UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Unsupervised text simplification using neural style transfer

Author:
Oleg KARIUK

Supervisor:
Dima KARAMSHUK

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Science*

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Oleg KARIUK, declare that this thesis titled, “Unsupervised text simplification using neural style transfer” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Unsupervised text simplification using neural style transfer

by Oleg KARIUK

Abstract

With the growing interdependence of the world economies, cultures and populations the advantages of learning foreign languages are becoming more than ever apparent. The growing internet and mobile phone user base provides significant opportunities for online language learning, the global market size of which is forecasted to increase by almost $17.9 bn during 2019-2023. One of the most effective ways to better oneself in a foreign language is through reading.
Graded readers — the books in which the original text is simplified to lower grades of complexity — make the process of reading in a foreign language less daunting. Composing a Graded reader is a laborious manual process. There are two possible ways to computerize the process of writing Graded readers for arbitrary input texts. The first one lies in utilizing a variation of supervised sequence-to-sequence models for text simplification. Such models depend on scarcely available parallel text corpora, the datasets in which every text piece is available in the original and simplified versions. An alternative unsupervised approach lies in applying neural style transfer techniques where an algorithm can learn to decompose a given text into vector representations of its content and style and to generate a new version of the same content in a simplified language style. In this work, we demonstrate the feasibility of applying unsupervised learning to the problem of text simplification by using cross-lingual language modeling. It allows us to improve the previous best BLEU score from 88.85 to 96.05 for the Wikilarge dataset in unsupervised fashion, and SARI score from 30 to 43.18 and FKGL from 4.01 to 3.58 for the Newsela dataset in semi-supervised one. Apart from that, we propose new penalties that provide more control during beam search generation.

Acknowledgements

I would first like to thank my thesis advisor Dima Karamshuk from Facebook Research. Whenever I was in difficulty or need advice on my research or writing he always was there for me. Dima is an endless source of brilliant ideas.

I would also like to thank the Ukrainian Catholic University and Oleksii Molchanovskyi for the Master Program that opened me up to the world of Data Science.

Finally, I must express my gratitude to my spouse and my son for providing me with absolute support and continuous encouragement. This accomplishment would not have been possible without their patience and humility. They sacrificed so much time that we could spend together. Thank you.

Contents

List of Figures

List of Tables

List of Abbreviations

BLEU	**B**ilingual **E**valuation **U**nderstudy
BPE	**B**yte **P**air **E**ncoding
CBT	**C**hildren's **B**ooks **T**est
DRESS	**D**eep **RE**inforcement **S**entence **S**implification model
EASSE	**E**asier **A**utomatic **S**entence **S**implification **E**valuation
FKGL	**F**lesch-**K**incaid **G**rade **L**evel
LSTM	**L**ong **S**hort-**T**erm **M**emory
MT	**M**achine **T**ranslation
NLP	**N**atural **L**anguage **P**rocessing
PBMT	**P**hrase **B**ased **M**achine **T**ranslation
RNN	**R**ecurrent **N**eural **N**etwork
SW	**S**imple**W**iki

List of Symbols

$BLEU$	BLEU score
$SARI$	SARI score
$FKGL$	FKGL score
F	F1 score
P	Precision
R	Recall

Chapter 1

Introduction

1.1 Motivation

Text simplification deals with the problem of rewriting complex texts into a simpler language which is easier to read and understand. The main goal of text simplification is to reduce the linguistic complexity of a text while preserving its original information and meaning. Key factors that help to improve the readability of texts are the vocabulary, the length of the sentences and the syntactic structures which are present in the text.

Text simplification is an important task that has numerous potential practical applications. Simplification techniques can be used to make reading easier for a broader range of readers, including:

- people with disabilities (Canning et al., 2000; Carroll et al., 1999; Inui et al., 2003);
- people with low-literacy (De Belder and Moens, 2010; Watanabe et al., 2009);
- language learners (Allen, 2009; Petersen and Ostendorf, 2007);
- non-experts (Elhadad and Sutaria, 2007; Siddharthan and Katsos, 2010).

Moreover, applying text simplification in a text pre-processing stage has been shown to improve the performance of many natural language processing (NLP) tasks, including:

- relation extraction, the task of finding a relevant semantic relation between two given target entities in a sentence (Miwa et al., 2010);
- syntactic parsing, the task of finding structural relationships between words in a sentence Jonnalagadda et al., 2009);
- semantic role labeling, the task of modeling the predicate-argument structure of a sentence (Vickrey and Koller, 2008);
- machine translation (Štajner and Popovic, 2016);
- text summarization (Margarido et al., 2008).

In this work, we focus on utilizing text simplification in *online language learning*. Among others, this can help to automate the laborious manual process of writing *Graded readers* — books in which language style has been intentionally simplified to make it more accessible for foreign language learners. Graded readers are commonly composed for various levels from beginners to advanced and are graded for vocabulary, the complexity of grammar structures and also by the number of words.

Text simplification models in the literature are commonly designed to simplify texts in three aspects:

1. **lexical**, which assumes replacing complex words with simpler equivalents (Candido et al., 2009; Glavaš and Štajner, 2015; Yatskar et al., 2010; Biran, Brody, and Elhadad, 2011; Devlin and Tait, 1998);
2. **syntactic**, which implies adjusting the structure of the sentences (Siddharthan, 2006; Filippova and Strube, 2008; Brouwers et al., 2014; Chandrasekar and Srinivas, 1997; Canning and Taito, 1999);
3. **semantic**, which assumes text paraphrasing (Kandula, Curtis, and Zeng-Treitler, 2010).

From the sentence perspective, simplification includes **splitting** (Siddharthan, 2006; Petersen and Ostendorf, 2007; Narayan and Gardent, 2014), **deletion and compression** (Rush, Chopra, and Weston, 2015; Clarke and Lapata, 2006; Filippova and Strube, 2008; Filippova et al., 2015; Knight and Marcu, 2002), and **paraphrasing** (Wubben, Bosch, and Krahmer, 2012; Nisioi et al., 2017; Specia, 2010; Wang et al., 2016; Coster and Kauchak, 2011).

Most of the recent text simplification systems are based on the variations of *sequence-to-sequence* (Seq2Seq) models that require parallel corpora for training (Kajiwara and Komachi, 2016; Scarton, Paetzold, and Specia, 2018; Zhang and Lapata, 2017). Unfortunately, the scarcity of parallel data limits the scalability of this approach in application to different languages, domains, and output styles. Moreover, the *Parallel Wikipedia Simplification* corpus, which has become the benchmark dataset for training and evaluating text simplification systems, is (a) prone to automatic sentence alignment errors, (b) contains a lot of inadequate simplifications and (c) poorly generalizes to other text styles (Xu, Callison-Burch, and Napoles, 2015).

In contrast to sequence-to-sequence models, *unsupervised learning algorithms* do not require labeled parallel corpora. In a nutshell, they can learn to decompose a given text into in vector representations of its content and its style and, further, generate the same content in a simplified language.

1.2 Goals of the master thesis

The focus of this current thesis is on unsupervised text simplification which has been significantly less studied in the literature. To this end, we aim to:

1. Attest the feasibility of applying unsupervised learning (i.e., neural style transfer) to the problem of text simplification by applying cross-lingual language modeling.
2. Conduct its comprehensive evaluation over a variety of datasets and metrics and in comparison to the existing supervised baselines.
3. Introduce beam search generation penalties for better control and results.
4. Investigate directions to improve the performance of the proposed approach through better architectures of the neural network and training regimes.

1.3 Structure of the thesis

In Chapter 2 we review the background and literature related to the task of text simplification. Chapter 3 focuses on the evaluation methodology and discusses the pros and cons of different existing evaluation metrics. In Chapter 4 we describe the datasets utilized in the rest of the thesis. Chapters 5 and 6 provide details on the methodology of our work and a detailed overview of the conducted experiments. Last but not least, in Chapter 7 we sum up our results and contributions and outline directions for future research.

Chapter 2

Background and Related work

In recent years, the problem of text simplification has often been addressed as the monolingual language-to-language *machine translation* from the original to simplified sentences. The existing machine translation models from the literature were modified to the particularities of the text simplification task.

Zhu, Bernhard, and Gurevych, 2010 proposed a model for sentence simplification via *tree transformation* based on the techniques from statistical machine translation. The model applies a sequence of simplification operations to perform splitting, dropping, reordering and word/phrase substitutions.

A variation of *phrase-based machine translation* (PBMT) with a dissimilarity component was proposed by Wubben, Bosch, and Krahmer, 2012. The proposed approach focuses on dissimilarity rather than deletion in the PBMT decoding stage, as simplification does not necessarily imply shortening. Outputs of the PBMT model are re-ranked according to their dissimilarity to the input sentence.

Narayan and Gardent, 2014 presented a hybrid approach to sentence simplification which combines *deep semantics and monolingual machine translation* to derive simple sentences from the complex ones. Their simplification model consists of a probabilistic model for splitting and dropping, a PBMT model for substitution and reordering and a language model learned on Simple English Wikipedia for fluency and grammaticality. The simplification process is split into two steps. Firstly, the probabilistic model performs sentence splitting and deletion operations, therefore, producing one or more intermediate simplified sentences. Secondly, simplified sentences are further simplified using the PBMT system.

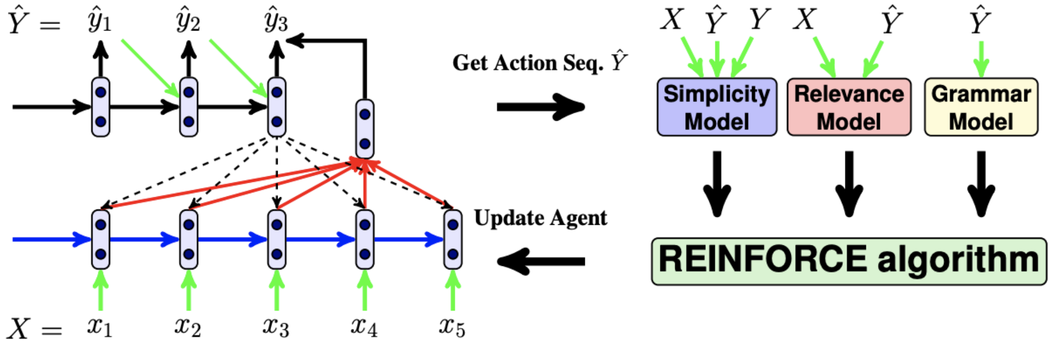
$\hat{Y} = \hat{y}_1 \; \hat{y}_2 \; \hat{y}_3$
Get Action Seq. $\hat{Y}$
X $\hat{Y}$ Y X $\hat{Y}$ $\hat{Y}$
Simplicity Model
Relevance Model
Grammar Model
Update Agent
REINFORCE algorithm
$X = x_1 \; x_2 \; x_3 \; x_4 \; x_5$

FIGURE 2.1: DRESS model. X is the complex sentence, Y is the reference sentence and $\hat{Y}$ simplification produced by the encoder-decoder model. Source: Zhang and Lapata, 2017.

