Understanding and Enhancing the Use of Context for Machine Translation

Marzieh Fadaee

Understanding and Enhancing the Use of Context for Machine Translation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op dinsdag 10 november 2020, te 12.00 uur

door

Marzieh Fadaee

geboren te Berlijn

Promotiecommissie

Promotor:		
	dr. C. Monz	Universiteit van Amsterdam
Co-promotor:		
	dr. A. Bisazza	Rijksuniversiteit Groningen
Overige leden:		
	dr. A. Birch	University of Edinburgh
	prof. dr. A.P.J. van den Bosch	Radboud Universiteit Nijmegen
	prof. dr. P.T. Groth	Universiteit van Amsterdam
	prof. dr. E. Kanoulas	Universiteit van Amsterdam
	dr. I. Markov	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Cover by Mostafa Dehghani, Painting “Series VIII: Picture of the Starting Point” by Hilma af Klint
Printed by Ipskamp printing, Amsterdam
ISBN: 978-94-6421-059-0

To those who see the world through words.

Acknowledgements

> "In the good mystery there is nothing wasted, no sentence, no word that is not significant. And even if it is not significant, it has the potential to be so - which amounts to the same thing." [1]

No word that is not significant; this is the main discussion of this book on the importance of context as well as an accurate description of how I (and this book) came to be. When I started my PhD the world was a different place. Then again, that can be said about every minute that goes by. What remains a constant is the impact of people around us, in the beginning of things and in the end. It has been an agonizingly fun ride! And I am very fortunate to have many amazing people around me that made it possible.

First, Christof. I came to Amsterdam with a promise and it turned out to be quite an experience. I owe my deepest gratitude to Christof for trusting me and taking me on this journey. Over the years he taught me how to research new areas, how to break down complicated questions to get to the heart of the problem, and how to never be happy with an answer and always push for more. I specially would like to thank him for his attentiveness when it counted. I enjoyed the walks, the talks, and making him exasperated with my wordplays.

I would like to thank Arianna for her guidance and for being a friend. Her critical thinking and questioning the things that I took for granted taught me how to explore novel corners of research and ask new questions. I will always be grateful to have her by my side, whether it was during discussions with Christof or the disappointment that was the Star Wars sequels.

Next, I would like to express my gratitude to Maarten for building a great research group at ILPS. From the first day of joining ILPS, I felt part of a team. I enjoyed the fun impromptu discussions, the group activities, and everything in between. Now that I've been away for some time, I see that what we had at ILPS was something truly special and Maarten is a big reason for that. Thanks for the guidance and for encouraging me to push my boundaries.

My sincere gratitude to Alexandra, Antal, Paul, Evangelos, and Ilya who were generous with their time to read this thesis and accepted to be part of my defense committee.

I would like to thank the MT members: Marlies, Ke, Katya, and Praveen, my brothers and sisters in arms. We shared laughs and tears, successes and failures, valuable and worthless discussions. And we all came out of it alive! What I got out of this PhD was more than I expected and I am glad to have met these people along the way. Thanks for making my PhD experience a fun and enjoyable one.

I am grateful for many current and former members at ILPS for our times together. From showing the ropes and sharing their wisdom, to asking intriguing questions, to fun sushi parties. Thanks Alexey, Amir, Anna, Artem, Boris, Chang, Chuan, Daan, Damien,

[1] Paul Auster, The New York Trilogy

Dan, David, David, Dilek, Eva, Evgeny, Fei, Harrie, Hendra, Hendrike, Hinda, Ivan, Jiahuan, Jie, Julia, Julien, Kaspar, Maarten, Maurits, Mostafa, Pengjie, Richard, Ridho, Rolf, Shangsong, Shaojie, Spyretta, Svitlana, Tobias, Trond, Vera, Xiaohui, Xiaojuan, Xinyi, Yangjun, Yifan, Zhaochun, and Ziming.

I particularly like to thank Anne, Ana, Bob, Christophe, Evangelos, Ilya, Isaac, Manos, and Tom, who are responsible for some of my good memories. Special thanks to Petra, who knows the answer to every question and is a delight to be around. I would also like to thank Mostafa for designing the cover and putting up with all of my typography suggestions. Cristina, thanks for being a ray of sunshine full of support and insight. We will always have Toronto.

Next, I want to thank my wonderful paranymphs: Maartje, my office-mate and friend who makes an excellent arrabiata pasta, and Nikos, my partner in crime, who makes every moment the opposite of boring. Thanks for agreeing to be by my side on my defense day.

I spent a summer in Aachen doing an internship at eBay which was an invaluable practice. Thanks, Shahram and Jose for the guidance and the interesting discussions, and Daniel, Leonard, Michael, Nicola, and Sivan for making it a nice experience. Parnia thanks for the talks, the hikes, and the Ash Reshtehs.

Moving to Amsterdam meant a fresh start and a new beginning. I was lucky to meet many amazing people that help me weather the gray days and savor the sunny ones. I am particularly thankful to Ali, Ali, Ameneh, Amin, Amirhosein, Arash, Aylar, Azad, Azadeh, Behrouz, Danial, Fahimeh, Fatemeh, Hester, Hoda, Hoda, Hojat, Hoora, Irene, Iris, Jasmijn, Jeyran, Mahdi, Mahdi, Mahdieh, Maria, Maryam, Marzieh, Marzieh, Masoud, Masoumeh, Mohammad, Mohammad Amin, Mohammad Hosein, Mohsen, Narges, Naser, Nasim, Nasrin, Parisa, Parisa, Parisa, Saba, Samira, Samira, Sara, Shayan, Shima, Vahid, Vivianne, Wilker, Zahra, and Zoheir. A particular thanks to a special group of people with whom I had many imaginative conversations: Ava, Behsa, Borna, NikAyeen, Pouya, Radin, Saba, Samin, Sarina, Sepehr, Sina, and Zeinab. They made me an optimist about the future.

My Algonquin Round Table, whom I am grateful to have in my life: Abbas, Ali, Arian, Atieh, Elaheh, Faezeh, Farzad, Ghazaleh, Louise, Mandana, Maryam, Mehdi, Safoora, Sara, Sara, Siavash. I met each of these people at different points in my life and something about each of them got stuck with me. Geographical determinism might be working against us now, but I cherish every time we got a chance to meet somewhere in this world.

My brothers Mohammad and Mohsen, thanks for doing what brothers do best: supporting me in times of need and humbling me the rest of the time. Fatemeh and Mahsa, I am thankful for the kind words and the support.

I am eternally grateful to my amazing parents. Baba, you showed me every day that there is always something new worth learning. Your dedication to diving blindly into new areas, languages, and technologies is always inspiring. You taught me to never stop fighting for what I want. Maman, you are the strongest woman I know and I appreciate

everything you have done for me. Every step I took has been a little bit easier because of the hard decisions you had to make. I followed my dreams because of you.

Lastly, Hamid. You are the kindest, smartest, most driven person I know. And yes I am still not a fan of using superlatives. It is truly earned in your case though. This would not have been possible without you. You were the best sounding board for everything that did (and did not) end up in this book. I am lucky to have your love, your support, and your wisdom by my side in this journey.

Marzieh
September 2020

> "Everything becomes essence; the center of the book shifts with each event that propels it forward. The center, then, is everywhere, and no circumference can be drawn until the book has come to its end." [2]

[2] Paul Auster, The New York Trilogy

Contents

— Il n'y a pas de hors-texte.

Jacques Derrida

1 Introduction

Neural networks learn patterns from data to solve complex problems. To understand and infer meaning in language, neural models have to learn complicated nuances.

Discovering distinctive linguistic phenomena from data is not an easy task. For instance, lexical ambiguity is a fundamental feature of language which is challenging to learn (Small et al., 2013). Even more prominently, inferring the meaning of rare and unseen lexical units is difficult with neural networks (Koehn and Knowles, 2017). For instance, Rios et al. (2018) provide an example where an English-German translation model translates the sentence "[. . .] *Hedge-Fund-* ***Anlagen*** *nicht zwangsläufig risikoreicher sind als traditionelle* ***Anlagen***" to "[. . .] *hedge fund assets are not necessarily more risky than traditional plants*". Here, the ambiguous word *'Anlagen'* is first translated correctly to *'assets'*, but then incorrectly to *'plants'* in the second occurrence.

To understand many of these phenomena, a model has to learn from a few instances and be able to generalize well to unseen cases. Natural language speakers typically learn the meanings of words by the *context* in which they are used. Miller (1985) states that:

> "*When subtle semantic distinctions are at issue, it is customary to remark that a satisfactory language understanding system will have to know a great deal more than the linguistic values of words.*"

Sentence and document-level context provide the possibility to go beyond lexical instances and study words in a broader context. Neural models use a sizable amount of data that often consists of contextual instances to learn patterns. However, the learning process is hindered when training data is scarce for a task (Kaiser et al., 2017, Edunov et al., 2018). Even with sufficient data, learning patterns for the long tail of the lexical distribution is challenging (Wang et al., 2017). To address these problems, one approach is to augment the training data (Sennrich et al., 2016b). Many strategies for data augmentation focus on increasing the amount of data to assist the learning process of data-driven neural models. While simply increasing the size of data is helpful, it is not entirely clear *where* the improvements come from and *how* neural models benefit

from the additional context with augmentation.

Arguably, it is important to understand the impact of new contexts to design augmentation models that exploit these contexts. This includes understanding what constitutes a beneficial context, and how to enhance the use of context in neural models. In this thesis, we focus on understanding certain potentials of contexts in a neural model, and design augmentation models to benefit from them.

We focus on machine translation as a prominent instance of the more general language understanding problems. In order to translate from a source language to a target language, a neural model has to understand the meaning of constituents in the provided context and generate constituents with the same meanings in the target language. This task accentuates the value of capturing nuances of language and the necessity of generalization from few observations (Li et al., 2020). Additionally, the lack of large amounts of labeled data is even more pronounced in machine translation in the form of bilingual corpora. This signifies the need for efficient and informed data augmentation models.

The main problem we study in this thesis is what neural machine translation models (NMT) learn from data, and how we can devise more focused contexts to enhance this learning. We believe that looking more in-depth into the role of context and the impact of data on learning models is essential to advance the Natural Language Processing (NLP) field. Understanding the importance of data in the learning process and how neural network models interact, utilize, and benefit from data can help develop more accurate NLP systems. Moreover, it helps highlight vulnerabilities and volatilities of current neural networks and provides insights into designing more robust models.

1.1 Research outline and questions

This thesis explores the role of context in language understanding and in particular, machine translation using recent advances in deep learning. We develop novel models and learning algorithms to examine the abilities of neural networks in learning from data. Specifically, we are interested in the importance of contextual cues in translating words and various ways we can use data to advance translation systems further.

Before investigating the role of context in the bilingual setting of machine translation, we ask ourselves a more general question about the impact of context in monolingual settings. As a preliminary investigation into this question, we look into ambiguous words where, by definition, context is the prominent factor in understanding word meaning. We study how document-level contexts as topics aid in distinguishing different meanings of a word.

Next, in more detail, we focus on the influence of context in the bilingual setting of machine translation. While recent advances in neural networks have been very successful in translation, the significance of different aspects of data is still largely unexplored. We investigate how the translation models exploit context to learn and

transfer meaning and show that manipulating data improves translation quality. In particular, our proposed models examine how different and diverse contexts resolve various obstacles of translation.

Lastly, we address the shortfalls of relying only on the observed context to learn word meaning and focus on particularly interesting cases. Neural networks optimize the learning process on the available data. We examine under which conditions the observed context in the training data is not enough for meaning inference and capturing various linguistic phenomena. Moreover, we raise questions about the learning abilities of current translation models and where they fail to capture the available information in data. With contextual modifications, we identify an underlying generalization problem in state-of-the-art translation models.

Concretely, we set out to answer the following research questions in this thesis:

Research Question 1: *Can document-level topic distribution help infer the meaning of a word?*

In this research question, we investigate whether using document-level context, as opposed to sentence-level only, has an impact on learning word representations. Word representations are abstract feature vectors that capture word meanings. To produce good representations, the learning model must capture various linguistic phenomena such as the ambiguity of the language. Notably, we tackle the problem of representing ambiguous words by defining multiple representations per word and using implicit topics of documents to distinguish between different meanings of a word. We divide this research question into three sub-questions and address them in Chapter 3:

RQ1.1 *To what extent can distributions over word senses be approximated by distributions over topics of documents without assuming these concepts to be identical?*

Modeling document topics is commonly used in different ways to address the challenging task of word sense disambiguation (Boyd-Graber et al., 2007, Li et al., 2010, Chaplot and Salakhutdinov, 2018). However, the topic of a document does not directly correspond to the senses of the words in that document. We investigate whether a document topic distribution is an informative signal to help distinguish between different senses of a word and how we can leverage this information to learn word representations. Next, we ask:

RQ1.2 *How can we exploit document-level topics to distinguish between different meanings of a word and learn the corresponding representations?*

To answer this question, we estimate document-topic distributions using unsupervised topic modeling techniques. We observe that the produced distribution over topics is different for different senses of an ambiguous word. We propose three variants of the Skipgram word embedding model (Mikolov et al., 2013a) to integrate topic distributions and learn multiple representations per word.

RQ1.3 *What are the advantages of using document-level topics in learning multiple representations per word?*

To further evaluate our models, we analyze the linguistic phenomena captured by topic-sensitive word representations. Namely, we show that different senses of a word are separated into different representations. We observe that the additional context of a document topic is most beneficial when the task is more complex. We find that these representations achieve improvements over the baselines for word similarity and lexical substitution tasks.

Having examined the effectiveness of learning word representations using auxiliary contextual information, we then investigate how the diversity of the context affects language understanding and transfer of meaning between two languages. Concretely we ask:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

In this research question, we choose machine translation as the task of interest. We investigate this question by diversifying the local context for different words and propose various data augmentation techniques with the new contexts. Additionally, we explore the influence of these synthetic contexts on translation quality. We divide this research question into four sub-questions and discuss them in Chapters 4 and 5 of this thesis.

RQ2.1 *How can we successfully augment the training data with diverse contexts for rare words?*

In this question, we are interested in translation of rare words in low-resource settings where the available data is scarce for one or both languages. The success of neural networks is partly due to their ability to learn from vast amounts of data efficiently. These models suffer significantly when sufficient data is not available (Ngo et al., 2019). Subsequently, even with adequate data, neural machine translation models have difficulty learning the meaning of rare words existing in the source language (Koehn and Knowles, 2017). Additionally, they are also not successful in generating rare words in the target language (Luong et al., 2015b). To answer this question, in Chapter 4, we propose a data augmentation technique that targets rare words and substitute them in new sentences with novel contexts. Leveraging a monolingual corpus, which is available in much larger quantities in comparison to a bilingual corpus, we create new contexts for rare words in the training data. We investigate how additional data can improve the learning and the generation of rare words. In Chapter 4, we show that by increasing the diversity of the contexts of rare words, we can achieve significant improvements in translation quality.

RQ2.2 *Do rare words benefit from augmentation via paraphrasing during test time?*

Diversifying context is only valid when both source and target sentences are modified, i.e., at training time when the model has access to the sentence pairs. During inference, only the source sentence is available and we use the reference sentence solely for evaluation. As a consequence, any changes to the source sentence have to be meaning-preserving so that we do not modify the reference translations. We propose a data augmentation technique at test time, focusing on *paraphrasing* rare and unknown words in the source sentence. In contrast to our previous approach where the goal was to diversify the context of rare words in the training data, here we substitute rare words with more common synonyms. In Chapter 4, we show that with paraphrasing rare words at test time, we gain improvements in translation quality.

RQ2.3 *Do signals from the NMT model help identify low-confidence words that could benefit from additional context?*

In the previous research questions, we identify rare words as words that can benefit from additional contexts. While the translation quality of these words improves with our proposed data augmentation technique, these are not the only words that suffer due to inadequacies in the training data. In Chapter 5, we expand our investigation in this direction. Rather than using features like frequency in the training data, we look into the *model* itself and where it struggles. We detect the words for which the model has low confidence during translation. We examine various approaches to identify these low-confidence words as signaled by the model and augment the training data accordingly. Hence, we ask:

RQ2.4 *How can we successfully apply data selection of monolingual data to diversify the contexts of low-confidence words?*

To generate new contexts and augment the training data, we propose targeted back-translation. Back-translation leverages monolingual data in the target language and a trained translation model to translate randomly selected sentences into the source language (Sennrich et al., 2016b). The automatically generated bilingual data, although noisy, is added to the training data and the translation model is trained on the augmented data. In Chapter 5, we identify words that can benefit from diverse context. We show that by back-translating sentences containing low-confidence words, we achieve improvements over the baselines.

Having demonstrated the advantages of using contextual cues in various forms to improve word representation learning and translation modeling, we come to the final research question of this thesis. Here, we investigate the shortcomings of relying on the observed context. Concretely we ask:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

While the success of neural networks in NLP is indisputable, it is well worth to ask whether neural networks have hidden vulnerabilities. In this research question, we also choose machine translation as the task of interest. We are interested in vulnerabilities of the translation models that can be exposed by looking into the data. In particular, we divide this research question into four sub-questions and discuss them in Chapters 6 and 7 of this thesis:

RQ3.1 *What are the challenges of idiom translation with neural models?*

Neural translation models struggle in handling idiosyncratic linguistic patterns. One of these patterns are idioms, which are semantic lexical units whose meaning is not merely a function of the meaning of its constituent parts. In Chapter 6, we look into idiomatic expressions in particular and why the translation of such phrases is a challenge. Furthermore, we automatically label parallel training and test data for idiomatic expressions using a bilingual dictionary of idioms. We assess whether the sentential context is enough for inferring idiomatic meanings and show that it is indeed not the case.

Next, we ask:

RQ3.2 *How is the translation quality of NMT influenced by idiomatic expressions?*

There is no explicit indicator in the data to signal whether a phrase should be translated literally or idiomatically in any given context. Researchers have shown that neural models can benefit from side constraints in data in various cases. For instance, Sennrich et al. (2016a) note that adding side constraints as unique tokens at the end of the source text help the model translate to the desired level of politeness. In Chapter 6, we investigate whether a similar technique is useful for the translation of sentences containing idiomatic expressions.

Finally, we look into other vulnerabilities of neural models which can be highlighted by contextual cues. Our next research question focuses on other cases where NMT models fail to generate a correct translation. To investigate this question, we first examined how to expose this shortcoming in translation models by asking:

RQ3.3 *How can contextual modifications during testing reveal a lack of robustness of translation models and affect the translation quality?*

In Chapter 7, we ask how receptive the translation models are to manipulations of data. While previous works have investigated the performance of neural models when encountering noise in the form of adversarial instances (Goodfellow et al.,

2015, Michel and Neubig, 2018, Belinkov and Bisk, 2018), we are interested in unexpected performance when the data is *not* noisy.
Next, we investigate the robustness of neural translation models by asking:

RQ3.4 *To what extent is a lack of robustness an indicator of a generalization problem in neural machine translation models?*

We propose an approach to generate contextual modifications in the test data, yielding semantically and syntactically correct sentences. Our new test data sheds light on volatile behaviour in current state-of-the-art translation models. In Chapter 7, we show that identifying this volatility is already achievable with extremely minor modifications. Our findings highlight unexpected but recurring patterns of errors and possible problems of generalization in neural translation models.

1.2 Main contributions

Here we summarize the main algorithmic and empirical contributions of this thesis to the field of natural language processing and in particular machine translation, as well as the constructed resources.

1.2.1 Algorithmic contributions

We develop novel learning algorithms and neural network models for investigating the influence of context in learning capacities of models.

1. We present a framework for learning multiple embeddings per word using topical context. With three variants of our model, we employ topical context in various ways and learn distinctions between different senses of the words (Chapter 3).

2. We introduce a data augmentation technique for generating new contexts for rare words in machine translation. Leveraging monolingual data, we propose a neural language model that given a sentence, suggests rare words to substitute into the given context. This new method can be applied to any low-resource language pair as long as there are monolingual data available in both languages (Chapter 4).

3. We introduce a novel method to identify difficult words, where the neural translation model has low prediction confidence. Leveraging this information, we improve upon an existing augmentation technique by replacing its random selection with targeted selection and specifically provide new contexts for low-confidence words (Chapter 5).

4. We propose a procedure to (i) automatically detect idiomatic expressions in sentences using a dictionary of idioms, and (ii) automatically annotate the bilingual data with the corresponding idioms (Chapter 6).

5. We introduce an effective technique to shed light on the lack of robustness of neural translation models. Our approach generates variants of the same sentences that differ slightly and are semantically and syntactically correct. We investigate the behaviour of the neural model in translating these variants by proposing metrics to identify volatile performance (Chapter 7).

1.2.2 Empirical contributions

We evaluate our proposed models on large scale data sets as well as controled experiments to validate our hypotheses. We provide empirical results for each research question asked in this thesis. More specifically:

1. We compare how different approaches of incorporating topical context affect the resulting representations. We assess the topic-sensitive word representations on word similarity and lexical substitution tasks and perform a qualitative analysis between different representations of a word (Chapter 3).

2. We evaluate the effectiveness of our first data augmentation approach in machine translation for two language directions: English→German and German→English. We simulate a low-resource setting by only using a subset of the available training data, while simultaneously being able to compute the upper bound of performance in case more data is available. Our approach successfully mitigates the problem of rare word translation, where sufficient bilingual training data is not available. We perform an analysis of the confidence of the translation model for both generating and translating rare words (Chapter 4).

3. We evaluate our second proposed data augmentation approach in machine translation for two language directions: English→German and German→English. We study the effects of previous data augmentation techniques on confidence and the learning capacity of the translation model. We compare various ways of identifying low-confidence words and show that targeted data augmentation using these words improves translation quality. We demonstrate that with diversifying contexts of difficult words, the confidence of the model in predicting these words and consequently the translation quality improve (Chapter 5).

