THE UNIVERSITY
of EDINBURGH

Universal Rewriting via Machine Translation

Jonathan Stephen Mallinson

Doctor of Philosophy

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

2021

Abstract

Natural language allows for the same meaning (semantics) to be expressed in multiple different ways, i.e. paraphrasing. This thesis examines automatic approaches for paraphrasing, focusing on three paraphrasing subtasks: unconstrained paraphrasing where there are no constraints on the output, simplification, where the output must be simpler than the input, and text compression where the output must be shorter than the input.

Whilst we can learn paraphrasing from supervised data, this data is sparse and expensive to create. This thesis is concerned with the use of transfer learning to improve paraphrasing when there is no supervised data. In particular, we address the following question: can transfer learning be used to overcome a lack of paraphrasing data? To answer this question we split it into three subquestions (1) No supervised data exists for a specific paraphrasing task; can bilingual data be used as a source of training data for paraphrasing? (2) Supervised paraphrasing data exists in one language but not in another; can bilingual data be used to transfer paraphrasing training data from one language to another? (3) Can the output of encoder-decoder paraphrasing models be controlled?

We address question 1 by developing, in Chapter 3, a Neural Machine Translation (NMT) pivoting approach, which uses two pre-trained NMT models to perform paraphrasing with no paraphrasing data. A source sentence is translated into multiple foreign pivots, these multiple pivots are then simultaneously translated back into the original language, producing a paraphrase. Chapter 4 extends this approach and addresses questions 1 and 3, where we train a sentence compression, with no sentence compression data. Instead, we train NMT models using variable disentanglement to separate the semantics of the sentence from the length of the output sentence in a controllable manner. In this way, a user can specify the length of the translation, and when combined with the pivoting technique a user can set the output length as shorter than the original length, creating a compression of the original sentence.

In Chapter 5 we further explore question 3, addressing the problem of the sparsity of simplification data, and the bespoke needs of simplification users. We develop a variable disentanglement approach, which separates the semantics of the source sentences, and the lexical and syntactic structure of the output simplification. A user is then able to control the simplification to produce a simplification that best suits their needs. Finally, in chapter 6 we answer question 2 where there exists simplification data in a high resource language but not in low resource languages. We develop a model

that uses task-specific transformer layers and a shared encoder which was trained using multi-task learning to both translate and simplify. By sharing encoding layers the model is able to transfer simplification data from one language to another.

Lay Summary

Natural language allows for the same meaning (semantics) to be expressed in multiple different ways, i.e. paraphrasing. This thesis examines automatic approaches for paraphrasing, focusing on three paraphrasing tasks: (1) paraphrasing, where the rewritten sentence should mean the same as the input sentence, but should look dissimilar, (2) sentence compression where the output sentence should capture the meaning of the input sentence but be shorter, and (3) sentence simplification where the output sentence should mean the same as the input sentence, but be easier to understand.

Automatic approaches to paraphrasing often use large amounts of example sentences; given the input what should the output be? However, example sentences aren't always available, and if they do available, they often only exist in English, meaning that automatic paraphrasing models can not be trained in other languages. In this thesis we explore alternative approaches for automatic paraphrasing models, which do not use paraphrasing examples. Instead we propose using transfer learning, a family of techniques that adapts example sentences from related tasks, where large amounts of example sentences do exist. We focus on transferring translation data, where the input sentence is in one language and the output sentence, which means the same, is in another language. We focus on translation data because it exists in large quantities across many languages and show how it can be used for automatic paraphrasing in multiple different languages.

Acknowledgements

I would like to thank my PhD supervisors, Mirella Lapata and Rico Sennrich, for their feedback and guidance which helped improve the content of my PhD work. I would also like to thank all my collaborators: Li Dong, John Weiting, Siva Reddy, and Kevin Gimpel as well as my internship hosts, from whom I learnt so much: Guillermo Garrido, Aliaksei Severyn, and Mohit Iyyer, and all my fellow interns! I also thank my examiners, Alexandra Birch and Chris Callison-Burch for taking their valuable time to assess this thesis. Their constructive feedback have helped strengthen it.

I would also like to thank my CDT cohort[α], my office[β], the lunch time study group[γ], friends that I distracted when I wanted a break from work[δ], my housemates[ε], and my thesis writing groups[ζ], who made my time as a PhD student amazing. Alexander Robertson[β], Anna Page[ε], Arthur Bražinskas[β], Artur Bekasov[δ], Ben Rozemberczki[δ], Borislav Ikonomov[α,γ], Carl Allen[δ,ε,ε], Charlie Nash[γ], Clara Vania[δ], Connie Crowe[ε], Deena Bardsley[ε], Ed Fincham[ε], Ida Szubert[β], Ivana Balažević[δ,ε], Jack Baker[ε], James Owers[α,β,γ,δ,ε,ζ], John[δ], Jozef Mokry[α,η], Kaitlyn Hawley[ε], Kate McCurdy[β], Magdalena Navarro[δ], Marco Damonte[β], Maša Močnik[ζ], Matt Chapman-Rounds[α,β,γ,ζ], Mattias Appelgren[α,γ,δ,ε], Naomi Saphra[γ,δ], Ryan Davies[α,γ,δ], Sabine Weber[δ], Sameer Bansal[δ], Simão Eduardo[α,γ,δ], Sorcha Gilroy[δ,ε], Soňa Mokrá[α,β], Spandana Gella[δ], Zack Hodari[α,γ,δ,ε].

Finally, I thank my sister, my mum, my dad, and my grandparents, for always being there for me. Thank you so much for your never-ending support throughout my entire PhD journey.

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.

Table of Contents

Chapter 1

Introduction

Natural language allows for the same meaning (semantics) to be expressed in multiple different ways, i.e. paraphrasing. Paraphrasing is an important task within NLP/NLG, it can be used to increase the robustness of existing NLP/NLG systems, allowing them to focus on the semantics of the language, rather than the surface form, the multiple different ways in which semantics can be expressed. Constrained forms of paraphrasing, where there are requirements on the surface form, have even more applications and include tasks such as sentence compression, simplification, grammatical error correction, and style transfer. These tasks can all improve accessibility for users, allowing users to have their grammatical mistakes automatically corrected, making text easier to read, or provide short summaries.

Automatic approaches for learning to paraphrase often require large amounts of supervised training data, this data is sparse, not existing in large quantities for all paraphrasing tasks, or does not exist in every language and the data can be expensive to create, often requiring expert annotators. In this thesis we use transfer learning to overcome this lack of supervised paraphrasing training data. In particular, we address the following question:

- Can transfer learning be used to overcome a lack of paraphrasing data?

In the remainder of the chapter, we will further introduce paraphrasing, its challenges, our solutions, and the central question of the thesis in more detail.

1.1 Paraphrasing

Paraphrasing can be broadly described as the task of using an alternative surface form to express the same semantic content (Madnani and Dorr, 2010). There are many pos-

sible ways to construct a paraphrase; individual words which have the same semantic content are often called lexical paraphrases or synonyms. For instance the words *excited* and *thrilled* mean the same things. Phrasal paraphrases on the other hand refer to small fragments of text, a few words, which have the same meaning. *Help out* and *lend a hand* are phrasal paraphrases of each other. Finally, sentential paraphrases are entire sentences which convey the same meaning, such as *The earthquake shook the city* and *The metropolis trembled due to the quake*. Whilst sentential paraphrases can be formed by replacing words or phrases within a sentence, this would limit the possible output surface forms. Within this thesis, we focus entirely on sentential paraphrasing, which allows for a wide range of lexical and syntactic transformations.

Paraphrasing can be refined into subtasks by adding constraints to the output surface form. For instance, by constraining the length of the output sentences, we can use paraphrasing to perform sentence compression. If, however, we add the constraint that the surface form is easily understandable, then the paraphrase is a simplification. Table 1.1 highlights the wide range of paraphrasing tasks, further motivating their study. Within this thesis, we focus on three types of paraphrasing: unconstrained paraphrasing (which we will often refer to as paraphrasing), sentence simplification, and sentence compression.

Unconstrained Paraphrasing generates an output paraphrase where the surface form differs from the source sentence, where there are no additional constraints. Table 1.2 shows several different paraphrasing examples, produced by phrasal paraphrasing and sentential paraphrasing. The examples highlight how the same source sentences can have multiple outputs, an additional challenge which is discussed in Section 1.4. While automatic approaches for unconstrained paraphrasing have been used directly by end-users, for example being used as a writing prompt tools (He et al., 2013), the main appeal stems from its application to a wide range of NLP problems, by making existing NLP models more robust, allowing models to focus on semantics rather than the surface form. There are two common ways to integrate paraphrasing; either through training data augmentation, or test data augmentation. In training data augmentation the training data is extended with paraphrases. For instance Wang et al. (2015) used paraphrasing to overcome a lack of semantic parsing training data. Using a semantic grammar they generated logical forms paired with canonical utterances. These utterances were then paraphrased to produce more realistic and varied sentences. A semantic parser was then trained from the paraphrases to the logical form.

Task	Description
Sentence Splitting	Single sentences are split into multiple sentences.
Input:	*John is an astronaut who went to the moon.*
Output:	*John is an astronaut. He went to the moon.*
Sentence Fusion	Multiple sentences are merged into a single sentence.
Input:	*John is an astronaut. He went to the moon.*
Output:	*John is an astronaut that went to the moon.*
Unconstrained Paraphrasing	The output sentence has a different surface form.
Input:	*'Cause everybody hates a tourist.*
Output:	*Nobody likes tourists.*
Sentence Compression	The output sentence uses fewer words than the input.
Input:	*He never gave up hope, he kept writing.*
Output:	*He kept writing.*
Style Transfer	The surface form is changed to match a particular style, e.g., politeness.
Input:	*Send me the data!!!*
Output:	*Could you please send me the data?*
Post-Editing	Translation mistakes found in the input sentence are fixed.
Input	*Allons tous à la plage.* $\rightarrow$ *Let's just all retire to a beach...*
Output:	*Let's all go to the beach.*
Grammatical Error Correction	Grammatical mistakes found in the input text are fixed.
Input:	*Run to the park is an good form of exercising.*
Output:	*Running in the park is a good form of exercise.*

Table 1.1: An overview of the different types of constrained paraphrasing tasks.

Test data augmentation provides, as input to the model, the source sentence and multiple paraphrases. Multiple paraphrases can be used to increase the confidence of the answer, for instance by voting on the correct output.

Source	There are a thousand ways to say I love you.
Phrasal	There are a **lot of** ways to say I **adore** you.
Sentential	I love you **can be said in** a thousand **different** ways.
Sentential	A **person can express the phrase** "I love you" **using many different surface forms**.
Source	The pigs ate the pie.
Phrasal	The pigs **devoured** the pie.
Sentential	The pie **was eaten** by the pigs.
Sentential	**There were** pigs**, and a** pie**, which was eaten by them**.

Table 1.2: Examples of paraphrasing at different linguistic levels. We see that there are many different paraphrases for the same source sentence. **Bold** indicates additional words not found in the source sentence.

Source	I took a couple of steps towards it, but the currents kept pushing the boat further and further away.
Extractive	I took steps towards it, but the currents kept pushing the boat away.
Abstractive	The boat kept being pushed away.
Abstractive	The boat kept being pushed further and further away.
Source	He repeatedly dived under the water, frantically searching for me.
Extractive	He repeatedly dived under the water, searching for me.
Abstractive	He kept searching for me.
Abstractive	He repeatedly dived under the water, to find me.

Table 1.3: Examples of abstractive and extractive sentence compressions.

Sentence Compression constrains the output paraphrase to be shorter than the input, producing a summary that retains the most important information while preserving its fluency. This task has attracted much attention due to its potential applications such as text summarization, (Jing, 2000; Madnani et al., 2007b; Woodsend and Lapata, 2010; Berg-Kirkpatrick et al., 2011; Chali et al., 2017), and displaying text on small-screens (Corston-Oliver, 2001). Approaches for generating compression can be divided into two categories (Mani, 2001): (1) Extractive compression, words and phrases are deleted, but no new words can be added and (2) Abstractive compression,

with no restrictions on how compressions can be generated, other than the output is shorter than the input. Table 1.3 provides examples of both extractive and abstractive sentence compressions, as well as highlighting how there are multiple possible compressions per source sentence, each of which provides different trade-offs between retaining meaning and producing a shorter output. In addition to extractive approaches being less flexible in the type of surface forms they can produce, it is not clear how they would be applicable for other paraphrasing subtasks, such as sentence fusion. As we wish our proposed solutions to be applicable to all paraphrasing tasks, we focus exclusively on abstractive compression.

Source	Parkes was a key location for the railway, serving as a hub for a great deal of passengers.
Simplification	Parkes was an important town for the railway, being used as a hub for many passengers.
Simplification	Parkes was a key location for the railway. It served as a hub for a great deal of passengers.
Source	Shakespeare has always been celebrated, his works have been praised by theatregoers and readers.
Simplification	Shakespeare's works have always been celebrated by theatregoers and readers.
Simplification	Everyone likes Shakespeare's plays.

Table 1.4: Examples of multiple simplifications for the same source sentence.

Sentence Simplification constrains the output to make sentences easier to read and understand whilst retaining most of their meaning. There are several groups of people who have low literacy skills and could benefit from simplifications, including children, second language learners, and people with autism, dyslexia and aphasia (Rello et al., 2013; Shewan and Canter, 1971). Sentence simplification is a varied process; it can include replacing complex words and phrases with a simpler paraphrase or simplifying the syntax, for instance splitting long sentences into multiple shorter sentences. Examples of different types of simplification can be found in Table 1.4, again we highlight how there are many possible simplifications per source sentence. Sentence simplification is not only applicable to end-users but can also be incorporated into other NLP systems, acting as a preprocessing step to make sentences easier for NLP systems, and has been used within: parsers (Chandrasekar et al., 1996), summarizers (Klebanov et al., 2004), semantic role labelers (Vickrey and Koller, 2008; Woodsend and Lapata, 2014), and machine translation (Mehta et al., 2020).

1.2 Automatic Approaches

Automatic approaches often treat paraphrasing as a text-to-text problem, with a source input sentence and a corresponding output target sentence. To solve this problem machine translation frameworks - either Statistical Machine Translation (SMT) or, more recently, encoder-decoder (Sequence-to-Sequence) models - are often used to learn these. These approaches train on supervised paraphrasing datasets, trying to learn a generalizable mapping between source and target sentences, allowing them to generalize to unseen sentences.

SMT, in an effort to reduce sparsity, decomposes the problem. Where instead of mapping between entire sentences, SMT maps between sub-parts of the source and target sentences. There are three commonly used decompositions: lexical, phrasal, and syntactic (Och and Ney, 2000). Lexical approaches learn a mapping between words in the source and target sentence, producing a probabilistic word translation table. To generate a target sentence, source words are individually translated, and a language model is used to ensure the output is fluent. Words, however, may not always be the best choice for the decomposition. Sometimes one source word can map into multiple target words or vice versa. Phrase-based models learn a probabilistic mapping, a phrase table, between small sequences of words. Syntax-based approaches differ, as they learn a joint syntactic grammar, consisting of syntactic fragments, applied to both the target and source side (Och and Ney, 2000).

Neural approaches do not decompose the problem; instead, they rely on the ability of a neural network to learn continuous features of the entire source sentence without needing preprocessing tools or syntactic information (e.g., part-of-speech tags, parse trees) to reduce sparsity. A common approach for neural models is encoder-decoder with attention, as popularised by Bahdanau et al. (2015). There are three major architectures, Recurrent Neural Networks (RNN) (Bahdanau et al., 2015; Sutskever et al., 2014), Convolutional Neural Networks (CNN) (Kalchbrenner et al., 2016; Gehring et al., 2017), and Transformers (Vaswani et al., 2017). For all approaches a network first *encodes* the source sentence into a sequence of latent representations, using the encoder. The decoder then *decodes* the entire target sentence word-by-word, *attending* to these latent representations.

SMT has previously been shown to be able paraphrase (Napoles-Cohen, 2019; Ganitkevic et al., 2018), However, within this thesis we exclusively use encoder-decoder models, they have has been shown to outperform SMT on a wide range of tasks and

datasets, including machine translation (Barrault et al., 2019), post-editing (Chatterjee et al., 2018), grammatical error correction[1], simplification[2], and summarization[3].

1.3 Challenges

Whilst text-to-text frameworks are a powerful approach for paraphrasing they require large amounts of supervised data. Neural approaches are particularly data-hungry, typically training on several hundred-thousand to several million sentence pairs. Training on fewer datapoints can lead to a significant decrease in performance, as demonstrated for sentence fusion (Malmi et al., 2019). This is problematic as paraphrasing data is often not available or not available in sufficient quantities. Within this thesis, we address this data shortage, focusing on two particular aspects.

Lack of data As each supervised paraphrasing dataset is language-specific, every language needs its own dataset. This results in the need for a huge amount of parallel data across many languages. While datasets can sometimes be automatically gathered by scraping the web, this requires language-specific heuristics to be developed for each language, as well as a website (or an equivalent) existing in each language. The alternative is to create the dataset using human annotators. This process is costly as annotation is expensive, of variable quality, requires an adequate annotator pool for each language, and may require the annotators to be in-domain experts.

Personalization Further exacerbating the issue over a lack of data, is the lack of the *right data*. As the examples demonstrate in Tables 1.2, 1.3, and 1.4, each source sentence can have multiple possible outputs, where the correct/preferred output is determined by the users. For example, for sentence compression, what is the desired length of the compression? What is acceptable to delete to produce a short output? The answers to these questions depend on what the users wish to do with the summary. For instance, if there is limited screen space, then the summary must fit within this limit. For simplification, what aspect is considered complex, syntactic or lexical? Which words does the user not understand? Which syntactic construction does the user not understand? The answers to all these questions are user-specific. Currently, the output surface form is determined solely by the data the model was trained on. This

[1] https://nlpprogress.com/english/grammatical_error_correction.html
[2] https://nlpprogress.com/english/simplification.html
[3] https://nlpprogress.com/english/summarization.html

is problematic if the data was scraped, as it may not match the requirements of the user. In the case that the data was created, this still requires that each user or subset of users creates their own bespoke training data, an expensive proposition.

1.4 Thesis Proposal

Within this thesis we propose using transfer learning and bilingual data to overcome the *lack of paraphrasing data*. Transfer learning is a family of approaches that can adapt training data from one task and apply it to a different task (Pan and Yang, 2009). Bilingual data is a good source for paraphrasing, as translation is a similar task to paraphrasing, in that the semantics are preserved between the input sentence and the output translation. Using bilingual data also comes with many advantages: (1) Bilingual datasets are parallel consisting of an input and output sentence, allowing them to be easily used within an encoder-decoder frameworks. (2) Bilingual datasets are large and exist in many languages. For instance, the open parallel corpus, provides over 2.5 billion bilingual sentence pairs, covering over 100 languages, and multiple different domains Tiedemann (2012). (3) Bilingual datasets are continuously growing, either through extracting translations from the web or through the efforts of translators producing additional parallel text (Koehn, 2005; Lison and Tiedemann, 2016). To address the *personalization* challenge we propose to use transfer learning to adapt datasets that exist already for paraphrasing tasks that have no data, where *no data* can be a complete lack of data or not personalised data as discussed previously.

In specifics this thesis addresses the following research question:

Can transfer learning be used to overcome a lack of paraphrasing data?

We further break this question down into three smaller subquestions:

1. No supervised data exists for a specific paraphrasing task. Can bilingual data be used as a source of training data for paraphrasing?

2. Supervised paraphrasing data exists in one language but not in another. Can bilingual data can be used to transfer paraphrasing training data from one language to another?

3. Can the output of encoder-decoder paraphrasing models be controlled?

The first question directly tackles the problem of a lack of supervised paraphrasing data. The second question assumes there exists some data in one language but not another. This is a common scenario, as often paraphrasing training data exists in English but not in other languages. By transferring data across languages we are able to take advantage of existing datasets. The third question relates to both the lack of data and lack of the right data. We show that by adding a control mechanism to a paraphrasing model we can perform other constrained paraphrasing tasks, for which we have no data.

To answer these questions, we extend three existing transfer learning techniques: pivoting (Ganitkevitch et al., 2013), cross-lingual learning, and variable disentanglement (Higgins et al., 2016). We first show how pivoting can be used to perform unconstrained paraphrasing with bilingual data, answering the first question. Secondly, we show how variable disentanglement can be used to control the output of encoder-decoder models, allowing users to produce personalized paraphrases, answering the third question. By combining pivoting and variable disentanglement, bilingual data can be used to perform specific paraphrasing sub-tasks, allowing users to produce personlized paraphrases, thereby also answering the third question. We then show how cross-lingual learning can be used to transfer paraphrasing data from a high resource language to a low resource language, thereby answering the second question. In Chapter 2 we discuss alternative approaches to dealing with a lack of data, including unsupervised learning, pre-training, and domain adapation.

Pivoting is a technique to overcome the lack of training data, by combining existing datasets through a common element to create a new dataset. Pivoting has been previously used for paraphrasing, combining two bilingual datasets to form a monolingual dataset (Bannard and Callison-Burch, 2005; Ganitkevitch et al., 2011; Madnani et al., 2007a; Callison-Burch, 2008; Pavlick et al., 2015). The intuition behind bilingual pivoting is that two English phrases e_1 and e_2 that translate to the same foreign phrase f can be assumed to have the same meaning.

Previous approaches used SMT phrase tables to model translation probabilities (Ganitkevitch et al., 2013). Since phrase tables are finite, it is possible to pivot over all possible, f, foreign phrases. We propose an approach which works with encoder-decoder models, allowing us to take advantage of the improvements that encoder-decoder models offers over SMT. In our framework, a source sentence is translated into a foreign sentence, before being translated back into the original language. In this

case, the possible pivots are comprised of all possible sentences, which is intractable to pivot over. Within Chapter 3 we explore multiple approaches to making this tractable.

Variable Disentanglement is an approach to control some variable of interest while leaving the remaining attributes unchanged (Higgins et al., 2016). In this thesis we use variable disentanglement to control the output surface form; for instance the length, lexical or syntax surface form of the generated paraphrases. Previous work has focused on how to disentangle the variable of interest from the rest of the representation. A common approach is to use an auto-encoder to encode a sentence into a vector. A discriminator is then trained to predict the variable of interest from this vector (John et al., 2018; Hu et al., 2017; Shen et al., 2017). The encoder must encode the semantics of the input while perturbing the discriminator, ensuring that the discriminator is not able to predict the variable of interest. The decoder is then conditioned on the encoded sentences and a representation of the variable of interest.

An alternative but related set of approaches, which we call *soft variable disentanglement*, does not focus on trying to disentangle the semantics from the feature of interest but instead provides the feature of interest as an input, often as side information. For instance Michel and Neubig (2018); Sennrich et al. (2016a) experimented with adding tags to the source/target sentence to indicate the gender of the speaker or the politeness level within machine translation systems. As such the model is trained using a triplet consisting of the source sentence x, the target y and the feature of interest $F(y)$. A simple formulation of variable disentanglement is that the model then learns to generate y from x and $F(y)$, as such:

$$P(y|\text{ENC}(x), \text{ENC}(F(y))) \tag{1.1}$$

where ENC is an encoding mechanism such as a feedfoward neural network or an RNN. As the model is given gold information about the output in $\text{ENC}(F(y))$, there is no need for this information to be encoded within $\text{ENC}(x)$, and the parameter budget would be better spent encoding the semantics. As such the model is encouraged to produce a disentangled representation. At test time the user can provide their preferred values for $F(y)$.

We apply variable disentanglement to models trained with general-purpose paraphrasing datasets allowing them to become user-centric, with no additional training. We show, in Chapters 4 and 5, how this approach can allow users to control the length, syntax and lexical items of the output. By combining variable disentanglement and

pivoting we can train models on bilingual data and use them for specific constrained paraphrasing tasks.

Cross-lingual learning attempts to embed multiple languages into the same space. This allows data points between languages to be compared and by providing a shared space, information can transferred between languages, e.g., between high and low resource languages. Prior work has focused on what to embed (Ruder et al., 2019): words, sentences, or documents as well as what level of supervision is used. We focus on supervised sentential embeddings, and transferring data between languages. Firat et al. (2016b); Johnson et al. (2017); Ha et al. (2016) have shown that translation data can be transferred across languages, while the model is trained on translation $x \leftrightarrow y$ and $x \leftrightarrow z$, (where x, y, z are different languages), it is able to translate $z \leftrightarrow y$, with no additional training data. We focus on transferring supervised paraphrasing data from a resource rich language x to a resource poor language y. In this case we have $x \leftrightarrow y$ and $x \rightarrow x'$, where $x \rightarrow x'$ is the supervised paraphrasing training data. We wish to learn paraphrasing for resource poor language $y \rightarrow y'$, noticing that, unlike the translation case, y' does not appear in the training data.

Whilst there has been limited research on transferring generation data across languages, there has been work on transferring classification data across languages (Zhou et al., 2016; Chen et al., 2019). Zhou et al. (2016) transferred supervised sentiment labels from English to sentences in Chinese using a dual objective, that embeds Chinese and English into the same space, and learns sentiment labels on the English data. We propose a modeling framework which transfers supervised paraphrasing data from English to another language (for which no supervised data exists). We train the model on paraphrasing and translation data, using a shared transformer encoder, which constructs language-agnostic and task-agnostic representations, with a combination of task-specific encoder layers added on top (e.g., for translation and paraphrasing). A language-specific Transformer is then used to decode the sentence, which is described in Chapter 5.

1.5 Thesis Structure

This thesis consists of six further chapters. The next chapter contains relevant background information, including information on encoder-decoder models, evaluation metrics, and alternative approaches for overcoming the lack of training data. The following

	Challenges			Approaches		
	Unsupervised	Transfer	Control	Pivoting	Disentanglement	Cross-lingual
Chapter 3	✓	✗	✗	✓	✗	✗
Chapter 4	✓	✗	✗	✓	✓	✗
Chapter 5	✗	✗	✓	✗	✓	✗
Chapter 6	✗	✓	✗	✗	✓	✓

Table 1.5: Outline of thesis, where each Chapter addresses a particular challenge using a particular technique including: *Pivoting*, variable *disentanglement*, and *Cross-lingual* learning. *Unsupervised*, means no paraphrasing training data is used, *transfer* means training data is provided in one language but not another, and *control*, the output is controllable by the user.

four chapters address the challenges and approaches outlined in Section 1.5, and the last chapter presents our conclusions and directions for future work.

Papers published during my PhD but not discussed within the thesis include Dong et al. (2017), Wieting et al. (2017), and Puduppully et al. (2019b). Within Dong et al. (2017) we proposed a model which uses paraphrasing for question-answering. The model receives as input questions and paraphrases of these questions. A neural scoring model then scores the paraphrases, assigning a higher score to those paraphrases which are most likely to produce the correct answer when used as input to the answering model. The paraphrases and their scores were then used as input to the answering model to produce the answer. The neural scoring model and answering model were trained end-to-end.

In Wieting et al. (2017) we used neural machine translation to generate sentential paraphrases via back-translation of bilingual sentence pairs. We found that the data quality of these paraphrases were stronger than prior work based on bitext, and on par with manually-written English paraphrase pairs. A further discussion of this work can be found in Chapter 3 (Section 3.2.1).

Puduppully et al. (2019b) was the University of Edinburgh's submission to the Document-level Generation and Translation Shared Task. This shared task required models to generate a summary of basketball games, and consisted of six tracks. In the first and second track an English or German summary was produced from structured data representing the basketball game. The third and fourth track consisted of translating a summary of the basketball game between English and German. The fifth and

sixth provided structured data, and a summary in English or German. The model generated the summary in the other language. For the first and second track we trained a multilingual German/English model using the content selection and planning approach of Puduppully et al. (2019a) with an added language tag prepended to the decoders' output to indicate the desired output language. For the third, fourth, fifth, and sixth track, we trained a transformer encoder-decoder translation model on WMT19 parallel data. Additionally, we created in-domain monolingual data by extracting basketball-related texts from monolingual sources. These sentences were then back-translated and added to the translation data.

Chapter 3: Paraphrasing with Neural Pivoting

This chapter demonstrates how neural pivoting can be used to overcome the lack of supervised unconstrained paraphrasing data (see Table 1.5). We introduce bilingual pivoting in the context of Neural Machine Translation and present a paraphrasing model based purely on neural networks. We propose and evaluate multiple variants of neural pivoting. Experimental results across tasks and datasets show that neural pivoting outperforms those obtained with conventional statistical machine translation based pivoting approaches.

Chapter 4: Sentence Compression with Neural Pivoting

Within this chapter, we show how neural pivoting and variable disentanglement can be used to generate sentence compressions from translation data (see Table 1.5). Compressions are obtained by translating a source string into a foreign language and then translating it back into the source while controlling the translation length. Using variable disentanglement we train translation models to separate the representation of the source sentence and the length of the target sentence. In this way, the user can control the length of the translation at test time. We release three cross-lingual sentence compression datasets in English, German and Czech. Experimental results on these datasets show that pivoting with variable disentanglement is an effective way of producing compressions.

Chapter 5: Controllable Simplification

Within this chapter we assume there exists general-purpose simplification data; however, this data is not tailored for specific users needs. We show how variable disentanglement can be used to tailor the output of a model trained on this data. We train models which allow users to control both the syntax and the lexical items which appear within the output (see Table 1.5). Empirical results show that this is an effective way of controlling the output simplicity level as well as producing good general-purpose simplifications.

Chapter 6: Zero-Shot Crosslingual Sentence Simplification

Within this Chapter, we demonstrate a cross-lingual approach which allows us to transfer supervised simplification data from one language to another language, where no such data exists (see Table 1.5). For this chapter, we assume simplification data exists in English but not in German. We propose a cross-lingual Transformer encoder-decoder model trained on bilingual and simplification data. To evaluate the performance of our model we construct a German sentence simplification evaluation set. Empirical results on our dataset and others, using both human and automatic metrics, show that our approach produces better simplifications than unsupervised and pivot-based methods.

The content of Chapter 3 was published in Mallinson et al. (2017), this work builds upon the work of my master's dissertation[4] Mallinson (2016), which introduced pivoting with neural machine translation. This thesis expands upon this, by refining the model, adding additional tasks, running new experiments, providing additional analysis, and rewriting the text. The contents of Chapter 4 was published in Mallinson et al. (2018), Chapter 5 in Mallinson and Lapata (2019) and Chapter 6 in Mallinson et al. (2020a).

1.6 Summary

This chapter has introduced and motivated the transfer learning for paraphrasing. We highlighted how a lack of data is problematic for paraphrasing models and proposed a general-purpose transfer learning solution to overcome this lack of data. We discussed

[4]Which was part of the PhD program.

four major contributions of this thesis: (1) Using bilingual data with encoder-decoder models to perform unconstrained paraphrasing (2) Introducing user control to paraphrase systems, allowing for personalized outputs. (3) By combining user control with pivoting, we show how bilingual data can be used to perform specific paraphrasing subtasks. (4) Introducing an approach that uses bilingual data to transfer supervised paraphrasing training data between languages. In the next chapter we introduce background information for the rest of the thesis, as well as discussing alternative approaches to transfer learning.

Chapter 2

Background

This chapter provides the background information needed for the rest of the thesis. We include a detailed description of neural encoder-decoder models, focusing on recurrent based models and Transformers. We describe the architecture, the training, and inference time decoding. We provide details on the automatic evaluation metrics used, describing the implementation and what they measure.

We also discuss the alternatives to transfer learning for overcoming a lack of paraphrasing data, including domain adaptation (Chu and Wang, 2018), pre-training (Devlin et al., 2019), and unsupervised learning (Artetxe et al., 2018; Lample et al., 2018). For related work on particular paraphrasing tasks, simplification, summarization, or unconstrained paraphrasing we refer the reader to the chapters following which tackle these tasks.

2.1 Encoder-decoder Models

As mentioned in Chapter 1, modern approaches (Zhang and Lapata, 2017; Nishihara et al., 2019) view paraphrasing as monolingual text-to-text rewriting, and employ the very successful neural encoder-decoder architecture (Bahdanau et al., 2015; Sutskever et al., 2014). In contrast to traditional methods (Khosmood, 2012; Bhagat and Hovy, 2013; Kozlowski et al., 2003; Dras, 1999), which target individual aspects of paraphrasing, such as passive to active voice, or lexical replacement, neural models have no special-purpose mechanisms for ensuring how to best paraphrase text. They rely on representation learning to implicitly capture paraphrase rewrites from data, i.e. examples of paraphrase sentence pairs.

The two most popular encoder-decoder architectures are Recurrent Neural Net-

works (RNN) with attention (Bahdanau et al., 2015; Sutskever et al., 2014), and Transformers (Vaswani et al., 2017). Both architectures are trained on sets of input sentences $x = (x_1, ...x_{|x|})$ and corresponding output sentences $y = (y_1, ..., y_{|y|})$, from which the models predict y given x, $P(y|x)$, and decompose the problem into:

$$P(y|x) = \prod_{j}^{|y|} P(y_j|y_{<j}, x) \tag{2.1}$$

The model predicts the output sentence one token at a time, j, conditioning on the source sentence and the previously generated words. Both architectures break the model down into two parts, the *encoder*, which produces representations of the source sentence and the *decoder* which sequentially generates the output, forming a new representation at each time step. While both RNN and Transformers follow this pattern, they differ in the way that they generate these representations.

2.1.1 Recurrent Neural Networks

Encoder RNN encoders sequentially encode source words one word at a time, x_i, into hidden state (h_i), by combining word embeddings, vector representation of a word, $e(x_i)$ and the previous hidden state:

$$h_i = \text{RNN}(x_i, h_{i-1}) = f(W_e e(x_i) + W_h h_{i-1}) \tag{2.2}$$

where f is any non-linear function, such as tanh or sigmoid, W is a matrix, and $W_e e(x_i)$ provides an embedding for the token x_i. The source sentence is then represented as a set of hidden states, $c = (h_1, ...h_{|x|})$. In practice, bidirectional RNN encoders are often used, where one RNN encodes the sentence left-to-right ($\overrightarrow{h}$), and another independent RNN encodes the sentence right-to-left ($\overleftarrow{h}$). These hidden representations are then concatenated together to form the representations of the sentence:

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h}_{|x|-i}] \tag{2.3}$$

Decoder A conditional RNN decoder is then used to generate the output sequence a word at a time. A decoder is initialised with the output of the encoder, which tries to represent the meaning of the sentence. However, encoding the meaning of an entire sentence into a single vector is an open problem (Conneau et al., 2018). "You can't cram the meaning of a whole **** sentence into a single **** vector!" - Raymond J. Mooney. Instead attention mechanisms have been introduced (Bahdanau et al., 2015;

Sutskever et al., 2014). In attention-based encoder-decoder models, the model assumes that at different steps of the generation the decoder should focus on different parts of the input. How much each part of the input should be focused on is determined by an attention mechanism (Bahdanau et al., 2015), which consists of three components: keys, query and values. In RNN encoder-decoder models, the encoder hidden states (h) are the keys, which are then queried by the decoder hidden state (s_j). Each key-query pair returns a score: $score(h_i, s_j)$. These scores are then normalised across all the keys using the softmax function:

$$\text{ATT}_{i,j} = \frac{\exp(\text{score}(i,j)}{\sum_{i'} \exp(\text{score}(i',j))} \tag{2.4}$$

The normalized scores are then combined with the values, in this case the source hidden state, to form a context vector at each time step of the decoder:

$$c_j = \sum_i \text{ATT}_{i,j} * h_i \tag{2.5}$$

Thus, attention produces a weighted average of the source sentence. There are many possible scoring functions that could be used, however popular choices include: dot product (Luong et al., 2015a):

$$\text{score}(i,j) = s_j^T h_i \tag{2.6}$$

the scaled dot product, where n is the size of the key vector (Vaswani et al., 2017):

$$\text{score}(i,j) = \frac{s_j^T h_i}{\sqrt{n}} \tag{2.7}$$

and additive (Bahdanau et al., 2015):

$$\text{score}(i,j) = v^\top \tanh(W[s_j; h_i]) \tag{2.8}$$

With attention defined, we can now define the decoder, with a focus on the RNN decoder architecture of Sennrich et al. (2017). An RNN decoder uses three inputs to update its hidden state s_j: the previous hidden state s_{j-1}, the source hidden states c and the previously predicted word y_{j-1}:

$$s_j = \text{RNN}(s_{j-1}, y_{j-1}, c) \tag{2.9}$$

The model performs three steps. First the *look step*, which produces an intermediate representation s'_j, using an RNN to combine the previous decoder hidden state and the previously predicted word:

$$s'_j = \text{RNN}_1(s_{j-1}, y_{j-1}) \tag{2.10}$$

The attention mechanism, ATT, is then used to compute the context vector c_j, where the encoder states c are they keys and values, and the query is s'_j.

The *update step* next generates the hidden representation s_j using another RNN. The inputs to this RNN are the intermediate representation s'_j and the context vector c_j:

$$s_j = \text{RNN}_2(s'_j, c_j) \tag{2.11}$$

Note that the two RNN blocks (RNN_1, RNN_2) are not individually recurrent, however the input to RNN_2 is the output of RNN_1, and the input of RNN_1 is the output of RNN_2, making the decoder recurrent overall.

Finally, the model performs the *generate step*; a softmax layer is applied to s_j to produce a distribution over the vocabulary:

$$P(y|x) = \prod_{j=1}^{|y|} p(y_j|s_j) \tag{2.12}$$

It should be noted that Sennrich et al. (2017), produce an intermediate representation by combining, using a feed-forward layers, s_j, c_j, and y_{j-1}. A softmax layer is applied to this representation. The model is trained using a negative log-likelihood loss and back propagation through time (BPTT) (Rumelhart et al., 1986; Werbos, 1988). Training RNN, however, can be difficult due to the exploding and vanishing gradient problems (Bengio et al., 1994). To remedy this Long Short-Term Memory Networks (LSTM) (Gers et al., 2000) and Gated Recurrent Units (GRU) (Cho et al., 2014) were proposed. As we use GRUs throughout this thesis, we will focus exclusively on them, however they are conceptually similar to LSTMs.

Gated Recurrent Units use two linear gates to control the flow of information at every update. The reset gate determines how much information from the previous timestep to forget:

$$r_i = \text{sigmoid}(W_e e(x_i) + W_1 h_{i-1}) \tag{2.13}$$

The hidden state is updated using the reset gate to form an intermediate representation:

$$\hat{h}_i = \tanh(W_e e(x_i) + W_3(r_i \odot h_{i-1}) \tag{2.14}$$

where $\odot$ is the Hadamard product. The update gate determines how much information from the previous timestep should be passed to the future, and is calculated as:

$$z_i = \text{sigmoid}(W_e e(x_i) + W_h h_{i-1}) \tag{2.15}$$

Finally the intermediate representation with the hidden state are combined:

$$h_i = (1 - z_i) \odot h_{t-1} + z_i \odot \hat{h}_i \tag{2.16}$$

The use of linear gates allows information to easily propagate into the future, for instance GRU can set the update gate to 0, meaning all the previous information is preserved, in contrast it would be hard for an RNN to learn a weight matrix that preserves previous hidden state. LSTMs follow a similar approach but include an additional gate as well as an additional state.

2.1.2 Transformers

Transformers have become an increasingly popular alternative to RNN, which replace recurrent connections with multi-headed self-attention. In self-attention, unlike in RNN where the previous hidden state is updated with a new word, the hidden state is updated by attending to all hidden states in the sequences; as such the hidden state is a weighted sum of all hidden states of the sentence. In this way transformers encode source sentences simultaneously, not sequentially, which allows the encoder to make use of both the left and right context. Additionally, by using attention rather than recurrences, transformers are better able to model long-distance dependencies, as information does not need to be passed through multiple recurrent steps.