Zhang and Lapata, 2017 proposed a *deep reinforcement sentence simplification model* (DRESS, fig. 2.1) with an encoder-decoder architecture implemented by recurrent neural networks (RNNs). To make the output simpler and grammatically correct while also preserving the meaning of the input, they trained the model in a *reinforcement learning* (RL) framework. It explores the space of possible simplifications while learning to maximize an expected reward function that encourages outputs that meets simplification constraints.

2.1 Simplification as a style transfer

Text simplification can be viewed as a form of *style transfer* (Wubben, Bosch, and Krahmer, 2012; Sulem, Abend, and Rappoport, 2018b), or *stylistic paraphrasing*, with the goal of rewriting a sentence such that we preserve the meaning but alter the style. Generating paraphrases targeting a more general interpretation of style was first attempted in (Xu et al., 2012). All of these works are based on statistical machine translation methods.

Recently, however, the advances in neural machine translation have started to be applied to general stylistic paraphrasing. The Shakespeare dataset (Xu et al., 2012) was used recently with a variation of Seq2Seq models (Jhamtani et al., 2017). They proposed to use a mixture model of *pointer network* and *Long Short-Term Memory* (LSTM) to transform a modern English text to a Shakespearean style English. The authors reported an improvement over statistical machine translation methods. Another impressive work in this direction uses a large set of Bible translations to transfer a prose style with an encoder-decoder recurrent neural network and *Moses* - a statistical machine translation system (Carlson, Riddell, and Rockmore, 2018). Table 2.1 gives some examples produced by this approach.

Source	**Target**	**Moses output**	**Seq2Seq**
Then Samuel gave him an account of everything, keeping nothing back. And he said, It is the Lord; let him do what seems good to him.	And Samuel told him every whit, and hid nothing from him. And he said, It is Jehovah: let him do what seemeth him good.	Then Samuel told him of all things not. And he said, It is Jehovah; let him do that which seemeth him good.	And Samuel told all things, and did not hold back. And he said, It is Jehovah; let him do what seemeth good to him.
And Jehovah saith, 'Judah doth go up; lo, I have given the land into his hand'.	And the Lord said, Judah is to go up: see, I have given the land into his hands.	And the Lord said, 'Judah will go up, see, I have given the land into his hand.'	And the Lord said, Judah will go up; see, I will give the land into his hand.

TABLE 2.1: Examples which show Moses and Seq2Seq Bible style transfer. Source: Carlson, Riddell, and Rockmore, 2018.

Unfortunately, the lack of appropriate training corpora has complicated the direct application of the style transfer approaches to text simplification.

2.2 Unsupervised style transfer

In contrast, unsupervised style transfer algorithms work with unlabeled datasets which are significantly cheaper and easier to obtain. However, they are also considerably less explored in the literature. Amongst the few exceptions, are:

Paetzold and Specia, 2016 who proposed an unsupervised lexical simplification technique that replaces complex words in the input with simpler synonyms, which are extracted and disambiguated using word embeddings.

Shen et al., 2017 who proposed to apply an *adversarial training* to unsupervised style transfer and introduced a refined alignment of sentence representations across text corpora. They build an encoder that takes a sentence and its original style indicator as input and maps it to a style-independent content representation that is passed to a style-dependent decoder. The key contribution of this approach is in applying discriminators both on the encoder representation and on the hidden states of the decoders to ensure that they have the same distribution.

Zhang et al., 2018 who proposed a two-stage joint training method to boost source-to-target and target-to-source style transfer systems using non-parallel text. They build bidirectional word-to-word style transfer systems in a statistical machine translation framework to generate a pseudo-parallel corpus and constructed two attention-based neural machine translation style transfer systems with the pseudo corpus. Then an iterative back-translation algorithm was employed to better leverage non-parallel text to jointly improve bidirectional neural machine translation based style transfer models.

Surya et al., 2019 who used unlabeled corpora containing simple and complex sentences to train the system based on the shared encoder and two decoders. They proposed a novel training scheme which allows the model to perform content reduction and lexical simplification simultaneously through proposed losses and de-noising.

In comparison with the above-mentioned unsupervised models, we explore a novel application of the architecture for cross-lingual language modeling to the task of text simplification. Our approach achieves superior BLEU and SARI results on the Wikilarge dataset. In addition, we conduct a more comprehensive evaluation and assess the system's performance from a wider variety of metrics (see Chapter 5 and 6 for details).

2.3 From LSTM to Transformers

More generally, the recent trend in natural language processing research has been around using *Transformer* neural network architectures which are based on *attention mechanisms* (Vaswani et al., 2017). Thus, Alec Radford and Sutskever, 2018, Howard and Ruder, 2018 and Devlin et al., 2019 investigated language modeling for pre-training Transformer encoders and demonstrated dramatic improvements on classification tasks from the GLUE benchmark (Wang et al., 2018). Ramachandran, Liu, and Le, 2017 showed that machine translation tasks can also gain significant improvements by utilizing language modeling pre-training.

Zhao et al., 2018 introduced a supervised sentence simplification model based on the Transformer architecture and proposed two approaches to integrating the Simple PPDB (Pavlick and Callison-Burch, 2016) knowledge base for simplification that contains 4.5 million paraphrase rules. The first one is the *Deep Memory Augmented Sentence Simplification* (DMASS) model. It has an augmented dynamic memory to record multiple key-value pairs for each rule in the Simple PPDB which helps to overcome the problem when the neural network focuses more on frequent rules and ignores rare rules. The second model, *Deep Critic Sentence Simplification* (DCSS), encodes the context and the output of each simplification rules into the shared parameters.

Mikolov, Le, and Sutskever, 2013, Faruqui and Dyer, 2014, Xing et al., 2015 and Ammar et al., 2016 investigated usage of small dictionaries to align *word representations* from different languages. The need for cross-lingual supervision was slashed by Smith et al., 2017 and completely removed by Conneau et al., 2018.

Numerous works on the text simplification task prove once again its importance. Recent advances in the field of NLP have been dictated by *Transfer Learning* methods with Transformer language models. They became the source of our inspiration for this work and we believe they can rise text simplification systems to a new level.

Chapter 3

Evaluation

It is widely accepted that text simplification can be performed by *splitting*, *deletion* and *paraphrasing* (Feng, 2008). The splitting operation breaks down a long sentence into shorter ones. Deletion gets rid of unimportant parts of a sentence. The paraphrasing operation includes reordering, lexical substitutions and syntactic transformations (Xu et al., 2016). The best method for determining the quality of simplification is through human evaluation. Traditionally, a simplified output is judged in terms of grammaticality, meaning preservation and simplicity. For training and comparing models, the most commonly used automatic metrics are:

- *BLEU*, to assess an extent to which the output differs from the references;
- *SARI*, to evaluate the quality of the output by comparing it against the input and references;
- *FKGL*, to estimate the readability of the output.

3.1 BLEU

BLEU (*Bilingual Evaluation Understudy*) is a precision-oriented metric that estimates the proportion of n-gram matches between a system's output and a reference (Papineni et al., 2002). It was one of the first metrics which had shown a high correlation with human judgments of quality and remains one of the most popular automated, inexpensive and language-independent metrics.

BLEU uses a modified precision to compare a candidate translation against multiple references. The reason for the modification is that machine translation systems can generate more words than there are in the references. A simple precision measure sums the number of candidate n-grams which appear in any reference and then divides it by the total number of n-grams in the candidate translation. This may result in a poor translation with high precision (Table 3.1).

Candidate:	`the the the the the the the.`
Reference 1:	`The cat is on the mat.`
Reference 2:	`There is a cat on the mat.`

TABLE 3.1: Example of poor machine translation output with high precision. Source: Papineni et al., 2002.

All seven words in the candidate translation appear in the references. Thus a unigram simple precision is:

$$P = \frac{m}{w_t} = \frac{7}{7} = 1$$

where m is a number of words from the candidate found in the references, and w_t is the total number of words in the candidate. This is an example of a perfect score given for a poor translation.

A simple modification solves this issue. To calculate modified unigram precision we first count the maximum number of times a word occurs in any single reference translation. Next, we clip the total count of each candidate word by its maximum reference count $Count_{clip} = min(Count, Max\_Ref\_Count)$, add these clipped counts up, and divide by the total (unclipped) number of candidate words (Papineni et al., 2002). In the above example, the modified unigram precision score would be:

$$P = \frac{min(Count, Max\_Ref\_Count)}{w_t} = \frac{min(7,2)}{7} = \frac{2}{7}$$

The modified n-gram precision is computed similarly for any n: all candidate n-gram counts and their corresponding maximum reference counts are collected. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate n-gram. The n which has the highest correlation with human judgments was found to be 4. The unigram scores account for the adequacy of the translation, while the longer n-gram account for the fluency (Papineni et al., 2002).

One of the problem with the modified n-gram precision is that it fails to enforce the proper translation length. A possible candidate translation for the above example might be `the cat` and the modified unigram precision would be:

$$P = \frac{1}{2} + \frac{1}{2} = 1$$

To overcome this problem a *multiplicative brevity penalty factor* is used. With the brevity penalty in place, a high-scoring candidate translation must match the reference translations in length, in word choice, and in word order (Papineni et al., 2002). If the total length of the translation corpus c is less then or equal to the total length of the reference corpus r, the brevity penalty is decaying exponential with r/c:

$$BP = e^{1-\frac{r}{c}}$$

Thus BLEU score is the geometric mean of the test corpus's modified precision scores multiplied by an exponential brevity penalty factor. Geometric average of the modified n-gram precisions, p_n, is calculated using n-grams up to N and positive weights w_n summing to 1:

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n)$$

Despite being widely used and considered to be an informative metric for text-to-text generation, including text simplification, BLEU is not well suited for assessing simplicity from a lexical point of view (Xu et al., 2016). Moreover, BLEU often negatively correlates with simplicity, essentially penalizing simpler sentences (Sulem, Abend, and Rappoport, 2018a).

3.2 SARI

SARI, introduced by Xu et al., 2016, compares system output against the references and against the input sentence. It measures how the simplicity of a sentence was improved based on the words added, deleted and kept by the system (fig. 3.1).