4. We conduct an empirical evaluation of translation models facing sentences that include an idiomatic expression. Using annotated training and test data, we demonstrate how the current neural translation models struggle with translating idioms. We show that even when we annotate them in the training data, translating these expressions is a challenge and the translation models require much broader knowledge to learn them (Chapter 6).

5. We show that fluctuations in translations of extremely similar sentences are more prominent than expected. These findings can be used to develop more robust models (Chapter 7).

1.2.3 Resource contributions

We release the resources of the proposed models in this thesis including source codes and annotated data. More specifically:

1. Chapter 3: We released the code for the proposed models where we use document topics to learn word representations.

2. Chapter 4: We released the code for targeted data augmentation of parallel corpora using language models.

3. Chapter 6: We released the annotations of idiomatic phrases in training, development, and test data. The bilingual corpora can be used for translation of English→German and German→English.

4. Chapter 7: We released a data set which contains multiple variants for each sentence pair in the standard WMT English↔German test data. We annotate the translations of these variants and label different types of errors. Additionally, we release the code for generating sentence variations of bilingual corpora for a more in-depth evaluation of translation quality.

1.3 Thesis overview

After this introductory chapter, the remainder of this thesis consists of a background chapter (Chapter 2), five research chapters (Chapters 3-7), and a concluding chapter (Chapter 8). Below we present a high-level overview of the main content of each of these chapters.

Chapter 2: Background provides an introduction to the neural machine translation (NMT) paradigm used in this thesis. We briefly review the core models, the training and test data required, and the learning and optimization strategies we employ. We also discuss different representation learning approaches. Additionally, we describe the basic experimental settings for our systems. Finally, we provide an overview of evaluation metrics used in this thesis.

Chapter 3: Representation learning using documental context introduces the concept of learning multiple representations per word to capture lexical ambiguity in a language. We first investigate the influence of document topics on distinguishing different meanings of a word, then propose various models to integrate topical information in representation learning, and finally analyze the performance of these contextual representations and compare them to single representations. Our findings in in this chapter provide answers to **RQ1**.

Chapter 4: Data augmentation for rare words focuses on the impact of additional context in influencing translation quality of rare words. Notably, we use language models to substitute rare words in existing bilingual contexts. We augment the translation model with the newly generated data and as a result, improve both the generation frequency and the translation quality of rare words. Our results in this chapter provide answers to **RQ2.1** and **RQ2.2** .

Chapter 5: Data augmentation based on model failure examines the influence of augmenting data with diverse context for difficult words on translation models. We first inspect the learning process of state-of-the-art translation models and identify where they are not confident in their predictions. After further analyzing the words that translation models have difficulties in learning, we introduce an augmentation approach to target these words. We improve upon an existing data augmentation approach by devising new contexts for low-confidence words. Our results in this chapter provide an answer to **RQ2.3** and **RQ2.4** .

Chapter 6: Analyzing idiomatic expressions investigates translation errors prevalent in current models. First, we identify multiword expressions that are syntactically or semantically idiosyncratic and challenging to translate. Next, we create a parallel corpus consisting of sentence pairs with idiomatic expressions. For this study, we introduce new error analysis measures to evaluate the translation quality of these expressions individually. We provide empirical answers to **RQ3.1** and **RQ3.2** in this chapter.

Chapter 7: Analyzing volatility investigates the robustness of state-of-the-art translation models to variants in source sentences. We propose an effective technique to generate modifications in test sentences while avoiding the introduction of semantic or syntactic noise. Investigating the translation outputs of different models on the modified test corpus reveals the extent of volatility that exists in translation models. We perform an analysis of robustness of our models to answer **RQ3.3** and **RQ3.4** .

Chapter 8: Conclusion concludes this thesis by revisiting the research questions and their corresponding answers. We also reflect on future research directions and on what the community can learn from the findings in this thesis.

1.4 Origins

The research presented in Chapters 3-7 of this thesis is based on a number of peer-reviewed publications. Below, we indicate the origins of each chapter.

Chapter 3 is based on Marzieh Fadaee and Arianna Bisazza and Christof Monz, "Learning Topic-Sensitive Word Representations", *In Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (ACL), (Fadaee et al., 2017b). Fadaee designed and carried out the experiments. All authors contributed to the discussion and text.

Chapter 4 is based on Marzieh Fadaee and Arianna Bisazza and Christof Monz, "Data Augmentation for Low-Resource Neural Machine Translation", *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, (Fadaee et al., 2017a). Fadaee designed the methods, performed the experiments and wrote most of the text. Bisazza and Monz contributed to the discussion and editing.

Chapter 5 is based on Marzieh Fadaee and Christof Monz, "Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation", *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Fadaee and Monz, 2018). Fadaee designed the methods, performed the experiments and wrote most of the text. Monz contributed to the discussion and editing.

Chapter 6 is based on Marzieh Fadaee and Arianna Bisazza and Christof Monz, "Examining the Tip of the Iceberg: A Data Set for Idiom Translation", *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, (Fadaee et al., 2018). Fadaee designed the methods, performed the experiments, and wrote the text. Bisazza and Monz contributed to the discussion and editing.

Chapter 7 is based on Marzieh Fadaee and Christof Monz, "The Unreasonable Volatility of Neural Machine Translation", *In Proceedings of the 4th Workshop on Neural Generation and Translation (WNGT)*, (Fadaee and Monz, 2020). Fadaee designed the methods, performed the experiments, and wrote the text. Monz contributed to the discussion and editing.

2 Background

Neural machine translation (NMT) is an end-to-end learning approach to machine translation that is based on neural networks. In contrast to traditional translation systems such as phrase-based machine translation (PBMT) (Koehn et al., 2003), all components of the neural translation model are trained jointly to maximize translation performance. In this chapter, we discuss the NMT paradigm and the properties of building a translation model.

The training data, in the format of parallel data, is a fundamental part of building NMT models. We first explain the training data used in the NMT paradigm in Section 2.1, followed by an overview of data preparation and building the translation vocabulary in Section 2.2. Next, we discuss different word representation models in Section 2.3. In the following sections, we review the two main NMT frameworks used in this thesis: recurrent neural networks (Section 2.4) and the transformer model (Section 2.5). Both models are classes of artificial neural networks and use large amounts of parallel data to learn a translation model. Finally, in Section 2.6, we describe the evaluation approaches used in the later chapters of this thesis.

2.1 Parallel and monolingual corpora

Neural models, and specifically neural translation models, rely heavily on training data. The primary training data for learning translation models are parallel corpora, which are aligned texts in two or more languages. These corpora are often paired at the sentence-level, ideally providing an exact translation of every sentence in the source and target language.

The quality of the translation system depends on the quality and the size of the training data. Acquiring good-quality parallel corpora requires manual translation by professional translators and as a result is expensive. Examples of available parallel corpora gathered by experts in the domain include Europarl (Koehn, 2005), which is the proceedings of the European Parliament published on the web, and JRC-Acquis (Steinberger et al., 2006), which is the total body of the European Union law applicable

in the EU Member States. Callison-Burch et al. (2007) gathered News Commentary corpora which consist of political and economic commentary crawled from the web site Project Syndicate. This data is extracted every year for the translation task of the WMT conference (Barrault et al., 2019).

Monolingual data, in comparison, are available in abundance for many languages. Phrase-based machine translation models use monolingual corpora in the target language (Koehn et al., 2003, Brants et al., 2007, Koehn et al., 2007) to improve the fluency of the generated translation (Lembersky et al., 2011). Monolingual parallel corpora of aligned complex-simple sentences are also used with pharse-based (Wubben et al., 2012, Kajiwara and Komachi, 2016) and neural (Zhang and Lapata, 2017) models to learn to simplify text. The monolingual News Crawl corpus from WMT and many available corpora in the Linguistic Data Consortium (LDC)[1] are examples of commonly utilized data in machine translation.

Vanilla NMT models typically do not use any monolingual data in their training. In Chapters 4 and 5 of this thesis, we address this matter by focusing on the use of monolingual data for NMT. Recently there have been studies that propose various approaches for incorporating information from monolingual data in the models (Domhan and Hieber, 2017, Burlot and Yvon, 2018, Currey and Heafield, 2019). Currey et al. (2017) created a parallel corpus from monolingual data in the target language by copying it so that each source sentence is identical to its corresponding target sentence. With this simple technique, they observe improvements on relatively low-resource language pairs. Another category of approaches is to translate sentences from monolingual data and augment the bitext with the resulting pseudo parallel corpora. This category of approaches is discussed in the following section.

2.1.1 Back-Translation in machine translation

In this section, we introduce the conventional method of generating synthetic data, namely back-translation and its effectiveness in PBMT and NMT. Back-translation uses an intermediate MT model, trained on parallel data, to translate target monolingual data into the source language. The result of back-translation is a parallel corpus where the source side is synthetic MT output while the target is actual text written by humans.

This technique is not bounded to neural networks, and prior to NMT models, it has been used in combination with PBMT. Schwenk (2008) proposes to translate large amounts of monolingual data with a PBMT system and use those as additional training data. They observe that this lightly-supervised training achieves improvements in translation quality. Rapp (2009) introduces the back-translation score as an alternative mean for the evaluation of PBMT models. He trains a translation model in both directions and evaluates the quality of the model by translating the target sentences back to the source language. The score is therefore computed by comparing the back-translated sentence to the original source sentence. As part of their experiments,

[1] https://www.ldc.upenn.edu/

Tiedemann et al. (2016) note that back-translating sentences from monolingual news data and augmenting the parallel training data improves the translation quality of a PBMT system. In these experiments, the models have to be re-tuned from scratch with the additional synthetic data.

In the framework of NMT, Sennrich et al. (2016b) show that back-translating sentences from monolingual data improves the performance of NMT models. This approach of augmenting the training data has since become common practice in training NMT models. Pham et al. (2017) experimented with using domain adaptation methods to select monolingual data for back-translation based on the cross-entropy between monolingual data and the in-domain corpus (Axelrod et al., 2015), but did not find any improvements over random sampling as initially proposed by Sennrich et al. (2016b).

Edunov et al. (2018) investigate back-translation in NMT at a large scale by adding hundreds of millions of back-translated sentences to the bitext. They study different methods for generating synthetic sentences and show that synthetic data based on sampling and noised beam search provides a stronger training signal than using pure beam. They observe that the generated corpora tend to stray away from the distribution of natural data. Brants et al. (2007) suggest a distributed language model infrastructure, which allows direct integration into the hypothesis-search algorithm. They observe that translation quality continues to improve with increasing language model size. Ueffing et al. (2007) use an iterative procedure that translates the monolingual source language data in each iteration and then re-trains the phrased-based translation model. They conclude that when bilingual training data are scarce, a PBMT system could be trained on a small amount of data and then iteratively improved by adding reliable translations of monolingual data to the training data.

He et al. (2016a) observe that any machine translation task has a dual task, for instance, English→French translation (primal) versus French→English translation (dual). They propose an approach based on reinforcement learning, where two agents, representing the primal and dual task, teach each other. The agents leverage monolingual data by translating it forward to the other language and then translate backward to the original language. Gulcehre et al. (2017) propose two methods, shallow and deep fusion, for integrating a neural language model into an NMT system. They observe improvements by combining the scores of a neural language model trained on target monolingual data with an NMT system.

2.2 Translation vocabulary

In translation models, the vocabulary of the source and target language is defined as what the model is exposed to during training. Word-level translation models are unable to translate or generate unseen words at inference. The number of words in the vocabulary can be remarkably large and training models on large vocabularies is computationally expensive.

An early practice was to limit the vocabulary to the K most frequent words, where K is often in the range of 30k (Bahdanau et al., 2015) to 80k (Sutskever et al., 2014). The tail of the vocabulary not included in this shortlist is mapped to a special token `[unk]` representing an unknown or out-of-vocabulary word. This method results in neural models that can be trained and tested within a reasonable amount of time, however, as a consequence of this simplification, the translation quality of the model suffers. Specifically, the performance decreases significantly when the translation of a source sentence requires many unknown words (Cho et al., 2014).

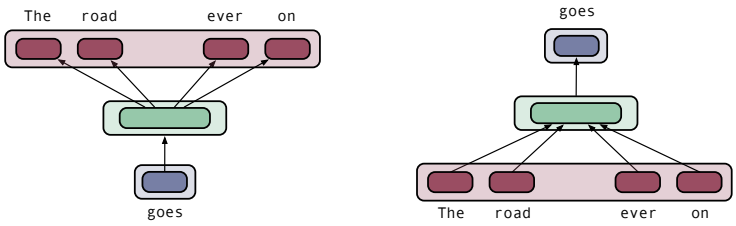
To address this issue, Jean et al. (2015) proposed an approximate training algorithm that can use a very large target vocabulary (vocabularies of 500,000 source and target words). They show that decoding the target sentence by sampling only a small subset of the whole vocabulary achieves competitive results without sacrificing too much speed. Luong et al. (2015b) proposed a copy mechanism that aligns the OOV words on both the source and the target side by learning to copy indices. Sennrich et al. (2016c) analyzed NMT models that work with subword units and observed that the majority of tokens are potentially translatable through smaller units. They modify Byte Pair Encoding (BPE) (Gage, 1994) to segment words into subword units, where each of which should be frequently observed in the corpus. While some segmentations correspond to correct morphemes, for many words that is not the case. For instance, the word *'quixotism'* would be segmented into *'quixot + -ism'* and the word *'sceptical'* would be segmented into *'scep + -tic + -al'*. This approach is very effective in generalization and is able to generate words not seen during training using these subword units.

In this thesis, we segment words during preprocessing using the BPE technique in all translation experiments unless stated otherwise. We refer to the subword units throughout the chapters as *tokens*.

2.3 Word representations

The first step in using neural models for text is to map the words in the vocabulary to *dense vectors* of real numbers. These vectors are somewhat similar to *sparse vectors* used in distributional semantics, where they represent meaning by capturing similarities between lexical units based on their distributional properties (Baroni et al., 2014, Baroni and Lenci, 2010). The context of the lexical unit is commonly used for the computation of dense and sparse embeddings. The intuition is that since similar words appear in similar contexts, they end up with similar embeddings (Firth, 1957). Computation of dense vectors is often a by-product of solving a natural language processing task such as language modeling or translation.

Word representations can be categorized into two groups (Wang et al., 2019b): *static embeddings* where a fixed vector is learned for each word in the vocabulary, and *dynamic embeddings* where vectors are dynamically calculated for each sentence. In the next sections, we discuss different approaches in each category.



(a) Skipgram model **(b)** CBOW model

Figure 2.1: Representation learning architectures proposed by Mikolov et al. (2013a). The CBOW model predicts the current word given the context, and the Skipgram model predicts the surrounding words given the current word

2.3.1 Static embeddings

Traditional word embedding techniques learn a global and static word embedding for every word in the vocabulary. Mikolov et al. (2013a) proposed two models for learning word representations: continuous Skipgram and continuous bag-of-words (CBOW). Both models use a feed-forward neural network architecture with the objective of modeling language, illustrated in Figure 2.1. This architecture does not include any non-linearity. The CBOW model has a projection layer which is shared between all words. This layer averages the input vectors. Next, using a classifier, the model predicts the word w_t given the context of words surrounding w_t in a fixed sized window: $[w_{t-c}, \ldots, w_{t+c}]$. The objective of the CBOW model is to maximize the following average log probability:

$$\frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}) \tag{2.1}$$

where T is the length of the sequence of training words and c is the context window size. The Skipgram model is similar to CBOW, but instead of predicting w_t, the model predicts the words within a fixed range surrounding w_t. The objective of the Skipgram model is to maximize the following average log probability:

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{\substack{-c \leqslant j \leqslant c \\ j \neq 0}} \log p(w_{t+j} \mid w_t) \tag{2.2}$$

where T is the length of the sequence of training words. Note that in both CBOW and Skipgram models, the context window includes both the past and the future.

Pennington et al. (2014) combined count-based matrix factorization and context-based Skipgram model together. The intuition is that meaning of words can be captured by the ratios of co-occurrence probabilities. They proposed a weighted least squares

model that trains on global word co-occurrence counts. They showed that the vector space learned from this model captures meaningful vector space substructures. While some syntactic and semantic features in language are captured by these word embeddings (Mikolov et al., 2013c), the dimensions are often not interpretable.

These models utilize surrounding words as context. However, word representations can capture different phenomena if the definition of context is different. Levy and Goldberg (2014) proposed to use dependency-based contexts, extracted from dependency parse-trees. They observed that these embeddings are less topical and exhibit more functional similarity than the original Skipgram embeddings.

Static embeddings for the most part learn a static matrix of embeddings for each word type and ignore capturing some nuances of language such as ambiguity. In Chapter 3, we address this issue by exploring document topics and learning multiple embeddings per word type to capture polysemy.

2.3.2 Dynamic embeddings

Static models generate out-of-context embeddings for word types and are simple and efficient to train and use. However, learning meaningful word representations has recently been elevated beyond this paradigm. Rather than learning static representations for word types, these models learn *dynamic* vectors for word instances in context using language modeling objectives. We denote this kind of embeddings as dynamic because instead of a static matrix of embeddings, they are obtained through the hidden states of a language model given the context.

Peters et al. (2018) proposed to use a bidirectional recurrent neural network to extract context-dependent representations. The learning objective is to predict the next word in a sequence, given the previous context words. Devlin et al. (2019) use a transformer architecture and define two new objectives for training: *masked language modeling*, and *next sentence prediction*. During masked language modeling, they mask a randomly selected word in a sentence, and the model has to predict that word given the context. The second objective gets two input sentences and predicts whether the second sentence is indeed the next sentence. The contextualized word embeddings are successful at downstream NLP tasks such as question answering and textual entailment (Zhang et al., 2019, Garg et al., 2020, Lan et al., 2020, Joshi et al., 2020).

While word vectors in neural translation models can be initialized with these static or dynamic word representations, they are often initialized randomly (Wu et al., 2016). One reason can be that with large-scale parallel data, these initial word representations will be forgotten during the training of the NMT model. Qi et al. (2018) showed that for low-resource language pairs and when languages are more similar, pre-trained embeddings can be effective. Lewis et al. (2020) recently proposed a denoising autoencoder model named BART for pretraining sequence-to-sequence models. They corrupt text with an arbitrary noising function and learn a model to reconstruct the original text. BART is effective when fine-tuned for text generation and translation but also works well for

comprehension tasks. The research presented in this thesis mostly predates the work mentioned in this section. We use *static* embeddings in Chapter 3 where we investigate the effect of having more than one representation per word type. As for the chapters on machine translation, we consider the most widespread setup where embeddings are initialized randomly before training on the parallel data.

2.4 Recurrent translation models

In this section, we discuss a category of neural models that are effective in modeling languages. Earlier developments of neural models addressing language modeling tasks incorporated the temporal structure of the language in the structure of the network (Jacquemin, 1994, Schmidhuber, 1993). A recurrent neural network (RNN) is an example of these sequential models (Rumelhart et al., 1988). RNNs are powerful models that achieve state-of-the-art results in a variety of tasks such as question answering (Garg et al., 2020), reading comprehension (Zhang et al., 2020), image semantic segmentation (Yuan et al., 2019), and speech recognition (Xiong et al., 2017).

RNN models are capable of modeling sequences of text with various length, while selectively passing on information between different time steps in the sequence. A long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) is an RNN structure that uses special LSTM units in addition to standard ones. These special units include a memory cell that can maintain information for long sequences. LSTM models address the *vanishing gradient problem* in the earlier RNN architecture where the weights and biases of the hidden layers are not updated effectively because the gradient decreases exponentially (Hochreiter, 1998).

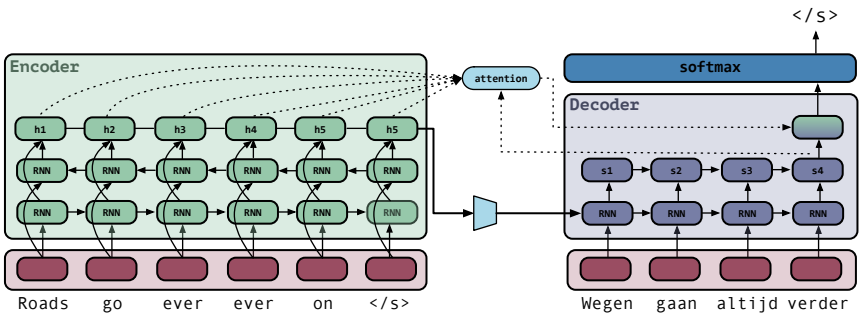


Figure 2.2: An illustration of an RNN encoder-decoder with attention.

Sutskever et al. (2014) and Cho et al. (2014) were among the firsts to employ RNNs to build an end-to-end machine translation model. Bahdanau et al. (2015) and Luong et al. (2015a) introduced an *attention mechanism* that achieved performance on par with traditional statistical models. In the following sections, we describe the RNN architecture with attention used in the NMT experiments in this thesis.

2.4.1 Model architecture

Neural machine translation models fall under a sequence-to-sequence framework where an encoder builds up a representation of the source sentence and a decoder generates the target translation. Both the encoder and the decoder can be recurrent neural models. Figure 2.2 illustrates this architecture which we will discuss in detail in this section. In order to train an NMT system, two sequences of tokens, $X = \big[x_1, \ldots, x_n\big]$ and $Y = \big[y_1, \ldots, y_m\big]$, are given in the source and target language, respectively. As discussed in Section 2.3, the input tokens are mapped to an embedding space. As a result, the source sequence is the input to the encoder as vectors: $\big[\mathbf{x}_1, \ldots, \mathbf{x}_n\big]$.