As shown in Figure 2.1 a Transformer is composed of an encoder and decoder, where both are made up of multiple identical transformer layers. In the encoder, a transformer layer is composed of two sublayers, a self-attention layer (Figure 2.1, Multi-Head Attention), which allows a hidden state to attend to all other hidden state; and a point-wise, fully connected feedforward layer (Figure 2.1, Feed Forward). A residual connection (He et al., 2016) is employed around each of the two sublayers (Figure 2.1, Add), followed by layer normalization (Ba et al., 2016) (Figure 2.1, Norm). A hidden encoder state at time-step i and layer l is defined as:

$$h_i^l = (h_i^{l-1} + \text{self-attention}(h_i^{l-1})) + f(h_i^{l-1} + \text{self-attention}(h_i^{l-1})) \tag{2.17}$$

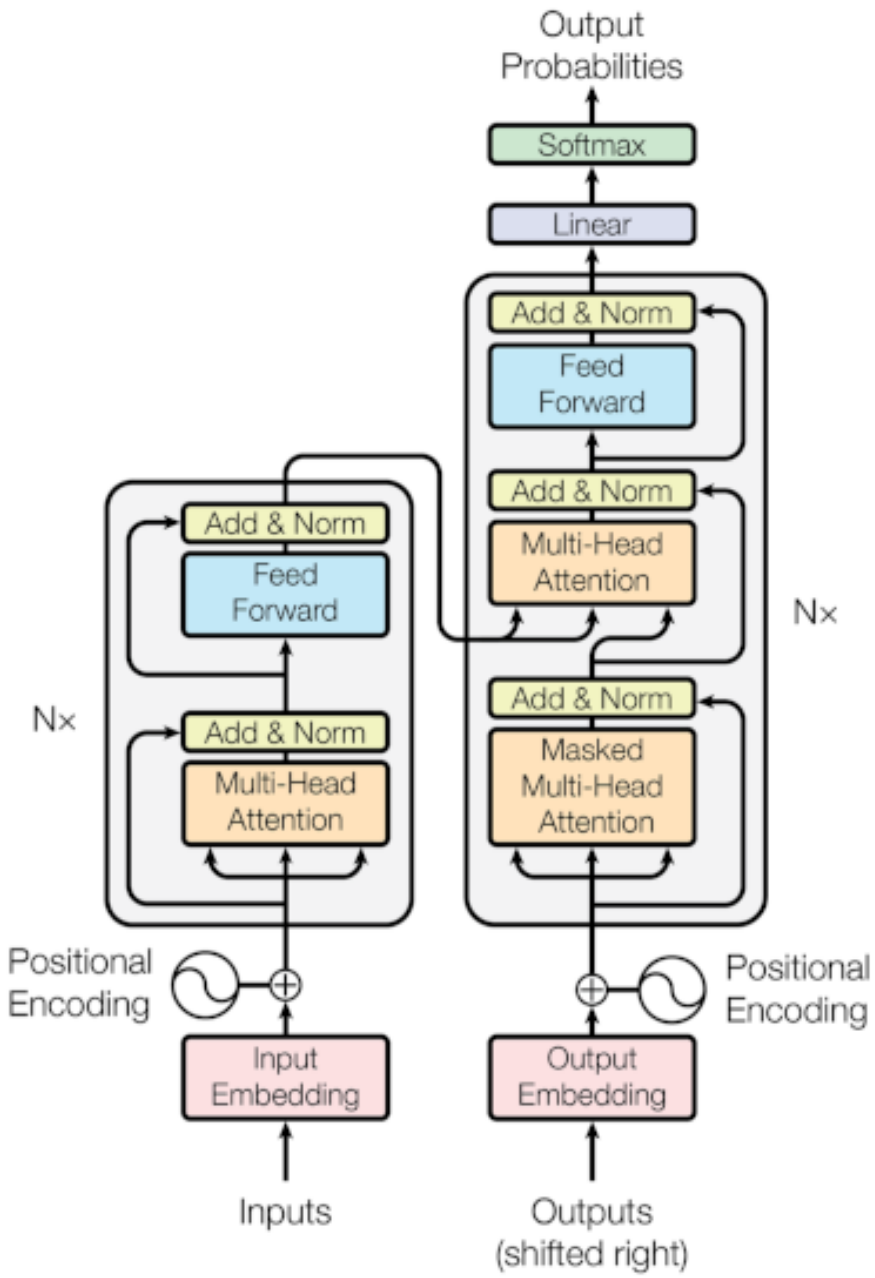

Figure 2.1: Diagram of Transformer encoder-decoder, as adapted (scaled) from Vaswani et al. (2017). The Transformer encoder-decoder is composed of two parts, on the left the encoder, and on the right the decoder. Both the encoder and decoder consist of N identical layers, which take in as input the output of the previous layer. An encoder layer consists of two sublayers: a self-attention mechanism and feedforward layer. After each sub-layer a residual connection is used and layer normalization is applied. The decoder has an additional sublayer which attends to the output of the encoder. A softmax layer is applied to the top of the decoder which predicts the output. The input to both the encoder and decoder are word embeddings and positional embeddings.

where f is a point-wise, fully connected feedfoward layer and the output of the encoder is the hidden states from the top transformer layer $h = (h_1^L, ...h_{|x|}^L)$.

For the self-attention layer, Transformers use a scaled dot product attention mechanism, where each source hidden state h_i^l is linearly transformed into a query vector q_l^i, a key k_i^l vector, and a value vector v_i^l. Attention is then performed between the query and keys, forming the output of the sublayer, a weighted average of the value vectors, as described in Section 2.1.1.

Transformers extend attention to multi-headed attention, which consists of multiple independent attention mechanisms, with separate query, key, and value matrices. The output of the independent attention mechanisms are then concatenated together to produce the output of the self-attention sublayer. Using multiple attention layers allows the attention mechanisms to focus on different aspects of output, such as syntactic or semantic phenomena.

For the input layer, h^0, word embeddings and positional embedding are added together and then used as the input to the model:

$$h_i^0 = W_e e(x_i) + W_{pos} e(i) \tag{2.18}$$

Unlike recurrent based approaches, where there are distinct timesteps, transformers have no distinct order. To remedy this, positional embeddings, representing a distinct timestep, are added to the word embeddings, where positional embeddings can be learnt or a fixed representation.

Transformer Decoder

Transformer decoders (Figure 2.1, second block) are tasked with producing the output sequence one word at a time, and they take as input the output of the encoder c, and the previously predicted words (Figure 2.1, Outputs (shifted right)):

$$P(y|x) = \prod_j^{|y|} P(y_j|y_{<j}, c) \tag{2.19}$$

Transformer decoders are similar to transformer encoders and consist of multiple transformer decoder layers. However they include a third sublayer (Figure 2.1, Masked Multi-Head Attention) before the self-attention layer, an attention layer which attends to the output of the encoder. For time step i in layer l, we define the hidden state s_i^l as[1]:

[1]For simplicity we exclude the residual connections.

$$s_j^l = f(\text{self-attention}(\text{attention}(s_j^{l-1}), c)) \tag{2.20}$$

Attention is also multi-headed and uses a scaled dot product. A mask is applied to self-attention in the decoder ensuring that the decoder can only attend to previous hidden states. A softmax layer is applied to the final layer of the decoder s_j^L which produces a distribution over the vocabulary. The first decoder layer is defined as:

$$s_j^0 = W_e e(x_{j-1}) + W_{pos} e(j) \tag{2.21}$$

Transformers are trained using a negative log-likelihood loss and BPTT.

2.1.3 Decoding

When decoding, the goal to find the most probable output y of a given source sentence x, argmmax $P(y|x)$. However the search space is the vocabulary size to the power of the maximum sequence length, with vocab size often being in the tens of thousands and max length, as measured by the number of tokens, often ranges from 32 to 128. It is therefore intractable to score all possible outputs. Instead, approximations are used; one approach is greedy decoding, where at every timestep the most probable word is chosen. However, a locally best decision does not guarantee that the sequence as a whole is the most likely. As such, a popular alternative is beam search, which compares probabilities of sequences. A fixed beam size is used and at each time step the N-most probable sequences are kept and the rest discarded.

To counter the effect of long sequences having a lower probability than short sequences, as at each time step the probability can never increase, length normalisation is usually applied:

$$P(y|x) * |Y|^{\alpha} \tag{2.22}$$

A higher α leads to long sentences being given a higher scores.

2.2 Evaluation Metrics

Within this section, we provide an overview of the automatic evaluation metrics used throughout the thesis. The majority of the metrics measure the token level overlap between the model output sentence g, and the references r. Additionally, some metrics

examine the overlap between the source sentence s and the output sentence. Metrics often can be computed at the sentence level, where scores from individual sentences are averaged, or at the corpus level, where the overlap is scored for the entire test set. Throughout this thesis we use corpus level metrics, unless otherwise stated.

BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) BLEU was originally developed for evaluating machine translation output, however, it is now commonly used to measure adequacy for paraphrasing tasks. It combines the n-gram precision, p_n, between the generated output, g, and the reference r, and a brevity penalty (BP):

$$\text{BLEU}(g,r) = BP \cdot \exp\big(\sum_{n=1}^{N} \frac{1}{N} \log p_n\big) \tag{2.23}$$

$$BP = \begin{cases} 1 & \text{if } |g| > |r| \\ e^{(1-|r|)/|g|} & \text{if } |g| <= |r| \end{cases} \tag{2.24}$$

where $|g|$ is the length (number of tokens), of the candidate output, $|r|$ is the length of the reference, and the standard choice is $N = 4$. BLEU was developed for use as a *corpus* level metric, which, when used as a sentence level metric, it is fairly common for $log(p_n = 0)$, leaving BLEU to be undefined. To avoid this, smoothing is often applied (Chen and Cherry, 2014). BLEU can be used with multiple references. For corpus level BLEU, output n-grams are matched against any reference n-gram and the reference length $|r|$ is set to the shortest reference length. When using sentence level BLEU, the output can be compared against individual references, the maximum or average BLEU is then reported.

self-BLEU While BLEU measures the distance between the model output and the reference, self-BLEU measures the distance between the output and the source, quantifying the amount of rewrites that the model has performed.

iBLEU (Sun and Zhou, 2012) combines BLEU and self-BLEU, rewarding model output g which is close to the reference r_n and penalising the output which is close to the source sentence s:

$$\text{iBLEU}(s,r,g) = \alpha \text{BLEU}(g,r) - (1-\alpha)\,\text{BLEU}(g,s) \tag{2.25}$$

The α parameter determines the importance of matching against the reference sentence, where a higher α de-emphasises the consequences of being close to the source sentence.

Copy is a corpus level metric which measures the percentage of sentences copied (with no changes made) from the source to the output, quantifying the extent to which a model performs any rewrites at all.

ROUGE-N (Lin, 2004a) ROUGE-N is a recall-based evaluation metric, measuring the n-gram recall between the generated output and the references. Unlike BLEU where the average [1-4]-gram overlap is used, only one N is chosen, it is therefore normal to report multiple ROUGE scores, i.e. ROUGE-1 and ROUGE-2. When there are multiple references the recall n-gram score is based on the proportion of reference that contain the n-gram. Although less common, ROUGE-N can use precision or F1 overlap between the output and the reference.

ROUGE-L (Lin, 2004a) measures the longest matching sequence of words between the generated output and the reference sentence, using the longest common subsequence (LCS) algorithm. Unlike ROUGE-N, ROUGE-L has no parameter to select (the choice of n-gram), and can compare arbitrarily long sequences. ROUGE-L is defined as:

$$\text{ROUGE-L}(g,r) = \frac{|\text{LCS}(g,r)|}{|r|} \tag{2.26}$$

where LCS is the longest common subsequence. When there are multiple references $\text{LCS}(g,r)$ is the union of all $\text{LCS}(g,r_i)$.

ROUGE-S (Lin, 2004a) measures skip-bigram matches between the generated output and the reference. A skip-gram is a type of n-gram where the words do not need to be in consequtive order, instead there can be additional words between them. ROUGE-S is calculated in the same way as ROUGE-N, where n-grams are replaced with skip-grams.

FKGL (Kincaid et al., 1975) The Flesch-Kincaid Grade Level index (FKGL) measures the readability of the output, by taking into account the average number of words per sentence and the average number of syllables per word and is calculated as follows:

$$\text{FGKL} = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \tag{2.27}$$

Scores start from -3.4, with no upper limit. A higher score indicates that the text is easier to read.

FKBLEU (Xu et al., 2016) One of the downsides of FKGL is that it only measures readability, and offers no guarantee that the output is semantically related to the input. To remedy this FKBLEU combines FKGL and iBLEU to measure both readability and adequacy. The resulting metric, FKBLEU, is defined as a geometric mean of the iBLEU and the FKGL difference between input and generated sentences:

$$\text{FKBLEU} = \text{iBLEU}(s,r,g)^{0.5} * \text{FKdiff}(g,s)^{0.5} \tag{2.28}$$

$$\text{FKdiff}(g,s) = \text{sigmoid}(\text{FKGL}(s) - \text{FKGL}(g)) \tag{2.29}$$

SARI (Xu et al., 2016) is calculated using the average of three rewrite operation scores: addition, copying, and deletion. It rewards addition operations when the system's output is not in the input but occurs in the references. Analogously, it rewards words deleted/retained if they are in both the system output and the references. For addition and copying it uses F1 4-gram overlap, however for deletion it uses 4-gram precision, encouraging models to be conservative when deleting. It is calculated as follows:

$$SARI(s,r,g) = (F1_{keep}(s,r,g) + F1_{add}(s,r,g) + P_{delete}(s,r,g))/3 \tag{2.30}$$

SARI supports multiple references, rewarding an operation proportional to the number of times that the operation appears within the reference.

TER (Translation Error Rate) (Snover et al., 2006) is a metric developed to determine the amount of Post-Editing required to correct a generated translation. It calculates the minimal number of edits required to transform the generated output to the reference. Edits include insertion, deletion, and substitution of single words as well as shifts of phrases (multiple words). A shift moves a phrase within the output to another location. All edits, including shifts of any length and distance have the same cost. TER

reports the number of each edit type, and by summing the number of edits is defined as:

$$\text{TER}(r_i, c) = \frac{|\text{edits}(r_i, c)|}{|r|} \tag{2.31}$$

where $|r_i|$ is the number of words in the reference, and the $|\text{edits}(r_i, c)|$ is the number of edits requires to change the reference into the generated output. It is calculated at a sentence level and when there are multiple references $|r|$ is the average reference length and minimum TER per reference is returned.

2.3 Alternatives to Transfer Learning

In this thesis, we use transfer learning to overcome a lack of paraphrasing data. However, there are several alternatives that we do not explore, which could have been used.

2.3.1 Unsupervised Learning

Throughout this thesis we have assumed the existence of supervised parallel training data that can be transferred to our primary task for which we have no data. However, in unsupervised learning, the approach requires no supervised training data from a related task, instead using only non-parallel data. Unsupervised natural language generation models have recently shown promising results on a variety of tasks. Artetxe et al. (2018); Lample et al. (2018), demonstrated how an *unsupervised* neural machine translation model can translate between two languages by training on non-parallel English and German. They train an encoder-encoder model using two objectives: (1) *denoising*, where a source sentence is noised and then the corresponding decoder is tasked with reconstructing the original sentence and (2) *on-the-fly back-translation*, which translates the sentence in inference mode; this translation is then encoded and the task is to reconstruct the original sentence. Surya et al. (2019) showed that with additional coverage loss how this approach could be applied to sentence simplification, where one dataset consists solely of simple sentences and another dataset consists solely of complex sentences. We further explain this approach in Chapter 5.

As discussed in Chapter 1 style transfer tasks, where the style of the output differs from the input has increasingly used unsupervised techniques. In this task an input text of style 1, must be written into style 2. For example taking an informal sentence and making it formal, or rewriting modern English into Shakespearean English. There

are rarely parallel style transfer datasets, however it is common for there to exist non parallel data in both styles. A common approach is variables disentanglement where the semantics of the source sentence and the style of the source sentence are separated. As mentioned in Chapter 1 (Section 1.4) most solutions use an adversarial approach, where discriminators try to predict the style and encoder is trained to perturb the discriminator, leading to a style agnostic representation of the source sentence (Fu et al., 2018; Shen et al., 2017; Zhao et al., 2018b).

2.3.2 Pre-Training

unsupervised pre-training on large text corpora has provided signifcant benefits to both NLG and NLP. With BERT (Devlin et al., 2019), and related models such as RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020) have shown significantly improved results on multiple NLU benchmarks such as GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), and SQuAD (Rajpurkar et al., 2016). These models were trained using a Masked Language Modelling MLM) objective, where a model uses the context words surrounding a [MASK] token to try to predict what the [MASK] word should be. For example, for the sentence *"Alaska is about [MASK] times large than New York"*, the model is trained to predict the missing token, *twelve*. Unlike encoder-decoder models described in Section 2.2, these approaches consist of an encoder, where a softmax layer is applied to the top of the encoder to predict the masked tokens.

Left-to-right language models, commonly called language models, are another popular pre-training approach. Unlike MLM, these models have no encoder, instead consisting solely of a decoder, where at every time-step they predict what the next word is, given all the previous words. Radford et al. (2019); Brown et al. (2020) with GPT-2 and GPT-3 respectively, showed that large language models trained on billions of tokens could achieve impressive results on NLU benchmarks.

As neither language models nor masked language models follow an encoder-decoder paradigm it is non-trivial to use these approaches within encoder-decoder models. However, there has been work in adapting them to an encoder-decoder framework. Liu and Lapata (2019) proposed initializing both the encoder and decoder with BERT and then training on a document summarization task. As BERT is more similar to an encoder rather than decoder, a separate learning rate was set for the encoder and decoder. Rothe et al. (2020) experimented with initializing the encoder and decoder, either randomly, with BERT, or with GPT-2. They evaluated on sentence fusion tasks

and document summarization tasks, showing BERT-2-random achieved impressive results. They also showed they were able to train the sentence fusion model with as few as a 4500 sentences. Mallinson et al. (2020b); Malmi et al. (2019) used BERT to initizalize a text-editing model, and on a sentence fusion tasks showed that a model was able to train on as few 450 sentence fusion examples.

In contrast to the encoder or decoder only approach, there have been several pre-trained encoder-decoder models. These models are trained to produce the output, or part of the output, sentence from the input sentence which has had noise or/and masking applied to it (Lewis et al., 2020; Raffel et al., 2019), achieving impressive results on a wide range of NLG tasks.

While pre-training models have been shown to compensate for a lack of large training generation datasets they still require training data in order to fine-tune the model. As such these approaches are orthogonal to this thesis where we assume no training data. Additionally, in this thesis we avoided using pre-trained models in part due to the recency in which the models have been released, but also because it ties our architectural decisions to those of the pre-trained models, often requiring the use of large models, making training expensive and slow.

2.3.3 Domain Adaptation

Domain adaptation is the ability to apply a model trained on a *source* domain to a different *target* domain. This is desirable as there may not be data in the target domain and domain mismatches between training and testing negativity impacting model performance (Koehn and Knowles, 2017; Kobus et al., 2017).

Domain adaptation and transfer learning both focus on adapting data, transferring learning focuses on how best to adapt data from another task, domain adaptation however focuses on how best to adapt data for the same task from another domain. This is particularly relevant to question 2 of the thesis, where we could consider general purpose data to be out-of-domain and personalized data, which rarely exists, could be considered in-domain data.

Domain adaptation can be broken down into four distinct approaches: data, training, architecture, and decoding (Chu and Wang, 2018). Data centric approaches adapt the data the model is being trained on. Moore and Lewis (2010) proposed extracting in-domain data from out-of-domain datasets by scoring sentences using a language model trained on the in-domain data and keeping only datapoints that receive a high

score. Alternatively silver in-domain parallel data can be created by back-translating, using a model trained on the inverse task, large amounts of in-domain monolingual data (Sennrich et al., 2016d).

Training objective approaches change how the model is trained. A common approach is train on out-of-domain data then fine-tuned on in-domain data (Luong and Manning, 2015). A related approach is to train on out-of-domain data, then fine-tuned on both of in-domain and out-of-domain data (Chu et al., 2017). Up-weighting can also be used, where the model is trained on both out-of-domain and in-domain data, but the model gives greater importance to the in-domain data (Wang et al., 2017).

Architecture-based approaches change the architecture of the model. Domain tags have been used to indicate which domain the output belongs to (Sennrich et al., 2016a). Britz et al. (2017) use a discriminator, which is trained to predict the domain of the source sentence, and the encoder is trained to perturb the discriminator.

Decoding-based approaches use custom decoding methods. Shallow fusion combines the probability from a language model trained on in-domain data with the probabilities produced by an encoder-decoder model trained on out-of-domain parallel data (Gulcehre et al., 2015).

2.4 Summary

Conclusion In this chapter, we introduced the background information needed for this thesis. We focused on encoder-decoder models and automatic evaluation metrics. In addition, we discussed alternatives to transfer learning for overcoming a lack of training data including unsupervised learning, domain adaptation, and pre-training.

Next Chapter In the next chapter, we introduce a transfer learning approach: Neural pivoting for unconstrained paraphrasing. We introduce unconstrained paraphrasing and show how two bilingual pre-trained, RNN translation models can perform monolingual paraphrasing with no paraphrasing data.

Chapter 3

Paraphrasing with Neural Pivoting

This chapter is based on Mallinson et al. (2017) which was published in EACL 2017 and answers the following question:

- No supervised data exists for a specific paraphrasing task. Can bilingual data be used as a source of training data for paraphrasing?

We consider the case of unconstrained paraphrasing and propose neural pivoting, an approach which requires no supervised paraphrasing training data. Instead, neural pivoting leverages bilingual corpora to find meaning-equivalent phrases in a single language by *pivoting* over a shared translation in another language, transferring supervising machine translation data over to paraphrasing. While previous pivoting approaches used Statistical Machine Translation (SMT), we show how pivoting can be performed using Neural Machine Translation (NMT) to produce sentential paraphrases. Our approach represents paraphrases in a continuous space, estimates the degree of semantic relatedness between text segments of arbitrary length, and generates candidate paraphrases for any source input. Experimental results across tasks and datasets show that neural paraphrases outperform those obtained with conventional phrase-based SMT pivoting approaches.

3.1 Introduction

Paraphrasing can be broadly described as the task of using an alternative surface form to express the same semantic content (Madnani and Dorr, 2010) and has been used in many NLP applications as discussed in Chapter 1 (Section 1.1). Historically, paraphrasing literature has focused on the automatic extraction of paraphrases from various

types of corpora consisting of parallel, non-parallel, and comparable texts (Deléger and Zweigenbaum, 2009; Barzilay and McKeown, 2001; Brockett and Dolan, 2005). However, these corpora are limited in size, domain, quality, and language. As such one of the most successful proposals does not use supervised paraphrasing data but instead employs bilingual parallel corpora to induce paraphrases based on techniques from phrase-based SMT (Koehn et al., 2003). As mentioned in Chapter 1 the intuition behind Bannard and Callison-Burch (2005) bilingual pivoting method is, that two English strings e_1 and e_2 that translate to the same foreign string f can be assumed to have the same meaning. They pivot over f to extract $\langle e_1, e_2 \rangle$ as a pair of paraphrases. Drawing inspiration from syntax-based SMT, several subsequent efforts (Callison-Burch, 2008; Ganitkevitch et al., 2011) extended this technique to syntactic paraphrases leading to the creation of PPDB (Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014), a large-scale paraphrase database containing over a billion paraphrase pairs in over 20 different languages.

Source	Paraphrases	
The general commanded his troops	The general spoke to his troops	✓
The soloist commanded attention	The soloist spoke to attention	×
She bought 10 stocks in Microsoft	She bought 10 shares in Microsoft	✓
She made stock for the soup	She made shares for the soup	×
he looked up the fact	he researched the fact	✓
he looked up at the sky	he researched at the sky	×

Table 3.1: Paraphrase examples which highlight the importance of a wider context. We see how lexical substitution can be appropriate in one context but not in another.

We revisit the bilingual pivoting approach from the perspective of NMT, an approach to machine translation based purely on neural networks (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015; Sutskever et al., 2014; Luong et al., 2015b; Vaswani et al., 2017). At its core, NMT uses a deep neural network trained end-to-end to maximize the conditional probability of a correct translation given a source sentence, using a bilingual corpus. In this chapter we introduce neural pivoting, an approach which ports the bilingual pivoting method to NMT and argue that it offers at least three advantages over conventional methods. Firstly, NMT has been shown to offer much higher quality translations than SMT. Secondly, our neural paraphrasing model learns continuous space representations for phrases and sentences (aka *embeddings*) that can

be usefully incorporated in downstream tasks such as recognizing textual similarity and entailment. Thirdly, the proposed model is able to score a pair of paraphrase candidates (of arbitrary length) and generate target paraphrases for a given source input. Due to the architecture of NMT, generation takes advantage of wider context compared to phrase-based approaches: target paraphrases are predicted based on the meaning of the source input and all previously-generated target words. Table 3.1 demonstrates the importance of using context when paraphrasing, where we see how lexical paraphrases in one context are appropriate, but not in other contexts.

In the remainder of the chapter, we survey existing paraphrasing datasets, introduce neural pivoting, and experimentally compare it to the phrase-based pivoting approach. We evaluate the model's paraphrasing capability both *intrinsically* in a paraphrase detection task (i.e., decide the degree of semantic similarity between two sentences) and *extrinsically* in a generation task. Across tasks and datasets our results show that neural paraphrases yield superior performance when assessed automatically and by people.

3.2 Background

The literature on paraphrasing is vast with methods varying according to the type of paraphrase being induced (lexical or structural), the type of data used (e.g., monolingual or parallel corpus), the underlying representation (surface form or syntax trees), and the acquisition method itself. We restrict ourselves to surveying existing supervised sentential paraphrasing corpora, generating paraphrase datasets with machine translation, and bilingual approaches.

3.2.1 Supervised Datasets

This section provides an overview of publicly available human-generated paraphrase datasets, with Table 3.3 providing an overview and Table 3.2 showing examples from these dataset. It should be noted that many of the existing paraphrasing datasets were created for either paraphrase identification, i.e. predicting if two sentences are paraphrases, or semantic similarity scoring, i.e. assigning a score on how semantically related two sentences are. These datasets can be converted into generation datasets by removing the negative examples.

Dataset	Source	Paraphrase
MSRP	David Gest has sued his estranged wife Liza Minelli for %MONEY% million for beating him when she was drunk.	Liza Minelli's estranged husband is taking her to court for %MONEY% million after saying she threw a lamp at him and beat him in drunken rages.
MSRP	Wynn paid $23.5 million for Renoir's "In the Roses (Madame Leon Clapisson)" at a Sotheby auction on Tuesday	Wynn nabbed Renoir's "In the Roses (Madame Leon Clapisson)" for $23.5 on Tuesday at Sotheby's
PIT-2015	Ezekiel Ansah wearing 3D glasses wout the lens	Wait Ezekiel ansah is wearing 3d movie glasses with the lenses knocked out
PIT-2015	Marriage equality law passed in Rhode Island	Congrats to Rhode Island becoming the 10th state to enact marriage equality
WikiAnswers	How big is the biggest mall?	most expensive mall in the world?
WikiAnsers	How many of india's population are muslim?	How many populations of muslims in india?
Quora	how to be a good geologist?	what should i do to be a great geologist?
Quora	why is creativity important?	why creativity is important?
Semeval	There are dogs in the forest.	The dogs are alone in the forest.
Semeval	A young person deep in thought.	A young man deep in thought.
Semeval	Un perro está con un juguete.	Un perro tiene un juguete.
Semeval	Una dama está cantando	Una dama cantando.
MTC	At least 12 people were killed in the battle last week.	At least 12 people lost their lives in last week's fighting
MTC	(Kuala Lumpur) Lien Hoe is expected to redeem the rest 65% of the bonds, whose tatal value is 53,810,000 lingji, at the end of this year by getting the loan finacing from banks.	(report from Kuala Lumpur) Lien Hoe expected that by way of bank loans, it may redeem the currently remaining 65% or 53,810,000 ringgit of its bonds, before the end of this year.
Book	There was once a Prince who wished to marry a Princess; but then she must be a real Princess.	ONCE upon a time there was a prince who wanted to marry a princess; but she would have to be a real princess.
Book	The head-master made a sign to us to sit down. Then, turning to the class-master, he said to him in a low voice–	The headmaster motioned us to be seated, then, turning to the teacher:

Table 3.2: Examples of paraphrases from existing datasets, highlighting the wide range in the types of rewrites and the varying quality of the paraphrases across datasets.

Dataset	Size	Method	Domain	Languages
MSRP	4K	Annotated	News	English
PIT-2015	4K	Annotated	Twitter	English
WikiAnswers	2.4M	User grouped	Questions	English
Quora	50K	User grouped	Questions	English
Books	45K	Automatic/Translation	Fiction	English
MTC	6K	Translation	News	English
Semeval English	6K	Annotated	Mixed	English
Semeval Spanish	1K	Annotated	Mixed	Spanish
Semeval Arabic	1K	Annotated	Mixed	Arabic

Table 3.3: Overview of Paraphrasing *datasets*, including the number of paraphrases (*size*), *method* in which they were collected, the *domain*, and the *languages* of the dataset.

MSR Paraphrase Corpus (MSRP, Dolan and Brockett (2005)) is a binary paraphrase classification dataset, which contains 6000 human-annotated sentence pairs, 68% of which are paraphrases. It was created by clustering related news articles. A classifier, using string similarity features, was then used to extract possible paraphrase pairs from these clusters, which were then hand annotated. As the classifier used string similarity features many paraphrases have high n-gram overlaps with the source sentence. In addition to being used for paraphrase identification, MSRP has been used as a paraphrase generation training set (Brad and Rebedea, 2017), and a test set (Roy and Grangier, 2019).

Twitter Paraphrase Corpus (PIT-2015, Xu et al. (2015a)) is a paraphrase identification dataset, which was created from Twitter's trending topic data. Multiple crowdworkers annotated pairs of tweets as paraphrases or not. This resulted in a paraphrase classification dataset containing 18,000 sentence pairs on 400 distinct topics, of which 30% were marked as paraphrases. In addition to being used for paraphrase detection it has also been used to extract idiomatic phrases (Pershina et al., 2015).

WikiAnswers (WikiAnswers, Fader et al. (2013)) contains questions taken from the question/answer website WikiAnswers[1]. To reduce duplicate questions, users grouped

[1] http://wiki.answers.com/

related questions together. WikiAnswers was then scraped to extract 2.4M distinct questions sets. Fader et al. (2013) sampled 100 sentence pairs, and found 55% of those sampled were valid paraphrases. Due to its size WikiAnswers has been used as training data for paraphrase generation models (Wang et al., 2019c; Liu et al., 2020; Li et al., 2019), as well as being used as a paraphrase evaluation set later in this chapter, semantic passers dataset (Berant and Liang, 2014), question answering dataset (Fader et al., 2013; Bernhard and Gurevych, 2008), and for paraphrase template extraction (Dong et al., 2017; Fader et al., 2013).

Quora[2] is a question/answer website, which released[3] a dataset of potential paraphrased questions collected from the site used for paraphrase identification. The corpus consists of 400,000 pairs of sentences which have been human annotated with binary labels, of which 12.5% are paraphrases. Due to Quora's size, it has been used extensively for training and evaluation paraphrase generation systems (Huang et al., 2019; Mao and Lee, 2019; Li et al., 2018; Wang et al., 2019c).

Barzilay and McKeown (2001) (Books) dataset consists of paraphrases extracted from multiple translations into English of the same fictional novels. Paraphrases are extracted using automatic sentence alignment, resulting in a corpus of 45000 paraphrases. 127 paraphrase pairs were sampled and 120 (94.5%) paraphrases were identified as being correct. This dataset was originally used for creating lexical and syntactic paraphrase rules; and within this chapter we use it for paraphrase generation evaluation.

The Multiple-Translation Chinese part 1-4 (MTC, Huang et al. (2002, 2003); Ma (2004, 2006)) contains news stories from different news agencies. These texts are translated into English by multiple different translation agencies, averaging 4 translations per source sentence. These multiple translations are aligned, under the assumption that translations of the same source sentence are paraphrases, to create a paraphrase corpus. As well as being used a multiple reference test set for translation, this has also been used for paraphrase metric evaluation (Weese et al., 2014) and as an evaluation sets by (Roy and Grangier, 2019) and us, as we describe in Section 3.4.6.

[2] http://www.quora.com)

[3] https://www.kaggle.com/c/quora-question-pairs

Semantic Textual Similarity (STS, Agirre et al. (2016)) is a semantic text evaluation task, where sentences are hand-annotated according to how related they are, giving a score between 1 and 6. In addition to being treated as a regression set, binary labels are also provided by thresholding the scores. The dataset was collected from a wide variety of domains and across three languages: English, Spanish, and Arabic, with the majority of the data being English. The Semantic Textual Similarity task has had multiple iterations, where each iterations contains the previous iteration data as well as new data. Later in this chapter we use an earlier iteration as a test set to evaluate paraphrase classification (Section 3.4.5). Additionally Roy and Grangier (2019) have also used STS as paraphrase generation test set.

As demonstrated in Table 3.3 there is limited supervised paraphrasing data, where only large dataset exists for English within the question domain, motivating an alternative approach to supervised learning. We propose an approach based on translation, which has been successfully used to create paraphrases datasets (Books & MTC), however, these datasets are small (see Table 3.3) as they required human generated translations. In the next section we demonstrate how human generated translations can be partially replaced with automatically generated translations and use these translations to generate paraphrase datasets.

3.2.2 Automatic construction of Paraphrasing Datasets

Since the completion of Mallinson et al. (2017) there have been several papers (Wieting et al., 2017; Wieting and Gimpel, 2018; Hu et al., 2019) which use a similar NMT pivoting approach to create paraphrase datasets automatically. These datasets were created by pairing references from bilingual dataset and a translation of the source sentence, using NMT. These approaches could by considered a special case of neural pivoting, where instead of using a NMT model to produce the intermediate foreign pivot, a human written reference is used instead. However, by using human references these approaches are not able to generate paraphrases for arbitrary sentences, as a human translation does not exist for all sentences. Instead they produce training data by decoding large amounts of bilingual data; a separate paraphrasing encoder-decoder model is then trained on these decoded and paired sentences, which can then be used to generate paraphrases for arbitrary sentences. This has the advantage that specialised neural pivoting models do not need to be used for paraphrasing; instead any model

can be used. However, decoding large numbers of sentences is computationally expensive, with each language needing millions of decoded sentences. Neural pivoting is therefore computationally cheaper, as it does not require decoding datasets for each language, but instead can use existing neural machine translation models.

Specifically, Wieting et al. (2017); Wieting and Gimpel (2018) used English-Czech translation data to create PARANMT, a large corpus of 50 million English paraphrases. This paraphrasing data was then used to train paraphrastic sentence embeddings, achieving state-of-the-art results on a wide range of semantic similarity tasks. Hu et al. (2019) proposed PARABANK, an 80 million paraphrase dataset, created using a similar approach to that of PARANMT. PARABANK, however, applied lexical constraints to the NMT decoder when generating translations, marking words to be kept or not. These constraints were used to ensure the paraphrases were sufficiently distinct from the input. Human annotators were asked to judge the paraphrases from PPDB, PARABANK, and PARANMT and found that 82% of PARABANKS were both grammatical and preserved meaning, compared to 73% of PARANMT, and 71% of PPDB.

3.2.3 Bilingual Pivoting

Source	Paraphrases
run	undertaken, ruled, turned, guaranteed, organised, organized
dog	puppy, doggie, doggy, lapdog, watchdog
significant quantity	large quantities, large quantity, substantial quantity
keenly	most strongly, deeply, strongly, eagerly, very

Table 3.4: Examples of PPDB phrase-based paraphrases.

Paraphrase extraction using bilingual parallel corpora was proposed by Bannard and Callison-Burch (2005). They first extract bilingual phrase tables and then obtain English phrase-based paraphrases by pivoting through foreign language phrases. Paraphrases for a given phrase are ranked using a paraphrase probability as defined by:

$$P(e_2|e_1) = \sum_f P(e_2|f)P(f|e_1) \tag{3.1}$$

using the translation model phrase probabilities $P(f|e_1)$ and $P(e_2|f)$ where f and e are the foreign and English strings, respectively. Several follow-up approaches have been proposed, including representing paraphrases via rules obtained from a synchronous

context-free grammar (Ganitkevitch et al., 2011; Madnani et al., 2007a) as well as labelling paraphrases with linguistic annotations such as CCG categories (Callison-Burch, 2008) and part-of-speech tags (Zhao et al., 2008). Pavlick et al. (2015) released PPDB 2.0, a dataset containing millions of lexical, phrasal and syntactic paraphrases. PPDB 2.0 used a supervised classifier to score how semantically related the automatically extracted paraphrases were, and provided additional automatic annotation, such as style information, of complexity and formality. Example paraphrasing rules from PPDB can be seen in Table 3.4.

In addition to extracting paraphrases, there has been much work on generating sentential paraphrases. Zhao et al. (2008); Ganitkevitch et al. (2011); Napoles et al. (2016) parametrise SMT systems with extracted paraphrase rules. These rules are then combined with other features, such as language model scores to generate paraphrases. Sun and Zhou (2012) trained two independent phrase-based SMT models, $e \rightarrow f$ and $f \rightarrow e$, which are then jointly fine-tuned for paraphrasing. A paraphrase is created by translating a source sentence into a single foreign pivot which is then backtranslated.

3.2.4 Neural Machine Translation

As discussion in Chapters 1 and 2, neural approaches have become the dominant approach to machine translation, as seen on the leader board of WMT 19 (Bojar et al., 2017). Central to this approach is an encoder-decoder architecture, where the encoder reads the source sequence into a sequence of continuous-space representations from which the decoder generates the target sequence one token at a time (Bahdanau et al., 2015; Sutskever et al., 2014). An attention mechanism (Bahdanau et al., 2015) is used to generate the region of focus during decoding.

We employed an encoder-decoder as the backbone of our paraphrasing model. In its simplest form our model exploits a one-to-one NMT architecture: the source English sentence is translated into k candidate foreign sentences and then back-translated into English. Inspired by multi-way machine translation, which has shown performance gains over single-pair models (Zoph and Knight, 2016; Dong et al., 2015; Firat et al., 2016a), we also explore an alternative pivoting technique which uses multiple languages rather than a single one. Our model inherits advantages from NMT such as a small memory footprint and conceptually easy decoding (implemented as beam search). Beyond paraphrase generation, we experimentally show that the representations learned by our model are useful in semantic relatedness tasks. Our model is

syntax-agnostic: paraphrases are represented on the surface level without knowledge of any underlying grammar. We capture paraphrases at varying levels of granularity: words, phrases or sentences, without having to *explicitly* create a phrase table.

3.3 Neural Pivoting

In this section we present PARANET, our **Para**phrasing model based on **Ne**ural Machine **T**ranslation. PARANET uses neural machine translation to first translate from English to a foreign pivot, which is then back-translated to English, producing a paraphrase. In the following, we briefly overview the basic encoder-decoder NMT framework and then discuss how it can be extended to paraphrasing.

3.3.1 NMT Background

In this chapter we restricted ourselves to Recurrent Neural Network (RNN) based neural machine translation models. However, we would like to note that our approach is applicable to all encoder-decoder models NMT models. In the neural encoder-decoder framework for MT (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015b), the encoder is used to compress the meaning of the source sentence into a sequence of vectors. The decoder, a conditional language model, generates a target sentence word-by-word. PARANET uses a bi-directional RNN, where each context vector h_i is the concatenation of the forward and the backward RNN's hidden states at time i.

The decoder is a conditional RNN language model that given the source sentence and the previously generated words, produces a probability distribution over the translation.

$$P(y|x) = \prod_{j}^{|y|} P(y_j|y_{<j}, x) \tag{3.2}$$

where j is the decoder time step, more details on neural machine translation models can be found in Chapter 2 (Section 2.1).

3.3.2 Pivoting

Pivoting is often used in machine translation to overcome the shortage of parallel data, i.e., when there is not a translation path from the source language to the target. Instead, pivoting takes advantage of paths through an intermediate language. The idea

of phrase-based pivoting dates back to at least da Fonseca and Carolino (1855), a Portuguese-English phrasebook, which is believed to have been created by translating a French-Portuguese phrasebook into English using only a French–English dictionary (Monteiro, 2004). More recently Kay (1997) observed that ambiguities in translating from one language into another may be resolved if a translation into some third language is available, and has met with success in traditional phrase-based SMT (Wu and Wang, 2007; Utiyama and Isahara, 2007) and in neural MT systems (Firat et al., 2016b).