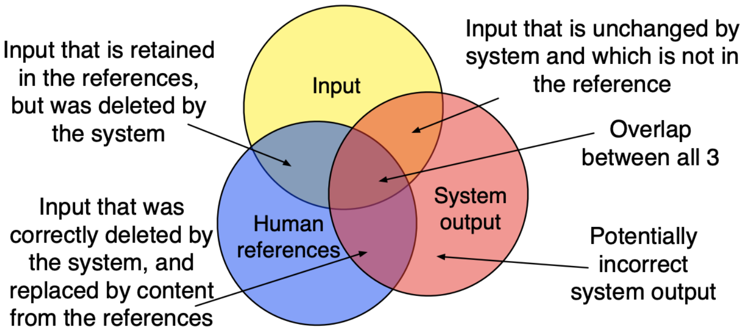
Input that is retained in the references, but was deleted by the system
Input
Input that is unchanged by system and which is not in the reference
Overlap between all 3
Input that was correctly deleted by the system, and replaced by content from the references
Human references
System output
Potentially incorrect system output

FIGURE 3.1: Metrics that evaluate the output of monolingual text-to-text generation systems. The different regions of this Venn diagram are treated differently with the SARI metric. Source: Xu et al., 2016.

SARI rewards addition operations, where system output O was not in the input I but occurred in any of the references R, i.e. $O \cap \bar{I} \cap R$. We define n-gram precision $p(n)$ and recall $r(n)$ for addition operations as follows (Xu et al., 2016):

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \bar{I})} \tag{3.1}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(R \cap \bar{I})} \tag{3.2}$$

where $\#g(\cdot)$ is a binary indicator of occurrence of n-grams g in a given set and

$$\#_g(O \cap \bar{I}) = \max\left(\#_g(O) - \#_g(I), 0\right)$$

$$\#_g(R \cap \bar{I}) = \max\left(\#_g(R) - \#_g(I), 0\right)$$

Example below (Table 3.2) demonstrates how the addition of word *now* is rewarded in both $p_{add}(n)$ and $r_{add}(n)$, but the addition of *you* in *Output 1* is penalized in $p_{add}(n)$.

Input:	`About 95 species are currently accepted.`
Reference 1:	`About 95 species are currently known.`
Reference 2:	`About 95 species are now accepted.`
Reference 3:	`95 species are now accepted.`
Output 1:	`About 95 you now get in.`
Output 2:	`About 95 species are now agreed.`
Output 3:	`About 95 species are currently agreed.`

TABLE 3.2: Example of sentence simplifications for SARI calculation. Source: Xu et al., 2016.

The *SARI* scores for these outputs are 0.2683, 0.7594, and 0.5890 respectively. The BLEU scores are 0.1562, 0.6435, and 0.6435. BLEU is unable to separate *Output 2* and *Output 3* because matching any of the references is rewarded in the same way.

SARI rewards keep operation, where n-grams are retained in both output and references. Number of such references matters. It bears in mind that some words or phrases don't require simplification:

$$p_{keep}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap O), \#_g I \cap R'\right)}{\sum_{g \in I} \#_g(I \cap O)} \tag{3.3}$$

$$r_{keep}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap O), \#_g(I \cap R')\right)}{\sum_{g \in I} \#_g(I \cap R')} \tag{3.4}$$

where

$$\#_g(I \cap O) = \min\left(\#_g(I), \#_g(O)\right)$$

$$\#_g(I \cap R') = \min\left(\#_g(I), \#_g(R)/r\right)$$

R' indicates the n-gram count over R with fractions. In the above example (Table 3.2) 2 out of the total $r = 3$ references contain the word `about`, thus its count is weighted by $2/3$.

For deletion, SARI calculates precision only. Deleting too many words decreases readability far more than not deleting:

$$p_{del}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap \bar{O}), \#_g I \cap \bar{R}'\right)}{\sum_{g \in I} \#_g(I \cap \bar{O})} \tag{3.5}$$

where

$$\#_g(I \cap \bar{O}) = \max\left(\#_g(I) - \#_g(O), 0\right)$$

$$\#_g(I \cap \bar{R}') = \max\left(\#_g(I) - \#_g(R)/r, 0\right)$$

Final SARI score calculates arithmetic average of n-gram precisions and recalls:

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del} \tag{3.6}$$

where

$$d_1 = d_2 = d_3 = 1/3$$

$$P_{operation} = \tfrac{1}{k} \textstyle\sum_{n=[1,\dots,k]} p_{operation}(n)$$

$$R_{operation} = \tfrac{1}{k} \textstyle\sum_{n=[1,\dots,k]} r_{operation}(n)$$

$$F_{operation} = \tfrac{2 \times P_{operation} \times R_{operation}}{P_{operation} + R_{operation}}$$

$$operation \in [del, keep, add]$$

where k is the highest n-gram order.

3.3 BLEU vs SARI

Xu et al., 2016 and Wubben, Bosch, and Krahmer, 2012 showed that *BLEU* does not demonstrate significant correlation with the simplicity scores rated by humans. In contrast, SARI achieves a much better correlation with human evaluations of simplicity. On the other hand, BLEU has a higher correlation on grammaticality and meaning preservation. BLEU gives a higher score to an output that is not too short and contains only n-grams that occur in references. When applied to monolingual tasks like simplification, it does not take into account any differences between the input and the references. Whereas SARI considers both precision and recall looking at the differences between the references and the input (Xu et al., 2016).

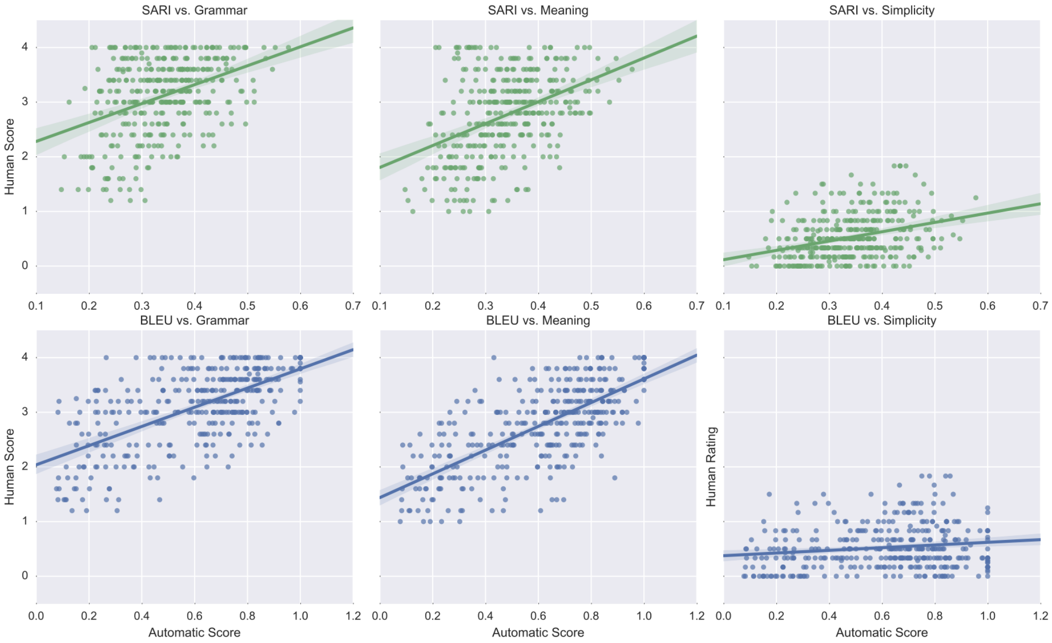

FIGURE 3.2: Scatter plots of automatic metrics against human scores for individual sentences. Source: Xu et al., 2016.

Scatter plots in fig. 3.2 highlights the correlation of human scores on meaning and grammar with BLEU and on simplicity with SARI. The outputs which are similar to the input get a high BLEU score. That is because for the monolingual simplification task, the more references are created the more n-grams of the input are included in the references. Outputs with few changes receive high grammar and meaning scores from humans as well, but it does not imply that they are good simplifications. Thus, BLEU prefers conservative systems that make few or no changes, while SARI penalizes them.

3.4 FKGL

Flesch-Kincaid Grade Level (FKGL) estimates the readability of text using cognitively motivated features (Kincaid et al., 1975). A lower value indicating higher readability. Commonly reported as measures of simplicity, FKGL relies on average sentence lengths and the number of syllables per word. Short sentences get low scores even if they have poor grammaticality or do not preserve meaning (Wubben, Bosch, and Krahmer, 2012). FKGL was developed by J. Peter Kincaid for the U.S. Navy in 1975. The Navy used the Flesch-Kincaid Grade score for assessing the difficulty of technical manuals.

The grade level is calculated with the following formula:

$$0.39\left(\frac{\#words}{\#sentences}\right) + 11.8\left(\frac{\#syllables}{\#words}\right) - 15.59 \tag{3.7}$$

A result is a number that corresponds to a US grade level. An FKGL score of 8 means that the reader needs at least a grade 8 level of reading to understand it. The FKGL coefficients were derived via multiple regressions applied to the reading compression test scores of 531 Navy personnel reading training manuals (Xu et al., 2016). The more words a sentence contains the more difficult it is. Similarly, words with many syllables are harder to read than words that use fewer.

3.5 Automatic sentence simplification evaluation

Alva-Manchego et al., 2019 introduced the *Easier Automatic Sentence Simplification Evaluation* (EASSE) framework[1], a Python package for automatic evaluation of the sentence simplification. EASSE provides a broad range of evaluation resources from standard automatic metrics (e.g. BLEU, SARI, FKGL) to quality estimations and comprehensive HTML reports on quantitative and qualitative assessments of a simplification output.

Using both the source sentence and the output simplification quality estimation, it brings additional insights into simplification systems which are not revealed by automatic metrics, e.g.:

- *compression ratio*, the length of the output sentence divided by the length of the input sentence;
- *proportion of exact matches* with the original sentences;
- average *proportion of added words*;
- average *proportion of deleted words*.

In this section, we found out that the evaluation of text simplification is not a simple task and requires multiple metrics for an accurate assessment. In this work, we will use BLEU, SARI, FKGL, Exact matches ratio, Addition, and Deletion ratios. In addition, in Section 5.5 we introduce *Compound Simplification Score* for comparing models with different BLEU, SARI and FKGL scores.

[1] https://github.com/feralvam/easse

Chapter 4

Datasets description

We conducted our experiments on three different simplification datasets, the summary statistics of which are presented in Table 4.1.