The encoder then encodes the input sequence into hidden states, where at time step t the hidden state is a function of the current input vector and the previous hidden state:

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}) \tag{2.3}$$

Function f adds non-linearities to the transformation of the input sequence to the output of the encoder. With a bidirectional architecture, two RNNs are run on the input sequence: one in forward and one in backward direction. The hidden state at time t is created by concatenating the forward and backward hidden states at each point in time, the input has access to the information on both sides. Note that the forward and backward hidden states are concatenated to create the top hidden states of the encoder, $\mathbf{h}_t$ as follows:

$$\mathbf{h}_t = \big[\overrightarrow{\mathbf{h}_t}^\intercal; \overleftarrow{\mathbf{h}_t}^\intercal\big]^\intercal, t = 1, \ldots, n \tag{2.4}$$

The decoder then generates the target translation one word at a time starting with the last hidden state of the encoder and the representation for the start-of-sentence symbol <s>. Each decoder hidden state $\mathbf{s}_t$ is computed as:

$$\mathbf{s}_t = g(\mathbf{s}_{t-1}, \mathbf{y}_{t-1}, \mathbf{c}_t) \tag{2.5}$$

where g is a transformation function that outputs a vocabulary-sized vector and $\mathbf{y}_{t-1}$ is the representation of the previously predicted token. $\mathbf{c}_i$ is the context vector for output at position i and is defined as:

$$\mathbf{c}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{h}_j \tag{2.6}$$

This context vector is recomputed at each time step. α_{ij} is the attention weight and it is computed for all source words at each time step i. We will discuss different approaches to computing attention weights in Section 2.4.3.

Next, the decoder predicts each target token y_t by computing the conditional proba-

bility:

$$p(y_t \mid y_{<t}, X) = \text{softmax}\left(\mathbf{s}_t\right) \tag{2.7}$$

This conditional probability is computed over the vocabulary of the target language which is fixed during training and testing. For token y_t, the conditional probability $p(y_t \mid y_{<t}, X)$ during training quantifies the difficulty of predicting that token in the context $y_{<t}$. The prediction loss of token y_t is the negative log-likelihood of this probability. During training on a parallel corpus $\mathbb{D}$, the cross-entropy objective function is defined as:

$$\mathcal{L} = \sum_{(X,Y)\in\mathbb{D}} \sum_{i=1}^{m} -\log p(y_i \mid y_{<i}, X) \tag{2.8}$$

The objective of this function is to improve the model's estimation of predicting target words given the source sentence and the target context. The model is trained end-to-end by minimizing the negative log-likelihood of the target words using stochastic gradient descent.

NMT systems often benefit from multiple layers of stacked RNNs during training (Wu et al., 2016). By increasing the number of parameters, the learning capability of the model also increases (Britz et al., 2017). Belinkov et al. (2017) show that different layers in the encoder capture different linguistic features, namely that higher layers capture semantics while lower layers tend to capture syntax. Encoding the input sequence in both directions also provides advantages (Luong et al., 2015a, Bahdanau et al., 2015). The backward layer in a recurrent model learns more about the semantics of words, whereas the forward layer encodes more of the local context (Ghader and Monz, 2019).

2.4.2 Inference

During inference, a trained model is given a source sentence and it generates the target translation word by word using a left-to-right beam search technique (Jelinek, 1998) This procedure was already adopted by pre-neural translation methods such as phrase-based translation models (Koehn et al., 2003). Generation of target words stops when a special end-of-sentence symbol $<\!/\mathrm{s}\!>$ is generated. At each step, the model computes a probability distribution over all words in the target language and chooses the most likely word:

$$\hat{Y} = \arg\max_{Y} p(Y \mid X) \tag{2.9}$$

Sutskever et al. (2014) showed that increasing the beam size beyond 2 does not improve the predictions significantly and even with a beam size of 1, the model performs well. With a large enough beam size, the best translation performance can be reached with the drawback of efficiency (Freitag and Al-Onaizan, 2017). It is common practice

Name	Proposed by	Alignment score
Additive	Bahdanau et al. (2015)	$\text{score}(\mathbf{s}_i, \mathbf{h}_t) = \mathbf{v}_a^\intercal \tanh(\mathbf{W}_a[\mathbf{s}_i, \mathbf{h}_t])$
Location-base	Luong et al. (2015a)	$\alpha_{n,t} = \text{softmax}(\mathbf{W}_a \mathbf{s}_i)$
General	Luong et al. (2015a)	$\text{score}(\mathbf{s}_i, \mathbf{h}_t) = \mathbf{s}_i^\intercal \mathbf{W}_a \mathbf{h}_t$
Dot-product	Luong et al. (2015a)	$\text{score}(\mathbf{s}_i, \mathbf{h}_t) = \mathbf{s}_i^\intercal \mathbf{h}_t$
Scaled dot-product	Vaswani et al. (2017)	$\text{score}(\mathbf{s}_i, \mathbf{h}_t) = \frac{\mathbf{s}_i^\intercal \mathbf{h}_t}{\sqrt{n}}$

Table 2.1: Different alignment scores in the literature used for creating the context vector.

to set beam size to around 5 to 10 (Wu et al., 2016, Edunov et al., 2018). Beam search decoding, even though effective, still suffers from *exposure bias*. Exposure bias results from the mismatch between how the models are trained and how they are used at inference (Wiseman and Rush, 2016, Ranzato et al., 2016). During training, the model is guided by the ground-truth target translation. However, at inference, target translations are not available and the model has to rely on its own predictions which can be wrong. Collobert et al. (2019) proposed a fully differentiable beam search decoder that can be used during training and eliminates this bias.

2.4.3 Attention mechanism

One of the shortcomings of the discussed models is that the translation quality decreases considerably as sentences become longer (Koehn and Knowles, 2017). One reason is that the source sentence is encoded into one *fixed length* vector and this vector is expected to be a complete and static representation of the source sentence. To address this problem, several works focus on learning a context vector with connections to the source sentence (Graves et al., 2014, Bahdanau et al., 2015, Luong et al., 2015a). This context vector regulates the alignment between the source and the target sentences and is a sum of the hidden states of the input, weighted by alignment scores. At each time step t, the model computes a variable-length alignment weight vector based on the current target state and all source inputs. Table 2.1 summarizes different approaches for computing alignment scores.

It is worth noting that while attention matches traditional word alignment at times, it often captures relations beyond that between the source and target sentence (Ghader and Monz, 2017, Koehn and Knowles, 2017).

2.5 Fully attention-based translation models

In the previous section, we discussed attention mechanisms where the model selectively attends to the source sequence to make predictions. Self-attention is a type of

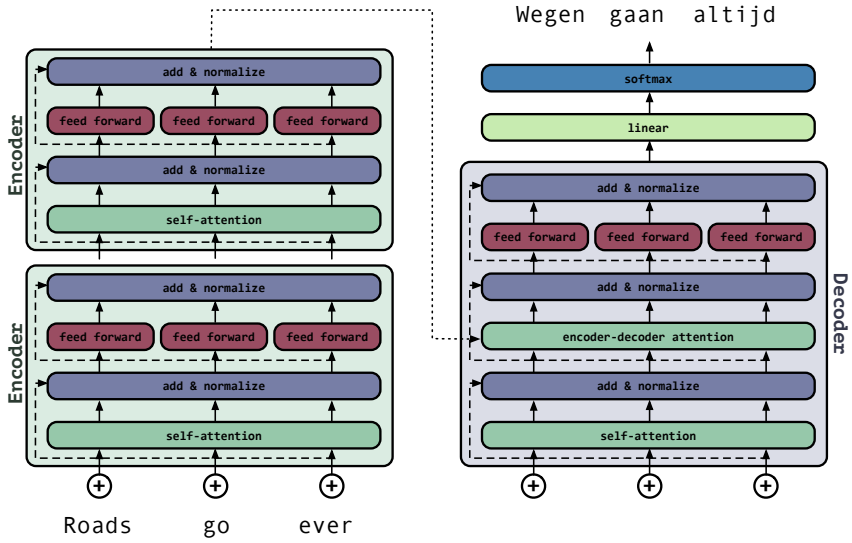


Figure 2.3: An illustration of a Transformer model introduced in Vaswani et al. (2017).

attention mechanism that connects different positions *within* a sequence to compute a representation. The Transformer model proposed by Vaswani et al. (2017) is a sequence-to-sequence model that relies solely on attention to encode the input and generate the output sequence. One of the main advantages of this architecture is that it can be trained with massive parallelization because it bypasses the recurrent dependency that exists in RNN models. The transformer model has been shown to perform quite well in bilingual and multilingual settings (Lakew et al., 2018) and has become the most common choice to implement NMT models (Edunov et al., 2018, Aharoni et al., 2019). In this section, we describe this architecture in more detail.

2.5.1 Model architecture

The transformer model is an encoder-decoder architecture without a sequential structure. The encoder is given an input sequence of tokens $X = \left[x_1, \ldots, x_n\right]$, and encodes it as a continuous representation $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right]$ based on the attention. The decoder then generates the output sequence $Y = \left[y_1, \ldots, y_m\right]$ token by token, given the representation $\mathbf{X}$ and the previously generated token. Figure 2.3 illustrates this architecture which we will discuss in detail below.

For every token x_i in the input sequence, we first create a query $\mathbf{q}_i$, a key $\mathbf{k}_i$, and a value $\mathbf{v}_i$ vector. Self-attention then uses *scaled dot product attention* (last row in Table 2.1) to compute the attention score of token x_i against other words in the input sequence. This attention has a scaling factor where n is the dimension of the source hidden states. To calculate representation $\mathbf{x}_i$, a softmax layer is then used to normalize

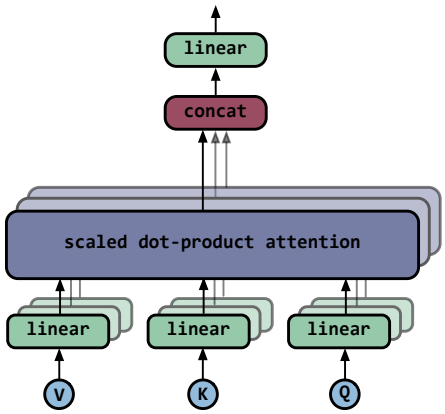


Figure 2.4: An illustration of self-attention in the Transformer model based on Vaswani et al. (2017).

the self-attention scores and multiplies it with $\mathbf{v}_i$. In practice, the attention is computed on matrices of inputs $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{K}^\intercal \mathbf{Q}}{\sqrt{d_k}})\mathbf{V} \tag{2.10}$$

where d_k is the embedding dimension of the key vectors which scales the dot product. The encoder is a stacking of identical layers each consisting of a *multi-head self-attention* layer and a point-wise fully connected feed-forward network. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrices are split up into multiple heads and the multi-head attention mechanism computes the attention in parallel. Each token in the sequence goes through the encoder independently. During encoding, there are dependencies between the paths of different tokens in the self-attention layer, but the feed-forward layer of each token does not have any dependencies. As a result words in the sequence can be processed in parallel. The independent attention outputs are then concatenated and linearly projected as follows:

$$\text{Multi-Head Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\text{head}_1, \ldots \text{head}_h)\mathbf{W}^o \tag{2.11}$$

where

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \tag{2.12}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ are weight matrices that map the input representations to the query, key, and value matrices. $\mathbf{W}^o$ is the linear transformation that generates the output. All weight matrices are learned during training of the model. Figure 2.4 illustrates this component. Similarly to the encoder, the decoder consists of a stack of identical layers,

as well as a third sub-layer, which computes multi-head attention over the output of the encoder stack. The self-attention layers in the decoder work slightly differently from the ones in the encoder. The computation of attention is *masked* before the softmax step to prevent looking to the future of the sequence during training.

Finally, there is a fully connected neural network that transforms the output of the stack of decoders into the target vocabulary vector. The softmax function turns the scores into probabilities and the word with the highest probability is generated (greedy decoding). Alternatively, decoding can be done using the beam search technique similar to the RNN models discussed in Section 2.4.2.

2.5.2 Residual connections

Another effective detail of the transformer architecture is the inclusion of residual connections (He et al., 2016b) to facilitate optimization. Residual connections connect the output of one layer with the input of an earlier layer. Every self-attention and feed-forward neural network in the encoder and the decoder stack has a residual connection around it and a normalization layer (Ba et al., 2016). This shortcut connection is particularly effective in training very deep architectures and mitigates the vanishing gradient problem.

2.5.3 Positional Encoding

As discussed earlier, the transformer model does not have a recurrent structure and can be trained with a high degree of parallelization. However, languages are structured sequentially and it is necessary to encode some form of word order in the sequence (Tran et al., 2018). To address this shortcoming, the transformer adds a positional encoding vector to every word in the input sequence. These embeddings model the position of each word, or the relative distance between different words in the input.

Vaswani et al. (2017) proposed sine and cosine functions of different frequencies to compute positional encodings:

$$\text{positional encoding}_{(i,\delta)} = \begin{cases} \sin(\frac{i}{10000^{2\delta'/d}}) & \text{if } \delta = 2\delta' \\ \\ \cos(\frac{i}{10000^{2\delta'/d}}) & \text{if } \delta = 2\delta' + 1 \end{cases} \tag{2.13}$$

where i is the position and $\delta = 1, \ldots, d$ is the dimension. They also experimented with learned positional embeddings similar to Gehring et al. (2017), by assigning each input token with a learned vector that encodes its absolute position, and observed similar results to the sinusoidal version.

2.6 Translation evaluation

We evaluate all translation experiments in this thesis using the BiLingual Evaluation Understudy metric, better known as BLEU (Papineni et al., 2002). This metric assesses the closeness of the generated translation to a human reference translation. It includes a *brevity penalty* (BP) to avoid preferring shorter translations. The BLEU score for n-grams up to length N is defined as:

$$\text{BLEU}_n = \text{BP} \,.\, \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.14}$$

where w_n is a weight assigned to the size of n-gram (often set uniformly to $1/N$). p_n is computed as:

$$p_n = \frac{\sum_{c \in \{candidates\}} \sum_{n\text{-gram} \in c} \text{count}_{clip}(n\text{-gram})}{\sum_{c' \in \{candidates\}} \sum_{n\text{-gram}' \in c'} \text{count}(n\text{-gram}')} \tag{2.15}$$

where:

$$\text{count}_{clip}(x) = \min(\text{count}(x), \textit{max\_ref\_count})(x)) \tag{2.16}$$

Here, $candidates$ are translation candidates, and *max_ref_count* is the largest count observed in the reference for that word. Scores are calculated over sentence pairs in the test set and the average BLEU is reported for the entire test set. Unless stated otherwise, in this thesis we compute case-sensitive BLEU up to and including n-grams of length 4.

We also use other evaluation metrics, namely METEOR and Translation Error Rate (TER), in some chapters of this thesis. METEOR is another metric to automatically evaluate translation quality (Banerjee and Lavie, 2005, Denkowski and Lavie, 2011, 2014). Similar to BLEU, this metric compares the translation output with a reference translation, however, it addresses some of the deficiencies of the BLEU metric. This is done by aligning the two sentences not only based on the exact match, but also on matching synonyms and paraphrases. METEOR has to be fine-tuned to achieve maximum correlation with human judgments (Agarwal and Lavie, 2008). TER is an easy-to-explain metric to compare translation output and manually created reference translation (Snover et al., 2006). It measures the number of edits required to change a translation output into one of the references. A higher score of TER is a sign of more post-editing effort and it may not always correlate with translation quality.

While all these metrics attempt to measure translation quality, they assume inexact models of permissible variations in translation and may not capture the precise quality of a system (Callison-Burch et al., 2006). However, they allow for systematic evaluation of incremental changes to a single system and are very inexpensive to perform. We

specifically choose BLEU because it is the most common metric and it makes it possible to compare various systems. In Chapters 6 and 7 of this thesis, we explore cases where the translation quality of NMT models are affected, but individual automatic metrics do not reflect this change in quality.

3

Topic-Sensitive Word Representations

3.1 Introduction and research questions

Word representations in the form of dense vectors, or word embeddings, capture semantic and syntactic information (Mikolov et al., 2013a, Pennington et al., 2014) and are widely used in many NLP tasks such as sentiment analysis (Tang et al., 2014, Yu et al., 2017), identifying multiword expressions (Salehi et al., 2015, Gharbieh et al., 2016), and translation (Zou et al., 2013, Artetxe et al., 2018a). These representation models are based on the assumption that the meaning of a word can be inferred from its textual context (Firth, 1957).

Currently, there are two categories of approaches to learning word representations (discussed in Section 2.3): *static embeddings* where a fixed vector is learned for each word in the vocabulary, and *dynamic embeddings* where vectors are dynamically calculated for each sentence. Dynamic embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) store the learned weights of the network, and use that to get word representations by computing them on the fly for a given context. As a result, these models can capture context-dependent characteristics of the language such as polysemy: in natural language, words usually have more than one meaning (or sense).

Like dynamic embeddings, the research presented in this chapter aims at overcoming the inability of static embeddings to capture polysemy. However, it predates dynamic embeddings. Before the advent of contextualized embedding approaches, most static representation models learned *one* fixed-length representation per word. However, this approach and how it is evaluated has some shortcomings.

Firstly, many tasks can benefit from using multiple representations per word to capture polysemy (Bengio et al., 2003, Reisinger and Mooney, 2010). Many intricate distinctions of word senses are lost when we use one embedding vector to capture multiple meanings. Additionally, this simplification of natural language unintentionally leads to more simplistic evaluation tasks. Most studies on static word embeddings used word similarity task to assess the accuracy of the static word representations where word pairs are ranked based on how similar or related they are. However, most of the word

similarity benchmarks present words out of context. The word pair '*bank*' and '*reef*' can have very different similarity scores depending on the context of the word *bank*. Finally, there is no clear and quantifiable definition of *similarity* and *relatedness* when comparing two words, and as a result, different benchmarks have different interpretations (Faruqui et al., 2016).

In this chapter, we seek to address these shortcomings and propose an approach for learning multiple static word representations per word. We aim to understand the role of a particular kind of context, namely document topics, for learning these representations. We analyze to what extent learning multiple topic-sensitive embeddings per word captures polysemy, which we believe is a necessary step towards further understanding the impact of context. Concretely, we ask:

Research Question 1: *Can document-level topic distribution help infer the meaning of a word?*

We first look at the importance of document-level context and how it can help to separate different meanings of the word. We study the integration of this topical information in learning word representation and evaluate the embeddings on a contextual task. Concretely, we ask:

RQ1.1 *To what extent can distributions over word senses be approximated by distributions over topics of documents without assuming these concepts to be identical?*

We introduce a model that uses a nonparametric Bayesian model, namely Hierarchical Dirichlet Process (HDP), to learn multiple topic-sensitive representations per word. Yao and Van Durme (2011) showed that HDP is effective in learning topics yielding state-of-the-art performance for sense induction. This approach learns the granularity of senses from the data and does not require heuristic parameter setting. The authors assumed that topics and senses are entirely interchangeable, and so they trained individual models per word. However, this assumption makes it difficult to scale to large data. In our approach, we do not hold the same assumption, which enables us to use HDP to model topics effectively using large unannotated training data. We aim to *approximate* the word senses with topics and further use this additional signal for training the embeddings of each topic-word pair separately.

RQ1.2 *How can we exploit document-level topics to distinguish between different meanings of a word and learn the corresponding representations?*

We propose three unsupervised, language-independent approaches to approximate senses with topics and learn multiple topic-sensitive embeddings per word. Our first model uses a hard topic labeling approach to learn representations. The second model jointly learns topic-labeled and generic representations for each word in

order to share statistical information between different meanings of a particular word. The third model uses topic distributions for each word following the notion that meanings of words are not mutually exclusive in a given context. We show that in the lexical substitution ranking task (McCarthy and Navigli, 2007) our models outperform two competitive baselines and perform comparably to the best-performing methods despite the fact that—unlike those methods—our approach does not use any syntactic information.

RQ1.3 *What are the advantages of using document-level topics in learning multiple representations per word?*

The process of learning topics and topic-sensitive representations is applied to the same corpus ensuring compatibility between the granularity of topics and diversity of meanings of word embeddings. By learning the granularity of topics from the corpus we do not use any external knowledge sources. As a result, this approach can be used for low-resource languages with no manually curated knowledge sources. Additionally, this approach broadens the contextual signals for learning more accurate representations when sentence-level context is not sufficient. We evaluate our representations on the contextual word similarity task and the lexical substitution task, both of which showcase the importance of learning multiple embeddings per word.

Organization. This chapter is organized as follows: In Section 3.2, we provide an overview of existing work on word sense disambiguation and static sense representations literature. Next, in Section 3.3, we introduce our representation models. We present experimental details and study different tasks to evaluate the representations in Section 3.4. In Section 3.5, we present a more in-depth analysis of the resulting representations. Finally, we discuss the conclusions and implications of this work in Section 3.6.

3.2 Related work

In this section, we discuss previous works that focus on learning multiple static word embeddings per word to capture polysemy, as well as related work in the area of word sense disambiguation.

3.2.1 Word sense disambiguation

Word sense disambiguation is the problem of determining which *sense* of a word is activated by the use of the word in a particular context (Ide and Véronis, 1998).

There have been several attempts to build repositories for word senses (Miller, 1995, Navigli and Ponzetto, 2010), but this is laborious and therefore limited to few languages.

Moreover, defining a universal set of word senses is challenging as polysemous words can exist at many levels of granularity (Kilgarriff, 1997, Navigli, 2012). For this reason, earlier work focuses on unsupervised sense induction, often following a Bayesian framework (Brody and Lapata, 2009, Lau et al., 2014).