In the case of paraphrasing, there is not a path from English to English. Instead, a path from English to French to English can be used. In other words, we translate a source sentence into a pivot language and then translate the pivot back into the source language. Pivoting using NMT ensures that the entire sentence is considered when choosing a pivot. The fact that contextual information is considered when translating allows for a more accurate pivoted sentence. It also places greater emphasis on capturing the meaning of the sentence, which is a key part of paraphrasing. Unlike Equation 3.1 where the pivots are a finite set of phrases, within NMT the pivots are entire sentences. As such there is an infinite set of pivots which can not be marginalised out.

A naive approach is one-to-one back-translation. The source English sentence E_1, is translated into a single French sentence F. Next, F is translated back into English, giving a probability distribution over English sentences, E_2. This translation distribution acts as the paraphrase distribution $P(E_2|E_1,F)$:

$$P(E_2|E_1,F) \approx P(E_2|F) \tag{3.3}$$

One-to-one back-translating offers an easy way to paraphrase, because existing NMT systems can be used with no additional training or changes. However, there are several disadvantages; for example the French sentence F must fully capture the exact meaning of E_1, as E_1 and E_2 are conditionally independent given F. Since there is rarely a clear one-to-one mapping between sentences in different languages, information about the source sentence can be lost, leading to inaccuracies in the paraphrase probabilities. To avoid this, we propose back-translating through multiple sentences within one and multiple foreign languages.

Multi-pivoting PARANET uses multiple foreign translations, as this helps to ensure that multiple aspects (semantic and syntactic) of the source sentence are captured.

Moreover, multiple pivots provide resilience against a single bad translation, which would prevent one-to-one back-translation from producing accurate paraphrase probabilities. We propose an approach analogous to that of PPDB (Equation 3.1), where foreign pivots are marginalised out. However, as there is a non-finite number of foreign pivots sentences this becomes intractable. Instead PARANET approximates this and pivots through the set of K-best translations $\mathcal{F} = \{F_1, ...F_K\}$ of E_1. Using Equation 3.1 and substituting in a NMT model (Equation 3.2), we get:

$$P(E_2|E_1) \approx \sum_{i=1}^{K} \Big(\prod_{j}^{|E_2|} P(E_{2j}|E_{2<j}, \mathcal{F}_i) P(\mathcal{F}_i|E_1) \Big) \tag{3.4}$$

Where k is the number of foreign pivots, if using a single language for a pivot this would be equal to the first translation models beam size. We note that this requires that we finish translating each pivot sentence before combining the probability distributions. This is problematic, as the pivots may translate to different outputs.

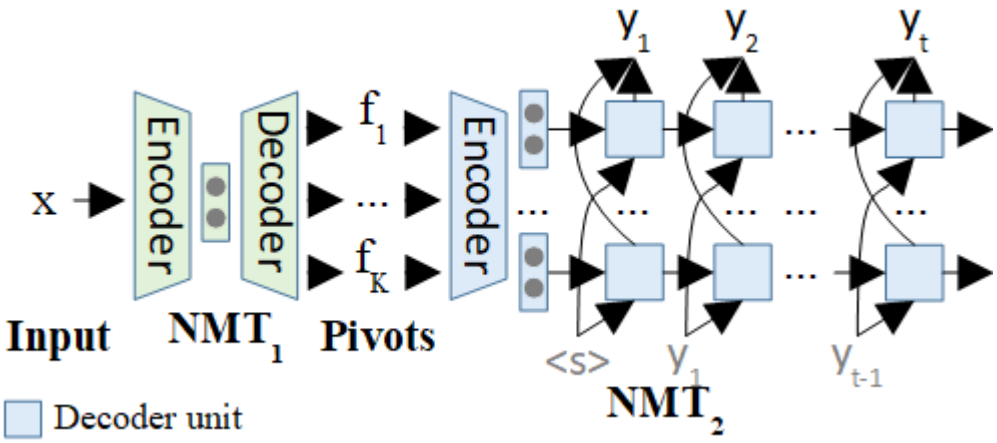

Figure 3.1: Overview of NMT-based paraphrase generation, as adapted from Dong et al. (2017). Source NMT (green) translates source sentence x into pivots $f_1 ... f_K$ which are then backtranslated by target NMT (blue) where K decoders jointly predict tokens at each time step.

Instead, we draw inspiration from the Firat et al. (2016b) late averaging approach, which averages the output probability of each translation at each time step. Each translation path individually computes the distribution over the target vocabulary $P(E_{2j}|E_{2<j}, \mathcal{F}_1)$ and $P(E_{2j}|E_{2<j}, \mathcal{F}_2)$, which is then averaged at each time step:

$$P(E_2|, \mathcal{F}_1, \mathcal{F}_2) \approx \prod_{j}^{|E_2|} \big(\lambda_1 P(E_{2j}|E_{2<j}, \mathcal{F}_1) + \lambda_2 P(E_{2j}|E_{2<j}, \mathcal{F}_2) \big) \tag{3.5}$$

where λ_1 and λ_2 are set to 0.5. We use this approach to provide a tokenwise approximation to Equation 3.4; for each output token we marginalise out the foreign pivots:

$$P(E_2|E_1) \approx \prod_{j}^{|E_2|} \Big(\sum_{i=1}^{K} P(\mathcal{F}_i|E_1) P(E_{2j}|E_{2<j}, \mathcal{F}_i) \Big) \tag{3.6}$$

In this case λ weights are set to the initial translation probabilities $P(\mathcal{F}_i|E_1)$, thus capturing the model's confidence in the accuracy of the translation. To ensure a probability distribution, we normalize the K-best list $\mathcal{F}$, such that the translation probabilities sum to one. An outline of our approach can be seen in Figure 3.1.

Multi-lingual Pivoting PARANET further expands on the multi pivot approach by pivoting not only over multiple sentences from one language, but also over multiple sentences from multiple languages. Multi-lingual pivoting has been recently shown to improve translation quality (Firat et al., 2016b), especially for low-resource language pairs. Here, we hypothesize that it will also lead to more accurate paraphrasing.

Multi-lingual pivoting requires a small extension to late-weighted combination. We illustrate with German as a second language. First, the source sentence is translated into a K-best list of French $\mathcal{F}^{Fr}$, and a K-best list of German $\mathcal{F}^{De}$. Late-weighted combination is then applied, producing $P(E_{2j}|E_{2<j}, \mathcal{F}^{Fr})$ and $P(E_{2j}|E_{2<j}, \mathcal{F}^{De})$. These two output distributions are averaged, producing a multi-sentence, multi-lingual paraphrase score:

$$\begin{aligned} \prod_{j}^{|y|} \frac{1}{2} \Big(& \sum_{i=1}^{K} P(E_{2j}|E_{2<j}, \mathcal{F}_i^{Fr}) P(\mathcal{F}_i^{Fr}|E_1) + \\ & \sum_{i=1}^{K} P(E_{2j}|E_{2<j}, \mathcal{F}_i^{De}) P(\mathcal{F}_i^{De}|E_1) \Big) \end{aligned} \tag{3.7}$$

This can be trivially generalized to multiple languages. In this chapter we use up to three.

3.3.3 PARANET Applications

The applications of PARANET are many and varied. We discuss some of these here and present detailed experimental evidence in Section 3.4.

Detection PARANET can be readily used for paraphrase detection (the task of analyzing two text segments and determining if they have the same meaning), by computing Equation (3.6). In addition, it can identify which linguistic units (sentences,

phrases, word) are considered paraphrases and to what extent. PARANET's explanatory power stems from the attention mechanism inherent in the NMT systems.

In encoder-decoder models, attention is used during each step of decoding to indicate which are the relevant source words. In our case, each word of the paraphrase attends to words within the pivot sentence and each word in the pivot sentence attends to words within the source sentence. By summing out the weighted pivot sentence, it is possible to see the attention from paraphrase to source:

$$\alpha(E_2^j, E_1^i, \mathcal{F}) = \sum_i^K \left(P(E_2|E_1, \mathcal{F}_i) \cdot \sum_m^{|F|} (\alpha_{m,j}^{F,E_2} \cdot \alpha_{i,m}^{E_1,F})\right) \tag{3.8}$$

where $\alpha_{i,m}$ is the attention weight source token at index i apply to target token at index m.

Examples are shown in Figure 3.2 where attention has successfully identified the semantically-equivalent parts of two sentences. It should be noted that recent work has disputed the notion that attention can be used for interpretability Jain and Wallace (2019); Wiegreffe and Pinter (2019). However, there has also been interest on increasing the interpretability of attention (Tutek and Šnajder, 2020) .

Generation Furthermore, PARANET can be readily used to perform text generation (via the NMT decoder) without additional resources or parameter estimation. It is able to generate paraphrases for words, phrases, and sentences. As PARANET was primarily trained on senential bilingual data it is best suited for generating entire sentences. However, additional word/phrase level bilingual data could have been trained on.

Paraphrastic Embeddings The successful use of word embeddings in various NLP tasks has provided further impetus to use paraphrases. Wieting et al. (2015) take the paraphrases contained in PPDB and embed them into a low-dimensional space using a RNN similar to Socher et al. (2013). In follow-up work (Wieting et al., 2016), learn sentence embeddings based on supervision provided by PPDB. In our approach, embeddings are learned as part of the model and are available for any-length segments making use of no additional machinery beyond NMT itself.

Data Augmentation As discussed in Chapter 1 (Section 1.1) paraphrasing has often been used for data augmentation. Within Dong et al. (2017) PARANET was integrated into a question answering framework, where multiple paraphrases of a question are given as input to the model. It was compared against a PPDB based system, where

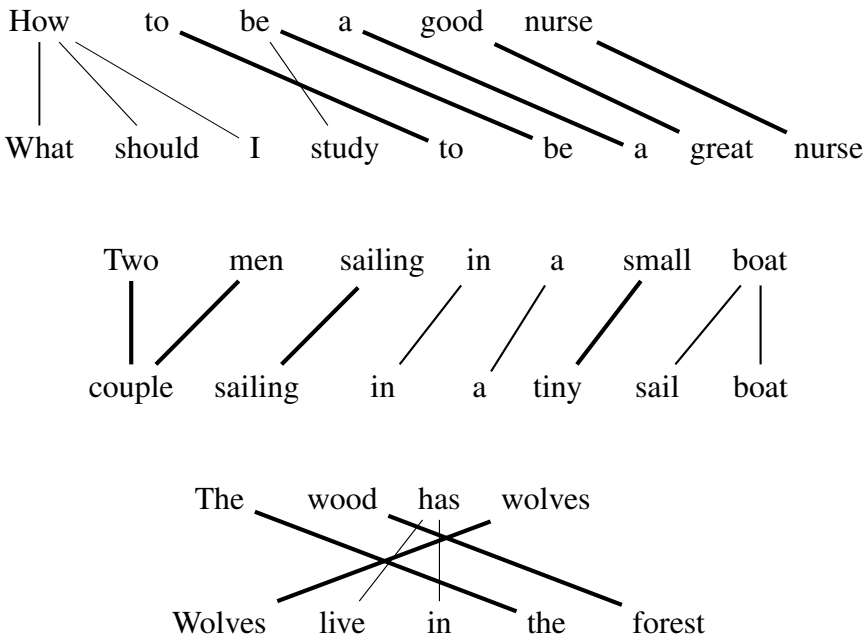

Figure 3.2: Paraphrase attention between two sentences. Line thickness indicates the strength of the attention.

words and phrase in the source sentence were paraphrased using PPDB. Across multiple question-answering datasets PARANET was shown to outperform the PPDB based approach.

3.4 Experiments

PARANET was evaluated in several ways: (a) we examined whether the paraphrases learned by our model correlate with human judgments of paraphrase quality; (b) we assessed PARANET in paraphrase and similarity detection tasks; and (c) in a sentence-level paraphrase generation task. We first present details on how PARANET and comparison models were trained and then discuss our results.

3.4.1 Neural Machine Translation Training

We used Groundhog[4] as the implementation of the NMT system for all experiments. We generally followed the settings and training procedure from previous work (Bahdanau et al., 2015; Sennrich et al., 2016c). As such, all networks have a hidden layer size of 1000, and an embedding layer size of 620. During training, we used Adadelta

[4] http://www.github.com/sebastien-j/LV_groundhog

(Zeiler, 2012), a minibatch size of 80, and the training set was reshuffled between epochs. We trained a network for approximately 7 days on a single GPU, then the embedding layer was fixed and training continued, as suggested in Jean et al. (2015), for 12 hours. Additionally, the softmax was calculated over a filtered list of candidate translations. Following Jean et al. (2015), we set the common vocabulary size as 10000 and 25 uni-gram translations, using a bilingual dictionary based on fast-align (Dyer et al., 2013).

In our experiments, we used up to six encoder-decoder NMT models (three pairs); English→French, French→English, English→Czech, Czech→English, English→German, German→English. All systems were trained on the available training data from the WMT15 shared translation task (4.2 million, 15.7 million, and 39 million sentence pairs for EN↔DE, EN↔CS, and EN↔FR, respectively). For EN↔DE and EN→CS, we also had access to back-translated monolingual training data (Sennrich et al., 2016c), which we also used in training. The data was pre-processed using standard pre-processing scripts[5] found in MOSES (Koehn et al., 2007). Words were split into sub-word units, following Sennrich et al. (2016d).

3.4.2 Statistical Machine Translation Training

Throughout our experiments we compare PARANET against a paraphrase model trained with a commonly used Statistical Machine Translation system (SMT), which we henceforth refer to as PARASTAT. Specifically, for each language pair used, an equivalent IBM Model 4 phrase-based translation model was trained. Additionally, an Operation Sequence Model (OSM) was included, which has been shown to improve the performance of SMT systems (Durrani et al., 2011). SMT translation models were implemented using both GIZA++ (Och and Ney, 2003) and MOSES (Koehn et al., 2007) and were trained using the same pre-processed bilingual data provided to the NMT systems. The SMT systems used a KenLM 5-gram language model (Heafield, 2011), trained on the mono-lingual data from WMT 2015. For all languages pairs, both KenLM and MOSES were trained using the standard settings.

Under the SMT models, paraphrase probabilities were calculated analogously to Equation (3.6):

$$P(E_2|E_1) \approx \sum_{i=1}^{K} P(E_2|\mathcal{F}_i)P(\mathcal{F}_i|E_1) \tag{3.9}$$

[5]including: Truecasing (truecase.perl), and corpus cleaning (clean-corpus-n.perl)

where $P(E_2|\mathcal{F}_i)$ and $(\mathcal{F}_i|E_1)$ are defined by the phrase-based translation model, and $\mathcal{F}$ denotes the K-best translations of E_1. This approach differs from PARANET in two ways: (1) the pivot probabilities are combined at the sentence level rather than at the token level and (2) the pivot sentences have to be combined outside of the decoder. This is due to the limitation of SMT, providing no easy ways in which to perform multi-source translation within the decoder.

Direction	F→E			E→F		
System	SMT	NMT	SOTA	SMT	NMT	SOTA
French	0.241	0.201	0.349 [6]	0.233	0.271	0.336 [7]
German	0.207	0.282	0.320 [8]	0.208	0.248	0.320 [9]
Czech	0.216	0.197	0.262 [10]	0.145	0.176	0.188 [11]

Table 3.5: BLEU scores (WMT 2015 test set) for SMT and NMT, and SOTA models (foreign to English (F→E) and English to foreign (E→F) directions).

BLEU scores for NMT and SMT systems, and the current state-of-the-art (SOTA) can be seen in Table 3.5. We note while NMT and SMT achieve comparable scores, they are below the current state-of-the-art systems. As such we would hope to see better performance with better machine translation models.

3.4.3 Correlation with Human Judgments

The PPDB 2.0 Human Evaluation dataset is a sample of paraphrase pairs taken from PPDB which have been human annotated for semantic similarity (Pavlick et al., 2015). 26,455 samples were taken from a range of syntactic categories, resulting in paraphrase candidates varying from single words to multi-word expressions. Each paraphrase pair was judged by five people on a 5-point scale. Ratings were then averaged giving each paraphrase pair a score between not related(1) and a paraphase(5).

Using this dataset we measure the correlation (Spearman ρ) between (length normalized) PARANET probabilities (Equation (3.6)) assigned to paraphrase pairs and human judgments. Figure 3.3 shows correlation coefficients for all language pairs us-

[6] Edunov et al. (2018)
[7] Marie et al. (2015)
[8] Peter et al. (2017)
[9] Peter et al. (2017)
[10] Ding et al. (2016)
[11] Bojar and Tamchyna (2015)

Score	5	4	3	2	1
Source	about 10	afflicts	uncivilized	advise	thank you
Paraphrase	roughly 10	effects	dirty	guess	yes/match?
Source	gladness	what then	drafting	preferably	should
Paraphrase	joy	what now	preprocessing	ever	protect
Source	5000.00	redifining	telescope	just now	sweet
Paraphrase	5000	restating	binoculars	doing what	uh

Table 3.6: Example pairs at each quality level, according to the average of 5 ratings assigned by annotators on MTurk, 5 being the most similar.

ing a single foreign pivot and 200 pivots. Across all language combinations[12] multiple pivots achieve better correlations, with the German, Czech pair performing best with $\rho = 0.53$. For comparison, Pavlick et al. (2015) report a correlation of $\rho = 0.41$ using Equation (3.1) and PPDB (Ganitkevitch et al., 2013). The latter contains over 100 million paraphrases and was constructed over several English-to-foreign parallel corpora including Europarl v7 (Koehn, 2005) which contains bitexts for the 19 European languages. The approach of Wieting et al. (2016) as discussed in section 3.2.3 reports a correlation of $\rho = 0.61$

Following Pavlick et al. (2015), we next developed a supervised scoring model. Specifically, we fit a decision tree regressor on the PPDB 2.0 dataset using the implementation provided in scikit-learn (Pedregosa et al., 2011). To improve accuracy and control overfitting we built an ensemble of regression trees using the Extra-Trees algorithm (Geurts et al., 2006) which fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset. In our experiments 1,000 trees were trained to minimize mean square error. The regressor was trained with the following basic features: sentence length, 1-4 gram string similarity, the paraphrase probability $P(E_2|E_1)$, the language model score $P(E_1)$, and the cosine distance of the sentence vectors, as calculated by the encoder. To address the problem of rare sentences receiving low probabilities regardless of the source sentence, we create an inverse weighting by $P(E_2|E_2)$, which approximates how difficult it is to recover E_2:

$$pscore(E_2, E_1) = \frac{P(E_2|E_1)}{P(E_2|E_1) + P(E_2|E_2)} \tag{3.10}$$

[12]When considering pivots in multiple languages, we collapse subwords into words, averaging the translation probability of the subwords. This is due to different language pairs using different subword vocabularies.

Two features reflect the alignment between candidate paraphrases. We built an alignment matrix according to Equation (3.8), and used the mean of the diagonal as a feature. The second feature is the number of unaligned words which we compute by calculating hard alignments between the two paraphrases.

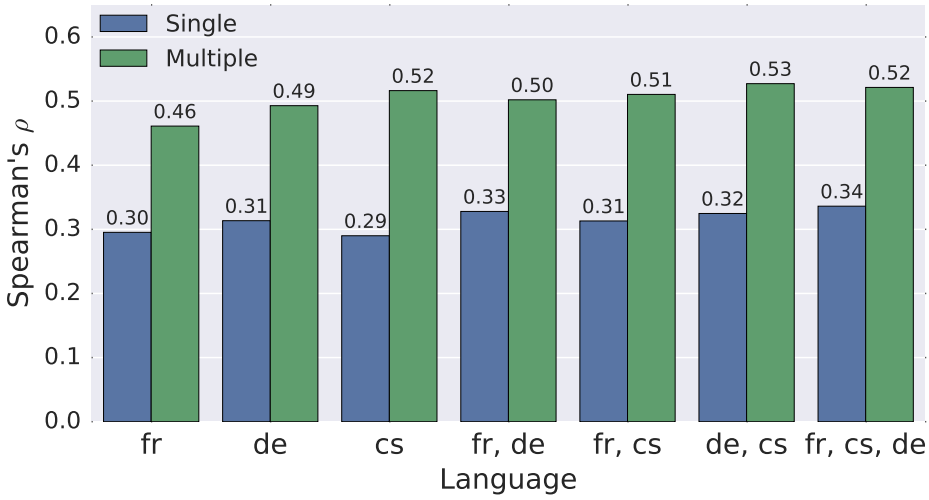

Figure 3.3: Correlation of PARANET predictions against human ratings for paraphrase pairs. Comparison using single and multiple (200) pivots, across language combinations.

Regressors varied with respect to how $P(E_2|E_1)$ was computed, keeping the string-based features the same. Equations 3.7 and (3.9) were used to calculate paraphrase probability for PARANET and PARASTAT, respectively. For both models beam search (with width set to 100) was used to generate the K-best list. For each language, the K-best list is the union of the 100-best list of E_1 and the 100-best list of E_2, giving a maximum of 200 pivot sentences per language. As set out in Pavlick et al. (2015), evaluation was done using cross validation: in each fold, we hold out 200 phrases. Table 3.4 presents results for PARANET and PARASTAT using different languages as pivots. PARANET outperforms PARASTAT across the board. Furthermore, despite using fewer features and pivot languages, it obtains a closer correspondence to human data compared to PPDB 2.0 (Pavlick et al., 2015).

3.4.4 Paraphrase Identification and Similarity

The SemEval-2015 shared task on Paraphrase and Semantic Similarity In Twitter (PIT), as discussed in Section 3.2.1, uses a training and development set of 17,790 sentence pairs and a test set of 972 sentence pairs. By design, the dataset contains colloquial

Model	PARASTAT	PARANET
fr	0.574	0.700
de	0.638	0.710
cz	0.564	0.713
de,fr	0.566	0.722
de,cz	0.640	0.731
fr,cz	0.569	0.724
fr, cz, de	0.633	**0.735**
PPDB 2.0	0.713	

Table 3.7: Correlation (Spearman ρ) of supervised models against human ratings for paraphrase pairs. Boldface indicates the best performing model.

Source	Paraphrase	Score	Binary
The Marlins just won 21 in 20 innings 20 INNINGS	The Marlins just beat the Mets 21 in 20 innings	5	✓
Apple Launches the all new MacBook Air	Also a new MacBook Air and Pro announced	4	✓
Sarah Palin is at the game are you pumped	sarah Palin at the IndyMia game	3	✗
New MacBook Air no new MacBook Pro	MacBook Air is the way to go	2	✗
The last rap battle in 8 mile though	But why were people watching the heat play when 8 mile is on	1	✗

Table 3.8: Examples of source-target pairs from PIT, with semantic scores and binary paraphrase labels.

sentences representing informal language usage and sentence pairs which are lexically similar but semantically dissimilar. The shared task consists of a (binary) paraphrase identification subtask (i.e., determine whether two sentences are paraphrases) and an optional semantic similarity task (i.e., determine the similarity between two sentences on a scale of 1–5, where 5 means completely equivalent and 1 not equivalent).

We trained a decision tree regressor on the PIT-2015 similarity dataset using the features described previously. Once trained, the decision tree regressor can be readily applied to the semantic similarity subtask. For the paraphrase detection subtask, we use the same model and apply a threshold (optimized on the validation set) such that those pairs that are over this threshold are deemed paraphrases.

Model	Similarity		Detection	
	PARASTAT	PARANET	PARASTAT	PARANET
fr	0.540	0.569	0.613	0.624
de	0.543	**0.571**	0.616	0.620
cz	0.547	0.569	0.620	0.622
de, fr	0.543	0.569	0.602	0.622
de, cz	0.540	0.570	0.606	0.615
fr, cz	0.546	0.568	0.600	**0.634**
fr, cz, de	0.539	0.568	0.596	0.620
random	0.017		0.266	
WTMF	0.350		0.536	
logistic reg	0.511		0.589	
ASOBEK	0.475		0.674	
MITRE	0.619		0.667	

Table 3.9: Paraphrase detection results (F1) and Semantic similarity results (Pearson) on the PIT-2015 data set. Boldface indicates the best performing paraphrasing model.

Tables 3.9 present our results on the two subtasks together with previously published results. We evaluate system performance on the detection task using F1 (the harmonic mean of precision and recall). For semantic similarity, system outputs are compared by Pearson correlation against human scores. The first block in the tables summarize results for PARANET and PARASTAT using different languages as pivots. The second block includes three baselines provided by the organizers of the shared task: a random baseline, a logistic regression baseline with minimal n-gram word overlap features; and a model which uses weighted matrix factorization (WTMF) and has access to dictionary definitions provided in WordNet, OntoNotes, and Wiktionary (Guo and Diab, 2012). The last two rows show the highest scoring systems: ASOBEK (Eyecioglu and Keller, 2015) ranked 1st in the identification subtask and MITRE (Zarrella et al., 2015) in the similarity subtask. ASOBEK uses knowledge-lean features based on word and character n-gram overlap, whereas MITRE is a combination of multiple systems including mixtures of string matching metrics, alignments using tweet-specific word representations, and recurrent neural networks. Since this work, the state-of-the-art for detection stands at F1 72.1, as reported in Lan et al. (2017).

As can be seen, PARANET achieves better similarity and detection scores than all

baselines and PARASTAT, for any combinations of languages. This is particularly impressive as the translation models were trained on very dissimilar data. Compared to the state of the art, PARANET fares worse; however our model was not particularly optimized on the PIT-2015 dataset which was merely used as a testbed for a fair comparison. It is thus reasonable to assume that taking into account more elaborate features (e.g., based on character embeddings) performance would be improved. The highest semantic similarity score is obtained with PARANET trained using German data. The highest scoring paraphrase detection model was PARANET trained on French and Czech data. Interestingly, using multiple pivot languages seems to offer small improvements in most cases. The languages selected as pivots in our experiments were somewhat ad-hoc. We expect to get more mileage if these are selected from the same language family or with more linguistic insight (e.g., morphologically rich vs. poor).

3.4.5 Semantic Textual Similarity

Source	Paraphrase	Score
A passenger train waiting in a station.	A passenger train sits in the station.	5
Whats in Feinstein's gun bill	#tgdn #pjnet Whats in Feinstein's bill?	4
Mall attackers used 'less is more' strategy	In Kenya, attackers used 'less is more' strategy	3
ALTHOUGH SATELLITE INTENSITY ESTIMATES FROM TAFB AND SAB ARE ONLY T.	SUBJECTIVE DVORAK INTENSITY ESTIMATES FROM TAFB AND SAB INCREASED TO KT.	2
exceed or surpass, go beyond, be greater than something	pass by, over, or under without making contact.	1
Death toll rises in Russia plane crash	Death toll rises to 39 in Italy coach crash	0

Table 3.10: STS examples with their corresponding scores.

As discussed in Section 3.2.1 the semantic textual similarity (STS) tasks requires systems to rate the degree of semantic equivalence between two text snippets. We present results on the Semeval-2015 English subtask which contains sentences from a wide range of domains, including newswire headlines, image descriptions, and answers from Q&A websites. The training/test sets consist of 11,250 and 3,000 sentence pairs, respectively. Sentence pairs are rated on a 1–5 scale, with 5 indicating they are completely equivalent.

We used the decision tree regressor with the same features described in the previous section. Again, we experimented with one, two, and three languages as pivots,

Model	PARASTAT	PARANET
fr	0.657	0.682
de	0.666	0.678
cz	0.649	0.688
de, fr	0.665	0.684
de, cz	0.662	0.687
fr, cz	0.654	**0.690**
fr, cz, de	0.658	0.689
Tokencos	0.587	
DLS@CU	0.801	

Table 3.11: Results on the Semeval-2015 semantic similarity dataset. Boldface indicates the best performing paraphrasing model.

and compared PARANET and PARASTAT directly. Our results are summarized in Table 3.11. The third block in the table presents a simple cosine-based baseline provided by the organizers (Tokencos) and the top-performing system (DLS@CU) which uses PPDB paraphrases to identify semantically-similar words and word2vec embeddings trained on approximately 2.8 billion tokens (Sultan et al., 2014).

PARANET outperforms PARASTAT on all languages and language combinations. Both systems outperform the Semeval baseline but are worse compared to the top scoring system. We see for PARANET Czech achieves the highest scores; this could be in part due to Czech non-strict word order, which allows for paraphrases that employ more movement.

3.4.6 Paraphrase Generation

Finally, we evaluated PARANET (and PARASTAT) in a paraphrase generation task. We created sentential paraphrases for three (parallel mono-lingual) datasets representative of different domains and genres: (a) the Multiple-Translation Chinese (MTC) corpus, we sampled 1,000 sentences for validation and testing, respectively (each source sentence had an average of 4 paraphrases); (b) the Jules Verne's Twenty Thousand Leagues Under the Sea novel (Leagues) as taken from the Book corpus, we sampled 500 sentences for validation/testing (each source sentence had one paraphrase); and (c) the Wikianswers corpus, we sampled 1,000 questions for validation/testing (each

question has on average 21 paraphrases). More information on these dataset can be seen in section 2.2.1 and examples can be seen in Table 3.12.

WikiAnswers	Who wrote the Winnie the Pooh books?
	Who is the author of winnie the pooh?
	What was the name of the authur of winnie the pooh?
	Who wrote the series of books for Winnie the poo?
	Who wrote the children's storybook 'Winnie the Pooh'?
	Who is poohs creator?
Leagues	"Electricity!" I exclaimed in some surprise.
	"Electricity?" I cried in surprise.
MTC	At least 12 people were killed in the battle last week
	At least 12 people lost their lives in last week's fighting
	Last week's fight took at least 12 lives
	The fighting last week killed at least 12

Table 3.12: Example paraphrase sets taken of the test sets.

In order to select the best paraphrase candidate for a given input sentence, PARASTAT was optimized on the training set using Minimum Error Training (MERT, Och and Ney (2003)). MERT integrates automatic evaluation metrics such as BLEU into the training process to achieve optimal end-to-end performance. Naively optimizing for BLEU, however, will result in a trivial paraphrasing system heavily biased towards producing identity (not rewriting the source sentence) "paraphrases". Instead we use iBLEU (described in Chapter 2 (Section 2.5)), applied to the second SMT model $P(E_2|\mathcal{F})$, which penalizes paraphrases which are similar to the source sentence and rewards those close to the target.

Dataset	Source	PARANET
Wikianswers	How many calories in a handful of strawberries?	The number of calories in a handful of strawberries.
Leagues	"Faith i should never have believed it," said Conseil.	"Faith, I never would have believed", Conseil said.
MTC	China expresses strong dissatisfaction over the Japanese leader's move this time.	China expresses a strong dissatisfaction over Japanese leader's move.

Table 3.13: Example paraphrases produced by PARANET.

PARANET relies on a relatively simple architecture which is trained end-to-end

with the objective of maximizing the likelihood of the training data. Since evaluation metrics cannot be straightforwardly integrated into this training procedure, we reranked the *k*-best paraphrases obtained from PARANET using a simple regression model which favours sentences which are dissimilar to the source. Specifically, we trained a decision tree regression model with iBLEU as the target variable using the same features described in Section 3.4.4. Examples of paraphrases generated by PARANET are shown in the Appendix A (Section A.1) and in Table 3.13.

System output was assessed automatically using iBLEU with human-written paraphrases as reference. In addition, we evaluated the generated text by eliciting human judgments via Amazon Mechanical Turk. We randomly selected 100 source sentences from each dataset and generated output with PARANET and PARASTAT (using German as a pivot). We also included a randomly selected human paraphrase as a gold standard. Workers (self-reported native English speakers) were asked to rank the three paraphrases from best to worst (ties were allowed) in order of semantic equivalence (does the paraphrase convey the same meaning as the source?) and fluency (is the description written in well-formed English?). Participants were explicitly told to give high ranks to output demonstrating a fair amount of paraphrasing and low ranks to trivial paraphrases (e.g., deletion of articles or punctuation). Instructions given to the workers can be found in Appendix A (Section A.2). We collected 5 responses per input sentence.

Model	PARASTAT	PARANET
French	22.6	29.9
German	28.2	29.5
Czech	28.0	29.1
Gold	59.9	

Table 3.14: Sentence level iBLEU scores Using plus one smoothing (Lin and Och, 2004), for PARASTAT and PARANET. Additionally we report a Human (Gold) score, which is calculated by randomly using one of the references as the generated paraphrasing, and removing this paraphrase from the reference set.

Table 3.14 summarizes our automatic results across the three datasets. We set $\alpha = 0.8$ for iBLEU as we experimentally found it offers the best trade-off between semantic equivalence and dissimilarity. As an upper-bound we also measure iBLEU amongst the gold paraphrases provided by humans. As the translation models had dif-

Model	Wikianswers	Leagues	MTC	All
PARASTAT	2.09	2.38	2.23	2.26
PARANET	**1.86**	1.94	**1.70**	**1.83**
Humans	2.17	**1.81**	2.0	2.0

Table 3.15: Mean Rankings given to paraphrases by human participants (a lower score is better).

ferent vocabularies, in part due to the creation of subwords, we only used one language as a pivot. Again, we observe that PARANET has a slight advantage over PARASTAT in terms of iBLEU, however both systems tend to paraphrase less compared to the gold standard. Table 3.15 shows the mean ranks given to these systems by human subjects. An Analysis of Variance (ANOVA) revealed a reliable effect of system type. Post-hoc Tukey tests showed that PARANET is significantly ($p < 0.01$) better than PARASTAT across datasets; PARANET is also significantly ($p < 0.01$) better than the the gold standard on both MTC and the Wikianswers dataset. We attribute this to the noisy nature of these two datasets which contain a wealth of paraphrases, a few of which are ungrammatical, or contain typos or abbreviations leading to low scores among humans.

3.5 Summary

Conclusion Within this chapter we set out to answer the question *Can bilingual data be used as a source of training data for paraphrasing?* To answer it we developed a transfer learning approach, pivoting, which uses bilingual data and NMT to perform unconstrained paraphrasing. Experimental results across several tasks (similarity prediction, paraphrase identification, and paraphrase generation) showed that NMT pivoting outperforms conventional paraphrasing methods.

Next chapter In the next chapter we expand the NMT pivoting approach, applying it to sentence compression. We remove the external reranking component from the model. Instead, we use variable disentanglement to control the output, specifically the length of the paraphrases. We show, in multiple languages, that by controlling the output length of a pivoting model we are able to perform sentence compression with bilingual data.

Chapter 4

Sentence Compression with Neural Pivoting

This chapter is based on Mallinson et al. (2018) which was published in EMNLP 2018 and answers the following questions:

- Can the output of encoder-decoder paraphrasing models be controlled?
- No supervised data exists for a specific paraphrasing task. Can bilingual data be used as a source of training data for paraphrasing?

We consider the constrained paraphrasing task, sentence compression. Within this chapter we add a controlability mechanism to the unconstrained pivoting paraphrasing approach of the previous chapter, allowing users to control the output length of the paraphrase. To do so we train translation models using variable disentanglement, where we separate the semantics of the sentence from the output length. At test time we allow the user to specify the target length, producing their ideal compression ratio. The approach is as follows; the source sentence is translated to a foreign pivot, this pivot is then translated back into the original language, while controlling for the length, thus producing a compression. By adding explicit length controls to our model, we also remove the need for the external reranker of the previous chapter. We further improve on the work of Chapter 3 by evaluating in multiple languages: English, French, and German. However, due to the lack of sentence compression data, we created and released[1] test data for these languages. Empirical results showed that a pivoting approach combined with variable disentanglement outperformed various supervised efforts.

[1]The dataset can be found at `https://github.com/Jmallins/MOSS`

4.1 Introduction

Source	The firefighters and the townspeople rallied together and continued to fight the fire.
Extractive	The firefighters and the townspeople continued to fight the fire.
Abstractive	Everyone fought the fire.
Source	Some people liked the Titanic and some people didn't.
Extractive	People liked the Titanic and people didn't.
Abstractive	Reviews for Titanic were mixed.

Table 4.1: Examples highlighting the difference between extractive and abstractive sentence compression, extractive approaches can only delete source tokens, whereas abstractive approach can perform any type of rewrite.

Sentence compression aims to produce a summary of a single sentence that retains the most important information while preserving its fluency. As mentioned in Chapter 1 (Section 1.1) there are many applications for sentence compression. Historically, research on sentence compression has focused on a simplification of the task where compressions are produced exclusively by deleting words from the input text, known as extractive sentence compression (Knight and Marcu, 2002; Riezler et al., 2003; Turner and Charniak, 2005; McDonald, 2006; Clarke and Lapata, 2008; Cohn and Lapata, 2009a), whereas more recently approaches have viewed sentence compression as a more general text rewriting problem, where all edit operations are used, known as abstractive sentence compression, (Galley and McKeown, 2007; Woodsend and Lapata, 2010; Cohn and Lapata, 2013). The examples in Table 4.1 show that abstractive approaches offer more flexibility, are more concise and produce more fluent compressions than extractive compressions.

Irrespective of how the compression task is formulated, much of the earlier work relies on syntactic information such as parse trees to help determine what to delete from a sentence. Recently there has been much interest in applying neural network models to sentence compression (Rush et al., 2015; Filippova et al., 2015; Chopra et al., 2016; Kikuchi et al., 2016; Zhou et al., 2017; Baziotis et al., 2019). Neural extractive sentence compression treats the task as a sequence labeling problem, where each word is marked to either be deleted or kept (Filippova et al., 2015). In contrast abstractive approaches use an encoder-decoder approach, as described in Chapter 2 (Section 2.2), which avoid the explicit use of syntax, which often require handwritten rules, and which would need to be constructed on a per-language basis. In this chapter we focus on encoder-decoder abstractive approaches, as it is not clear how extractive

approaches would transfer over to other paraphrasing tasks, which require a wide range of rewriting operations, and not just deletion, as demonstrated in Table 4.2.

Sentence Fusion	John trained to be an astronaut for 10 years. John went to the moon.	
Extractive	John trained to be an astronaut for 10 years went to the moon.	✗
Abstractive	John trained to be an astronaut for 10 years and went to the moon.	✓
Simplification	John adored Mary.	
Extractive	John adored Mary.	✗
Abstractive	John loved Mary.	✓

Table 4.2: Examples of an extractive approaches failing to be appropriate for the constrained paraphrasing tasks, simplification and sentience fusion.

Neural network-based approaches are data-driven, relying on the ability of recurrent architectures to learn continuous features without recourse to preprocessing tools or syntactic information (e.g., part-of-speech tags, parse trees). In order to achieve good performance they require large amounts of training data, in the region of millions of long-short sentence pairs[2]. However, there is a lack of sentence compression data, particularly for languages other than English.

This chapter addresses the paucity of data for sentence compression models. We argue that *bilingual* corpora are a rich source for learning a variety of rewrite rules across languages. Bilingual data is particularly suited for sentence compression as it inherently has length variations, for instance the French sentence *J'ai fait un voyage à Paris quand j'étais beaucoup plus jeune* could be translated to *I went on a trip to Paris when I was a lot younger* or to the shorter translation *I took a trip to Paris when I was much younger*. Baker et al. (1993) reported that translated text can be be more explicit than the original sentence, less ambiguous, syntactically simpler, and avoids repetitions, all of which can result in length variations between the source and target. Graham et al. (2019) noted that there exists a length difference between bilingual parallel sentences if the sentence appeared was original or if it was the translation. In addition length variations between different translations of the same source sentence can be found within Creutz (2018), a dataset which consisted of paraphrases extracted from bilingual subtitle, as described in Chapter 3 (Section 3.2.1).

Bilingual data and existing neural machine translation (NMT) models (Sutskever et al. 2014; Bahdanau et al. 2015) can be easily adapted to the compression task

[2]Rush et al. (2015) use approximately four million training instances and Filippova et al. (2015) two million.

through the bilingual pivoting of Chapter 3 coupled with methods which decode the output sequence to a desired length (e.g., subject to language and genre requirements). We obtain compressions by translating a source string into a foreign language and then back-translating it into the source while controlling the translation length using variable disentanglement (Kikuchi et al., 2016). Our model can be trained for any language as long as a bilingual corpus is available, and can perform arbitrary rewrites. We also demonstrate that models trained on multilingual data perform well out-of-domain.

