove the additional complexity of fine-tuning after meta-learning has motivated efforts into meta-learning for domain generalisation (Wang et al., 2021b). This approach to meta-learning modifies MAML-based methods to directly simulate a domain shift by introducing a secondary 'meta-test' step after meta-training. This step optimises for cross-task generalisation conditioning out-of-distribution performance (meta-test) on in-task few-shot learning (meta-train). This simulation improves the model's capability to reason about novel domains without additional fine-tuning during inference. This appeals to our case study by offering a few-shot optimisation strategy in a single training stage. Additionally, it is desirable in our scenario to produce a singular multilingual model, rather than a multilingual initialisation requiring monolingual fine-tuning for target languages.

Wang et al. (2021a) propose Domain Generalisation MAML (DG-MAML), a methodology for generalisation in semantic parsing wherein the meta-train and meta-test steps of MAML explicitly target generalisation across different domains. This follows the setup of MAML in Section 5.1.2, but samples different tasks for meta-training and meta-testing. A 'task' in DG-MAML is a set of paired data examples relevant to a specific database (i.e., the domain). A single step of DG-MAML might use meta-training on a database of baseball statistics, and meta-testing will evaluate generalisation to a database about bird migration. This simulates the same domain transfer objective from MAML in a framework specific to semantic parsing. DG-MAML modifies the MAML objective to Equation (5.13), where τ_{TRAIN} and τ_{TEST} are semantic parsing tasks from different databases (i.e., different databases, tables and column structures with different grounding environments).

$$\min_{\theta} \mathbb{E}_{\tau_{\text{TEST}}} \left[\ell_{\tau_{\text{TEST}}} \left(U^{k}_{\tau_{\text{TRAIN}}} (\theta) \right) \right] \tag{5.13}$$

Wang et al. (2021a) also propose Domain Generalisation First-Order MAML (DG-FMAML): a simplified first-order variant, competitive with DG-MAML with reduced memory and computation demands. The objective of DG-FMAML is to minimise the loss on τ_{TEST} after one update on τ_{TRAIN} as Equation (5.14). This updates θ by simulating a domain shift without requiring higher-order gradients.

$$\min_{\theta} \mathbb{E}_{\tau_{\text{TEST}}} \left[U^{k=1}_{\tau_{Train}} (\theta_1) - \theta_1 \right] \tag{5.14}$$

The similarities between Equation (5.4) and Equation (5.14) highlight that Reptile and DG-FMAML are closely related. The former excludes a meta-testing step to run multiple updates over some τ_{TRAIN} task and the latter can be interpreted as Reptile with $k = 1$ and an additional meta-test loss. In this chapter, we reconcile these algorithmic similarities into a single optimisation process. For cross-lingual transfer, we view each language as a task to optimise generalisation from the source language (meta-training on English) to target languages during meta-testing.

5.1.5 DRAKON: Meta-Learning for Cross-lingual Transfer

Our method, DRAKON, integrates the *solution manifold* optimisation from Reptile with the out-of-domain generalisation from DG-FMAML. This yields a meta-learning algorithm capable of cross-lingual adaptation from English to target languages as a single training process without additional fine-tuning. DRAKON extends and unifies Reptile and DG-FMAML by combining practices from each algorithm. DRAKON extends Reptile by augmenting the meta-training step with out-of-distribution (i.e., cross-lingual) generalisation. DRAKON extends DG-FMAML by replacing the single step on τ_{TRAIN} with a Reptile step over k batches. We show that DRAKON aligns with the τ_{TEST} step with an approximation of the global training trajectory rather than a single batch.

In terms of meta-learning tasks, our approach to cross-lingual transfer uses English paired training data as the meta-training task, τ_{TRAIN}, and few-shot samples of each target language as the meta-testing task, τ_{TEST}. Meta-training optimises the semantic parsing task loss, using English, and meta-testing regularises this optimisation to additionally promote cross-lingual similarity in gradient steps.

5.1.5.1 Meta-Training a Semantic Parser

Meta-training uses the Reptile inner-loop described in Equation (5.4). The high-resource task is often referred to as the "support" task used to condition optimisation for another

"target" task. We sample k batches of English paired training data, $\mathcal{B}^S = \{(x_{\text{EN}}, y)\}^k$. For each of k batches: we generate predictions, compute losses, calculate gradients and adjust parameters using some optimiser (see illustration in Figure 5.1). After k successive optimisation steps: the initial weights in this episode, θ_1, are optimised to θ_k. This process, as Equation (5.15), matches the Reptile process in Equation (5.4). Equation (5.16) shows the meta-train update, $\nabla_{\tau_{\text{TRAIN}}}$ as the Reptile difference step between initial and final parameters.

$$\theta_k = U^k_{\tau_{\text{TRAIN}}}(\theta_1) \tag{5.15}$$

$$\nabla_{\tau_{\text{TRAIN}}} = \theta_k - \theta_1 = U^k_{\tau_{\text{TRAIN}}}(\theta_1) - \theta_1 \tag{5.16}$$

Optimisation using $\nabla_{\tau_{\text{TRAIN}}}$ corresponds to optimisation towards the solution manifold for monolingual semantic parsing. While MAML is typically a multi-task objective during meta-training, we follow Kedia et al. (2021) highlighting the generalisation benefit of single-task meta-training. Our intuition is that this optimisation target represents the task solution for any high-resource language.

5.1.5.2 Meta-Testing for Cross-lingual Generalisation

DRAKON combines meta-training using Reptile with meta-testing evaluation on a τ_{TEST} sampling data from target languages. The τ_{TEST} task, or "target" task, predicts loss on batches from τ_{TEST} conditioned on the meta-training described above. Meta-testing samples a target language, $l \sim L$, and the target batch, $\mathcal{B}^T = (x_l, y)$, to compute loss $\ell_{\tau_{\text{TEST}}}$ at θ_k. Equation (5.17) expresses the meta-test loss as the minibatch loss over $\mathcal{B}^T$ with respective meta-test gradient, $\nabla_{\tau_{\text{TEST}}}$, as Equation (5.18).

$$\ell_{\tau_{\text{TEST}}} = \frac{1}{\|\mathcal{B}^T\|} \sum_{(x,y)\in\mathcal{B}^T} \ell(x, y) \tag{5.17}$$

$$\nabla_{\tau_{\text{TEST}}} = \nabla \ell_{\tau_{\text{TEST}}} (= g_{\tau_{\text{TEST}}}) \tag{5.18}$$

We evaluate the parser on a sampled target language for cross-lingual generalisation conditioned on the Reptile update at θ_k. The gradient of the meta-test loss can be expressed in the same notation as Equation (5.10) as $g_{\tau_{\text{TEST}}}$. We show in Section 5.1.6 that this gradient comprises target loss at θ_k and additional terms maximising the inner product between gradients of different languages (similar to Equation (5.11)).

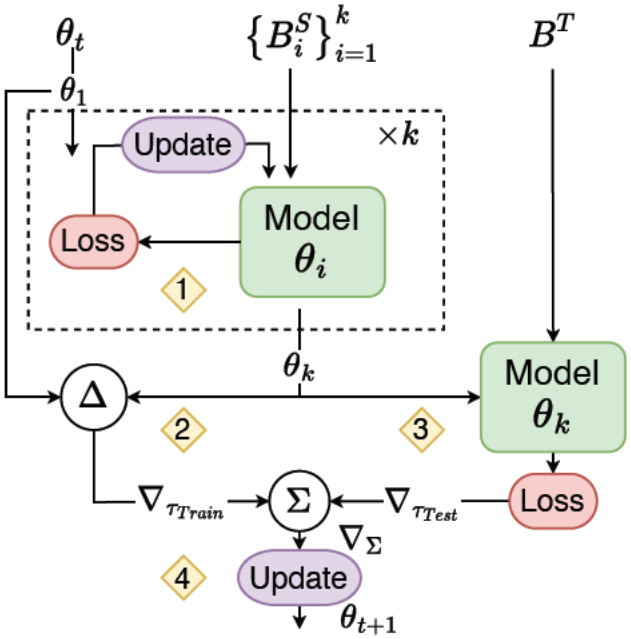

Figure 5.1: One episode of DRAKON optimisation. (1) Run k iterations of gradient descent over k support batches to learn θ_k, (2) compute $\nabla_{\tau_{\text{TRAIN}}}$, the difference between θ_k and θ_1, (3) find the loss, $\ell_{\tau_{\text{TEST}}}$, on the target batch using θ_k and (4) compute the final gradient update from $\nabla_{\tau_{\text{TRAIN}}}$ and $\nabla_{\tau_{\text{TEST}}}$.

A single episode of DRAKON is outlined in Figure 5.1 and formally defined in Algorithm 1. We repeat this process over T episodes to train model $f(x, \theta)$ to convergence. Without the meta-test gradient, $\nabla_{\tau_{\text{TEST}}}$, the method is equivalent to Reptile and requires later fine-tuning for task-specific utility. DRAKON is equivalent to DG-FMAML if $k = 1$. DG-FMAML will combine target batch gradients with individual support batches, but this may approximate the global training trajectory poorly with smaller local gradient updates and no cross-batch generalisation during meta-training. This omits the benefit of the Reptile update using $k > 1$ batches during meta-training. Our intuition is that DRAKON using the Reptile update for meta-training overcomes this batch-level 'noise' in learning. The meta-test step can now optimise an inner product between the target gradient and the approximate global training trajectory. This improves the similarity in loss minimisation for target languages further than DG-FMAML without requiring more data. Results in Section 5.3 confirm this intuition that DRAKON improves over DG-FMAML by improving the meta-training phase.

Concerning the data efficiency of DRAKON, our approach surpasses DG-FMAML by exploiting the asymmetric data requirements between meta-train and meta-test steps in Algorithm 1: one batch of a target language is required for k batches of the

Algorithm 1 DRAKON

Require: Number of training episodes, T, and number of inner Reptile steps, k.

Require: Support data sample, $\mathcal{S}_{\text{EN}}$

Require: Target data samples, $\mathcal{S}_l$, for each language l in target languages $L = \{l_1, \ldots, l_L\}$.

Require: Inner learning rate, α, outer learning rate, β

Require: Inner optimiser, $U(\theta_i, \alpha, \nabla)$, updating θ_i according to step size α and gradient ∇.

Require: Outer optimiser, $V(\theta_t, \beta, \nabla)$, updating θ_t according to step size β and gradient ∇.

1: Initialise $\theta_{t=1}$, the vector of initial parameters
2: **for** $t \leftarrow 1$ **to** T **do**
3: Sample K support batches $\{\mathcal{B}^S\}_{k=1}^{K}$ from $\mathcal{S}_{\text{EN}}$.
4: Sample target language l from L languages.
5: Sample target batch $\mathcal{B}^T$ from $\mathcal{S}_l$.
6: **for** $i \leftarrow 1$ **to** k [Inner Loop] **do**
7: $\ell_{\tau_{\text{TRAIN}}} = \frac{1}{\|\mathcal{B}_i^S\|} \sum_{(x,y)\in\mathcal{B}^{S_i}} \ell(f(x,\ y))$
8: $\theta_{i+1} \leftarrow U(\theta_i,\ \alpha,\ \nabla\ell_{\tau_{\text{TRAIN}}})$
9: **end for**
10: Meta-Training gradient: $\nabla_{\tau_{\text{TRAIN}}} = \theta_k - \theta_1$
11: Meta-Test step: $\ell_{\tau_{\text{TEST}}} = \frac{1}{\|\mathcal{B}^T\|} \sum_{(x,y)\in\mathcal{B}^T} \ell(f(x,\ y))$
12: Total gradient: $\nabla_\Sigma = \nabla_{\tau_{\text{TRAIN}}} + \nabla\ell_{\tau_{\text{TEST}}}$
13: Update $\theta_{t+1} \leftarrow V(\theta_t,\ \beta,\ \nabla_\Sigma)$
14: **end for**

source language assuming the epochs across different data sources are synchronised. DG-FMAML requires as much target task data as support task data with the same assumptions. For example, if $k = 10$ then using this $\frac{1}{k}$ proportionality requires 10% of target-language data relative to support. Multilingual training on L languages requires a smaller $\frac{1}{Lk}$ sample per language for equivalent synchronisation. We demonstrate in Section 5.3 that we can use a smaller $< \frac{1}{Lk}$ quantity per target language to increase sample efficiency i.e., the target task epochs can be shorter than support task epochs. This improves the data efficiency of DRAKON without penalty to parser accuracy.

5.1.6 Gradient Analysis of DRAKON

We now analyse the generalisation behaviour within DRAKON using a Taylor Series approximation of gradients. Figure 5.2 outlines the gradients for a single DRAKON step. Following Nichol et al. (2018), we recall the gradient in a single step of the inner loop (Algorithm 1 Line 8) as Equation (5.12). This g_i comprises the loss minimising gradient

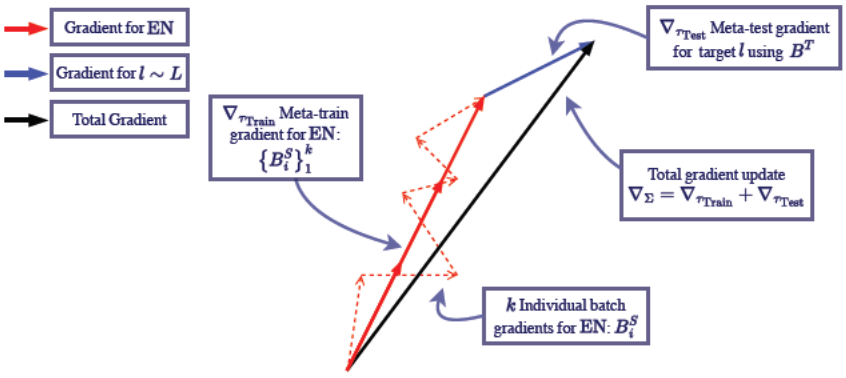

Figure 5.2: Gradients for a single step of learning with DRAKON. Meta-training runs k steps of gradient descent on batches of English data (dashed red). The Meta-train gradient, $\nabla_{\tau_{\text{TRAIN}}}$, is initial and final parameter difference (solid red). The Meta-test step, $\nabla_{\tau_{\text{TEST}}}$, is one step of target language paired data conditioned on the meta-training in this episode (solid blue). The total gradient, ∇_Σ, is a combination of both steps (black).

as an initial component, $\hat{g}_i$, and the product of the Hessian at i with the gradient of prior inner loop steps.

$$g_i = \bar{g}_i - \alpha \bar{H}_i \sum_{j=1}^{i-1} \bar{g}_j + O\left(\alpha^2\right) \tag{5.12}$$

$$\nabla_{\tau_{\text{TRAIN}}} = \theta_k - \theta_1 = \sum_{i=1}^{k} g_i \tag{5.19}$$

Equation (5.16) expresses the meta-train update as the difference between initial and final parameters. This is also equivalent to Equation (5.19) as the sum of each gradient component from θ_1 to θ_k. $\nabla_{\tau_{\text{TRAIN}}}$ comprises both gradients of k steps minimising loss and each step's respective additional terms maximising the inner product of inter-batch gradients.

The meta-test step can be simplified as a $k+1^{\text{th}}$ batch of Reptile training computed on a different data distribution ignored for $\nabla_{\tau_{\text{TRAIN}}}$. Given this framing, we can express the meta-test gradient, $g_{\tau_{\text{TEST}}}$ as Equation (5.20) using the same Taylor expansion. The second term in Equation (5.20), $\bar{H}_{\tau_{\text{TEST}}} \nabla_{\tau_{\text{TRAIN}}}$, highlights how DRAKON improves cross-lingual transfer. This inner product increases the gradient similarity between the target language gradient and *the complete training trajectory*. This improves how cross-lingual capability is learned by maximising the similarity between source and target language gradients. This additional optimisation contribution is explicitly absent from DG-FMAML without approximating the global training direction.

$$g_{\tau_{\text{TEST}}} = \bar{g}_{\tau_{\text{TEST}}} - \alpha \bar{H}_{\tau_{\text{TEST}}} \nabla_{\tau_{\text{TRAIN}}} + O\left(\alpha^2\right) \tag{5.20}$$

As an example, we can express the total gradient update, ∇_Σ, when $k = 2$ as Equation (5.21). Within the parentheses are the intra-task and cross-lingual gradient products as components promoting fast learning across multiple axes of generalisation. We describe the effect of $g_{\tau_{\text{TEST}}}$ as *cross-lingual regularisation* of the solution manifold learned from meta-training.

$$\begin{aligned} \nabla_\Sigma &= g_1 + g_2 + g_{\tau_{\text{TEST}}} \\ &= \bar{g}_1 + \bar{g}_2 + \bar{g}_{\tau_{\text{TEST}}} \\ &- \alpha\left(\bar{g}_2 \bar{g}_1 + \bar{g}_{\tau_{\text{TEST}}}[\bar{g}_1 + \bar{g}_2]\right) + O\left(\alpha^2\right) \end{aligned} \tag{5.21}$$

The critical hyperparameter in DRAKON is the number of inner-loop steps, k, representing a trade-off between the Reptile step complexity and target step frequency. At small k, the Reptile approximation of the global training trajectory may be suboptimal, leading to poor overall learning with frequent $g_{\tau_{\text{TEST}}}$ steps encouraging cross-lingual generalisation. At large k, an improved Reptile trajectory approximation (with higher k) incurs fewer target batch steps per epoch. In this case, insufficient cross-lingual regularisation may now limit target language performance. We observe an empirical trade-off in setting k which we discuss further in Section 5.3. Similar to DG-FMAML, DRAKON does not require additional fine-tuning as the cross-task generalisation is explicit within the algorithm. DRAKON can be naively interpreted as integrating the Reptile meta-training and target language fine-tuning stages. However, the actual advantage of DRAKON is the additional inner product terms in Equation (5.21) promoting cross-lingual generalisation.

5.2 Experiments

5.2.1 Datasets and Few-shot Sampling

MultiATIS++SQL We use MultiATIS++SQL now assuming partial access to the training split of MultiATIS++SQL. Section 2.3.1 details a complete description of MultiATIS++SQL with input-output examples shown in Table 2.5. We use the MultiATIS++SQL multilingual test set for evaluating cross-lingual transfer from English

(EN) to French (FR), Portuguese (PT), Spanish (ES), German (DE), and Chinese (ZH). For text-to-SQL type semantic parsing, there are no standard practices for few-shot data selection. Therefore, we build few-shot samples by randomly sampling some percentage of data and using the target language examples in this group with the remaining examples in English. We examine few-shot sample ratios at 1%, 5%, and 10% of the existing data. With a total training sample of 4473 examples, these sampling rates correspond to 45, 224, and 447 examples in each few-shot percentage respectively. We report the average of five runs to minimise variation from random sampling.

MTOP We also evaluate the MTOP dataset using English as the support task and target languages as the target task. Section 2.3.2 details a complete description of MTOP with input-output examples shown in Table 2.6. We use the MTOP multilingual test set for evaluating cross-lingual from English (EN) to French (FR), Spanish (ES), German (DE), Hindi (HI), and Thai (TH). Unlike for MultiATIS++SQL, there is an existing rationale for few-shot transfer of spoken-language understanding (SLU) dialogue semantic parsing. SCIEM accuracy is discussed further in Section 2.3. We follow the *Samples-per-Intent-and-Slot* (SPIS) strategy from Chen et al. (2020) adapted to our cross-lingual scenario. SPIS randomly selects examples and keeps data that mention any slot and intent value (e.g., "`IN:`" and "`SL:`") with fewer than some rate in the existing subset. Sampling stops when all slots and intents have a minimum frequency of the sampling rate. Practically, an SPIS rate of 1, 5, 10 approximately equates to 284 (1.8%), 1,125 (7.2%), and 1,867 (11.9%) examples for MTOP in each target language. Similar to MultiATIS++SQL, we report the average of five runs.

The SPIS sampling approach approximately normalises the frequency of semantic categories (intent or slot) and ensures all categories are similarly represented in the few-shot sample. Therefore the model will observe most semantic categories a similar number of times to not inherit any label biases in the low-resource domain (or language). This can be interpreted as creating an approximately uniform prior distribution over labels for the target language.

5.2.2 Experimental Setting

We use the DRAKON algorithm to train an encoder-decoder parser following the model design discussed in Section 2.2.1. We compare to previous methods in Chapters 3 to 4 and various robust baselines for few-shot cross-lingual transfer.

5.2.2.1 Setting and Comparison

MULTILINGUAL **Gold** A multilingual Transformer is trained on the union of all professionally translated data. As in Chapters 3 to 4, this represents the **upper bound** for our model using all available data without few-shot constraints.

EN **Only** A monolingual Transformer is trained on only English training data. This model is evaluated on the target language test data with no translation. This is a **zero-shot** baseline for Chapter 4.

TRANSLATE TEST A monolingual Transformer is trained on source English data ($\mathcal{S}_{\text{EN}}$). Machine translation is used to translate test data from target languages into English. Logical forms are predicted from translated data using the English model. This is the **silver-standard** baseline method for Chapter 3. As in Chapter 4, we report this baseline using OPUS translation (Tiedemann and Thottingal, 2020).

TRANSLATE TRAIN Machine translation is used to translate English training data into each target language as described in Chapter 3. A monolingual Transformer is trained on translated training data and logical forms are predicted using this model. This is the **silver-standard** baseline method for Chapter 3. As in Chapter 4, we report this baseline using OPUS translation (Tiedemann and Thottingal, 2020)

FATES Our proposal for machine translation from Chapter 3 using multiple encoders and multiple MT engines for cross-lingual semantic parsing without gold training data. This is the best **silver-standard** method from Chapter 3.

ZEUS Our zero-shot multi-task model from Chapter 4 using auxiliary data for cross-lingual latent representation alignment. We note that ZEUS already performs close to the upper bound but few-shot sampling may be more data efficient than leveraging large corpora for auxiliary tasks. This is the best **zero-shot** method from Chapter 4.

Train-EN∪All A Transformer is trained on English data and samples from all target languages together in a single stage i.e., $\mathcal{S}_{\text{EN}} \cup \mathcal{S}_L$. This is superior to training without English (on $\mathcal{S}_L$ only), we contrast to this approach for a more competitive comparison. This is a baseline **few-shot** method.