	WikiLarge	**Newsela**	**News Crawl/SW-CBT**
Source (monolingual)	291,402	81,705	1,500,000
Target (monolingual)	291,402	76,073	1,489,778
Train set	5,000	5,000	-
Validation set	2,000	1,500	-
Test set	359	1,500	-
Vocab source	41,303	33,316	43,222
Vocab target	39,912	22,405	49,118
Compression ratio	0.98	0.76	1.21
Sentence splits	1.09	1.01	0.99
FKGKL (source)	9.51	8.51	7.89
FKGKL (target)	6.33	2.86	5.46

TABLE 4.1: Datasets.

4.1 WikiLarge

The *Parallel Wikipedia Simplification* (PWKP) corpus introduced by Zhu, Bernhard, and Gurevych, 2010 has become a benchmark for training and evaluating text simplification models. It constitutes a collection of parallel sentences from the English Wikipedia[1] and Simple English Wikipedia[2]. Simple English Wikipedia is an online encyclopedia aimed at children and adults who are learning the English language. Its articles contain fewer words and simpler grammar than those in English Wikipedia.

WikiLarge is a Wikipedia corpus constructed by Zhang and Lapata, 2017. It is a combination of three datasets:

- PWKP (Zhu, Bernhard, and Gurevych, 2010), the dataset described above;
- aligned sentence pairs from Kauchak, 2013;
- aligned and revisioned sentence pairs from Woodsend and Lapata, 2011.

Originally it had 296,402 sentence pairs but we took 5,000 pairs for machine translation step during our model training (see Chapter 6 for details). For validations and tests, we used datasets created by Xu et al., 2016. They consist of complex sentences

[1] https://en.wikipedia.org/

[2] https://simple.wikipedia.org/

from the WikiSmall dataset aligned with simplifications provided by *Amazon Mechanical Turk* [3]. Each original sentence in the dataset has 8 simplified references. See Table 4.1 for details.

4.2 Newsela

Newsela dataset was introduced by Xu, Callison-Burch, and Napoles, 2015. The authors argued that Wikipedia as a simplification data resource is sub-optimal because it is prone to automatic sentence alignment errors, contains a large proportion of inadequate simplifications and it generalizes poorly to other text genres.

Newsela is a platform that provides reading materials for classroom usage[4]. On request, they provide a corpus that includes thousands of news articles professionally leveled to different reading complexities. For every original sentence (Version 0) there are 4 or 5 simplified versions (Version 5 or 6 being the simplest).

We used the most contrast article versions: 0-level for a source dataset and 4-level for a target dataset. For the machine translation step, for the test, and for the validation datasets we used parallel complex-simple pairs provided by Xu, Callison-Burch, and Napoles, 2015. See Table 4.1 for details.

4.3 News Crawl and SimpleWiki with Children's Books Test

To test the performance of our model on a corpus of different styles and sizes we collected our own datasets for training and used the Wikilarge and the Newsela sets for the machine translation step, test and validation.

As a source "complex" monolingual dataset we used 1,500,000 sentences from the WMT 2014 News Crawl[5], a dataset consisting of text crawled from online news. For target "simple" dataset we combined sentences from SimpleWiki (SW)[6] with the Children's Books Test (CBT) from Hill et al., 2015. The CBT is built from children books freely provided by Project Gutenberg [7]. After removing duplicates from the SW-CBT dataset, the resulting target monolingual dataset contains 1,489,778 sentences. See Table 4.1 for details.

4.4 Data Pre-processing

For data pre-processing we used a script provided by XLM model[8]. It uses Moses[9] to replaces Unicode punctuation, normalize it, remove non-printing characters and tokenize the data. Then it uses fastBPE[10] to apply 60,000 BPE (Byte Pair Encoding) codes[11] to monolingual and parallel test data. These BPE codes were learned during the training of the pre-trained XLM model which we use for our experiments. Finally, the script generates the same shared vocabulary through the BPE codes to improve the alignment of embedding spaces across languages.

[3] https://www.mturk.com/

[4] https://newsela.com/

[5] http://statmt.org/wmt14/training-monolingual-news-crawl/

[6] https://dumps.wikimedia.org/simplewiki/latest/

[7] https://gutenberg.org/

[8] https://github.com/facebookresearch/XLM/blob/master/get-data-nmt.sh

[9] http://www.statmt.org/moses/

[10] https://github.com/glample/fastBPE

[11] https://dl.fbaipublicfiles.com/XLM/codes_enfr

In this chapter, we described Wikilarge and Newsela datasets which became benchmarks for the evaluation of text simplification systems. Furthermore, we introduced our own monolingual dataset based on News Crawl, SimpleWiki and Children Books Test. Vocabulary size, compression ratio, and FKGL score prove once again the high quality of the Newsela dataset.

Chapter 5

Methodology

In this chapter, we outline the methodology for our experiments. First of all, we describe the model we use. Then we review beam search generation and proposed penalization. In the end, we introduce the cross-validation technique we use and reveal details on the training process.

5.1 Unsupervised Machine Translation

Lample et al., 2018b and Artetxe et al., 2018 have proposed unsupervised Machine Translation (MT) which relies on monolingual (i.e., non parallel) corpora only. The authors have defined four key principles required for training such models:

- MT system initialization;
- language modeling;
- iterative back-translation (Sennrich, Haddow, and Birch, 2016);
- shared encoder latent representations.

Building on this idea, Lample et al., 2018a introduced an UnsupervisedMT[1] model that outperforms previous approaches and is easier to train and tune.

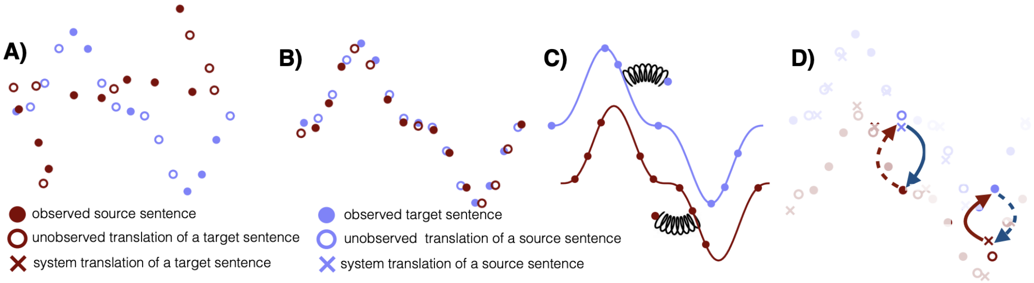
A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
× system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
× system translation of a source sentence

FIGURE 5.1: Toy illustration of the three principles of unsupervised MT. Source: Lample et al., 2018a

Fig. 5.1 demonstrates the usage of the above-mentioned principals. A) There are two monolingual corpora. B) **Initialization**. The two distributions are aligned by performing word-by-word translation. C) **Language modeling**. A language model is learned independently in each domain to infer the structure in the data. D) **Back-translation**. Starting from an observed source sentence they use the current source → target model(dashed arrow), yielding a potentially incorrect translation (blue cross

[1] https://github.com/facebookresearch/UnsupervisedMT

near the empty circle). Starting from this (back) translation, they use the target → source model (continuous arrow) to reconstruct the sentence in the original language. The discrepancy between the reconstruction and the initial sentence provides an error signal to train the target → source model parameters. The same procedure is applied in the opposite direction to train the source → target model (Lample et al., 2018a).

5.2 XLM

Based on the ideas of aligning the distributions of sentences in different languages, Lample and Conneau, 2019 reduced the need for parallel data. They introduced supervised and unsupervised approaches for cross-lingual language models (XLMs[2]) training based on Transformers' architecture Vaswani et al., 2017. The unsupervised method relies on monolingual corpora only, whereas the supervised one leverages parallel data. The XLM model achieves a better performance than the original BERT[3] on all GLUE tasks[4].

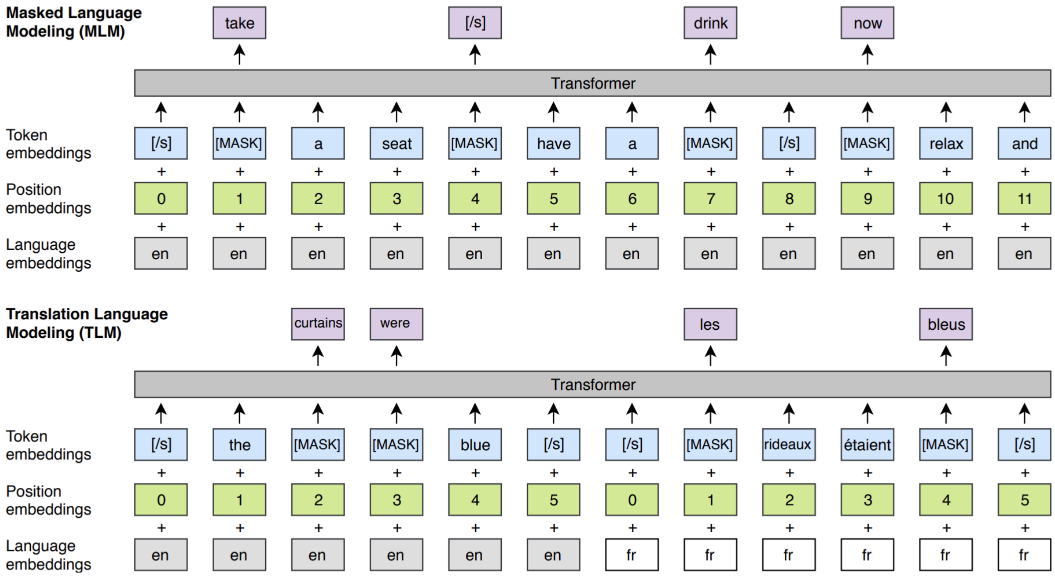
Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr

FIGURE 5.2: Cross-lingual language model pretraining. Source: Lample and Conneau, 2019

The unsupervised **cross-lingual text representations** are obtained with the help of *Causal Language Modeling* (CLM) and *Masked Language Modeling* (MLM) training objectives. During training, they process all languages with the same shared vocabulary created through Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2015). CLM is a Transformer language model trained to predict the probability of a word given the previous words in a sentence $P(w_t|w_1, \ldots, w_{t-1}, \Theta)$. MLM assumes random sampling of 15% of the BPE tokens from the text streams, replacing them by a [MASK] token 80% of the time, by a random token 10% of the time, and keeping them unchanged 10% of the time (fig. 5.2).