Although our approach does not employ compression corpora for training, for evaluation purposes, we create MOSS, a new **M**ultilingual C**o**mpre**ss**ion dataset for English, French, and German. MOSS is a *parallel* corpus containing documents from the European Parliament Proceedings, TED talks, news commentaries, and the EU bookshop. Each document is written in English, French, and German, and compressed by native speakers of the respective language who process a document at a time. We obtain five compressions per document leading to 2,000 long-short sentence pairs per language. Like previous related resources (Clarke and Lapata, 2008; Cohn and Lapata, 2013; de Loupy et al., 2010) our corpus is curated manually; however it differs from Toutanova et al. (2016) in that it contains compressions for individual sentences.

There has been relatively little interest in compressing languages other than English, perhaps in part due to a lack of training data. A few models have been proposed for Japanese (Hori and Furui, 2004; Hirao et al., 2009; Harashima and Kurohashi, 2012), including a neural network model (Hasegawa et al., 2017) which repurposes Filippova and Altun's data construction method for Japanese. There is a compression corpus available for French (de Loupy et al., 2010); however, we are not aware of any modelling work on this language.

Our contributions are three-fold: (1) a novel application of bilingual pivoting to sentence compression; (2) corroborated by empirical results showing that our model scales across languages and text genres without additional supervision over and above what is available in the bilingual parallel data; (3) and the release of a multilingual, multi-reference compression corpus which can be effectively used to gain insight in the compression task and facilitate further research in compression modeling.

4.2 Compression Datasets

Within this section, we discuss publicly available sentence compression datasets. Table 4.3 provides an overview of existing datasets and Table 4.4 shows examples from these

datasets.

Dataset	Size	Method	Domain	Type	Languages
Ziff-Davis	1K	Automatic	Products	Extractive	English
Cohn and Lapata (2013)	1K	Manual	News	Abstractive	English
Clarke and Lapata (2008)	2K	Human	Spoken/News	Extractive	English
Gigaword	4M	Automatic	Headlines	Abstractive	English
Filippova and Altun (2013)	250K	Automatic	Headlines	Extractive	English
Toutanova et al. (2016)	26K	Human	Varied	Abstractive	English
de Loupy et al. (2010)	8K	Human	News	Extractive	French

Table 4.3: Overview of sentence compression *dataset*s, including the number of paraphrases (*size*), *method* in which they were collected, the *domain*, the *type* (extractive or abstractive), and the *languages* of the dataset.

Cohn and Lapata (2013) collected newspaper articles from the American News Text corpus and the British National Corpus. Annotators were asked to compress the source sentences *"while preserving the most important information and ensuring the compressed sentences remained grammatical and preserving meaning"*. Additionally, annotators were told to ensure that the resulting (compressed) document was coherent.

Ziff-Davis corpus (Knight and Marcu, 2002) contains 1000 sentences-compression pairs extracted from news articles on computer products and corresponding abstracts. Sentences from the abstracts were automatically aligned against sentences in the full article to produce source-compression pairs. This dataset has previously been used for training and evaluating sentence compression models (Knight and Marcu, 2002; Cohn and Lapata, 2009b).

Clarke and Lapata (2008) created two manual compression corpora where sentences were extracted from written (1500 sentences) and spoken (1000 sentences) sources. Annotators were asked to *"delete any words they deemed unnecessary, provided their deletions, preserved the most important information in the source sentence and the compressed sentence remained grammatical"*. Additionally, annotators could leave a sentence unchanged. This dataset has previously been used for training and evaluating sentence compression models (Clarke and Lapata, 2008; Cohn and Lapata, 2009b).

Dataset	Source	Compression
Ziff-Davis	Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.	Arborscan produces very large dxf files.
Clarke and Lapata (2008)	The aim is to give councils some control over the future growth of second homes	The aim is to give councils control over the growth of homes.
Cohn and Lapata (2013)	Bad weather dashed hopes of attempts to halt the flow during what was seen as a natural lull in the lava's momentum.	The weather prevented attempts to stop the lava flow
Gigaword	a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted, a to p judiciary official said tuesday.	iranian-american academic held in tehran released on bail
Gigaword	ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .	european mediterranean ministers gather for landmark conference by julie bradford
Filippova and Altun (2013)	Country star Sara Evans has married former University of Alabama quarterback Jay Barker.	Country star Sara Evans has married
Filippova and Altun (2013)	Intel would be building car batteries, expanding its business beyond its core strength, the company said in a statement	Intel would be building car batteries
Toutanova et al. (2016)	Think of all the ways everyone in your household will benefit from your membership N/A in Audubon.	Imagine how your household will benefit from your Audubon membership.
Toutanova et al. (2016)	Will the administration live up to its environmental promises? Can we save the last of our ancient forests from the chainsaw?	Can the administration keep its promises? Can we save the last of our forests from loss?
de Loupy et al. (2010)	Les banques françaises n'ont pas publié de chiffres précis sur leur exposition à Lehman Brothers mais ont diffusé des messages au marché laissant entendre clairement que celle-ci était limitée et bénéficiait, pour ce qui est du risque de contrepartie sur des transactions de marché, de sûretés sous forme de collatéral.	Les banques françaises n'ont pas publié de chiffres sur leur exposition à Lehman Brothers mais ont diffusé des messages laissant entendre que celle-ci était limitée et bénéficiait de sûretés.

Table 4.4: Examples from sentence compression datasets.

Gigaword (Rush et al., 2015) Since large scale compression datasets do not occur naturally, they must be somehow approximated. Rush et al. (2015) pair headlines with the first sentence of a news article, under the assumption that the headline will be shorter, as headlines-first sentence pairs have been shown to be semantically similar (Dorr et al., 2003). Using the Annotated English Gigaword corpus (Napoles et al., 2011), they create 4 million sentence-compression pairs. Gigaword has been used extensively for the training and evaluation of sentence compression systems (See et al., 2017; Paulus et al., 2018; Kouris et al., 2019). Although large, headlines are syntactically quite different from *normal* sentences. For example, they may not have a main verb, they may not contains determiners or not appear as full sentences, limiting their use as a general purpose simplification system (Filippova and Altun, 2013).

Whilst this approach could be used to construct sentence compression datasets for many languages, the training corpus construction process must be repeated and reconfigured for new languages and domains (e.g., many headline-first sentence pairs are spurious and need to be filtered using language and domain specific heuristics).

Filippova and Altun (2013) Similarly to Gigaword, Filippova and Altun (2013) extract first sentences and headlines. However, they syntactically transform the headline ensuring it is extractive and a more natural sentence. They sampled sentence-compression pairs from the dataset, and found them to be grammatical and meaning-preserving. This dataset has been used for training and evaluating sentence compression models (Cífka et al., 2018; Filippova and Altun, 2013). This approach has also been adapted for Japanese (Hasegawa et al., 2017), however the resulting dataset has not been publicly released.

Toutanova et al. (2016) crowdsourced a large compression corpus which contains manual compressions for single and multiple sentences (26,000 source-compression pairs). Source sentences were taken from four domains: Newswire, Letters, Journal, and Non-fiction. The sentences were first compressed to a minimum reduction of 25% by five annotators, producing five compressions. A separate set of annotators were then used to remove bad compressions. This dataset has been used for both evaluation and training of sentence compression models (Toutanova et al., 2016; Mallinson et al., 2020b). Toutanova et al. (2016) also used this dataset to correlate sentence compression evaluation metrics and human judgements.

de Loupy et al. (2010) created a manual sentence compression dataset for French. News articles from 20 topics were compressed sentence-by-sentence by four annotators. Annotators could only delete words. As far as we are aware this corpus has not been used for any sentence compression tasks.

From Table 4.3, we note that the only large scale sentence compression datasets are automatically constructed for headline-first sentence pairs and only exist in English, motivating the need for alternative approaches to supervised learning.

4.3 Neural Pivot Compression

In our pivot-based sentence compression model an input sequence is first translated into a foreign language, and then back into the source language. We use the neural pivoting approach of Chapter 3. However, unlike Chapter 3, we use variable disentanglement to parameterize our translation models with a length feature, which allows us to produce compressed output. In the next section we define two variants, either performing compression in one step or alternatively in two steps which affords more flexibility.

4.3.1 NMT Background

We now briefly describe the relevant parts of an encoder-decoder model. For a more detailed description of RNN encoder-decoder models we refer the reader to Chapter 2 (Section 2.2). An encoder takes in a source $x = (x_1, ..., x_{T_x})$ of length T_x and the decoder generates a target sequence $y = (y_1, ..., y_{T_y})$ of length T_y. Let h_i be the hidden state of the source symbol at position i, obtained by concatenating the forward and backward encoder RNN hidden states, $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$. We deviate from previous work (Bahdanau et al., 2015; Sutskever et al., 2014) in that we initialize the decoder (s_0) with the average of the hidden states:

$$s_0 = tanh(W_{init} \frac{\sum_{i=1}^{T_x} h_i}{T_x}) \tag{4.1}$$

where W_{init} is a learnt parameter. Our decoder is a conditional recurrent neural network, specifically a Gated Recurrent Unit (GRU, (Cho et al., 2014)).

4.3.2 Length Control

To be able to produce compressed sentences, we use variable disentanglement to separate the semantics of the source sentence and the length of the target sentence. The model is parameterized with a length vector which informs the model of the target output length. Our approach is similar to the *LenInit* model of Kikuchi et al. (2016), who also parameterize a model with length information, using a scaling function, as we will explain later. However we differ from Kikuchi et al. (2016) in two respects, (1) we use a GRU instead of an LSTM and (2) we apply this technique to bilingual data, not sentence compression data. The hidden state of the decoder consists of the average of the encoder's hidden states but also a length vector LV, a learnt parameter, which is scaled by the desired target length $T_{y'}$. We therefore rewrite Equation (4.1) as follows:

$$s_0 = tanh\left(W_{init}\left[\frac{\sum_{i=1}^{T_x} h_i}{T_x}; LV \cdot T'_y\right]\right) \tag{4.2}$$

We now define our NMT model as:

$$P(y|x, T'_y) = \prod_{j}^{T_y} P(y_j|y_{<j}, x, T'_y) \tag{4.3}$$

During training, the target length is set to $T'_y = T_y$, i.e. the model is given the true output length, which We consider this a form of soft variable disentanglement.

At test time, where we do not know the gold target length, the target length generally varies according to the domain, genre, and language at hand. We determine the target length experimentally based on a small validation set.

4.3.3 Pivoting

Chapter 3 showed how pivoting could successfully be used to perform unconstrained paraphrasing, where we defined the probability of generating a paraphrase E_2 from E_1 through a k-best list of intermediate pivots $\mathcal{F}$:

$$P(E_2|E_1) \approx \prod_{j}^{T_{E_2}} \Big(\sum_{i=1}^{K} P(\mathcal{F}_i|E_1) P(E_{2j}|E_{2<j}, \mathcal{F}_i) \Big) \tag{4.4}$$

To ensure the model produces compressed output, we extend the pivoting approach in two ways, dual step and single step compression, a compairson can be seen in Figure 4.1. In *single step* compression, one of the translation models is parameterized with

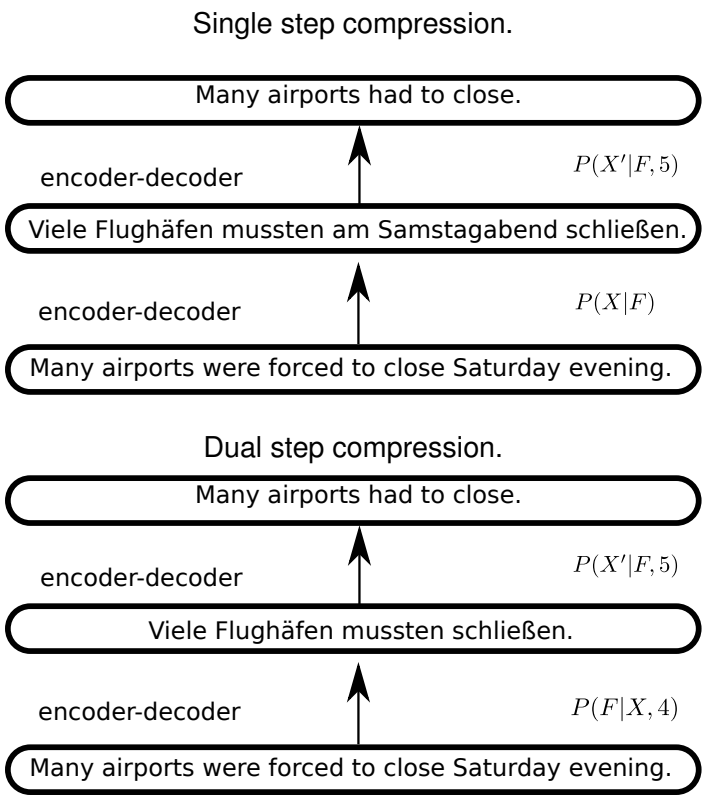

Figure 4.1: Single and dual step compression.

length information:

$$P(E_2|E_1,T'_{E_2}) \approx \prod_{j}^{T_{E_2}} \big(\sum_{i=1}^{K} P(\mathcal{F}_i|E_1) P(E_{2j}|E_{2<j},T'_{E_2},\mathcal{F}_i) \big) \tag{4.5}$$

In this approach we parameterize the final translation model with length information. Whilst we could have parameterized the first model, this would only allow us to control the intermediate pivot translation, which would have no guarantees on the final output. In *dual-step* compression, we parameterize both translation models with length information:

$$P(E_2|E_1,T'_{E_2},T'_{\mathcal{F}}) \approx \prod_{j}^{T_{E_2}} \big(\sum_{i=1}^{K} P(\mathcal{F}_i|E_1,T'_{\mathcal{F}}) P(E_{2j}|y_{<j},T'_{E_2},\mathcal{F}_i) \big) \tag{4.6}$$

We use the heuristic below to set the length of the intermediate translation:

$$T'_{\mathcal{F}} = T'_{E_2} + \alpha(T_{E_1} - T'_{E_2}) \tag{4.7}$$

where T_{E_1} is the length of E_1 and a high value for α results in the majority of the compression happening in the first translation operation. The α value is determined experimentally based on a small validation set.

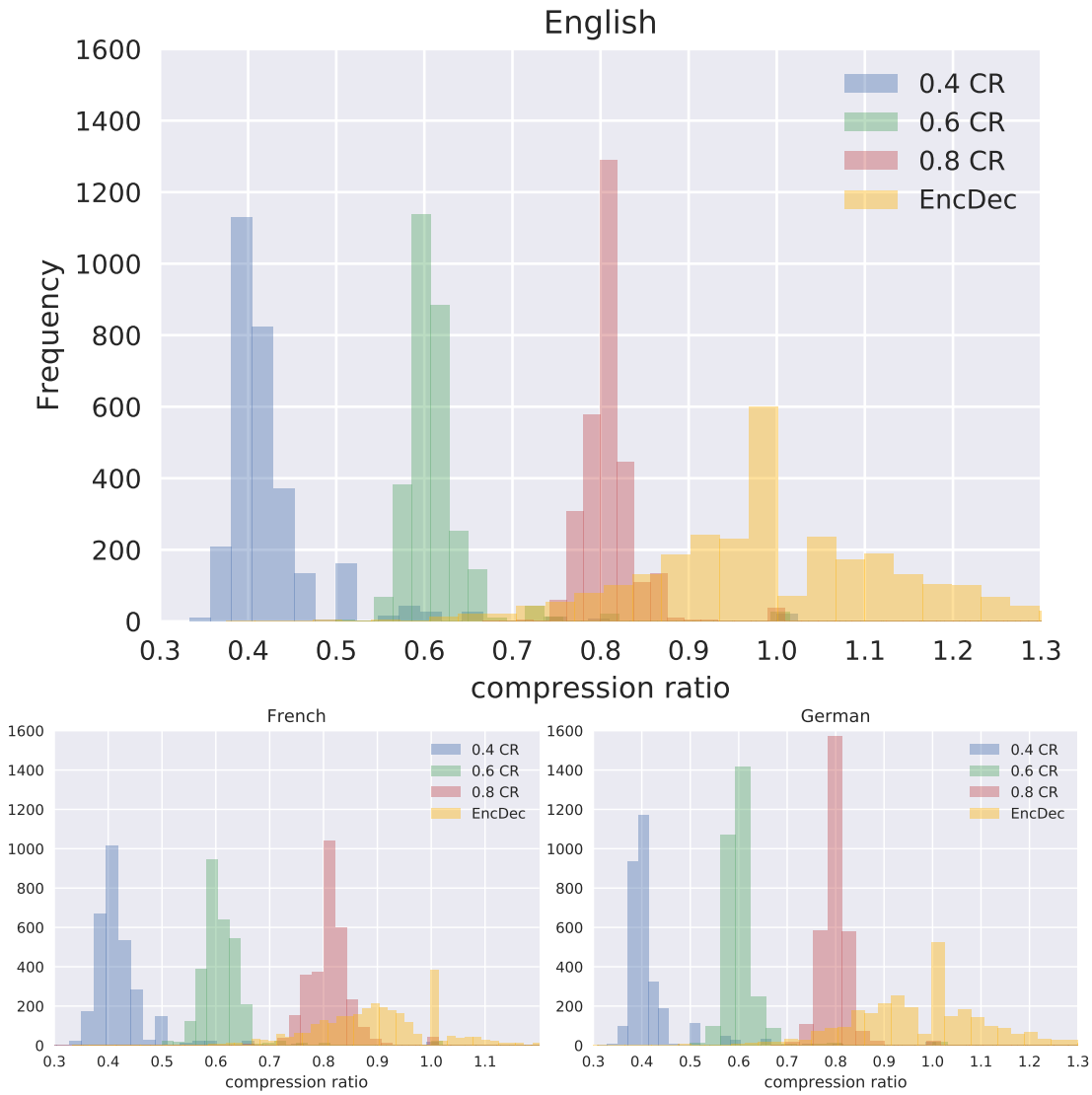

Figure 4.2: Histograms of output lengths at three compression rates (CR) compared to a vanilla encoder-decoder system which does not manipulate output length. German is used as pivot for English, and English as pivot for French and German.

In Figure 4.2 we illustrate how the pivot-based model sketched above can successfully control the output of the generated compressions. We show the output of a *single-step* compression model on three languages initialized with varying compression rates, which refers to the percentage of words retained from the source sentence in the compression. (See Section 4.5 for details on how the models were trained and tested). The compression rate (CR) is used to determine the length parameter of Equation (4.3):

$$T'_{E_2} = T_{E_1} \cdot CR \tag{4.8}$$

Figure 4.1 shows how the output length varies compared to a vanilla encoder-

decoder system which uses pivoting to backtranslate the source language as explained in Chapter 3. We can see that the majority of sentences are generated with length close to the desired compression rate. However, we found that dual-compression performs better when the system is expected to drastically compress the source sentence (e.g., in a headline generation task). Imposing a high compression ratio from the start tends to produce unintelligible text. The model attempts to reduce the length of the source at all costs, even at the expense of being semantically faithful to the input. Performing two moderate compressions in succession reduces both length and content conservatively and as a result produces more meaningful text.

4.4 The MOSS Dataset

For evaluation purposes, we created a multilingual sentence compression corpus in English, German, and French. The corpus was collated from existing document and sentence aligned multilingual datasets which vary both in terms of topic and genre. We sampled five documents each from:

1. Europarl, the European Parliament Proceedings Parallel Corpus (Koehn, 2005), has been used extensively in machine translation research. It contains the minutes of the European parliament and is a spoken corpus of formulaic nature; speakers take part in debating various issues concerning EU policy (e.g., taxation, environment).
2. The TED parallel Corpus (Cettolo et al., 2012) contains transcripts in multiple languages of short talks devoted to "spreading powerful ideas on a variety of topics ranging from science to business and global issues".
3. The EU bookshop corpus (Skadiņš et al., 2014) contains publications from European institutions covering a variety of topics such as refugees, gender equality, and travel.
4. The News Commentary Parallel Corpus contains articles downloaded from Project Syndicate, an international media organization that publishes commentary on global topics (e.g., economics, world affairs).

We obtained compressions using the Crowdflower platform[3]. Crowdworkers were given instructions that explained the task and defined sentence compression with the

[3]Now known as Appen `http://www.appen.com`

English	French	German
On the very day that the earthquake struck, the European Council asked the High Representative and the Commission to mobilise all appropriate assistance.	Le jour même du tremblement de terre, le Conseil européen a demandé à la haute représentante et à la Commission de mobiliser toute l'aide appropriée.	Am gleichen Tag, an dem das Erdbeben ausbrach, ersuchte der Europäische Rat die Hohe Vertreterin und die Kommission um die Mobilisierung aller angemessenen Hilfe.
Assistance was mobilized on the very day of the earthquake.	*Le Conseil européen a demandé à la haute représentante et à la Commission de mobiliser l'aide.*	*Europa erbrachte Hilfe noch am selben Tag.*
We're at a tipping point in human history, a species poised between gaining the stars and losing the planet we call home.	L'histoire humaine est à un tournant. Notre espèce hésite à toucher les étoiles ou à perdre la planète qui est la sienne.	Wir stehen vor einem historischen Wendepunkt: zwischen dem Griff nach den Sternen und dem Verlust unseres Heimatplaneten.
We're at tipping point in human history, poised between gaining the stars and losing the Earth.	*L'humanité est à un tourt. Notre espèce a envie des étoiles ou à perdre sa planète.*	*Wir sind vor einem historischen Wendepunkt: zwischen dem Griff nach Sternen und Verlust unseres Planeten.*
Surveys undertaken by the World Bank in developing countries show that when poor people are asked to name the three most important concerns they face good health is always mentioned.	Les enquêtes menées par la Banque mondiale dans les pays en développement montrent que, quand on demande aux populations pauvres de nommer les trois défis les plus importants qu'ils rencontrent, leur "bonne santé" fait toujours partie de cette liste.	Umfragen der Weltbank in Entwicklungsländern zeigen, wenn man Arme nach den drei wichtigsten Anliegen fragt, die sie beschäftigen, wird "Gesundheit" immer genannt.
World Bank surveys in developing countries show poor people always name good health as an important concern.	*Quand on demande aux populations pauvres de nommer les trois défis les plus importants qu'ils rencontrent, leur "bonne santé" fait toujours partie de la liste.*	*Umfragen in Entwicklungsländern zeigen, dass bei Armen das wichtigste Anliegen Gesundheit ist.*

Table 4.5: Examples of compressions from the MOSS corpus. Sentences shown (in order of appearance) from Europarl, TED, and News Commentary corpora.

aid of examples. They were asked to compress while preserving the most important information, ensuring the sentences remained grammatical and meaning preserving. Annotators were encouraged to use any rewriting operations that seemed appropriate, e.g., to delete words, add new words, substitute them, or reorder them. Annotation proceeded on a document-by-document basis, line-by-line. Crowdworkers compressed the first twenty lines of each document and we elicited five compressions per document. Example compressions are shown in Table 4.5.

English	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	27.29	17.48	0.64	0.45	0.11	10.66	1.72	0.45
TED	10.64	8.12	0.76	0.34	0.02	2.57	1.02	0.15
News	19.17	14.22	0.74	0.38	0.14	5.39	1.91	0.43
Books	20.52	16.12	0.78	0.32	0.11	4.50	1.54	0.38
All	19.41	13.99	0.73	0.37	0.10	5.78	1.55	0.35

French	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	29.40	23.48	0.79	0.43	0.83	7.04	2.90	0.38
TED	6.16	5.11	0.83	0.44	0.03	1.35	1.33	0.04
News	27.52	21.95	0.79	0.37	0.14	6.37	3.06	0.50
Books	22.32	18.48	0.83	0.36	0.52	4.21	1.79	0.20
All	21.35	17.26	0.81	0.40	0.38	4.74	2.27	0.28

German	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	24.53	16.87	0.69	0.38	0.10	8.70	1.14	0.18
TED	5.36	4.55	0.85	0.24	0.02	0.76	0.53	0.10
News	23.48	16.49	0.70	0.45	0.13	8.39	2.15	0.47
Books	19.83	14.97	0.75	0.50	0.52	5.66	2.89	0.34
All	18.30	13.22	0.75	0.39	0.19	5.88	1.68	0.27

Table 4.6: MOSS statistics across corpora and languages: length of source (SL) and target sentence (TL), compression rate (CR), TER scores, and the average (per sequence) number of insertions (Ins), deletions (Del), substitutions (Sub), and shifts (Shft).

Table 4.6 presents various statistics on our corpus. Europarl contains the longest

sentences across languages (see column SL), TED contains the shortest sentences, while the other two corpora are somewhere in between. We also observe that crowdworkers compress the least when it comes to TED (see column CR), which is not surprising given the brevity of the utterances. Overall, French speakers seem more conservative when shortening sentences compared to English and German. In general, compression rates are genre-dependent; they range from 0.64 (for English Europarl) to 0.85 (for German TED). We also examined the degree to which crowdworkers paraphrase the source sentence using Translation Edit Rate (TER; Snover et al., 2006), a measure commonly used to automatically evaluate the quality of machine translation output, where a higher score means the output is more different, further described in Chapter 2 (Section 2.2). We used TER to compute the (average) number of edits required to change a long sentence to shorter output. We also report the number of edits by type, i.e., the number of insertions, substitutions, deletions, and shifts needed (on average) to convert long to short sentences. We observe that crowdworkers perform a fair amount of rewriting across corpora and languages. The most frequent rewrite operations are deletions followed by substitutions, shifts, and insertions.

4.5 Experimental Setup

Neural Machine Translation Training Nematus[4] (Sennrich et al., 2017) was used as the machine translation system for all our experiments. We generally used the default settings and training procedures as specified within Nematus. All networks have a hidden layer size of 1,000, and an embedding layer size of 512. In addition, layer normalization (Ba et al., 2016) was used. During training we used ADAM (Kingma and Ba, 2015), a minibatch size of 80, and the training set was reshuffled between epochs. We also employed early stopping using BLEU on their respective WMT validation set. We used up to four encoder-decoder NMT models in our experiments. German training/test data was taken from the WMT16 shared task and French from the WMT14 shared task. The training data was 4.2 million and 39 million sentence pairs for en-de, and en-fr, respectively. We also used back-translated monolingual training data, from the news domain, (Sennrich et al., 2016c) in training for the German systems. BLEU

[4]The Theano branch was used.
[5](Zhang et al., 2019)
[6]`https://www.deepl.com/press.html`
[7](Sennrich et al., 2016b)
[8](Macháček and Bojar, 2013)

Languages	US	SOTA
English → French	27.03	40.58 [5]
French → English	29.14	45.90 [6]
English → German	29.34	34.20 [7]
German → English	31.19	40.20 [8]

Table 4.7: BLEU translation scores for our pivot based models (US) and the current state-of-the-art (SOTA).

scores[9] on the respective WMT test sets for our models and the current state-of-the-art can be seen in Table 4.7. The data was pre-processed using standard scripts found in MOSES (Koehn et al., 2007). Rare words were split into sub-word units, using byte pair encoding (BPE; Sennrich et al. 2016d). The BPE operations are shared between language directions.

We experimented with various model variants using one or multiple pivots. The compression rate (see Equation (4.3)) was tuned experimentally on the validation set which consisted of one document from each domain (20 source sentences; 100 compression-pairs). Compression rates varied from 0.55 to 0.85 and were broadly comparable to those shown in Table 4.6.

Comparison Systems We compared our model against a GRU sequence-to-sequence attention-based model (seq2seq). This model was trained on a monolingual dataset extracted from the Annotated English Gigaword corpus (Napoles et al., 2011), as described in Section 4.2. The dataset consists of approximately 4 million pairs of the first sentence from each source document and its headline. We also trained *LenInit* (Kikuchi et al., 2016) on the same corpus which is conceptually similar to sequence-to-sequence model but additionally controls the output length using a length embedding vector (as described in Section 4.3.2).[10] Unfortunately, we could not train these models for French or German, since there are no monolingual sentence compression datasets available at a similar scale.

An obvious workaround is to translate Gigaword into French and German and then train compression models on the translated data, where source and target sentences are independently translated. As the quality of the translation is relatively poor, we also

[9] BLEU scores were calculated using mteval-v13a.pl.

[10] We used our own implementation of LenInit which on DUC-2004 obtained ROUGE scores similar to those published in Kikuchi et al. (2016).

propose a pivot-based method, where at test time we translated German or French into English, compressed it with seq2seq and LenInit trained on the Gigaword corpus, and then translated the compressions back to French or German. As such the probability of the compression is:

$$P(F_2|F_1) \approx P(F_2|E')P(E'|E)P(E|F_1) \tag{4.9}$$

where $P(E|F_1)$ is the translation probability from foreign sentence (F) into English (E), $P(E'|E)$ is the probability of generating the compression (E') and $P(F_2|E')$ is the translation probability for translating the compression back into the original language. For simplicity, instead of using multiple pivots, we used a single translation and a single compression.

Finally, we include a prefix (*Pfix*) baseline which does not perform any rewriting but simply truncates the source sentence so that it matches the compression ratio of the validation set.

4.6 Results

	RS-R	D2-R	R2-F1
Pfix	45.38	47.57	33.67
seq2seq	18.29	23.55	15.60
LenInit	17.90	19.64	11.18
$SP_{\mathcal{L},de}$	**34.60**	**37.97**	**22.67**
$SP_{\mathcal{L},fr}$	27.42	32.34	19.29
$MP_{\mathcal{L},de}$	28.71	34.70	19.06
$MP_{\mathcal{L},fr}$	20.74	27.50	13.89
Gold	76.60	71.68	42.89

English (a)

	RS-R	D2-R	R2-F1	RS-R	D2-R	R2-F1
Pfix	60.33	62.44	53.37	56.28	50.78	45.84
seq2seq	13.84	18.00	9.74	5.72	12.95	5.21
$seq2seq_{en}$	16.39	22.08	13.17	9.43	14.78	6.79
LenInit	9.91	14.52	8.08	4.91	11.77	2.87
$LenInit_{en}$	20.08	24.41	13.06	13.19	18.67	7.65
$SP_{\mathcal{L},en}$	**43.38**	**46.17**	**35.07**	**38.19**	**38.54**	**31.15**
$MP_{\mathcal{L},en}$	31.55	37.88	26.59	23.62	29.13	17.36
Gold	74.42	80.00	52.13	76.01	77.48	48.36

French (b) German (c)

Table 4.8: Automatic evaluation on MOSS; S/MP: single/multiple pivot models; $\mathcal{L}$: length parameter; pivot languages: English (en), French (fr), German (de); seq2seq and LenInit (Kikuchi et al., 2016) are sequence-to-sequence models trained on Gigaword; Gold is inter-annotator agreement.

MOSS Evaluation We assessed model performance using three automatic metrics which represent different aspects of the compression task and have been found to cor-

relate well with human judgments (Toutanova et al., 2016; Clarke and Lapata, 2006). These include a recall metric based on skip bi-grams, any pair of words in a sequence allowing for gaps of size four[11] (RS-R); a recall metric based on bi-grams of dependency tree triples (D2-R); and bi-gram ROUGE (R2-F1). We used the Stanford neural network parser (Chen and Manning, 2014) to obtain dependency triples. RS-R and D2-R have been shown to strongly correlate with combined grammar and meaning human judgements, and meaning human judgements (Toutanova et al., 2016). R2-F1 correlates strongly with grammar human judgements. More information on these evaluation metrics can be seen in Chapter 2 (Section 2.3).

Table 4.8(a) reports results on English with a model which controls the output length ($\mathcal{L}$) and uses either a single pivot (SP; $K = 1$) or multiple pivots (MP; $K = 10$). We experimented with French (fr) or German (de) as pivot languages. All pivot-based models perform compression in a single step (see Section 4.3.3). As can be seen, models which use a single pivot are better than those using multiple ones (German is better than French; see SP_{de} vs SP_{fr}). We found that the use of multiple pivots resulted in more accurate semantics at the cost of compression.

Overall, pivot-based models outperform seq2seq and LenInit. This is perhaps to be expected since these models are tested on out-of-domain data with different vocabulary and writing conventions; MOSS does not contain any newspaper articles. Unfortunately, it is not possible to train seq2seq and LenInt on in-domain data as compression data only exists for the headlines-first sentences pairs. As an upper bound, we also report how well humans agree with each other, treating one (randomly selected) reference as system output and computing how it agrees with the rest (row Gold in Table 4.8). All models lag significantly behind human performance on this task.

Tables 4.8(b) and 4.8(c) report results on French and German, respectively. For these languages, we obtained best results with English as pivot, using a single-step compression model. Seq2seq and LenInit perform poorly when trained directly on translations of Gigaword into French and German; their performance improves considerably when they are trained on Gigaword and used to compress English translations of French or German (seq2seq_{en}, LenInit_{en}). Again, we observe that our models ($\text{SP}_{\mathcal{L},en}$, $\text{MP}_{\mathcal{L},en}$) outperform the comparison systems across all metrics and that using a single pivot yields better compressions. Example compressions are given in Table 4.9 where we show output produced by seq2seq and SP for each language (see the Appendix B.1 for more examples). Finally, notice that automatic scores for the prefix baseline across

[11] We add a begin-of-sentence marker at the start of the candidate and reference sentences.

languages are misleadingly high, since it simply repeats the source sentence up to a fixed length without performing any rewriting.

	English	French	German
SEQ2SEQ	Europe urged to help quake victims.	Le Conseil Européen demande une aide pour les victimes du tremblement de terre.	Europäischer Rat sucht Hilfen für Quiz-Opfer.
SP	The European Council called on the High Representative and the Commission to mobilise all appropriate assistance.	Le Conseil Européen a demandé au Haut Représentant et à la Commission de mobiliser l'assistance.	Am selben Tag forderte der Europäische Rat die Hohe Vertreterin und die Kommission auf, jede Hilfe.
SEQ2SEQ	Advance for Sunday July a new look at the world.	Un tournant pour le tournant.	Die Stars der Stars und die Stars.
SP	We are at a turning point in human history and losing the planet we call home.	L'histoire de l'humanité est à la croisée des chemins et de l'histoire.	Zwischen dem Griff der Sterne und dem Verlust unseres Planeten stehen wir vor.
SEQ2SEQ	Poor people ask to name the three most important concerns.	Les enquêtes de la Banque mondiale révèlent que la santé fait toujours partie de la liste.	Weltbank-Umfragen zeigen arme Menschen in Entwicklungsländern.
SP	Polls conducted by the World Bank show that when poor people are asked to mention the three main concerns.	Les enquêtes menées par la Banque mondiale dans les pays en développement montrent que, lorsqu'on demande aux pauvres de nommer les trois plus grands éfis.	Wenn man die Armen nach den drei Hauptanliegen fragt, werden sie gefordert.

Table 4.9: System output for the example source sentences in Table 4.5.

We also elicited human judgments through the Crowdflower platform. We asked crowdworkers to rate the grammaticality of the target compressions and whether they preserved the most important information from the source. For both questions, they used a five-point rating scale where a high number indicates better performance. Full instructions can be found in Appendix B (Section B.2). We randomly selected 25 sen-

Models	English			French			German		
	Imp	Gram	Avg	Imp	Gram	Avg	Imp	Gram	Avg
Pfix	2.72	2.98	2.85	2.73	2.89	2.80	3.17	2.96	3.06
LenInit	2.51	3.0	2.75	1.82	2.62	2.22	2.10	3.25	2.67
$SP_{\mathcal{L}}$	**3.27**	**3.69**	**3.48**	**3.48**	**3.60**	**3.54**	**3.30**	**3.87**	**3.59**
Ref	3.47	3.80	3.63	4.05	4.14	4.10	3.97	4.26	4.10

Table 4.10: Mean ratings elicited by humans on MOSS; Avg is the average rating of grammaticality and importance.

tences from each corpus from the test portion of MOSS, i.e., 100 long-short sentence pairs per language. We compared compressions generated by our model ($SP_{\mathcal{L}}$), with seq2seq models for the three languages, the prefix baseline, and (randomly selected) gold-standard reference (Ref) compressions from MOSS. All systems used the length parameter to allow comparisons with approximately the *same* compression rates. We collected five ratings per compression. Our results are summarized in Table 4.10. We show mean ratings for grammaticality (Gram), importance (Imp) and their combination (column Avg). Across languages our model (SP_L) significantly ($p < 0.05$) outperforms comparison systems (Pfix, seq2seq) on both dimensions of grammaticality and importance (significance tests were performed using a student t-test). All systems are significantly worse ($p < 0.05$) than the human reference compressions.

	SL	TL	CR	TER	Ins	Del	Sub	Shft
English	19.41	12.31	0.63	0.65	0.10	6.68	2.14	0.44
French	21.35	14.98	0.70	0.67	0.29	5.71	3.36	0.61
German	18.30	12.51	0.68	0.67	0.16	6.38	2.94	0.50

Table 4.11: Statistics of model output ($SP_{\mathcal{L}}$) on MOSS (aggregated across domains): length of source (SL) and target (TL), compression rate (CR), TER scores, and the average number of insertions (Ins), deletions (Del), substitutions (Sub), and shifts (Shft).

Finally, in Table 4.11 we analyze the output of our best model ($SP_{\mathcal{L}}$) using the same statistics we applied to the human compressions (see Table 4.6). As can be seen, the model generally compresses more aggressively and applies more edits than the crowdworkers (both compression rates and TER scores are higher for all three languages). Although the rate of insertions and deletions is similar to humans, substitutions and

shifts happen to a greater extent for our model, indicating that it performs a good amount of paraphrasing.

DUC-2004 Evaluation Besides MOSS, we evaluated our model on the benchmark DUC-2004 task-1 dataset. In this task, the aim is to create a very short summary (75 bytes) for a document. The evaluation set consists of 500 source documents (from the New York Times and Associated Press Wire services) each paired with four human-written (reference) summaries. We follow previous work (Rush et al., 2015; Chopra et al., 2016) in compressing the first sentence of the document and presenting this as the summary. To make the evaluation unbiased to length, the output of all systems is cut off after 75-characters and no bonus is given for shorter summaries.

Models	RS-R	D2-R	R2-F1	R1-R	R2-R	RL-R
Pfix	15.25	15.59	5.38	20.42	5.86	18.07
SP$_{\mathcal{L},de}$	**12.93**	**13.89**	**4.97**	**20.70**	**5.35**	**18.35**
SP$_{\mathcal{L},fr}$	12.06	12.18	4.42	19.77	4.75	17.40
MP$_{\mathcal{L},fr}$	10.38	11.85	3.70	18.67	4.03	16.20
MP$_{\mathcal{L},de}$	11.06	13.26	4.30	19.10	4.69	16.84
Gold	16.41	18.12	7.72	26.95	7.72	22.79
seq2seq				25.03	8.40	22.35
ABS Rush et al. (2015)				26.55	7.06	22.05
ABS+ Rush et al. (2015)				28.18	8.49	23.81
RAS Chopra et al. (2016)				28.97	8.26	24.06
LenInit[12] Kikuchi et al. (2016)				25.87	8.27	23.24
LenEmb Kikuchi et al. (2016)				26.73	8.40	23.88

Table 4.12: DUC-2004 results (75 char length cap); results for comparison systems are taken from their respective papers.

Our results are shown in Table 4.12. To compare with existing methods, we also report ROUGE (Lin, 2004b) unigram and bigram overlap (Lin, 2004b) and the longest common subsequence (ROUGE-L)[13]. We employed a dual-step compression model (see Section 4.3) because preliminary experiments showed that it was superior to single-stage variants. We compared single and multiple pivot models against

[12]Our LenInit implementation obtains R1-R 29.26, R2-R 9.56, and RL-R 25.70

[13]We used ROUGE version 1.5.5 with the original DUC-2004 ROUGE parameters.