Train-EN→FT-All Similar to Train-EN∪All, but in successive stages. A model is trained on $\mathcal{S}_{\text{EN}}$ and then fine-tuned on target samples, $\mathcal{S}_L$. This is similar to the meta-learning paradigm using simpler Adam optimisation for the first stage. This is a baseline **few-shot** method.

Reptile-EN→FT-All Initial training uses Reptile (Nichol et al., 2018) on English support data followed by fine-tuning on target samples, $\mathcal{S}_L$. This is a typical usage of Reptile for training a low-resource multi-domain parser (Chen et al., 2020). As mentioned above, this method is the same constituent meta-training and meta-testing steps as DRAKON split into distinct training phases. This is a baseline **few-shot** method.

DG-FMAML We compare to DG-FMAML (Wang et al., 2021a) in Section 5.3.3 in our analysis and study of tuning the k hyperparameter. As previously mentioned, DG-FMAML is a special case of DRAKON when $k = 1$. This is a baseline **few-shot** method.

5.2.2.2 Model Training

As in previous chapters, we follow the Transformer encoder-decoder setup from Section 2.2.2. The encoder, Q_ϕ, is pre-trained using the encoder parameters from the MBART50 pre-trained model (Tang et al., 2021). This model has a single Transformer decoder, G_ψ, trained from scratch.

We focus on the algorithmic contribution of DRAKON by fixing the same model for all experiments. This isolates the contribution of the training algorithm only. Experimental hyperparameters were tested at the few-shot rate of 1% for MultiATIS++SQL and then applied to all experiments. For the inner optimisation loop in Algorithm 1, we use Stochastic Gradient Descent (SGD) with a learning rate, α, of 1×10^{-4}. Using an adaptive optimiser in the inner loop (e.g., Adam) was observed to degrade performance. As we required to reset the inner-loop optimiser for each inner loop (i.e., after the k outer step), we found no benefit to an adaptive optimiser which struggle to learn adaptive learning statistics over $< k$ steps. In contrast, the outer optimisation loop uses the Adam-based optimisation setting, defined in Section 2.2.2, where the optimiser learns adaptive learning statistics over the outer-loop steps defined by total updates e.g., Equation (5.21). We did not experiment with learning different normalisation layers for inner and outer loops as our network does not use batch-level normalisation. We suggest this may be necessary for other architectures. We also did not weight decay

during learning as both were observed to damage the meta-training gradient update. We optimise the episode length to $k = 10$ and discuss this setting in Section 5.3.3. We train DRAKON for 100 epochs over the inner-loop data with early stopping after 50 epochs using a validation loss criterion. We identify that DRAKON requires a minimum of 50 epochs of training to produce superior cross-lingual transfer outcomes.

5.3 Results

We train a parser using the DRAKON algorithm described in Section 5.1 in the experimental settings outlined in Section 5.2. As DRAKON is an algorithm contribution, we focus on evaluating DRAKON to verify the utility of our method compared to training methods with the same model and data. This differs from Chapters 3 to 4 detailing modelling contributions for the same problem.

5.3.1 Is a Few-shot Transfer Methodology Competitive?

We compare DRAKON to upper- and lower-bounds in Table 5.3. As in previous chapters, the upper-bound is 'MULTILINGUAL Gold' training on all available data. The lower bounds are machine-translation methods using OPUS (Tiedemann and Thottingal, 2020), and the zero-shot 'EN Only' baseline.

Lower Bound Baselines: We compare DRAKON across three sampling rates to translation and zero-shot baselines in Table 5.2. For both MultiATIS++SQL and MTOP, a few-shot method with the smallest sample size (1% or 1 SPIS respectively) surpasses both translation methods and zero-shot transfer from English.

For MultiATIS++SQL, DRAKON @1% demonstrates a +12.4% and +13.2% improvement relative to TRANSLATE TRAIN and TRANSLATE TEST respectively. Similarly for MTOP, the DRAKON @1 SPIS parser improves over translation by 33.5% and 6.2% for TRANSLATE TRAIN and TRANSLATE TEST. DRAKON is superior to synthetic data even with very few gold examples. The DRAKON @1% improves over zero-shot 'EN only' by $\geq$ 21.7% and $\geq$ 22.2% for MultiATIS++SQL and MTOP respectively. Unsurprisingly, few-shot transfer supersedes zero-shot transfer without representation alignment. For either dataset, increasing the sample size further improves accuracy gain over all baselines. Analysing the performance gap between DRAKON and baselines, we identify that correctly interpreting entities is the primary benefit of DRAKON. Baseline

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
TRANSLATE TRAIN OPUS	—	56.8	39.1	51.8	60.4	59.6	53.5
TRANSLATE TEST OPUS	—	57.7	58.1	58.3	58.8	50.9	56.8
EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
DRAKON @1%	73.8±0.3	70.4±1.8	70.8±0.7	68.9±2.3	69.1±1.2	68.1±1.2	69.5±1.1
DRAKON @5%	74.4±1.3	73.0±0.9	71.6±1.1	71.6±0.7	71.1±0.6	69.5±0.5	71.4±1.3
DRAKON @10%	75.8±1.3	74.2±0.2	72.8±0.6	72.1±0.7	73.0±0.6	72.8±0.5	**73.0**±0.8

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
TRANSLATE TRAIN OPUS	—	24.4	23.1	32.7	22.4	9.5	22.4
TRANSLATE TEST OPUS	—	44.9	63.1	39.1	47.1	54.2	49.7
EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
DRAKON @1 SPIS	71.8±2.0	59.0±1.7	61.1±1.1	59.4±1.6	56.1±1.4	43.9±2.3	55.9±6.9
DRAKON @5 SPIS	72.1±1.9	65.4±2.6	67.0±4.6	65.3±2.6	63.0±0.1	49.8±0.8	62.1±7.0
DRAKON @10 SPIS	72.5±0.4	65.6±0.5	67.5±0.8	65.8±0.6	63.8±1.1	50.6±1.1	**62.7**±6.9

(b) MTOP

Table 5.2: Comparisons between DRAKON to lower-bounds for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy. We compare DRAKON to machine-translation baselines and the zero-shot transfer performance from training only on English. DRAKON must surpass these lower-bound baselines to justify the few-shot approach. For few-shot methods, we report the average over five different few-shot data splits $\pm$ the standard deviation across runs. The significant best result is bolded.

methods can mistranslate or omit fluent references to entities in target languages. While this is inconsequential for some NLU tasks (Conneau et al., 2018b), this error is often critical for semantic parsing. DRAKON improves how observed, or unobserved, entities are modelled during cross-lingual transfer. We expand on this discussion as error analysis in Section 5.3.5.

Despite overall improvement, we note that the standard deviation for DRAKON @1 SPIS overlaps with TRANSLATE TEST despite significant improvement overall. TRANSLATE TEST is the strongest baseline for both datasets by modelling the task entirely in English. As translation into English is often superior to translation from English (Moghe et al., 2023b), we anticipate the highest quality translation from this baseline. We identify in Table 5.4 that data quality and our algorithmic improvements to training are required to no longer overlap with TRANSLATE TEST. We note that DRAKON does not surpass the silver-standard LLM-based methods we discuss in Chapter 3. Further improvements to sample-efficient few-shot generalisation are needed to compete with larger models using synthetic data. We revisit this comparison in Chapter 6 where our contribution is more competitive.

Upper Bounds and Prior Methods We compare DRAKON to the 'MULTILINGUAL Gold' upper-bound, FATES from Chapter 3, and ZEUS from Chapter 4 in Table 5.3. At higher sampling rates, DRAKON approaches the upper bound at only -0.2% or -3.3% difference for MultiATIS++SQL or MTOP respectively. Considering 'MULTILINGUAL Gold' uses $\geq 10\times$ the data of DRAKON with similar performance — this result highlights the benefit of our optimisation strategy targeting cross-lingual transfer beyond simply adding more data. We visualise this comparison in Section 5.3.4.

DRAKON performs above FATES even at the smallest sampling ratios: $+2.0\%$ for MultiATIS++SQL and $+18.8\%$ for MTOP. We infer that small gold data samples are more valuable than larger samples of "silver-standard" data from machine translation for our case study. In contrast, we find that DRAKON does not improve on ZEUS at the smallest sampling ratio and requires additional data (i.e., 5% rate for MultiATIS++SQL or 5 SPIS rate for MTOP[2]) for significantly improved parsing than ZEUS. At a surface level, this raises that zero-shot transfer can be superior to few-shot transfer. However, the best setting of ZEUS requires $> 50,000$ target language utterances for latent representation alignment; whereas DRAKON uses only hundreds of annotated examples

[2]We henceforth refer to the sampling ratio of X% or X SPIS, for MultiATIS++SQL and MTOP respectively, as X% for clarity.

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
MULTILINGUAL Gold	74.9	74.2	73.0	70.4	74.6	73.7	73.2
FATES (best)	74.9	70.5	69.2	62.4	68.7	66.5	67.5
ZEUS (best)	74.4	72.3	69.7	68.5	69.0	69.2	69.7
DRAKON @1%	73.8±0.3	70.4±1.8	70.8±0.7	68.9±2.3	69.1±1.2	68.1±1.2	69.5±1.1
DRAKON @5%	74.4±1.3	73.0±0.9	71.6±1.1	71.6±0.7	71.1±0.6	69.5±0.5	71.4±1.3
DRAKON @10%	75.8±1.3	74.2±0.2	72.8±0.6	72.1±0.7	73.0±0.6	72.8±0.5	**73.0**±0.8

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
MULTILINGUAL Gold	75.5	69.7	72.4	67.9	65.5	54.6	66.0
FATES (best)	69.7	44.7	45.7	49	32.9	13.4	37.1
ZEUS (best)	77.5	66.2	67.4	64.2	59.4	47.7	61.9
DRAKON @1 SPIS	71.8±2.0	59.0±1.7	61.1±1.1	59.4±1.6	56.1±1.4	43.9±2.3	55.9±6.9
DRAKON @5 SPIS	72.1±1.9	65.4±2.6	67.0±4.6	65.3±2.6	63.0±0.1	49.8±0.8	62.1±7.0
DRAKON @10 SPIS	72.5±0.4	65.6±0.5	67.5±0.8	65.8±0.6	63.8±1.1	50.6±1.1	**62.7**±6.9

(b) MTOP

Table 5.3: Comparisons between DRAKON to upper-bound training and the best methods from previous Chapters for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy. We compare between (a) training on gold-standard translations in all target languages (MULTILINGUAL Gold); (b) FATES from Chapter 3; and the ZEUS zero-shot model from Chapter 4.For few-shot methods, we report the average over five different few-shot data splits $\pm$ the standard deviation across runs. The significant best result is bolded.

for similar performance. We could therefore consider DRAKON as more data-efficient than ZEUS, and potentially more economical depending on the annotation costs of the few-shot sample.

Our findings compared to baselines and previous chapters further highlight the importance of data *quality*. We observe that small samples of gold annotated translations are more valuable for training than either larger quantities of synthetic data or alternative gold data from auxiliary tasks. Few-shot cross-lingual transfer yields the best outcomes with pragmatic costs for our case study.

5.3.2 Is DRAKON the Best Few-Shot Generalisation Strategy?

We compare DRAKON to established learning algorithms in Table 5.4 for MultiATIS++SQL and MTOP. This evaluates the DRAKON algorithm relative to other methods with identical models and data resources. Across comparison methods, we observe that single-stage training ('Train-EN∪All') underperforms compared to two-stage training ('Train-EN→FT-All' or 'Reptile-EN→FT-All') at every sampling ratio. We suggest that English data may "dominate" the few-shot sample in this approach for poorer target language accuracy. The strongest comparison is the 'Reptile-EN→FT-All' model, highlighting that meta-learning is a competitive baseline for single-task optimisation.

DRAKON significantly improves cross-lingual generalisation across all languages at equivalent and lower sample sizes. For MultiATIS++SQL at 1%, DRAKON improves by an average +15.7% over the closest comparison, 'Reptile-EN→FT-All', and +27.1% for the weakest 'Train-EN∪All' comparison. Similarly at 5%, we find +9.8% gain, and at 10%, we find +8.9% relative to the 'Reptile-EN→FT-All' competitor. Similarly for MTOP, the benefit of DRAKON is +6.6% at 1%, +2.2% at 5% and +1.2% at 10% over 'Reptile-EN→FT-All'. The benefit of DRAKON is broadly weaker for MTOP than for MultiATIS++SQL. This can be attributed to more challenging cross-lingual transfer for MTOP where utterance tokens are inserted into the logical form (see Section 2.3.2).

Contrasting across sample sizes — the most accurate version of DRAKON uses @10% or 10 SPIS sampling for MultiATIS++SQL and MTOP respectively. However, the 5% sampling ratio performs more similarly to 10% than the lowest sampling ratio. Sampling at 5% is only −1.6% or −0.6% below the best performance for MultiATIS++SQL and MTOP respectively. This benefit is smaller than the +1.9% or +6.2% improvement between 1% and 5%. The decreasing marginal benefit of additional data highlights a trade-off between the value of high-quality data for parsing and the

	Training Algorithm	EN	FR	PT	ES	DE	ZH	TARGET AVG.
@1%	Train-EN∪All	69.7±1.4	44±3.5	42.2±3.7	38.3±6.8	45.8±2.6	41.7±3.6	42.4±2.8
	Train-EN→FT-All	71.2±2.3	53.3±5.2	49.7±5.4	56.1±2.7	52.5±6.7	39.0±4.0	50.1±6.6
	Reptile-EN→FT-All	73.2±0.7	58.9±4.8	54.8±3.4	52.8±4.4	60.6±3.6	41.7±4.0	53.8±7.4
	DRAKON	73.8±0.3	70.4±1.8	70.8±0.7	68.9±2.3	69.1±1.2	68.1±1.2	**69.5**±1.1
@5%	Train-EN∪All	67.3±1.6	55.2±4.5	54.7±4.5	44.4±4.5	55.8±2.9	52.3±4.3	52.5±4.7
	Train-EN→FT-All	69.2±1.9	58.9±5.3	54.8±5.4	52.8±4.5	60.6±6.5	41.7±9.5	53.8±7.4
	Reptile-EN→FT-All	69.5±1.8	65.3±3.8	61.3±6.0	59.6±2.6	64.9±5.1	56.9±9.2	61.6±3.6
	DRAKON	74.4±1.3	73.0±0.9	71.6±1.1	71.6±0.7	71.1±0.6	69.5±0.5	**71.4**±1.3
@10%	Train-EN∪All	65.7±1.9	61.5±1.7	62.1±2.3	53.7±3.2	62.7±2.3	60.6±2.4	60.1±3.7
	Train-EN→FT-All	67.4±1.9	63.8±5.8	60.3±5.3	59.6±4.0	64.5±6.5	58.4±6.4	61.3±2.7
	Reptile-EN→FT-All	72.8±1.8	66.3±4.2	64.6±4.9	62.3±6.4	66.6±5.0	60.7±3.6	64.1±2.6
	DRAKON	75.8±1.3	74.2±0.2	72.8±0.6	72.1±0.7	73.0±0.6	72.8±0.5	**73.0**±0.8

(a) MultiATIS++SQL

	Training Algorithm	EN	FR	ES	DE	HI	TH	TARGET AVG.
@1 SPIS	Train-EN∪All	71.5±8.6	44.2±5.2	44.8±2.1	44.3±5.5	42.1±10.2	31.8±6.4	41.4±5.5
	Train-EN→FT-All	66.8±4.0	45.8±5.7	45.7±8.2	45.6±5.5	44.5±8.5	34.4±3.3	43.2±4.9
	Reptile-EN→FT-All	70.7±2.2	52.2±0.7	53.9±5.2	53.1±1.9	50.1±5.8	37.3±7.1	49.3±6.9
	DRAKON	71.8±2.0	59.0±1.7	61.1±1.1	59.4±1.6	56.1±1.4	43.9±2.3	**55.9**±6.9
@5 SPIS	Train-EN∪All	71.8±4.7	53.7±3.3	54.2±5.9	55.3±4.3	52.6±3.2	41.4±2.9	51.5±5.7
	Train-EN→FT-All	71.2±2.9	58.6±7.3	60.2±4.2	59.1±6.0	55.3±7.4	43.8±10.7	55.4±6.8
	Reptile-EN→FT-All	71.9±2.7	63.2±0.7	65.2±6.6	63.2±3.1	60.7±3.3	47.1±1.3	59.9±7.3
	DRAKON	72.1±1.9	65.4±2.6	67.0±4.6	65.3±2.6	63.0±0.1	49.8±0.8	**62.1**±7.0
@10 SPIS	Train-EN∪All	69.3±1.6	60.0±2.3	60.9±3.3	61.2±2.2	59.3±3.4	45.2±1.3	57.3±6.8
	Train-EN→FT-All	71.4±3.4	61.2±2.4	63.4±1.1	61.6±3.6	58.7±2.3	46.1±2.9	58.2±6.9
	Reptile-EN→FT-All	71.8±4.5	64.5±2.4	66.1±3.4	64.8±6.5	62.0±7.2	50.2±6.6	61.5±6.5
	DRAKON	72.5±0.4	65.6±0.5	67.5±0.8	65.8±0.6	63.8±1.1	50.6±1.1	**62.7**±6.9

(b) MTOP

Table 5.4: Comparisons between DRAKON and few-shot training algorithms outlined in Section 5.2.2 for (a) MultiATIS++SQL and (b) MTOP. We compare each algorithm on identical data splits across three realistic low-resource sampling scenarios for each dataset. Our results demonstrate the benefit of DRAKON compared to adjacent methods using the same data resources. For few-shot methods, we report the average over five different few-shot data splits ± the standard deviation across runs. The significant best result is bolded.

diminishing results of additional annotation. A 5% sample is sufficient to improve on our methods from previous chapters and approximate the most accurate model with less data.

Comparing the benefit of greater sampling across methods, DRAKON is more 'stable': offering accurate parsing at lower sampling rates and marginal increases from more data. This contrasts with other methods of gaining +17.7%, +11.2% or +10.3% improvement between @1% and @10% on MultiATIS++SQL for 'Train-EN∪All', 'Train-EN→FT-All', and 'Reptile-EN→FT-All' respectively. As every method improves with more data, the higher gain for each comparison highlights the inadequacy with few samples, rather than the benefit of additional examples. Notably, the improvement of DRAKON over 'Reptile-EN→FT-All', which is the same optimisation split into different training stages, highlights how our regularised manifold learning approach is superior to learning a manifold through meta-training and later fine-tuning this to target languages.

Across languages at 1% sampling, DRAKON improves primarily for languages dissimilar to English to better minimise the cross-lingual transfer gap. For MultiATIS++SQL ZH, we see that DRAKON @1% is +26.4% above the closest baseline. This contrasts with the smallest gain of +8.5% MultiATIS++SQL DE. Our improvement also yields less variability across target languages—the standard deviation across languages for DRAKON @1% is $\sigma = 1.1$, compared to $\sigma = 2.8$ for 'Train-EN∪All' or $\sigma = 7.4$ for 'Reptile-EN→FT-All'.

5.3.3 Which Hyperparameters are Critical for DRAKON?

We report ablations on DRAKON as a case study on MultiATIS++SQL in Figure 5.3. We vary the inner loop size hyperparameter k at 5% sampling for Figure 5.3(a) and vary the sampling ratio with fixed $k = 10$ for Figure 5.3(b). As previously discussed, the 5% sampling rate is sufficient to perform above all baselines and methods from previous chapters. All discussed findings extend to the MTOP dataset but are omitted for brevity.

Influence of k on Performance In Figure 5.3(a), we study the influence of hyperparameter k (the inner-loop size in Algorithm 1 or the batches approximating the *solution manifold* trajectory) on target language accuracy. When $k = 1$, DRAKON is equivalent to DG-FMAML (Wang et al., 2021a) and approximates the training trajectory using a single batch. We observe $k = 1$ is suboptimal across all target languages and suggest that DRAKON is broadly an improvement to DG-FMAML. We empirically observe

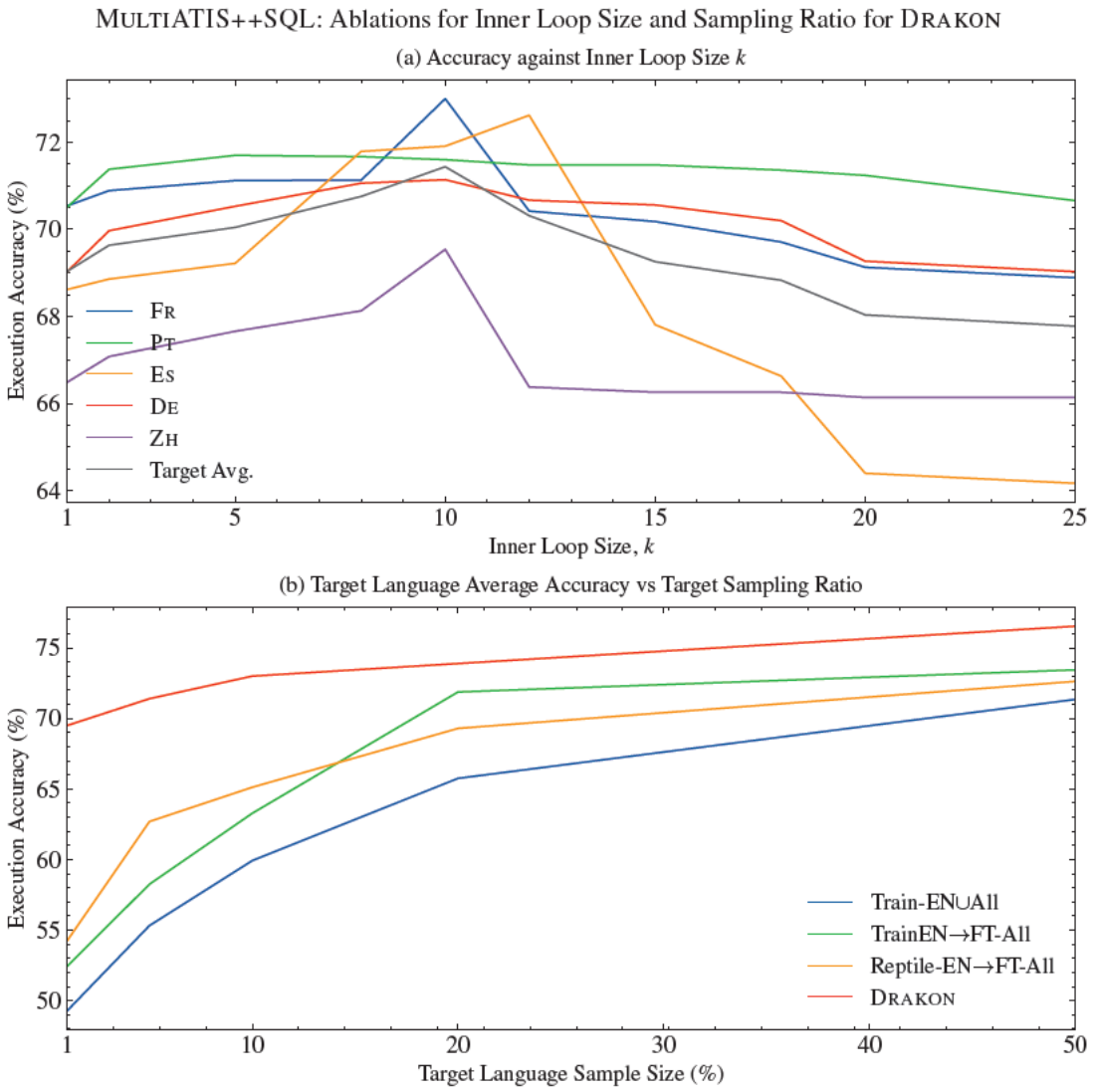

Figure 5.3: Ablation experiments on MultiATIS+SQL sampling for (a) accuracy against inner loop size k at 5% sampling, and (b) average target language accuracy against few-shot sample size relative to support dataset from 1% to 50%. For (a), the $k = 1$ case is equivalent to DG-FMAML (Wang et al., 2021a).

that increasing k benefits performance by encouraging cross-lingual generalisation with the gradient trajectory instead of a single batch. However, as discussed in Section 5.2, increasing k also decreases the frequency of the outer step within an epoch—leading to poor cross-lingual transfer at high k. We identify a stable operating regime for setting k around $k = 10$ where performance is approximately similar. Given this setting of k, the target sample size must be 10% of the support sample size for synchronised training epochs. However, Table 5.4 identifies DRAKON as the most robust algorithm for "over-sampling" smaller target samples for resource-constrained cross-lingual transfer.