Since both the CLM and MLM only require monolingual data, they cannot be used to utilize parallel data. *Translation Language Modeling* (TLM) leverages parallel corpora to improve cross-lingual pre-training. It extends the BERT MLM approach

[2] https://github.com/facebookresearch/XLM
[3] https://github.com/google-research/bert
[4] https://github.com/facebookresearch/XLM#i-monolingual-language-model-pretraining-bert

by using parallel sentences (fig. 5.2). TLM randomly masks words in both the source and target sentences. To predict a word masked in a source sentence, the model can either attend to surrounding source words or to the target translation, encouraging the model to align the source and target representations. The target context can be used if the source one is not sufficient to guess the masked source words (Lample and Conneau, 2019).

To the best of our knowledge, cross-lingual language modeling has not been applied before for the task of text simplification. Following this approach, we used the XLM model for our experiments.

We included the following 6 steps into the model training: CLM, TLM, *Parallel Classification* (PC), *Denoising Auto-Encoder* (AE), *Machine Translation* (MT) and *Back-translation* (BT). During the PC step, the model predicts if pairs of sentences are mutual translations of each other. AE and MT steps are similar with the only difference that for AE step the model uses mono language sentences and add noise before masking and encoding. The BT step, described in the previous section, is similar to Lample et al., 2018a. MT is a supervised machine translation step. In our experiments, we first consider the settings with no supervision (i.e., by excluding MT and TLM steps) and later added an MT step trained on a small parallel corpus to attest the extent to which a little supervision can help with the simplification problem. It is worth noting that the addition of TLM step, which also relies on parallel corpora, had marginal impact on the performance of the models.

The importance of each step can be weighted by a coefficient but we did not see any improvement when changing the values of the coefficients and used the default lambdas of 1 for every step.

5.3 Beam Search Generation

Beam Search is a common technique to improve decoding performance. Instead of decoding the most probable words in a greedy fashion, it generates an output sentence by keeping a fixed number (specified by beam size parameter) of hypotheses with the highest log-probability at each step. The approach explores a set of candidate hypotheses until the sentence is fully decoded and selects the one with the highest log-probability at the end (Fig. 5.3).

Such a decoding strategy based on scoring provides us with additional control over sentence generation. To manage the exact matches ratio, length and simplicity (FKGL-based) of a hypothesis we added three types of score penalties.

Length penalty (LP) favors shorter or longer hypothesis depending on λ_{length} parameter:

$$LP = \lambda_{length} \times \exp(length(hypothesis))$$

Exact matches penalty (EMP) uses cosine similarity between input and hypothesis to restrict copying of input:

$$EMP = \lambda_{exact\_matches} \times \exp(cosine\_similarity(input, hypothesis))$$

FKGL penalty (FKGLP) encourages hypothesis with lower FKGL score:

$$FKGLP = \lambda_{FKGL} \times \exp(FKGL(hypothesis))$$

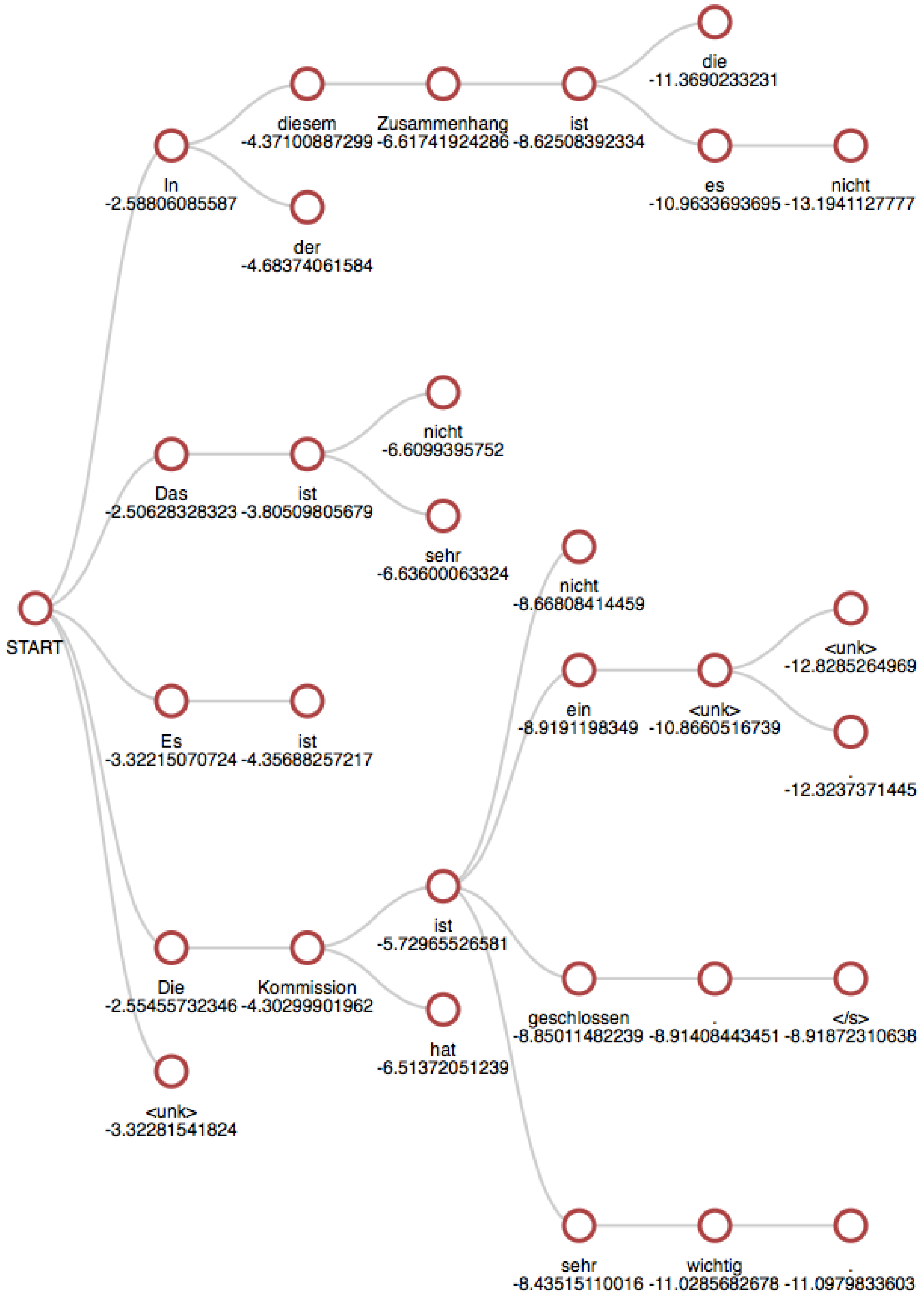

FIGURE 5.3: Visualisation of beam search of width 5. Source: OpenNMT.

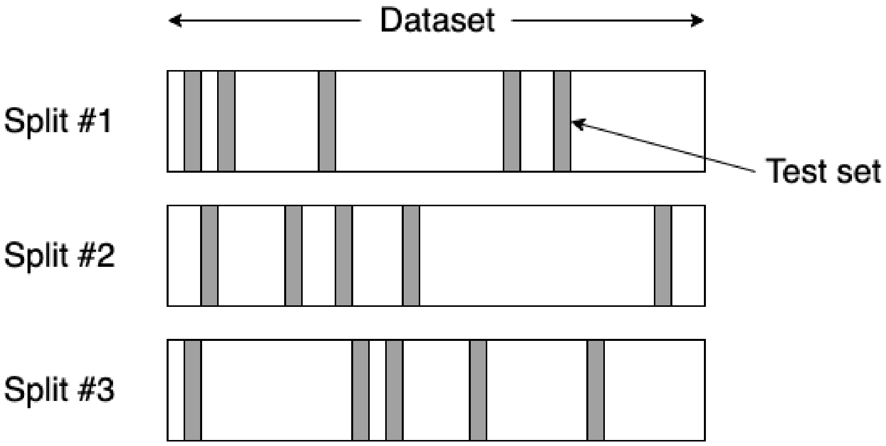

FIGURE 5.4: Repeated random subsampling. Grey areas are test data and white ares are train data.

We demonstrate that together these beam search penalties make it possible to improve the decoding results of a model after training.

5.4 Random Subsampling Validation

We choose random sub-sampling validation (Dubitzky, 2007) for assessing how our models will generalize to an independent data set and for eliminating statistical errors due to dataset split. This approach belongs to *non-exhaustive cross validation* methods. It does not compute all possible splits of the dataset but creates multiple random splits into training and test data (Fig. 5.4). For each such split, we train a model on training data and evaluate using test data. The final results are then averaged over the splits. The advantage of this method is that the proportion of the train/test split is not dependent on the number of partitions. The disadvantage is that test sets may overlap and some examples may never be selected. To overcome this possible problem we repeat this procedure 10 times.

5.5 Training Details

We trained our models on the Nvidia Tesla V100 GPU card and used the same hyper-parameters across datasets except for beam search penalty coefficients.

Both encoder and decoder have **embedding layer of size 1024**, **6 attention layers** with **8 heads** and *GELU* activation function, and regularized with an **attention dropout rate of 0.1**.

We used *Adam* optimizer with learning rate decay based on the inverse square root of the update number. **Learning rate** was set to **0.0001**, the first momentum coefficient was set to 0.9 and the second momentum coefficient to 0.98.

We examined our XLM models performance based on a range of simplification metrics discussed in Chapter 3. To determine which set of scores is better we introduced a new *Compound Simplification Score* (CSS). Since BLEU and SARI take values from 0 to 100 (the higher the better) and FKGL takes values from 0 to 10 on our datasets (the lower the better) we defined it as follows:

$$CSS = \frac{BLEU}{100} + \frac{SARI}{100} + \frac{100 - FKGL \times 10}{100}$$

We used CSS as our stopping criteria, i.e., 5 epochs of non-increasing number. We used different epoch sizes for different datasets based on their sizes: 80,000 sentences for Newsela, 150,000 for Wikilarge and 200,000 for News/SW-CBT datasets. On average a single epoch took 15 minutes on Newsela, 35 minutes on Wikilarge and 40 minutes on News/SW-CBT datasets. All our models on all the datasets converged within 10-15 epochs.

Chapter 6

Experiments

We performed our experiments on the XLM model described in Section 5.2. The baseline unsupervised XLM models trained on Newsela and Wikilarge datasets gave us encouraging results in comparison with the supervised models from the literature. Further improvements achieved by adding limited supervised Machine translation (MT) step, new monolingual corpora and modified beam search generation led us to the best SARI result on Newsela dataset and the best BLEU result on Wikilarge dataset.