Source	King Norodom Sihanouk has declined requests to chair a summit of Cambodia's top political leaders, saying the meeting would not bring any progress in deadlocked negotiations to form a government.
$SP_{\mathcal{L},de}$	King Norodom Sihanouk has refused to chair Cambodia summit.
Gold	Sihanouk refuses to chair Cambodian political summit at home or abroad.
Source	Cambodia's ruling party responded Tuesday to criticisms of its leader in the U.S. Congress with a lengthy defense of strongman Hun Sen's human rights record.
$SP_{\mathcal{L},de}$	Cambodia's ruling party responded Tuesday to criticism of its leader in the US.
Gold	Cambodian party defends leader Hun Sen against criticism of U.S. House.
Source	The Swiss government has ordered no investigation of possible bank accounts belonging to former Chilean dictator Augusto Pinochet, a spokesman said Wednesday.
$SP_{\mathcal{L},de}$	Swiss government ordered no inquiry into possible bank accounts of former Chilean dictator Augusto.
Gold	Switzerland joins charges against Pinochet but avoids bank probe.

Table 4.13: System output for DUC-2004.

seq2seq and existing compression models, ABS and ABS+ (Rush et al., 2015), two encoder-decoder models trained on the English Gigaword. ABS+ applies minimum error rate (MERT) training as a copying mechanism. LenEmb and LenInit include a length parameter (Kikuchi et al., 2016), whereas RAS uses a convolutional based recurrent neural network architecture. Since the completion of this work there has been continued interest in this test set and the state-of-the-art achieves a Rouge-1: 32.57, Rouge-2: 11.63, and Rouge-L: 28.24 (Takase and Kobayashi, 2020), which combines a Transformer and an approach which reduces the number parameters needed for word embeddings. We also report how well DUC-2004 abstractors agree with each other (row Gold in Table 4.12). Example compressions are given in Table 4.13, where we show output produced by $SP_{\mathcal{L},de}$ and a human reference (see the Appendix B (Section B.1) material for further examples).

Using automatic metrics we see that our model generally performs worse compared to these systems and that German is the best pivot for English. Although the objective of this chapter is not to obtain state-of-the-art scores on this evaluation set, it is interesting to see that our model is able to compress out of domain. We do not have access to headline-first sentence pairs, while all comparison systems do. We also elicited human judgments on the compressions of 100 lead sentences whose documents were randomly selected from the DUC-2004 test set. We compared the prefix baseline, our

model ($SP_{\mathcal{L},de}$), ABS+ (Rush et al., 2015), LenEmb (Kikuchi et al., 2016), Topiary (Zajic et al., 2004), and a randomly selected reference. Topiary came top in almost all measures in the DUC-2004 evaluation; it first compresses the lead sentence using linguistically motivated heuristics and then enhances it with topic keywords. Crowdworkers rated grammaticality and importance, using a five-point scale; we collected five ratings per compression.

Models	Grammaticality	Importance	Average
Pfix	3.03	2.93	2.98
$SP_{\mathcal{L},de}$	3.37	3.22	3.29
Topiary	3.05	3.15	3.10
ABS+	**3.67**	**3.23**	**3.45**
LenEmb	3.14	3.08	3.09
Ref	3.62	3.27	3.45

Table 4.14: Mean ratings elicited by humans on DUC-2004; Avg is the average rating of grammaticality and importance.

As shown in Table 4.14 ABS+ has the lead with our system following. In terms of grammaticality, ABS+ and $SP_{\mathcal{L},de}$ are not significantly different from the gold standard nor from each other (Pfix, Topiary, and LenEmb are significantly worse than Gold; $p < 0.05$). In terms of importance, pairwise differences between systems and the gold standard are not significant. Overall, we observe that $SP_{\mathcal{L},de}$ performs comparably to ABS+ even though it was not trained on any compression specific data. Inspection of the system output reveals that our model performs more paraphrasing than comparison systems (a conclusion also confirmed by the statistics in Table 4.11).

4.7 Summary

Conclusion In this chapter we set out to answer the question: *Can the output of encoder-decoder paraphrasing models be controlled?* We showed that by using variable disentanglement we were able to control the length of a translation, whilst leaving the semantics intact. When paired with the pivoting approach of Chapter 3 we showed that we are able to use bilingual data to perform sentence compression, thus answering: *Can bilingual data be used as a source of training data for sentential NMT paraphrasing models?* Empirical results across three languages showed that our approach

outperformed supervised baselines.

Next chapter In the next chapter, we expand upon the controllability aspect of this chapter, we propose an approach which allows users to control the lexical choices and syntax of the output of a sequence-to-sequence model. We demonstrate this approach by introducing a controllable simplification model, which is trained on a generic supervised simplification dataset, but allows for controlling the output simplicity.

Chapter 5

Controllable Simplification

This chapter is based on Mallinson and Lapata (2019) and answers the following question:

- Can the output of encoder-decoder paraphrasing models be controlled?

In the previous chapter we saw how the output length of sequence-to-sequence models can be controlled via variable disegentlament; we expand upon this approach in this chapter. In doing so we focus on the constrained paraphrasing task of sentence simplification. Whilst previous approaches have been able to simplify sentences for a homogeneous audience, in this chapter we introduce an approach which allows for personalised simplifications. We argue that different users have different simplification needs (e.g., dyslexics vs. non-native speakers). We propose CROSS, a **C**ont**RO**llable **S**entence **S**implification model, which provides fine grain control of both the *level* of simplicity and the *type* of the simplification. We achieve this by enriching a Transformer-based architecture with syntactic and lexical constraints which we implement using variable disegentlament, and training on readily-available general purpose simplification data. Empirical results on two benchmark datasets show that constraints are key to successful simplification, offering flexible generation output.

5.1 Introduction

As discussed previously (Chapters 1), sentence simplification aims to reduce the linguistic complexity of a text whilst retaining most of its meaning. In this chapter, we propose a user-centric simplification model which draws on the advantages of the sequence-to-sequence architecture but can also *explicitly* model rewrite operations,

such as lexical and syntactic simplifications, and as a result generate output according to specifications. Although many simplification systems (e.g., (Zhu et al., 2010; Kauchak, 2013; Zhang and Lapata, 2017; Palmero Aprosio et al., 2019; Zhao et al., 2018a)) are intended for general purpose usage, different target populations may have different needs (Siddharthan, 2014). For instance, whether or not the syntax should be simplified depends on the reader: those affected by aphasia benefit from simpler syntax, while dyslexics have trouble processing long and infrequent words (Rello et al., 2013; Shewan and Canter, 1971). Unfortunately, simplification training datasets that target these different user groups are not available. It is therefore beneficial to have a model which can be trained on general purpose datasets and then be easily adapted for particular users or user group without being redesigned or retrained every time from scratch.

Our simplification model adopts the Transformer architecture (Vaswani et al., 2017) which has become the de facto standard and state-of-the-art in machine translation (Bojar et al., 2016) and relies entirely on self-attention to compute representations of its input and output without using recurrent or convolutional neural networks. Our innovation is to enrich a Transformer-based sequence-to-sequence model with syntactic and lexical constraints which allow the user to control both the *level* of simplicity and the *type* of simplification. We enable the model to make decisions about which words or syntactic structures to replace by enriching the training data with explicit information pertaining to lexical substitution and syntactic simplification. For example, we can mark words as to *keep* or *substitute*, or append a high-level level syntactic description (a template) to the source and target sentence. At test time, the user provides their constraints and the decoder must first decode the syntax of the target sentence before decoding the lexical tokens.

We evaluate our system on two publicly-available datasets collected automatically from Wikipedia (Woodsend and Lapata, 2011; Kauchak, 2013; Zhu et al., 2010) and human-authored news articles (Xu et al., 2015b) and report results using automatic and human evaluation. By comparing our constrained model against non-constrained variants we show that constraints are key to successful simplification, offering generation flexibility and controllable output. Our contributions in this chapter are three-fold: (1) we show that adding lexical and syntactic constraints to a Transformer produces state-of-the-art simplification results; (2) these constraints allow users to adapt the model to their personal needs; and (3) we conduct a comprehensive evaluation and comparison study which highlights the merits and shortcomings of various recently

proposed simplification models on two datasets.

5.2 Background

In this section, we provide relevant background information on existing approaches to simplification, controlling sequence-to-sequence models, and existing simplification datasets.

5.2.1 Modelling

One of the first neural network approaches to simplification was presented in Zhang and Lapata (2017), an encoder-decoder LSTM, trained with reinforcement learning, to optimize for grammaticality, simplicity, and adequacy (*DRESS*), and its extension, DRESS-Ls, which has an additional lexical simplification component. Dong et al. (2019) use a Programmer-Interpreter (Reed and de Freitas, 2016), which receives as input the source sentence and applies a sequence of edit operations (add, delete, keep). Kriz et al. (2019) propose adapting the loss function to give greater importance to simple words and to rerank a diverse set of simplifications according to fluency, adequacy, and simplicity.

Translation data, in the form of paraphrases, has also been incorporated into simplification models leading to significant improvements. Guo et al. (2018) use multi-task learning to augment the limited amount of simplification training data. In addition to training on complex-simple sentence pairs, their model employs paraphrases, created automatically using machine translation. Zhao et al. (2018a) introduces *DMASS*, an augmented Transformer-based simplification model with lexical rules obtained from Simple PPDB (Pavlick and Callison-Burch, 2016), a database of paraphrase rules, automatically annotated with simplicity scores.

In recent years there has been increased interest in controlling the output of simplification models. Bingel et al. (2018) notably acknowledge the fact that there is no one-size-fits-all solution to text simplification and develop a tool which can be personalized to a user's needs and adapted over time. Their system decides whether a word (in context) poses difficulty to the reader and suggests lexical substitutions. Scarton and Specia (2018) train a sequence-to-sequence model on Newsela, attaching tags which specify the grade level of the output sentences. Nishihara et al. (2019) expand upon this work by weighting the loss function to favour the generation of certain words. Since

the completion of this work, Martin et al. (2020a) proposed a model, *ACCESS*, which allows users to provide a high level specification of the output sentence, including: the amount of character overlap between the source and output, as measured by Levenshtein distance; character length ratio between source sentence and target sentence (compression level); lexical simplicity as measured by word frequency; and syntactic complexity as measured by the maximum depth of the dependency tree of the source divided by that of the target.

Previous works outside of simplification on controllability have focused on controlling the length and content of summaries (Kikuchi et al., 2016; Fan et al., 2018), politeness in machine translation (Sennrich et al., 2016a), and style (Ficler and Goldberg, 2017). Iyyer et al. (2018) propose a paraphrasing approach where users can control the paraphrase syntax by providing a syntactic template.

Our work draws inspiration from Grangier and Auli (2018) who post-edit the output of machine translation under the assumption that a human modifies a sentence by marking tokens they would like the system to change. Our model also controls simplification by taking as input both the sentence and change markers for it. However, we allow for a wider spectrum of rewrite operations than Grangier and Auli (2018) who focus solely on deletion and do not take syntax into account.

5.2.2 Datasets

In this section we provide details on the publicly-available English simplification datasets, in the next chapter we provide details on non-English datasets. Table 5.1 provides an overview of existing datasets and Table 5.2 shows examples from these datasets.

Dataset	Size	Method	Domain	Audience	Languages
WikiSmall	89K	Aligned	Wikipedia	Everyone	English
WikiLarge	400K	Aligned	Wikipedia	Everyone	English
Mturk	2K	Human	Wikipedia	N/A	English
Newsela	94K	Aligned	News	Children (8-18)	English

Table 5.1: Overview of sentence simplification *dataset*s, including the number of paraphrases (*size*); *method* in which they were collected; the *domain*; the target *audience* of the simplifications, where N/A indicates no target users were indicated when creating the dataset; and the *languages* of the dataset.

Dataset	Source	Simplification
WikiSmall	Genetic engineering has expanded the genes available to breeders to utilize in creating desired germlines for new crops.	New plants were created with genetic engineering.
WikiSmall	Every rhombus is a parallelogram, and a rhombus with right angles is a square.	A rhombus with all angles equal is called a square.
WikiLarge	The Great Dark Spot is regarded as a hole in the methane cloud deck of Neptune.	The Great Dark Spot is thought to represent a hole in the methane cloud deck of Neptune.
WikiLarge	1. Allessandro (“Sandro”) Mazzola, born 8 November 1942, used to be an Italian football player.	Alessandro (”Sandro”) Mazzola (born 8 November 1942) is an Italian former football player.
Mturk	admission to tsinghua is extremely competitive.	admission to tsinghua is highly competitive.
Mturk	he also completed two collections of short stories entitled the ribbajack & other curious yarns and seven strange and ghostly tales.	he also wrote two books of short stories called, the ribbajack & other curious yarns and seven strange and ghostly tales.
Newsela	One of the readers was Mohammed Baghdadi, 32, a manager at the Ministry of Trade.	Mohammed Baghdadi is 32.
Newsela	Weariness frays their voices, but they’re still on the freedom highway.	Their voices sound tired.

Table 5.2: Examples of sentence simplifications pairs from available datasets.

WikiSmall (Zhu et al., 2010) is a parallel simplification corpus constructed by automatically aligning sentences between English Wikipedia and simple Wikipedia. Simple Wikipedia is written by amateurs and *"is for everyone! That includes children and adults who are learning English."*. When creating articles authors are instructed to *"Use Basic English vocabulary and shorter sentences."*. The training set contains 89,042 sentence pairs. This dataset has been used extensively for both training and evaluating simplification models (Zhang and Lapata, 2017; Guo et al., 2018; Zhao et al., 2018a; Kriz et al., 2019). However, Xu et al. (2015b) highlighted several problems. They found that the alignments were prone to errors, that a large proportion of simplifications were inadequate and that the data generalized poorly to other text genres.

WikiLarge (Zhang and Lapata, 2017) is a large (296,402 sentence pairs) corpus which consists of a mixture of three Wikipedia simplification datasets collated by Zhu et al. (2010), containing WikiSmall, the two other Wikipedia simplification datasets Woodsend and Lapata (2011), and Kauchak (2013). Thus simplifications are taken from both English Wikipedia and Simple Wikipedia. WikiLarge has been used extensively as a training dataset for simplification (Zhang and Lapata, 2017; Guo et al., 2018; Zhao et al., 2018a; Kriz et al., 2019).

Mturk (Xu et al., 2016) is used as a test set for WikiLarge (Zhang and Lapata, 2017). It consists of 359 sentences, taken from Wikipedia, which were then simplified using Mechanical Turk to create eight reference simplifications per source sentence. Turkers were asked to produce a simpler version of the sentence. They were instructed to produce a *"simpler version while preserving its meaning, without losing any information or splitting sentence."* and to *"reduce the number of difficult words or idioms, simplify complex phrasing and make the sentence more straight forward."* Manual inspection by the authors was done to remove *bad* workers.

Newsela is a simplification corpus comprising of news articles written by Newsela's professional editors in English and Spanish. Each news article is written at five different simplicity levels (5-0), corresponding to grade levels 3 (8 year olds) to 12 (18 year olds). To assist editors in writing at the correct simplicity level the Lexile framework (Lennon and Burdick, 2004) was used. The Lexile framework is a proprietary readability metric, which combines word frequency information and sentence length.

The English subset of Newsela consists of 23,130 articles. Simplification pairs are produced by aligning sentences between articles of different simplicities, such that the source comes from more complex articles than the target. Unfortunately, due to the restrictive license of Newsela, there are no publicly-available alignments of these sentences; instead many have been proposed. For English, Zhang and Lapata (2017) produced 94K sentence pairs which have been used by many researchers for both training and testing (Zhang and Lapata, 2017; Guo et al., 2018; Zhao et al., 2018a; Kriz et al., 2019). They align all sentences from more complex to less complex, removing sentences paired from 0–1, 1–2, and 2–3. Others, Alva-Manchego et al. (2017); Scarton et al. (2018); Štajner and Saggion (2018) have proposed their own alignments which only align sentences between adjacent simplicity levels (i.e. 0-1, 1-2, 2-3, and 3-4), whereas Scarton and Specia (2018), generate alignments between all versions (i.e., 0-{1,2,3,4}, 1-{2,3,4}, 2-{3,4}, and 3-{4}). Jiang et al. (2020) use a neural CRF alignment model, which they found produced a larger and a higher quality dataset.

From Table 5.1 we see that target audiences for existing simplification datasets are very broad, ranging from no explicitly specified target audience, to everyone, to children aged 8-18. As a result large groups of users, such as second language learners, have no dedicated simplification datasets, thus requiring them to use general purpose simplification data.

5.3 Model Description

In this chapter we propose an approach which allows us to train on general purpose simplification but produce an output which targets individual users. The main idea is to control the output of a neural encoder-decoder model using constraints in both the encoder and decoder. The model still learns how to simplify from data, i.e., pairs of source (complex) and target (simple) sentences which are additionally annotated with change markers (e.g., indicating which words to replace, which syntactic constructs to delete) and takes these into account while generating simplifications.

5.3.1 Transformer

We will first define a basic encoder-decoder model for sentence simplification and then explain how to add constraints. Given a complex sentence $x = (x_1, x_2, ..., x_{|x|})$,

Tokens	take the square root of the variance .
Linearized	ROOT(take OBJ(DET(the) AMOD(square) NMOD(variance CASE(of) DET(the))) PUNCT(.))
Template	OBJ(AMOD(d0) DET(d0) NMOD(d1)) PUNCT( )
Input/Output	OBJ(AMOD(d0) DET(d0) NMOD(d1) OBJ) PUNCT( ) ||| take the square root of the variance .
Constraints	ROOT(OBJ, PUNCT), OBJ(AMOD, DET, NMOD), PUNCT()

Table 5.3: Example of source sentence with *linearized* parse, *template*, constraints extracted from the template, and input provided to our model (for training). To convert from a linearized parse to a template, first the dependents are ordered, then the opening and closing brackets are matched together (excluded for brevity). Finally, we remove levels lower than 2 and instead replace them with the *d** token which represents the maximum depth of the child.

our model learns to predict its simplified target $y = (y_1, y_2, ..., y_{|y|})$. Inferring the target y given source x can be modeled as a sequence-to-sequence learning problem. We adopt Transformer's multi-layer and multi-head attention architecture; further details can be found in Chapter 2 (Section 2.1.2) and Vaswani et al. (2017). The Transformer encoder has n layers, which transform the input sequentially: $X^{l+1} = L_l(X^l)$, where the first layer encodes the source embeddings:

$$X_i^0 = W_e e(x_i) + W_{pos} e(i) \tag{5.1}$$

where e is the word embedding matrix and $e_{pos,i}$ are positional embeddings.

The output representations of the full Transformer, $X^N = L_{1:N}(X^0)$. The decoder is composed of a stack of identical layers. In addition to self-attention the decoder attends to the source sentence X^N.

5.3.2 Lexical Constraints

Lexical substitution, the replacement of complex words with simpler alternatives, is an integral part of sentence simplification and has been the subject of much previous work (Specia et al., 2012; Paetzold and Specia, 2017; Lee and Yeung, 2018; Yatskar et al., 2010; Devlin, 1999; Inui et al., 2003; Kaji et al., 2002). We enrich the encoder of the Transformer with lexical constraints, by adding indicator features to each word embedding, specifying if the token should be kept. We employ three indicator types:

1. The token should be replaced; during training this is set if the token does not appear in the target sentence;

2. The token should be kept; during training this is set if the token is in the target sentence;

3. There is no preference for the token to be kept or replaced; during training half of all tokens are randomly assigned this value.

We differ from Grangier and Auli (2018) in that we have a third no preference type and we apply constraints more flexibly, as we mark tokens to be kept as long as their stems match, which allows for greater syntactic changes, for instance *run* can be changed to *running*.

Indicator features are added to the word embedding and positional encoding, as seen in the equation below:

$$X_i^0 = W_e e(x_i) + W_{pos} e(i) + W_c e(cw_i) \tag{5.2}$$

where cw_i are indicator features learnt during training. In this way we add soft constraints to the encoder. During training the indicator variables are true, therefore the model can learn to rely on them, and not encode this information itself, we consider this a form of soft variable disentanglement. In addition we also add hard constraints to the decoder. We restrict the generation of complex words marked with delete; during decoding we use constrained beam search, where complex words are given zero probability (Post and Vilar, 2018).

At *test time*, the user can control the model's output simply by (1) striking out tokens they wish to discard; (2) marking tokens they want to keep; or (3) leaving tokens unmarked. These could be words that an aphasic reader has trouble understanding, or a second language learner is not familiar with. For example in the sentence "*Dextromethorphan ~~occurs~~ as a* ***white powder***", *occurs* should be replaced and *white powder* should be preserved.

We use a fairly inexpensive approach to learn a list of complex words from training data, which can marked for replacement. To do so we follow Moore and Lewis (2010) who use language models to filter out-of-domain data, we use two uni-gram language models, one trained on complex sentences and one trained on simple sentences, to produce a list of words ordered by their relative simplicity:

$$\text{Complexity}(word) = \frac{P(word|complex)}{P(word|simple)} \tag{5.3}$$

Using Equation (5.3), we order all words in the training set with Complexity($word$)>1 and take the first N words to produce the complex list (e.g., *cavalier*, *offbeat*, *insofar*).

Additionally we experimented with using a simplification dictionary[1] provided by the Wikipedia editor "SpencerK" (Spencer Kelly). Due to the limited size of this dictionary, we combine it with an automatically created simplification dictionary, learnt from the training data. Word alignments, produced using GIZA++ (Och and Ney, 2003), were used to create phrase tables, which we treat as a simplification dictionary (*abandon* → *leave*, *replenished* → *filled*, *fraudulent* → *fake*. However, preliminary results showed this performed worse than using a list of complex words.

5.3.3 Syntactic Constraints

Syntactic simplification aims to reduce the syntactic complexity of a text while preserving its meaning and information content. Although the bulk of previous work has focused on sentence splitting, namely rewriting a complex sentence into multiple simpler sentences (Carroll et al., 1999; Chandrasekar et al., 1996; Vickrey and Koller, 2008; Sulem et al., 2018b; Siddharthan, 2004), other operations which reduce syntactic complexity involve rendering passive voice into active, simplifying relative clauses and coordination, as well reordering constituents or deleting them.

Syntax is introduced to our model by annotating the complex source and simplified target with high level syntactic descriptions (aka templates). Templates are induced from the training corpus by parsing source and target sentences with a universal dependencies parser (Straka, 2018). An example of a parse can be seen in Table 5.3. Dependency parses are further linearized and we extract a template corresponding to the top two levels of the parse. We differ from Iyyer et al. (2018), in that our templates are based on dependency parsers in contrast to constituent tree, and templates are prepended to the front of the target sentence. As such our model produces a probability distribution over possible templates, allowing us to choose a relevant template whilst still being able to reject complex templates.

The annotation process described above renders the model syntax-aware. Analogously to the lexical constraints, a globally constraint variant of beam search is used at test time and syntactic indicator features (i.e., replace, keep, indifference) are added to

[1] http://www.spencerwaterbed.com/soft/simple/about.html

the encoder. To reduce sparsity, a Markovian assumption is applied to the templates. Each constraint consists of one parent and its children as found within the template (see Table 5.3 for examples). Unlike lexical constraints, which are applied at the token level, syntactic constraints are applied at the rule level. At test time, the user provides a list of constraints the system must adhere to. The list is used to mark the input syntax and to constrain the decoder's output. For example, applying the constraint *Root(nsubj nmod nmod advcl)* $\rightarrow$ Root(nsubj nmod advcl) to the source sentence "*She remained in the United States until 1927 when she and her husband returned to France.*" produces the simplification "*She remained in the USA until she returned to France with her husband in 1927.*". We follow the lexical approach and generate a list of complex rules which the output must avoid (see Table 5.4).

WikiLarge
Root(cop, det, nsubj, punct, vocative)
Root(avmod, cop, det, parataxis, punct)
Root(aux, cop, det, nsubj, parataxis, punct)

Newsela
Root(cop, nsubj, onl, punct)
Root(iobj, nsubj,punct,xcomp)
Root(advmod, cop, csubj, punct)

Table 5.4: Examples of complex syntactic root rules as taken from Newsela and WikiLarge.

5.3.4 Constraint Combination

Lexical and syntactic constraints can be easily combined by merging the two sets of constraints provided by the user. In this case six indicator features are used, three for the lexical constrains and three for the syntactic constraints.

5.4 Experimental Setup

Datasets We experimented with two simplification datasets: (1) Newsela using the splits and alignments of Xu et al. (2015b) and (2) WikiLarge (Zhang and Lapata, 2017).

Model Configuration For both datasets we used the Transformer as implemented within OpenNMT-py (Klein et al., 2017). The encoder and decoder consist of 8 layers with a hidden dimension of size 500. Word embeddings, size 500, were initialized randomly and shared between the encoder and decoder. We used ten attentional heads and a copy mechanism (See et al., 2017). The network was optimized using Adam (Kingma and Ba, 2015), and SARI (Xu et al., 2016) was used for early stopping. The vocabulary size was limited to the 50,000 most frequent tokens, the remaining tokens were replaced with an UNK token.

Lexical Constraints Configuration At test time, we explored two approaches to applying the constraints to the encoder. For WikiLarge, simple tokens were marked with keep, and complex tokens were marked with replace. We included approximately $\sim$12,000 most complex words. For Newsela simple tokens were marked with indifference and complex tokens were marked with replace. We included approximately $\sim$7,000 most complex words. In both approaches, all function words were marked with indifference.

Syntactic Constraints Configuration At test time, complex syntactic rules were marked with the replace indicator and all other rules were always marked with the keep indicator. For Newsela, we include approximately 29% of the rules, whereas for WikiLarge we include approximately 13% of the rules.

Evaluation Metrics As there is no single agreed-upon metric for simplification (Alva-Manchego et al., 2020; Sulem et al., 2018a); we evaluate the models outputs using a combination of five automatically generated scores, which have been used previously in the literature Xu et al. (2016); Zhang and Lapata (2017). These metrics have been previously shown to correlate with human judgments of simplification quality (Xu et al., 2016)[2] and essentially quantify: a) whether the output is similar to the gold standard reference (*Target*-based, *T*); b) whether the output is similar to the source (*Source*-based, *S*); and c) whether the output is simple on its own, with no regard to preserving the meaning of the original sentence (*Readability*-based, *R*). We indicate

[2]However we note, that these correlations were calculated on a specific English test set using sentence level metrics.

the type of each metric using superscripts. We report BLEU^{T3}, SARI^{T,S4}, FKGL^{R5}, S-BLEU^{S}, and Copy^{S}. We included copy and S-BLEU to highlight the amount of rewriting a models does. More details on these evaluation metrics can be found in Chapter 2 (Section 2.3). We also evaluated system output by eliciting human judgments via Amazon's Mechanical Turk. Native English speakers (self reported) were asked to rate simplifications on three dimensions: Grammaticality (is the output grammatical and fluent?), Meaning Adequacy (to what extent is the meaning expressed in the original sentence preserved in the output, with no additional information added?), and Simplicity (is the output a simpler version of the input?). Full instructions can be found in Appendix C (Section C.2). The ratings were obtained using a five point Likert scale. 100 sentences were randomly sampled from the test set[6], each sample received five ratings, resulting in 500 judgments per test set.

5.5 Results

Our first suite of experiments compares our approach against the state-of-the-art simplification models aiming to show that our model can also function as a general-purpose simplification system. There is no point in having a controllable model if it cannot generate adequate simplifications on its own. Our second suite of experiments examines how the simplicity level can be manipulated.

Automatic Evaluation Table 5.5 summarizes our automatic evaluation results[7] on WikiLarge and Newsela. We compared our model against three well-established non-neural models: PBMT-R (Wubben et al., 2012), a phrase-based machine translation model, SBMT-SARI (Xu et al., 2016), a syntax-based translation model trained on PPDB and which is then tuned using SARI, and Hybrid (Narayan and Gardent, 2014), a model which performs sentence splitting and deletions and then simplifies with PBMT-R. We also compare against various neural simplification models: (a) the three LSTM-based models reported in (Zhang and Lapata, 2017), namely EncDecA, an encoder-decoder model with attention, DRESS and DRESS-Ls; (b) DMASS (Zhao

[3]We used `multi-bleu-detok.perl` to calculate corpus-level BLEU.

[4]We used corpus level SARI with precision for deletion operator.

[5]following Zhang and Lapata (2017) we used a corpus level FKGL, however we ensured that a newline indicated a sentence break.

[6]We used the same samples as Zhang and Lapata (2017).

[7]As our automatic metrics differ from previous papers we re-calculate all scores for all available simplification models.

et al., 2018a), (c) a vanilla transformer-based encoder-decoder model without any constraints.

We report results for several variants of our model which we call CROSS as a shorthand for **C**ont**RO**llable **S**entence **S**implification. CROSS-Lex contains lexical constraints only, CROSS-Syn focuses solely on syntactic simplifications, while CROSS is the full model with both types of constraints. We also include two strong baselines, repeating the source sentence (Source) and truncating the source sentence to the first *N* words, as determined by the validation set (Truncate).

Results on WikiLarge are mixed, with no model being best for every metric. We see that SBMT-SARI achieves the highest SARI, with minimal copying and a moderate S-BLEU. Of the two existing state-of-the-art models, DRESS-Ls and DMASS, we see that DRESS-Ls achieves a moderate SARI score and a moderate S-BLEU, however, it has a high Copy score. This suggests that DRESS-Ls is very polar, applying high amounts of rewriting to some sentences and keeping others completely unchanged. DMASS achieves the second highest SARI score and Copy is low, however, S-BLEU is high suggesting it produces modest changes consistently.

CROSS achieves a slightly worse SARI than the baseline Transformer, however this is in part due to the Transformer's high Copy and high S-BLEU. In contrast, CROSS achieves a low S-BLEU and Copy score similar to that of the references. CROSS has a lower SARI compared to CROSS-Lex, however, it has a better S-BLEU and Copy. CROSS outperforms CROSS-Syn with a better SARI and Copy score. The results also show that *standard* encoder-decoder models (EncDecA, Transformer) produce outputs which are highly similar to the input, highlighting the importance of constraining the output.

We next consider the Newsela dataset. We see that DRESS-Ls achieves the highest SARI, however, it also has the highest level of copying and a moderately high S-BLEU. DMASS, on the other hand, achieves a low SARI, but with a low amount of copying and a low S-BLEU. Also notice that the Truncate baseline has the highest BLEU score, outside of the DRESS models. The Transformer achieves a moderate SARI, however, it also has a high Copy and high S-BLEU. CROSS achieves a low SARI which in part can be explained by its high level of rewriting as seen in the low S-BLEU and Copy. We see that CROSS-Lex has a higher SARI compared to CROSS but worse S-BLEU and Copy scores. CROSS-Syn and CROSS both have very similar scores, however, CROSS-Syn performs more rewrites.

WikiLarge	SARI	BLEU	FKGL	S-BLEU	Copy
Reference	N/A	N/A	8.24	63.92	16.2%
Source	26.31	**99.37**	9.54	100.00	100%
Truncate	35.62	99.32	9.54	95.48	**0%**
PBMT-R	40.30	81.02	8.40	74.95	09.7%
Hybrid	27.59	48.69	4.72	**30.57**	03.1%
SBMT-SARI	**40.75**	73.01	7.53	67.93	10.6%
EncDecA	39.58	89.00	8.61	83.81	40.7%
DRESS	35.45	77.32	**6.76**	56.96	21.5%
DRESS-Ls	36.08	80.35	6.90	60.21	26.2%
Transformer	36.21	81.51	8.73	76.33	36.2%
DMASS	40.35	79.68	7.45	70.82	15.6%
CROSS-Lex	38.82	70.70	7.92	65.62	10.6%
CROSS-Syn	33.89	64.98	7.98	68.88	19.9%
CROSS	36.07	64.64	7.46	56.11	15.6%

Newsela	SARI	BLEU	FKGL	S-BLEU	Copy
Reference	N/A	N/A	3.43	17.81	0%
Source	11.97	20.79	8.61	100.00	100%
Truncate	36.92	21.54	5.57	62.54	**0%**
PBMT-R	41.23	17.62	7.96	75.29	05.9%
Hybrid	35.37	10.87	4.14	19.96	03.3%
EncDecA	42.98	21.17	5.48	52.54	15.7%
DRESS	42.85	22.65	4.20	39.69	11.3%
DRESS-Ls	**43.26**	**23.66**	4.36	42.72	14.5%
Transformer	42.21	19.90	4.77	40.05	11.6%
DMASS	37.36	07.51	3.84	**11.15**	01.1%
CROSS-Lex	41.56	18.88	3.81	33.98	06.8%
CROSS-Syn	38.12	14.30	3.48	21.35	05.1%
CROSS	37.57	12.68	**3.51**	26.55	05.6%

Table 5.5: Automatic evaluation on WikiLarge and Newsela test set. We also report the average FKGL, S-BLEU, and Copy of all references (Reference).

WikiLarge	Gram	Mean	Simp	AVG	Min
Reference	4.01*	4.13**	3.56**	3.90**	3.16*
DRESS-Ls	4.32**	3.97**	3.14	3.81**	2.80
DMASS	3.69	3.21	2.57**	3.16	2.29**
Transformer	3.91	3.63	3.04**	3.53	2.72**
CROSS-Lex	3.72	3.41	3.18	3.43	2.80
CROSS-Syn	3.54	2.22	2.46**	3.07*	2.15**
CROSS	3.61	3.37	3.13	3.37	2.84

Newsela	Gram	Mean	Simp	AVG	Min
Reference	4.11**	3.73**	3.88**	3.91**	3.47**
DRESS-Ls	3.33*	2.98**	2.93	3.08**	2.45
DMASS	2.05**	1.55**	1.74**	1.78**	1.39**
Transformer	2.88**	2.47**	2.70	2.68**	2.00**
CROSS-Lex	3.07**	2.89**	2.95	2.97**	2.45
CROSS-Syn	3.60	3.37	2.89	3.27	2.31
CROSS	3.54	3.41	2.91	3.28	2.29

Table 5.6: Human evaluation on WikiLarge and Newsela. Models significantly different from CROSS are marked with * $(p < 0.05)$ and ** $(p < 0.01)$. Significance tests were performed using a student *t*-test.

Human Evaluation The results of our human evaluation are presented in Table 5.6. We follow previous approaches and report Grammaticality, Meaning Adequacy, and Simplicity individually and combined (AVG is the average of the three dimensions). In addition, we include a new metric *Minimum*, which is the (average) minimum value of Grammaticality, Meaning Adequacy, and Simplicity per sentence. We include Minimum because we argue that a simplification is only as good as its weakest dimension. We note that it is trivial to produce a sentence that is perfectly adequate and fluent, by simply repeating the source sentence. It is also easy to produce a simple sentence if we do not care about adequacy. We evaluated CROSS (and CROSS-Lex, CROSS-Syn variants) against the two state-of-the-art models DMASS and DRESS-Ls as well a Transformer baseline. We also elicited judgments on the gold standard Reference as an upper bound.

Human evaluation on WikiLarge (top half in Table 5.6) shows that both DRESS-Ls and CROSS achieve highest scores for Minimum. CROSS significantly outperforms all other models for both Min and Simplicity. Transformer achieves a higher score for both Grammaticality and Meaning compared to CROSS. However, this can be explained due to the high Copy score, which guarantees high Grammaticality and Adequacy scores. This can also in part explain the high Grammaticality and Meaning Adequacy scores for DRESS-Ls. CROSS-Syn achieves lower scores compared to CROSS-Lex, suggesting that syntactic changes are not as important for WikiLarge.

Human evaluation on Newsela (second half of Table 5.6) shows that all CROSS variants are better than related Transformer and DMASS models across all metrics. CROSS and DRESS-Ls both achieve the highest Minimum scores. For all other metrics, CROSS is better or the same than all other models. CROSS and CROSS-Syn achieve similar results, both outperforming CROSS-Lex. This suggests that syntactic simplifications are more prominent in Newsela compared to WikiLarge.

Analysis of Model Output We further analyzed the simplifications produced by CROSS to gain insight on the types of simplifications it generates. We sampled 100 sentences (50 from each test set) and classified the simplifications into two categories, namely lexical (Lex) or syntactic (Syn). For syntactic simplifications we further marked whether these pertained to common changes, i.e., passive to active voice (Voice), past tense to present or past perfect (Tense), and sentence splitting (Split). Table 5.7 shows a breakdown of these phenomena for CROSS, the baseline Transformer model, and the references. As can be seen, CROSS performs similar simplifications to the references,

	Lex	Syn	Voice	Tense	Split	All
Reference	35%	10%	7%	5%	6%	41%
Transformer	9%	1%	1%	0%	0%	9%
CROSS	29%	8%	8%	2%	0%	35%

Table 5.7: Proportion of simplifications on a 100 sentence sample from the WikiLarge and Newsela test sets. Examining Lexical simplifications (Lex), syntactic simplifications (Syn), passive to active voice (Voice), past tense to present or past perfect (Tense), and sentence splitting (Split).

WikiLarge	Gram	Mean	Simp	AVG	Min	FKGL
XSimple	3.30	3.09	3.06*	3.15	2.84	6.96
Simple	3.24	3.11	2.87	3.08	2.77	7.46

Newsela	Gram	Mean	Simp	AVG	Min	FKGL
XSimple	3.46**	2.88**	3.11**	3.15	2.33**	2.91
Simple	3.89	3.59	2.53	3.34	2.10	3.51

Table 5.8: Human evaluation on varying simplicity of model output. Ratings that are significantly different are marked with * $(p < 0.05)$ and ** $(p < 0.01)$. Significance tests were performed using a student *t*-test.

Source	In its pure form, Dextromethorphan occurs as a white powder.
Reference	Dextromethorphan is a white powder in its pure form.
Simple	In its pure form, Dextromethorphan **is like** a white powder.
XSimple	Dextromethorphan can be found as white powder.
Source	The Pentagon is poised to spend billions to build a new stealth bomber, a top secret project that could bring hundreds of jobs to the wind-swept desert communities in Los Angeles County's northern reaches.
Reference	Mission to build the secret warplane.
Simple	The Pentagon secret project that could bring hundreds of jobs to the desert-swept communities in Los Angeles County.
XSimple	It could also bring hundreds of jobs.
Source	The United States is about to spend billions of dollars to build a top-secret warplane.
Reference	Mission to build the secret warplane
Simple	The United States is about spend dollars to build a top-secret warplane.
XSimple	The United States is about to build a warplane.

Table 5.9: Example outputs, including both simple and eXtra simple.

and substantially more syntactic changes compared to the Transformer.

Controllability A central claim of this chapter is that CROSS can be adapted to user needs. We test this claim, by experimenting with varying the simplicity level of the output. Specifically, we sampled 100 complex source sentences (with FKGL score of 11 or higher) from the WikiLarge and Newsela test sets and produced two sets of outputs, one with our general-purpose system which produces a moderate amount of simplification (Simple), and another one where we forced the model to simplify more drastically, extra simple (XSimple). This was achieved by increasing the number of lexical and syntactic constraints the model must adhere to. Specifically, we include the 12,000 most complex words for Newsela, and the 18,000 most complex tokens for Wikilarge. We also increased the number of complex syntactic constraints to approximately 40% for Newsela and 25% for WikiLarge. We note that very restrictive constraints can lead to a loss of meaning, where the output simplification shows no resemblance to the input sentence.

Results in Table 5.8 show that CROSS is able to successfully alter the simplicity level of the output. For both datasets we see that participants perceive differences between the output of the simple and XSimple models (this is also reflected in the FKGL which is lower for XSimple). For WikiLarge, all scores apart from simplification do not differ significantly. For Newsela, we see that XSimple sentences are

significantly less adequate and grammatical. However, on average Simple and XSimple sentences do not significantly differ, showing a trade-off between simplicity and adequacy/grammaticality. Examples of system output are shown in Table 5.9 and in Appendix C (Section C.1).

5.6 Summary

Conclusion In this chapter we set out to answer: *No supervised data exists for a specific paraphrasing task. Can bilingual data be used as a source of training data for paraphrasing?* To answer this question we developed a simplification model using variable disentanglement, which can jointly or individually control the syntax and lexical choice of its output. Experiments showed that our constraint-aware model produces state-of-the-art simplification results. We further showed that by adjusting these constraints we can control the level of simplification of the output.

Next Chapter In the next chapter, we expand upon the idea of adapting existing simplification data. We propose an approach which adapts simplification data from one language and applies it to other languages. This is ideal for situations where there exists data in a high resource language, but not in a low resource language.