Performance with Larger Sampling Ratios We consider a wider range of target data sample sizes between 1% to 50% in Figure 5.3(b). Baseline approaches converge in performance between 71.0% and 73.5% at 50% target sample size. The comparative benefit of DRAKON maintains at higher sample sizes with an accuracy of 76.5%. The benefit of DRAKON is still greatest at low sample sizes; however, we maintain a +2.6% gain over the closest system at 50%. While low sampling is the most economical, the consistent benefit of DRAKON suggests an overall benefit using DRAKON for cross-lingual optimisation.

5.3.4 Visualising Latent Representation Similarity

Analysis of DRAKON in Section 5.1.6 presupposes that first-order meta-learning creates a dense high-likelihood sub-region in the parameters (i.e., *solution manifold*). Under these conditions, representations of target languages should cluster close to representations for the support task, given the optimisation combining the gradients for both tasks. This should allow for rapid adaptation with minimal samples. This contrasts with methods without meta-learning, lacking guidance on representation density or gradient similarity. However, metrics in Tables 5.2 to 5.4 do not directly study if this intended effect arises. To this end, we visualise the latent encodings of the Multi-ATIS++SQL test inputs following Section 2.4.2. In Figure 5.4, we compare to other few-shot methodologies, and in Figure 5.5, we compare to our prior methods and the upper/lower bounds. We identify that DRAKON produces less monolingual clustering artefacts than FATES or ZEUS. We conjecture that cross-lingual similarity is a proxy for manifold alignment—our goal is accurate cross-lingual transfer using closely aligned representations from source and target languages (Xia et al., 2021).

We quantitatively analyse the relationship between latent representations in Table 5.5.

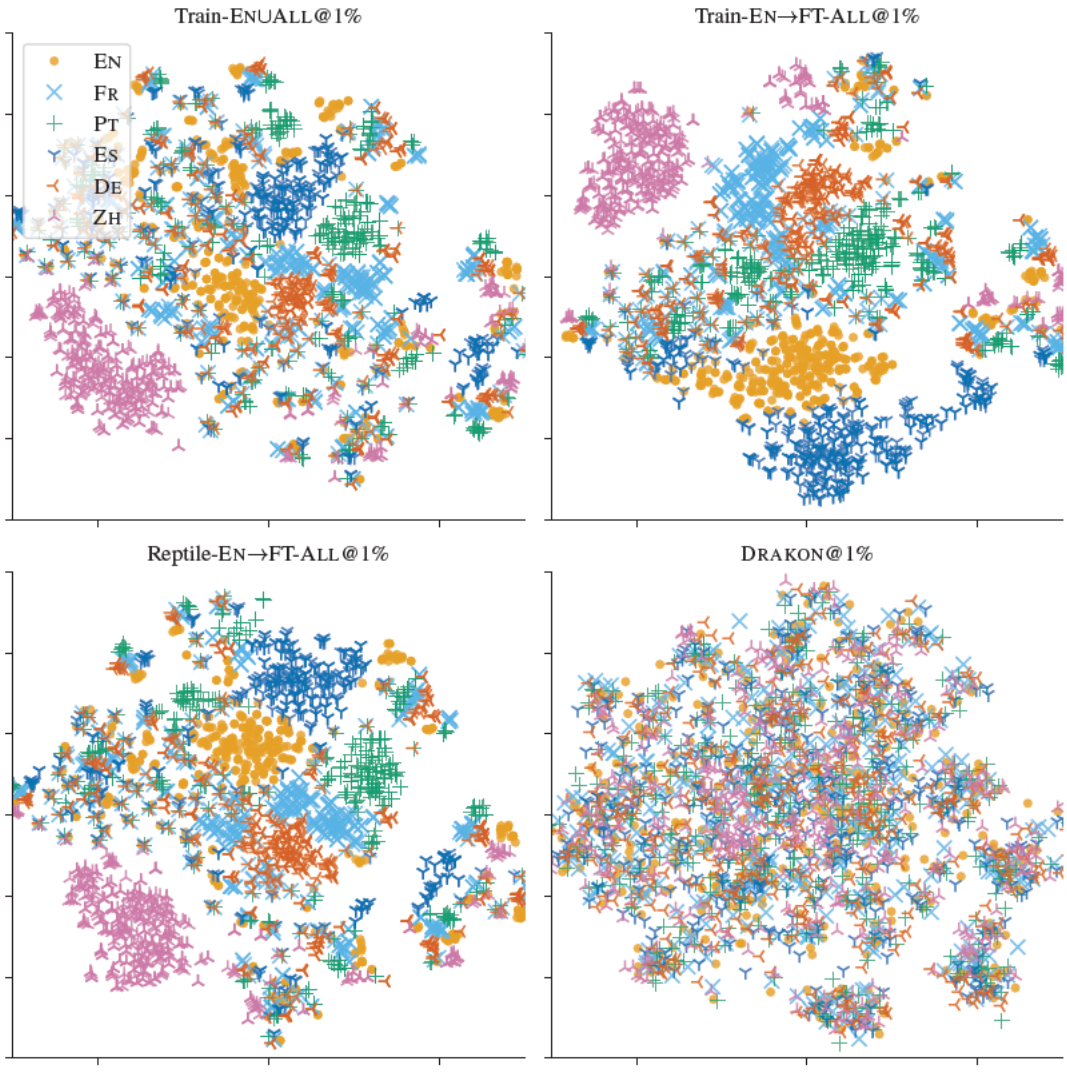

Figure 5.4: Visualisation of MultiATIS++SQL encodings (test set; 50% random parallel sample) using t-SNE for few-shot training algorithms (@1% sampling from Table 5.4). We identify the regularised parameter manifold improves cross-lingual transfer with improved cross-lingual latent representation similarity using DRAKON.

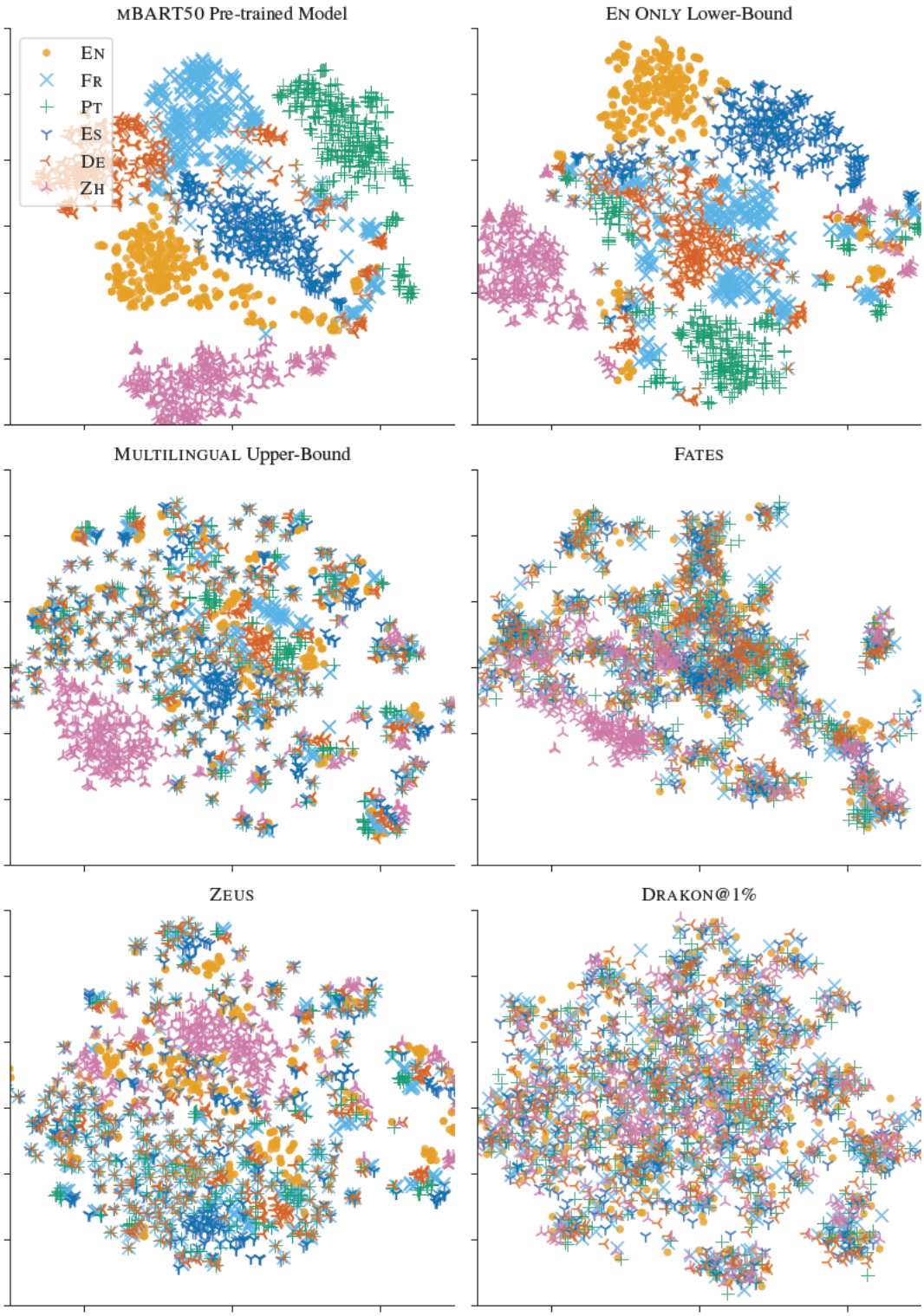

Figure 5.5: Visualisation of MultiATIS++SQL encodings (test set; 50% random parallel sample) using t-SNE. We compare the original MBART50 pre-trained model, the EN-ONLY zero-shot lower bound, MULTILINGUAL training upper bound, FATES from Chapter 3, ZEUS from Chapter 4, and DRAKON. DRAKON improves cross-lingual similarity more than our previous proposals using a few-shot sample of target language data.

Method Type	Method	Cosine (↑)	Top-1	Top-5	Top-10
Pre-trained Model	MBART50	0.576	0.521	0.745	0.796
Lower Bound	EN Only	0.364	0.669	0.775	0.964
Upper Bound	MULTILINGUAL Gold	0.698	0.784	0.981	0.991
Machine Translation	FATES	0.670	0.720	**0.957**	**0.992**
Zero-shot	ZEUS	0.760	**0.832**	0.944	0.971
Few-shot @1% sampling	Train-EN∪ALL	0.470	0.634	0.835	0.877
	Train-EN→FT-ALL	0.541	0.714	0.898	0.922
	Reptile-EN→FT-ALL	0.673	0.702	0.920	0.946
	DRAKON	**0.844**	0.797	0.949	0.963

Table 5.5: Average similarity between encodings of English and target languages for MultiATIS++SQL. Cosine similarity evaluates average distance between encodings of parallel sentences. Top-k evaluates if the parallel encoding is ranked within the k most cosine-similar vectors (higher (↑) is better). Best excluding the upper-bound is bold.

We observe that some few-shot baselines have a weaker cosine similarity but stronger nearest neighbour similarity compared to the pre-trained MBART50 model. Our visualisation and these results support that fine-tuning modifies the global structure of the latent space with potentially detrimental effects on representation similarity. DRAKON reports the strongest cosine similarity and Top-k similarity for all $k = \{1, 5, 10\}$ across few-shot methods. DRAKON is also competitive to ZEUS in most metrics, but is not strictly improving compared to our best zero-shot method. This is similar to the variation in performance observed in Table 5.3. Notably, DRAKON reports a higher cosine similarity than the 'MULTILINGUAL Gold' upper bound. We suggest that this reflects the benefit of promoting cross-lingual similarity during training, similar to the findings of Chapter 4. These findings support that DRAKON legitimately learns some manifold structure during training and this *regularised manifold* improves cross-lingual representation alignment within the model.

5.3.5 Error Analysis

We examine the improvement from DRAKON, studying where our method improves over few-shot and translation baselines as an error analysis. Similar to previous chapters,

EN	Show me all flights from San Jose to Phoenix
FR	Me montrer tous les vols de San Josè á Phoenix
×	SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'SAN FRANCISCO' AND flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'PHILADELPHIA';
✓	SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'SAN JOSE' AND flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'PHOENIX';

Figure 5.6: Contrast between SQL from a French input from MultiATIS++SQL for Train-EN∪All and DRAKON. The entities “San José” and “Phoenix” are not observed in the 1% sample of French data but are mentioned in the English support data. The Train-EN∪All approach fails to connect attributes seen in English when generating SQL from French inputs (×). Training with DRAKON better leverages support data to generate accurate SQL from other languages (✓).

we sample 50 MultiATIS++SQL test examples where DRAKON correctly predicted the outputs but the baselines failed. Similar to Figure 5.4, we consider the results of models using 1% sampling as the worst-case performance.

Entity Resolution Accurate semantic parsing requires sophisticated entity handling to translate mentioned proper nouns from utterance to logical form. In our few-shot sampling scenario, *most* entities will appear in the English support data (e.g., "Denver" or "American Airlines"), and *some* will be mentioned within the target language sample (e.g. "Mineápolis" or "Nueva York" in Spanish). These samples are unlikely to include all possible entities from random sampling. Effective cross-lingual learning must "connect" these entities from support to the target task, such that these names can be parsed when predicting SQL from the target language. As shown in Figure 5.6, the failure to recognize entities from support data, for inference on target languages, is a critical failing of all models besides DRAKON. The improvement in cross-lingual similarity using DRAKON expresses a specific improvement in entity recognition. Compared to the worst performing model, Train-EN∪All, 56% of improvement accounts for handling entities absent from the 1% sample but present in the 99% English support data. While DRAKON can generate accurate SQL, other models are limited in expressivity to fall back on using seen entities from the 1% sample. This notably accounts for 60% of improvement in parsing Chinese, with minimal orthographic overlap to English, indicating that DRAKON better leverages support data without reliance on token similarity. In 48% of improved parses, entity mishandling is the *sole error* — highlighting how limiting poor entity interpretation is for our task.

Modifier Phrases Our model also improves handling of novel *modifiers* (e.g. "on a weekday", "round-trip") absent from target language samples. Modifiers are often realized as additional sub-queries and filtering logic in SQL outputs. Comparing DRAKON to Train-EN∪All, 34% of improvement is related to modifier handling. Less capable systems fall back on modifiers observed from the target language sample or ignore them entirely to generate inaccurate SQL. While DRAKON better links parsing knowledge from English to target languages— the overall task is not solved. Outstanding errors in all languages primarily relate to query complexity. Future improvement can address parsing complex natural language expressions from multiple languages (often in some language-specific vernacular) into accurate SQL.

5.4 Related Work

Meta-learning in NLP has become a prominent toolkit for generalisation to new tasks and domains using fewer examples (Wang et al., 2021b; Lee et al., 2021; Hedderich et al., 2021; Zhao et al., 2021). The formulation of "learning to learn" iteratively improves both the relevant task(s) and the speed at which these tasks are mastered. The Model-Agnostic Meta-learning (Finn et al., 2017, MAML) methodology has been widely studied as a form of meta-learning only modifying the training process for any model. MAML optimises a model to rapidly learn multiple known tasks to efficiently learn an unseen task at test time with minimal examples.

Gu et al. (2018) propose an early demonstration of MAML within NLP for translation of languages with scarce labelled data. Gu et al. (2018) train a neural machine translation model on high-resource language bitext using MAML. A model trained using MAML can rapidly learn translation into a low-resource language or language pairs with little parallel data. Similarly, Nooralahzadeh et al. (2020) propose X-MAML as a learning framework for cross-lingual meta-learning. X-MAML defines a two-stage optimisation episode sampling high- and low-resource languages to iterate meta-train and meta-test steps. X-MAML is similar to DRAKON, but our method differs in the usage of Reptile to efficiently estimate the meta-train gradient using *multiple* batches of the high-resource language (i.e., support task).

Model-based meta-learning has also been proposed for cross-lingual transfer. This differs from MAML in learning additional parameters for the "learning to learn" goal. Xia et al. (2021) propose a 'representation transformation network' to transform the contextual embedding output of XLM-Roberta (Conneau et al., 2020) from low-resource languages into the embedding space for a high-resource language for sequence labelling and classification tasks. This allows the low-resource language to benefit from the pre-existing informative representation structures learned from languages with more data (Singh et al., 2019b). For similar tasks, Xu et al. (2021) propose another model-based method meta-learning layer-wise learning rates to selectively freeze model components. Layers are frozen dependent on activation magnitudes to determine which layers are the most or least useful during simulated zero-shot cross-lingual transfer.

Within semantic parsing, MAML-type learning has demonstrated accurate parsing on unseen domains and datasets. Huang et al. (2018) hypothesise that the data distributions for each MAML step should be similar to improve generalisation. They propose to construct meta-train and meta-test steps for MAML by sampling a meta-

test step and using a 'relevance function' to construct a meta-train batch of similar queries. This benefits generalisation to unseen tables in WikiSQL (Zhong et al., 2017). Conklin et al. (2021) propose an alternative strategy: sampling *easy* and *hard* batches for the meta-train and meta-test MAML steps respectively. Explicitly constructing a more challenging distribution shift demonstrates broad improvement in compositional generalisation benchmarks. As previously discussed, Wang et al. (2021a) propose Domain-Generalisation MAML (DG-MAML) and Domain-Generalisation First-Order MAML (DG-FMAML) for adaptation to unseen SQL tables in the Spider dataset (Yu et al., 2018b). DG-MAML applies the same principles as Conklin et al. (2021) using different databases to simulate a distribution shift and ultimately improve generalisation to an unseen SQL table during inference. Chen et al. (2020) highlight the benefit of training with Reptile for task-oriented semantic parsing. Chen et al. (2020) identify that training on high-resource domains using Reptile improves the sample efficiency of the final fine-tuning stage for adaptation to low-resource domains. This is approximately equivalent to the 'Reptile-EN→FT-All' comparison discussed in Section 5.2.2. However, our results show that integrating training on high-resource languages with adaptation to target languages improves task performance and eliminates a secondary training phase. DRAKON, originally presented in Sherborne and Lapata (2023), is the only method to integrate Reptile-based meta-train steps with out-of-distribution (i.e., cross-lingual) meta-test steps for a multilingual model without further fine-tuning.

Since the original publication of this work, **follow up research** has furthered investigation into very low-resource cross-lingual transfer. (Agrawal et al., 2023) use language language models for question answering with only five examples in target languages. Wu et al. (2023) also investigate how improving your meta-training tasks can better optimise cross-lingual transfer with fewer examples during cross-lingual meta learning.

5.5 Summary

In this chapter, we shift focus to a few-shot cross-lingual semantic parsing framework to eliminate the data quality issues observed in prior chapters. We evaluate meta-learning for sample-efficient and model-agnostic few-shot cross-lingual transfer. We propose DRAKON to address a hypothesis that the approximate *solution manifold* outcome from meta-learning can be regularised for cross-lingual transfer using periodic training steps on target languages as regularisation. By deriving the underlying gradients from

training with DRAKON, we find that our method optimises the product of both intra- and inter-language gradients between batches.

Empirical evidence over MultiATIS++SQL and MTOP identifies that the few-shot paradigm is more sample-efficient and accurate than our previously proposed techniques and all baselines. Ablations and analysis identify an optimal trade-off between training episode length (i.e., meta-test step frequency) for overall learning and cross-lingual regularisation outcomes. We also observe that data sampling and episode length are largely decoupled for improved sample efficiency using fewer target language examples. Revisiting the analysis of the latent space also identifies that DRAKON develops representations which follow a semantically distributed latent structure. Quantitative analysis further supports that the cross-lingual latent similarity improves beyond prior methods. While DRAKON demonstrates the smallest cross-lingual transfer gap seen yet in this thesis; there may exist alternative methods which can better exploit the lowest sampling ratios where DRAKON was weakest. In the next chapter, we revisit the motivation of representation alignment to propose an alternative few-shot strategy with improved data efficiency at the lowest sampling rates.