6.1 Comparison models

We compared the performance of our model against multiple others mentioned in Chapter 2. *PBMT-R* is phrase-based machine translation system with a re-ranking post-processing step proposed by Wubben, Bosch, and Krahmer, 2012. *Hybrid* is a simplification model that includes a probabilistic model for splitting and dropping and a PBMT-R model for substitution and reordering (Narayan and Gardent, 2014). *SBMT-SARI* is a syntax-based translation model trained on PPDB (Ganitkevitch, Van Durme, and Callison-Burch, 2013) and trained with SARI (Zhang and Lapata, 2017). *EncDecA*, a basic attention-based encoder-decoder model, *DRESS*, a deep reinforcement learning model, *DRESS-LS*, a linear combination of DRESS and the lexical simplification model, all of them were introduced in Zhang and Lapata, 2017. *DMASS+DCSS* is a combination of DMASS and DCSS models from Zhao et al., 2018 (Section 2.3).

The BLEU, SARI and FKGL results for the above-mentioned models were taken from Zhang and Lapata, 2017 and Zhao et al., 2018. For the models introduced in Chapter 5 we also report Exact matches, Addition and Deletion ratios which provide additional insights into performance of simplification systems.

6.2 Unsupervised approach

Our baseline XLM models were trained following completely unsupervised approach. The model trained on Newsela dataset showed good results on all metrics apart from FKGL. It also has high Exact matches ratio and low Addition and Deletion ratios (Table 6.1) suggesting that the model chose a conservative strategy of copying source sentences in many cases.

The model trained on Wikilarge dataset showed an excellent BLEU score. Such a "good" performance is explained by high Exact matches ratio of 0.93 (Table 6.2). Due to the nature of Wikilarge dataset (Section 4.2) a model can just duplicate the input and it will obtain a very high BLEU score.

	BLEU	**SARI**	**FKGL**	**Exact Match**	**Add.**	**Del.**
PBMT-R	18.19	15.77	7.59	-	-	-
Hybrid	14.46	30.00	4.01	-	-	-
EncDecA	21.7.0	24.12	5.11	-	-	-
DRESS	23.21	27.37	4.13	-		
DRESS-LS	**24.30**	26.63	4.21	-	-	-
DMASS+DCSS	-	27.28	5.17	-	-	-
XLM	16.97	19.32	10.52	0.46	0.04	0.05
XLM (News/SW-CBT)	18.50	12.94	10.36	0.97	0.00	0.00
XLM (News/SW-CBT, penalized beam)	18.34	13.49	8.46	0.72	0.01	0.00
XLM (MT)	19.44	**43.18**	4.18	0.09	0.12	0.53
XLM (MT, News/SW-CBT)	19.33	39.60	5.57	0.21	0.10	0.45
XLM (MT, News/SW-CBT, penalized beam)	16.56	42.24	**3.58**	0.04	0.1	0.6
Output = Source	18.52	12.78	10.36	1.00	0.00	0.00
Output = Target	100.00	100.00	4.18	0.00	0.19	0.61

TABLE 6.1: Automatic evaluation on Newsela test set. Source: Zhang and Lapata, 2017, Zhao et al., 2018.

	BLEU	**SARI**	**FKGL**	**Exact Match**	**Add.**	**Del.**
PBMT-R	81.11	38.56	8.33	-	-	-
Hybrid	48.97	31.40	**4.56**	-	-	-
SBMT-SARI	73.08	39.96	7.29	-	-	-
EncDecA	88.85	35.66	8.41	-	-	-
DRESS	77.18	37.08	6.58	-	-	-
DRESS-LS	80.12	37.27	6.62	-	-	-
DMASS+DCSS	-	**40.42**	7.18	-	-	-
UNTS+10K	76.13	35.29	-	-	-	-
XLM	94.83	28.30	9.75	0.76	0.02	0.01
XLM (News/SW-CBT)	**96.91**	28.00	9.94	0.93	0.00	0.00
XLM (News/SW-CBT, penalized beam)	94.95	30.03	9.82	0.45	0.01	0.03
XLM (MT)	92.66	30.99	9.68	0.73	0.02	0.02
XLM (MT, News/SW-CBT)	96.05	29.44	9.81	0.89	0.01	0.02
XLM (MT, News/SW-CBT, penalized beam)	76.93	35.63	7.74	0.3	0.04	0.26
XLM (MT, Newsela)	3.63	31.80	6.24	0.01	0.17	0.44
Output = Source	97.41	27.32	9.90	1.00	0.00	0.00
Output = Target	68.87	40.83	8.33	0.00	0.19	0.21

TABLE 6.2: Automatic evaluation on Wikilarge test set. Source: Zhang and Lapata, 2017, Zhao et al., 2018, Surya et al., 2019.

	BLEU	SARI	FKGL	Matches	Add.	Del.
XLM (MT) #0	18.25	43.33	4.17	0.07	0.14	0.53
XLM (MT) #1	20.27	43.39	4.11	0.08	0.11	0.53
XLM (MT) #2	19.09	43.23	3.91	0.08	0.12	0.55
XLM (MT) #3	18.58	43.30	3.94	0.08	0.14	0.56
XLM (MT) #4	20.78	43.44	4.37	0.07	0.12	0.53
XLM (MT) #5	17.35	42.96	3.97	0.08	0.14	0.56
XLM (MT) #6	19.22	42.77	4.15	0.08	0.12	0.54
XLM (MT) #7	20.07	43.36	4.30	0.10	0.11	0.52
XLM (MT) #8	20.23	42.52	4.82	0.13	0.11	0.47
XLM (MT) #9	20.58	43.45	4.06	0.10	0.11	0.52
Mean	19.44	43.18	4.18	0.09	0.12	0.53
Variance	1.28	0.10	0.07	0.00	0.00	0.00

TABLE 6.3: Repeated random sub-sampling validation on Newsela train and test sets.

6.2.1 Adding larger monolingual corpus

Since Newsela and Wikilarge are not very large datasets, a possible option to improve the performance of the model was to train the baseline XLM model on larger monolingual corpora. Therefore, we trained the baseline model on the SW-CBT dataset which has about 1,500,000 sentences (described in Section 4.3) and evaluated it on the Newsela and Wikilarge test sets. The resulting *XLM (News/SW-CBT)* model has also featured a tendency to copy the input which, as in the case with XLM model, we have regularized later in this chapter by introducing a supervised MT step (Table 6.1).

6.2.2 Adding larger monolingual corpus and penalized beam search

For Newsela, adding penalization worked well and reduced FKGL from 10.36 to 8.46 points and the Exact matches ratio from 0.97 to 0.72 with a slight improvement of SARI. As for Wikilarge, SARI increased from 28 to 30.03, FKGL dropped from 9.94 to 9.82 points, Exact matches ratio plunged from 0.93 to 0.45. The BLEU score worsened a little for both datasets.

6.3 Adding limited supervision

By adding an MT step with just 5,000 parallel sentences, we helped the model trained on Newsela dataset to learn to remove redundant information. This has dramatically improved the performance of the model. The **SARI score skyrocketed from 23.29 to 43.18** points, **FKLG dropped from 10.39 to 4.18** and Deletion ratio increased almost 7-fold alongside with the three-times drop in Exact matches ratio (Table 6.1).

To ensure that the obtained results are not due to a statistical error we conducted a repeated random sub-sampling validation. We created 10 random splits of the dataset into training and test data (Table 6.3). The mean scores over the splits gave a Deletion ratio of 0.53 out of the best possible 0.61 points[1]. Along with this, we obtained the best SARI score of 43.18 among all simplification models known to us.

For Wikilarge, additional MT step gave a little improvement in terms of SARI and FKGL scores but reduced the BLEU result (Table 6.2).

[1] Best possible is achieved when simplified sentences equal target ones.

6.3.1 Adding larger monolingual corpus

XLM model trained on SW-CBT dataset with an MT step and evaluated on both Newsela and Wikilarge demonstrated a worsening of almost all metrics (Tables 6.1 and 6.2) with clear commitment to copy source sentences.

6.3.2 Adding monolingual corpus and penalized beam search

To address the issue with the input copying by the XLM (MT, News/SW-CBT) model, we again used the beam search penalties. For Newsela, this drastically reduced the Exact matches ratio from 0.21 to 0.04 and **FKGL from 5.57 to 3.58**. Thus we obtained **much better FKGL score than the previous best result of 4.01 points by the Hybrid model** (Table 6.1).

As for Wikilarge, this additional regularisation markedly improved all metrics except of BLEU. A dramatically improved Deletion ratio had negative effect on it. BLEU doesn't encourage shorter sentences (Section 3.1) and, hence it reduced its score from 96.05 to 78.01 (Table 6.2). Table 6.9 presents some good examples of improvements in comparison with the XLM (MT) model.

6.4 Trained on Newsela, evaluated on Wikilarge

Since we obtained good result on Newsela dataset, we decided to evalute Wikilarge test set on XLM (MT) model trained on Newsela dataset. XLM (MT, Newsela) model received extremely low BLEU score of 3.63 due to increased Deletion ratio of 0.44 points. More importantly, XLM (MT, Newsela) model was able to get low enough FKGL score (Table 6.2).

6.5 Newsela outcomes

In the Table 6.4 we can see how different models simplify a sentence from the Newsela dataset. Our XLM (MT) model is the only one which replaced `variety of skills` with `range of skills` as the target version did, but it was unable to make the sentence shorter. The worst simplification seems to be provided by Hybrid model. Even though it is the shortest one it does not make any sense.

The best simplifications according to SARI are presented in Table 6.5. In the first example the model made a simplification exact to target, while in the second example is was very close.

Sometimes XLM (MT) model overdo it with the sentence compression. A good example of such behavior is presented in Table 6.6.

One of the key properties of good simplification models is their ability to split long sentences into smaller ones. XLM (MT) tries to do that but could not boast of much success (Table 6.7).

On the other hand, for some sentences XLM (MT) simplifies better by making an output sentence longer than the input one. (Table 6.8).

Newsela contains high quality simplifications created by professional editors, thus it is not easy to teach a model to do right simplifications. It is not enough just to copy the input (but we will see that this may be a good strategy on Wikilarge dataset). The target simplifications contain a large ratio of addition (0.19) and deletion (0.61). We believe that the larger corpora based on Newsela articles may remarkably improve the results.

Source	There's just one major hitch: the primary purpose of education is to develop citizens with a wide variety of skills.
Target	The purpose of education is to develop a wide **range** of skills.
PBMT-R	It's just one major hitch: the purpose of education is to **make people** with a wide variety of skills.
Hybrid	one hitch the purpose is to develop citizens.
EncDecA	The **key** of education is to develop **people** with a wide variety of skills.
DRESS	There's just one major hitch: the **main goal** of education is to develop **people** with **lots of** skills.
DRESS-LS	There's just one major hitch: the **main goal** of education is to develop citizens with **lots of** skills.
XLM (MT)	There's just one **big** hitch: the primary purpose of education is to develop citizens with a wide **range** of skills.