Yao and Van Durme (2011) show that a nonparametric Bayesian model, such as the Hierarchical Dirichlet Process (HDP), is effective in learning topics and can yield state-of-the-art results in sense induction. The advantage of nonparametric methods is that they learn the granularity of topics from the data and do not require to fix the number of senses per word a priori. By using HDP for sense induction, they assume that topics and senses are interchangeable and train a topic model for each target word with a sampled number of context instances. However, inference with HDP does not scale to large corpus sizes due to the complexities of the model (Jordan, 2011, Paisley et al., 2015).

3.2.2 Static sense representations

We discussed in Section 2.3.1 that the most commonly used approaches learn exactly one embedding per word (Mikolov et al., 2013a, Pennington et al., 2014). However, even before the advent of dynamic embeddings, several studies have focused on learning multiple embeddings per word due to the ambiguous nature of language (Qiu et al., 2016). Huang et al. (2012) cluster word contexts and use the average embedding of each cluster as word sense embeddings, which yields improvements on a word similarity task. Neelakantan et al. (2014) propose two approaches, both based on clustering word contexts: In the first, they fix the number of senses manually, and in the second, they use an ad-hoc greedy procedure that allocates a new representation to a word if existing representations explain the context below a certain threshold.

Li and Jurafsky (2015) used a Chinese Restaurant Process (CRP) model to distinguish between senses of words and train vectors for senses, where the number of senses is not fixed. They change the Skipgram model (Mikolov et al., 2013a) to perform sense induction and sense embedding updates simultaneously. They use two heuristic approaches for assigning senses in a context: 'greedy' which assigns the locally optimum sense label to each word, and 'expectation' which computes the expected value for a word in a given context with probabilities for each possible sense.

3.3 Topic-Sensitive representations

In this section, we introduce our approach to learn topic-sensitive word representations based on the Skipgram model proposed by Mikolov et al. (2013a). We previously mentioned that inference with HDP does not scale to large corpus sizes (Jordan, 2011, Paisley et al., 2015). Here, we describe our proposed models to learn topics from a corpus using HDP (Teh et al., 2006, Lau et al., 2014) in a way that is applicable to large corpora.

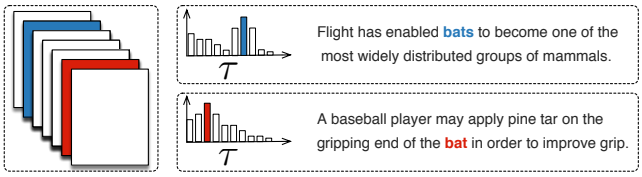


Figure 3.1: An example of sentences including the word 'bat' from two documents with different topic distribution.

The main advantage of this model compared to non-hierarchical methods like the Chinese Restaurant Process (CRP) is that each document in the corpus is modeled using a mixture model with topics shared between all documents (Teh et al., 2005, Brody and Lapata, 2009). HDP yields two sets of distributions that we use in our methods: (i) distributions over topics for words in the vocabulary, and (ii) distributions over topics for documents in the corpus.

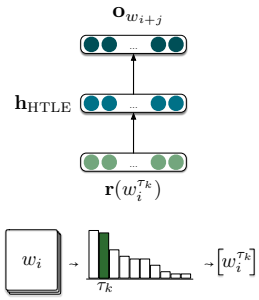
Similarly to Neelakantan et al. (2014), we use neighboring words to detect the meaning of the context, however, we also use the two HDP distributions. By doing so, we take advantage of the topic of the document beyond the scope of the neighboring words, which is helpful when the immediate context of the target word is not sufficiently informative. We modify the Skipgram model (Mikolov et al., 2013a) to obtain multiple topic-sensitive representations per word type using topic distributions.

Additionally, the context vectors of a word type with multiple topics are shared in our model. This is especially beneficial for infrequent words, where a rare sense of the word can use the contextual information of other senses of the word. We assume that meanings of words can be determined by their contextual information and use the distribution over topics to differentiate between occurrences of a word in different contexts, i.e., documents with different topics (see example in Figure 3.1). We propose three different approaches illustrated in Figures 3.2 and 3.3: two methods with hard topic labeling of words and one with soft labeling. In the following sections, we discuss each of these model variants in detail.

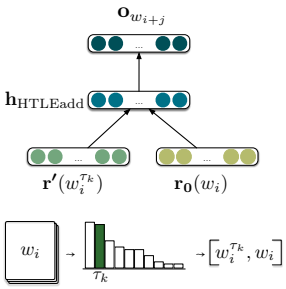
3.3.1 Hard topic-labeled representations

In the hard-labeling approach, we assign exactly one topic to each word based on sampling from the topic distribution. We use the trained HDP model to label every word in the training data with the chosen topic ID.

Our first model variant (Figure 3.2 (a)) considers each word-topic pair as a separate vocabulary entry. To reduce sparsity on the context side and share the word-level information between similar contexts, we use topic-sensitive representations for target words (input to the Skipgram network) and standard, i.e., unlabeled, word representations for context words (output to the Skipgram network). Note that this results in different input



(a) HTLE: Hard-topic labeled representations model



(b) HTLEadd: Hard topic-labeled representations plus generic word representations model

Figure 3.2: Illustrations of two proposed models with hard-labeling topics in this chapter.

and output vocabularies. The training objective is then to maximize the log-likelihood of context words w_{i+j} given the target word-topic pair w_i^τ:

$$\mathcal{L}_{HardT\text{-}SG} = \frac{1}{I}\sum_{i=1}^{I}\sum_{\substack{-c\leqslant j\leqslant c\\ j\neq 0}} \log p(w_{i+j} \mid w_i^\tau) \tag{3.1}$$

where I is the number of words in the training corpus, c is the context size and τ is the topic assigned to w_i by HDP sampling. $\mathbf{o}_{w_{i+j}}$ is the context (i.e., output) representation for the word w_i. Note that w_i is an occurrence of word w in context $[i-c, i+c]$.

The embedding of a word in context $\mathbf{h}(w_i)$ is obtained by simply extracting the row of the input lookup table ($\mathbf{r}$) corresponding to the HDP-labeled word-topic pair:

$$\mathbf{h}_{\text{HTLE}}(w_i) = \mathbf{r}(w_i^\tau) \tag{3.2}$$

A possible shortcoming of the HTLE model is that the representations are trained separately and information is not shared between different topic-sensitive representations of the same word. To address this issue, we introduce a model variant that learns multiple topic-sensitive word representations and generic word representations simultaneously (Figure 3.2 (b)). In this variant (HTLEadd), the target word embedding is obtained by adding the word-topic pair representation ($\mathbf{r}'$) to the generic representation of the corresponding word ($\mathbf{r_0}$):

$$\mathbf{h}_{\text{HTLEadd}}(w_i) = \mathbf{r}'(w_i^\tau) + \mathbf{r_0}(w_i) \tag{3.3}$$

This representation captures both the generic and the contextual meaning of the

word.

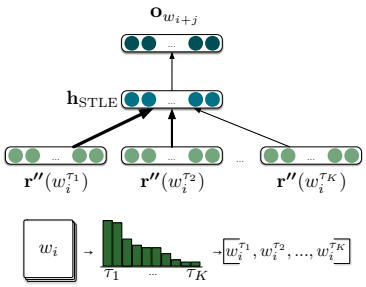


Figure 3.3: Illustration of our soft topic-labeled representation model (STLE).

3.3.2 Soft topic-labeled representations

The model variants above rely on the hard labels resulting from HDP sampling. As a soft alternative to this, we can directly include the topic distributions estimated by HDP for each document, see Figure 3.3. Since the topics are not clearly separated, every identified topic of a word can contribute to the learning process proportional to its value. Specifically, for each update, we use the topic distribution to compute a weighted sum over the word-topic representations ($\mathbf{r''}$):

$$\mathbf{h}_{\text{STLE}}(w_i) = \sum_{k=1}^{T} p(\tau_k \mid d_i)\, \mathbf{r''}(w_i^{\tau_k}) \tag{3.4}$$

where T is the total number of topics, d_i the document containing w_i, and $p(\tau_k \mid d_i)$ the probability assigned to topic τ_k by HDP in document d_i. The training objective for this model is:

$$\mathcal{L}_{SoftT\text{-}SG} = \frac{1}{I} \sum_{i=1}^{I} \sum_{\substack{-c \leqslant j \leqslant c \\ j \neq 0}} \log p(w_{i+j} \mid w_i, \tau) \tag{3.5}$$

where τ is the topic of document d_i learned by HDP. The STLE model has the advantage of directly applying the distribution over topics in the Skipgram model. Also, for each instance, we update all topic representations of a given word with non-zero probabilities, which has the potential to reduce the sparsity problem.

3.3.3 Embeddings for polysemous words

The representations obtained from our models are expected to capture the meaning of a word in different topics. We now examine whether these representations can distinguish between different word senses. Table 3.1 provides examples of nearest neighbors. For

Table 3.1: Nearest neighbors of three examples in different representation spaces using cosine similarity. **word2vec** and **GloVe** are pre-trained embeddings from (Mikolov et al., 2013a) and (Pennington et al., 2014), respectively. **SGE** is the Skipgram baseline and **HTLE** is our topic-sensitive Skipgram (cf. Equation (3.2)), both trained on Wikipedia. τ_k stands for HDP-inferred topic k.

	Pre-trained		**Trained on Wikipedia**		
	word2vec	**Glove**	**SGE**	**HTLE:** τ_1	**HTLE:** τ_2
bat	bats	bats	uroderma	ball	vespertilionidae
	batting	batting	magnirostrum	pitchout	heran
	Pinch_hitter_Bray…	Bat	sorenseni	batter	hipposideros
	batsman	catcher	miniopterus	toss-for	sorenseni
	batted	fielder	promops	umpire	luctus
	Hawaiian_hoary	hitter	luctus	batting	coxi
	Lelands.com…	outfield	micronycteris	bowes	kerivoula
	yelled_Cheater	hitting	hipposideros	straightened	natterer
	wicketkeeper_Andr…	batted	chaerephon	fielder	nyctophilus
	lefthanded_batter	catchers	pteronotus	flies	artibeus
jaguar	jaguars	jaguars	electramotive	ford	wiedii
	Macho_B	xk8	vk66de	bmw	puma
	panther	xj6	viper	chevrolet	margay
	lynx	xjs	id66	honda	tapirus
	rhino	panther	xj666	porsche	jaguarundi
	lizard	xkr	roadster	multimatic	yagouaroundi
	tapir	xj8	saleen	monza	vison
	tiger	mercedes	siata	nissan	concolor
	leopard	Jaguar	enetered	xj	tajacu
	Florida_panther	porsche	chevrolet	dodge	tayassu
appeal	appeals	appeals	court	court	sfa
	appealing	appealed	appeals	case	steadfast
	appealed	appealing	appealed	appeals	lackadaisical
	Appeal	Appeal	carmody	appealed	assertions
	rehearing	court	upheld	decision	lack
	apeal	decision	verdict	proceedings	symbolize
	Appealing	conviction	jaruvan	disapproves	fans
	ceasing_hostilities…	plea	affirmed	ruling	attempt
	ruling	sought	appealable	upholding	unenthusiastic
	Appeals	dismiss	battin	carmody	cancellation

Table 3.2: Statistics of the degree of polysemy in Wordnet and HTLE.

	Wordnet	**HTLE**
Degree of polysemy	2.08	4.79
Single [sense/representation] words	26,755	21,490

comparison, we include our own baseline, i.e., embeddings learned with Skipgram on our corpus, as well as Word2Vec (Mikolov et al., 2013b) and GloVe embeddings (Pennington et al., 2014) pre-trained on large data.

In the first example, the word *bat* has two different meanings: animal or sport device. We can see that the nearest neighbors of the baseline and pre-trained word representations either center around one primary, i.e., most frequent, meaning of the word, or looks like a mixture of different meanings. The topic-sensitive representations, on the other hand, correctly distinguish between the two different meanings. A similar pattern is observed for the word *jaguar* and its two meanings: car or animal. The last example, *appeal*, illustrates a case where topic-sensitive embeddings are not clearly detecting different meanings of the word, despite having some correct words in the lists. Here, the meaning *attract* does not seem to be captured by any embedding set.

These observations suggest that topic-sensitive representations capture different word senses to some extent. To quantify the degree of polysemy in our embeddings, we compare it to Wordnet (Miller, 1995). Wordnet is a manually curated lexical database of English that interlinks different senses of the words through conceptual-semantic and lexical relations. As a result, words that are found near one another in the network are semantically disambiguated. Table 3.2 shows statistics for Wordnet and our proposed embedding method. We observe that the *degree of polysemy* of our embeddings is more than double that of Wordnet. Note that while in our models we do not explicitly specify the desired number of senses per word type, the hyperparameters have an impact on it: γ manages the variability of the global sense distribution and α manages the variability of each word type's selection of senses. These hyperparameters are discussed in Section 3.4.1.

Moreover, we observe that not every topic-sensitive word representation corresponds to a distinct and unique sense. In our experiments, we see that at times, multiple embeddings capture the same sense of the word. However, they are also in close proximity in the embedding space and end up being very similar. To provide a systematic validation of our approach, we now investigate whether topic-sensitive representations can improve tasks where polysemy is a known issue.

3.4 Evaluation

In this section, we present the setup for our experiments and empirically evaluate our approach on the context-aware word similarity and lexical substitution tasks.

3.4.1 Experimental setup

All word representations are learned on the English Wikipedia corpus containing 4.8M documents (approximately 1 billion tokens). Preprocessing of the training data includes lowercasing, removing stop words, and removing words occurring less than 100 times. The topics are learned on a 100K-document subset of this corpus using the HDP implementation of Teh et al. (2006). HDP has two hyperparameters, γ and α, which control the variability of the global topic distribution and each word's choice of topics, respectively. We do not tune these parameters and following the literature, we put gamma priors $\mathrm{Gamma}(1, 1)$ and $\mathrm{Gamma}(1, 0.1)$ on hyperparameters γ and α respectively. These parameters encourage skewed topic distributions which are typically observed in natural languages (Gale et al., 1992). Once the topics have been learned, we run HDP on the whole corpus to obtain the word-topic labeling (Section 3.3.1) and the document-level topic distributions (Section 3.3.2). We train each model variant with window size $c = 10$ and different embedding sizes (100, 300, 600) with random initialization. All model variants in this chapter are trained on the same training data with the same settings, following suggestions by Mikolov et al. (2013a) and Levy et al. (2015).

Table 3.3: Word similarity benchmarks for intrinsic evaluation of word representations.

Data set	Word pairs	Reference
RG	65	Rubenstein and Goodenough (1965)
WS353	353	Finkelstein et al. (2001)
MTurk287	287	Radinsky et al. (2011)
MEN	3000	Bruni et al. (2012)
RW	2034	Luong et al. (2013)
SimLex	999	Hill et al. (2015)

3.4.2 Word similarity task

The most popular intrinsic evaluation of static word representations is the word similarity task. In this task, a list of pairs of words with their similarity scores judged by human annotators is provided. The goal is to measure how well the word vector representations capture the notion of word similarity by ranking the word pairs according to their

similarity scores. Table 3.3 provides a list of benchmarks with the number of word pairs in each data set that we use for evaluation.

Similar to most previous approaches (Radinsky et al., 2011, Hassan and Mihalcea, 2011, Yih and Qazvinian, 2012), we use Spearman's ρ (rank correlation coefficient) to assess the monotonic relationship between the model's ranking of word pairs and the gold standard's ranking.

Our models learn multiple embeddings per word, but these benchmarks do not include any context to help distinguish between the vectors. Therefore we employ several techniques for selecting and combining the representation vectors:

- **Max**: computes the pairwise similarity between the nearest topic-sensitive embeddings of the word pair.
- **Mean**: computes the pairwise similarity between the means of all topic-sensitive embeddings of each word.
- **wMean**: computes the pairwise similarity between the weighted means of all topic-sensitive embeddings of each word. The weights are defined according to the frequency of each topic.

Table 3.4 provides the results for the word similarity experiments. We observe slight improvements in different settings, but there is no clear indication that one model performs best across all data sets. It is also clear that each data set has a different level of difficulty, and because of the differences in quality of the word pairs and the definition of *similarity* for the annotators, they are not analogous.

Table 3.4: Spearman's rank correlation performance on word similarity tasks. All vectors are 100-dimensional.

		RG	**WS353-rel**	**WS353-sim**	**MTurk287**	**MEN**	**RW**	**SimLex**
SGE		0.77	0.44	0.69	0.66	**0.71**	**0.38**	0.29
HTLE	Max	0.68	0.20	0.46	0.42	0.51	0.15	0.20
	Mean	0.65	0.29	0.61	0.62	0.57	0.35	0.22
	wMean	0.48	0.32	0.40	0.55	0.50	0.08	0.10
HTLEadd	Max	**0.81**	0.30	0.57	0.56	0.63	0.24	0.22
	Mean	0.77	0.42	0.67	**0.68**	0.69	0.36	0.28
	wMean	0.60	0.36	0.48	0.63	0.57	0.13	0.14
STLE	Max	0.74	0.43	0.69	0.67	0.69	0.20	0.30
	Mean	0.71	**0.45**	**0.69**	0.67	0.67	0.23	**0.30**
	wMean	0.65	0.43	0.64	0.65	0.68	0.16	0.24

One of the main concerns of using these benchmarks, in general, is that the notion of word similarity is subjective and there is no clear division between similarity and

relatedness (Faruqui et al., 2016, Torabi Asr et al., 2018). As a result, some data sets penalize representation models that consider two related words as 'not similar', while others do not. For instance, in MEN (Bruni et al., 2012), the guidelines did not distinguish between similarity and relatedness and gave examples of both similarity (e.g., *"car-automobile"*), and relatedness (e.g., *"wheels-car"*) as valid options to the annotators. The instructions for the SimLex data set (Hill et al., 2015), however, included guidelines for the annotators with examples of related pairs (e.g., *"car-tyre"*) that are *not* to be labeled similar.

Faruqui et al. (2016) evaluated several issues of the word similarity task. These include low correlation with extrinsic evaluation, no consideration of polysemy, absence of statistical significance, and semantic versus task-specific embeddings. To specifically address the lack of context to identify polysemous words, Huang et al. (2012) proposed the Stanford contextual word similarity data set (SCWS). In the following subsection, we evaluate our embeddings using this data set.

3.4.3 Context-Aware word similarity task

As mentioned before, there are multiple test sets available for intrinsic evaluation of embeddings, but in almost all of them word pairs are considered out of context. To evaluate our static embeddings intrinsically, we use the SCWS data set (Huang et al., 2012). To the best of our knowledge, this was the only word similarity data set considering word context at the time our models were developed. Note that more recently, instead of intrinsic evaluations, the performance of dynamic contextual embeddings is typically evaluated on downstream NLP tasks (Peters et al., 2018, Devlin et al., 2019).

The SCWS data set contains word pairs and their respective contexts with average human ratings indicating the similarity of the target words. Table 3.5 presents examples of word pairs and their contexts in SCWS. To evaluate our models on SCWS, we run HDP on the data treating each word's context as a separate document. We compute the similarity of each word pair as follows:

$$\text{Sim}(w_1, w_2) = \cos(\mathbf{h}(w_1), \mathbf{h}(w_2)) \tag{3.6}$$

where $\mathbf{h}(w_i)$ refers to any of the topic-sensitive representations defined in Section 3.3. Note that w_1 and w_2 can refer to the same word.

We compare our models to various baselines: The Skipgram model (SGE), the context-aware Skipgram model (SGE + context), and the best-performing multi-sense embeddings model per word type (MSSG) (Neelakantan et al., 2014). The context-aware Skipgram baseline (SGE + context) computes the average pairwise cosine similarity between a target word in each context with every word in the opposing context.

For MSSG we use the best performing similarity measure (avgSimC) as proposed

Table 3.5: Examples from SCWS data set. Each example includes the word pair (identical or non-identical), their corresponding contexts, and the average human score between 0 and 10 to indicate the similarity.

Word pair	*bitter, bitter*
context$_1$	It has an aromatic, warm and slightly bitter taste.
context$_2$	AK - a very common beer name in the 1800s - was often referred to as a "mild bitter beer" interpreting "mild" as "unaged".
Human score	6.0
Word pair	*bitter, resentful*
context$_1$	Named for the tattoos they decorated themselves with and bitter enemies of encroaching Roman legions, the Picts fired Howard's imagination and crystallized in him a love for barbarians and outsiders from civilization who lived lives of great hardship and struggle but also great freedom and verve.
context$_2$	Legge-Bourke had been hired by Prince Charles as a young companion for his sons while they were in his care, and Diana was extremely resentful of Legge-Bourke and her relationship with the young princes.
Human score	9.0
Word pair	*bitter, taste*
context$_1$	This practice began during the Prohibition as a means of covering the bitter taste.
context$_2$	Once it has decayed, it leaves no taste or odor in drinking water.
Human score	7.0

by Neelakantan et al. (2014):

$$\text{avgSimC}(w_1, w_2) = \sum_{j=1}^{K} \sum_{i=1}^{K} P(w_1, c_1, i) P(w_2, c_2, j) d(\mathbf{v}(w_1, i), \mathbf{v}(w_2, j)) \tag{3.7}$$

where $P(w, c, k)$ is the probability that w takes the k-th sense given context c. $\mathbf{v}(w, k)$ is the embedding for word w with assigned sense k. $d(\mathbf{v}(w_1, i'), \mathbf{v}(w_2, j'))$ is the similarity measure between the given embeddings $\mathbf{v}(w_1, i')$ and $\mathbf{v}(w_2, j')$. avgSimC measures the similarity between each pair of senses by how well each sense fits the context at hand.