Chapter 6

Optimal Transport for Cross-lingual Posterior Alignment

In previous chapters, we proposed methods for developing a parser using machine translation (Chapter 3), auxiliary tasks (Chapter 4), and meta-learning (Chapter 5). At present, our findings identify the value of gold-standard target language data for accurate and generalisable cross-lingual semantic parsing. Chapter 5 suggests that few-shot methods are superior to other approaches, in addition to the utility of meta-learning for cross-lingual adaptation. Sampling the target language distribution directly mitigates the challenges of *fluency* from Chapter 3 and *domain relevance* from Chapter 4. Contingent on economical annotation costs, we assume a few-shot sampling from the true data distribution is the best strategy for our engineer's case study. In this chapter, we now propose an improved few-shot cross-lingual transfer method surpassing DRAKON in both efficiency and accuracy.

A critical component of our prior motivation and analyses is the notion of *cross-lingual representation alignment*. Section 2.4.2 details our intuition where if a multilingual encoder can map *different inputs* to the same *latent representation*, a decoder can deterministically map this *latent representation* to the same *logical form*. In Chapter 4, this informs our objective of encouraging latent representation similarity using auxiliary tasks. In Chapter 5, we observe improved representation similarity by improving cross-lingual gradient similarity via meta-learning. A caveat of prior chapters is that cross-lingual representation alignment is *encouraged* but not strictly *enforced*. Auxiliary tasks and cross-lingual meta-learning encourage cross-lingual gradient alignment *implicitly* within respective methods. Cross-lingual representation alignment is realistically a *side-effect* of our techniques, not the intended objective. Furthermore, we have

not validated the *causality* of this relationship by studying the alignment as a secondary outcome after task improvement.

Prior chapters were *implicit* in targeting this alignment, and in this chapter, we propose to make this goal *explicit*. This chapter considers the hypothesis that **explicit cross-lingual representation alignment improves the transfer of task knowledge to target languages**. We consider directly manipulating representation alignment as the optimisation objective for cross-lingual transfer. We introduce a method for explicit latent representation alignment titled MINOTAUR: **Min**imising **O**ptimal **T**ransport distance for **A**lignment **U**nder **R**epresentations. MINOTAUR introduces a latent variable within an encoder-decoder model defining the latent space as a probability distribution. Interpreting the latent space as probabilistic enables measurement of representation similarity as a closed-form *divergence* between distributions. MINOTAUR optimises cross-lingual representation alignment by penalising this divergence measured between languages. As the latent representations are complex distributions, MINOTAUR introduces an alignment objective with both coarse- and fine-grained divergence penalties to optimise global and local alignment. We define the model and latent alignment objective using Optimal Transport (Monge, 1781; Villani, 2008) in Section 6.1. We define our experiments and comparisons in Section 6.2. Similar to Chapter 5, we follow a few-shot transfer strategy but define an alternative episodic training loop in Section 6.2.2 using parallel data. Our results in Section 6.3 demonstrate that MINOTAUR is more sample-efficient than DRAKON, producing more accurate target language parsing with fewer target language examples. A deeper investigation analyses the contribution of each component of MINOTAUR and identifies where parallel data is necessary, and unnecessary, for cross-lingual transfer. Finally, we examine the latent representations from MINOTAUR targeting explicit alignment, rather than methods encouraging an alignment side-effect.

6.1 Problem Formulation

6.1.1 Revisiting Alignment for Cross-lingual Transfer

Before we introduce the contributions of this chapter, we revisit the framework of representation alignment from Section 2.4.2. We outlined the desirable property of representation alignment in Equations (2.23) to (2.25), where utterances in different languages, x_{EN} and x_l, are encoded to similar latent representations. Provided these

encodings are sufficiently similar, we expect that the encoding of any language will predict the same semantically equivalent logical form.

$$x_{\text{EN}} \neq x_l \qquad \text{Equivalent semantics in languages English and } l \qquad (2.23)$$

$$Q_\phi(x_{\text{EN}}) \approx Q_\phi(x_l) \qquad \text{Approximately similar latent encodings} \qquad (2.24)$$

$$\hat{y}_{\text{EN}} = \hat{y}_l \qquad \text{Parse to predict the same logical form} \qquad (2.25)$$

Chapter 4 proposes to encourage this alignment using auxiliary tasks with multilingual data. We observe improved representation similarity using this method. However, we argue that this strategy to encourage alignment (i.e., Equation (2.24)) *does not directly optimise for this effect*. Analysing Equation (2.24), the notion of *approximately similar encodings* is not precisely defined. Chapter 4 observes improved representation alignment without an adequate similarity criterion. We define the maximum tolerable similarity as η, the maximum permissible error between encodings producing the same predicted y. Given η, the similarity condition can be expressed as Equation (6.1) where we expect η to be some small number (e.g., $\eta = 10^{-8}$).

$$|Q_\phi(x_{\text{EN}}) - Q_\phi(x_l)| < \eta \qquad \text{i.e., } Q_\phi(x_l) \text{ is } < \eta \text{ close to } Q_\phi(x_{\text{EN}}) \qquad (6.1)$$

While definable, this expression of representation alignment is heuristic. A scalar η is inflexible to how different languages have varying similarities and does not provide a useful basis to learn the desirable alignment. Optimising this form of objective (i.e., minimising η) has been approached using contrastive learning for adequate representations (Wu and Dredze, 2020). Methods of this constrative form are discussed in Section 6.4. Recently, Wieting et al. (2023) raised that contrastive learning can be avoided using a variational model as we propose here.

We instead propose to introduce some notion of *structure* to the latent representation space to (i) define representation similarity as a divergence between probability distributions, and (ii) compute representation similarity in closed-form expressions. Our method directly optimises this latent similarity to improve cross-lingual semantic parsing. Alignment in the input (at x) requires fine-grained labels of how tokens map between languages (i.e., word alignment). Alignment in the output (at y) may not prove sufficient to encourage cross-lingual transfer. This is evidenced by our methods for alignment using multi-task learning in Chapter 4, and meta-learning in Chapter 5. In either case, optimising similarity between outputs (i.e., losses or gradients at y) did not

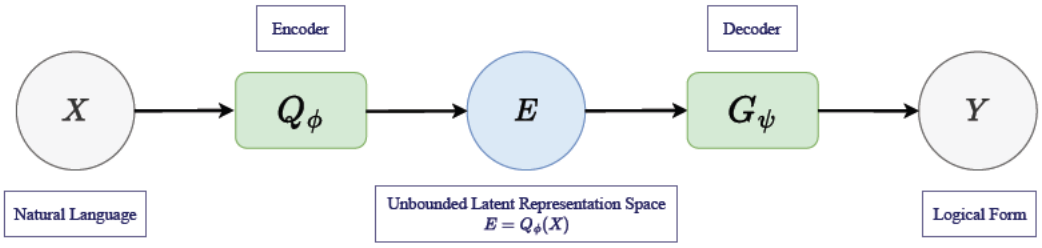

Figure 6.1: A typical Sequence-to-Sequence Encoder-Decoder model without the variational reparameterisation. The encoder, Q_ϕ, converts utterance inputs, X, into some continuous latent encoded representation, E. The decoder, G_ψ, predicts a logical form output, Y, conditioned upon this latent representation. The latent representation in this model is embedded in some arbitrary high-dimensional space without any bound or structure.

wholly mitigate the cross-lingual transfer gap. Given the additional expense of aligning at x, and our results of aligning at y—we now consider an approach explicitly aligning in the continuous latent space between the encoder and decoder.

Section 6.1.2 defines an augmented encoder-decoder model using latent variables to provide the aforementioned structure to the encoding space. Section 6.1.3 defines the Optimal Transport problem framework we exploit for cross-lingual transfer within the latent variable encoder-decoder. Finally, Section 6.1.4 defines how we can explicitly learn cross-lingual similarity using latent structure without an arbitrary tolerance factor.

6.1.2 Augmenting the Encoder-Decoder with Latent Variables

The Transformer semantic parser defined in Section 2.2.1 is outlined in Figure 6.1. This framework underpins the model of previous chapters. However, we now augment this model to impose explicit structure within the latent space.

Our augmentations follow the variational auto-encoder framework (Kingma and Welling, 2014; Rezende et al., 2014) introducing latent variable, Z, between encoder and decoder. Within this framework, encoder Q_ϕ represents inputs from domain $\mathcal{X}$ as a continuous latent variable Z, $Q_\phi : \mathcal{X} \rightarrow \mathcal{Z}$. Decoder G_ψ predicts outputs Y conditioned on samples from the latent space, $G_\psi : \mathcal{Z} \rightarrow \mathcal{Y}$. Encoder Q_ϕ now includes the additional subnetwork Q_σ, defined below, which we assume implicit within Q_ϕ unless stated otherwise ($\sigma \in \phi$).

The critical change in the model is the reparameterisation of the encoder output. Where the typical model in Figure 6.1 outputs contextual dense embeddings in some

arbitrary high-dimensional space, the augmented encoder predicts *a probability distribution* to sample the latent variable. For input sequence $x = \{x_1, \ldots, x_T\}$ and latent dimensionality d, the encoder produces a conditional distribution of latent encodings, $\mathbf{z} = \{z_1, \ldots, z_T\}$, parameterised as a sequence of T mean states $\mu \in \mathbb{R}^{T \times d}$, and variance matrix $\Sigma \in \mathbb{R}^{d \times d}$ shared for all states. Σ is a diagonal matrix embedding variance $\sigma^2 \in \mathbb{R}^d$ output from the encoder. Σ does not contain any co-variance terms. The encoder, Q_ϕ, outputs μ, Σ. Z is then sampled from the normal distribution parameterised by μ and Σ as Equation (6.2). The encoder approximates the posterior distribution $Q_\phi(Z|X)$ for latent variable Z conditioned on observed input X. We use the Gaussian reparameterisation trick (Kingma and Welling, 2014) to sample Z using random variable ε as Equation (6.3) where $\odot$ is the elementwise Hadamard product. This technique allows differentiation through the random node $\mathbf{z}$ by sampling a random ε where the gradient is inconsequential.

$$\mathbf{z} = Q_\phi(x) \sim \mathcal{N}(\mu, \Sigma) \tag{6.2}$$

$$\mathbf{z} = \mu + \Sigma \odot \varepsilon, \ \varepsilon \sim \mathcal{N}(0, I) \tag{6.3}$$

Finally, an output sequence $\hat{y}$ is predicted by the decoder through autoregressive generation conditioned on $\mathbf{z}$. The decoder is unchanged compared to the model in Section 2.2.1. Figure 6.2 describes our augmented encoder-decoder. Within the original alignment objective in Equation (6.1), the exemplar latent encoding from English, $Q_\phi(x_{\text{EN}})$, is now sampled from distribution $\mathcal{N}(\mu_{\text{EN}}, \Sigma_{\text{EN}})$, with associated density and divergence in probability space. Rather than ensuring we are η similar to this encoding, the alignment can be expressed as a similarity between probability densities. We formally define this expression in Section 6.1.4.

Our model diverges from the standard variational auto-encoder framework by lacking any pooling. A typical variational model compresses information into a single vector bottleneck, $z \in \mathbb{R}^{1 \times d}$. This produces informative representations useful for unconditional generation (i.e., generating an output by sampling the latent space with no input). However, our setup requires only conditional generation from an input (i.e., $\mathcal{X} \rightarrow \mathcal{Y}$). Practically, we observed in early experiments that pooling for a single vector z over T time steps weakens overall performance due to information loss from aggregating representations. We therefore omit pooling in our model to predict latent sample $\mathbf{z} \in \mathbb{R}^{T \times d}$ with the same dimensionality as the original encoder output.

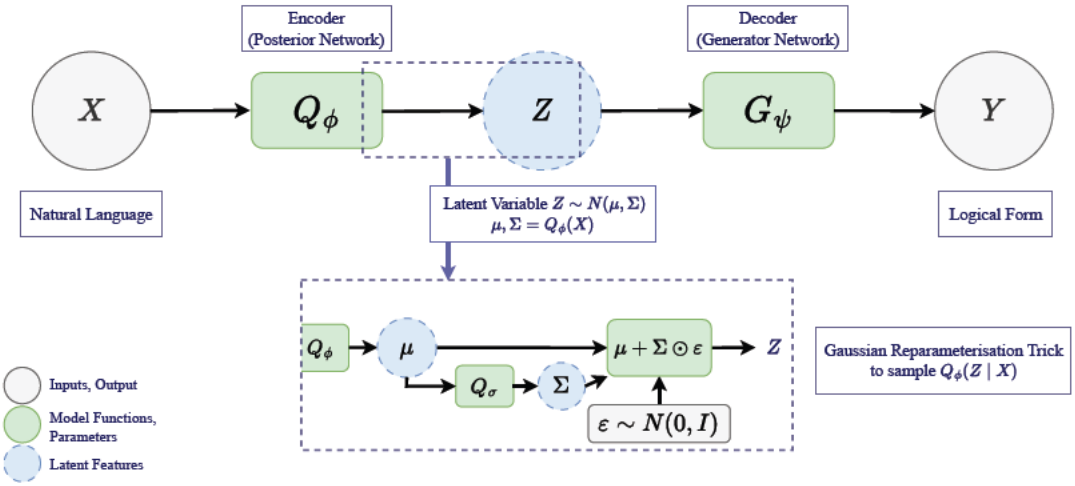

Figure 6.2: A Sequence-to-Sequence Encoder-Decoder model with a variational reparameterisation for the explicit latent variable Z. The encoder, $Q_{\phi,\,\sigma}$, approximates the posterior distribution over Z and the decoder, G_ψ, is the generator network predicting logical form Y conditioned on Z. Z is a sample from the Gaussian distribution predicted by the posterior network (Kingma and Welling, 2014). MINOTAUR uses Z to align representations from different languages by aligning the posterior distributions.

Predicting Variance for Gaussian Distributions The only new parameters for the model outlined in Figure 6.2 is the subnetwork, Q_σ, for predicting the variance term, Σ, for sampling from latent variable Z. We follow the multi-head pooling layer from Liu and Lapata (2019) to combine $\mu \in \mathbb{R}^{T\times d}$ states into a single variance vector, $\sigma^2 \in \mathbb{R}^{1\times d}$, embedded in variance matrix $\Sigma \in \mathbb{R}^{d\times d}$. This network adapts multi-head attention to average representations over time with each "head" similarly learning feature importance across time. For each time t in T, Q_σ projects each encoder output μ_t using learned parameter $W_K \in \mathbb{R}^{d\times d}$ as Equation (6.4). K is equivalent to the 'key' in multi-head attention. We produce a distribution over T time steps, $\alpha_{1,\ldots,T}$, using Equation (6.5). Finally, the output σ^2 is a weighted sum of inputs as Equation (6.6) using additional learned parameter $W_V \in \mathbb{R}^{d\times d}$, equivalent to the 'value' in multi-head attention.

$$a_t = {W_K}^T \mu_t \tag{6.4}$$

$$\alpha_t = \frac{\exp a_t}{\sum_{t' \in T} \exp a_{t'}} \tag{6.5}$$

$$\sigma^2 = \sum_{t' \in T} \alpha_{t'} W_V \mu_{t'} \tag{6.6}$$

We set the dimension of parameters in Q_σ as equivalent to the latent variable

dimensionality d (practically $d = 1024$) such that the number of heads is arbitrary given the product of the number of heads and head dimension is d (# heads $\times$ $d_{\text{head}} = d$). Therefore, we omit the head indexing notation in Equations (6.4) to (6.6). Finally, the variance matrix for Z is σ^2 embedded in a diagonal matrix as Equation (6.7).

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix} \tag{6.7}$$

The additional network Q_σ modifies the encoder to predict both parameters required to sample from a Gaussian distribution for Z. We predict a mean, μ, for the distribution at each time step in T, but predict one Σ for the sequence of encodings. Practically, we observed no benefit to predicting a unique variance for each time step. As further discussed in Section 6.1.4, using a single Σ for each encoded sequence also simplifies computation for representation alignment.

6.1.3 Kantorovich Transportation Problem

The alignment objective, first discussed in Section 2.4.2, can now consider alignment within the latent variable domain, replacing the unbounded representation space between encoder and decoder. Whereas the non-parametric similarity in Equation (6.1) is difficult to optimise, the difference between parametric states is a stable, well-defined measurement. Our intuition is that introducing latent structure allows explicit representation alignment between languages. We can directly optimise a model to imitate the latent representation of English for each target language without auxiliary losses (Chapter 4) or gradient alignment (Chapter 5). Given the semantic parser defined above, we now define a framework for this objective using Optimal Transport.

Given the parser definition, we now define how we use the latent variable Z, for cross-lingual transfer via Optimal Transport. Our parser derives from the *Wasserstein Auto-Encoder* (WAE) from Tolstikhin et al. (2018) as an alternative form of the variational model. The objective function of a WAE is to minimise the *transportation cost* moving probability from one distribution to another. This can be described using the Kantorovich form (Kantorovich, 1958) of the Optimal Transport problem (Monge, 1781) in Equation (6.8). Given two distributions P_X, P_Y, the objective is to find a *transportation plan*, $\Gamma(X, Y)$, within the set of all joint distributions, $P(X \sim P_X, Y \sim P_Y)$, to map probability mass from P_X to P_Y with minimal cost. Within Equation (6.8), T_c ex-

presses the problem of finding a plan which minimises transportation cost measured by cost function $c(X, Y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$:

$$T_c(P_X, P_Y) := \inf_{\Gamma \in (X \sim P_X,\ Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)] \tag{6.8}$$

The WAE is proposed as an auto-encoder where the output reconstructs the input (i.e., P_Y approximates P_X), however, in our setting P_X is the natural language input distribution and P_Y is the logical form output distribution. X and Y are realisations of the same underlying semantics.

As output Y is conditioned on only Z, inputs and outputs are conditionally independent given the latent variable, $y \perp\!\!\!\perp x \mid \mathbf{z}$. This allows a factorised reformulation of the transportation plan using Bayes rule i.e., $\Gamma(X, Y) \rightarrow \Gamma(Y|X)P_X$. We now consider a non-deterministic mapping from X to Y under observed P_X. Tolstikhin et al. (2018, Theorem 1) identifies how to *factor* this mapping through latent variable Z, leads to Equation (6.9). Equation (6.9) expresses the transportation cost, c, from X to Y through sampled approximate posterior $Q(Z|X)$, from domain of possible posteriors, $\mathcal{Q}$, and observed distribution P_X. Tolstikhin et al. (2018, Theorem 1) identify that solving the original form of Equation (6.8) requires marginal posterior, $Q(Z)$, to match the prior distribution $P(Z)$ exactly. We follow prior work in setting $P(Z) \sim \mathcal{N}(0, I)$. This constraint is relaxed by introducing a non-negative regularisation penalty on $Q(Z)$ in Equation (6.9) as $\mathbb{D}(Q(Z), P(Z))$. We discuss similarities of this regularisation to the variational ELBo below.

$$T_c(P_X, P_Y) = \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q_\phi(Z|X)}[c(Y, G(Z))] + \alpha\mathbb{D}(Q(Z), P(Z)) \tag{6.9}$$

Equation (6.9) is now a minimisable objective: identify the probabilistic encoder, $Q_\psi(Z|X)$, and decoder $G_\psi(Z)$ which minimises cost function c; subject to regularisation on the divergence $\mathbb{D}$ between the marginal posterior $Q(Z)$ and prior $P(Z)$. In our setting, the cost function c is the cross-entropy loss between gold outputs Y and predicted outputs.

Optimising a variational model typically requires optimising the evidence-lower bound (ELBo) which minimises the divergence between observed conditional posterior, $Q(Z|X)$, and prior distribution $P(Z)$. While formally robust and appropriate, this objective can practically drive $Q(Z|X)$ to zero assuming the prior is an isotropic Gaussian distribution. This undesirable problem is referred to as *posterior collapse* where Z is no

longer semantically informative. The WAE instead regularises the marginal distribution, $Q(Z) = \mathbb{E}_{X \sim P_X}[Q(Z|X)]$, against the prior distribution. Marginal regularisation allows individual conditional posterior samples to remain non-zero i.e., informative. This avoids posterior collapse to ensure that our latent features in Z are accurately describing X.

For the divergence between marginal posterior and prior distribution, we follow Tolstikhin et al. (2018) and Wang and Wang (2019) in using Maximum Mean Discrepancy (Gretton et al., 2012, MMD). MMD provides an unbiased estimate of $\mathbb{D}(Q(Z), P(Z))$ using the batch of approximate posterior samples $Q(Z|X)$ and randomly sampled $P(Z)$. Equation (6.10) defines MMD using some kernel $k: \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{R}$, defined over a reproducible kernel Hilbert space, $\mathcal{H}_k$:

$$\text{MMD}_k(P, Q) = \| \int_{\mathcal{Z}} k(z, \cdot)\, dP - \int_{\mathcal{Z}} k(z, \cdot)\, dQ \|_{\mathcal{H}_k} \tag{6.10}$$

Informally, MMD minimise the distance between the "means" of the features P and Q estimated over n samples projected through kernel k. We set n to the batch size in our experiments. Equation (6.11) defines the MMD estimate over observed $\mathbf{p}$ and $\mathbf{q}$ using the heavy-tailed *inverse multiquadratic* (IMQ) kernel k:

$$MMD_k(\mathbf{p}, \mathbf{q}) = \frac{1}{n_p(n_p - 1)} \sum_{z' \neq z} k(p_z, p_{z'}) \frac{1}{n_q(n_q - 1)} \sum_{z' \neq z} k(q_z, q_{z'}) - \frac{2}{n_p n_q} \sum_{z,\, z'} k(p_z, q_{z'}) \tag{6.11}$$

We define the IMQ kernel in Equation (6.12) below where $C = \frac{d}{5}$ for dimensionality d and S is defined in Equation (6.13).

$$k(p, q) = \sum_{s \in S} \frac{s \cdot C}{s \cdot C + \|p - q\|_2^2} \tag{6.12}$$

$$S = [0.1, 0.2, 0.5, 1, 2, 5, 10] \tag{6.13}$$

The objective in Equation (6.9) defines the Optimal Transport problem for transforming probability mass from the input distribution (i.e., natural language) to the output distribution (i.e., logical forms). We use this framework with the model, defined in Section 6.1.2, and the WAE optimisation defined above to learn the optimal transportation plan in predicting accurate logical forms from natural language inputs. We now describe how to exploit this to transfer an optimal transportation plan to target languages.