TABLE 6.4: System outputs on Newsela dataset. Source: Zhang and Lapata, 2017.

Source	Florida sees more stranded whales than **another** state, **followed by California**.
Target	Florida sees more stranded whales than **any other** state.
XLM (MT)	Florida sees more stranded whales than **any other** state.
Source	Sage Kotsenburg, one of White's Olympic **teammates, called the modified course "sick" - a compliment, in this world.**
Target	Sage Kotsenburg **is** one of White's Olympic **teammates**.
XLM (MT)	Sage Kotsenburg **is** one of White's Olympic athletes.

TABLE 6.5: Best simplifications on Newsela dataset according to SARI.

Source	**Making the site even more significant, they say, is the fact that Carr's team has also uncovered artifacts and other elements from two later historic structures sandwiched over the Tequesta village at the site - a well and artifacts from Fort Dallas, a mid-19th century military fortification used during two of the Seminole Indian wars, and brick column bases and other traces of Flagler's hotel, which prompted the founding of the city of Miami.**
Target	Carr's team has **found other** artifacts **there. Two building were** later **constructed on top of** the village.
XLM (MT)	**The** team **found some important pieces.**

TABLE 6.6: Simplification with the most compression on Newsela dataset.

Source	**Entering, for instance, museum-goers** will cross a **water feature** to **recall** the **experience of slaves crossing the ocean** to **come to** America.
Target	**Museum-goers** will **enter the building across** a **body of water.**
XLM (MT)	**Visitors** will cross a **waterway** to **see** the **story. Visitors will walk a waterway** to America.

TABLE 6.7: Simplification with sentence split on Newsela dataset.

Source	The **notion** that Snowden had no **option** but to leak is indefensible.
Target	**But, the** notion that Snowden had no **choice** but to leak **secrets** is indefensible.
XLM (MT)	The **idea** that Snowden had no **choice** but to leak **the information** is indefensible.

TABLE 6.8: Simplification that is longer than the source on Newsela dataset.

Source	Brighton is a city in Washington county, Iowa, United States.
Target	Brighton is a city **of** Iowa **in the** United States.
XLM (MT)	Brighton is a city in Washington county, Iowa, United States.
XLM (MT, penalized beam)	Brighton is a city **of** Iowa **in the** United States.
Source	Despina was discovered in late July, 1989 from the images taken by the Voyager 2 probe.
Target	Despina was **found** in late July, 1989 from the images taken by the Voyager 2 probe.
XLM (MT)	Despina was discovered in late July, 1989 from the images taken by the Voyager 2 probe.
XLM (MT, penalized beam)	Despina was **found** in late July, 1989 from the images taken by the Voyager 2 probe.

TABLE 6.9: Penalized beam search helps to overcome a problem when the system copies input on Wikilarge dataset.

6.6 Summary

We conducted our experiments following unsupervised and semi-supervised methods, i.e., by adding supervised MT step trained on parallel corpora. We attested two strategies to improve performance. The first one lied in training the model on a large monolingual corpus with penalized beam search generation. The second one consisted of adding limited supervision through a Machine translation step trained on 5,000 parallel sentences.

For the Newsela, the first strategy improved BLEU and FKGL scores but had a negative impact on SARI and Exact matches ratio. The MT step, in its turn, dramatically improved all the metrics giving the best SARI and FKGL scores among all the models known to us.

As for Wikilarge, the SW-CBT dataset makes it possible to obtain an unprecedented BLEU score while slightly reducing other metrics. With respect to the second strategy, the most noticeable improvement was reached by XLM (MT, News/SW-CBT, penalized beam) model (Tables 6.1 and 6.2).

In general, we noticed that adding a large monolingual SW-CBT dataset had a positive impact on BLEU scores, while MT step highly improves SARI and FKGL results.

Chapter 7

Conclusion

7.1 Summary of contributions

In this work we considered the task of sentence simplification in an unsupervised and semi-supervised fashion and made the following contributions:

1. To the best of our knowledge, our work is the first attempt to apply cross-lingual language modeling to the text simplification problem.
2. We introduced regularisation penalties for beam search generation to control exact matches, length and FKGL score of a simplified sentence. This gave us an increase of SARI by 2.64 and FKGL by 1.99 points on the Newsela dataset and improved SARI by 6.19 and FKGL by 2.07 points on the Wikilarge dataset with the semi-supervised approach.
3. In comparison to previous work in this direction, we have conducted a more comprehensive evaluation by using a larger variety of simplification metrics.
4. We collected a brand new 1,500,000 sentences monolingual dataset and applied it to unsupervised training steps which yielded an additional increase of 1.53 points in BLEU score on the Newsela dataset and 2.08 points on the Wikilarge dataset with the unsupervised approach.

Overall, we developed two approaches for text simplification using cross-lingual language modeling. The first one is completely unsupervised. The second one is semi-supervised, which uses a small parallel corpus of 5,000 sentences in addition to a much large monolingual one. The unsupervised approach gave us the best BLEU score on the Wikilarge dataset, whereas the semi-supervised demonstrated the best SARI and FKGL scores on the Newsela dataset, therefore improving the state-of-the-art results by a margin of 9.08%, 43.93%, and 10.72%, correspondingly.

7.2 Directions for future research

One of the most promising directions for future research we see in devising larger and higher quality monolingual datasets. Specifically, we believe that this research will benefit from the new text corpora with a broader variety of FKGL grades between the source and the target sentences, better-simplified vocabularies in the output and a sufficient amount of training examples with sentence splitting (i.e., when a single complex sentence is split into multiple simpler ones) which often provide a better simplification output.

Another interesting related problem lies in generating simplifications with tuneable grade levels. There are multiple ways to achieve this, for instance, by training separate models for different grade levels; weighing or constraining the proposed FKGL penalty by the required output grade level; etc. We also consider controlling the output "simplified" vocabulary by either introducing a penalty for utilizing less commonly used words or relaxing the constraint on using shared representations for simplified and original languages in the decoding architecture.

Last but not least, we believe that the future research in this direction will benefit from a better evaluation of the grammaticality of the generated simplifications by either conducting a human review of the output or by analyzing its semantic decomposition.

Bibliography

Alec Radford Karthik Narasimhan, Tim Salimans and Ilya Sutskever (2018). “Improving language understanding by generative pre-training”. In: URL: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/languageunderstandingpaper.pdf.

Allen, David (Dec. 2009). “A study of the role of relative clauses in the simplification of news texts for learners of English”. In: *System* 37, pp. 585–599. DOI: 10.1016/j.system.2009.09.004.

Alva-Manchego, Fernando et al. (Nov. 2019). “EASSE: Easier Automatic Sentence Simplification Evaluation”. 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): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, pp. 49–54. URL: https://www.aclweb.org/anthology/D19-3009.

Ammar, Waleed et al. (2016). “Massively Multilingual Word Embeddings”. In: *ArXiv* abs/1602.01925.

Artetxe, Mikel et al. (2018). “Unsupervised neural machine translation”. In: *Proceedings of the Sixth International Conference on Learning Representations*.

Biran, Or, Samuel Brody, and Noémie Elhadad (June 2011). “Putting it Simply: a Context-Aware Approach to Lexical Simplification”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 496–501. URL: https://www.aclweb.org/anthology/P11-2087.

Brouwers, Laetitia et al. (Apr. 2014). “Syntactic Sentence Simplification for French”. In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 47–56. DOI: 10.3115/v1/W14-1206. URL: https://www.aclweb.org/anthology/W14-1206.

Candido Jr., Arnaldo et al. (2009). “Supporting the Adaptation of Texts for Poor Literacy Readers: A Text Simplification Editor for Brazilian Portuguese”. In: *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. EdAppsNLP ’09. Boulder, Colorado: Association for Computational Linguistics, pp. 34–42. ISBN: 978-1-932432-37-4. URL: http://dl.acm.org/citation.cfm?id=1609843.1609848.

Canning, Y. and J. Taito (Jan. 1999). “Syntactic simplification of newspaper text for aphasic readers”. In:

Canning, Yvonne et al. (2000). “Cohesive Generation of Syntactically Simplified Newspaper Text”. In: *Proceedings of the Third International Workshop on Text, Speech and Dialogue*. TDS ’00. London, UK, UK: Springer-Verlag, pp. 145–150. ISBN: 3-540-41042-2. URL: http://dl.acm.org/citation.cfm?id=647238.720905.

Carlson, Keith, Allen Riddell, and Daniel Rockmore (Oct. 2018). “Evaluating prose style transfer with the Bible”. In: *Royal Society Open Science* 5, p. 171920. DOI: 10.1098/rsos.171920.

Carroll, John et al. (June 1999). “Simplifying Text for Language-Impaired Readers”. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Norway: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E99-1042.

Chandrasekar, Raman and Bangalore Srinivas (1997). “Automatic induction of rules for text simplification”. In: *Knowl.-Based Syst.* 10.3, pp. 183–190. DOI: 10.1016/S0950-7051(97)00029-4. URL: https://doi.org/10.1016/S0950-7051(97)00029-4.

Clarke, James and Mirella Lapata (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 the 44th Annual Meeting of the Association for Computational Linguistics*. ACL-44. Sydney, Australia: Association for Computational Linguistics, pp. 377–384. DOI: 10.3115/1220175.1220223. URL: https://doi.org/10.3115/1220175.1220223.

Conneau, Alexis et al. (2018). “Word Translation Without Parallel Data”. In: *International Conference on Learning Representations (ICLR)*.

Coster, William and David Kauchak (June 2011). “Simple English Wikipedia: A New Text Simplification Task”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 665–669. URL: https://www.aclweb.org/anthology/P11-2117.

De Belder, Jan and Marie-Francine Moens (Jan. 2010). “Text simplification for children”. In:

Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*.

Devlin, Siobhan and John Tait (1998). “The use of a psycholinguistic database in the simplification of text for aphasic readers”. In:

Dubitzky Werner; Granzow, Martin; Berrar Daniel (2007). *Fundamentals of data mining in genomics and proteomics.* Springer Science & Business Media, p. 178.

Elhadad, Noemie and Komal Sutaria (2007). “Mining a Lexicon of Technical Terms and Lay Equivalents”. In: *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. BioNLP ’07. Prague, Czech Republic: Association for Computational Linguistics, pp. 49–56. URL: http://dl.acm.org/citation.cfm?id=1572392.1572402.