Table 3.6 provides the Spearman's correlation scores for different models against the human ranking. We see that with dimensions 100 and 300, two of our proposed models obtain slight improvements over the baseline. However, for higher dimensions (embedding size 600), the MSSG model is the best performing system.

The main advantage of having multiple embeddings per word for different meanings is in comparing pairs of identical word with ambiguous meanings. With multiple representations per word, we can have a better estimation of similarities for identical words, given that we detect different senses correctly. Spearman's rank correlation between the two systems is based on the average differences between the two ranks of

Table 3.6: Spearman's rank correlation performance for the Word Similarity task on SCWS (Huang et al., 2012).

	Dimension		
Model	100	300	600
SGE + context (Mikolov et al., 2013a)	0.59	0.59	0.62
MSSG (Neelakantan et al., 2014)	0.60	**0.61**	**0.64**
HTLE	**0.63**	0.56	0.55
HTLEadd	0.61	**0.61**	0.58
STLE	0.59	0.58	0.55

each observation. To further understand the performance of our models, we look into the rank differences for the two types of word pairs in SCWS (identical and non-identical) and the two types of embeddings (assigning the same topic or not). The former explicitly evaluates the performance of our models on identical words, and the latter evaluates the impact of the topic-labeling step. This results in four categories for comparison.

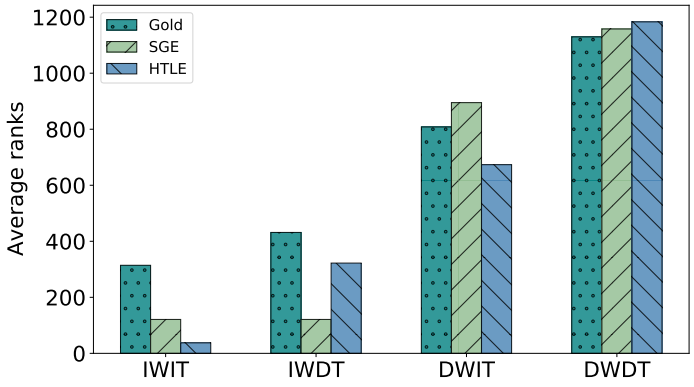


Figure 3.4: Average absolute rank of the baseline embeddings and the HTLE embeddings in four categories, where being closer to the gold rankings is better. The categories are marked with a combination of these labels: **I**: identical, **D**: different. **W**: word, **T**: topic. For instance `IWDT` is the category of word pairs where ***i****dentical* ***w****ords have* ***d****ifferent* ***t****opics*. Examples of the categories are presented in Table 3.5.

Since the difficulty of each category is different, we expect different performances from the models. Figure 3.4 shows the average absolute rank of the baseline embeddings and the HTLE embeddings with 300 dimensions in these four categories. The gold rank, which is the ranking of all word pairs in the data set according to human judgments, is

also shown for each category. The best-performing model is the one that is the closest to the gold ranking. Note that the gold rank naturally increases when words are different in comparison to when they are the same.

One can see that for identical words, labeled with different topics, `IWDT`, the rank assigned by topic-sensitive embeddings is much closer to the gold ranking than the one produced by the baseline. The average rank is also still higher when considering both categories of identical words: `IWDT` and `IWIT`. This indicates that the estimation of the similarity scores of identical words is notably more accurate in our model. However, for non-identical word pairs (`DWIT` and `DWDT`), the rank difference is higher for topic-sensitive embeddings and since the evaluation set consists of mostly non-identical word pairs, the correlation with gold ranking decreases in total in comparison with baseline word embeddings.

3.4.4 Lexical substitution task

Continuing our evaluation of word representations, in this section, we explore the lexical substitution task. This task requires one to identify the best replacements for a word in a sentential context. The replacements should be both semantically compatible with the word, and syntactically correct in the context. For example:

sentence	**substitutions**
The sun was bright.	*luminous, colorful*
He was bright and independent.	*intelligent, clever, smart*

The presence of many polysemous target words makes this task more suitable for evaluating sense embeddings. Following Melamud et al. (2015), we pool substitution candidates from different instances and rank them by the number of annotators that selected them for a given context. We use two evaluation sets: LS-SE07 (McCarthy and Navigli, 2007), and LS-CIC (Kremer et al., 2014). The Concept-in-Context (LS-CIC) set for the lexical substitution task is a large-scale corpus constructed by crowdsourcing (Kremer et al., 2014) and contains a more extensive set of words. The main difference between LS-CIC and LS-SE07 is that the former was constructed as a large-scale "all-words" corpus, while LS-SE07 mostly includes ambiguous words.

Unlike previous work (Szarvas et al., 2013, Kremer et al., 2014, Melamud et al., 2015), we do not use any syntactic information in our models, motivated by the fact that high-quality parsers are not available for most languages. The evaluation is performed by computing the Generalized Average Precision (GAP) score (Kishida, 2005). Given a gold standard of size R of ranked candidates, the GAP score is defined as:

$$\text{GAP} = \frac{\sum_{i=1}^{n} I(x_i)p_i}{R'} \qquad R' = \sum_{i=1}^{R} I(y_i)\overline{y}_i \tag{3.8}$$

where x_i is a binary variable symbolizing whether the i-th candidate as ranked by the model is in the gold standard and n is the number of candidates to be ranked. $I(x_i)$ is

one if x_i is larger than zero, and otherwise, it is zero. $\overline{y}_i$ is the average weight of the ideal ranked list in the gold standard.

In order to rank substitution candidates, we compute the similarity between the target word and each candidate similar to Melamud et al. (2015) but adapt it to include topic distributions as well as context words for word embeddings. We run HDP on the evaluation set and compute the similarity between target word w_t and each substitution w_s using two different inference methods in line with how we incorporate topics during training. We define the first method, Sampled *(Smp)*, as:

$$\cos(\mathbf{h}(w_s^\tau), \mathbf{h}(w_t^{\tau'})) + \frac{\sum_c \cos(\mathbf{h}(w_s^\tau), \mathbf{o}(w_c))}{C}, \tag{3.9}$$

and define the second method, Expected *(Exp)*, as:

$$\sum_{\tau,\tau'} p(\tau)\, p(\tau') \cos(\mathbf{h}(w_s^\tau), \mathbf{h}(w_t^{\tau'})) + \frac{\sum_{\tau,c} \cos(\mathbf{h}(w_s^\tau), \mathbf{o}(w_c))\, p(\tau)}{C}, \tag{3.10}$$

where $p(\tau)$ and $p(\tau')$ are the topic probabilities, $\mathbf{h}(w_s^\tau)$ and $\mathbf{h}(w_t^{\tau'})$ are the representations for substitution word s with topic τ and target word t with topic τ' respectively (see Section 3.3), w_c are context words of w_t taken from a sliding window of the same size as the embeddings, $\mathbf{o}(w_c)$ is the context (i.e., output) representation of w_c, and C is the total number of context words. Note that these two methods are consistent with how we train HTLE and STLE. The *Smp* method, similar to HTLE, uses the HDP

Table 3.7: GAP scores on LS-SE07 and LS-CIC sets. For SGE + CONTEXT we use the *context* embeddings to disambiguate the substitutions. Improvements over the best baseline (MSSG) are marked $^{\blacktriangle}$ at $p < .01$ and $^{\triangle}$ at $p < .05$.

		LS-SE07			**LS-CIC**		
		Dimension			Dimension		
Model	**Infer.**	100	300	600	100	300	600
SGE		36.2	40.5	41.1	30.4	32.1	32.3
SGE + context	n/a	36.6	40.9	41.6	32.8	36.1	36.8
MSSG		37.8	41.1	42.9	33.9	37.8	39.1
HTLE		39.8$^{\blacktriangle}$	42.5$^{\blacktriangle}$	43.0$^{\blacktriangle}$	32.1	32.7	33.0
HTLEadd	Smp	39.4$^{\triangle}$	41.3$^{\blacktriangle}$	41.8	30.4	31.5	31.7
STLE		35.2	36.7	39.0	32.9	32.3	33.9
HTLE		**40.3**$^{\blacktriangle}$	**42.8**$^{\blacktriangle}$	**43.4**$^{\blacktriangle}$	36.6$^{\blacktriangle}$	**40.9**$^{\blacktriangle}$	**41.3**$^{\blacktriangle}$
HTLEadd	Exp	39.9$^{\blacktriangle}$	41.8$^{\blacktriangle}$	42.2	35.5$^{\triangle}$	37.9$^{\triangle}$	38.6
STLE		38.7$^{\triangle}$	41.0	41.0	**36.8**$^{\blacktriangle}$	36.8	37.1

model to assign topics to word occurrences during testing. The *Exp* method, similar to STLE, uses the HDP model to learn the probability distribution of topics of the

context sentence and uses the entire distribution to compute the similarity. Both of these inference methods can be used with either model.

For the context-aware Skipgram baseline (SGE + context), we compute the similarity as follows:

$$\mathrm{Sim}(w_s, w_t) = \cos(\mathbf{h}(w_s), \mathbf{h}(w_t)) + \frac{\sum_c \cos(\mathbf{h}(w_s), \mathbf{o}(w_c))}{C} \tag{3.11}$$

This computation uses the similarity between the candidate word and all words in the context, as well as the similarity between target and candidate words. We also report results for the baseline Skipgram model (SGE) without using the provided context as well as the MSSG model. The MSSG baseline uses the best performing similarity measure (Equation 3.7) as proposed by Neelakantan et al. (2014) for a context-aware comparison.

Table 3.7 shows the GAP scores of our models and the baselines. We use the nonparametric rank-based Mann-Whitney-Wilcoxon test (Sprent and Smeeton, 2016) to check for statistically significant differences between runs. We observe that all models using multiple embeddings per word perform better than SGE. Our proposed models outperform both SGE and MSSG in both evaluation sets, with more pronounced improvements in the LS-CIC data set. Note that we do not require any syntactic information and only focus on the semantic aspect of the task. We further observe that our *Exp* method is more robust and performs better for all embedding sizes. Moreover, we can see a decrease in GAP for the model variant HTLEadd compared to HTLE. By including a generic representation for each word, different topic representations drift close to each other and obtain a more general meaning of the word as well as the topic-specific meaning. Such representations are not beneficial for this task.

Table 3.8 shows the GAP scores broken down by the main word classes: noun, verb, adjective, and adverb. With 100 dimensions, our best model (HTLE) yields improvements across all POS tags, with the largest improvements for adverbs and smallest improvements for adjectives.

When increasing the dimension size of embeddings, the improvements hold up for all POS tags apart from adverbs. It can be inferred that larger dimension sizes capture semantic similarities for adverbs and context words better than other parts-of-speech categories. Additionally, we observe for both evaluation sets that the improvements are preserved when increasing the embedding size. It should also be noted that the distribution of POS tags in the test set is approximately uniform except for adverbs of which there are fewer instances.

Our findings confirm Li and Jurafsky (2015)'s observations to some extent: Higher dimension embeddings capture part of the information on semantic relations that models with multiple embeddings per word capture. In our experiments, SGE with 600 dimensions performs better than HTLE with 100 dimensions. Given a relevant semantic task, this advantage of having multiple embeddings per word can be observed more

Table 3.8: GAP scores on the candidate ranking task on LS-SE07 for different part-of-speech categories.

	Model	Noun	Verb	Adjective	Adverb
Dim=100	SGE	33.1	29.2	31.7	38.2
	SGE + context	37.2	31.6	37.1	42.2
	HTLE	**42.4**	**33.9**	**38.1**	**49.7**
	STLE	42.0	33.1	38.1	47.2
Dim=300	SGE	39.0	33.8	36.4	50.1
	SGE + context	39.2	35.0	39.0	**55.4**
	HTLE	**44.9**	**37.0**	**41.0**	50.9
	STLE	42.7	37.0	39.9	50.2
Dim=600	SGE	39.1	34.3	36.9	52.8
	SGE + context	39.7	35.7	39.9	**56.2**
	HTLE	**45.2**	**37.2**	**42.1**	51.9
	STLE	44.0	37.1	41.5	51.0

appropriately regardless of dimension.

3.5 Qualitative analysis

In this section, we discuss some properties of our representations and provide an analysis of the semantic information captured by topic-sensitive embeddings in the lexical substitution evaluation. Table 3.9 illustrates the performance of different models by providing examples from the lexical substitution task LS-SE07. In Table 3.9 (a), for the word *'bright'*, we observe that our model captures the meaning of the word in the context *('talented')* and provides a sensible substitution ranking, but the GAP scores are low. This is due to the gold substitution list being incomplete and the highest-ranking words of our model, despite matching the context, are not in the gold ranking. SGE and MSSG however, rank the candidates belonging to a different sense (*'shiny'*) higher.

Additionally, Table 3.9 provides an example of two different contexts for the same word *'rich'*. In Table 3.9 (b.1), the topic of the sentence for the word *'rich'* was learned correctly, but it is misleading for the substitution because the meaning of the word changes in the local context and SGE and MSSG perform better than our model and rank the correct substitution *'wealthy'* higher. However, for the same word (*'rich'*) in a different context, Table 3.9 (b.2), the topic-sensitive model obtains a more accurate substitution ranking and SGE fails to identify the meaning of the word (*'wealthy'*) in the context. Example (c) in Table 3.9 provides an instance for substituting the word *'fixing'* in which we achieve a higher GAP score by using SGE. Here, both MSSG

Table 3.9: Examples of word substitution rankings and respective GAP scores. The gold rank includes substitution words and annotators' votes. The models are word embeddings with context (SGE), MSSG (Neelakantan et al., 2014), and our topic-sensitive model (HTLE). Target words in the contexts and correct words in the rankings are bold.

	Substitution instance	**GAP**
(a)	During the siege, George Robertson had appointed Shuja-ul-Mulk, who was a ***bright*** boy only 12 years old and the youngest surviving son of Aman-ul-Mulk , as the ruler of chitral.	
Gold	intelligent (3), clever (3), smart (1)	
SGE	*shining, luminous, vibrant, brilliant, vivid, colourful, gleam, light, sharp,* ***smart****, . . .*	12.02
MSSG	*vivid, shining, luminous, brilliant, colourful, vibrant, gleam, sharp, light, talented, . . .*	10.34
HTLE	*brilliant, gifted, talented, capable, sharp,* ***intelligent****, clear, vivid, colourful, shining, . . .*	16.00
(b.1)	Trees on Anmyeondo used to be thick and lush to the extent of prompting a saying, "you can become ***rich*** with an axe", but now only few trees are left due to reckless deforestation since the time of Korea's liberation from Japanese colonial rule.	
Gold	wealthy (5)	
SGE	***wealthy****, abundant, vibrant, lush, abounding, lavish, ample, valuable, . . .*	100
MSSG	*lush, abundant,* ***wealthy****, vibrant, abounding, valuable, lavish, ample, . . .*	33.33
HTLE	*abundant, valuable, lush, abounding, vibrant, ample, high, significant, . . .*	7.69
(b.2)	Africas central problems in the WTO revolve around the imbalances and biases created by ***rich*** countries in the Uruguay round agreement (URA) .	
Gold	wealthy (5)	
SGE	*abundant, lush,* ***wealthy****, abounding, vibrant, valuable, ample, lavish, . . .*	33.33
MSSG	*abundant,* ***wealthy****, vibrant, lush, abounding, lavish, ample, valuable, . . .*	50
HTLE	***wealthy****, abundant, valuable, vibrant, significant, abounding, ample, . . .*	100
(c)	I feel I can get a lot more done as a selectman by being innovative and ***fixing*** the problems we have with cash flows, because they occur every year.	
Gold	resolve (2), solve (1), mend (1), repair (1)	
SGE	***resolve****,* ***mend****,* ***repair****, heal, cure, improve, correct, stick, do, patch, . . .*	79.45
MSSG	*do,* ***mend****, heal, cure,* ***resolve****,* ***repair****, correct, improve, stick, patch, . . .*	29.04
HTLE	*do, improve, heal,* ***repair****, cure,* ***resolve****,* ***mend****, determine, stick, . . .*	21.72

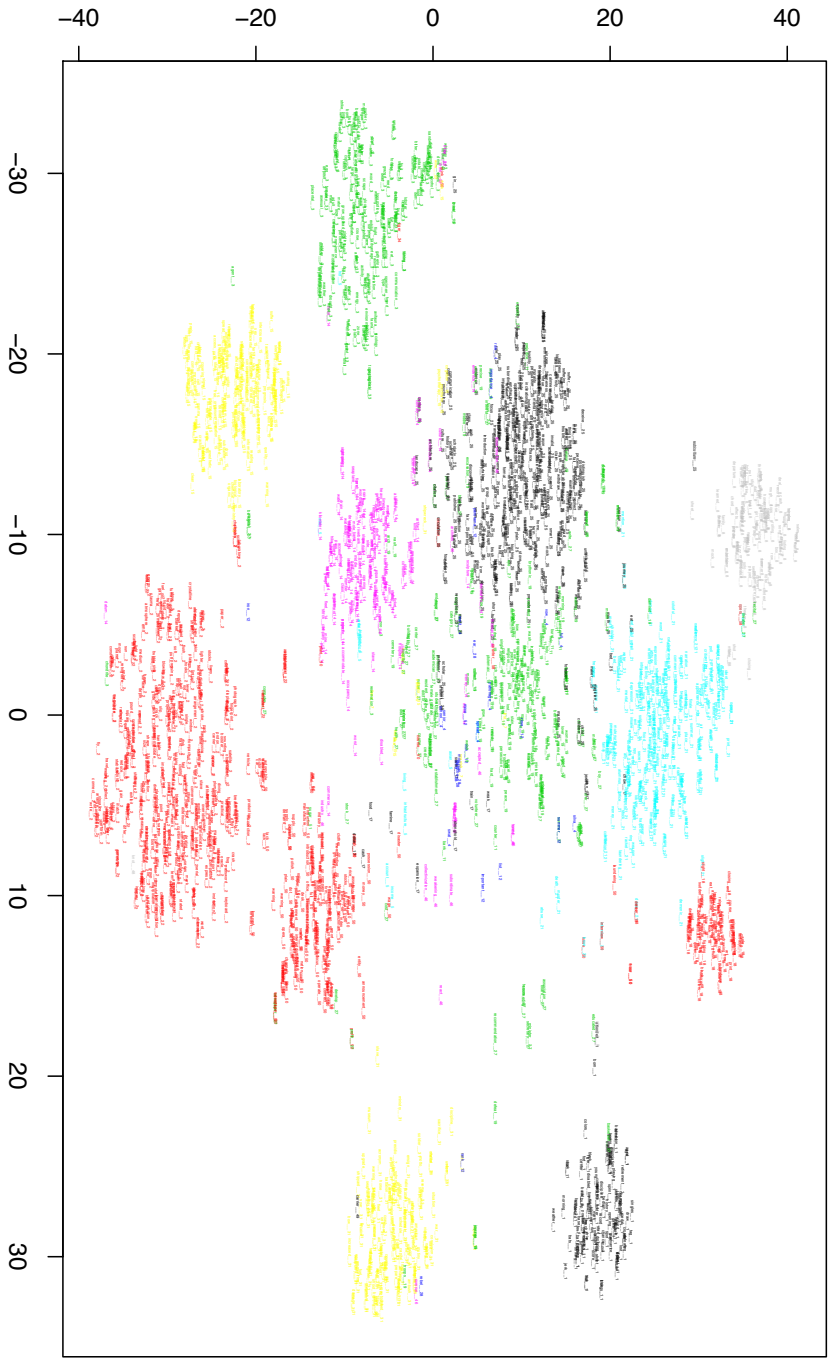


Figure 3.5: Visualizing a subset of word-topic pairs using t-SNE to showcase topic assignment separations. Colors distinguish topics. We observe that words that are labeled the same topics end up in the same clusters.

and our model detect an inaccurate sense for the context (*'heal'*) and rank the words accordingly.

These examples show that when we learn multiple embeddings per word, we can disambiguate words in context to a greater degree. Comparing the MSSG and the HTLE model, we see a slight difference in results. One distinction between the MSSG model and the HTLE model is their definition of context. The former uses a context window of length 10 (Neelakantan et al., 2014) and the latter uses a context window of length 10 as well as the document-level (in this case the complete sentence) topical information. While both models perform similarly, HTLE can be more effective when requiring larger contexts for disambiguation.

Figure 3.5 uses t-SNE (van der Maaten and Hinton, 2008) to visualize the embeddings of a subset of word-topic pairs in the vocabulary with colors distinguishing between topics. This figure gives us a basic understanding of the vector space and the distribution of topics. We observe that in general, words are closer (more similar) to other words from the same topic, rather than the same words in different topics. Additionally, we observe that in some cases, the same word with different topic labels ends up with almost overlapping embeddings. This indicates that while we assign different topic labels to a word in different documents, as long as the sense of the word is the same, the embeddings we learn turn out very similar.

3.6 Conclusion

Studying word embeddings is a good medium for getting an understanding of the impact of context in preserving word meaning. In this chapter, we have explored how document-level context can be useful to learn a more informed word representation. We asked:

RQ1.1 *To what extent can distributions over word senses be approximated by distributions over topics of documents without assuming these concepts to be identical?*

We introduced a model that uses a hierarchical Dirichlet process to learn topic distributions over documents. We observed that these distributions distinguish between senses of words. This method exploits the document-level context of words and does not require annotated data or linguistic resources. Using this information, we asked:

RQ1.2 *How can we exploit document-level topics to distinguish between different meanings of a word and learn the corresponding representations?*

We approximated the word senses with topics and further used this additional signal for training the embeddings of each topic-word pair separately. Our first model hard-labeled words with topics to learn representations. The second model

jointly learned topic-labeled and generic representations for each word in order to share statistical information between different meanings of a particular word. The third model used topic distributions for each word following the notion that meanings of words are not mutually exclusive in a given context.