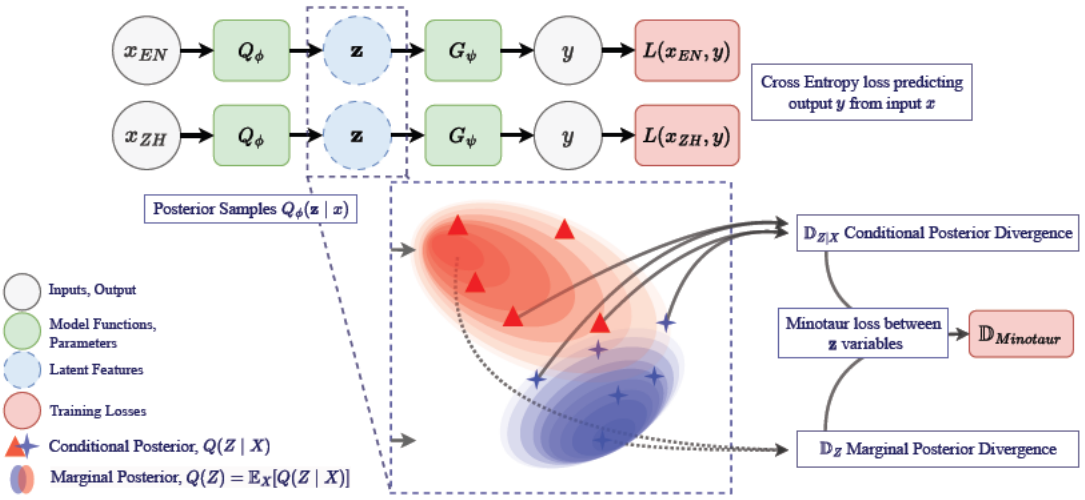

Figure 6.3: MINOTAUR proposes to penalise the divergence between the sampled latent variable states. Two forward passes on parallel data from English and a sampled target language (e.g., ZH) are run to compute the transportation cost from X to Y (i.e., $\ell(x,y)$). MINOTAUR computes an additional loss between $\mathbf{z}$ states to minimise latent divergence between conditional and marginal posterior distributions.

6.1.4 MINOTAUR: Explicit Alignment for Cross-lingual Transfer

Consider learning the optimal mapping from English utterances, x_{EN}, to logical forms, y, through latent variable Z in Equation (6.8). The optimisation in Equation (6.9) converges on an optimal transportation plan Γ^*_{EN} with associated minimum cost.[1]

Now consider an input x_l from language l which is semantically equivalent to x_{EN}. The representation alignment objective from Section 2.4.2 is now a minimal and intuitive condition that parallel x encodes to equivalent $\mathbf{z}$; equivalent $\mathbf{z}$ subsequently yields equivalent y. We propose a straightforward extension of learning Γ^*_{EN} to learn to map different x to the same latent variable state $\mathbf{z}$. Our proposal is to bootstrap the transportation plan for target language l (i.e., $\Gamma^*_l(X_l, Y)$) by only aligning Z in a few-shot learning scenario. We hypothesise that explicit alignment, through directly inducing representation alignment, produces generalisable cross-lingual transfer using a small sample of each target language.

We introduce MINOTAUR: **Min**imising **O**ptimal **T**ransport distance for **A**lignment **U**nder **R**epresentations. MINOTAUR proposes to explicitly align representations by matching latent variables between languages. A typical WAE builds an informative latent structure in Z using only the divergence penalty between the marginal posterior

[1] Γ_* is implicit within the model parameters.

and the sampled prior distribution. MINOTAUR proposes to additionally penalise the divergence *between approximate posteriors from different languages*. We describe our explicit goal as *posterior alignment*, differing from Chapter 4 as alignment is now the direct optimisation objective function. Figure 6.3 outlines one training step of MINOTAUR using parallel input batches. A single step computes the typical cross-entropy loss, with additional loss explicitly optimising for similarity between sampled latent variables. This optimises the transportation plan for language l to match the plan for English, $\Gamma^*_l = \Gamma^*_{\text{EN}}$, transferring the learned capabilities from high-resource English with only a few training examples.

Given parallel inputs x_{EN} and x_l in English and language l, with equivalent LF ($y_{\text{EN}} = y_l$), their latent encodings are given by Equation (6.14) and Equation (6.15) respectively.

$$\mu_{\text{EN}}, \Sigma_{\text{EN}} = Q_\phi(x_{\text{EN}}), \mathbf{z}_{\text{EN}} \sim \mathcal{N}(\mu_{\text{EN}}, \Sigma_{\text{EN}}) \tag{6.14}$$

$$\mu_l, \Sigma_l = Q_\phi(x_l), \mathbf{z}_l \sim \mathcal{N}(\mu_l, \Sigma_l) \tag{6.15}$$

As mentioned in Section 6.1.2, the posterior samples ($\mathbf{z}_{\text{EN}}, \mathbf{z}_l \in \mathbb{R}^{T \times d}$) are complex structures. Our interpretation of *complexity* is that each $\mathbf{z}$ sample will encode relevant information about the inputs adhering to some local and global structure. We make no further assumptions about the information structure beyond the Gaussian reparameterisation. This complexity contrasts to a *simple* structure such as the isotropic Gaussian prior distribution. To accurately estimate divergence between complex posteriors, we follow Mathieu et al. (2019) in using a decomposed alignment signal minimising both marginal posterior divergence and conditional posterior divergence. We interpret this as coarse-grained and fine-grained divergence respectively. Our intuition is that fine-grained divergence encourages similarity at a *token* level, optimising for similarity between conditional posterior samples. In contrast, we intuit that coarse-grained divergence optimises for similarity at a *language* level. Estimating the marginal $Q(Z)$, from batches of tokens in each language, will estimate the language-level distribution over Z. Therefore, we expect that penalising this marginal divergence optimises for more global similarity. For approximate posterior samples from Equations (6.14) to (6.15), we can express the MINOTAUR alignment objective as Equation (6.16).

$$\mathbb{D}_{\text{MINOTAUR}}(\mathbf{z}_{\text{EN}}, \mathbf{z}_l) = \alpha_P \mathbb{D}_Z\left(Q_\phi(\mathbf{z}_{\text{EN}}), Q_\phi(\mathbf{z}_l)\right) + \beta_P \mathbb{D}_{Z|X}\left(Q_\phi(\mathbf{z}_{\text{EN}}|x_{\text{EN}}) \| Q_\phi(\mathbf{z}_l|x_l)\right) \tag{6.16}$$

Within Equation (6.16), $\mathbb{D}_Z$ is a divergence penalty between *marginal* distributions to align global structure, and $\mathbb{D}_{Z|X}$ is the divergence penalty between *conditional* posteriors to align local structure, and. Each divergence uses some scalar weighting, similar to Equation (6.9), as (α_P, β_P) respectively.

Similar to the prior alignment, we use the MMD distance to align marginal posteriors using Equation (6.10) (i.e., marginal posteriors over Z between languages). For the two sampled batches in Figure 6.3, MMD estimates a single marginal distribution, $Q(Z)$, for each language using all sampled $\mathbf{z}$ states from each utterance in a batch. MMD is estimated between two $Q(Z)$ to return a single distance as a differentiable objective. We require a sufficiently large batch size to ensure accurate estimation of the marginal distribution: we observed our batch size of 16 as sufficient in our experiments.

For conditional posterior alignment, we consider closed-form solutions to estimate divergence. This exact expression is a core benefit of the Gaussian reparameterisation expressing the latent Z as a parametric statistic. This approach is easy to compute, numerically stable and an accurate estimator of divergence between high-dimensional Gaussians (Takatsu, 2011). We primarily use the L_2 Wasserstein distance, W_2, as the Optimal Transport-derived minimum transportation cost between Gaussian distributions. Equation (6.17) expresses the L_2 Wasserstein distance between parametric distributions $\mathbf{p}$ and $\mathbf{q}$ where each mean is μ, covariance is $\Sigma = \text{Diag}\{\sigma_i^2, \ldots, \sigma_n^2\}$, encodings have dimensionality d and $\text{Tr}\{\}$ is the matrix trace function. Practically, this distance is similar to Euclidean (L_2 norm) distance between distribution means with an additional penalty to match covariances.

$$W_2(\mathbf{p}, \mathbf{q}) = \|\mu_\mathbf{p} - \mu_\mathbf{q}\|_2^2 + \text{Tr}\{\Sigma_\mathbf{p} + \Sigma_\mathbf{q} - 2\left(\Sigma_\mathbf{p}^{\frac{1}{2}}\Sigma_\mathbf{q}\Sigma_\mathbf{p}^{\frac{1}{2}}\right)^{\frac{1}{2}}\} \tag{6.17}$$

We also consider the Kullback-Leibler Divergence (KL) between two Gaussian distributions as Equation (6.18). Minimising KL is equivalent to maximising the mutual information between distributions as an information-theoretic goal of semantically aligning the encoded information in posteriors. While not a well-defined distance, KL is a common divergence used in variational modelling serving as a practical baseline. Section 6.3 demonstrates that W_2 is superior to KL in all cases.

$$\text{KL}(\mathbf{p}\|\mathbf{q}) = \frac{1}{2}\left(\log\left(\frac{|\Sigma_\mathbf{q}|}{|\Sigma_\mathbf{p}|}\right) - d_{p,q} + \text{Tr}\{\Sigma_\mathbf{q}^{-1}\Sigma_\mathbf{p}\} + (\mu_\mathbf{q} - \mu_\mathbf{p})^T \Sigma_\mathbf{q}(\mu_\mathbf{q} - \mu_\mathbf{p})\right) \tag{6.18}$$

In Equations (6.16) to (6.18), we express $\mathbb{D}_{Z|X}$ between singular $\mathbf{p}$ and $\mathbf{q}$ distributions (i.e., between individual tokens) for notation clarity. However, the approximate posterior $\mathbf{z}$ contains T samples for each input utterance of length T. To align conditional posteriors, we minimise the *mean of all pairwise conditional posterior divergence*. For $\mathbf{z}_{\text{EN}}$ and $\mathbf{z}_l$, the final conditional posterior divergence is expressed as Equation (6.19). Minimising this mean divergence makes no assumptions about which tokens from each input *should* be made more or less similar. We also observed no empirical benefit in minimising only divergences between encodings from word-aligned parallel tokens.

$$\mathbb{D}_{Z|X}\left(Q_\phi\left(\mathbf{z}_{\text{EN}}|x_{\text{EN}}\right)\|Q_\phi\left(\mathbf{z}_l|x_l\right)\right) = \frac{1}{T_{\text{EN}}T_l}\sum_{i\in T_{\text{EN}},j\in T_l}\mathbb{D}_{Z|X}\left(\mathbf{z}_{\text{EN}i}\|\mathbf{z}_{lj}\right) \tag{6.19}$$

Given the full definition of the MINOTAUR alignment loss, we can now express the Optimal Transport problem from Equation (6.9) in a familiar form as Equation (6.20). Equation (6.20) expresses the transportation cost, T_c, for a single $(x,\ y)$ pair during training: the cross-entropy between predicted and gold y sampling an approximate posterior $Q_\phi\left(\mathbf{z}|x\right)$, and WAE divergence penalty on the marginal posterior distribution.

$$\ell(x,y) = \mathbb{E}_{Q(\mathbf{z}|x)}\left[-\sum_i y_i\left(\log G_\psi\left(\mathbf{z}\right)\right)_i\right] + \alpha\mathbb{D}\left(Q_\phi(\mathbf{z}), P(\mathbf{z})\right) \tag{6.20}$$

To introduce posterior alignment, we augment Equation (6.20) with the MINOTAUR loss from Equation (6.16), for Equation (6.21) using parallel examples $(x_{\text{EN}}, y_{\text{EN}})$ and (x_l, y_l). Unlike DRAKON, MINOTAUR requires parallel data for conditional posterior alignment. We investigate this requirement further in Section 6.3.

$$\ell_\Sigma = \ell(x_{\text{EN}}, y_{\text{EN}}) + \ell(x_l, y_l) + \mathbb{D}_{\text{MINOTAUR}}\left(\mathbf{z}_{\text{EN}}, \mathbf{z}_l\right) \tag{6.21}$$

Similar to Chapter 5, we construct an episodic training loop for MINOTAUR where we periodically augment the standard cross-entropy optimisation with MINOTAUR posterior alignment. This algorithm is described as Algorithm 2, using MINOTAUR loss every K steps for few-shot induction of cross-lingual alignment. If K is small then MINOTAUR loss dominates learning to degrade overall performance. However, if K is too large then MINOTAUR is insufficient to optimise posterior alignment. Optimally setting K requires empirical validation. We find that using $K = 10$ to match Chapter 5 is sufficient in our experiments.

Algorithm 2 MINOTAUR Training Algorithm

Require: Number of training episodes, T, and duration of training episode, K.

Require: Support data sample, $\mathcal{S}_{\text{EN}}$

Require: Target data samples, $\mathcal{S}_l$, for each language l in target languages $L = \{l_1, \ldots, l_L\}$.

Require: Learning rate, α

Require: Optimiser, $U(\theta_i, \alpha, \nabla)$, updating θ_i according to step size α and gradient ∇.

Initialise $\theta_{t=1}$, the vector of initial parameters
for $t \leftarrow 1$ **to** T **do**
  Sample K training batches $\{\mathcal{B}^S\}_{k=1}^{K}$ from $\mathcal{S}_{\text{EN}}$.
  Sample target language l from L languages.
  Sample target batch $\mathcal{B}^T$ from $\mathcal{S}_l$.
  Retrieve batch of examples $\mathcal{B}'^S$ from $\mathcal{S}_{\text{EN}}$ parallel to $\mathcal{B}^T$.
  for $i \leftarrow 1$ **to** K **do**
    Forward pass for $\mathcal{B}^S$: $\ell, \{\mathbf{z}_{\text{EN}}\}_{i=1}^{||\mathcal{B}'^S||}$ from Equation (6.20).
    $\theta_{i+1} \leftarrow U(\theta_i,\ \alpha,\ \nabla\ell)$
  end for
  Forward pass for $\mathcal{B}'^S$: $\ell^S, \{\mathbf{z}_{\text{EN}}\}_{i=1}^{||\mathcal{B}'^S||}$ from Equation (6.20).
  Forward pass for $\mathcal{B}^T$: $\ell^T, \{\mathbf{z}_l\}_{i=1}^{||\mathcal{B}^T||}$ from Equation (6.20).
  MINOTAUR loss between approximate posteriors $\mathbf{z}_{\text{EN}}$ and $\mathbf{z}_l$ from Equation (6.16).
  Total loss: $\ell_\Sigma = \ell^S + \ell^T + \mathbb{D}_{\text{MINOTAUR}}\left(\{\mathbf{z}_{\text{EN}}\}_{i=1}^{||\mathcal{B}'^S||}, \{\mathbf{z}_l\}_{i=1}^{||\mathcal{B}^T||}\right)$ from Equation (6.21).
  Update $\theta_{t+1} \leftarrow U(\theta_t,\ \alpha,\ \nabla\ell_\Sigma)$
end for

6.2 Experiments

6.2.1 Datasets and Few-shot Sampling

MultiATIS++SQL We use the MultiATIS++SQL dataset similar to Chapter 5. Section 2.3.1 details a complete description of MultiATIS++SQL with input-output examples shown in Table 2.5. We use the MultiATIS++SQL multilingual test set for evaluating cross-lingual transfer from English (EN) to French (FR), Portuguese (PT), Spanish (ES), German (DE), and Chinese (ZH). MINOTAUR is evaluated as a few-shot cross-lingual transfer problem randomly sampling small samples of data from target languages: French, Portuguese, Spanish, German, and Chinese. We examine few-shot sample ratios at 1%, 5%, and 10% of the existing English data. With a total training sample of 4473 examples, these sampling rates correspond to 45, 224, and 447 examples in each few-shot percentage respectively. We report the average of five runs to minimise

variation from random sampling.

MTOP We also evaluate the MTOP dataset following the sampling method from Chapter 5. Section 2.3.2 details a complete description of MTOP with input-output examples shown in Table 2.6. We use the MTOP multilingual test set for evaluating cross-lingual from English (EN) to French (FR), Spanish (ES), German (DE), Hindi (HI), and Thai (TH). We evaluate a few-shot cross-lingual transfer problem sampling data to five target languages: French, Spanish, German, Hindi, and Thai. Sampling follows the same *Samples-per-Intent-and-Slot* (SPIS) strategy which ensures a minimum coverage for each semantic category in logical forms across target languages. An SPIS rate of 1, 5, 10 approximately equates to 284 (1.8%), 1,125 (7.2%), and 1,867 (11.9%) examples for MTOP in each target language. Similar to MultiATIS++SQL, we report the average of five runs.

6.2.2 Experimental Setting

We train the model defined in Section 6.1.2 using the training algorithm outlined in Section 6.1.4. Our primary comparisons for few-shot cross-lingual transfer are the baselines and DRAKON method from Chapter 5. For MTOP, we compare two methods for "silver-standard" data generation: "Translate-and-Fill" (Nicosia et al., 2021, TaF) which generates training data using MT, and CLASP (Rosenbaum et al., 2022) which uses MT and prompting to generate multilingual training data. We previously compared these techniques in Chapter 3, and now revisit these approaches as our techniques are more competitive.

6.2.2.1 Setting and Comparison

MULTILINGUAL **Gold** A multilingual Transformer is trained on the union of all professionally translated data. This is the same upper bound as in Chapter 5. We compare to an unmodified Transformer model (SEQ2SEQ) and the latent-variable model described in Section 6.1.2 (WAE) trained with all multilingual data. This method is an **upper bound** to MINOTAUR.

EN ONLY A monolingual Transformer is trained on only English training data. This model is evaluated on the target language test data with no translation. This is a **zero-shot** baseline for Chapter 4.

TRANSLATE TEST A monolingual Transformer is trained on source English data ($\mathcal{S}_{\text{EN}}$). Machine translation translates test data from target languages into English to predict logical forms from translated inputs. We use the OPUS translation system (Tiedemann and Thottingal, 2020) similar to prior chapters. This is the lower bound for Chapter 3. This method is an **silver standard lower bound** to MINOTAUR.

TRANSLATE TRAIN Machine translation translates English training data into each target language as described in Chapter 3. A monolingual Transformer is trained on translated training data and logical forms are predicted using this model. This is the baseline method for Chapter 3. As above, we use OPUS translation (Tiedemann and Thottingal, 2020) for this method. This method is an **silver standard lower bound** to MINOTAUR.

FATES Our proposal for machine translation from Chapter 3 using multiple encoders and multiple MT engines for cross-lingual semantic parsing without gold training data. This method is a **syntehtic data** comparison from Chapter 4.

ZEUS Our zero-shot multi-task model from Chapter 4 using auxiliary data for cross-lingual latent representation alignment. We note that already performs close to the upper bound but few-shot sampling may be more data efficient than leveraging large corpora for auxiliary tasks. This method is a **zero-shot** comparison from Chapter 4.

DRAKON Our few-shot meta-learning method from Chapter 5 which established more accurate parsing at all sampling levels than all methods mentioned above. We compare to DRAKON for both datasets at all sampling rates for a direct comparison between methods. DRAKON and MINOTAUR are trained with identical data samples to contrast if MINOTAUR can use the same data more efficiently for target language parsing. This method is a **few-shot** comparison from Chapter 5.

6.2.2.2 Model Training

We generally inherit the same model setup as previous chapters to train the latent variable model defined in Section 6.1.2. The encoder, Q_ϕ, is similarly pre-trained using the encoder parameters from the MBART50 pre-trained model (Tang et al., 2021). This model has a single Transformer decoder, G_ψ, trained from scratch. The new

subnetwork for predicting variance, Q_σ, is an additional multi-head attention layer with dimensionality $d = 1024$ trained from scratch.

Similar to Chapter 5, experimental hyperparameters were tested at the few-shot rate of 1% for MultiATIS++SQL and then applied to all experiments. We follow the batch size of 16 and an episode length of $K = 10$ steps. The scalar hyperparameters for MINOTAUR, (β_P, α_P), are set to $(0.5, 0.01)$ respectively. These were optimised via linear search within the range $[0.001, 1]$. We train for a maximum of 10 epochs with early stopping as measured by validation loss.

6.3 Results

We train the MINOTAUR parser from Section 6.1 for comparison to the systems outlined in Section 6.2.2. Our results contrast MINOTAUR to lower-bounds, upper-bounds and related work in a similar structure to Chapter 5. We generally observe MINOTAUR validates the hypothesis that explicit representation alignment improves cross-lingual transfer. MINOTAUR is also more sample efficient than prior contributions and requires less training. Our ablations identify that MINOTAUR is also capable without parallel data and produces a more semantically distributed latent representation space.

6.3.1 Is a Few-shot Transfer Methodology Competitive?

Similar to Chapter 5, we compare MINOTAUR to lower-bound baselines using OPUS machine translation and zero-shot transfer from English to target languages. The upper-bound is 'MULTILINGUAL Gold' training on all target language data i.e., 100% translation. We compare to 'MULTILINGUAL Gold' training either the unmodified Transformer parser (SEQ2SEQ) or the latent variable parser (WAE).

Lower Bound Baselines We compare MINOTAUR to lower-bound methods in Table 6.1. MINOTAUR @1% is significantly above both translation and zero-shot baselines. The largest contribution to this gain is improvement in languages distant from English which benefits more from additional target language data. Compared to zero-shot transfer, MINOTAUR @1% improves by +31.5% for MultiATIS++SQL ZH, and +42% or +36.3% for MTOP HI and TH respectively. Similar languages benefit less from few-shot transfer as zero-shot performance is greater. The gain for FR is a lesser +12.4% for MultiATIS++SQL or +29.9% for MTOP. MINOTAUR also performs above

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
TRANSLATE TRAIN OPUS	—	56.8	39.1	51.8	60.4	59.6	53.5
TRANSLATE TEST OPUS	—	57.7	58.1	58.3	58.8	50.9	56.8
EN Only	77.2	61.3	42.5	46.5	50.2	38.5	47.8
MINOTAUR @1%	73.0±0.4	73.7±0.6	71.4±0.9	71.0±0.5	70.4±1.3	70.0±0.9	71.3±1.4
MINOTAUR @5%	77.0±1.0	73.9±1.4	72.8±1.1	71.1±0.6	72.8±2.0	72.3±0.6	72.6±1.0
MINOTAUR @10%	79.8±0.4	75.6±1.8	75.4±0.8	73.2±1.7	76.8±1.5	72.5±0.7	**74.7**±1.8

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
TRANSLATE TRAIN OPUS	—	24.4	23.1	32.7	22.4	9.5	22.4
TRANSLATE TEST OPUS	—	44.9	63.1	39.1	47.1	54.2	49.7
EN Only	72.4	42.0	43.9	46.8	23.1	12.8	33.7
MINOTAUR @1 SPIS	79.5±0.4	71.9±0.2	72.3±0.1	68.4±0.3	65.1±0.1	49.1±4.3	65.4±9.5
MINOTAUR @5 SPIS	77.7±0.6	72.0±0.6	73.6±0.3	69.1±0.5	68.2±0.5	52.1±3.4	67.0±8.6
MINOTAUR @10 SPIS	80.2±0.4	72.8±0.5	74.9±0.1	70.0±0.7	68.6±0.5	54.7±2.5	**68.2**±7.9

(b) MTOP

Table 6.1: Comparisons between MINOTAUR to lower-bounds for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy. We compare MINOTAUR to machine-translation baselines and the zero-shot transfer performance from training only on English. MINOTAUR must surpass these lower-bound baselines to justify the few-shot approach. For few-shot methods, we report the average over five different few-shot data splits $\pm$ the standard deviation across runs. The significant best result is bolded.