Chapter 6

Zero-Shot Crosslingual Sentence Simplification

This chapter is based on Mallinson et al. (2020a) which was published in EMNLP 2020 and answers the following question:

- Can bilingual data be used to transfer sentential paraphrasing training data from one language to another?

We consider the constrained paraphrasing task of sentence simplification. In the previous chapter we trained a sequence-to-sequence model on large amounts of parallel simplification data. However, large scale parallel paraphrasing data often only exists in English. Therefore we propose a zero-shot modelling framework which transfers simplification knowledge from English to another language (for which no parallel simplification corpus exists). A shared transformer encoder constructs language-agnostic representations, with a combination of task-specific encoder layers added on top (e.g., for translation and simplification) and language-specific decoders. Empirical results using both human and automatic metrics show that our approach produces better simplifications than unsupervised and pivot-based methods.

6.1 Introduction

As discussed in Chapters, 1 and 5, sentence simplification aims to reduce the linguistic complexity of a text whilst retaining most of its meaning. Modern approaches (Zhang and Lapata, 2017; Nishihara et al., 2019; Dong et al., 2019; Martin et al., 2020a) view the simplification task as monolingual text-to-text rewriting and employ the very suc-

cessful encoder-decoder neural architecture (Bahdanau et al., 2015; Sutskever et al., 2014).

As seen in Chapter 5 (Section 5.2) large-scale parallel datasets exist for English as well as Spanish (Agrawal and Carpuat, 2019), however there is a limited amount of simplification data for other languages. For example, Klaper et al. (2013) automatically aligned 7,000 complex-simple German sentences[1], and Brunato et al. (2015) released 1,000 complex-simple Italian sentences. But data-driven approaches to simplification, in particular popular neural models, require significantly more training data to achieve good performance, making these datasets better suited for testing/development purposes.

Unsupervised approaches (Surya et al., 2019; Artetxe et al., 2018; Zhao et al., 2020) which forego the use of parallel corpora are an appealing solution to overcoming the paucity of data. However, in this chapter we argue that better simplification models can be obtained by taking advantage of existing complex-simple data in a high-resource language and bilingual data in a low-resource language (i.e., a language for which no parallel simplification corpus exists). Drawing inspiration from the success of machine translation (Firat et al., 2016b; Blackwood et al., 2018; Johnson et al., 2017), we propose a modeling framework which transfers simplification knowledge from English to another language while generalizing across language and task barriers during training.

The backbone of our model is an encoder-decoder transformer (Vaswani et al., 2017) trained using multi-task learning to either translate, autoencode, simplify, or language model in both high- and low-resource languages. Regardless of the task or language, we employ the same base encoder on top of which *task-specific* transformer layers are added, while *language-specific* transformer decoders are used to generate the output sequence. Since the same base encoder is used for all tasks and languages, the model learns task- and language-agnostic representations. A beneficial side-effect is that the proposed architecture can be trained using one language and tasked to simplify another.

As simplifications for multiple languages can be produced within the *same* model, our approach is more scalable compared to pivot-based methods, as described in Chapter 4. The latter would first translate the complex sentence into a high-resource language, apply a monolingual simplification model, and then translate back the output to the original language. We avoid having to train multiple models and make *multiple* hops, where each hop can add noise, and instead develop a *one-hop* crosslingual zero-

[1]This dataset has not been publicly released.

shot approach. We evaluate our model using English as our high-resource language and German as our low-resource language on two test sets from different domains, and with different end-users in mind. These include TextComplexityDE (Naderi et al., 2019), a recently-created corpus of German Wikipedia sentences deemed complex by second language German learners. We also release a second dataset which contains manual simplifications of articles taken from GEOlino[2], a popular children's magazine. Empirical results using both human and automatic metrics show that our approach produces better simplifications than both unsupervised and pivot-based methods.

Our contributions in this chapter are threefold: (1) a cross-lingual model architecture which allows the transfer of simplification knowledge from high- to low-resource languages, alleviating the paucity of training data for monolingual simplification; (2) a comprehensive evaluation framework using automatic metrics and human judges; and (3) the release of a dataset in German which we hope will facilitate further research in automatic simplification[3].

6.2 Background

Within this section we cover background information on simplification, crosslingual generation, and current simplification datasets.

Simplification Chapter 5 (Section 5.5) provides details on existing state-of-the-art models for supervised simplification, including using bilingual data to supplement simplification data. In this section we survey semi-supervised and unsupervised approaches. Martin et al. (2020b) trains a controllable paraphrasing model on general-purpose paraphrases datasets and at test time the model is constrained to output simple paraphrases. Artetxe et al. (2018); Lample et al. (2018) demonstrate how an *unsupervised* neural MT model can be trained by optimizing two objectives: *denoising* and *on-the-fly back-translation*, as discussed in Chapter 2 (Section 2.3). Surya et al. (2019) extend this approach further by adding two losses, which they show result in better simplifications: (1) an *adversarial loss* using a discriminator which tries to determine if the source sentence is complex or simple, and (2) a *diversification loss*, where a classifier is trained to determine if the source sentence was encoded using the complex or simple encoder. Zhao et al. (2020) also extend Artetxe et al. (2018), by using

[2] https://www.geo.de/geolino

[3] https://github.com/Jmallins/ZEST-data

simple PPDB to create noise, replacing simple phrases with complex phrases. Reinforcement learning is further used to reward the fluency, adequacy and simplicity of simplifications.

Crosslingual Generation Cross-lingual transfer learning based approaches have originated in machine translation. Dong et al. (2015) translate from one source language to multiple target languages (one-to-many) adding a separate decoder for each. Follow-on work (Luong et al., 2015a; Firat et al., 2016a) perform translation with multiple encoders and decoders (many-to-many). Johnson et al. (2017) and Ha et al. (2016) train multilingual models where all languages share encoder and decoder parameters, and language tags (appended to the source sentence) are used to specify the target.

Multilingual models are also capable of translating between unpaired languages, thereby performing zero-shot translation (Firat et al., 2016b; Johnson et al., 2017; Ha et al., 2016). Blackwood et al. (2018) propose sharing all parameters but the attention mechanism, while Lu et al. (2018) develop a shared "interlingua layer" between the language-specific encoders and decoders.

While zero-shot approaches are effective for translating between unpaired languages, they do not consider the case where there exists *no parallel* data for a language. For simplification, we assume that there is no parallel corpus in the low-resource language (e.g., complex-simple German). Furthermore, preliminary results showed that zero-shot translation approaches (Johnson et al., 2017) which append a tag in the source sentence — this tag would indicate the simplification task in our case — perform poorly, basically resulting in the source sentence being copied over with no changes made. We suspect this is due to tags not providing a sufficiently strong guidance to the model. We circumvent this by replacing tags with task-specific transformer encoder layers which are added on top of the base encoder. By using transformer layers instead of tags, we are able to better enforce the desired behaviour, simplification, when performing zero-shot simplification, as the encoded information is forced to take into account the simplification layer.

This proposed architecture allows us to transfer supervision signals across languages and is potentially useful for other generation tasks, including question generation (Kumar et al., 2019) and sentence compression (Shen et al., 2018; Duan et al., 2019a).

Our approach of task-specific transformer layers is similar to recent work on adapters (Houlsby et al., 2019). Adapters provide a lightweight alternative to fine-tuning neu-

Dataset	Size	Method	Domain	Languages
WikiSmall	89K	Aligned	Wikipedia	English
WikiLarge	400K	Aligned	Wikipedia	English
Mturk	2K	Human	Wikipedia	English
Newsela	94K	Aligned	News	English
Newsela	150K	Aligned	News	Spanish
Brunato et al. (2015)	1K	Human	Mixed	Italian
SIMPITIKI	1K	Human	Wikipedia	Italian
PaCCSS–IT	63K	Aligned	Web	Italian
Caseli et al. (2009)	2K	Human	News	Portugese

Table 6.1: Overview of sentence simplification *dataset*s, including the number of paraphrases (*size*), *method* in which they were collected, the *domain* and the *languages* of the dataset. English dataset are included for comparison.

ral networks, and consist of adding intermediate parameters between already trained transformer layers. The entire network is frozen apart from these new parameters, thus allowing the model to be adapted to a particular task without training the entire network. Duan et al. (2019b) train a generic varational auto-encoder; they then use adapters to change attributes of the output text, such as sentiment topic, or length. Task-specific adapters could have been used as an alternative to task specific layers for ZEST. This would have allowed for quicker training.

6.2.1 Simplification Datasets

Within this section, we discuss existing non-English sentence simplification datasets (For English simplification dataset we refer the reader to Chapter 5 (Section 5.2)), restricting ourselves to datasets which are currently publicly available. Table 6.1 provides an overview of existing datasets and Table 6.2 shows examples from these datasets.

Newsela As mentioned in the previous chapter Newsela is a simplification corpus comprising news articles written by Newsela's professional editors and exists in both English and Spanish. For Spanish there exists two sets of splits/alignments; Agrawal and Carpuat (2019), which produced 150K complex-to-simple pairs, and Palmero Aprosio et al. (2019) which produce 56K pairs.

Dataset	Source	Simplification
SIMPITIKI	Romani conquistarono la Valle Camonica nel 16 a.C. tramite i</del>l proconsole dell'Illiria Publio Silio Nerva.	I Romani conquistarono la Valle Camonica nel 16 a.C. <ins>con un'azione de</ins>l proconsole dell'Illiria Publio Silio Nerva.
PaCCSS-it	MIl risultato , purtroppo , è sotto gli occhi di tutti .	I risultati sono sotto gli occhi di tutti .
Caseli et al. (2009)	– Elas atacam mais nos locais em que o pessoal das excursões lava mãos e pratos após as refeições – diz Adalberto Silva, presidente da associação dos moradores da localidade.	– Elas atacam mais nos locais em que o pessoal das excursões lava mãos e pratos após as refeições – diz Adalberto Silva. Adalberto Silva é presidente da associação dos moradores da localidade.

Table 6.2: Examples of sentence simplifications pairs as directly taken from the available datasets.

SIMPITIKI (Tonelli et al., 2016) Aligns revised sentences and their corresponding original sentences from the Italian Wikipedia, if the revision notes indicates a simplification has occurred. 4,356 sentence pairs were originally produced; these were then manually curated into 1,166 pairs. SIMPITIKI has been used a as part of larger Italian simplification dataset (Palmero Aprosio et al., 2019) and to evaluate Italian simplification systems (Scarton et al., 2017).

Brunato et al. (2015) created two Italian simplification subcorpora, the first contains 32 short novels for children, which were then simplified. The second is composed of 24 texts produced and simplified by teachers. For both corpora sentences were aligned across documents. It has been used as part of a larger Italian simplification dataset (Palmero Aprosio et al., 2019).

PaCCSS-it (Brunato et al., 2016) contains 63,000 complex-to-simple Italian sentence pairs automatically extracted from the Web. A subset of sentences were manually annotated. These annotations were used to train a classifier, which scored the extracted pairs which is then used to filter the corpus. Scarton et al. (2017) further filtered this corpus to produce higher quality simplification pairs.

Caseli et al. (2009) is a Brazilian Portuguese simplification dataset composed of 104 texts from the Zero Hora newschapter, paired with manually created simplifications. It

has been used for training sentence splitting models (Gasperin et al., 2009).

As can be seen in Table 6.1 English has by far the largest simplification dataset, with Spanish Newsela being fairly large, however it comes with a restrictive licence. The table also serves to highlight that most languages have no simplification dataset at all. With no or little simplification data it becomes hard to train neural models, motivating this chapter, where we transfer simplification data from English, where we see there is lots of data and apply it to German where there is no data. We note that this pattern of there being large amounts of data in English, but little in other languages is repeated for many NLG tasks.

6.3 Zero-shot Simplification

We first define a basic encoder-decoder Transformer, for more complete details we refer the reader to Chapter 2 (Section 2.1), before adapting it for zero-shot crosslingual simplification with multi-task learning.

6.3.1 Encoder-Decoder

Given a source sentence $x = (x_1, x_2, ..., x_{|x|})$, our model learns to predict target $y = (y_1, y_2, ..., y_{|y|})$, where y could be a translation (e.g., from English to German) or a simplification (e.g., from complex to simple English). Inferring target y given source x can be modeled as a sequence-to-sequence learning problem (Bahdanau et al., 2015). Our approach adopts the Transformer's multi-layer and multi-head attention encoder-decoder architecture (Vaswani et al., 2017). The Transformer encoder has n layers (denoted L_l for layer l), which transform the input sequentially: $X^{l+1} = L_l(X^l)$. For details regarding the Transformer layer, we refer the reader to Chapter 2 (Section 2.1.2) and Vaswani et al. (2017). The output representations of the full Transformer, $X^N = L_{1:N}(x)$. The decoder is composed of a stack of identical layers. In addition to self-attention the decoder attends to the source sentence x^N. Encoder and decoder stacks are trained to minimize the cross-entropy loss of y given x:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|y|} \log p(y_i | y_{<i}, X^N; \theta) \tag{6.1}$$

Task	Source Language	Target Language	Target Domain
Auto-encoding	HR	HR	complex
Auto-encoding	LR	LR	simple
Auto-encoding	LR	LR	complex
Auto-encoding	HR	HR	simple
Translate	HR	LR	complex
Translate	LR	HR	complex
LM	None	HR	complex
LM	None	HR	simple
LM	None	LR	complex
LM	None	LR	simple
Simplify	HR	HR	simple

Table 6.3: Training tasks and their instantiations.

6.3.2 Multi-task Learning

We define a multi-task crosslingual setup where the model is trained on four basic tasks; namely translation, autoencoding, language modeling, and simplification. We train on different instantiations of these tasks depending on the *source language* which can be high-resource (HR; e.g., English) or low-resource (LR; e.g., German), the *target language* (which is again HR or LR), and the *output domain* which can be simple or complex. We assume we only have monolingual simplification data in the high-resource language and that we have bilingual translation data only in the complex domain. Table 6.3 has a breakdown of the tasks we consider, with a more detailed description given below.

Simplification is the backbone of the model and consists of a complex source sentence which must be transformed into a simple sentence, while still retaining the original meaning. We assume we only have parallel training data in the high-resource language (see last row in Table 6.3).

Translation consists of a source sentence, which must be translated into the target language while retaining the meaning of the source. By training on translation data, our model learns language-agnostic representations which are helpful for simplifying

in the low-resource language.

Autoencoding We also include an autoencoding task, i.e., translation between the same language. As it is trivial to autoencode with attention, we apply source token dropout, where randomly-selected source tokens are replaced with a special DROP token (Lample et al., 2018). We apply this dropout to all tasks (translation, autoencoding, and simplification). Additionally, this task allows us to incorporate monolingual non-parallel simple data from the low-resource language.

Language Modeling has no source sentence; instead the decoder must learn to predict the next token based on its history. This task also allows us to incorporate monolingual non-parallel simple data from the low-resource language (see rows LM in Table 6.3).

Domains We define two domains: the simple domain which consists of text that is easy to read, and the complex domain where text has not been explicitly written for ease of reading. Introducing domains to the model allows us to further inject knowledge about monolingual non-parallel simple sentences from the low-resource language. We use the target audience of the data to determine if it is simple or complex (e.g., if the text comes from Simple Wikipedia or a children's book it is representative of simple language). In practice, there often exist only limited amounts of non-parallel simple sentences in the low-resource setting, highlighting the difficulty of the task.

6.3.3 Crosslingual Training

With the tasks defined, we explain how the model is able to switch among them. We propose a modular encoder, where different encoder layers are used for different tasks; an outline of this can be seen in Figure 6.1. For every task we use the same k base transformer encoder layers, where k is a hyper-parameter. Each task $\mathcal{T}$ (simplification, translation, language modeling), has additional t dedicated transformer layers $L^{\mathcal{T}}_{1:t}$, which are applied on top of the base k layers, $L^{\mathcal{T}}_{1:t}(L_{1:k}(X^0))$, replacing the need for task tags, which we found the model was ignoring when performing zero-shot simplification. Each domain $\mathcal{D}$ (simple/complex), also has d additional dedicated transformer layers $d^{\mathcal{D}}_{1:d}$ applied on top of the task specific layers, again replacing the use of tags to better enforce the output domain constraint. The final representation of the

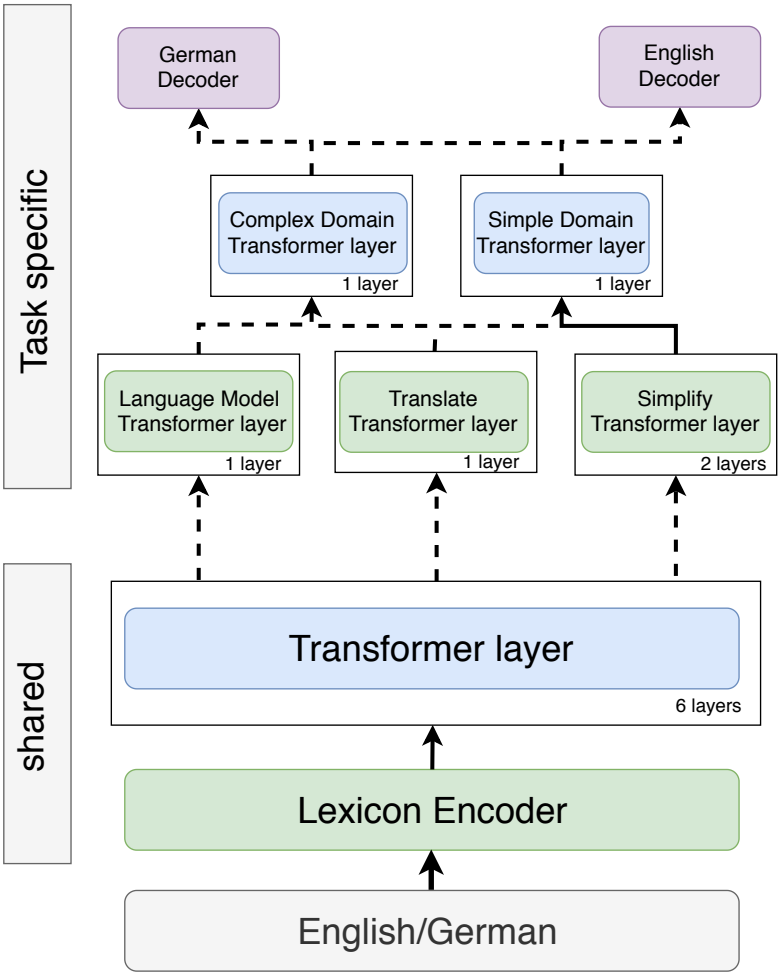

Figure 6.1: Architecture of our crosslingual encoder-decoder model. The *lexicon Encoder* transforms words into word embeddings. Solid lines indicate mandatory paths, dotted lines indicate possible paths.

source sentence x is:

$$x^N = L^{\mathcal{D}}_{1:d}(L^{\mathcal{T}}_{1:t}(L_{1:k}(X^0))) \tag{6.2}$$

Training is done end-to-end to minimize the negative log-likelihood; for each mini-batch we specify the task, domain and output language (O):

$$\mathcal{L}_{\text{CE}} = -\log P(y|y_{<j}, X^N; \theta, \{\mathcal{D}, \mathcal{T}, O\}) \tag{6.3}$$

$\mathcal{D}$ and $\mathcal{T}$ determine the choice of dedicated Transformer layers. We use a dedicated Transformer decoder for each output language O to encourage the model to learn a language agnostic representation. All text is splt into subwords using SentencePiece (Kudo and Richardson, 2018), resulting in a shared vocabulary between LR and HR. This allows for word embeddings to be shared between the encoder and the decoders.

We further force representations to be language agnostic, by employing a DISCriminator (Ganin and Lempitsky, 2015), a feed-forward network trained to distinguish HR and LR from the hidden representations. The encoder is then trained to perplex the discriminator. Specifically, we add two discriminators to our model; one determines the language of the source sentence (I) using $L_{1:k}(X^0)$, and the other predicts the target language using the output of the encoder x^N. In this way we ensure the input to the simplification transformer layers is language agnostic as well as the output. The discriminator is trained to minimize the binary cross-entropy loss (BCE) between its predictions and the ground truth:

$$\begin{aligned}\sum_{i=1}^{|X|} \text{BCE}(I, \text{DISC}(L_{1:k}(X^0)_i; \theta_{d_I}) + \\ \text{BCE}(O, \text{DISC}(X^N_i); \theta_{d_O})\end{aligned} \tag{6.4}$$

where θ_{d_I} and θ_{d_O} are the parameters of the two discriminators. The encoder is trained using an adversarial loss, to perturb the discriminator:

$$\begin{aligned}\mathcal{L}_{\text{ADV}} = -\sum_{i=1}^{|X|} \text{BCE}(I, \text{DISC}(L_{1:k}(X^0)_i; \theta) + \\ \text{BCE}(O, \text{DISC}(X^N_i); \theta)\end{aligned} \tag{6.5}$$

The adversarial loss is combined, and optimized simultaneously, with the cross-entropy loss to produce the training objective of the entire model.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ADV}} \tag{6.6}$$

where λ moderates the degree to which the encoder should perturb the discriminators. A high value for λ can cause the encoder to not encode any information regarding the source input.

To perform simplification in the low-resource language at test time, the base encoder is used with the simplification stack which is subsequently decoded with the LR decoder. To perform crosslingual simplification, the decoder can simply be changed to the HR decoder.

6.4 Experimental Setup

Training Set Our training data is summarized in Table 6.5. For all experiments we assume that English is the high-resource language and German is the low-resource language. Simplification data in English is taken from WikiLarge (Zhang and Lapata, 2017), described in Section 5.2.1. English-German bilingual data is taken from the WMT19 news translation task. Complex monolingual non-parallel data uses one side of the WMT19 translation data. Simple English non-parallel data uses sentences extracted from simple Wikipedia, a simplified version of Wikipedia. Simple German non-parallel data uses sentences scraped from GEOLino (Hancke et al., 2012). Examples of the training data can be seen in Table 6.4.

Test Set We evaluated our model on two German simplification datasets, each targeting different users. TextComplexityDE (Naderi et al., 2019) consists of sentences from Wikipedia, which were considered complex by second language German learners. These sentences were then simplified by a native German speaker and checked by these learners. In addition, we created a test set from GEOlino. We extracted 20 articles[4] from three categories: nature, physics, and people. A trained German linguist then simplified the articles, sentence by sentence, to be understandable for children aged between 5–7 years. Examples from the test set can be seen in Table 6.7. Our simplifying instructions can be found in the Appendix D (Section D.1).

Table 6.6 shows various descriptive statistics on our test sets. GEOlino is larger and consists of both single and multiple source sentences. The FRE readability metric (more information can be found in Chapter 2 (Section 2.1)) shows that both the source and target sentence are very simple. We also see moderate amounts of sentence splitting (the number of sentences per instance increases in the simplified target).

[4]Articles were limited to 20 sentences. Half the articles were reserved for a validation set.

Dataset	Source	Target
WMT19	SAN FRANCISCO – Es war noch nie leicht , ein rationales Gespräch über den Wert von Gold zu führen .	SAN FRANCISCO – It has never been easy to have a rational conversation about the value of gold .
WMT19	Ich erkläre die am Freitag , dem 17. Dezember unterbrochene Sitzungsperiode des Europäischen Parlaments für wiederaufgenommen , wünsche Ihnen nochmals alles Gute zum Jahreswechsel und hoffe , daß Sie schöne Ferien hatten .	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999 , and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period .
WikiLarge	Seventy-five defencemen are in the Hall of Fame , more than any other current position , while only 35 goaltenders have been inducted .	Seventy-five defencemen are in the Hall of Fame , more than any other current position , while only 35 goaltenders have been inducted .
GEOlino	Verteilt die Himbeeren in einem Eiswürfelbereiter und füllt sie mit Wasser auf. Stellt sie für mindestens 4 Stunden ins Gefrierfach.	
GEOlino	Seine Mutter möchte ihn also etwas bremsen.	
Simple Wikipedia	It is found in the region Pays de la Loire in the Vendée department in the west of France.	
Simple Wikipedia	For example, Mohandas Gandhi was a very influential person, because the things he did and said changed many peoples lives, and many people believe he has even influenced the world.	

Table 6.4: Examples from the training datasets. Further examples from WikiLarge can be seen in Table 6.1.

	Source	Target	Size
WikiLarge	$\text{English}_{\text{C}}$	$\text{English}_{\text{S}}$	300K
WMT19	$\text{English}_{\text{C}}$	German_{C}	6.0M
GeoLino	—	German_{S}	200K
Wikipedia	—	$\text{English}_{\text{S}}$	1.4M

Table 6.5: Training data used in our experiments; monolingual corpora shown under Target; indices are shorthands for Complex and Simple language.

	TextComplexityDE		GEOlino	
	Source	Target	Source	Target
Length	28.66	29.23	15.68	15.05
Sents	1.09	2.17	1.13	1.55
FRE	28.53	49.3	62.87	68.73
Size	122		663	
TER	67.95		24.12	
Insertions	3.20		0.43	
Deletions	3.17		1.08	
Substitution	9.10		1.54	
Shifts	1.70		0.18	

Table 6.6: Descriptive statistics of test set, including: *Size*, the number of instances; *Length*, average number of words; *Sents*, average number of sentences per instance; average *Flesch Reading Ease (FRE)* score (higher is simpler); *TER*, the translation error rate, measuring distance between source and target and is composed of 4 parts, for which we report the average number of: insertions, deletions, Substitutions and shifts.

Dataset	Simple	Complex
TextComplexityDE	Wegen dieser leichten Vergänglichkeit wurde ‚Seifenblase' zu einer Metapher für etwas, das zwar anziehend, aber dennoch inhalts- und gehaltlos ist.	Weil Seifenblasen nicht lange halten, wurden sie zu einem sprachlichen Ausdruck für etwas, das anziehend aber inhaltslos ist.
TextComplexityDE	Als Gründe dafür, dass Väter ihre Arbeitszeit relativ selten für die Familienarbeit reduzieren, werden u. a. finanzielle Nachteile aufgrund von Gehaltsunterschieden zwischen Männern und Frauen, fehlende Teilzeitstellen für höhere Positionen sowie eine Profitorientierung der Konzerne, die auf familiäre Bedürfnisse der Angestellten keine Rücksicht nehme, genannt.	Väter reduzieren ihre Arbeitszeit relativ selten für die Familienarbeit. Das kann verschiedene Gründe haben. - Sie fürchten finanzielle Nachteile aufgrund von Gehaltsunterschieden zwischen Männern und Frauen - es gibt keine Teilzeitstellen für höhere Positionen - Konzerne nehmen auf familiäre Bedürfnisse der Angestellten keine Rücksicht, da sie profitorientiert organisiert sind
GEOlino	Das Licht der Sonne bestimmt außerdem, wann und wie lange Tiere (und, bis auf Ausnahmen, auch Menschen) aktiv sind beziehungsweise schlafen: Vögel beispielsweise beginnen bei einer bestimmten morgendlichen Helligkeit mit ihrem Gesang.	Das Licht der Sonne bestimmt, wann und wie lange Tiere und Menschen aktiv sind oder schlafen. Vögel beginnen bei einer bestimmten Helligkeit zu singen.
GEOlino	Wie viele andere Tiere kommunizieren Elefanten in Tonlagen, die das menschliche Ohr gar nicht wahrnehmen kann.	Elefanten machen Töne in Tonlagen, die das menschliche Ohr gar nicht wahrnehmen kann.

Table 6.7: Examples from the GEOlino and TextComplexityDE test sets.

TextComplexityDE is more complex, with the source sentences having the lowest FRE score. The target simplifications, while noticeably simpler than the source, are still more complex than GEOlino. We also observe a significant amount of sentence splitting in this dataset. TextComplexityDE also has a significantly higher Translation Error Rate (TER), showing more of each type of operation. However, we should note that GEOLino has approximately the same TER as the WikiLarge test set (25.85).

Evaluation As there is no single agreed-upon metric for simplification (Alva-Manchego et al., 2020; Sulem et al., 2018a); we evaluate the models outputs using a combination of four automatically generated scores, which essentially quantify: a) whether the output is similar to the gold standard reference (*Target*-based, *T*); b) whether the output is similar to the source (*Source*-based, *S*); and c) whether the output is simple on its own, with no regard to preserving the meaning of the original sentence (*Readability*-based, *R*). We indicate the type of each metric using superscripts. BLEU^{T}[5], iBLEU^{T,S}, SARI^{T,S}[6], and FRE-BLEU^{T,S,R}[7], these metrics have been previously shown to correlate with human judgments of simplification quality (Xu et al., 2016)[8], where FRE-BLEU, is newly introduced, and combines metric which reference the target, source and simplicity level. FRE-BLEU (Xu et al., 2016) combines the difference in FKGL of the source and the output and the iBLEU score. FKGL is a shorthand for the Flesch-Kincaid Grade Level readability score which was originally developed for English but has **not** been ported to German. So instead we use the Flesch Reading Ease readability test which has been ported for German (FRE; Amstad 1978) and adapt FK-BLEU to use the difference in FRE. Calculated as $\text{FRE} = 180 - \text{ASL} - (58.5 \cdot \text{ASW})$ where ASL is the average sentence length and ASW the average number of syllables per word. More information on these metrics can be found in Chapter 2 (Section 2.2).

We also evaluated system output by eliciting human judgments via Amazon's Mechanical Turk. Native German speakers (self reported) were asked to rate simplifications on three dimensions: *Grammaticality* (is the output grammatical and fluent?), *Meaning Adequacy* (to what extent is the meaning expressed in the original sentence preserved in the output, with no additional information added?), and *Simplicity* (is the

[5]Scores were calculated at the corpus level using multi-bleu-detok.perl

[6]Corpus level SARI scores were used.

[7]Sentence length was determined using mosses sentence splitter and syllables were counted using hunspell de_DE dictionary.

[8]However we note, that these correlations were calculated on a specific English test set using sentence level metrics.

output a simpler version of the input?). Ratings were obtained using a five point Likert scale. We randomly sampled 100 source sentences from each test set (GEOlino and TextComplexityDE), each sample received five ratings, resulting in 500 judgments per test set. Annotator instructions can be found in the Appendix D (Section D.3).

Model Parameters For all experiments, the base encoder stack consists of six transformer layers, the simplification stack has two layers, the decoder stack six layers, and all other stacks are a single transformer layer. Each layer has a hidden dimension of size 512 and an inner dimension size of 2,048. Word embeddings, size 512, were initialized randomly and shared between the encoder and both decoders. We used eight attentional heads. Dropout was set to 0.1; source word dropout was also set to 0.1. The discriminator consists of a four layer feedforward network with dropout set to 0.1. The networks were optimized using Adam (Kingma and Ba, 2015). Multi-tasking was performed by alternating batches of different tasks. We select a minibatch from a task with a probability inversely proportional to the training loss of the task. This is due to the differing amount of training data, allowing for quicker overfitting on the smaller datasets. Additionally, some tasks such as autoencoding are easier than other tasks, leading to quicker overfitting. One model was selected using the average FRE-BLEU score across all development sets.

All text was preprocessed using the UDPipe tokenization script (Straka, 2018) and truecasing was applied. SentencePiece was subsequently applied to the text, with a SentencePiece vocabulary size of 50,000 and a sampling size of $l = \infty$ and a smoothing parameter of $\alpha = 0.25$ (Kudo, 2018).

6.5 Results

Automatic Evaluation Table 6.8 summarizes our automatic evaluation results. We compare our **ZE**ro-shot cro**S**slingual **S**entence simplifica**T**ion model, which we call ZEST, against multiple unsupervised and supervised baselines. We compare against two unsupervised neural MT models. Surya et al. (2019) (U-NMT) and Surya et al. (2019) (U-SIMP), as described in Section 5.2. Both models were trained on the non-parallel simple and complex German datasets.

We additionally include a supervised baseline based on *pivoting*, which requires three independently trained models: a complex source German sentence is first translated to English (de $\rightarrow$ en); it is then simplified (complex en $\rightarrow$ simple en), before

Models	FRE-BLEU	iBLEU	BLEU	SARI
ZEST	**36.04**	**12.99**	**21.11**	**41.12**
Pivot	28.44	8.09	11.50	38.64
U-SIMP	29.95	8.97	15.03	37.40
U-NMT	26.63	7.09	11.72	35.97

(a) TextComplexityDE

Models	FRE-BLEU	iBLEU	BLEU	SARI
ZEST	62.37	44.72	58.68	39.09
Pivot	39.54	17.81	22.92	27.94
U-SIMP	59.53	**46.33**	**61.10**	**40.00**
U-NMT	**62.57**	39.50	52.02	35.22

(b) GEOlino

Table 6.8: Results using automatic evaluation metrics; best scores for each metric are **boldfaced**.

	TextComplexityDE		GEOlino	
Model	FRE-BLEU	SARI	FRE-BLEU	SARI
ZEST	36.04	41.11	**62.37**	39.09
−ADV	**36.81**	40.47	60.61	**40.98**
−LM	35.46	41.26	57.29	40.33
−AE	35.56	41.60	57.66	36.49
−LM−AE	35.39	**41.71**	55.37	35.42

Table 6.9: Ablation study examining the impact of removing the adversarial (ADV) loss, and then additionally removing the language modeling loss (LM), and autoencoding loss (AE), separately, then together.

	EN→DE	DE→EN
ZEST	32.11	30.90
Pivot	34.15	31.72
SOTA	44.9	42.8

Table 6.10: BLEU scores on the WMT19 test set.

translating it back to German (en → de). All three models consist of a transformer with eight encoder/decoder layers and were trained using the same data as employed in our approach (see Table 6.5; WMT19 and WikiLarge). Results on the WMT19 test for ZEST, Pivot, and the state-of-the-art (Bojar et al., 2017) can be seen in Table 6.10. With regard to English simplification (complex en → simple en), the Pivot system achieved a SARI score of 36.30 on the WikiLarge test, and ZEST 37.78. On the same test set, (Zhang and Lapata, 2017), a commonly used baseline simplification system, obtains a SARI score of 37.26, whereas the state-of-the-art system achieves 41.70 (Martin et al., 2020a). Whilst our systems achieve simplification results below state-of-the-art, it is possible to incorporate the improvements of these models into ZEST, which we leave for further work.

The results in Table 6.8 show that ZEST obtains the highest results for all metrics on TextComplexityDE. U-SIMP achieves the second best FRE-BLEU score, while Pivot achieves the second best SARI. Overall, U-NMT produces the worst results. Results on the GEOlino dataset are more mixed, with no model achieving the highest score across all metrics. ZEST does well across all metrics, scoring the second highest for every metric, whereas the scores for U-SIMP and U-NMT spike on different metrics. U-NMT achieves the best FRE-BLEU score, however, on other metrics it is the second lowest. In contrast U-SIMP has a low FRE-BLEU score but for all other metrics it scores the highest. Pivot receives the lowest scores across all metrics. Example outputs can be seen in Table 6.12 and in the Appendix D (Section D.2).

We further examined the impact different loss functions have on the performance of ZEST and these results are presented in Table 6.9. We see that training only on simplification and translation data (−LM−AE) significantly damages the performance of the model, producing the lowest FRE-BLEU scores and the lowest SARI score on GEOlino. We observed that removal of the autoencoding loss (−AE) led to sentences which strayed too far from the source sentence, thereby losing meaning, whereas removal of the language modeling loss (−LM) led to sentences being too close to the

Models	Mean	Gram	Simp	AVG	Min
Reference	4.35**	4.54**	3.81*	4.23**	3.60**
U-SIMP	2.67**	2.87**	2.80**	2.78**	2.22**
Pivot	3.65**	4.13	**3.67**	3.82*	3.18
ZEST	**4.05**	**4.15**	3.63	**3.94**	**3.23**

(a) TextComplexityDE

Models	Mean	Gram	Simp	AVG	Min
Reference	4.73**	4.75**	3.79**	4.42**	3.69**
U-SIMP	4.19*	4.30**	3.22*	3.90*	3.08**
Pivot	3.69**	3.76**	3.25*	3.45**	2.83**
ZEST	**4.38**	**4.57**	**3.44**	**4.13**	**3.24**

(b) GEOlino

Table 6.11: Mean ratings given to simplifications by human participants; highest ratings for each system are **boldfaced**. Models significantly different from ZEST are marked with $^{*}(p < 0.05)$ and $^{**}(p < 0.01)$. Significance tests were performed using a student *t*-test.

source sentence, resulting in too little simplification. The inclusion of the adversarial loss ($-$ADV) showed a small overall increase in FRE-BLEU and a small decrease in SARI.

Human Evaluation Table 6.11 summarizes the results of the human evaluation. We elicited judgments for three systems, namely ZEST, U-SIMP, and the Pivot-based approach. We also included the gold standard Reference as an upper bound (see the Appendix D (Section D.3) for examples of sentence pairs shown to crowdworkers). We report mean ratings for Meaning adequacy, Grammaticality and Simplicity, their combined average (AVG), and their (average) Minimum value. As mentioned in Chapter 5, we include Minimum because we argue that a simplification is only as good as its weakest dimension.

On TextComplexityDE, ZEST is significantly better than the unsupervised approach across all dimensions. It is on par with Pivot in terms of Grammaticality, Simplicity, and Minimum (ratings are not significantly different). However, ZEST is significantly better in terms of Meaning adequacy, and on average. On GEOlino, ZEST is significantly better against all comparison models on all dimensions. Perhaps unsurprisingly,

Complex	Von hier aus erhaltet ihr einen **[eindrucksvollen]**[1] Rundum-Blick über die ganze Schlucht **[hinweg]**[2] bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
U-SIMP	Von hier eraus ihr haltet einen umfassenden Rundum-Blick über die ganze bis **[hinweg]**[2] hinweg zu hin zu ihren **[5000]**[3] Meter hohen Kraterwände.
Pivot	Von hier genießen Sie einen **[beeindruckenden]**[1] Rundumblick über die gesamte Schlucht bis **[zu]**[2] den 500 m hohen Kraterwänden.
ZEST	Von hier aus erhaltet ihr einen Rundum-Blick auf die ganze Schlucht.

(a) GEOlino

Complex	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] **[Erdwärmenutzung]**[6] **[im]**[2] großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
U-SIMP	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] Erdwärmemer **[im]**[2] großen Stil immer auf eine andere Abkühlung des Gesteins**[)]**[7].
Pivot	Dabei handelt es sich nur um eine **[Verdoppelung]**[6] **[des weltweiten Energiebedarfs]**[5], **[was]**[5] **[bedeutet]**[2], **[dass]**[5] die großflächige **[geothermische]**[8] **[Nutzung]**[6] immer einer lokalen **[Kühlung]**[6] **[des Gesteins]**[4] entspricht.
ZEST	Das bedeutet, **[dass]**[5] Erdwärme im großen Stil immer auf eine lokale Abkühlung **[des]**[2] Gesteins hinausläuft.

(b) TextComplexityDE

Table 6.12: Examples of system output and simplification violations: (1) word has 13+ letters; (2) sentence has 12+ words; (3) high number; (4) genitive; (5) subordinate clauses; (6) abstract words; (7) special characters; (8) difficult/foreign words.

across datasets, participants perceive gold standard simplifications as superior to the output of all comparison models.

Model	lex	SC	RC	pas	suj	gen	spl
Reference	38.7	11.9	10.5	06.8	16.2	12.2	35.5
U-SIMP	41.2	**18.7**	04.3	04.6	07.7	08.0	**03.2**
Pivot	44.9	17.3	07.8	**06.7**	11.6	**14.6**	**03.2**
ZEST	**51.9**	08.3	**11.8**	04.9	**13.0**	5.8	02.3

Table 6.13: Proportion of simplifications on 100 sentences. Simplifications include: lexical (lex), subordinate clause (SC), relative clause (RC), passive voice (pas) subjunctive (suj), genitive (gen), and sentence splitting (spl).

Error analysis We further analysed the types of simplifications produced by each system. We sampled 100 source sentences (50 from each dataset) and elicited judgments from annotators. Guidelines given to the annonators can be found in Appendix D (Section D.4) The annotators were asked to indicate the types of simplification which occurred, including: lexical substitutions, passive to active voice, splitting sentence into multiple sentences, and rewriting to avoid subordinate clauses, relative clauses, the subjunctive mood, and the genitive case. The results in Table 6.13 show that ZEST performs a wide variety of simplifications and produces the largest number of lexical simplifications. While all models produce more lexical substitutions than the references, the references split sentences frequently, whereas in all cases, the models split the sentence minimally. The Pivot model simplifies genitives the most while U-SIMP simplifies subordinate clauses most. ZEST produces the largest number of lexical simplifications, and simplifications related to relative clauses and subjunctives.