Lastly, we investigated the effectiveness of these embeddings by asking:

RQ1.3 *What are the advantages of using document-level topics in learning multiple representations per word?*

When the evaluation tasks require less contextual information, the performance of our model was similar to the baselines. We evaluated word embeddings on the word similarity task and observed slight improvements under different settings. However, there was no clear indication that one model performs best across all data sets. Next, we evaluated the embeddings in a more context-aware setting. Using the SCWS data set, where context is available for word pairs, we saw that two of our models obtained improvements over the baseline. However, with higher dimensions, the MSSG model (Neelakantan et al., 2014) was the best performing system. Finally, we showed that in the lexical substitution ranking task (McCarthy and Navigli, 2007) our models outperformed two competitive baselines and performed comparably to the best-performing methods even though —unlike those methods—our approach did not use any syntactic information.

Taken together, these questions answered:

Research Question 1: *Can document-level topic distribution help infer the meaning of a word?*

Our experiments showed that we can use topic distribution over documents to improve the learning of word representations. With our proposed approach, we obtained improvements in the lexical substitution task without using any syntactic information. Our HTLE model which learns representations by hard-labeling topics to target words and learning individual embeddings achieved the best performance. We observed that topic-sensitive representations capture different senses of the words to some extent and work best when context is available.

It is worth mentioning that the methods proposed in this chapter predate more powerful neural models, such as transformers as well as complex language modeling objectives to learn dynamic contextual embeddings (Peters et al., 2018, Radford et al., 2019, Devlin et al., 2019). These models incorporate sentence-level context (and at times multiple sentences) in the computation of each word representation and have been shown to be very effective at capturing different meanings of words.

In this chapter, we studied how *static embeddings* can benefit from larger contextual cues, namely the topic of the document. As a byproduct of our models, we also learned

representations for topics and our visualization of these embeddings suggested that words belonging to the same topic are indeed clustered together. Embeddings that integrate informative priors such as topics are more interpretable (Koç et al., 2018) and can be used to advance our understanding of what word embeddings capture and represent (Hurtado Bodell et al., 2019).

4

Data Augmentation for Rare Words

4.1 Introduction and research questions

In the previous chapter, we observed the impact of context on learning word representations, in particular, modeling polysemy which is a challenging phenomenon in language. In the following chapters of this thesis, we investigate other challenges in learning word meaning. One medium to evaluate language understanding is machine translation. A machine translation system needs to understand the meaning in the source language and transfer it into the target language.

The quality of a neural machine translation system depends substantially on the availability of sizable parallel corpora. To train NMT models with reliable parameter estimations, these networks require numerous instances of sentence translation pairs with words occurring in diverse contexts, which is typically not available for low-resource language pairs. As a result, NMT falls short of producing good-quality translations for these language pairs (Zoph et al., 2016, Koehn and Knowles, 2017, Gu et al., 2018b, Ngo et al., 2019). The solution is to either revise the learning models (Östling and Tiedemann, 2017), or to provide more training data by manual annotation (Melamed, 1998) or to perform automatic data augmentation (Sennrich et al., 2016b, Wang et al., 2018). Since manual annotation of data is time-consuming, data augmentation for low-resource language pairs is a more viable approach.

In computer vision, data augmentation techniques are widely used to increase robustness and improve the learning of objects with a limited number of training examples. In image processing, the training data is augmented by, for instance, horizontally flipping, random cropping, tilting, and altering the RGB channels of the original images (Krizhevsky et al., 2012, Chatfield et al., 2014). Since the content of the new image is still the same, the label of the original image is preserved (see Figure 4.1: top). While data augmentation has become a standard technique to train deep networks for image processing, it is not common practice for training networks for NLP applications such as machine translation.

In this chapter, we address the challenge of translation of low-resource language

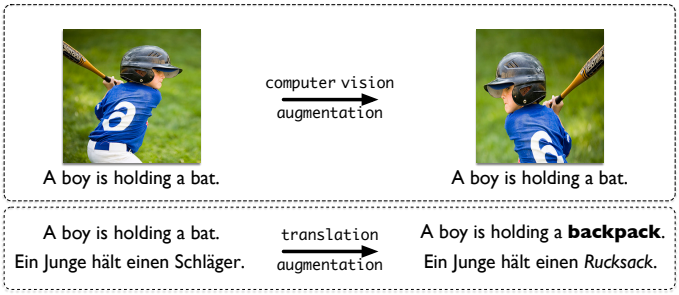


Figure 4.1: Top: flip and crop, two label-preserving data augmentation techniques in computer vision. Bottom: Altering one sentence in a parallel corpus requires changing its translation.

pairs where the primary obstacle is the lack of sufficient training data. Motivated by the success of data augmentation in computer vision, we investigate in this chapter whether NMT can benefit from data augmentation as well. Concretely, we ask:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

Research has shown that there is a strong correlation between the size of the training data and the quality of neural models (Halevy et al., 2009). To investigate this relation in machine translation, we compare how the translation and generation of a word changes by adding diverse contexts to the training data. In this chapter, we focus on low-resource language pairs, simulating a low-resource setting as done in the literature (Marton et al., 2009, Duong et al., 2015), to examine the effects of the lack of data on translation quality. In particular, we look into the translation and generation of rare words, thus asking:

RQ2.1 *How can we successfully augment the training data with diverse contexts for rare words?*

The impact of training data scarcity on translation quality is especially noticeable for rare words (Sennrich et al., 2016c). We demonstrate that parameter estimation of rare words is challenging in NMT, and it is further exacerbated in a low-resource setting. We investigate the effects of additional context, generated automatically, on both *translating* and *generating* rare words. We achieve this by proposing a simple yet effective approach that augments the training data by altering existing sentences in the parallel corpus, similar in spirit to the data augmentation approaches in computer vision (see Figure 4.1). First, we propose a weaker notion of label preservation that allows altering both source and target sentences at the same time as long as they remain translations of each other.

Next, we examine augmentation during *test time* by exploring a stronger notion of label preservation, and we ask:

RQ2.2 *Do rare words benefit from augmentation via paraphrasing during test time?*

For the augmentation process in this scenario to be possible, any change to a sentence in one language must preserve the meaning of the sentence. It is essential because we aim for not altering reference translations during evaluation. We hypothesize that it should be useful to alter the source sentence containing a rare or out-of-vocabulary word by paraphrasing it with a more common word. In addition, we also investigate the performance of different paraphrasing resources.

Organization. This chapter is organized as follows: After reviewing the previous work (Section 4.2), we present our main data augmentation model in Section 4.3. In Section 4.4, we introduce the general experimental setup, followed by a detailed description of the results of the translation experiments in Section 4.5. We analyze the effectiveness of the model further in Section 4.6. Next, we discuss augmentation at inference and propose a meaning-preserving method in Section 4.7. Finally, we conclude in Section 4.8 with an outlook of future work.

4.2 Previous work

Relevant previous work for the work described in this chapter involves two research topics: First, we briefly review the literature on image data augmentation. Second, we discuss researches studying challenges in translation of low-resource language pairs.

4.2.1 Data augmentation in computer vision

Neural models learn best when massive data is available (Halevy et al., 2009). As a result, data augmentation has become one of the staple preprocessing steps in image classification (Krizhevsky et al., 2012, Huang et al., 2019, Cubuk et al., 2019), image generation (Kynkäänniemi et al., 2019, Karras et al., 2019), and object detection (Singh et al., 2018, Liu et al., 2016). Data augmentation approaches address the overfitting problem in neural models from the perspective of the training data. Extensive research in computer vision has been done over the years on different techniques of data manipulation. There are various techniques to manipulate image data. For instance:

- **Geometric transformations:** These changes are simple alterations that are applied to the images in the training data—for instance, flipping, rotation, and cropping.
- **Mixing images:** These transformations are done by combining multiple images either from the same class or sampled from the entire training data—for instance,

averaging pixel values of multiple images or randomly cropping and patching images. Inoue (2018) show that by overlaying images and augmenting data, they achieve significant improvements in classification accuracy on CIFAR-100 data set.

Note that the augmentation techniques in computer vision have hardly any concerns about whether the image still *remains* an equivalent image after the alteration. That is not the case for augmentation of sentences, where any alteration should be mindful of generating semantically and syntactically correct sentences.

4.2.2 Low-Resource translation

Parallel data, which is the primary source of learning for machine translation models, is constructed manually and is not available in abundance for every language pair (see Section 2.1). Neural translation models especially suffer from the lack of sufficient parallel data for training. Koehn and Knowles (2017) experiment with different corpora sizes and show that their NMT system only outperforms their phrase-based machine translation system when more than 100 million words of parallel data are available. However, Sennrich and Zhang (2019) show that NMT models are highly sensitive to hyperparameters such as BPE vocabulary size. They observe strong improvements by adapting system parameters to low-resource settings.

Additionally, in a low-resource setting, the problem of translating rare words is more pronounced. Both Sutskever et al. (2014) and Bahdanau et al. (2015) observe that NMT models tend to translate sentences with many rare words more poorly than sentences containing mostly frequent words. Several recent approaches have targeted the low-resource obstacle in machine translation in different ways. Based on their approach and viewpoint on addressing this problem, current research can be categorized into four main groups:

Leveraging monolingual data We discussed several approaches to leverage monolingual data in translation in Section 2.1.1. These approaches address the problem by leveraging data resources other than the limited parallel corpora. Sennrich et al. (2016b) propose a method to back-translate sentences from monolingual data and augment the bitext with the resulting pseudo-parallel corpora. While this approach is successful in improving the translation quality, it is not effective in very low-resource settings where the back-translation model cannot be trained to a sufficient level of quality (Abdulmumin et al., 2020). This approach is further discussed in Chapter 5. Currey et al. (2017) create a parallel corpus from monolingual data in the target language by copying it so that each source sentence is identical to its corresponding target sentence. With this simple technique, they observe improvements on relatively low-resource language pairs such as English↔Turkish and English↔Romanian.

Re-designing the model These approaches target the model itself and propose changes to the standard neural translation models (Costa-jussà and Fonollosa, 2016, Sennrich et al., 2016c). Östling and Tiedemann (2017) propose to learn sentence reordering during translation of low-resource language pairs by introducing more local dependencies. They use word alignments to provide supervision to the reordering model. Lee et al. (2017) introduce a fully character-level translation model that maps a character sequence in a source language to a character sequence in a target language. They observed significant improvements in the translation of morphologically rich languages where the word-level NMT models fail to translate rare and out-of-vocabulary words. Previous approaches which propose different segmentations of the input sequence are effective; however, they present a different set of challenges: With longer sequences, the model requires information to be retained over longer temporal spans. Moreover, since the meaning of a word is not a compositional function of its characters, the model must learn to memorize many character sequences as higher-level linguistic abstractions. (Cherry et al., 2018).

Cross-lingual transfer learning These strategies use models trained on high-resource language pairs to transfer various parameters and components to the low-resource language pair. Zoph et al. (2016) propose to train a high-resource language pair first and then transfer some of the learned parameters to the low-resource pair to initialize and constrain training. Gu et al. (2018a) show that sharing lexical and sentence-level representations across multiple source languages aid in the translation of low-resource languages. They also use monolingual embeddings along with seed parallel data from all languages to build a universal representation. Cross-lingual approaches are particularly valuable for multilingual translation learning, where a single NMT model learns to translate between multiple languages (Firat et al., 2016, Johnson et al., 2017, Blackwood et al., 2018, Aharoni et al., 2019). While these approaches are very impressive in translating between language pairs not seen during training, this paradigm cannot outperform the individual models trained on bilingual corpus in many cases (Johnson et al., 2017).

Unsupervised learning These studies focus on zero-resource learning, where there are no parallel corpora available for a language pair (Yang et al., 2018, Artetxe et al., 2018a,b, 2019). Lample et al. (2018a) propose a model that takes sentences from monolingual data in two different languages and maps them into the same latent space. The model learns to translate without using any parallel data by reconstructing both languages from the shared feature space. Lample et al. (2018b) address the challenge of only having access to monolingual corpora in each language. They use a smoothed n-gram language model (phrase-based model) as a data-driven prior to denoising sentences and automatically generate the parallel data by iterative back-translation (neural model). These approaches are effective; however, the pseudo sentences used for training are usually of low quality as translation mistakes accumulate during training. Additionally,

while they perform well between languages that are from the same branch, they perform poorly between distant languages (Sun et al., 2020).

4.3 Data augmentation for rare words

In this section, we propose a novel approach for data augmentation of parallel corpora. Specifically, we use a Bidirectional RNN model trained on monolingual data to introduce completely new contexts for rare words in the bitext. In our approach we use sentences from the training data as starting points and use the probability distribution of the output layer of an RNN to insert words into new contexts: Given a source and target sentence pair (S, T), we want to alter it in a way that we obtain new contexts for these words while diversifying as much as possible the training examples. A number of ways to do this can be envisaged, as for example paraphrasing (parts of) S or T, or altering both and preserve the semantic equivalence between S and T. We explore both approaches in this chapter.

We choose to focus on a subset of the vocabulary that we know to be poorly modeled by our baseline NMT system, namely words that occur rarely in the parallel corpus. Thus, the goal of our data augmentation technique is to provide novel contexts for rare words. To achieve this we search for contexts where a common word can be replaced by a rare word and consequently replace its corresponding word in the other language by that rare word's translation:

original pair	**augmented pair**
$S : s_1, \ldots, s_i, \ldots, s_n$	$S' : s_1, \ldots, s'_i, \ldots, s_n$
$T : t_1, \ldots, t_j, \ldots, t_m$	$T' : t_1, \ldots, t'_j, \ldots, t_m$

where t_j is a translation of s_i and word-aligned to s_i, and t'_j is the translation of s'_i. Plausible substitutions are those that result in a fluent and grammatical sentence but do not necessarily maintain its semantic content. As an example, the rare word *motorbike* can be substituted in different contexts:

Sentence [original **substituted**]	Plausible
My sister drives a [car **motorbike**]	yes
My uncle sold his [house **motorbike**]	yes
Alice waters the [plant **motorbike**]	no (semantics)
John bought two [shirts **motorbike**]	no (syntax)

Implausible substitutions need to be ruled out during data augmentation. To this end, rather than relying on linguistic resources which are not available for many languages, we rely on LSTM language models (LM) trained on large amounts of monolingual data in both forward and backward directions.

Our data augmentation method involves the following steps:

Targeted words selection: Following common practice, our NMT system limits its vocabulary V to the v most common words observed in the training corpus. We select the words in V that have fewer than R occurrences and use this as our targeted rare word list V_R.

Rare word substitution: If the LM suggests a rare substitution in a particular context, we replace that word and add the new sentence to the training data. Formally, given a sentence pair (S, T) and a position i in S, we compute the probability distribution over V by the forward and backward LMs and select rare word substitutions $\mathcal{C}$ as follows:

$$\overrightarrow{\mathcal{C}} = \{s'_i \in V_R : \text{topK}\, P_{\textit{ForwardLM}_S}(s'_i \mid s_1^{i-1})\} \tag{4.1}$$

$$\overleftarrow{\mathcal{C}} = \{s'_i \in V_R : \text{topK}\, P_{\textit{BackwardLM}_S}(s'_i \mid s_n^{i+1})\} \tag{4.2}$$

$$\mathcal{C} = \{s'_i \mid s'_i \in \overrightarrow{\mathcal{C}} \wedge s'_i \in \overleftarrow{\mathcal{C}}\} \tag{4.3}$$

where s_i^j are context words from position i to j and topK returns the K words with highest conditional probability according to the context. The selected substitutions s'_i, are used to replace the original word and generate a new sentence.

Translation selection: Using automatic word alignments[1] trained over the bitext, we replace the translation of word s_i in T by the translation of its substitution s'_i. Following a common practice in statistical MT (Koehn et al., 2007), the optimal translation t'_j is chosen by multiplying direct and inverse lexical translation probabilities with the LM probabilities of the translation in context:

$$t'_j = \underset{t \in \textit{trans}(s'_i)}{\arg\max}\, P(s'_i \mid t) P(t \mid s'_i) P_{\textit{ForwardLM}_T}(t \mid t_1^{j-1}) P_{\textit{BackwardLM}_T}(t \mid t_n^{j+1}) \tag{4.4}$$

If no translation candidate is found because the word is unaligned or because the language model probability is less than a certain threshold, the augmented sentence is discarded. This reduces the risk of generating sentence pairs that are semantically or syntactically incorrect.

We use the described steps of targeting and substituting words in source and target sentences to augment the training data:

Sampling: We loop over the original parallel corpus multiple times, sampling substitution positions i in each sentence and making sure that each rare word gets augmented at most N times so that a large number of rare words can be affected. We stop when no new sentences are generated in one pass over the training data.

[1] We use fast-align (Dyer et al., 2013) to extract word alignments and a bilingual lexicon with lexical translation probabilities from the low-resource bitext.

Augmentation: Assuming the original parallel data is $\mathcal{D}$, our data augmentation process can be represented by the following mapping:

$$\phi : \mathcal{D} \mapsto \mathcal{A} \tag{4.5}$$

where $\mathcal{A}$ is the modified set with new contexts for rare words built from sentence pairs in $\mathcal{D}$. Note that some sentences in $\mathcal{D}$ may not be augmented because of the randomness of the sampling step and shortage of substitution suggestions from the language model. As the final step, the training data is expanded as the union of the original data and the augmented data:

$$\mathcal{D}' = \mathcal{D} \cup \mathcal{A} \tag{4.6}$$

Our proposed method is demonstrated in Figure 4.2 with an example: Given the English sentence '*I had been told that you would not be speaking today.*', we randomly sample the position of the word '*not*' and explore the suggestions of the language model that fit the context. We select the word '*voluntarily*' because it is a rare word in the low-resource setting of our experiments and the English language model has high confidence in substituting it in the sentence. Correspondingly, we explore the translation candidates of the word '*voluntarily*' to make comparable changes to the German sentence. From the translation candidates, we choose the word that yields the

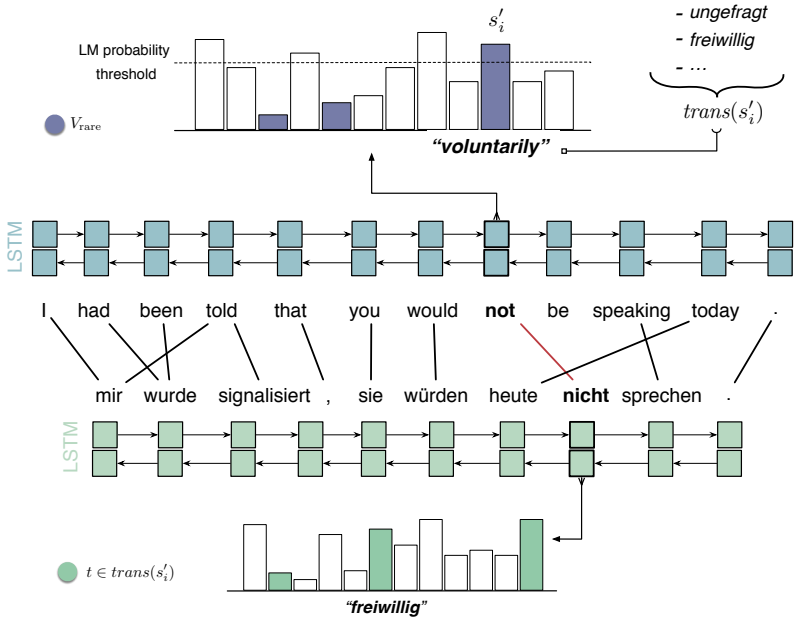


Figure 4.2: A visual representation of our proposed mechanism for generating new sentence pairs.

most fluent sentence according to the German language model. Therefore the word '*freiwillig*' is selected to substitute '*nicht*' in the German sentence. The newly generated sentence pairs are then added to the training data.

Table 4.1 provides several examples resulting from our augmentation procedure. While using a large LM to substitute words with rare words mostly results in grammatical sentences, this does not mean that the meaning of the original sentence is always preserved. Note that meaning preservation is not an objective of the proposed approach in this section and we will explore this further in Section 4.7.

Table 4.1: Examples of augmented data with highlighted Original words and **substituted** words.

	Original sentence pair	**Synthetic sentence pair**
(a)	SRC: I had been told that you would not be speaking today. TGT: mir wurde signalisiert, sie würden heute nicht sprechen.	SRC: I had been told that you would **voluntarily** be speaking today. TGT: mir wurde signalisiert, sie würden heute **freiwillig** sprechen.
(b)	SRC: the present situation is indefensible and completely unacceptable to the commission. TGT: die situation sei unhaltbar und für die kommission gänzlich unannehmbar.	SRC: the present situation is **confusing** and completely unacceptable to the commission. TGT: die situation sei **verwirrend** und für die kommission gänzlich unannehmbar.
(c)	SRC: ...agree wholeheartedly with the institution of an ad hoc delegation of parliament on the turkish prison system. TGT: ...ad-hoc delegation des parlaments für das regime in den türkischen gefängnissen voll und ganz zustimmen.	SRC: ...agree wholeheartedly with the institution of an ad hoc delegation of parliament on the turkish **missile** system. TGT: ...ad-hoc delegation des parlaments für das regime in den türkischen **flugwaffen** voll und ganz zustimmen.
(d)	SRC: cancellation fees are not subject to judiciary mitigation. TGT: stornogebühren unterliegen nicht dem richterlichen mäßigungsrecht.	SRC: cancellation fees are **generally** subject to **western** mitigation. TGT: stornogebühren unterliegen **allgemein** dem **westlichen** mäßigungsrecht.