TRANSLATE TEST as the most competitive baseline by average +14.5% and +15.7% for MultiATIS++SQL and MTOP respectively. Unlike for DRAKON, the standard deviation for MINOTAUR @1% does not overlap with TRANSLATE TEST, further indicating that this method is superior to baselines. We further analyse the improvement over baselines as error analysis in Section 6.3.4.

Upper Bounds and Prior Methods MINOTAUR is compared to the upper bound and adjacent methods in Table 6.2. Comparing between 'MULTILINGUAL Gold' methods, the WAE approach is significantly −0.3% weaker for MultiATIS++SQL to SEQ2SEQ but significantly surpasses the simpler SEQ2SEQ model for MTOP by +8.3%. We interpret this as sanity check validation that the introduction of the latent variable structure, as the only difference between SEQ2SEQ and WAE is not highly detrimental to overall parsing accuracy.

We observe that posterior alignment within multilingual modelling allows a few-shot model to improve on some upper-bound systems, exceeding expectations for a data-scarce scenario. For MultiATIS++SQL, MINOTAUR performs significantly above both 'MULTILINGUAL Gold' methods with $> 5\%$ sampling. Similarly for MTOP, MINOTAUR significantly improves on 'MULTILINGUAL Gold' (SEQ2SEQ) with > 1 SPIS sampling. The 'MULTILINGUAL Gold' (WAE) method is a stronger upper bound for MTOP which our results are not competitive against. This highlights a future improvement in cross-lingual semantic parsing to reach parity between English and target languages on this dataset.

MINOTAUR is competitive with some silver-standard methods using $> 5\times$ larger models for MTOP. We compare to 'Translate-and-Fill' (Nicosia et al., 2021, TaF) and CLASP (Rosenbaum et al., 2022) in Table 6.2(b). Both methods use data generation from large pre-trained models to produce a training dataset for fine-tuning. We significantly outperform CLASP by an average >3.0% at all sampling rates and TaF using mT5-large (Xue et al., 2021) by $> 1.6\%$ at $> 1\%$ sampling. However, MINOTAUR requires 10% sampling to significantly improve upon TaF using mT5-XL by +0.02%. Our model has only $\sim$ 200 million parameters whereas CLASP uses the 500 million parameter AlexaTM-500M (FitzGerald et al., 2022), mT5-large has 700 million parameters and mT5-XL has 3.3 billion parameters. Relative to model size, our approach offers improved computational efficiency and faster training for comparable results. Our efficacy using gold data and a smaller model, compared to silver-standard data in larger models, suggests a quality trade-off constrained by computation as a potential

	EN	FR	PT	ES	DE	ZH	TARGET AVG.
MULTILINGUAL Gold (SEQ2SEQ)	74.9	74.2	73.0	70.4	74.6	73.7	73.2
MULTILINGUAL Gold (WAE)	73.7	74.4	72.3	71.7	74.6	71.4	72.9
FATES (best)	74.9	70.5	69.2	62.4	68.7	66.5	67.5
ZEUS (best)	74.4	72.3	69.7	68.5	69.0	69.2	69.7
DRAKON @1%	73.8±0.3	70.4±1.8	70.8±0.7	68.9±2.3	69.1±1.2	68.1±1.2	69.5±1.1
DRAKON @5%	74.4±1.3	73.0±0.9	71.6±1.1	71.6±0.7	71.1±0.6	69.5±0.5	71.4±1.3
DRAKON @10%	75.8±1.3	74.2±0.2	72.8±0.6	72.1±0.7	73.0±0.6	72.8±0.5	73.0±0.8
MINOTAUR @1%	73.0±0.4	73.7±0.6	71.4±0.9	71.0±0.5	70.4±1.3	70.0±0.9	71.3±1.4
MINOTAUR @5%	77.0±1.0	73.9±1.4	72.8±1.1	71.1±0.6	72.8±2.0	72.3±0.6	72.6±1.0
MINOTAUR @10%	79.8±0.4	75.6±1.8	75.4±0.8	73.2±1.7	76.8±1.5	72.5±0.7	**74.7**±1.8

(a) MultiATIS++SQL

	EN	FR	ES	DE	HI	TH	TARGET AVG.
MULTILINGUAL Gold (SEQ2SEQ)	75.5	69.7	72.4	67.9	65.5	54.6	66.0
MULTILINGUAL Gold (WAE)	81.3	75.7	77.2	72.8	71.6	74.4	74.3
TaF mT5-large (Nicosia et al., 2021)	83.5	71.1	69.6	70.5	58.1	57.5	65.4
TaF mT5-XL (Nicosia et al., 2021)	85.9	74.0	71.5	72.4	61.9	60.2	68.0
CLASP (Rosenbaum et al., 2022)	84.4	72.6	68.1	66.7	58.1	—	—
FATES (best)	69.7	44.7	45.7	49	32.9	13.4	37.1
ZEUS (best)	77.5	66.2	67.4	64.2	59.4	47.7	61.9
DRAKON @1 SPIS	71.8±2.0	59.0±1.7	61.1±1.1	59.4±1.6	56.1±1.4	43.9±2.3	55.9±6.9
DRAKON @5 SPIS	72.1±1.9	65.4±2.6	67.0±4.6	65.3±2.6	63.0±0.1	49.8±0.8	62.1±7.0
DRAKON @10 SPIS	72.5±0.4	65.6±0.5	67.5±0.8	65.8±0.6	63.8±1.1	50.6±1.1	62.7±6.9
MINOTAUR @1 SPIS	79.5±0.4	71.9±0.2	72.3±0.1	68.4±0.3	65.1±0.1	49.1±4.3	65.4±9.5
MINOTAUR @5 SPIS	77.7±0.6	72.0±0.6	73.6±0.3	69.1±0.5	68.2±0.5	52.1±3.4	67.0±8.6
MINOTAUR @10 SPIS	80.2±0.4	72.8±0.5	74.9±0.1	70.0±0.7	68.6±0.5	54.7±2.5	**68.2**±7.9

(b) MTOP

Table 6.2: Comparisons between MINOTAUR to upper-bound training and the best methods from previous Chapters for (a) MultiATIS++SQL execution accuracy and (b) MTOP SCIEM accuracy. We compare between (a) training on gold-standard translations in all target languages (MULTILINGUAL Gold); (b) comparison methods using silver-standard data for MTOP only; (c) FATES from Chapter 3; (d) ZEUS from Chapter 4; and (e) DRAKON at equivalent sampling rates from Chapter 5. Few-shot methods report the average over five different few-shot data splits ± the standard deviation across runs.

future study.

Compared to our proposals from previous chapters, MINOTAUR is the most accurate cross-lingual semantic parser proposed in this thesis. MINOTAUR is significantly superior to both FATES and ZEUS at all sampling rates on MultiATIS++SQL and MTOP. We do not observe the same trend as the previous chapter where DRAKON performed marginally below ZEUS at low sampling rates. MINOTAUR can better exploit few-shot samples to perform significantly above our optimised zero-shot model with increased data efficiency and without large auxiliary task corpora.

Comparing average performance to DRAKON strategy, MINOTAUR is significantly more accurate at every few-shot sampling rate for MultiATIS++SQL and MTOP. MINOTAUR improves on DRAKON with performance more similar to DRAKON at *at a higher sample rate*. For example, MINOTAUR @1% is significantly improves on DRAKON @1% sampling, and is insignificantly similar to DRAKON @5% sampling ($p = 0.15$) for MultiATIS++SQL. MINOTAUR at the lowest sampling rate is actually significantly superior to the DRAKON at the highest sampling rate for MTOP. By extension, MINOTAUR is also significantly improved above all comparison methods from Chapter 5. These methods ('Train-EN∪All', 'Train-EN→FT-All', and 'Reptile-EN→FT-All') all performed worse than DRAKON, and therefore are also poorer than MINOTAUR. We also note that MINOTAUR requires < 10 epochs to train whereas DRAKON reports > 50 training epochs for weaker average accuracy. These comparisons demonstrate that explicit alignment offers both more accurate parsing and improved efficiency from data and training perspectives.

While we report that MINOTAUR is superior to DRAKON in target language average accuracy, we identify that MINOTAUR benefits comparatively less from additional samples for distant languages. For MultiATIS++SQL, ZH is the only individual language where DRAKON is superior to MINOTAUR at 10% sampling. Increasing the sampling rate from 1% to 10% improves accuracy by +4.7% for DRAKON compared to +2.5% for MINOTAUR. Similarly for MTOP, DRAKON improves by +7.7% with additional samples for HI and TH. MINOTAUR demonstrates a lesser +5.6% and +3.5% for HI and TH respectively. While MINOTAUR is generally an improved parser, we suggest that this contrast in data efficiency for distant target languages is owed to how the explicit alignment observes representations from distant language inputs. This distance between samples is generally larger between distant languages and smaller for related languages. Therefore, MINOTAUR may be weaker for languages with larger initial divergences to optimise. DRAKON can ignore this effect by promoting similarity at the gradient

$\mathbb{D}_{Z\|X}$	$\mathbb{D}_Z$	EN	FR	ES	DE	HI	TH	TARGET AVG.
KL	—	78.3±0.9	70.6±0.4	73.1±1.0	67±0.2	66.6±4.6	48.0±1.8	65.1±9.9
W_2	—	78.6±1.0	72.1±0.4	74.3±1.4	68.7±1.1	67.4±4.4	53.2±1.9	67.1±8.3
—	MMD	78.7±0.9	72.3±0.5	74.3±0.4	68.8±0.7	67.5±0.9	53.3±1.6	67.2±8.2
KL	MMD	78.4±1.8	71.8±5.0	73.3±2.2	68.5±6.5	67.3±3.1	54.0±4.3	67.0±7.7
W_2	MMD	80.2±0.4	72.8±0.5	74.9±0.1	70.0±0.7	68.6±0.5	54.7±2.5	**68.2**±7.9

Table 6.3: MTOP SCIEM Accuracy sampling 10 SPIS ablating alignment methods between conditional-only ($\mathbb{D}_{Z|X}$), marginal-only ($\mathbb{D}_Z$) and joint alignment ($\mathbb{D}_{Z|X} + \mathbb{D}_Z$). We contrast beween Kullback-Leibler Diverence (KL), L_2-Wasserstein distance (W_2), and Maximum Mean Discrepancy (MMD). The joint method using L_2-Wasserstein distance is empirically optimal and significantly above all other methods.

level regardless of the input language. We observe this effect in Figure 6.4, where MINOTAUR struggles to align ZH comparably to more similar languages.

6.3.2 How does MINOTAUR Enable Cross-lingual Transfer?

We report ablations of MINOTAUR on MTOP at 10 SPIS sampling for a case study in how each component of the system influences performance. We generally observe the effect of each ablation extends to additional sampling rates across both datasets.

Which Alignment Feature is Important for MINOTAUR? Table 6.3 considers each function for cross-lingual alignment outlined in Section 6.1.4 as an individual or composite alignment loss. The best approach, used in all other reported results, minimises the L_2 Wasserstein distance (W_2) for conditional posterior divergence and MMD for marginal posterior divergence. The closest competing method is '$\mathbb{D}_Z$ only' to align only marginal posteriors without any control on conditional posteriors. For conditional posteriors, W_2 is significantly superior to the Kullback-Leibler Divergence (KL) for conditional and joint cases. The W_2 distance directly minimises the Euclidean L_2 distance when variances of different languages are equivalent. This, in turn, is more similar to the Maximum Mean Discrepancy function (the best singular objective) which minimises the distance between the approximate means of each marginal distribution. We observe that minimising marginal posterior divergence with MMD is preferable to combining MMD with KL, which underperforms across our all experiments. The W_2 + MMD approach significantly outperforms all other combinations. The performance of MMD, com-

Alignment Measurement	EN	FR	ES	DE	HI	TH	TARGET AVG.
MMD	77.5±1.6	69.6±3.8	70.7±4.8	66.3±4.1	61.7±7.3	53.2±6.6	64.3±7.1
KL	77.9±1.4	69.8±3.7	70.9±4.5	66.5±3.7	62.1±7.0	52.1±5.9	64.3±7.6
W_2	77.1±1.4	69.2±4.0	70.3±4.7	65.8±3.8	61.7±6.8	52.5±6.9	63.3±7.2

Table 6.4: MTOP SCIEM Accuracy sampling 10 SPIS using non-parametric alignment without Z. Here the encoder output, $E_\phi(X)$ is directly input into decoder $G_\psi\left(E_\phi(X)\right)$. We contrast beween Kullback-Leibler Diverence (KL), L_2-Wasserstein distance (W_2), and Maximum Mean Discrepancy (MMD). All approaches significantly underperform relative to Table 6.3.

pared to methods for computing $\mathbb{D}_{Z|X}$, highlights that minimising divergence between *marginal posteriors* is the primary contributor for alignment. Minimising divergence between *conditional posteriors* is a weaker additional contribution. A potential factor for this contrast may be that conditional posterior divergence is more challenging to accurately estimate and optimise. Whereas marginal posterior divergence benefits from estimation using a batch of samples, the conditional posterior alignment appears more sensitive to the measurement of distance. As mentioned in Section 6.1.4, we did not observe a benefit in modifying the computation of conditional posterior divergence beyond an average between all pairwise divergences across parallel inputs. Future investigation is needed to improve estimating and minimising conditional posterior divergence for cross-lingual transfer.

Can MINOTAUR Function Without Latent Variables? We additionally examine if the alignment technique of MINOTAUR is effective for an unmodified Transformer encoder-decoder without latent variable feature. We report a variant of MINOTAUR using three divergence metrics between unbound encoder states in a similar training routine as our main results. The key model difference is the output of the encoder is not parametrically described as a probability distribution.

In Table 6.4, we contrast between MMD, KL divergence (i.e., $\sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$) and Euclidean L_2 distance as tractable metrics between encoder outputs. We observed that cosine distance, a reasonable additional comparison, was insufficiently stable during training due to numerical underflow errors. We omit this distance from Table 6.4. Regardless of distance, we broadly observe that any technique is significantly weaker than a respective counterpart in Table 6.3. This contrast suggests the smooth curvature and structure of the Gaussian parameterisation (Reizinger and Huszár, 2023) contribute

to effective cross-lingual alignment. Another contribution of the latent variable approach is some degree of non-determinism within the encoding space through random sampling. The augmented model can assign semantic meaning to some latent neighbourhood, with respective mean and probability density, rather than the exact point output by the deterministic encoder. We conjecture that the structure of similar meaning as a *neighbourhood*, rather than a *exact point*, in the latent space will support how task knowledge can be transferred between languages. Practically, these non-parametric approaches are also challenging to implement. The lack of closed-form expressions of distance (e.g., Equation (6.18) or Equation (6.17)) leads to numerical underflow instability during training. Even using MMD, which does not require an exact solution, fared poorer without the latent variable structure.

Does MINOTAUR Require Parallel Data? The methodology for MINOTAUR and training algorithm in Algorithm 2 mandate that MINOTAUR requires parallel inputs from different languages to align equivalent semantics across conditional posteriors. To study if MINOTAUR is robust to cross-lingual alignment without this requirement, we develop an ablation on MINOTAUR which optimises alignment between *non-parallel* inputs (i.e., x_l is not a translation of x_{EN} and outputs are not equivalent). We intuitively expect parallelism is needed to successfully learn the similarity between representations with equivalent semantics.

Table 6.5 shows that data parallelism is surprisingly *not required* using '$\mathbb{D}_{Z|X}$ only' to align marginal distributions *only*. The '$\mathbb{D}_{Z|X}$ only' and '$\mathbb{D}_{Z|X} + \mathbb{D}_Z$' techniques significantly under-perform relative to equivalent methods using parallel data. This is anticipated as alignment between conditional posterior samples from inequivalent inputs *should likely not be aligned*. Therefore, forcing this alignment could add unnecessary noise in training. However, the '$\mathbb{D}_{Z|X}$ only' method is significantly above other methods with the closest performance to the parallel equivalent. Parallel and non-parallel methods for '$\mathbb{D}_Z$ only' alignment differs by the smallest 0.3%. We note that this difference is still significant. MINOTAUR using only conditional posterior alignment is $\geq 2.4\%$ weaker without parallel data. The minimal difference between marginal posterior alignment supports our interpretation that MMD aligns at a 'language level'. We conjecture that marginal posterior alignment should not require parallel data, provided sufficient samples to estimate $Q(Z)$ for each language.

		EN	FR	ES	DE	HI	TH	TARGET AVG.
Parallel Data	$\mathbb{D}_{Z\|X}$ only	78.6±1.0	72.1±0.4	74.3±1.4	68.7±1.1	67.4±4.4	53.2±1.9	67.1±8.3
	$\mathbb{D}_Z$ only	78.7±0.9	72.3±0.5	74.3±0.4	68.8±0.7	67.5±0.9	53.3±1.6	67.2±8.2
	$\mathbb{D}_{Z\|X} + \mathbb{D}_Z$	80.2±0.4	72.8±0.5	74.9±0.1	70±0.7	68.6±0.5	54.7±2.5	**68.2**±7.9
Non-parallel Data	$\mathbb{D}_{Z\|X}$ only	78.9±0.2	67.3±5.6	68.3±3.8	64.6±4.9	59.4±3.4	54.0±3.6	62.7±6.0
	$\mathbb{D}_Z$ only	77.6±1.6	71.5±0.4	72.9±1.1	68.4±0.5	67.2±0.8	54.5±0.9	66.9±7.3
	$\mathbb{D}_{Z\|X} + \mathbb{D}_Z$	78.8±0.2	70.9±0.5	71.9±0.4	67.9±0.5	64.5±0.6	540±1.4	65.8±7.2

Table 6.5: MTOP SCIEM Accuracy sampling 10 SPIS using parallel data inputs (above), and non-parallel data inputs (below) between languages in MINOTAUR. We sample English input, x_{EN}, and an input in language l, x_l which is *not* a translation of x_{EN} for Equation (6.21). This approach weakens individual posterior alignment but identifies that MMD is the least sensitive to input parallelism. The significant best result is bolded.

6.3.3 Visualising Latent Representation Similarity

We study the representation space learned by MINOTAUR for MultiATIS++SQL at 1% sampling using the same t-SNE method from Section 2.4.2. For MINOTAUR, we average the sampled $\mathbf{z}$ states for each input sentence for comparison to previous chapters. We consider if explicit representation alignment emerges from training with MINOTAUR, and how this alignment contrasts to MBART50 (Tang et al., 2021) and methods from Chapters 3 to 5.

We visualise the contrasting alignment from different methods in Figure 6.4 using t-SNE. Table 6.6 also shows a quantitative comparison between methods. We identify that MINOTAUR produces the visually clearest representation alignment showing representations from different languages mapping to the same *semantic* cluster. MINOTAUR uses the latent variable structure to organise different input utterances to a latent distribution with the same semantics. MINOTAUR also produces less monolingual clustering artefacts compared to MBART50 and prior methods. We observe that MINOTAUR does not align ZH (pink) as successfully as other languages, echoing the challenge in cross-lingual alignment between distant languages observed in Chapter 3 and Chapter 4. We estimate this is an artefact of how explicit alignment is minimised. However, our explicit method overall demonstrates the closest alignment compared to previous chapters.

We quantitatively verify the validity of MINOTAUR alignment by analysing the high-dimensional similarity between encodings in Table 6.6. MINOTAUR produces representations which are more cosine-similar than any prior method, with a 63%

Method Type	Method	Cosine (↑)	Top-1	Top-5	Top-10
Pre-trained Model	MBART50	0.576	0.521	0.745	0.796
Lower Bound	Train-EN Only	0.364	0.669	0.775	0.964
Upper Bound	MULTILINGUAL Gold	0.698	0.784	0.981	0.991
Machine Translation	FATES	0.670	0.720	0.957	0.992
Zero-shot	ZEUS	0.760	0.832	0.944	0.971
Few-shot @1% sampling	Train-EN∪ALL	0.470	0.634	0.835	0.877
	Train-EN→FT-ALL	0.541	0.714	0.898	0.922
	Reptile-EN→FT-ALL	0.673	0.702	0.920	0.946
	DRAKON	0.844	0.797	0.949	0.963
	MINOTAUR	**0.941**	**0.874**	**0.994**	**0.998**

Table 6.6: Average similarity between encodings of English and target languages for MultiATIS++SQL. Cosine similarity evaluates average distance between encodings of parallel sentences. Top-k evaluates if the parallel encoding is ranked within the k most cosine-similar vectors (higher (↑) is better). Best excluding the upper-bound is bold.

improvement in similarity compared to the original MBART50 pre-trained model. Considering the Top-k ranked similarity, Table 6.6 identifies that $> 99\%$ of representations learned by MINOTAUR have a parallel utterance within the five closest representations. We interpret this as the representation space using MINOTAUR is more *semantically distributed* relative to MBART50 or our previous methods. Representations from MINOTAUR for a given utterance are closer to semantic equivalents than an arbitrary utterance in the same language. This technique offers the best solution to our overarching hypothesis of learning cross-lingual semantic parsing via accurate and robust representation alignment.