Faruqui, Manaal and Chris Dyer (2014). “Improving Vector Space Word Representations Using Multilingual Correlation”. In: *EACL*.

Feng (2008). “Text simplification: A survey.” In: *Text simplification: A survey*. The City University of New York, Technical Report.

Filippova, Katja and Michael Strube (June 2008). “Dependency Tree Based Sentence Compression”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. Salt Fork, Ohio, USA: Association for Computational Linguistics, pp. 25–32. URL: https://www.aclweb.org/anthology/W08-1105.

Filippova, Katja et al. (Sept. 2015). “Sentence Compression by Deletion with LSTMs”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 360–368. DOI: 10.18653/v1/D15-1042. URL: https://www.aclweb.org/anthology/D15-1042.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch (June 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. Atlanta, Georgia: Association for Computational Linguistics, pp. 758–764. URL: https://www.aclweb.org/anthology/N13-1092.

Glavaš, Goran and Sanja Štajner (July 2015). "Simplifying Lexical Simplification: Do We Need Simplified Corpora?" 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)*. Beijing, China: Association for Computational Linguistics, pp. 63–68. DOI: 10.3115/v1/P15-2011. URL: https://www.aclweb.org/anthology/P15-2011.

Hill, Felix et al. (2015). "The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations". In: *CoRR* abs/1511.02301.

Howard, Jeremy and Sebastian Ruder (July 2018). "Universal Language Model Finetuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://www.aclweb.org/anthology/P18-1031.

Inui, Kentaro et al. (2003). "Text Simplification for Reading Assistance: A Project Note". In: *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*. PARAPHRASE '03. Sapporo, Japan: Association for Computational Linguistics, pp. 9–16. DOI: 10.3115/1118984.1118986. URL: https://doi.org/10.3115/1118984.1118986.

Jhamtani, Harsh et al. (Sept. 2017). "Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models". In: *Proceedings of the Workshop on Stylistic Variation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 10–19. DOI: 10.18653/v1/W17-4902. URL: https://www.aclweb.org/anthology/W17-4902.

Jonnalagadda, Siddhartha et al. (June 2009). "Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder, Colorado: Association for Computational Linguistics, pp. 177–180. URL: https://www.aclweb.org/anthology/N09-2045.

Kajiwara, Tomoyuki and Mamoru Komachi (Dec. 2016). "Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1147–1158. URL: https://www.aclweb.org/anthology/C16-1109.

Kandula, Sasikiran, Dorothy Curtis, and Qing Zeng-Treitler (Nov. 2010). "A Semantic and Syntactic Text Simplification Tool for Health Content". In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010, pp. 366–70.

Kauchak, David (2013). "Improving Text Simplification Language Modeling Using Unsimplified Text Data". In: *ACL*.

Kincaid, J. Peter et al. (1975). "Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel". In:

Knight, Kevin and Daniel Marcu (July 2002). "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". In: *Artificial Intelligence* 139, pp. 91–107. DOI: 10.1016/S0004-3702(02)00222-9.

Lample, Guillaume and Alexis Conneau (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Lample, Guillaume et al. (2018a). "Phrase-Based & Neural Unsupervised Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lample, Guillaume et al. (2018b). "Unsupervised machine translation using monolingual corpora only". In: *International Conference on Learning Representations (ICLR)*.

Margarido, Rafael et al. (Oct. 2008). "Automatic summarization for text simplification: Evaluating text understanding by poor readers". In: *Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. DOI: 10.1145/1809980.1810057.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation". In: *ArXiv* abs/1309.4168.

Miwa, Makoto et al. (2010). "Entity-focused Sentence Simplification for Relation Extraction". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, pp. 788–796. URL: http://dl.acm.org/citation.cfm?id=1873781.1873870.

Narayan, Shashi and Claire Gardent (June 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)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 435–445. DOI: 10.3115/v1/P14-1041. URL: https://www.aclweb.org/anthology/P14-1041.

Nisioi, Sergiu et al. (July 2017). "Exploring Neural Text Simplification Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 85–91. DOI: 10.18653/v1/P17-2014. URL: https://www.aclweb.org/anthology/P17-2014.

Paetzold, Gustavo H. and Lucia Specia (2016). "Unsupervised Lexical Simplification for Non-native Speakers". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, pp. 3761–3767. URL: http://dl.acm.org/citation.cfm?id=3016387.3016433.

Papineni, Kishore et al. (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.

Pavlick, Ellie and Chris Callison-Burch (Aug. 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)*. Berlin, Germany: Association for Computational Linguistics, pp. 143–148. DOI: 10.18653/v1/P16-2024. URL: https://www.aclweb.org/anthology/P16-2024.

Petersen, Sarah E. and Mari Ostendorf (2007). "Text simplification for language learners: a corpus analysis". In: *SLaTE*.

Ramachandran, Prajit, Peter Liu, and Quoc Le (Sept. 2017). "Unsupervised Pretraining for Sequence to Sequence Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 383–391. DOI: 10.18653/v1/D17-1039. URL: https://www.aclweb.org/anthology/D17-1039.

Rush, Alexander M., Sumit Chopra, and Jason Weston (Sept. 2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: https://www.aclweb.org/anthology/D15-1044.

Scarton, Carolina, Gustavo Paetzold, and Lucia Specia (May 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). URL: https://www.aclweb.org/anthology/L18-1553.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural Machine Translation of Rare Words with Subword Units". In: *ArXiv* abs/1508.07909.

— (Aug. 2016). "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)*. Berlin, Germany: Association for Computational Linguistics, pp. 86–96. DOI: 10.18653/v1/P16-1009. URL: https://www.aclweb.org/anthology/P16-1009.

Shen, Tianxiao et al. (2017). "Style Transfer from Non-parallel Text by Cross-alignment". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 6833–6844. ISBN: 978-1-5108-6096-4. URL: http://dl.acm.org/citation.cfm?id=3295222.3295427.

Siddharthan, Advaith (2006). "Syntactic Simplification and Text Cohesion". In: *Research on Language and Computation* 4.1, pp. 77–109. ISSN: 1572-8706. DOI: 10.1007/s11168-006-9011-1. URL: https://doi.org/10.1007/s11168-006-9011-1.

Siddharthan, Advaith and Napoleon Katsos (2010). "Reformulating Discourse Connectives for Non-expert Readers". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Los Angeles, California: Association for Computational Linguistics, pp. 1002–1010. ISBN: 1-932432-65-5. URL: http://dl.acm.org/citation.cfm?id=1857999.1858142.

Smith, Samuel L. et al. (2017). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". In: *CoRR* abs/1702.03859. arXiv: 1702.03859. URL: http://arxiv.org/abs/1702.03859.

Specia, Lucia (2010). "Translating from Complex to Simplified Sentences". In: *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*. PROPOR'10. Porto Alegre, RS, Brazil: Springer-Verlag, pp. 30–39. ISBN: 3-642-12319-8, 978-3-642-12319-1. DOI: 10.1007/978-3-642-12320-7_5. URL: http://dx.doi.org/10.1007/978-3-642-12320-7_5.

Štajner, Sanja and Maja Popovic (2016). "Can Text Simplification Help Machine Translation?" In: *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pp. 230–242. URL: https://www.aclweb.org/anthology/W16-3411.

Sulem, Elior, Omri Abend, and Ari Rappoport (2018a). "BLEU is Not Suitable for the Evaluation of Text Simplification". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 738–744. DOI: 10.18653/v1/D18-1081. URL: https://www.aclweb.org/anthology/D18-1081.

— (June 2018b). "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)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 685–696. DOI: 10.18653/v1/N18-1063. URL: https://www.aclweb.org/anthology/N18-1063.

Surya, Sai et al. (July 2019). "Unsupervised Neural Text Simplification". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2058–2068. DOI: 10.18653/v1/P19-1198. URL: https://www.aclweb.org/anthology/P19-1198.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vickrey, David and Daphne Koller (June 2008). "Sentence Simplification for Semantic Role Labeling". In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, pp. 344–352. URL: https://www.aclweb.org/anthology/P08-1040.

Wang, Alex et al. (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://www.aclweb.org/anthology/W18-5446.

Wang, Tong et al. (2016). "Text Simplification Using Neural Machine Translation". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, pp. 4270–7271. URL: http://dl.acm.org/citation.cfm?id=3016387.3016551.

Watanabe, Willian Massami et al. (2009). "Facilita: Reading Assistance for Low-literacy Readers". In: *Proceedings of the 27th ACM International Conference on Design of Communication*. SIGDOC '09. Bloomington, Indiana, USA: ACM, pp. 29–36. ISBN: 978-1-60558-559-8. DOI: 10.1145/1621995.1622002. URL: http://doi.acm.org/10.1145/1621995.1622002.

Woodsend, Kristian and Mirella Lapata (2011). "Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming". In: *EMNLP*.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer (July 2012). "Sentence Simplification by Monolingual Machine Translation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1015–1024. URL: https://www.aclweb.org/anthology/P12-1107.

Xing, Chao et al. (2015). "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation". In: *HLT-NAACL*.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles (2015). "Problems in Current Text Simplification Research: New Data Can Help". In: *Transactions of the Association for Computational Linguistics* 3, pp. 283–297. DOI: 10.1162/tacl_a_00139. URL: https://www.aclweb.org/anthology/Q15-1021.

Xu, Wei et al. (Dec. 2012). "Paraphrasing for Style". In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, pp. 2899–2914. URL: https://www.aclweb.org/anthology/C12-1177.

Xu, Wei et al. (2016). "Optimizing Statistical Machine Translation for Text Simplification". In: *Transactions of the Association for Computational Linguistics* 4, pp. 401–415. DOI: 10.1162/tacl_a_00107. URL: https://www.aclweb.org/anthology/Q16-1029.

Yatskar, Mark et al. (June 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*. Los Angeles, California: Association for Computational Linguistics, pp. 365–368. URL: https://www.aclweb.org/anthology/N10-1056.

Zhang, Xingxing and Mirella Lapata (Sept. 2017). "Sentence Simplification with Deep Reinforcement Learning". In: *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, pp. 584–594. DOI: 10.18653/v1/D17-1062. URL: https://www.aclweb.org/anthology/D17-1062.

Zhang, Zhirui et al. (2018). "Style Transfer as Unsupervised Machine Translation". In: *ArXiv* abs/1808.07894.

Zhao, Sanqiang et al. (2018). "Integrating Transformer and Paraphrase Rules for Sentence Simplification". In: *arXiv preprint arXiv:1810.11193*.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych (Aug. 2010). "A Monolingual Tree-based Translation Model for Sentence Simplification". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, pp. 1353–1361. URL: https://www.aclweb.org/anthology/C10-1152.