Crosslingual Simplification We next explore how different tasks can be combined with no additional training data. We illustrate how our model can be used to tackle the task of both simplifying *and* translating. We now assume that the source complex sentence is in English and the simplified output sentence is in German. As there currently exist no crosslingual German simplification test sets, for evaluation purposes we hand-translated 100 complex sentences from each of the German test sets into English.

Models	FRE-BLEU	iBLEU	BLEU	SARI
ZEST	31.82	10.26	14.29	41.11
Pivot	32.72	10.71	15.19	41.60

(a) TextComplexityDE

Models	FRE-BLEU	iBLEU	BLEU	SARI
ZEST	43.65	19.17	25.00	34.62
Pivot	42.61	18.29	23.78	34.43

(b) GEOlino

Table 6.14: Crosslingual, simplifying English into German, automatic results.

Results[9] can be seen in Table 6.14 and example output in Table 6.15 and the Appendix D (Section D.2). For comparison, we provide the results of Pivot, which requires two independently-trained models: a complex source English sentence is first simplified (complex en $\rightarrow$ simple en), and then translated into German (en $\rightarrow$ de). While the results show that ZEST and Pivot are comparable, the fact that we can train our model on single tasks and then recombine the task-specific layers to allow zero-shot transfer to unseen task combinations opens up exciting new opportunities for future work.

6.6 Summary

Conclusion In this chapter we set out to answer the question: *Can bilingual data be used to transfer sentential paraphrasing training data from one language to another?* As such we developed a general approach for transferring generation data from high- to low-resource languages using language agnostic transformers combined with task specific transformer layers. Experimental results on transferring simplification knowledge from English to German showed that our approach was able to produce significantly better German simplifications than unsupervised and pivot-based approaches. In addition to zero-shot simplification, we showed that our model can generate German simplifications given English input, without any additional training.

Next Chapter The next chapter contains our conclusion, where we summarize our contributions of this thesis and provide possible future direction for research.

[9] Both SARI and FRE-BLEU are monolingual evaluation metrics, as such we use the original German source sentence.

EN Source	The mountain is the watershed on whose flanks the catchment areas of the Pacific Ocean, the Atlantic Ocean over the Gulf of Mexico, and the Arctic Ocean over Hudson Bay, meet.
DE Source	Der Berg ist der Wasserscheidepunkt an dessen Flanken sich die Einzugsgebiete des Pazifischen Ozeans, des Atlantischen Ozeans über den Golf von Mexiko und des Arktischen Ozeans über die Hudson Bay berühren.
EN ZEST	Der Berg ist der Weckschatz, auf dessen Flanken die Fanggebiete des pazifischen Ozeans, des Atlantischen Ozeans über dem Golf von Mexiko, und des Arktischen Ozeans über Hudson Bay, treffen.
DE ZEST	Der Berg ist der Wasserscheidepunkt an dem sich die Einzugsgebiete des Pazifiks, des Atlantischen Ozeans, des Golfs von Mexiko und des Arktischen Ozeans über die Hudson Bay treffen.

(a) TextComplexityDE

EN Source	Without the radiation energy of the sun, plant photosynthesis would not work.
DE Source	Ohne die Strahlungsenergie der Sonne würde die pflanzliche Photosynthese nicht funktionieren.
EN ZEST	Ohne die Strahlungsenergie der Sonne, Pflanzen Photosynthese würde nicht funktionieren.
DE ZEST	Ohne die Strahlungsenergie der Sonne würde die Pflanze nicht funktionieren.

(b) Geolino

Table 6.15: Examples of crosslingual simplification (EN Source → DE ZEST); for comparison, we also show the output of a monolingual system (DE Source → DE ZEST).

Chapter 7

Conclusions and Future work

In this thesis we developed sentential paraphrasing models for a wide range of tasks, users, and languages. We showed how transfer learning, an approach where training data from a related secondary task is adapted for a primary task, can be used to overcome the lack of paraphrasing data. Our motivation stems from the benefits that sentential paraphrasing models offer end-users as well as the ability for these models to improve the robustness of existing NLP frameworks and the acute lack of sentential paraphrasing data, hindering the training of these models. Data is often not available, or only available in select high-resource languages, or the data is not well suited for the target user. As such we developed transfer learning techniques, to adapt adjacent data for paraphrasing and in doing so we answer the following question:

Can we perform paraphrasing tasks with neural sequence-to-sequence models without task-specific paraphrasing data using transfer learning?

To help us answer this question we break it down into three further questions:

1. No supervised data exists for a specific paraphrasing task. Can bilingual data be used as a source of training data for paraphrasing?
2. Supervised paraphrasing data exists in one language but not in another. Can bilingual data can be used to transfer paraphrasing training data from one language to another?
3. Can the output of encoder-decoder paraphrasing models be controlled?

In Chapter 3 we answered question 1, by proposing an unconstrained paraphrasing model, which trained on no paraphrasing data. Instead, it adapted bilingual translation

data, which is available in large quantities and across many languages. To use bilingual data for monolingual paraphrasing we introduced neural pivoting; an approach which combines two pre-trained neural machine translation models. A source sentence is first translated into several *foreign* pivot sentences, these pivots are then simultaneously translated back into the original language, producing a paraphrase. We showed, using human and automatic evaluation, that neural pivoting produces semantically preserving and grammatical paraphrases.

In Chapter 4 we address questions 1 and 3, we extended neural pivoting with a control mechanism to perform sentence compression, again using only bilingual data. To do so, a length control mechanism was added to the translation models, which allowed users to specify the output length. Neural machine translation models were trained using latent variable disentanglement, where the length of the output translation was disentangled from the semantics of the translation. We showed that this approach allowed us to control the output of the neural machine translation model, answering question 3. To produce monolingual compression, a source sentence is first translated into multiple *foreign* pivots, these *foreign* pivots are then simultaneously translated back into the original sentence whilst controlling the target length. We demonstrated, in multiple languages, that this approach is effective for producing compression without compression data, answering question 1. As there is a lack of sentence compression data outside of English, we also produced and released multilingual sentence compression test sets, which we used in the evaluation of our model.

In Chapter 5, we further addressed question 3, by extending the controllability aspect of Chapter 4. We showed how syntax and the lexical-choice of a simplification can be disentangled from its semantics. We trained an encoder-decoder model on supervised simplification data using variable disentanglement, separating the semantics from the lexical and syntactic surface form, allowing a user to control the high-level syntactic output and lexical choices of the models. By adding a control mechanism our model can be trained on general-purpose simplification data whilst producing simplification tailored to a user's specific needs, negating the need to create bespoke training data for every user. We showed that this approach is able to produce good general purpose simplifications, and that by controlling the syntax and lexical choices we are able to adjust the simplicity level of the output.

In Chapter 6 we continued to explore sentence simplification, however, we focused on question 2, where simplification data exists in English, a high-resource language, but not in a low-resource language, German. To overcome this lack of data in German

we proposed an encoder-decoder architecture which learns language-agnostic simplification operations. The model was trained on translation data and English simplification data. It consisted of a shared encoder, which learns a language-agnostic representation of the sentence, this representation was then used as the input to a simplification transformer layer, trained solely on English simplifications. To evaluate our model, we produced and released a German simplification test corpus. We showed, using automatic and human evaluation, that our model can produce German simplifications from English data.

Future work

In this thesis, we tackled several questions regarding transfer learning, showing how existing data can be repurposed. However, there are several avenues which remain open for exploration. For instance, the models presented in this thesis were developed for a subset of paraphrasing task, in a small number of languages and in a mix of domains. It is worth studying and exploring how to improve the proposed models' scalability in terms of supporting more domains, languages, and paraphrasing tasks. Avenues for future research about the scalability for paraphrasing are many and varied. We discuss some promising directions below.

More Languages and Tasks In this thesis, we explored three paraphrasing tasks: unconstrained paraphrasing, sentence compression, and sentence simplification. However, as discussed in Chapter 1 (Section 1.1), there are many other paraphrasing tasks, such as sentence fusion, sentence splitting, and grammar correction, all of which have different data conditions and different challenges associated with them, which could lead to different architectures from those proposed in this thesis. For instance, in Chapter 4 we used variable disentanglement to control the length of the outputs for sentence compression, whereas for simplification, in Chapter 5, we used variable disentanglement to control the outputs syntax and lexical choices. From this we see that the different tasks required disentangling different variables, which we would assume this would hold true for future tasks. Instead of developing a set of controllable variables per task, a broad set of variables to disentangle could be developed which would be applicable for many paraphrasing tasks.

Additionally, throughout this thesis, we have focused on three languages: English, French, and German. These languages were chosen due to the availability of annota-

tors[1]. However, it is worth exploring other languages, particularly those outside the Indo-European language family, as these languages often suffer the most from a lack of data and therefore would benefit the most from transfer learning.

Dataset Creation In Chapter 3 (Section 3.5) we discussed how neural pivoting, and bilingual data can be combined to create large-scale paraphrasing training datasets. Once such datasets are created, standard models can be trained on them, removing the need for specialized architectures. A similar approach could also be used to create sentence compression datasets using the model proposed in Chapter 4. Combining bilingual data and length controlled translation model, the foreign sentence from the bilingual data is translated, while controlling the output length to ensure that it is shorter than the source sentence. The translation and the source sentence are then paired.

Similarly, in Chapter 6, we proposed ZEST, a model which transferred simplification data from English to German, which could also be used to create simplification datasets. Given a corpus of non-parallel German sentence, ZEST would generate simplifications, which then would be paired with the original sentences. ZEST can be trained with bilingual data from many languages, and therefore is able to produce simplifications and datasets in many languages.

Domain Adaptation In Chapters 3, 5, and 6, we used machine translation models trained on WMT data. While this data is collected from multiple sources, the two largest sources, as determined by the number of parallel sentences, are news and parliamentary proceedings. This leads to the models being trained on data which is far away from the desired target domain. For instance, in Chapter 5 models were trained WMT data and evaluated on Wikipedia sentences. Work within machine translation has highlighted the significant detrimental effects of training within one domain and testing in another domain (Koehn and Knowles, 2017; Kobus et al., 2017). Fortunately, as discussed in Chapter 2 (Section 2.3.1), work within machine translation has also shown ways in which to minimize the effect of differing domains, such as back-translation (Sennrich et al., 2016d) or domain control (Kobus et al., 2017). Additionally there exist significant amounts of translation data, from many domains, which were not included within the WMT datasets. In the future, more varied translation data could be

[1]Although the models developed can be applied in any languages, the reality of finding annotators creating dataset for evaluation meant that our experiments were conducted on exclusively European languages

used, as well as using existing machines translation approaches to overcome domain mismatches.

Semi-Supervised Learning Within this thesis we have focused on unsupervised learning, as we worked under the assumption that there was no available training data, often resulting in the need for us to create small scale datasets which we have used as the development and test set. However, instead of using this small amount of data as a development set we could instead use it as supervised training data, augmenting the large amount of unsupervised data the models are trained on. For instance, in Chapter 4 we propose unsupervised sentence compression; future work could focus on taking advantage of the limited amount of supervised training data we already have by combining it with the unsupervised data. In Chapter 6 we introduce a multi-task learning setup for simplification, when there is no German simplification data available. However, if we had a small amount of German simplification data, this could be added as an additional task to the multi-task setup.

Document/Paragraph Level Paraphrasing Within this thesis we have focused on sentence-to-sentence paraphrasing, which is simpler, computationally cheaper, and easier to create datasets for than document/paragraph level rewriting. However, restricting rewrites to individual sentences limits the amount of context available to model, and it limits the flexibility of the range of possible rewrites. As such we propose future work could examine paragraph or document level paraphrasing, which has been successfully applied to machine translation tasks Läubli et al. (2018); Junczys-Dowmunt (2019); Miculicich et al. (2018). Not only would working at the document/paragraph level offer greater context, but it would also allow the model to reorder the document, merge sentences, or even drop sentences entirely, where all three of these operations are applicable to both simplification, and compression.

For instance, pivoting approaches of Chapters 3 and 4 could use document level machine translation models instead of sentence level models, thus performing document level paraphrasing. The simplification models in Chapters 5 and 6 are trained on simplification created by aligning sentences between two documents; hence, instead of using aligned sentences the entire documents could be used to train the simplification models.

Bibliography

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., and Specia, L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., and Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *6th International Conference on Learning Representations,*

ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Ba, L. J., Kiros, R., and Hinton, G. E. (2016). Layer normalization. *CoRR*, abs/1607.06450.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, California.

Baker, M. et al. (1993). Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussÃ, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MÃller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Baziotis, C., Androutsopoulos, I., Konstas, I., and Potamianos, A. (2019). SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *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 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

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, Baltimore, Maryland. Association for Computational Linguistics.

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA. Association for Computational Linguistics.

Bernhard, D. and Gurevych, I. (2008). Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–52, Columbus, Ohio. Association for Computational Linguistics.

Bhagat, R. and Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Bingel, J., Paetzold, G., and Søgaard, A. (2018). Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Blackwood, G., Ballesteros, M., and Ward, T. (2018). Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O. and Tamchyna, A. (2015). CUNI in WMT15: Chimera strikes again. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 79–83, Lisbon, Portugal. Association for Computational Linguistics.

Brad, F. and Rebedea, T. (2017). Neural paraphrase generation using transfer learning. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 257–261, Santiago de Compostela, Spain. Association for Computational Linguistics.

Britz, D., Le, Q., and Pryzant, R. (2017). Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

Brockett, C. and Dolan, W. B. (2005). Support vector machines for paraphrase identification and corpus construction. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language*

Processing, pages 351–361, Austin, Texas. Association for Computational Linguistics.

Brunato, D., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2015). Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA. Association for Computational Linguistics.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii. Association for Computational Linguistics.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C., and Aluísio, S. M. (2009). Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, 41:59–70.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Chali, Y., Tanvee, M., and Nayeem, M. T. (2017). Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–424, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Chatterjee, R., Negri, M., Rubino, R., and Turchi, M. (2018). Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Chen, X., Awadallah, A. H., Hassan, H., Wang, W., and Cardie, C. (2019). Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chopra, S., Auli, M., and Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chu, C. and Wang, R. (2018). A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Cífka, O., Severyn, A., Alfonseca, E., and Filippova, K. (2018). Eval all, trust a few, do wrong to none: Comparing sentence generation models. *arXiv preprint arXiv:1804.07972*.

Clarke, J. and Lapata, M. (2006). Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 377–384, Sydney, Australia. Association for Computational Linguistics.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *J. Artif. Intell. Res.*, 31:399–429.

Cohn, T. and Lapata, M. (2009a). Sentence compression as tree transduction. *J. Artif. Intell. Res.*, 34:637–674.

Cohn, T. and Lapata, M. (2013). An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology*, 4(3):1–35.

Cohn, T. A. and Lapata, M. (2009b). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Corston-Oliver, S. (2001). Text Compaction for Display on Very Small Screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 89–98, Pittsburgh, Pennsylvania.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evalu-*

ation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

da Fonseca, J. and Carolino, P. (1855). *O novo guia da conversação, em portuguez e inglez: ou, Escolha de dialogos familiares sôbre varios assumptos*. Va. J.P. Aillaud, Monlon e Ca.

de Loupy, C., Guégan, M., Ayache, C., Seng, S., and Moreno, J.-M. T. (2010). A French human reference corpus for multi-document summarization and sentence compression. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Deléger, L. and Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10, Singapore. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.

Devlin, S. (1999). *Simplifying Natural Language for Aphasic Readers*. PhD thesis, University of Sunderland.

Ding, S., Duh, K., Khayrallah, H., Koehn, P., and Post, M. (2016). The JHU machine translation systems for WMT 2016. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 272–280, Berlin, Germany. Association for Computational Linguistics.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017). Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Dong, Y., Li, Z., Rezagholizadeh, M., and Cheung, J. C. K. (2019). EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 1–8.

Dras, M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text*. Macquarie University Sydney.

Duan, X., Yin, M., Zhang, M., Chen, B., and Luo, W. (2019a). Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Duan, Y., Xu, C., Pei, J., Han, J., and Li, C. (2019b). Pre-train and plug-in: Flexible conditional text generation with variational auto-encoders. *arXiv preprint arXiv:1911.03882*.

Durrani, N., Schmid, H., and Fraser, A. (2011). A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Eyecioglu, A. and Keller, B. (2015). Twitter paraphrase identification with simple overlap features and SVMs. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 64–69, Denver, Colorado. Association for Computational Linguistics.

Fader, A., Zettlemoyer, L., and Etzioni, O. (2013). Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Fan, A., Grangier, D., and Auli, M. (2018). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Ficler, J. and Goldberg, Y. (2017). Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.

Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Firat, O., Cho, K., and Bengio, Y. (2016a). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F. T., and Cho, K. (2016b). 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, Austin, Texas. Association for Computational Linguistics.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Galley, M. and McKeown, K. (2007). Lexicalized Markov grammars for sentence compression. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 180–187, Rochester, New York. Association for Computational Linguistics.

Ganin, Y. and Lempitsky, V. S. (2015). Unsupervised domain adaptation by backpropagation. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Ganitkevic, J. et al. (2018). *Large-Scale Paraphrasing for Text-to-Text Generation*. PhD thesis, Johns Hopkins University.

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. European Language Resources Association (ELRA).

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language*

Processing, pages 1168–1179, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Gasperin, C., Specia, L., Pereira, T., and Aluísio, S. (2009). Learning when to simplify sentences for natural text simplification. *Proceedings of ENIA*, pages 809–818.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. 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 1243–1252. PMLR.

Gers, F. A., Schmidhuber, J., and Cummins, F. A. (2000). Learning to forget: Continual prediction with LSTM. *Neural Comput.*, 12(10):2451–2471.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

Grangier, D. and Auli, M. (2018). QuickEdit: Editing text & translations by crossing words out. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 272–282, New Orleans, Louisiana. Association for Computational Linguistics.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guo, H., Pasunuru, R., and Bansal, M. (2018). Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference*

on Computational Linguistics, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Guo, W. and Diab, M. (2012). A simple unsupervised latent semantics based approach for sentence similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 586–590.

Ha, T., Niehues, J., and Waibel, A. H. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Harashima, J. and Kurohashi, S. (2012). Flexible Japanese sentence compression by relaxing unit constraints. In *Proceedings of COLING 2012*, pages 1097–1112, Mumbai, India. The COLING 2012 Organizing Committee.

Hasegawa, S., Kikuchi, Y., Takamura, H., and Okumura, M. (2017). Japanese sentence compression with a large training dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Vancouver, Canada. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

He, W., Liu, Z., and Liu, T. (2013). Paraphrase based similar expression generation. In *Intelligence Computation and Evolutionary Computation*, pages 369–374. Springer.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., and Lerchner, A. (2016). Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*.

Hirao, T., Suzuki, J., and Isozaki, H. (2009). A syntax-free approach to Japanese sentence compression. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 826–833, Suntec, Singapore. Association for Computational Linguistics.

Hori, C. and Furui, S. (2004). Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems*, E87-D(1):15–25.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Hu, J. E., Rudinger, R., Post, M., and Durme, B. V. (2019). PARABANK: monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6521–6528. AAAI Press.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596. PMLR.

Huang, S., Graff, D., Doddington, G., Consortium, L. D., et al. (2002). *Multiple-translation Chinese corpus*. Linguistic Data Consortium, University of Pennsylvania.

Huang, S., Graff, D., Walker, K., Miller, D., Ma, X., Cieri, C., and Doddington, G. (2003). Multiple-translation chinese (mtc) part 2 ldc2003t17. web download. *Philadelphia: Linguistic Data Consortium*.

Huang, S., Wu, Y., Wei, F., and Luan, Z. (2019). Dictionary-guided editing networks for paraphrase generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence*

Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6546–6553. AAAI Press.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16, Sapporo, Japan. Association for Computational Linguistics.

Iyyer, M., Wieting, J., Gimpel, K., and Zettlemoyer, L. (2018). Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. (2020). Neural crf model for sentence alignment in text simplification. *arXiv preprint arXiv:2005.02324*.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Sixth Applied Natural Language Processing Conference*, pages 310–315, Seattle, Washington, USA. Association for Computational Linguistics.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. (2018). Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multi-

lingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Junczys-Dowmunt, M. (2019). Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. *arXiv preprint arXiv:1907.06170*.

Kaji, N., Kawahara, D., Kurohashi, S., and Sato, S. (2002). Verb paraphrase based on case frame alignment. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 215–222, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A., and Kavukcuoglu, K. (2016). Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria.

Kay, M. (1997). The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3–23.

Khosmood, F. (2012). Comparison of sentence-level paraphrasing approaches for statistical style transformation. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer ....

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. (2016). Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria. Association for Computational Linguistics.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text simplification for information-seeking applications. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 735–747. Springer.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Kouris, P., Alexandridis, G., and Stafylopatis, A. (2019). Abstractive text summarization based on deep learning and semantic content generalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5092, Florence, Italy. Association for Computational Linguistics.

Kozlowski, R., McCoy, K. F., and Vijay-Shanker, K. (2003). Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the second international workshop on Paraphrasing*, pages 1–8.

Kriz, R., Sedoc, J., Apidianaki, M., Zheng, C., Kumar, G., Miltsakaki, E., and Callison-Burch, C. (2019). Complexity-weighted loss and diverse reranking for sentence simplification. In *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 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing:*

System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G., and Jyothi, P. (2019). Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Lan, W., Qiu, S., He, H., and Xu, W. (2017). A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*.

Lee, J. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lennon, C. and Burdick, H. (2004). The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Li, Z., Jiang, X., Shang, L., and Li, H. (2018). Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Li, Z., Jiang, X., Shang, L., and Liu, Q. (2019). Decomposable neural paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3403–3414, Florence, Italy. Association for Computational Linguistics.

Lin, C.-Y. (2004a). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, C.-Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Liu, X., Mou, L., Meng, F., Zhou, H., Zhou, J., and Song, S. (2020). Unsupervised paraphrasing by simulated annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 302–312, Online. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., and Sun, J. (2018). A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Luong, M.-T., Pham, H., and Manning, C. D. (2015a). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Luong, T., Pham, H., and Manning, C. D. (2015b). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Ma, X. (2004). Multiple-translation chinese (mtc) part 3. *Linguistic Data Consortium*.

Ma, X. (2006). Multiple-translation chinese (mtc) part 4. *Linguistic Data Consortium*.

Macháček, M. and Bojar, O. (2013). Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Madnani, N., Fazil Ayan, N., Resnik, P., and Dorr, B. (2007a). Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Madnani, N., Zajic, D., Dorr, B., Ayan, N. F., and Lin, J. (2007b). Multiple alternative sentence compressions for automatic text summarization. In *Proceedings of the 2007 Document Understanding Conference (DUC-2007) at NLT/NAACL*, page 24.

Mallinson, J. (2016). PARANET:Bilingual Encoder-Decoder Paraphrasing. Master's thesis, University of Edinburgh.

Mallinson, J. and Lapata, M. (2019). Controllable sentence simplification: Employing syntactic and lexical constraints. *arXiv preprint arXiv:1910.04387*.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *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. (2018). Sentence compression for arbitrary languages via multilingual pivoting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2453–2464, Brussels, Belgium. Association for Computational Linguistics.

Mallinson, J., Sennrich, R., and Lapata, M. (2020a). Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Mallinson, J., Severyn, A., Malmi, E., and Garrido, G. (2020b). FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

Malmi, E., Krause, S., Rothe, S., Mirylenka, D., and Severyn, A. (2019). Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Mani, I. (2001). *Automatic summarization*, volume 3. John Benjamins Publishing.

Mao, H.-R. and Lee, H.-Y. (2019). Polly want a cracker: Analyzing performance of parroting on paraphrase generation datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5960–5968, Hong Kong, China. Association for Computational Linguistics.

Marie, B., Allauzen, A., Burlot, F., Do, Q.-K., Ive, J., Knyazeva, E., Labeau, M., Lavergne, T., Löser, K., Pécheux, N., and Yvon, F. (2015). LIMSI@WMT'15 : Translation task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 145–151, Lisbon, Portugal. Association for Computational Linguistics.

Martin, L., de la Clergerie, É., Sagot, B., and Bordes, A. (2020a). Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2020b). Multilingual unsupervised sentence simplification. *arXiv preprint arXiv:2005.00352*.

McDonald, R. (2006). Discriminative sentence compression with soft syntactic constraints. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–304, Trento, Italy.

Mehta, S., Azarnoush, B., Chen, B., Saluja, A., Misra, V., Bihani, B., and Kumar, R. (2020). Simplify-then-translate: Automatic preprocessing for black-box machine translation. *arXiv preprint arXiv:2005.11197*.

Michel, P. and Neubig, G. (2018). Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.

Monteiro, G. (2004). English as she is spoke: 150 years of a classic. *Luso-Brazilian Review*, 41(1):191–198.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Naderi, B., Mohtaj, S., Ensikat, K., and Möller, S. (2019). Subjective assessment of text complexity: A dataset for german language. *CoRR*, abs/1904.07733.

Napoles, C., Callison-Burch, C., Ganitkevitch, J., and Van Durme, B. (2011). Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 84–90, Portland, Oregon. Association for Computational Linguistics.

Napoles, C., Callison-Burch, C., and Post, M. (2016). Sentential paraphrasing as black-box machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 62–66, San Diego, California. Association for Computational Linguistics.

Napoles-Cohen, C. (2019). *Monolingual Sentence Rewriting as Machine Translation: Generation and Evaluation*. PhD thesis, Johns Hopkins University.

Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Nishihara, D., Kajiwara, T., and Arase, Y. (2019). Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Statistical machine translation. In *5th EAMT Workshop: Harvesting Existing Resources*, Ljubljana, Slovenia. European Association for Machine Translation.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Paetzold, G. and Specia, L. (2017). Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40, Valencia, Spain. Association for Computational Linguistics.

Palmero Aprosio, A., Tonelli, S., Turchi, M., Negri, M., and Di Gangi, M. A. (2019). Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural*

Language Generation, pages 37–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Paulus, R., Xiong, C., and Socher, R. (2018). A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Pavlick, E. and Callison-Burch, C. (2016). Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany. Association for Computational Linguistics.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pershina, M., He, Y., and Grishman, R. (2015). Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. Association for Computational Linguistics.

Peter, J.-T., Guta, A., Alkhouli, T., Bahar, P., Rosendahl, J., Rossenbach, N., Graça, M., and Ney, H. (2017). The RWTH Aachen University English-German and German-English machine translation system for WMT 2017. In *Proceedings of the Second*

Conference on Machine Translation, pages 358–365, Copenhagen, Denmark. Association for Computational Linguistics.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Puduppully, R., Dong, L., and Lapata, M. (2019a). Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Puduppully, R., Mallinson, J., and Lapata, M. (2019b). University of edinburgh's submission to the document-level generation and translation shared task. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 268–272.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Reed, S. E. and de Freitas, N. (2016). Neural programmer-interpreters. 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*.

Rello, L., Baeza-Yates, R., Dempere-Marco, L., and Saggion, H. (2013). Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proc. INTERACT*.

Riezler, S., King, T. H., Crouch, R., and Zaenen, A. (2003). Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for

Lexical-Functional Grammar. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 197–204.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Roy, A. and Grangier, D. (2019). Unsupervised paraphrasing without translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6033–6039, Florence, Italy. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Scarton, C., Paetzold, G., and Specia, L. (2018). Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Scarton, C., Palmero Aprosio, A., Tonelli, S., Martín Wanton, T., and Specia, L. (2017). MUSST: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Scarton, C. and Specia, L. (2018). Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A. V. M., Mokry, J., et al. (2017). Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016d). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shen, S., Chen, Y., Yang, C., Liu, Z., and Sun, M. (2018). Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 26(12):2319–2327.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. S. (2017). Style transfer from non-parallel text by cross-alignment. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances*

in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6830–6841.

Shewan, C. M. and Canter, G. J. (1971). Effects of vocabulary, syntax, and sentence length on auditory comprehension in aphasic patients. *Cortex*, 7(3).

Siddharthan, A. (2004). Syntactic simplification and text cohesion". in research on language and computation. *Research on Language and Computation*, 4(1):77–109.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Citeseer.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Štajner, S. and Saggion, H. (2018). Data-driven text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 19–23, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Sulem, E., Abend, O., and Rappoport, A. (2018a). Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Sulem, E., Abend, O., and Rappoport, A. (2018b). Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Sultan, M. A., Bethard, S., and Sumner, T. (2014). DLS@CU: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland. Association for Computational Linguistics.

Sun, H. and Zhou, M. (2012). Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Surya, S., Mishra, A., Laha, A., Jain, P., and Sankaranarayanan, K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

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. 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 3104–3112.

Takase, S. and Kobayashi, S. (2020). All word embeddings from one embedding. *arXiv preprint arXiv:2004.12073*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Tonelli, S., Aprosio, A. P., and Saltori, F. (2016). Simpitiki: a simplification corpus for italian. *Proc. of CLiC-it*.

Toutanova, K., Brockett, C., Tran, K. M., and Amershi, S. (2016). A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

Turner, J. and Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 290–297, Ann Arbor, Michigan. Association for Computational Linguistics.

Tutek, M. and Šnajder, J. (2020). Staying true to your word:(how) can attention become explanation? *arXiv preprint arXiv:2005.09379*.

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vickrey, D. and Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Wang, R., Utiyama, M., Liu, L., Chen, K., and Sumita, E. (2017). Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Wang, S., Gupta, R., Chang, N., and Baldridge, J. (2019c). A task in a suit and a tie: paraphrase generation with semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7176–7183.

Wang, Y., Berant, J., and Liang, P. (2015). Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

Weese, J., Ganitkevitch, J., and Callison-Burch, C. (2014). Paradigm: Paraphrase diagnostics through grammar matching. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 192–201.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Towards universal paraphrastic sentence embeddings. In Bengio, Y. and LeCun, Y., editors, *4th Interna-*

tional Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

Wieting, J. and Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Wieting, J., Mallinson, J., and Gimpel, K. (2017). Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Woodsend, K. and Lapata, M. (2010). Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden. Association for Computational Linguistics.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Woodsend, K. and Lapata, M. (2014). Text rewriting improves semantic role labeling. *Journal of Artificial Intelligence Research*, 51:133–164.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Dolan, B. (2015a). SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International*

Workshop on Semantic Evaluation (SemEval 2015), pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Napoles, C. (2015b). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, California. Association for Computational Linguistics.

Zajic, D., Dorr, B., and Schwartz, R. (2004). Bbn/umd at duc-2004: Topiary. In *In Proceedings of the NAACL Workshop on Document Understanding*, pages 112–119, Boston, MA.

Zarrella, G., Henderson, J., Merkhofer, E. M., and Strickhart, L. (2015). MITRE: Seven systems for semantic similarity in tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 12–17, Denver, Colorado. Association for Computational Linguistics.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Zhang, B., Titov, I., and Sennrich, R. (2019). Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhao, S., Meng, R., He, D., Saptono, A., and Parmanto, B. (2018a). Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

Zhao, S., Wang, H., Liu, T., and Li, S. (2008). Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL-08: HLT*, pages 780–788, Columbus, Ohio. Association for Computational Linguistics.

Zhao, Y., Bi, W., Cai, D., Liu, X., Tu, K., and Shi, S. (2018b). Language style transfer from sentences with arbitrary unknown styles. *arXiv preprint arXiv:1808.04071*.

Zhao, Y., Chen, L., Chen, Z., and Yu, K. (2020). Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *AAAI*, pages 9668–9675.

Zhou, Q., Yang, N., Wei, F., and Zhou, M. (2017). Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.

Zhou, X., Wan, X., and Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

Appendix A

Paraphrasing with Neural Pivoting

A.1 Paraphrase examples

Tables A.1–A.3 show examples of PARANET output on the Wikianswers, Leagues, and MTC datasets.

	Wikianswers
Source.	How many calories in a handful of strawberries?
PARANET.	The number of calories in a handful of strawberries.
Source.	Beauty is not in the eye of the beholder.
PARANET.	Beauty is not in the mind of the viewer.
Source.	What is the importance of employee satisfaction in an organization?
PARANET.	What is the significance of staff satisfaction at an organisation?
Source.	What is the difference between electrical power and electrical energy?
PARANET.	What is the difference between electrical energy and electrical power?
Source.	How many high tides happen at a given coast in any 24 hour period?
PARANET.	How many high tides occur on a certain coast in 24 hours?
Source.	What is a beverage that starts with the letter p?
PARANET.	What is a drink that begins with the letter p?
Source.	What Swiss mathematician and teacher was responsible for instituting the use of the symbol for π in mathematical notation?
PARANET.	What Swiss mathematicians and teachers were responsible for the introduction of the symbol for π in math notation?
Source.	How do you make a pina colada?
PARANET.	How do you do a Pina colada?
Source.	What is the difference between a captain and a skipper?
PARANET.	What is the difference between being a captain and skipper?

Table A.1: Example PARANET paraphrases on the Wikianswers test set.

Leagues	
Source.	“Faith i should never have believed it,” said Conseil.
PARANET.	“Faith, I never would have believed”, Conseil said.
Source.	“I owed myself this revenge!” Said the Captain to the Canadian.
PARANET.	“I am indebted to this revenge!” the captain told the Canadian.
Source.	“Well, sir, you will only get your deserts.”
PARANET.	“Well, sir, you are only getting your deserts.”
Source.	“That’s what I’ve been telling you Ned.”
PARANET.	“That’s what I said, Ned.”
Source.	Very much embarrassed, after having vainly exhausted our speaking resources, I knew not what part to take, when Conseil said: “if master will permit me I will relate it in German.”
PARANET.	It was very embarrassing that I had used up our speaking time, and I did not know what to do, as Conseil said: “If the Masters allow me, I shall refer to German.”
Source.	Almost every day the panels in the lounge were open for some hours, and our eyes never tired of probing the mysteries of the underwater world.
PARANET.	Almost every day, the panels opened in the lounge for a few hours, and our eyes never tired, the secrets of the underwater world.
Source.	I bowed, and desired to be conducted to the cabin destined for me.
PARANET.	I bow to and wish I headed to the cabin for me.
Source.	I had one question left to address to Captain Nemo.
PARANET.	I had a question left to Captain Nemo.
Source.	“I have not the foggiest notion, Professor Aronnax.”
PARANET.	I have no idea, Professor Aronnax.

Table A.2: Example PARANET paraphrases on the Leagues test set.

	MTC
Source.	China expresses strong dissatisfaction over the Japanese leader's move this time.
PARANET.	China expresses a strong dissatisfaction over Japanese leader's move.
Source.	We will accelerate the drafting of telecommunications legalization, amend the law of post and the regulations governing wireless telecommunications.
PARANET.	We will speed up the design of telecommunications, change the law and regulations governing wireless telecommunication.
Source.	Liu said: the poverty-stricken areas are badly hit in the first stage of this year's floods and many counties and cities are listed as the poorest ones in the country.
PARANET.	Liu said: poverty-stricken areas are hit hard in the first phase of this year's flooding and many counties and towns are listed as the poorest in the country.
Source.	(London, AP) The British government is working on resolving the increasingly serious problems of street crimes and will strengthen patrolling police.
PARANET.	London, AP The British government is working to resolve the increasingly serious problems of street crime and will strengthen patrols.
Source.	Kida said that the dead killed by the heat wave were mostly old people with heart diseases.
PARANET.	Kida said the dead by heatwave were mostly old people with heart disease.
Source.	Growth of Mobile Phone users in Mainland China to Slow Down.
PARANET.	Growth of Mobile Phone users in Mainland China on Slow Down.
Source.	The survey report said that in the first six months of last year 18 sandstorms struck Beijing and they all originated from Inner Mongolia where 60% of the land is desert.
PARANET.	The report said that 18 sandstorms struck Beijing in the first six months of last year, and they were all from Inner Mongolia, where 60% of the desert is desert.
Source.	The World Cup co-host by Japan and South Korea, will inaugurate on May 31.
PARANET.	The World Cup, co-host Japan and South Korea, will open on May 31.
Source.	Two days ago, President Bush seemed opposed to this idea when he held talks with Sharon.
PARANET.	Two days ago President Bush opposed this idea when he talks to Sharon.
Source.	Russia Faces Population Crisis.
PARANET.	Russia's demographics problem.
Source.	Computer Crimes Cost US billions of Dollars Last Year.
PARANET.	Computer Crimes Cost American Billions of Dollars.
Source.	However, many sports associations in Chile hope to cooperate with China not just for the table tennis alone.
PARANET.	However, many sports federations in Chile are hoping to collaborate with China, not only for the table tennis players.

Table A.3: Example PARANET paraphrases on the MTC test set.

A.2 Evaluation instructions

In this task, you will look at at a series of sentences and their corresponding paraphrases, which have been created by different computer programs. Each sentence will be presented alongside three paraphrase sentences. Paraphrases are alternative ways to convey the same information. Your task is to rank these paraphrases from Best to Worst using a 1-3 rating scale, where 1 is best and 3 is worst (ties are allowed). There are no "correct" answers, so whatever choice seems appropriate to you is a valid response.

In general, you should give a paraphrase a high rank (i.e., 1) if it faithfully captures the same information (gist) as the sentence, and is grammatical (i.e., written in well-formed English). You should give a paraphrase a low rank (i.e., 3) if it is false, i.e., the paraphrase has a different meaning or if it is missing information which is important in the sentence. Furthermore, you should give a low rank to those paraphrases which are too similar to the original sentence e.g., deletion of a single world or just changing the punctuation.

Appendix B

Sentence Compression with Neural Pivoting

B.1 Sentence Compression Examples

In the following tables we provide sample outputs from all domains and languages for our strongest pivot based model, the strongest baseline model and the **Gold** human produced compressions. Where **SP**: single pivot models; $\mathcal{L}$: length parameter; **seq2seq**: encoder-decoder model trained on Gigaword; pivot languages: English (**en**), French (**fr**), German (**de**).

MOSS TED
Source: It was very cold. **seq2seq**: The cold in the cold of the cold. **SP**$_{\mathcal{L},de}$: It was cold. **Gold**: It was freezing.
Source: and the thermometer on the front porch read minus 40 degrees - **seq2seq**: a guide to the summer **SP**$_{\mathcal{L},de}$: and thermometers on front read minus 40 - **Gold**: Thermometer in the front reads -40
Source: Because I had an early flight to Europe the next morning, **seq2seq**: Early flight to Europe **SP**$_{\mathcal{L},de}$: Because I had an early flight to Europe. **Gold**: Because I was flying to Europe the next morning,
Source: Your physician says, "You know, I think you have some depression. **seq2seq**: You know about depression **SP**$_{\mathcal{L},de}$: Your doctor says: I think you have depression. **Gold**: The doctor thinks you have depression.
Source: So I quickly ran around and tried all the other doors and windows, **seq2seq**: All the doors open for windows **SP**$_{\mathcal{L},de}$: So I ran quickly and tried all other doors and windows, **Gold**: I sped round trying every door and window.

MOSS News
Source: The reaction of the European parliament and, I hope, of the European union, will be clear. **seq2seq**: EU parliament's reaction to EU parliament will be clear **SP**$_{\mathcal{L},de}$: How the European parliament will be clear of the EU **Gold**: I hope that the European union will be clear.
Source: It offers a vital opportunity for the leaders of both the industrialised world and the emerging economies to work together on a common agenda for immediate economic stability and longer-term recovery. **seq2seq**: Emerging economies to work together on economic stability. **SP**$_{\mathcal{L},de}$: Emerging economies offer vital opportunity for emerging economies to work together on economic stability and recovery efforts. **Gold**: It is an important opportunity for the leaders of the industrialised countries and emerging economies.
Source: We have reiterated the most important principles in our report which should underscore our foreign policy. **seq2seq**: The most important principles of foreign policy. **SP**$_{\mathcal{L},de}$: A look at the world's foreign policy. **Gold**: Our report reaffirms our foreign policy principles.
Source: The prime minister will make a statement to the european parliament on the preparations for next week's G20 summit . **seq2seq**: prime minister to make statement to european parliament on preparations for summit **SP**$_{\mathcal{L},de}$: Prime minister to make statement on preparations for next european parliament. **Gold**: The prime minister will make a statement next week.