6.3.4 Error Analysis

Similar to Chapter 5, we contrast the model predictions at 1% sampling between baselines and MINOTAUR. We specifically analyse the MultiATIS++SQL examples where MINOTAUR improves on DRAKON, and analyse the improvement over baselines similar to previous chapters.

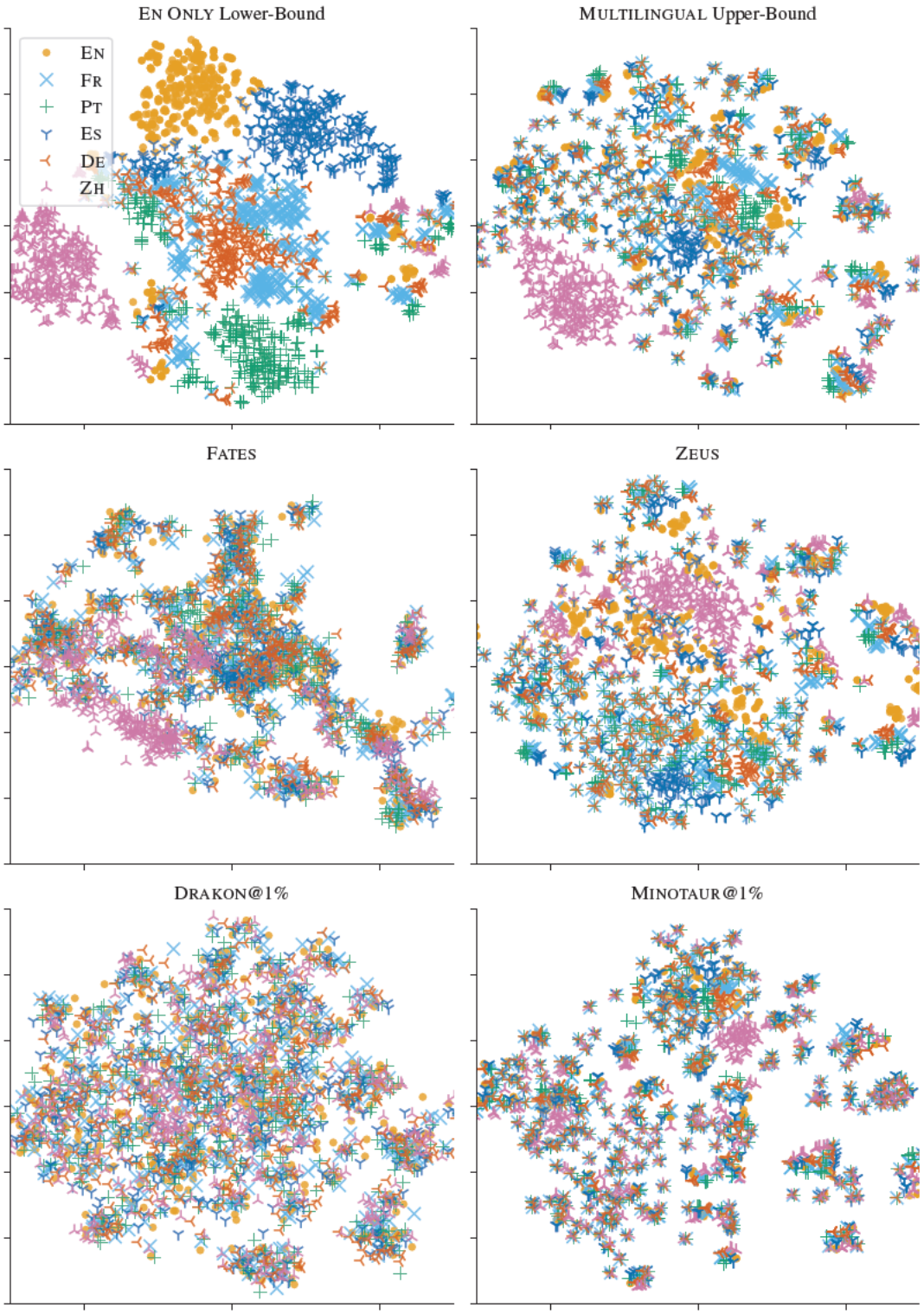

Figure 6.4: Visualisation of MultiATIS++SQL encodings (test set; 50% random parallel sample) using t-SNE (van der Maaten and Hinton, 2008). We compare the original MBART50 pre-trained model, the EN-ONLY zero-shot lower bound, MULTILINGUAL training upper bound, FATES from Chapter 3, ZEUS from Chapter 4, DRAKON from Chapter 5 and MINOTAUR. Compared to MBART50 and our previous proposals, MINOTAUR organises the latent space to be more *semantically distributed* across languages without monolingual separability. We observe improved semantic clusteringshown quantitatively in Table 6.6.

Multi-Word Expressions The primary improvement arises from improved handling of multi-word expressions and language-specific modifiers. For example, adjectives in English are often multi-word adjectival phrases in French (e.g., "cheapest" → "le moins cher" or "earliest" → "à plus tot"). Improved handling of this error type accounts for an average of 54% of improvement across languages, with the highest in French (68%) and lowest in Chinese (38%). We estimate that a combination of marginal and conditional posterior alignment in MINOTAUR benefits this specific case where semantics are expressed in varying numbers of words between languages. While this could be similarly approached using fine-grained token alignment labels, MINOTAUR improves transfer in this context without additional annotation.

This is a similar pattern to the improvement in handling 'Modifier Phrases' in Chapter 5. We identify that MINOTAUR makes similar improvements as DRAKON but makes further improvement in recognising these multi-word modifiers in target languages. While this analysis is prominent for French, it is unclear why the transfer to Chinese is weaker. A potential interpretation is that weaker transfer of multi-word expressions to Chinese could be related to poor sub-word tokenisation. Most sub-word tokenisation of Chinese decomposes into single character subwords (Hofmann et al., 2022). Therefore, multi-word expressions may be inadequately decomposed into constituent semantic units in the pre-trained vocabulary. Sub-optimal tokenisation of logographic or information-dense languages is an ongoing debate (Hofmann et al., 2022; Si et al., 2023) and requires further study beyond the scope of this thesis.

Executability Translation-based models and zero-shot systems often generate malformed, non-executable SQL. Additional improvement from MINOTAUR is observed from a 24% reduction in generating ill-formed SQL executed within a database. Syntactic correctness is critical when a parser encounters a rare entity or unfamiliar linguistic construction, and this improvement highlights how our model can better navigate inputs from languages minimally observed during training. We observed a similar improvement for ZEUS in Chapter 4. We suggest that both methods of representation alignment limit the encoder from generating representations which the decoder cannot interpret. By limiting the generation of 'malformed' representations, erratic predictions from unfamiliar encodings are reduced. This leaves only semantic errors within syntactically well-formed SQL to be addressed.

6.4 Related Work

Optimal Transport (OT) in Natural Language Processing is often used as an unsupervised solution to matching distributions from disparate domains. This framework proves useful when the input and output domains are known (e.g., P_X and P_Y) but the joint distribution, $P(X,Y)$, is unknown or intractable to estimate. OT manifests in tasks requiring some alignment or discrete mapping between domains such as cross-lingual embedding alignment or bilingual lexicon induction. Estimating this cross-domain coupling often employs Sinkhorn distances (Cuturi, 2013; Feydy et al., 2019) to iteratively optimise the coupling as an online minimisation sub-problem during training. Sinkhorn distances interpolate between Wasserstein distances and Maximum Mean Discrepancy. This is similar to MINOTAUR but requires an additional inner loop of iterative convergence to estimate the cross-domain coupling. MINOTAUR avoids using Sinkhorn distances by introducing the parametric posterior distribution to exploit closed-form divergences.

Zhang et al. (2017) use the Sinkhorn formulation of Optimal Transport to align static embedding matrices from different languages. This method differs from geometric methods, e.g., Conneau et al. (2017), by modelling the embeddings as distributions to compute a Sinkhorn distance between known embedding pairs. Zhang et al. (2017) use the Sinkhorn distance between cross-lingual token pairs to initialise the embedding mapping, and subsequently use geometric alignment to complete the mapping between embedding spaces. Alvarez-Melis and Jaakkola (2018) extend this method for bilingual lexicon induction by modelling the complete embedding matrices as distributions, rather than token-level pairings. Alvarez-Melis and Jaakkola (2018) use the Gromov-Wasserstein distance (Mémoli, 2011), an extension of the Wasserstein distance useful for comparing metric spaces directly, to minimise the distance between the complete embedding space across languages. This generalises the technique of Zhang et al. (2017) and removes the need for an additional geometric solution. Alqahtani et al. (2021) propose an alternative extension for contextual (i.e., BERT style) embeddings to learn word-alignment pairings from bitext corpora in an unsupervised fashion. Applying the Sinkhorn distance transformation to contextual embeddings demonstrates significant improvement in cross-lingual transfer from English for entailment and question answering.

Methods for explicit alignment often employ contrastive learning to "pull" similar representations closer and "push" dissimilar representations apart. Wu and Dredze (2020) propose a contrastive similarity loss, extending Alqahtani et al. (2021), using

cosine distance to improve cross-lingual transfer by fine-tuning on additional bitext. This method demonstrates improvement on a range of sequence-labelling tasks and is less computationally intensive than the above Sinkhorn-based methods. Recently within spoken-language understanding, Qin et al. (2022) and Liang et al. (2022) propose similar multi-level explicit alignment methods. These methods approximately learn an utterance-, label- and word-level alignment using contrastive learning to improve representation quality for cross-lingual transfer. While these techniques improve performance, they require expensive fine-grained annotations of word- and label-alignment across utterances. Wieting et al. (2023) demonstrate that using the same variational representation space removes the requirement for contrastive learning when learning embeddings for multilingual information retrieval. Wieting et al. (2023) optimise a single probabilistic representation space for both monolingual separability and multilingual representation similarity. This is shown to improve multilingual feature sharing and retain useful monolingual features, with an overall improvement in cross-lingual passage retrieval for question answering. While similar to MINOTAUR, this approach requires parallel data to learn any cross-lingual similarity. We show in Section 6.3 that MINOTAUR can succeed without this requirement.

Within semantic parsing, prior work validates that latent structure can improve performance and provide some interpretability of the latent space. Kočiský et al. (2016) demonstrate an early method of using a variational latent space to improve semantic parsing. The intermediate state of a recurrent neural network parser is parameterised as a Gaussian distribution, enabling robust sampling from the latent space to generate auxiliary logical forms. These samples are used for additional self-supervised auto-encoder training. This semi-supervised training improves parsing in domains with minimal data where logical forms are expensive to annotate. Yin et al. (2018) extend this approach using a tree-based hierarchical structure to model the latent posterior as the logical form. This reframes the task entirely as an auto-encoding objective with the logical form as the latent state intermediary between encoder and decoder. While both methods demonstrate the utility of latent structure in semantic parsing, neither is applied in a cross-lingual setting. MINOTAUR is the first to jointly model semantic parsing and cross-lingual transfer as an Optimal Transport task. This work is the most recent and novel to be published from this thesis leading to lesser recent follow up research. Recently, efforts such as Li et al. (2024) have further validated the importance of latent representation alignment within generative tasks to improve cross-lingual transfer.

6.5 Summary

This chapter proposes a second few-shot cross-lingual semantic parsing approach using probabilistic latent variables to learn cross-lingual representation alignment. In Section 6.1, we propose MINOTAUR: a method for transferring task knowledge to a target language only by aligning latent representations from different languages. MINOTAUR uses an Optimal Transport-based joint coarse- and fine-grained divergence penalty encouraging latent encodings of target languages to match the equivalent latent representation for English. We periodically augment the typical cross-entropy sequence generation loss with the MINOTAUR loss to contribute latent representation alignment jointly with learning the parsing function. Similar to Chapter 4, our intuition is that matching the latent representation for English in target languages enables target language parsing using a few annotated examples.

We validate our hypothesis that explicit representation alignment enables sample-efficient cross-lingual transfer. In Section 6.3, we identify why MINOTAUR generally outperforms DRAKON, ZEUS, and FATES as the most accurate target language parser. Notably, MINOTAUR is more accurate on average than DRAKON at every few-shot sampling level for both MultiATIS++SQL and MTOP. MINOTAUR is also more sample efficient than DRAKON in demonstrating superior accuracy with smaller samples. We observe that MINOTAUR is beneficially due to aligning marginal distributions between languages using batch-level estimation of each distribution. Ablations identify where MINOTAUR requires parallel data and the importance of the latent variable structure within the model for stability. We highlight that MINOTAUR aligning using only marginal distributions relaxes the parallel data constraint for alignment without parallel task-specific in target languages. Further analysis identifies that MINOTAUR can produce a more semantically distributed latent space with representations clustering by *meaning* rather than *source language*. MINOTAUR is the best-performing system in this thesis; additionally producing the most language-agnostic latent representations by directly targeting representation alignment.

Chapter 7

Conclusions and Future Work

7.1 Contributions

In this thesis, we consider the task of cross-lingual semantic parsing with limited resources. In Chapter 1, we outline a case study of an engineer prototyping expanding a semantic parser human-computer interface to support multiple additional languages beyond English. We propose strategies for this objective oriented towards data-efficient cross-lingual transfer requiring minimal additional data collection or annotation. Our modelling contributions focus on learning representation alignment between latent representations from different languages within the parser. This thesis reveals that accurate cross-lingual semantic parsing is possible by composing samples of target language data within models explicitly optimising for both parsing and cross-lingual transfer. Furthermore, we highlight the varying merits of machine translation, data borrowed from adjacent tasks, and few-shot sampling of the target language data distribution. The contributions in this thesis are commercially valuable, proposing economical strategies for assistant technologies to support more languages, and novel for cross-lingual transfer, proposing new methods for cross-lingual modelling and optimisation.

Our overarching hypothesis considers if **cross-lingual semantic parsing is enabled by aligning representations between English and target languages**. Under this view, we consider strategies to learn this alignment contingent upon sampling various true or synthetic data distributions. Globally, our results validate this hypothesis as we observe broad improvement in parsing in step with improved cross-lingual representation alignment. Between Chapters 3 to 6, techniques producing greater representation alignment also produce more accurate cross-lingual transfer. For our case study, compromising

between data quality, budget, model complexity and computation resources is necessary to choose the best parser.

In Chapter 3, we considered the hypothesis that **machine translation can adequately approximate natural language for cross-lingual transfer in semantic parsing**. This investigation studied whether machine translation can be used as an intermediary component in the parsing pipeline for representation alignment. Sufficient accuracy in machine translation offers a low-cost strategy for cross-lingual semantic parsing without requiring expert data annotation. We observe this hypothesis to be invalid, as machine-translated data is insufficiently natural to generalise to gold data from fluent target language speakers. Translating the test data in target languages to English provides a robust baseline. However, errors in interpreting entities and unnatural phrasing limit the performance without training on target language data directly. When translating data from English to target languages, we observed similar issues but identified improved solutions using data augmentation or model ensembling. Our model, FATES, combines multiple parallel encoders learning from multiple machine translation sources to increase the syntactic diversity for target language training data. This yields improvements in parsing all target languages and moderate improvement in representation alignment. FATES is competitive for some datasets and languages with the lowest cost of all methods proposed in this thesis.

In Chapter 4, we considered the hypothesis that **auxiliary multilingual tasks and data can induce cross-lingual representation alignment without target-language training data**. Given the deficiency with machine translation, we evaluated borrowing target language data from adjacent tasks to learn cross-lingual representation alignment. If auxiliary tasks produce generalisable representation alignment, then semantic parsing data for target languages is unnecessary for transfer. We propose ZEUS, a multi-task model combining monolingual semantic parsing with three auxiliary tasks using multilingual data. Our results validate the hypothesis that multilingual auxiliary tasks improve representation alignment for zero-shot cross-lingual semantic parsing. ZEUS improves upon FATES in representation similarity and parsing without introducing unnatural data from machine translation. Each auxiliary task in ZEUS contributes by fine-tuning the encoder to each target language or minimising the language-specificity of any latent encoding. Furthermore, we observe that variation in surface form diversity and style contributes significantly to parsing performance. The optimal ZEUS model trained auxiliary tasks using text in question-style syntax from all target languages. However, we identify this improvement is insufficient to compete with the upper-bound

of fully supervised training (i.e., 100% translation) due to a domain mismatch between auxiliary task data and parsing data.

In Chapter 5, we considered the hypothesis that **meta-learning improves few-shot cross-lingual transfer by promoting gradient-level cross-lingual regularisation of task-specific training**. We identified how machine translation and auxiliary tasks are challenged by fluency and domain relevance in data sources. To counter this, we considered if a few examples from target languages are more valuable for cross-lingual transfer without suffering the same issues. This few-shot approach finally considers the value of sampling the true target language data distribution, and how many samples we need for accurate transfer. We considered a meta-learning approach optimised for data-efficient few-shot cross-lingual transfer subject to translating $\leq 10\%$ of the English data. By 'learning to learn', meta-learning offers a promising strategy for rapidly learning to parse any target language from a few examples. Chapter 5 proposed DRAKON, a meta-learning algorithm optimised for cross-lingual transfer within semantic parsing. DRAKON optimises a meta-gradient for the task and aligns this gradient with a few-shot samples from the target languages. This gradient-level alignment integrates task-specific optimisation with cross-language task transfer to optimise in a mutually beneficial direction for English and target language gradients. Our results validate our hypothesis, observing improved representation alignment and more accurate cross-lingual transfer. DRAKON integrates the meta-training and fine-tuning stages of meta-learning into a single optimisation stage, surpassing all adjacent few-shot strategies in performance and representation alignment.

Finally, Chapter 6 considered the hypothesis that **explicit cross-lingual representation alignment improves the transfer of task knowledge to target languages**. We considered directly optimising the representation alignment objective, rather than studying this as an outcome from previous modelling contributions. Chapter 6 augmented a parser with a latent variable between the encoder and decoder. This model redefines the latent representation over a probability distribution. We proposed to minimise the divergence between probabilistic latent variables in a few-shot learning routine. We expect that minimising latent variable divergence directly optimises cross-lingual representation alignment within the parser. Our method, MINOTAUR, uses Optimal Transport to minimise a divergence penalty between latent variables from different languages. MINOTAUR measures divergence between latent variables at coarse and fine granularities to align both global and local representation structures. We observe that MINOTAUR surpasses all other methods including baselines, few-shot meta-learning

with DRAKON, FATES using machine translation, and zero-shot ZEUS using auxiliary tasks. This validates our hypothesis that explicit representation alignment surpasses methods encouraging implicit alignment to improve cross-lingual transfer. We also analytically identified that MINOTAUR learns the closest representation alignment with the clearest semantic similarity between representations with equivalent meaning from different languages.

Across our data sampling strategies, we find that **gold data is more valuable than silver data** for cross-lingual transfer in semantic parsing. In Chapter 3, we discussed the limitations of machine translation generally as errors in producing natural translations. We identified entity mishandling, unnatural phrasing, and erroneous disambiguation as major challenges to improving how machine translation can be exploited for our task. However, FATES demonstrated that providing multiple machine translation sources partially mitigates these errors using more diverse data. In Chapter 4, we identified that domain and syntax similarity is desirable for auxiliary task data. Zero-shot parsing suffered when auxiliary data poorly resembled the parsing test data e.g., using declarative sentence text from arbitrary sources. These limitations motivated our study of few-shot data sampling in Chapter 5 and Chapter 6. We observe that using a small sample of domain-relevant target language data enabled a model to learn improved cross-lingual similarity without distractors of unnatural or irrelevant data. Both DRAKON and MINOTAUR generally improved upon methods without few-shot sampling. Auditing of data quality is also no longer required as the few-shot samples are natural and relevant by design.

Considering our original case study in Chapter 1, our findings indicate that few-shot sampling of the target language is the most reliable strategy for accurate cross-lingual semantic parsing. We propose that the expense of modest expert annotation is more valuable than using cheaper sampling approaches producing poorer parsing. Few-shot sampling is best integrated with the MINOTAUR model, proving more data-efficient and requiring fewer training resources than DRAKON. The MINOTAUR parser demonstrated the most accurate parsing with the fewest resources to best satisfy the constraints of the case study. However, we identify that DRAKON is more accurate for languages highly dissimilar to English. We conjecture that this is due to the implicit methodology able to disregard language similarity during learning. Therefore, DRAKON could be a preferred strategy for prototyping a parser for distant target languages. Additionally, FATES and ZEUS provide competitive strategies for prototyping a parser if few-shot data sampling is beyond the available budget. The contributions of this thesis are as follows:

MultiATIS++SQL: In Section 2.3, we create a new split of the ATIS dataset Hemphill et al. (1990); Dahl et al. (1994) for executable cross-lingual transfer. This allows the evaluation of grounded cross-lingual transfer from English to French, German, Portuguese, Spanish, German and Simplified Chinese.

FATES: We propose a multi-encoder model for improving cross-lingual semantic parsing with silver-standard data sources.

ZEUS: We propose a multi-task model for learning representation alignment from English-language semantic parsing and multilingual auxiliary tasks.

DRAKON: We propose a meta-learning algorithm for few-shot cross-lingual semantic parsing. Our algorithm efficiently combines task-specific and cross-lingual gradients for cross-lingual transfer.

MINOTAUR: We propose an explicit representation alignment methodology for few-shot cross-lingual semantic parsing. MINOTAUR uses a latent-variable encoder-decoder model with representation alignment using Optimal Transport for data-efficient cross-lingual transfer.

We hope that our contributions to modelling, optimisation and data sampling motivates further study of achieving cross-lingual transfer within semantic parsing and challenging future tasks.