In our experiments, two translation data augmentation (TDA) setups are considered: only one word per sentence can be replaced ($\text{TDA}_{r=1}$) or multiple words per sentence can be replaced, with the condition that any two replaced words are at least five positions apart ($\text{TDA}_{r\geqslant 1}$). The latter incurs a higher risk of introducing noisy sentences but has the potential to positively affect more rare words within the same amount of augmented data. We evaluate both setups in the following section.

4.4 Data and experimental setup

In this section, we describe the experimental settings. To simulate a low-resource setting we randomly sample 10% of the English↔German WMT15 training data and report results on newstest 2014, 2015, and 2016 (Bojar et al., 2016). For reference, we also provide the result of our baseline system on the full data.

As NMT system, we use a 4-layer attention-based encoder-decoder model as described in Section 2.4 trained with hidden dimension 1000, and batch size 80 for 20 epochs. NMT models often limit their vocabularies to be the top K most frequent words in each language because of the computationally intensive nature of the softmax. In all experiments, the NMT vocabulary is limited to the 30K most common words in both languages. Note that our proposed data augmentation method does not introduce new words to the vocabulary. In all experiments, we preprocess source and target language data with Byte Pair Encoding (BPE) (Sennrich et al., 2016c) using 30K merge operations. In the non-label-preserving augmentation experiments, BPE is performed after data augmentation.

For the LMs needed for data augmentation, we train 2-layer LSTM networks in forward and backward directions on the monolingual data provided for the same task (3.5B and 0.9B tokens in English and German, respectively) with embedding size 64 and hidden size 128. We set the rare word threshold R to 100, and top K words to 1000. These values are determined heuristically from the training data. Another question we want to investigate is whether rare word substitution is more effective in the source or the target language. Therefore in the experiments, we augment the source side in English→German, and target side in German→English translation. In all experiments, we use the English LM for the rare word substitutions. Since our first approach is not label-preserving, we only perform augmentation during training and do not alter source sentences during testing. In Section 4.7, we also alter source sentences during testing while preserving labels.

We also compare our approach to Sennrich et al. (2016b) by back-translating monolingual data and adding it to the parallel training data. Specifically, we back-translate sentences from the target side that are not included in our low-resource baseline with two settings: keeping a one-to-one ratio of back-translated versus original data $(1:1)$ following the authors' suggestion, or using three times more back-translated data $(1:3)$. We measure translation quality by single-reference case-insensitive BLEU (Papineni et al., 2002) computed with the `multi-bleu.perl` script from Moses.

4.5 Results

In this section, we discuss the results on the translation task and evaluate the effectiveness of our approach in a simulated low-resource NMT scenario. We repeat the sampling and substitution step iteratively until we reach the desired corpus size for each

Table 4.2: Translation performance (BLEU) on German-English WMT test sets (2014, 2015, and 2016) in a simulated low-resource setting. Back-translation refers to the work of Sennrich et al. (2016b). Statistically significant improvements are marked $^{\blacktriangle}$ at the $p < .01$ and $^{\triangle}$ at the $p < .05$ level, with the first superscript referring to baseline and the second to back-translation$_{1:1}$.

		De-En					
Model	**Data**	**WMT14**		**WMT15**		**WMT16**	
Full data (ceiling)	3.9M	21.1		22.0		26.9	
Baseline	371K	10.6		11.3		13.1	
Back-translation$_{1:1}$	731K	11.4	(+0.8)$^{\blacktriangle}$	12.2	(+0.9)$^{\blacktriangle}$	14.6	(+1.5)$^{\blacktriangle}$
Back-translation$_{1:3}$	1.5M	11.2	(+0.6)	11.2	(–0.1)	13.3	(+0.2)
TDA$_{r=1}$	4.5M	11.9	(+1.3)$^{\blacktriangle,-}$	13.4	(+2.1)$^{\blacktriangle,\blacktriangle}$	15.2	(+2.1)$^{\blacktriangle,\blacktriangle}$
TDA$_{r\geqslant 1}$	6M	**12.6**	(+2.0)$^{\blacktriangle,\blacktriangle}$	**13.7**	(+2.4)$^{\blacktriangle,\blacktriangle}$	**15.4**	(+2.3)$^{\blacktriangle,\blacktriangle}$
Oversampling	6M	11.9	(+1.3)$^{\blacktriangle,-}$	12.9	(+1.6)$^{\blacktriangle,\triangle}$	15.0	(+1.9)$^{\blacktriangle,-}$

Table 4.3: Translation performance (BLEU) on English-German WMT test sets (2014, 2015, and 2016) in a simulated low-resource setting. Back-translation refers to the work of Sennrich et al. (2016b). Statistically significant improvements are marked $^{\blacktriangle}$ at the $p < .01$ and $^{\triangle}$ at the $p < .05$ level, with the first superscript referring to baseline and the second to back-translation$_{1:1}$.

		En-De					
Model	**Data**	**WMT14**		**WMT15**		**WMT16**	
Full data (ceiling)	3.9M	17.0		18.5		21.7	
Baseline	371K	8.2		9.2		11.0	
Back-translation$_{1:1}$	731K	9.0	(+0.8)$^{\blacktriangle}$	10.4	(+1.2)$^{\blacktriangle}$	12.0	(+1.0)$^{\blacktriangle}$
Back-translation$_{1:3}$	1.5M	7.8	(–0.4)	9.4	(+0.2)	10.7	(–0.3)
TDA$_{r=1}$	4.5M	10.4	(+2.2)$^{\blacktriangle,\blacktriangle}$	11.2	(+2.0)$^{\blacktriangle,\blacktriangle}$	13.5	(+2.5)$^{\blacktriangle,\blacktriangle}$
TDA$_{r\geqslant 1}$	6M	**10.7**	(+2.5)$^{\blacktriangle,\blacktriangle}$	**11.5**	(+2.3)$^{\blacktriangle,\blacktriangle}$	**13.9**	(+2.9)$^{\blacktriangle,\blacktriangle}$
Oversampling	6M	9.7	(+1.5)$^{\blacktriangle,\triangle}$	10.7	(+1.5)$^{\blacktriangle,-}$	12.6	(+1.6)$^{\blacktriangle,-}$

experiment. In our various experiments, we successfully augment between 72% to 81% of targeted rare words. All translation results are displayed in Table 4.2 and Table 4.3 for German→English and English→German experiments, respectively.

First, we observe that the low-resource baseline performs much worse than the full data system, re-iterating the importance of sizable training data for NMT. Next, we observe that both back-translation and our proposed TDA method significantly improve translation quality. However, TDA obtains the best results overall and significantly outperforms back-translation for all test sets. This is an important finding considering that our method involves only minor modifications to the original training sentences and does not involve any costly translation process, while the back-translation approach augments with novel target sentences. Improvements are consistent across both translation directions, regardless of whether rare word substitutions are applied to the source or to the target side. We also observe that altering multiple words in a sentence performs slightly better than altering only one word. This indicates that addressing more rare words is preferable even though the augmented sentences are more likely to be noisy.

To verify that the gains are actually due to the rare word substitutions and not just to the repetition of part of the training data, we perform a final experiment where each sentence pair selected for augmentation is added to the training data *unchanged* (Oversampling row in Tables 4.2 and 4.3). Surprisingly, we find that this simple form of sampled data replication outperforms both baseline and back-translation systems,[2] while $\text{TDA}_{r\geqslant 1}$ remains the best performing system overall.

Table 4.4: Average length of German→English translation systems, along with the average length of human reference translations (bottom line). Predominantly, we favour longer translations that are close to human reference translations, i.e., models with higher % Ref ratio.

	De-En			
	WMT14	**WMT15**	**WMT16**	**%Ref**
Baseline	19.9	19.2	19.9	0.88
$\text{TDA}_{r=1}$	21.4	20.4	21.2	**0.94**
$\text{TDA}_{r\geqslant 1}$	21.0	20.0	20.8	0.92
Reference	23.0	22.2	21.9	1.00

We also observe that the system trained on our augmented data tends to generate longer translations, which is favoured. Tables 4.4 and 4.5 provide the average length of the translation outputs of different systems, along with the average length of human reference translations. Averaging over all test sets and language pairs, the length of translations generated by the baseline is 0.89 of the average reference length, while for $\text{TDA}_{r=1}$ and $\text{TDA}_{r\geqslant 1}$ it is 0.94 and 0.93, respectively. We attribute this effect to the

[2]Note that this effect cannot be achieved by simply continuing the baseline training for up to 50 epochs.

Table 4.5: Average length of English→German translation systems, along with the average length of human reference translations (bottom line). Predominantly, we favour longer translations that are close to human reference translations, i.e., models with higher % Ref ratio.

	En-De			
	WMT14	**WMT15**	**WMT16**	**%Ref**
Baseline	19.8	19.4	18.9	0.91
$\text{TDA}_{r=1}$	20.5	20.2	19.9	**0.95**
$\text{TDA}_{r\geqslant 1}$	20.7	20.0	19.8	**0.95**
Reference	21.4	20.8	21.6	1.00

ability of the TDA-trained system to generate translations for rare words that were left untranslated by the baseline system.

4.6 Further analysis

In this section, we further analyze our findings and discuss the results of our proposed models. Our goal is to understand the impact of the introduced diverse contexts on the learning capabilities of the neural translation model.

4.6.1 Target words

A desired effect of our method is to increase the number of correct rare words generated by the NMT system at test time. To illustrate the impact of augmenting the training data by creating contexts for rare words on the *target* side, Table 4.6 provides an example for German→English translation.

We see that the baseline model is not able to generate the rare word *'centimetres'* as a correct translation of the German word *'zentimeter'*. However, this word is not rare in the training data of the $\text{TDA}_{r\geqslant 1}$ model after our augmentation and is generated during translation. Table 4.6 also provides several instances of augmented training sentences targeting the word *'centimetres'*. Note that even though some augmented sentences are rather unusual (e.g., *'the speed limit is five centimetres per hour'*), the NMT system still benefits from the new context for the rare word and is able to generate it during testing.

Figure 4.3 demonstrates that this is indeed the case for many words: the number of rare words occurring in the reference translation ($V_R \cap V_{ref}$) is three times larger in the TDA system output than in the baseline output. One can also see that this increase is a direct effect of TDA as most of the rare words are not 'rare' anymore in the augmented data, i.e., they were augmented sufficiently often to occur more than 100 times (see the hatched pattern in Figure 4.3). Note that in our experiments, we did not use any

Table 4.6: An example from WMT14 illustrating the effect of augmenting rare words on generation at test time. The translation of the baseline does not include the rare word *centimetres*, however, the translation of our TDA model generates the rare word and produces a more fluent sentence. Instances of the augmentation of the word *centimetres* in training data are also provided.

Example from WMT14

SRC	der tunnel hat einen querschnitt von 1,20 meter höhe und 90 zentimeter breite.
REF	the tunnel has a cross-section measuring 1.20 metres high and 90 centimetres across.
Baseline	the wine consists of about 1,20 m and 90 of the canal.
$\text{TDA}_{r \geqslant 1}$	the tunnel has a `unk` measuring meters 1.20 metres high and 90 **centimetres** wide.

*Examples from the training data displaying augmentations for the word **'centimetres'***

	Original data	Augmented data
(a)	the average speed of cars and buses is therefore around 20 kilometres per hour.	the average speed of cars and buses is therefore around 20 **centimetres** per hour.
(b)	grab crane in special terminals for handling capacities of up to 1,800 tonnes per hour.	grab crane in special terminals for handling capacities of up to 1,800 **centimetres** per hour.
(c)	all suites and rooms are very spacious and measure between 50 and 70 m.	all suites and rooms are very spacious and measure between 50 and 70 **centimetres**.
(d)	all we have to do is lower the speed limit everywhere to five kilometers per hour.	all we have to do is lower the speed limit everywhere to five **centimetres** per hour.
(e)	just 9.5 litres of water per minute flow through the innovative `unk` shower system, whereas the hansgrohe `unk` 85 green hand spray manages with only six litres per minute.	just 9.5 litres of water per minute flow through the innovative `unk` shower system, whereas the hansgrohe `unk` 85 green hand spray manages with only six **centimetres** per minute.

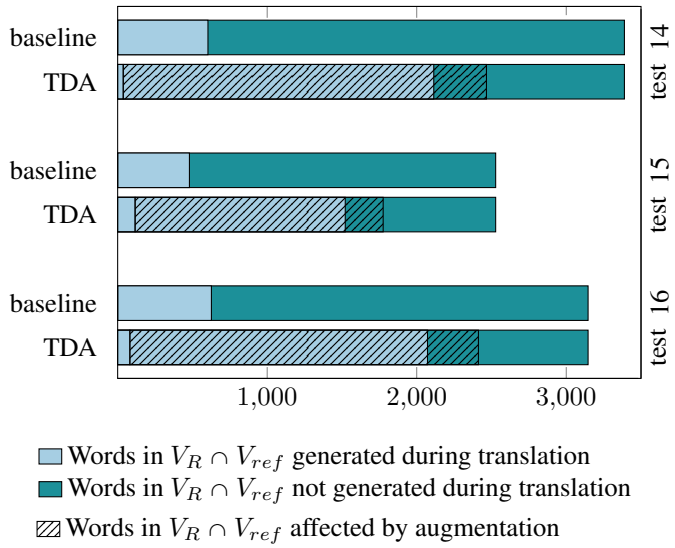


Figure 4.3: Effect of TDA on the number of unique rare words generated during De→En translation. V_R is the set of rare words targeted by $\text{TDA}_{r \geqslant 1}$ and V_{ref} the reference translation vocabulary.

information from the evaluation sets.

4.6.2 Source words

To gauge the impact of augmenting the contexts for rare words on the *source* side, we examine normalized attention scores of these words before and after augmentation. When translating English→German with our TDA model, the attention scores for rare words on the source side are on average 8.8% higher than when translating with the baseline model. This suggests that having more accurate representations of rare words increases the model's confidence to attend to these words when encountered during test time.

4.6.3 Negative examples

Table 4.7 provides examples of cases where augmentation results in incorrect sentences. In the first example, the sentence is ungrammatical after substitution (*'of / yearly'*), which can be the result of choosing substitutions with low probabilities from the English LM topK suggestions.

Errors can also occur during translation selection, as in the second example where *'betraut'* is an acceptable translation of *'entrusted'* but would require a rephrasing of the German sentence to be grammatically correct. Problems of this kind can be attributed to the German LM, but also to the lack of a more suitable translation in the lexicon

Table 4.7: Examples of incorrectly augmented data with highlighted Original words and **substituted** words.

	Original sentence pair	Synthetic sentence pair
(a)	SRC: registered users will receive the unk newsletter free of charge. TGT: registrierte user erhalten zudem regelmäßig den markenticker newsletter.	SRC: registered users will receive the unk newsletter free **yearly** charge. TGT: registrierte user erhalten zudem regelmäßig **jährlich** markenticker newsletter.
PROBLEM:	*Substitution results in syntactically incorrect source and target sentences.*	
(b)	SRC: the personal contact is essential to us. TGT: persönliche kontakt ist uns sehr wichtig.	SRC: the personal contact is **entrusted** to us. TGT: persönliche kontakt ist uns sehr **betraut**.
PROBLEM:	*The German sentence is grammatically incorrect.*	
(c)	SRC: unk unk wishes you very pleasant holiday. TGT: unk unk wünscht ihnen einen erholsamen urlaub.	SRC: unk unk wishes you very **crazy** holiday. TGT: unk unk wünscht ihnen einen **verrückt** urlaub.
PROBLEM:	*German substitution has the wrong inflection.*	
(d)	SRC: consumers are currently being deliberately misled. TGT: die verbraucher werden gegenwärtig bewusst getäuscht.	SRC: **schools** are currently being **widely** misled. TGT: **schulen** verbraucher werden gegenwärtig **weithin** getäuscht.
PROBLEM:	*Substituted with the wrong German word because of wrong alignment.*	

extracted from the bitext. Interestingly, this noise seems to affect NMT only to a limited extent.

4.6.4 Word segmentation

BPE (Sennrich et al., 2016c) is an essential preprocessing step in NMT to address the problem of rare and unknown words in the training data and we use it in all experiments. Although crucial, it is not very effective in a low-resource setting (Ngo et al., 2019). We suspect that this is caused by the scarcity of data, which results in inaccurate word splits with possibly rare subword units. This can be observed in the example in Table 4.6. In the experiments, the English and German words *'centi|metres'* and *'zenti|meter'* are both split into two subword units. Still, the baseline model fails to translate it correctly. This further stresses the importance of data augmentation with diverse contexts. Even though BPE segmentation is successful in rare word translations, our proposed approach yields additional improvements.

4.7 Meaning-Preserving augmentation

In the previous section, we proposed a model with a weak notion of label preservation that allows modifying both source and target sentences at the same time as long as they remain translations of each other. As a result, we alter and augment the training data, and the test data remains unchanged. This approach improves translation quality by better translating rare words because of the additional contexts during training. However, it does not address the problem of translating out-of-vocabulary (OOV) words during testing. To specifically target OOV words at test time, we propose a stronger notion of label preservation to only alter source sentences with paraphrases. Note that at test time, we only have access to source sentences and keep the reference sentences unchanged.

In this section, we investigate how we can benefit from external lexical resources to address the problem of translating unknown words. We define OOV as words not listed in the 30k most common words in source and target vocabulary. While BPE is an effective approach in addressing the OOV translation problem, we do not use it in these experiments. To benefit from external knowledge resources, we substitute OOVs with synonym words obtained from these resources that exist in our vocabulary. We experiment with three different resources to alter source sentences with paraphrases:

PPDB proposed by Ganitkevitch et al. (2013). This Paraphrase Database is an automatically extracted database from parallel corpora containing millions of paraphrases in multiple languages.

Wordnet proposed by Miller (1995). WordNet is a manually created large lexical database of English and includes relations between words and groups of cognitive

synonyms. We use GermaNet (Hamp and Feldweg, 1997, Henrich and Hinrichs, 2010) which is a lexical-semantic resource similar to Wordnet for the German language.

CBOW embeddings proposed by Mikolov et al. (2013b). We use the Continuous Bag of Words (CBOW) model to identify the words most similar to each OOV word and interpret that as the synonym.

HTLE embeddings proposed in Chapter 3. We use the multiple topic-sensitive representations per word to identify synonyms according to the context of the word. In contrast to the previous resources, this method provides context-sensitive synonyms which means the same word in different contexts has different synonym substitutions.

We substitute the OOV words in source sentences with synonyms that already exist in the NMT vocabulary. An example of the substitution is shown in Table 4.8 where the original word '*fateful*' is OOV and is replaced by the `unk` symbol in the original training data. Each paraphrase resource suggests a substitution for the target word in the sentence. We also experiment with targeting low-frequency words in the test data. Similar to OOVs, we substitute words that are rare in our training data (frequencies less than 100) with synonyms during test time.

original src	He said Lamb made the *fateful* 911 call sometime after that.
NMT input	He said Lamb made the `[unk]` 911 call sometime after that.
+PPDB	He said Lamb made the **disastrous** 911 call sometime after that.
+Wordnet	He said Lamb made the **fatal** 911 call sometime after that.
+CBOW	He said Lamb made the **tragic** 911 call sometime after that.
+HTLE	He said Lamb made the **critical** 911 call sometime after that.

Table 4.8: Examples of paraphrase modification. The out-of-vocabulary word *fateful* in the source sentence is substituted with synonyms obtained from different lexicon resources.

Tables 4.9 and 4.10 provide the results for the translation of German→English and English→German, respectively. Note that the HTLE embeddings are available only in English and so we only use this approach in the English→German translation experiments to augment English sentences on the source side. Overall the results show improvements over the baseline. BLEU scores using PPDB and Wordnet are very similar for all experiments. Improvements using CBOW and HTLE are also similar, however, HTLE is the most effective method out of the four approaches. This indicates, to a certain degree, the importance of context-aware substitutions in data augmentation.

Finally, we look into the rate a neural model generates the `unk` symbol before and after augmentation. When a source sentence contains several rare and OOV words, the translation model tends to use the `unk` symbol to represent these words. As a

Table 4.9: Translation performance (BLEU) on German-English WMT news test sets (2014, 2015, and 2016). `OOV` signifies out-of-vocabulary words. `rare` words are selected with the frequency threshold of 100.

			De-En		
Model	**+lexical DB**	**Target**	**WMT14**	**WMT15**	**WMT16**
Baseline			19.3	20.1	24.9
Subs	PPDB	OOV	21.2	22.2	26.9
	GermaNet	OOV	20.3	21.9	25.2
	CBOW	OOV	21.2	22.4	27.0
Subs	PPDB	OOV + rare	21.3	22.3	26.9
	GermaNet	OOV + rare	20.5	22.0	25.4
	CBOW	OOV + rare	**21.4**	**22.5**	**27.2**

Table 4.10: Translation performance (BLEU) on English-German WMT news test sets (2014, 2015, and 2016). `OOV` signifies out-of-vocabulary words. `rare` words are selected with the frequency threshold of 100.

			En-De		
Model	**+lexical DB**	**Target**	**WMT14**	**WMT15**	**WMT16**
Baseline			15.9	17.6	20.0
Subs	PPDB	OOV	17.2	18.5	21.8
	Wordnet	OOV	17.2	18.5	21.7
	CBOW	OOV	17.3	18.6	22.0
	HTLE	OOV	17.5	18.6	22.2
Subs	PPDB	OOV + rare	17.2	18.7	21.9
	Wordnet	OOV + rare	17.1	18.5	21.8
	CBOW	OOV + rare	17.4	18.7	22.2
	HTLE	OOV + rare	**17.9**	**19.1**	**22.5**

result, the model performs poorly and produces several `unk` symbols in the translation output. We investigate the fluency of the translation outputs by observing the rate of the generation of the `unk` symbol. We observe that in our experiments the number of `unk` symbols generated in the translation output drops. Table 4.11 provides statistics on the generation of `unk` token. Surprisingly, the significant differences in the number of `unk` symbols in the translation outputs are not entirely reflected in the BLEU scores.