MOSS EUPar
Source: Mr President, I believe that the level of interest in this matter far outweighs the crowd of people that have hurried. **seq2seq**: It's time for President **SP**$_{\mathcal{L},de}$: Mr President, I believe that interest in this matter far exceeds the crowds of people in this House this afternoon. **Gold**: Mr. President, the level of interest in this matter outweighs the crowd of people who hurried into the Chamber, which I regret for both groups.
Source: Mr President, Guinea-Conakry needs the support of the international community. **seq2seq**: President calls for international community to support international community **SP**$_{\mathcal{L},de}$: Mr President, Guinea needs international support. **Gold**: They need international support, Mr President.
Source: Following the death of President Conté, a military junta seized power. **seq2seq**: Military junta seizes power in ivory coast. **SP**$_{\mathcal{L},de}$: A military junta took power. **Gold**: A military junta seized power.
Source: Indeed, I believe that the Union's foreign policy is worthy of attention. **seq2seq**: The union 's foreign policy. **SP**$_{\mathcal{L},de}$: I think Union foreign policy deserves attention. **Gold**: Unions foreign policy is worth attention

DUC-2004
Source: King Norodom Sihanouk has declined requests to chair a summit of Cambodia's top political leaders, saying the meeting would not bring any progress in deadlocked negotiations to form a government. **SP**$_{\mathcal{L},de}$: King Norodom Sihanouk has refused to chair Cambodia summit. **Gold**: Sihanouk refuses to chair Cambodian political summit at home or abroad
Source: Cambodia's ruling party responded Tuesday to criticisms of its leader in the U.S. Congress with a lengthy defense of strongman Hun Sen's human rights record. **SP**$_{\mathcal{L},de}$: Cambodia's ruling party responded Tuesday to criticism of its leader in the US. **Gold**: Cambodian party defends leader Hun Sen against criticism of U.S. House

MOSS Books
Source: The establishment of Europol was agreed in the Maastricht Treaty on European Union of 7 February 1992. **seq2seq**: Europol agrees in Maastricht Treaty on 1992. **SP**$_{\mathcal{L},de}$: The Maastricht Treaty was agreed on 7 February 1992. **Gold**: The establishment of Europol was agreed in the Maastricht Treaty on European Union of 02/07/1992.
Source: We assist partner countries in developing quality education and training systems and in putting them into practice. **seq2seq**: Helping partner countries in developing quality education. **SP**$_{\mathcal{L},de}$: We support partner countries in developing quality and training systems. **Gold**: We assist partner countries in developing quality education and training and implementing these.
Source: We work on behalf of the European Union institutions, particularly the European Commission. **seq2seq**: The EU commission on behalf of EU institutions. **SP**$_{\mathcal{L},de}$: We work on behalf of the European Commission. **Gold**: We work with European union institutions, even the European Commission.

DUC-2004
Source: The Swiss government has ordered no investigation of possible bank accounts belonging to former Chilean dictator Augusto Pinochet, a spokesman said Wednesday. **SP**$_{\mathcal{L},de}$: Swiss government ordered no inquiry into possible bank accounts of former Chilean dictator Augusto. **Gold**: Switzerland joins charges against Pinochet but avoids bank probe
Source: Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean dictator has diplomatic immunity is ridiculous. **SP**$_{\mathcal{L},de}$: Britain has defended its arrest of Augusto Pinochet: Chile claims absurd. **Gold**: Britain defends, Chile condemns, arrest of Pinochet in London

MOSS TED
Source: Die Bedrohung hat einen merkwürdigen Ursprung: das organisierte Verbrechen. **seq2seq**$_{en}$: Organisierte kriminalität ist eine bedrohung für das organisierte verbrechen. **SP**$_{\mathcal{L},en}$: Diese Bedrohung beruht auf organisierten Verbrechen. **Gold**: Die Bedrohung hat einen Ursprung: das organisierte Verbrechen.
Source: Das organisierte Verbrechen ist stets auf der Suche nach solchen Möglichkeiten und wird immer wieder fündig. **seq2seq**$_{en}$: Das organisierte verbrechen der kriminalität **SP**$_{\mathcal{L},en}$: Organisierte Kriminalität sucht solche Möglichkeiten immer noch immer wieder. **Gold**: Das organisierte Verbrechen ist sucht stets nach solchen Möglichkeiten und wird immer wieder fündig.
Source: Aber die meisten Menschen sehen die Dinge anders. **seq2seq**$_{en}$: Die meisten menschen sehen dinge anders, aber die meisten menschen sehen sich anders. **SP**$_{\mathcal{L},en}$: Die meisten sehen anders. **Gold**: Menschen sind intressenlos.
Source: Die Desillusionierung mit der Politik und den Politikern ist ebenfalls ein Schlüsselelement unserer Sehnsucht nach vom Schicksal begünstigten Führern. **seq2seq**$_{en}$: Politische parteien in der nähe der politik **SP**$_{\mathcal{L},en}$: Die Enttäuschung über Politik und Politik ist auch ein Schlüsselelement unserer langjährigen Machthaber. **Gold**: Die Desillusionierung der Politik und den Politikern ist ein Schlüsselelement unserer Sehnsucht nach vom Schicksal begünstigten Führern.

MOSS News
Source: Dass es mir wirklich egal war, ob mich **seq2seq**$_{en}$: Es ist zeit für mich. **SP**$_{\mathcal{L},en}$: Dass mir wirklich egal war **Gold**: Dass es mir wirklich egal war, ob mich
Source: Zwischen dem Griff nach den Sternen **seq2seq**$_{en}$: Es ist zeit für mich. **SP**$_{\mathcal{L},en}$: Zwischen den Sternen **Gold**: Zwischen dem Griff nach den Sternen
Source: Brach ich in mein eigenes Haus ein. **seq2seq**$_{en}$: Wie man ein eigenes haus ist. **SP**$_{\mathcal{L},en}$: Ich brach ins Haus. **Gold**: Brach ich in mein Haus ein.

MOSS Books
Source: Der Verwaltungsrat setzt sich aus je einem Vertreter der einzelnen EU-Mitgliedstaaten zusammen und hat die Aufgabe, die Aktivitäten von Europol zu kontrollieren. **seq2seq**$_{en}$: EU will Europol kontrollieren , die aktivitäten von europol kontrollieren. **SP**$_{\mathcal{L},en}$: Der Vorstand besteht aus einem Vertreter der einzelnen EU-Mitgliedstaaten und hat die Aufgabe, Europol zu kontrollieren. **Gold**: Der Verwaltungsrat setzt sich je einem Vertreter der einzelnen EU-Mitgliedstaaten zusammen und hat die Aufgabe,die Aktivität von Europol zu kontrollieren.
Source: Darüber hinaus stärkt er Europa sowohl politisch als auch wirtschaftlich. **seq2seq**$_{en}$: ein blick auf europa , um die europäische union zu stärken. **SP**$_{\mathcal{L},en}$: Außerdem stärkt Europa politisch und wirtschaftlich. **Gold**: Darüber hinaus stärkt er Europa politisch und wirtschaftlich.
Source: Die Europäische Union will eine Vorreiterrolle bei der Durchsetzung eines langfristigen Entwicklungsmodells spielen. **seq2seq**$_{en}$: EU will eine vorreiterrolle bei der durchsetzung. **SP**$_{\mathcal{L},en}$: Die Europäische Union will bei der Umsetzung eines langfristigen Entwicklungsmodells. **Gold**: viele veränderungen zur verbesserung der zukunft für europa dar.
Source: Dies ist umso wichtiger, als derzeit alle europäischen Staaten über den wirtschaftlichen Aufschwung und die Überwindung der Finanzkrise diskutieren. **seq2seq**$_{en}$: Europäische Länder suchen Wege zur Überwindung der Finanzkrise. **SP**$_{\mathcal{L},en}$: Umso wichtiger ist jetzt, dass alle europäischen Staaten die wirtschaftliche Erholung und Bewältigung der Finanzkrise diskutieren. **Gold**: Dies ist umso wichtiger, als derzeit alle Staaten über wirtschaftlichen Aufschwung und Überwindung der Fizkrise diskutieren.

MOSS EUPar
Source: Die Geschwindigkeit, mit der er geholfen hat, die G20 als neues Instrument einer internationalen wirtschaftlichen Governance zu entwickeln, ist sowohl ermutigend als auch beeindruckend. **seq2seq**$_{en}$: G20 begrüßt neues instrument für die entwicklung der G20. **SP**$_{\mathcal{L},en}$: Die Geschwindigkeit, mit der er die G20 als neues Instrument der internationalen Wirtschaftsregierung entwickelt. **Gold**: Die Geschwindigkeit, mit der er half, die G20 als neues Instrument einer internationalen wirtschaftlichen Goverce zu entwickeln, ist ermutigend und beeindruckend.
Source: Die Minister hatten zu der Sache bereits beim ECOFIN-Frühstück am 15. März eine erste Aussprache. **seq2seq**$_{en}$: Das frühstück mit den ministern. **SP**$_{\mathcal{L},en}$: Die Minister hatten bereits am 15. März eine erste Aussprache. **Gold**: Minister hatten zum frühstück aussprache.
Source: Herr Präsident, Guinea-Conakry braucht die Unterstützung der internationalen Völkergemeinschaft. **seq2seq**$_{en}$: Guinea: böhne braucht internationale unterstützung. **SP**$_{\mathcal{L},en}$: Herr Präsident! Guinea braucht die Unterstützung der internationalen. **Gold**: Guinea braucht internationale unterstützung.
Source: Angesichts dieser Verletzung der zuvor von der Junta zugesagten Verpflichtungen organisierte die Opposition eine Demonstration, die von der Präsidentengarde brutal unterdrückt wurde. **seq2seq**$_{en}$: Opposition verschiebt Protest angesichts der Zusage der Junta **SP**$_{\mathcal{L},en}$: Die Opposition organisierte eine Demonstration, die vom Präsidentenamt brutal niedergeschlagen worden war. **Gold**: Die Opposition eine Demonstration, die von der Präsidentengarde brutal unterdrückt wurde.

MOSS EUPar
Source: auteure. - Monsieur le Président, je pense que l'intérêt dépasse largement la foule qui s'est précipitée dans cet hémicycle cet après-midi, ce que je regrette une fois de plus pour mon groupe et pour les autres. **seq2seq**$_{en}$: Au lendemain de la semaine. **SP**$_{\mathcal{L},en}$: Auteur. - Monsieur le Président, je crois qu'il existe un grand intérêt qui va bien au-delà de la foule, précipité dans cette Assemblée cet après-midi. **Gold**: auteure. - Monsieur le Président, je pense que l'intérêt dépasse la foule qui s'est précipitée dans cet hémicycle cet après-midi, ce que je regrette pour mon groupe et pour les autres.
Source: Nous demandons la mise en place d'un gouvernement de transition pour préparer les élections présidentielles et législatives, la junte s'étant définitivement mise au ban de la communauté internationale. **seq2seq**$_{en}$: Myanmar junte interdit à la communauté internationale. **SP**$_{\mathcal{L},en}$:Nous appelons à la création d'un gouvernement de transition pour la préparation des élections présidentielles et parlementaires avec la junte. **Gold**: Que soit mis en place un gouvernement de transition pour les présidentielles et législatives, la junte étant définitivement au ban de la communauté internationale.
Source: Il n'y a eu que quelques lignes ici et là, comme s'il n'y avait pas eu des centaines de personnes tuées, des actes terribles de torture et des actes effrayants de viol destinés à annihiler la dignité humaine. Au Moyen-Orient, tout le monde attend l'arrivée d'un nouvel Anouar Sadat dans le monde arabe. **seq2seq**$_{en}$: Il est temps de sauver des vies humaines dans la dignité humaine. **SP**$_{\mathcal{L},en}$: Il y avait quelques lignes ici et là, des centaines de personnes tuées, des actes de torture et des actes de viol effroyables visant à détruire la dignité humaine. **Gold**: Juste quelques lignes ci et là, comme s'il n'y avait pas eu des centaines de morts, des actes terribles de torture et de viol pour annihiler la dignité humaine.

Books MOSS
Source: Il repose sur le succès du marché unique etcontribue largement à la stabilité économique requise pour une croissance plus intense. **seq2seq**$_{en}$: La stabilité économique du marché unique **SP**$_{\mathcal{L},en}$: Le succès du marché unique contribue grandement à la stabilité économique nécessaire à la croissance. **Gold**: Reposant sur la marché unique, il contribue largement à la stabilité économique nécessaire à la croissance.
Source: Il renforce également l'Europe sur les plans politique et économique. **seq2seq**$_{en}$: L'Europe crise politique crise politique **SP**$_{\mathcal{L},en}$: Elle renforce également politiquement et économiquement l'Europe. **Gold**: l'Europe sur les plans politique et économique.
Source: Les nouveaux États membres se sontégalement engagés à entériner l'UEM et l'euro dans leurs traités d'adhésion. Au Moyen-Orient, tout le monde attend l'arrivée d'un nouvel Anouar Sadat dans le monde arabe. **seq2seq**$_{en}$: Les nouveaux Etats membres se sont engagés à soutenir l "émeu euro **SP**$_{\mathcal{L},en}$: Les nouveaux États membres s'engagent également à approuver l'UEM et les billets en euros. **Gold**: Les États membres se sont également engagés à entériner l'UEM et l'euro dans leurs traités d'adhésion.

TED
Source: Cette aspiration à trouver des hommes ou des femmes providentiels en ces temps de mondialisation vient de trois facteurs. **seq2seq**$_{en}$: Le monde de la mondialisation. **SP**$_{\mathcal{L},en}$: Cette aspiration à trouver des hommes ou des femmes provient de trois facteurs de mondialisation. **Gold**: Cette tendance vers des hommes et femmes providentiels en ce temps mondialisé vient de trois causes.
Source: Dans un premier temps, il s'agit de la complexité et de la vulnérabilité de notre monde. **seq2seq**$_{en}$: Un regard sur l'avenir du monde **SP**$_{\mathcal{L},en}$: Premièrement, nous parlons de la complexité et de la vulnérabilité du monde. **Gold**: Premièrement, la complexité et la vulnérabilité du monde.

MOSS News
Source: Pourtant, pour la plupart, nous ne voyons pas les choses ainsi. **seq2seq**$_{en}$: N'y voyez pas les choses qui ne voient pas les choses. **SP**$_{\mathcal{L},en}$: Pourtant, nous ne voyons pas encore les choses. **Gold**: Cependant, la majorité n'est pas du même avis.
Source: La plupart des gens chercheront toujours instinctivement de « grands hommes » pour figures historiques, les hommes (et les femmes) qui semblent forger les événements grâce à leur vision politique, leur charisme personnel et la force de leurs positions morales. **seq2seq**$_{en}$: Les hommes sont à la recherche de chiffres historiques. **SP**$_{\mathcal{L},en}$: La plupart des gens vont toujours chercher instinctivement les hommes et les femmes qui sont dépistés par leur vision politique, leur charisme personnel et la force de leur moral. **Gold**: La population tend à se rattacher à de grands hommes politique, à des personnes qui utilisent leur personnalité, leurs avis pour modifier l'histoire.
Source: Par la simple force de leur conviction et de leur personnalité, de telles figures, croyons-nous, peuvent faire avancer les choses, tout en apportant une lueur d'espoir dans un univers autrement détaché et impersonnel. Au Moyen-Orient, tout le monde attend l'arrivée d'un nouvel Anouar Sadat dans le monde arabe. **seq2seq**: Un regard sur l'avenir de notre pays. **SP**$_{\mathcal{L},en}$: Par leur seule conviction et personnalité, nous croyons que de telles figures pourraient faire avancer les choses, et apporter une lueur d'espoir dans un univers détaché et impersonnel. **Gold**: Par leur seule conviction et personnalité, nous croyons que de telles figures pourraient faire avancer les choses, et apporter une lueur d'espoir dans un univers détaché et impersonnel.

B.2 Evaluation instructions

In this experiment you will be asked to judge how well a given sentence compresses the meaning of another sentence. You will see a series of sentences together with their compressed versions. Some sentence compressions will seem perfectly OK to you, but others will not. All compressed versions were generated automatically by a computer program.

Your task is to judge how good a compressed sentence is according to two criteria:(a) grammaticality,and (b) importance. The grammaticality judgement is based on whether the sentence is understandable. The importance judgement relates to how well the compression preserves the most important information of the original and whether it is adequately compressed. Both judgements are rated on scales from 1 (poor) to 5 (good).

A compression with a low grammaticality score is one that is almost impossible to understand. Compressions should receivelow importance scoresif they miss out important information from the original sentence. Or do NOT remove any unneeded information from the original sentence even though it is evident that it can be omitted without drastic information loss.

A good compression is one that is readily comprehensible andretains the most important information from the original sentence. Good sentence compressions should receive high grammaticality and importance scores.

For example, if you were asked to rate the following compression:

- Nonetheless, FBI director Louis Freeh has today ordered a change - this is being reported by the New York Times - ordering new restrictions on the sharing of confidential information with the White House.

- Nonetheless, FBI director ordered change new restrictions sharing confidential information with White House.

This sentence would probably receive a low grammaticality score(for example, 1 or 2) as it is difficult to understand. However it should receive ahigh score for importance (for example, 4 or 5) as it is possible to get the gist of the original.

Now, consider the following compression of the same sentence:

- Nonetheless, FBI director Louis Freeh has today ordered a change - this is being reported by the New York Times - ordering new restrictions on the sharing of confidential information with the White House.

- FBI director Louis Freeh has today ordered a change - this is being reported by the New York Times.

you would give the compression a higher grammaticality score(for example, 4 or 5) but alow importance score (for example, 1 or 2). The compression preserves the least important information (the fact that the New York Times is reporting).

You will be presented with the original sentence and its corresponding compression. Once you read both sentences, please make your grammaticality and importance judgement. Simply select a number between 1 and 5 by clicking the appropriate button.

There are no 'correct' answers, so whatever numbers seem appropriate to you are a valid response. While you are deciding a number for a compression, try to ask the following questions:

Does the compressed sentence preserve the most important bits of information from the original sentence? Is the compressed sentence easy to understand? Has the compressed sentence removed information you deem not to be very important to the original sentence? Does the compressed sentence seem fluent? Use high numbers if the answer to the above questions is 'yes', low numbers if it is 'no', and intermediate numbers for sentences that are understandable, yet not entirely accurate or natural compressions of the original sentence.

Appendix C

Controllable Simplification

C.1 Simplification Examples

We show simplification examples created by our model and comparison systems on WikiLarge and Newsela (see Tables C.3–C.4).

Complex	It is ~~situated~~ at the coast of the Baltic Sea, where it encloses the city of Stralsund.
Reference	It is **located** at the coast of the Baltic Sea where it **surrounds** the city of Stralsund.
DRESS-Ls	It is **situated** at the coast of the Baltic Sea.
Transformer	It is **situated** at the coast of the Baltic Sea, where it encloses the city of Stralsund.
DMASS	It is **located** at the shore of the Baltic Sea, where it is located at the shore of the borough of stralsund.
CROSS-Lex	It is **at** the coast of the Baltic Sea, where it encloses the city of Stralsund.
CROSS-Syn	It is **located** at the coast of the Baltic Sea.
CROSS	It is **found** at the coast of the Baltic Sea, near the city of Stralsund.
Complex	In 1987 Wexler was ~~inducted~~ into the Rock and Roll Hall of Fame.
Reference	In 1987 Wexler was inducted into the Rock and Roll Hall of Fame.
DRESS-Ls	In 1987 Wexler was inducted into the Rock and Roll Hall of Fame.
Transformer	In 1987, Wexler was inducted into the Rock and Roll Hall of Fame.
DMASS	In 1987 Wexler was inducted into the Rock and Roll Hall of Fame.
CROSS-Lex	In 1987 Wexler was **added into** the Rock and Roll Hall of Fame.
CROSS-Syn	**He** was inducted into the Rock and Roll Hall of Fame in 1987.
CROSS	**He** was **added into** the Rock and Roll Hall of Fame in 1987.

Table C.1: System output on WikiLarge. Lexical items indicated for replacement are marked with a ~~strike out~~. We show the source Complex sentence and the simplified Reference as well as output from DRESS-Ls, a simple Transformer, DMASS, and three variants of our model; CROSS is the full system, CROSS-Lex applies only lexical constraints, while CROSS-Syn only syntactic ones. Substitutions are shown in **bold**.

Complex	In its pure form, Dextromethorphan occurs as a white powder.
Reference	Dextromethorphan **is** a white powder in its pure form.
DRESS-Ls	In its pure form, Dextromethorphan occurs as a white powder.
Simple	In its pure form, Dextromethorphan **is like** a white powder.
XSimple	Dextromethorphan **can be found** as white powder.
Complex	After graduation he returned to Yerevan to teach at the local Conservatory and later he was appointed artistic director of the Armenian Philarmonic Orchestra.
Reference	After graduation he returned to Yerevan to teach at the local Conservatory and was later appointed artistic director of the Armenian Philarmonic Orchestra.
DRESS-Ls	After graduation he returned to Yerevan to teach at the local Conservatory.
Simple	After graduation he returned to Yerevan to teach at the local Conservatory and later he **became** artistic director of the Armenian Philarmonic Orchestra.
XSimple	after graduation he returned to Yerevan to teach at the Conservatory and later he **became** artistic director of the Armenian Philarmonic Orchestra.

Table C.2: System output on WikiLarge for varying simplicity levels. We show the source Complex sentence and the simplified Reference as well as output from DRESS-Ls, and two variants of our model Simple and XSimple. Substitutions are shown in **bold**.

Complex	He thinks the new ~~stealth~~ bomber program would ~~ultimately~~ cost $90 billion.
Reference	He thinks the new stealth bomber program would **actually** cost $90 billion.
DRESS-Ls	He thinks the **new combat number** would cost $90 billion.
Transformer	He thinks the new bomber program would cost $90 billion.
DMASS	The new $ sinkhole program would be used this year.
CROSS-Lex	He thinks the new program would cost $90 billion.
CROSS-Syn	He thinks the program would cost $90 billion for the new stealth bomber.
CROSS	He thinks the program would cost $90 billion for the new bomber.
Complex	But then he heard the radio : a ~~massive~~ glacier had crashed down the mountain.
Reference	A huge glacier of ice had just crashed down the mountain.
DRESS-Ls	But then he heard the radio: A massive glacier had crashed down the mountain.
Transformer	But then he heard the radio: A massive glacier had crashed down the mountain.
DMASS	Then he heard the radio even though a **huge piece** had crashed down the mountain.
CROSS-Lex	But then he heard the radio: A **huge** glacier had crashed down the mountain.
CROSS-Syn	But then he heard a glacier crash down the mountain.
CROSS	But then he heard a glacier crash down the mountain.

Table C.3: System output on Newsela. We show the source Complex sentence and the simplified Reference as well as output from DRESS-Ls, a simple Transformer, DMASS, and three variants of our model; CROSS is the full system, CROSS-Lex applies only lexical constraints, while CROSS-Syn only syntactic ones. Substitutions are shown in **bold**. Lexical items indicated for replacement are marked with a ~~strike out~~.

Complex	The Pentagon is poised to spend billions to build a new stealth bomber, a top secret project that could bring hundreds of jobs to the wind-swept desert communities in Los Angeles County's northern reaches.
Reference	Mission to build the secret warplane.
DRESS-Ls	The Pentagon is trying to spend billions to build a new drone.
Simple	The Pentagon secret project that could bring hundreds of jobs to the desert-swept communities in Los Angeles County.
XSimple	It could also bring hundreds of jobs.
Complex	The United States is about to spend billions of dollars to build a top-secret warplane.
Reference	Mission to build the secret warplane.
DRESS-Ls	The United States is about to spend billions of dollars to build a secret bomb.
Simple	The United States is about spend dollars to build a top-secret warplane.
XSimple	The United States is about to build a warplane.

Table C.4: System output on Newsela for varying simplicity levels. We show the source Complex sentence and the simplified Reference as well as output from DRESS-Ls, and two variants of our model Simple and XSimple. Substitutions are shown in **bold**.

C.2 Evaluation Instructions

In this task you will read a series of sentences and their simpler versions created by a computer program. The program performs simplification by removing content but also by changing the structure and wording of the sentences so that they are easier to read. Please read all the sentences carefully, this should take you about 2 minutes + 30 seconds for the bonus (if you do the task very very quickly your hit will be rejected). You will be asked to judge three aspects of the simplifications: (1) is the simple version grammatical and fluent? if so you should give it a high Grammaticality score. (2) To what extent is the meaning expressed in the original sentence preserved in the simple version, with no additional information added? If most of the meaning is preserved you should give it a high meaning score. (3) is the proposed simplification a simpler version of the original sentence? If so you should give it a high Simplicity score.

In some cases the computer program will chose to add information which is not in the orginal sentence. If this is the case then you should probably rate it lower in terms of meaning.

In some cases, the simple sentence will be a copy of the original sentence, if this is the case, you should give it a 5 for meaning. For the Simplicity score, you should consider if you could make the original sentence simpler? if you can't make the sentence simpler, you should give the simplification a high simple score.

In the end, you will be asked to provide comments if you provide an insightful comment we will pay an additional $0.15 bonus. Bad comments include: Sentence 3 was not simpler or didn't contain all the information. The comment should say something which is not inferable from the scores you give.

Appendix D

Zero-Shot Crosslingual Sentence Simplification

D.1 Dataset construction examples

The following was provided for the simplification dataset construction.

This annotation experiment is concerned with simplification. You will be presented with a document. Your task is to read each sentence and simplify it such that children aged between 5 and 7 can understand it. The simplified version should be grammatical and retain all the important information of the original sentence.

In producing simplifications, you are free to delete words, add new words, substitute them, or reorder them. In addition, you might find it useful to change a complex sentence into multiple simple sentences.

To help you with the simplification task, we have produced a set of guidelines which you can follow. However, not all guidelines will always be applicable, so if you believe you can produce a simpler versionn then you may ignore the guidelines. We split the guidelines into two sections: word-level and sentence-level guidelines.

Word-level Guidelines

1. Special characters are not allowed, with the exception of: full stops, question marks, exclamation marks, quotation marks, and Mediopunkts (used to indicate compound splitting).

2. Numbers should be written as digits and not words.

3. The word *ein* (‘one’) should only be written with a 1 when it represents a number, not when it takes the role of an indefinite article.

4. Roman numerals must be avoided.

5. Large numbers, percentages and year dates should be used sparsely.

6. Use easy, short and well-known words. In case a difficult word is needed, it should be explained using simple words. For a list of simple words, please consult this dictionary: `https://hurraki.de/wiki/Hauptseite`.

7. Technical terms, foreign words and abbreviations should be avoided. Common acronyms like CD or *WC* may be used if their full forms (compact disc, water closet) are less common.

Sentence-level Guidelines

1. Coordinate and subordinate clauses are forbidden and should be transformed into independent main clauses. Main clauses should preferably contain active voice, and present, or past perfect tense. The subject-verb-object (SVO) word order should be chosen, unless another word order is more understandable.

2. Nominalizations and passive constructions are forbidden.

3. Attributive genitives should also be avoided. If possible, the genitive attribute should be transferred into a prepositional phrase using *von* (‘of’).

4. Negation should be avoided. If needed, it is better to formulate a sentence with *nicht* (‘not’) instead of *kein* (‘no’).

5. Transparent metaphors like leichte Sprache may be used if they can be easily understood. More complex metaphors and idioms should be replaced by literal expressions.

6. Split complex sentences into multiple simple sentences at semicolons and dashes. Also split sentences after colons if the segment after the colon is a complete sentence and not just an enumeration.

7. If a subordinate conjunction is found, split the sentence at the conjunction; edit and rephrase both resulting segments to form independent sentences. Add suitable connectives that express the intended rhetorical relation and restore word order.

8. Rephrase concessive clauses with subjunctions like *obwohl* ('although') the connective *trotzdem* ('however').

9. Analogously, rephrase consecutive clauses starting with *sodass* ('so that') using *deshalb* ('therefore').

10. Rephrase final clauses using the modal verb *wollen* ('want') and the connective *deshalb* ('therefore'). Since the subject is not mentioned overtly in German final clauses containing *um zu* ('in order to'), it has to be retrieved from the main clause.

11. Split coordinate clauses at coordinating conjunctions (e.g., *und* ('and'), *oder* ('or'), *aber* ('but'), *dennoch* ('however')). The second clause can start with *und* ('and') and *oder* ('or') to emphasize that they are linked to the previous sentence.

12. Replace appositions by sentences in which the noun phrase referred to by the apposition forms the subject (X) and the apposition itself becomes the predicative noun (Y), yielding an X is Y structure.

Final Remarks

The annotation will proceed on a document-by-document basis. In simplifying individual sentences you should ensure that:

- You have preserved all important information in the original sentence.
- The sentences are understandable to children aged 5 to 7.
- You did not render the resulting document incoherent or unreadable.
- You have preserved the grammaticality of the simplified sentences.

D.2 Simplification Examples

In Table D.1 we present examples of simplifications from both GEOlino and TextComplexityDE. We show the input Complex sentence, the Reference simplification, and the output of our model, ZEST, and two comparison systems U-SIMP and Pivot (see Section 5 in the main paper for more details). To provide some further insight on what the models are doing we have annotated words and phrases in the examples which constitute violations of simple language according to our guidelines above and those provided in `https://hurraki.de/pruefung/pruefung.htm`.

Table D.2 and D.3 contains additional examples, without annotation, while Table D.4 presents crosslingual simplicity examples. Specifically, we show model output in German (DE ZEST) when the input is complex English (EN Source), and for comparison German output (DE ZEST) when the input is complex German (DE Source).

Complex	Von hier aus erhaltet ihr einen **[eindrucksvollen]**[1] Rundum-Blick über die ganze Schlucht **[hinweg]**[2] bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
Reference	Von hier aus erhaltet ihr einen Rundum-Blick über die ganze Schlucht. Ihr seht hier bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
U-SIMP	Von hier eraus ihr haltet einen umfassenden Rundum-Blick über die ganze bis **[hinweg]**[2] hinweg zu hin zu ihren **[5000]**[3] Meter hohen Kraterwände.
Pivot	Von hier genießen Sie einen **[beeindruckenden]**[1] Rundumblick über die gesamte Schlucht bis **[zu]**[2] den 500 m hohen Kraterwänden.
ZEST	Von hier aus erhaltet ihr einen Rundum-Blick auf die ganze Schlucht.

(a) GEOlino

Complex	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] **[Erdwärmenutzung]**[6] **[im]**[2] großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
Reference	Das ist nur etwa das Doppelte **[des Energiebedarfs der Welt]**[4]. Das bedeutet, **[dass]**[5] die **[Benutzung]**[6] von Erdwärme immer dazu führt, **[dass]**[5] an **[sich]**[2] diesen Stellen das Gestein abkühlt.
U-SIMP	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] Erdwärmemer **[im]**[2] großen Stil immer auf eine andere Abkühlung des Gesteins**[)]**[7].
Pivot	Dabei handelt es sich nur um eine **[Verdoppelung]**[6] **[des weltweiten Energiebedarfs]**[5], **[was]**[5] **[bedeutet]**[2], **[dass]**[5] die großflächige **[geothermische]**[8] **[Nutzung]**[6] immer einer lokalen **[Kühlung]**[6] **[des Gesteins]**[4] entspricht.
ZEST	Das bedeutet, **[dass]**[5] Erdwärme im großen Stil immer auf eine lokale Abkühlung **[des]**[2] Gesteins hinausläuft.

(b) TextComplexityDE

Table D.1: Examples of system output and simplification violations (https://hurraki.de/pruefung/pruefung.htm): (1) word has 13+ letters; (2) sentence has 12+ words; (3) high number; (4) genitive; (5) subordinate clauses; (6) abstract words; (7) special characters; (8) difficult/foreign words.

Source	Zusätzlich kommt in der Forschung ein breites Spektrum verschiedener Isotopenanalysen zum Einsatz, deren jüngste Entwicklungen eine bis vor kurzem unerreichbare Messgenauigkeit ermöglichen.
Reference	Außerdem werden in der Forschung viele verschiedene Isotopenanalysen eingesetzt. Ihre jüngste Entwicklung ermöglichen ermöglichen eine bis vor kurzem unerreichbare Messgenauigkeit.
U-SIMP	Zusätzlich kommt in der Forschung ein großes verschiedene verschiedenen Gitopenempfindliche zum Einsatz, deren jüngsten Jahren eine bis vor kurzem brennende Messoberfläche ermöglichen.
Pivot	Darüber hinaus wird in der Forschung ein breites Spektrum unterschiedlicher Isotopanalysen eingesetzt.
ZEST	Zusätzlich kommt in der Forschung ein breites Spektrum verschiedenster Isotopen-Analysen zu Einsatz.
Source	Das ist nur etwa das Doppelte des Weltenergiebedarfs, was bedeutet, dass Erdwärmenutzung im großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
Reference	Das ist nur etwa das Doppelte des Energiebedarfs der Welt. Das bedeutet, dass die Benutzung von Erdwärme immer dazu führt, dass an sich diesen Stellen das Gestein abkühlt.
U-SIMP	Das ist nur etwa das Doppelte des Weltenergiebedarfs, was bedeutet, dass Erdwärmemer im großen Stil immer auf eine andere Abkühlung des Gesteins).
Pivot	Dabei handelt es sich nur um eine Verdoppelung des weltweiten Energiebedarfs, was bedeutet, dass die großflächige geothermische Nutzung immer einer lokalen Kühlung des Gesteins entspricht.
ZEST	Das bedeutet, dass Erdwärme im großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.

(a) TextComplexityDE

Table D.2: Simplification examples from TextComplexityDE and GEOlino.

Source	Tiere tauschen mittels Duftmarken viele verschlüsselte Botschaften untereinander aus.
Reference	Tiere tauschen mit ihrem Geruch viele Botschaften untereinander aus.
U-SIMP	Tiere tauschen Hilfe Duftmarken viele verschlüsselte Botschaften untereinander aus.
Pivot	Tiere tauschen viele verschlüsselte Nachrichten mit Duftmarken aus.
ZEST	Tiere tauschen mit Duftmarken viele verschlüsselte Botschaften aus.
Source	Der wiederum war überlebenswichtig für alle Landwirtschaft betreibenden Kulturen.
Reference	Der war wichtig für alle Kulturen, die Landwirtschaft betreiben.
U-SIMP	Der wiederum war überlebenswichtig für alle Landwirtschaft ben Kulturen.
Pivot	Sie war wiederum lebenswichtig für alle landwirtschaftlichen Kulturen.
ZEST	Der wiederum war für alle Landwirtschaft wichtig.

(b) GEOlino

Table D.3: Simplification examples from TextComplexityDE and GEOlino.

EN Source	The mountain is the watershed on whose flanks the catchment areas of the Pacific Ocean, the Atlantic Ocean over the Gulf of Mexico, and the Arctic Ocean over Hudson Bay, meet.
DE Source	Der Berg ist der Wasserscheidepunkt an dessen Flanken sich die Einzugsgebiete des Pazifischen Ozeans, des Atlantischen Ozeans über den Golf von Mexiko und des Arktischen Ozeans über die Hudson Bay berühren.
Reference	Der Berg markiert die Grenze zwischen den Gebieten des Pazifischen Ozeans, des Atlantischen Ozean und des Arktischen Ozeans.
EN ZEST	Der Berg ist der Weckschatz, auf dessen Flanken die Fanggebiete des pazifischen Ozeans, des Atlantischen Ozeans über dem Golf von Mexiko, und des Arktischen Ozeans über Hudson Bay, treffen.
DE ZEST	Der Berg ist der Wasserscheidepunkt an dem sich die Einzugsgebiete des Pazifiks, des Atlantischen Ozeans, des Golfs von Mexiko und des Arktischen Ozeans über die Hudson Bay treffen.

(a) TextComplexityDE

EN Source	Without the radiation energy of the sun, plant photosynthesis would not work.
DE Source	Ohne die Strahlungsenergie der Sonne würde die pflanzliche Photosynthese nicht funktionieren.
Reference	Ohne die Energie der Sonne würde die Photosynthese von den Pflanzen nicht funktionieren.
EN ZEST	Ohne die Strahlungsenergie der Sonne, Pflanzen Photosynthese würde nicht funktionieren.
DE ZEST	Ohne die Strahlungsenergie der Sonne würde die Pflanze nicht funktionieren.

(b) Geolino

Table D.4: Examples of crosslingual simplification (EN Source → DE ZEST); for comparison, we also show the output of a monolingual system (DE Source → DE ZEST).

D.3 Evaluation Instructions

In this task you will read a series of sentences and their simpler versions created by a computer program. The program performs simplification by removing content but also by changing the structure and wording of the sentences so that they are easier to read. Please read all the sentences carefully, this should take you about 2 minutes + 30 seconds for the bonus (if you do the task very very quickly your hit will be rejected). You will be asked to judge three aspects of the simplifications: (1) is the simple version grammatical and fluent? if so you should give it a high Grammaticality score. (2) To what extent is the meaning expressed in the original sentence preserved in the simple version, with no additional information added? If most of the meaning is preserved you should give it a high meaning score. (3) is the proposed simplification a simpler version of the original sentence? If so you should give it a high Simplicity score.

In some cases the computer program will chose to add information which is not in the orginal sentence. If this is the case then you should probably rate it lower in terms of meaning.

In some cases, the simple sentence will be a copy of the original sentence, if this is the case, you should give it a 5 for meaning. For the Simplicity score, you should consider if you could make the original sentence simpler? if you can't make the sentence simpler, you should give the simplification a high simple score.

In the end, you will be asked to provide comments if you provide an insightful comment we will pay an additional $0.10 bonus (pay out aims to be done within a week).

D.4 Simplification Analysis

Below we list the simplification phenomena we asked annotators to analysis.

Passive voice Does the complex sentence use the passive voice and the simple sentence use active voice?

Active voice means that a sentence has a subject that acts upon its verb. This is when the subject of a sentence performs the verb's action, Passive voice means that a subject is a recipient of a verb's action. This is when the subject is acted on by the verb.

- Complex: Als der Manager gefeuert wurde.
- Simple: Sie hat den Manager entlassen

Subordinate clause Does the complex sentence contain a subordinate clause and the simple sentence does not?

a sentence with 2 or more verbs and the second one is dependent on the first one and often indicated by words like dass, ob, weil.

- Complex: Sie ist nicht in die Schule gekommen, weil sie erkältet war.
- Complex: Ich mag, dass das Taschentuch so weich ist.
- simple:Sie war erkältet. Deshalb ist sie nicht zur Schule gekommen.
- Simple: Das Taschentuch ist weich. Ich mag das.

Relative clause Does the complex sentence contain a Relative clause and the simple sentence does not?

A relative clause is a type of subordinate clause that starts with the following: der, die, dass

Genitive Does the complex sentence use the genitive case whereas the simple sentence does not?

The genitive case is the case that shows possession. The genitive case is used with the genitive prepositions and some verb idioms. In simple German, von plus the dative often replaces the genitive.

- Complex: Der Sattel des Fahrrads.
- Simple: Der Sattel von dem Fahrrad.
- Complex:Der Sattel deines Fahrrads.
- Simple: Der Sattel von deinem Fahrrad.

Subjunctive Was the subjunctive mood used in the complex sentence and not in the simple sentence?

The subjunctive mood (Konjunktiv), is used to express unreal situations such as wishes, hypothetical situations and unreal conditional clauses, or to repeat what people say in indirect speech.

- Complex: Sie sagte, morgen könnte es regnen.
- Simple: Morgen regnet es vielleicht.
- Complex:Morgen könnte es regnen.
- Simple: Morgen regnet es vielleicht.

Sentence splitting Has the complex sentence been split into multiple sentences?

- Complex: Tim liebt Tina und Tina liebt Tim.
- Simple: Tim liebt Tina. Tina liebt Tim.
- Complex: Er hat nicht nur Hunde, sondern auch Katzen.
- Simple: Er hat Hunde. Er hat Katzen.