7.2 Future Work

Larger language models Chapter 6 presents a trade-off between model size and data quality wherein large language models (LLMs) can better exploit lower-quality data (e.g., from machine translation or prompt-based generation). We observe that these methods perform competitively to smaller models using gold-standard data such as MINOTAUR. We discuss this as a dichotomy, given the shortcomings of combining silver-standard data with smaller models from Chapter 3. However, the inverse case using an LLM with gold-standard data is a potential future investigation. We do not consider this approach as we judge the required computation, training and annotation expense as beyond feasible for our case study. Recent work validates the potential for few-shot cross-lingual transfer across multiple natural language understanding tasks excluding semantic parsing (Tanwar et al., 2023; Asai et al., 2023). However, there is an

ongoing debate between fine-tuning or in-context learning as best practice. In-context-learning (ICL) has the potential to be more efficient without requiring additional training but can suffer for languages less prevalent during pre-training (Ruder et al., 2023). Fine-tuning is more reliable for these lower-resource languages, but the computation required to robustly fine-tune an LLM is beyond the resources of many researchers. We raise that further study is needed to combine LLMs with cross-lingual semantic parsing using gold data. Future work must address this compromise between fine-tuning and ICL, and investigate how to yield accurate parsing when long SQL statements and serialised databases can consume much of the ICL context window. Methods such as latent variable alignment with MINOTAUR are potentially applicable at an LLM scale to improve parsing with latent structure. We raise that an LLM-driven cross-lingual parser can likely surpass the accuracy of parsers outlined in this thesis. However, the greater challenge is surpassing our methods' data and computation efficiency while retaining any performance gain. Future work could consider parameter-efficient fine-tuning (e.g., low-rank adaptation Hu et al. (2022)) for integrating cross-lingual transfer with semantic parsing.

Advancing Cross-lingual Alignment We study cross-lingual representation alignment within the context of semantic parsing. However, many additional tasks might benefit from cross-lingual representation alignment e.g., retrieval-oriented tasks, machine translation or cross-lingual summarisation. Our methods for latent similarity in cross-lingual transfer can be broadened to enable more general task transfer. For a retrieval example, an aligned representation space in a Retrieve-and-Generate system (Lewis et al., 2020c, RAG) could condition upon a new language by only aligning the embedding space for retrieving context. Wieting et al. (2023) propose latent variable representation alignment for cross-lingual question answering using bitext. MINOTAUR using marginal posterior alignment only could improve learning a representation space for languages where bitext is scarce or entirely unavailable. With parallel data available, MINOTAUR could be decomposed in incremental stages to learn representation alignment from parallel and non-parallel data. First, coarse-grained alignment could use non-parallel data, which is easier to source without demanding annotation. Second, any parallel data could be used for coarse- and fine-grained alignment similar to Chapter 6 to further improve latent similarity. This could maximise representation similarity by exploiting all available corpora for a target language.

Bibliography

Agrawal, P., Alberti, C., Huot, F., Maynez, J., Ma, J., Ruder, S., Ganchev, K., Das, D., and Lapata, M. (2023). QAmeleon: Multilingual QA with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Ahmad, W. U., Zhang, Z., Ma, X., Chang, K.-W., and Peng, N. (2019). Cross-lingual dependency parsing with unlabeled auxiliary languages. In Bansal, M. and Villavicencio, A., editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382, Hong Kong, China. Association for Computational Linguistics.

Alqahtani, S., Lalwani, G., Zhang, Y., Romeo, S., and Mansour, S. (2021). Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov-Wasserstein alignment of word embedding spaces. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

Ananiadou, S., McNaught, J., and Thompson, P. (2012). *The English Language in the Digital Age*. META-NET White Paper Series. Springer Nature, United States. EC FP7 PSP Project METANET4U.

Andreas, J., Vlachos, A., and Clark, S. (2013). Semantic parsing as machine translation. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 47–52, Sofia, Bulgaria. Association for Computational Linguistics.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019). The missing ingredient in zero-shot neural machine translation. *ArXiv preprint*, abs/1903.07091.

Artetxe, M., Goswami, V., Bhosale, S., Fan, A., and Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2020). Translation artifacts in cross-lingual transfer learning. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Asai, A., Kudugunta, S., Yu, X. V., Blevins, T., Gonen, H., Reid, M., Tsvetkov, Y., Ruder, S., and Hajishirzi, H. (2023). Buffet: Benchmarking large language models for few-shot cross-lingual transfer.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Beaulieu, A. (2009). *Learning SQL*. O'Reilly Media, Sebastopol, CA, 2 edition.

Berant, J. and Liang, P. (2014). Semantic Parsing via Paraphrasing. In *Proceedings of*

the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1415–1425, Stroudsburg, PA, USA.

Blasi, D., Anastasopoulos, A., and Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Chandu, K. R., Bisk, Y., and Black, A. W. (2021). Grounding 'grounding' in NLP. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Chang, T., Tu, Z., and Bergen, B. (2022). The geometry of multilingual language model representations. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. (2021). Evaluating large language models trained on code.

Chen, X., Ghoshal, A., Mehdad, Y., Zettlemoyer, L., and Gupta, S. (2020). Low-resource domain adaptation for compositional task-oriented semantic parsing. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics.

Christiansen, M. H. and Kirby, S. (2003). *Language evolution / edited by Morten*

H. Christiansen, Simon Kirby. Studies in the evolution of language ; 3. Oxford University Press, Oxford.

Cieri, C., Maxwell, M., Strassel, S., and Tracey, J. (2016). Selection criteria for low resource language programs. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Conklin, H., Wang, B., Smith, K., and Titov, I. (2021). Meta-learning to compositionally generalize. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3322–3335, Online. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018a). XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018b). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Culbertson, J., Schouwstra, M., and Kirby, S. (2020). From the World to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, 9:696–717.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., and Shriberg, E. (1994). Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Dankers, V., Lucas, C., and Titov, I. (2022). Can transformer be too compositional? analysing idiom processing in neural machine translation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J.,

Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dong, L. and Lapata, M. (2016). Language to logical form with neural attention. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43, Berlin, Germany. Association for Computational Linguistics.

Dong, L. and Lapata, M. (2018). Coarse-to-fine decoding for neural semantic parsing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017a). Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017b). Learning to paraphrase for question answering. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Dou, L., Gao, Y., Pan, M., Wang, D., Che, W., Zhan, D., and Lou, J.-G. (2023). Multispider: Towards benchmarking multilingual text-to-sql semantic parsing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., and Zhou, D. (2023). Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Duong, L., Afshar, H., Estival, D., Pink, G., Cohen, P., and Johnson, M. (2017a). Multilingual semantic parsing and code-switching. In Levy, R. and Specia, L., editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389, Vancouver, Canada. Association for Computational Linguistics.

Duong, L., Afshar, H., Estival, D., Pink, G., Cohen, P., and Johnson, M. (2017b). Multilingual Semantic Parsing And Code-Switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 379–389, Vancouver, Canada.

Eberhard, D., Simons, G., and Charles (2019). Languages of the World. *Ethnologue: Languages of the World. Twenty-Second Edition*, 22.

Edunov, S., Ott, M., Ranzato, M., and Auli, M. (2020). On the evaluation of machine translation systems trained with back-translation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR.

Finegan-Dollak, C., Kummerfeld, J. K., Zhang, L., Ramanathan, K., Sadasivam, S., Zhang, R., and Radev, D. (2018). Improving text-to-SQL evaluation methodology. In

Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman Vural, F. T., and Cho, K. (2016). Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Stroudsburg, PA, USA.

FitzGerald, J. G. M., Ananthakrishnan, S., Arkoudas, K., Bernardi, D., Bhagia, A., Bovi, C. D., Cao, J., Chada, R., Chauhan, A., Chen, L., Dwarakanath, A., Dwivedi, S., Gojayev, T., Gopalakrishnan, K., Gueudre, T., Hakkani-Tür, D., Hamza, W., Hueser, J., Jose, K. M., Khan, H., Liu, B., Lu, J., Manzotti, A., Natarajan, P., Owczarzak, K., Oz, G., Palumbo, E., Peris, C., Prakash, C. S., Rawls, S., Rosenbaum, A., Shenoy, A., Soltan, S., Harakere, M., Tan, L., Triefenbach, F., Wei, P., Yu, H., Zheng, S., Tur, G., and Natarajan, P. (2022). Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 2893–2902, New York, NY, USA. Association for Computing Machinery.

Fodor, J. A. (1975). *The Language of Thought*, volume 87. Harvard University Press.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.*, 20(3–4):121–136.

Gal, Y. and Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29:*

Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 1019–1027.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Ganitkevitch, J. and Callison-Burch, C. (2014). The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4276–4283, Reykjavik, Iceland.

Garmash, E. and Monz, C. (2016). Ensemble Learning for Multi-Source Neural Machine Translation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Gu, J., Wang, Y., Chen, Y., Li, V. O. K., and Cho, K. (2018). Meta-learning for low-resource neural machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Gupta, S., Shah, R., Mohit, M., Kumar, A., and Lewis, M. (2018). Semantic parsing for task oriented dialog using hierarchical representations. In Riloff, E., Chiang,

D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Harris, Z. (1960). *Structural Linguistics*. Phoenix books. University of Chicago Press.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Hershcovich, D., Aizenbud, Z., Choshen, L., Sulem, E., Rappoport, A., and Abend, O. (2019). SemEval-2019 task 1: Cross-lingual semantic parsing with UCCA. In May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., and Mohammad, S. M., editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1–10, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., and Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Herzig, J. and Berant, J. (2018). Decoupling structure and lexicon for zero-shot semantic parsing. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1619–1629, Brussels, Belgium. Association for Computational Linguistics.

Hofmann, V., Schuetze, H., and Pierrehumbert, J. (2022). An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(09):5149–5169.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Huang, P.-S., Wang, C., Singh, R., Yih, W.-t., and He, X. (2018). Natural language to structured query generation via meta-learning. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 732–738, New Orleans, Louisiana. Association for Computational Linguistics.

Huang, X. (1990). A machine translation system for the target language inexpert. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Iyer, S., Konstas, I., Cheung, A., Krishnamurthy, J., and Zettlemoyer, L. (2017). Learning a neural semantic parser from user feedback. In Barzilay, R. and Kan, M.-Y.,

editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Jia, R. and Liang, P. (2016). Data recombination for neural semantic parsing. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Jones, B., Johnson, M., and Goldwater, S. (2012). Semantic parsing with Bayesian tree transducers. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496, Jeju Island, Korea. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kamath, A. and Das, R. (2019). A survey on semantic parsing. In *Proceedings of the 1st Conference on Automated Knowledge Base Construction, AKBC*, Amherst, MA, USA.

Kann, K. (2023). Keynote talk. Third Workshop on Multilingual Representation Learning.

Kantorovich, L. (1958). On the translocation of masses. *Management Science*, 5(1):1–4.

Kate, R. J., Wong, Y. W., and Mooney, R. J. (2005). Learning to transform natural to formal languages. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, pages 1062–1068. AAAI Press.

Kedia, A., Chinthakindi, S. C., and Ryu, W. (2021). Beyond reptile: Meta-learned dot-product maximization between gradients for improved single-task regularization. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 407–420, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Keung, P., Lu, Y., Salazar, J., and Bhardwaj, V. (2020). Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ko, W.-J., El-Kishky, A., Renduchintala, A., Chaudhary, V., Goyal, N., Guzmán, F., Fung, P., Koehn, P., and Diab, M. (2021). Adapting high-resource NMT models to translate low-resource related languages without parallel data. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 802–812, Online. Association for Computational Linguistics.

Kočiský, T., Melis, G., Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., and Hermann, K. M. (2016). Semantic parsing with semi-supervised sequential autoencoders. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1078–1087, Austin, Texas. Association for Computational Linguistics.

Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors (2022). *Proceedings of*

the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kollar, T., Berry, D., Stuart, L., Owczarzak, K., Chung, T., Mathias, L., Kayser, M., Snow, B., and Matsoukas, S. (2018). The Alexa meaning representation language. In Bangalore, S., Chu-Carroll, J., and Li, Y., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 177–184, New Orleans - Louisiana. Association for Computational Linguistics.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., and Steedman, M. (2011). Lexical generalization in CCG grammar induction for semantic parsing. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Lee, H.-y., Vu, N. T., and Li, S.-W. (2021). Meta learning and its applications to natural language processing. In Chiang, D. and Zhang, M., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 15–20, Online. Association for Computational Linguistics.

Lee, K.-i., Kim, S., and Jung, K. (2023). Weakly supervised semantic parsing with execution-based spurious program filtering. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6894, Singapore. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020a). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lewis, P., Oguz, B., Rinott, R., Riedel, S., and Schwenk, H. (2020b). MLQA: Evaluating

cross-lingual extractive question answering. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020c). Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Li, C., Wang, S., Zhang, J., and Zong, C. (2024). Improving in-context learning of multilingual generative language models with cross-lingual alignment. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., and Mehdad, Y. (2021). MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Li, Z., Guo, J., Liu, Q., Lou, J.-G., and Xie, T. (2022). Exploring the secrets behind the learning difficulty of meaning representations for semantic parsing. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3616–3625, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Li, Z., Qu, L., Cohen, P., Tumuluri, R., and Haffari, G. (2023). The best of both worlds: Combining human and machine translations for multilingual semantic parsing with active learning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9511–9528, Toronto, Canada. Association for Computational Linguistics.

Liang, S., Shou, L., Pei, J., Gong, M., Zuo, W., Zuo, X., and Jiang, D. (2022). Label-aware multi-level contrastive learning for cross-lingual spoken language understanding. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9903–9918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Y. and Lapata, M. (2019). Hierarchical transformers for multi-document summarization. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Longpre, S., Lu, Y., and Daiber, J. (2021). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Mallinson, J., Sennrich, R., and Lapata, M. (2020). Zero-shot crosslingual sentence simplification. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4402–4412. PMLR.

Matthews, P. H. P. H. (2014). *The concise Oxford dictionary of linguistics / P.H. Matthews.* Oxford University Press, Oxford, 3rd ed. edition.

McCloskey, M. and Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165.

Mémoli, F. (2011). Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487.

Moghe, N., Razumovskaia, E., Guillou, L., Vulić, I., Korhonen, A., and Birch, A. (2023a). Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.

Moghe, N., Sherborne, T., Steedman, M., and Birch, A. (2023b). Extrinsic evaluation of machine translation metrics. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704.

Mooney, R. J. (2007). Learning for Semantic Parsing. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 311–324, Berlin, Heidelberg. Springer Berlin Heidelberg.

Moradshahi, M., Campagna, G., Semnani, S., Xu, S., and Lam, M. (2020). Localizing open-ontology QA semantic parsers in a day using machine translation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *ArXiv preprint*, abs/1803.02999.

Nicosia, M., Qu, Z., and Altun, Y. (2021). Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Packard, J. L. (2000). *The morphology of Chinese : a linguistic and cognitive approach*. Cambridge University Press, Cambridge.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Perez-Beltrachini, L., Jain, P., Monti, E., and Lapata, M. (2023). Semantic parsing for conversational question answering over knowledge graphs. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2507–2522, Dubrovnik, Croatia. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Philippy, F., Guo, S., and Haddadan, S. (2023). Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Qin, L., Chen, Q., Xie, T., Li, Q., Lou, J.-G., Che, W., and Kan, M.-Y. (2022). GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics.

Radford, A. (2009). *An Introduction to English Sentence Structure*. Cambridge University Press.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raman, K., Naim, I., Chen, J., Hashimoto, K., Yalasangi, K., and Srinivasan, K. (2022). Transforming sequence tagging into a Seq2Seq task. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11856–11874, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., and Lapata, M. (2016). Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

Reddy, S., Täckström, O., Petrov, S., Steedman, M., and Lapata, M. (2017). Universal semantic parsing. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Reizinger, P. and Huszár, F. (2023). SAMBA: Regularized autoencoders perform sharpness-aware minimization. In *Fifth Symposium on Advances in Approximate Bayesian Inference*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org.

Riley, P., Caswell, I., Freitag, M., and Grangier, D. (2020). Translationese as a language in "multilingual" NMT. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Rosenbaum, A., Soltan, S., Hamza, W., Damonte, M., Groves, I., and Saffari, A. (2022). CLASP: Few-shot cross-lingual data augmentation for semantic parsing. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 444–462, Online only. Association for Computational Linguistics.

Ruder, S., Clark, J. H., Gutkin, A., Kale, M., Ma, M., Nicosia, M., Rijhwani, S., Riley, P., Sarr, J.-M. A., Wang, X., Wieting, J., Gupta, N., Katanova, A., Kirov, C., Dickinson, D. L., Roark, B., Samanta, B., Tao, C., Adelani, D. I., Axelrod, V., Caswell, I., Cherry, C., Garrette, D., Ingle, R., Johnson, M., Panteleev, D., and Talukdar, P. (2023). Xtreme-up: A user-centric scarce-data benchmark for under-represented languages.

Schucher, N., Reddy, S., and de Vries, H. (2022). The power of prompt tuning for low-resource semantic parsing. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 148–156, Dublin, Ireland. Association for Computational Linguistics.

Shao, B., Gong, Y., Qi, W., Duan, N., and Lin, X. (2020). Multi-level alignment pretraining for multi-lingual semantic parsing. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3246–3256, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sherborne, T. and Lapata, M. (2022). Zero-shot cross-lingual semantic parsing. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Sherborne, T. and Lapata, M. (2023). Meta-learning a cross-lingual manifold for semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67.

Sherborne, T., Xu, Y., and Lapata, M. (2020). Bootstrapping a crosslingual semantic parser. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for*

Computational Linguistics: EMNLP 2020, pages 499–517, Online. Association for Computational Linguistics.

Shi, P., Zhang, R., Bai, H., and Lin, J. (2022). XRICL: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5248–5259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Si, C., Zhang, Z., Chen, Y., Qi, F., Wang, X., Liu, Z., Wang, Y., Liu, Q., and Sun, M. (2023). Sub-Character Tokenization for Chinese Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 11:469–487.

Singh, J., McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2019a). XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Singh, J., McCann, B., Socher, R., and Xiong, C. (2019b). BERT is not an interlingua and the bias of tokenization. In Cherry, C., Durrett, G., Foster, G., Haffari, R., Khadivi, S., Peng, N., Ren, X., and Swayamdipta, S., editors, *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Søgaard, A. (2022). Should we ban English NLP for a year? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Su, Y. and Yan, X. (2017). Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark.

Susanto, R. H. and Lu, W. (2017a). Neural architectures for multilingual semantic parsing. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 38–44, Vancouver, Canada. Association for Computational Linguistics.

Susanto, R. H. and Lu, W. (2017b). Neural Architectures for Multilingual Semantic Parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44, Stroudsburg, PA, USA.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Takatsu, A. (2011). Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Tanwar, E., Dutta, S., Borthakur, M., and Chakraborty, T. (2023). Multilingual LLMs are better cross-lingual in-context learners with alignment. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Thomason, J., Padmakumar, A., Sinapov, J., Walker, N., Jiang, Y., Yedidsion, H., Hart, J., Stone, P., and Mooney, R. (2020). Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., and Forcada, M. L., editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. (2018). Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

van der Goot, R., Sharaf, I., Imankulova, A., Üstün, A., Stepanović, M., Ramponi, A., Khairunnisa, S. O., Komachi, M., and Plank, B. (2021). From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Wang, B. (2021). *Generalization challenges in semantic parsing*. PhD thesis, School of Informatics, Edinburgh, UK.

Wang, B., Lapata, M., and Titov, I. (2021a). Meta-learning for domain generalization in semantic parsing. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pages 366–379, Online. Association for Computational Linguistics.

Wang, J., Lan, C., Liu, C., Ouyang, Y., and Qin, T. (2021b). Generalizing to unseen domains: A survey on domain generalization. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wang, P. Z. and Wang, W. Y. (2019). Riemannian normalizing flow on variational Wasserstein autoencoder for text modeling. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 284–294, Minneapolis, Minnesota. Association for Computational Linguistics.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Survey*, 53(3).

Wang, Z. and Hershcovich, D. (2023). On evaluating multilingual compositional generalization with translated datasets. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1687, Toronto, Canada. Association for Computational Linguistics.

Wei, X., Weng, R., Hu, Y., Xing, L., Yu, H., and Luo, W. (2021). On learning universal representations across languages. In *International Conference on Learning Representations*.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Wieting, J., Clark, J., Cohen, W., Neubig, G., and Berg-Kirkpatrick, T. (2023). Beyond contrastive learning: A variational generative model for multilingual retrieval. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12044–12066, Toronto, Canada. Association for Computational Linguistics.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wilks, Y. and Fass, D. (1992). The preference semantics family. *Computers & Mathematics with Applications*, 23(2):205–221.

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. AI-TR. M.I.T. Project MAC.

Wu, L., Guo, Z., Cui, B., Tang, H., and Lu, W. (2023). Good meta-tasks make a better cross-lingual meta-transfer learning for low-resource languages. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7431–7446, Singapore. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2020). Do explicit alignments robustly improve multilingual encoders? In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xia, M. and Monti, E. (2021). Multilingual neural semantic parsing for low-resourced languages. In Ku, L.-W., Nastase, V., and Vulić, I., editors, *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 185–194, Online. Association for Computational Linguistics.

Xia, M., Zheng, G., Mukherjee, S., Shokouhi, M., Neubig, G., and Awadallah, A. H. (2021). MetaXL: Meta representation transformation for low-resource cross-lingual learning. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Xu, W., Haider, B., Krone, J., and Mansour, S. (2021). Soft layer selection with meta-learning for zero-shot cross-lingual transfer. In Lee, H.-Y., Mohtarami, M., Li, S.-W., Jin, D., Korpusik, M., Dong, S., Vu, N. T., and Hakkani-Tur, D., editors, *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 11–18, Online. Association for Computational Linguistics.

Xu, W., Haider, B., and Mansour, S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yang, J., Fancellu, F., Webber, B., and Yang, D. (2021). Frustratingly simple but surprisingly strong: Using language-independent features for zero-shot cross-lingual semantic parsing. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5848–5856, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yin, P., Zhou, C., He, J., and Neubig, G. (2018). StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–765, Melbourne, Australia. Association for Computational Linguistics.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., and Radev, D. (2018a). Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., and Radev, D. (2018b). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Zelle, J. M. (1995). *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers*. PhD thesis, Department of Computer Sciences, The University of Texas at Austin, Austin, TX.

Zelle, J. M. and Mooney, R. J. (1996). Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 1050–1055.

Zettlemoyer, L. S. (2009). *Learning to Map Sentences to Logical Form*. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Zettlemoyer, L. S. and Collins, M. (2005). Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pages 658–666, Arlington, Virginia, United States.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Earth mover's distance minimization for unsupervised bilingual lexicon induction. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Zhao, M., Zhu, Y., Shareghi, E., Vulić, I., Reichart, R., Korhonen, A., and Schütze, H. (2021). A closer look at few-shot crosslingual transfer: The choice of shots matters. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

Zhao, Y., Chen, L., Chen, Z., and Yu, K. (2020). Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *The Thirty-Fourth*

AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 9668–9675. AAAI Press.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2SQL: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.