Table 4.11: The impact of paraphrase augmentation on the generation of `unk` tokens in the translation output. Reductions are computed in comparison with the baseline model. Lower number of `unks` is better.

	Baseline	PPDB	Wordnet	CBOW	HTLE
Number of `unks`	4931	4851	4857	3018	**3003**
Reduction in number of `unks`	-	1.62%	1.5%	38.8%	**39.1**%

4.8 Conclusion

In this chapter, we investigated the impact of diverse contexts on the translation of rare words. The quality of an NMT system depends substantially on the availability of sizable parallel corpora, which is only available for a limited number of languages and domains. The translation is particularly erroneous for low-frequency words; with only a few instances in the training data, the model has difficulties learning to translate these words. While the challenges of translating low-resource language pairs have been studied extensively, the impact of artificially generated contexts on the translation quality of words has hardly been studied. We addressed this issue in this chapter and investigated the effect of additional context in learning word translations, by asking:

RQ2.1 *How can we successfully augment the training data with diverse contexts for rare words?*

Our experiments showed that by providing more diverse contexts for rare words, we improve the estimation of the model and subsequently increase the number of times the model generates these words correctly. We have proposed an effective approach to augment the training data of neural machine translation for low-resource language pairs. We generated new sentence pairs containing rare words in new contexts by leveraging language models trained on large amounts of monolingual data. Our approach augments the data by diversifying the sentences of the parallel corpora, changing both source and target sentences. We showed that this approach leads to generating rare words more often during translation and thus improves translation quality. We observed substantial improvements in simulated low-resource English→German and German→English settings.

Having observed the impact of additional *training* data on the translation of rare words, we looked into how we can perform augmentation during *testing* and asked:

RQ2.2 *Do rare words benefit from augmentation via paraphrasing during test time?*

To answer this question, we first explained why our previously proposed method is not viable at test time. We do not have access to the reference translations during inference and as a result we only accept alterations to the source sentence that

keep the meaning of the sentence unchanged. We introduced a substitution method to replace both rare and out-of-vocabulary words in the source sentences with their paraphrases, using several knowledge resources. We gained improvements in BLEU scores over the baselines. However more interestingly, we significantly reduced the number of `unk` in the target output.

In summary, our extensive studies of the rare word translation challenge partially answered the following question:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

To answer this question, we examined the effect of the availability of data, and rare words in particular, on translation quality. We found that translating and generating rare words is a challenging task for NMT models. With the proposed data augmentation approach, we diversified and increased the contexts of rare words. We improved the translation quality by augmenting the data with these new sentence pairs.

In this chapter, we looked into the long tail of words where statistical models have difficulties learning. We continue our investigation into this question in the next chapter by examining whether the trained model itself can identify words that will benefit from the addition of diverse contexts.

5 Data Augmentation Based on Model Failure

5.1 Introduction and research questions

In the previous chapter, we have observed that the availability of large-scale training data is essential for sequence-to-sequence neural models to achieve good translation quality. We have shown that neural machine translation models benefit from data augmentation for rare words. By using a combination of parallel and synthetic data, neural models learn to translate more effectively. In this chapter, we study a more general approach to augmentation by identifying and targeting words that are most difficult to learn by the model.

Previous approaches have focused on leveraging monolingual data, which is available in much larger quantities than parallel data (Lambert et al., 2011). Sennrich et al. (2016b) proposed back-translation of monolingual target sentences to the source language and adding the synthetic sentences to the parallel data (discussed in Section 2.1.1). In this approach, a reverse model trained on parallel data is used to translate sentences randomly sampled from target-side monolingual data into the source language. This synthetic parallel data is then used in combination with the actual parallel data to re-train the model. This approach yields state-of-the-art results even when large amounts of parallel data are already available and has become common practice in NMT (Sennrich et al., 2017, García-Martínez et al., 2017, Ha et al., 2017). Generally speaking, back-translation mitigates the problem of overfitting and fluency by exploiting additional data in the target language (Sennrich et al., 2016b).

An important question for this technique is how to select the monolingual data in the target language that is to be back-translated into the source language in order to obtain the best possible translation quality. Earlier studies have explored to what extent data selection of parallel corpora can benefit translation quality (Axelrod et al., 2011, van der Wees et al., 2017), but such selection techniques have not been investigated in the specific context of back-translation.

Motivated by the success of back-translation in NMT, we investigate in this chapter whether back-translation can benefit from a more insightful data selection approach, i.e., *targeted* sampling. In particular, we explore what words benefit from the generation of additional context and how this information can help us develop more creative data selection methods and improve translation quality. To this end, we ask:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

We partially examined this research question in the previous chapter. In this chapter, we investigate whether model failures are good indicators for choosing new contexts. Methods similar to back-translation have a trained model at their disposal and use it to generate new contexts. We conduct a series of analyses on the learning process of neural translation models with synthetic data. Signals from a pre-trained model can show us where the model is struggling, which can be beneficial in selecting data for augmentation. So we ask:

RQ2.3 *Do signals from the NMT model help identify low-confidence words that could benefit from additional context?*

We review the influence of additional contexts generated by the back-translation approach on the learning process of NMT. Observing the loss function of the model during training, we study the changes in the prediction of every word in the vocabulary. Our findings show that it is mostly words that are difficult to predict in the target language that benefit from additional back-translated data. These low-confidence words have high prediction loss during training when the translation model converges. Leveraging this information, we explore alternatives to random sampling to specifically target these words, thus asking:

RQ2.4 *How can we successfully apply data selection of monolingual data to diversify the contexts of low-confidence words?*

We propose alternatives to the random sampling approach with a focus on increasing occurrences of low-confidence words in the training data. Our proposed approach is twofold: (i) identifying difficult *words* and sampling to increase occurrences of these words, and (ii) identifying *contexts* in which these words are difficult to predict and sample sentences similar to the difficult contexts. We then analyze various ways of identifying difficult words and augmenting the training data. Our investigations show that targeted sampling of monolingual data improves the translation quality of NMT models compared to standard back-translation.

Organization. This chapter is organized as follows: In Section 5.2, we provide an overview of existing work on data selection for machine translation. Next, in Section 5.3,

we describe our data and experimental setup. We study different aspects of the back-translation method in Section 5.4. In Section 5.5, we present a more in-depth analysis of the impact of the new contexts generated by back-translation on the prediction power of our NMT model. Next in Section 5.6, we propose a targeted sampling approach for selecting new contexts for the training data. We also provide experimental results and analyze the impact of different sampling methods. In Section 5.7, we propose an alternative data selection approach with context-aware sampling and provide qualitative results in Section 5.8. Finally, we discuss the conclusions and implications of this work in Section 5.9.

5.2 Related work

In this section, we provide an overview of work related to this chapter on data selection methods in MT.

5.2.1 Data selection in machine translation

Before the emergence of neural models, several previous studies in PBMT have focused on choosing which portion of the parallel corpora to use for training (Moore and Lewis, 2010, Wang et al., 2013). For instance, Axelrod et al. (2011) computed cross-entropy scores for sentence pairs using a 4-gram language model trained on a pseudo in-domain corpus. They sorted the sentence pairs based on this criterion and augmented the training data with the topK sentence pairs that are most relevant to the target domain.

With the development of neural models in machine translation, most works greedily use all available training data for a given language pair. However, it is unlikely that all data is equally useful for creating the best-performing system. Additionally, when the domain of the training and testing data is different, it is essential to carefully select the portion of the data that is most helpful for training. Various data selection methods have been proposed to address the problem of domain adaptation in MT (Silva et al., 2018, Wang et al., 2019a). These methods aim to reduce the model size and result in shorter training times by using a subset of the available data while maintaining high performance. For instance, van der Wees et al. (2017) introduced dynamic data selection, where they vary the selected data subsets during each training epoch. The ranking criteria are based on bilingual cross-entropy differences similar to Axelrod et al. (2011). They significantly reduce the training data size by only using parts of the data which are most relevant to the translation task at hand.

Poncelas et al. (2019a) use two transductive data selection methods, infrequent n-gram recovery and feature decay algorithms, to select a subset of sentence pairs from synthetic training data. This approach ensures that the selected sentence pairs share n-grams with the test set. Poncelas et al. (2019b) investigate whether combining the two paradigms of NMT and PBMT in generating synthetic data contributes to translation quality. In their proposed approach, they randomly select source sentences from the

PBMT synthetic data and the NMT synthetic data with a one-to-one ratio. They show that mixing PBMT and NMT back-translated data further improves over using each type of data alone.

5.3 Data and experimental setup

In this chapter, we conduct several experiments to evaluate the impact of synthetic context on translation quality. For the translation experiments, we use the English↔German WMT17 training data and report results on WMT news test sets 2014, 2015, 2016, and 2017 (Bojar et al., 2017). As NMT system, we use a 2-layer attention-based encoder-decoder LSTM model described in Section 2.4 implemented in OpenNMT (Klein et al., 2017). We train this model with embedding size 512, hidden dimension size 1024, and batch size of 64 sentences. We preprocess the training data with joint Byte Pair Encoding (BPE) using 32K merge operations (Sennrich et al., 2016c).

We compare the results to Sennrich et al. (2016b) by back-translating random sentences from the monolingual data and combine them with the parallel training data. To lessen the arbitrary effect of random sampling, we perform random selection and re-training three times and report the averaged outcomes for the three models. In all experiments, the sentence pairs are shuffled before each epoch. We measure translation quality by single-reference case-sensitive BLEU (Papineni et al., 2002) computed with the `multi-bleu.perl` script from Moses.

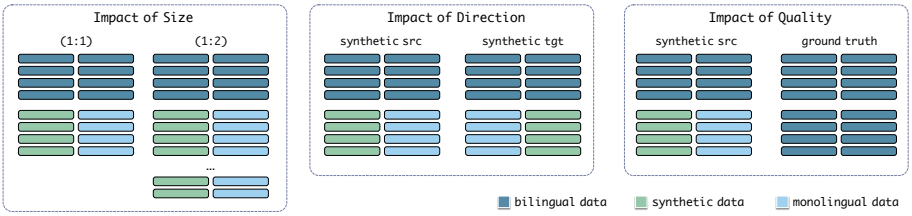


Figure 5.1: Illustration of three set of experiments analyzing different impacts of synthetic data as additional training data: *size* (left), *direction* (middle), and *quality* (right).

5.4 Analyzing back-translation with random sampling

In this section, we investigate different aspects and modeling challenges of integrating the back-translation method into the NMT pipeline. We are interested in investigating the impact of synthetic data on translation quality with random sampling augmentation (see Figure 5.1).

5.4.1 Impact of synthetic data size

One selection criterion for using back-translation is the ratio of real to synthetic data. Sennrich et al. (2016b) showed that higher ratios of synthetic data lead to decreases in translation performance. In order to investigate whether improvements in translation quality increase with higher ratios of synthetic data, we perform three experiments with different sizes of synthetic data (see Figure 5.1: left). Figure 5.2 shows the perplexity as a function of training time for different sizes of synthetic data.

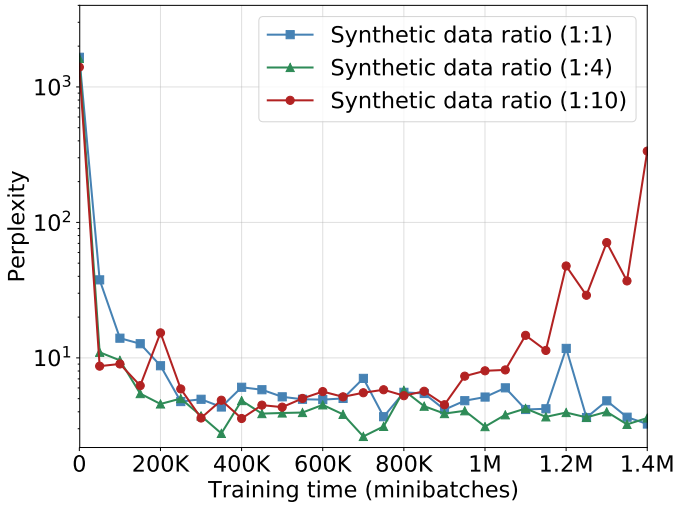


Figure 5.2: Training plots for systems with different ratios of $(real : syn)$ training data, showing perplexity on development set.

We find that all systems perform similarly in the beginning and converge after observing increasingly more training instances. However, the model with the ratio of (1:10) synthetic data becomes increasingly biased towards the noisy data after 1M instances. Decreases in performance with more synthetic than real data is also in line with findings of Poncelas et al. (2018). Comparing the systems using ratios of 1:1 and 1:4, we see that the latter achieves lower perplexity during training. Table 5.1 shows the performance of these systems on the German→English translation task. The BLEU scores show that the translation quality does not improve linearly with the size of the synthetic data. The model trained on 1:4 real to synthetic ratio of training data achieves the best results, however, the performance is close to the model trained on a ratio of 1:1.

Similar to the study in this section, Edunov et al. (2018) showed that it is possible to achieve the best results with a large-scale model trained on a 1:50 ratio of real to

Table 5.1: German→English translation quality (BLEU) of systems with different ratios of *real:syn* data.

	Size	WMT14	WMT15	WMT16	WMT17
Baseline	4.5M	26.7	27.6	32.5	28.1
+ synthetic (1:1)	9M	28.7	29.7	36.3	30.8
+ synthetic (1:4)	23M	29.1	30.0	36.9	31.1
+ synthetic (1:10)	50M	22.8	23.6	29.2	23.9

synthetic data. However, it is crucial to upsample the real data during training so that training batches contain on average an equal amount of real and synthetic data.

5.4.2 Impact of translation direction

Adding monolingual data in the target language to the training data has the potential benefit of introducing new contexts and improving the fluency of the translation model. The automatically generated translations in the source language introduce new contexts for the source words and, despite not being perfect, improve the quality of the final re-trained model.

Monolingual data is available in large quantities for many languages. The decision of the direction of back-translation is subsequently not based on the monolingual data available, but on the advantage of having more fluent source or target sentences.

Table 5.2: English→German translation quality (BLEU) of systems using forward and reverse models for generating synthetic data.

	Size	WMT14	WMT15	WMT16	WMT17
Baseline	4.5M	21.2	23.3	28.0	22.4
+ synthetic tgt	9M	22.4	25.3	29.8	23.7
+ synthetic src	9M	24.0	26.0	30.7	24.8

Lambert et al. (2011) showed that adding synthetic source and real target data achieves improvements in traditional phrase-based machine translation. Similarly, in previous works in NMT, back-translation is applied to monolingual data in the target language. Zhang and Zong (2016) proposed a self-learning algorithm to generate synthetic data from monolingual source sentences. During re-training, they distinguish between real and synthetic data by freezing the parameters of the decoder for the synthetic data.

We perform a small experiment to compare the impact of translation direction and where to incorporate monolingual data (see Figure 5.1: middle). Table 5.2 shows that in both directions, the performance of the translation system improves over the baseline. This is in contrast to the findings of Lambert et al. (2011) for PBMT systems where they show that using synthetic target data does not lead to improvements in translation quality.

Still, when back-translating target monolingual data, BLEU scores in the target language are higher than when translating monolingual data in the source language. This indicates the importance of having fluent sentences in the target language.

5.4.3 Impact of quality of the synthetic data

One selection criterion for back-translation is the quality of the synthetic data. Khayrallah and Koehn (2018) studied the effects of noise in the training data on a translation model and discovered that NMT models are less robust to many types of noise than PBMT models. In order for the NMT model to learn from the parallel data, the data should be fluent and close to the manually generated translations. However, automatically generating sentences using back-translation is not as accurate as manual translations.

To investigate the *oracle gap* between the performance of manually created and back-translated sentences, we perform a simple experiment using the existing parallel training data (see Figure 5.1: right). In this experiment, we divide the parallel data into two parts, train the reverse model on the first half of the data, and use this model to back-translate the second half. The manually translated sentences of the second half are considered as ground truth for the synthetic data.

Table 5.3 shows the BLEU scores of the experiments. As to be expected, re-training with additional parallel data yields higher performance than re-training with additional synthetic data. However, the differences between the BLEU scores of these two models are surprisingly small. This indicates that performing back-translation with a reasonably

Table 5.3: German→English translation quality (BLEU) of systems using synthetic source and human generated source data.

	Size	**WMT14**	**WMT15**	**WMT16**	**WMT17**
Baseline	2.25M	24.3	24.9	29.5	25.6
+ synthetic src	4.5M	26.0	26.9	32.2	27.5
+ ground truth	4.5M	26.7	27.6	32.5	28.1

good reverse model already achieves results that are close to a system that uses additional manually translated data. This is in line with findings of Sennrich et al. (2016b) who

observed that the same monolingual data translated with three translation systems of different quality and used in re-training the translation model yields similar results.

5.5 Back-Translation and token prediction loss

In the previous section, we observed that using back-translated data yields almost the same improvements as gold parallel data with the same target side. However, there is a limit in learning from synthetic data, and with higher ratios of synthetic data the model biases too much towards the synthetic data.

In this section, we investigate the influence of the sampled sentences on the model. In Chapter 4, we showed that targeting specific words during data augmentation improves the generation of these words in the right context. Specifically, adding synthetic sentences containing those words to the training data has an impact on the prediction probabilities of individual words. In this chapter, we further examine the effects of the back-translated synthetic data on the prediction of target tokens.

As mentioned in Section 2, the objective function of training an NMT system is to minimize $\mathcal{L}$ by minimizing the prediction loss, $-\log p(y_t \mid \mathbf{y}_{<t}, \mathbf{s}_n)$, for each target token in the training data, where:

$$p(y_t \mid y_{<t}, \mathbf{s}_n) = \text{softmax}\left(\mathbf{W}_o \widetilde{\mathbf{h}}_t\right) \tag{5.1}$$

Here, $\widetilde{\mathbf{h}}_t$ is the top hidden layer of the decoder and $\mathbf{W}_o$ is the output weight matrix. The addition of monolingual data in the target language improves the estimation of the probability $p(Y)$ and consequently, the model generates more fluent sentences.

Sennrich et al. (2016b) show that by using back-translation, the system with target-side monolingual data reaches a lower perplexity on the development set. This is expected since the domain of the monolingual data is news and therefore similar to the domain of the development set. To investigate the model's accuracy independently from the domains of the data, we collect statistics of the target token prediction loss during training.

Figure 5.3 shows the changes of token prediction loss when training is close to converging and the weights are verging on being stable. The values are sorted in decreasing order by the tokens' mean prediction losses of the system trained on real parallel data (before augmentation). We observe an effect similar to distributional smoothing (Chen and Goodman, 1996): First, we observe that the prediction loss increases slightly for most tokens (red). Next, we spot an irregular pattern in *decrease* of prediction loss (blue). The largest decrease in loss occurs for tokens with high prediction loss values when trained on the parallel data only. This indicates that by randomly sampling sentences for back-translation, the model improves its estimation of tokens that were originally more difficult to predict, i.e., tokens that had a high prediction loss.

Note that we compute the token prediction loss, without updating the weights, in

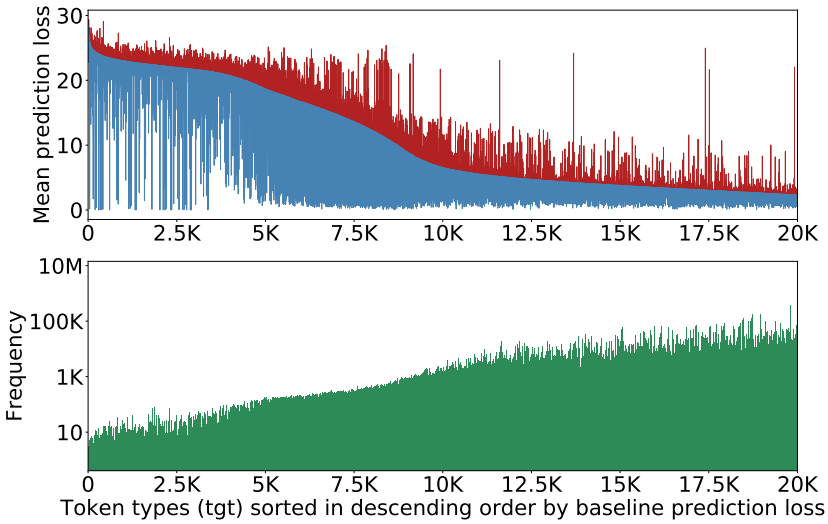


Figure 5.3: Top: Changes in mean prediction loss after re-training with synthetic data sorted by mean prediction loss of the baseline system (x-axis). Decreases and increases in values are marked blue and red, respectively. Bottom: Frequencies (log) of target tokens in the baseline training data. Note that data points in both plots (x-axis) represent *token* types in the vocabulary.

just one pass over the training corpus with the final model and as a result, the loss is not biased towards the order of presentation of the training sentences.

This finding motivates us to further explore sampling criteria for back-translation that contribute considerably to the parameter estimation of the translation model. We propose that by oversampling sentences containing difficult-to-predict tokens, we can maximize the impact of using the monolingual data. After translating sentences containing such tokens and including them in the training data, the model becomes more robust in predicting these tokens. In the next two sections, we propose several methods of using the target token prediction loss to identify the most beneficial sentences for back-translation and re-training the translation model.

5.6 Targeted sampling based on model failure

One of the main benefits of using synthetic data is getting a better estimation of words that were originally difficult to predict as measured by their high prediction losses during training. In this section, we propose four variations of how to identify these words and perform sampling to target these words. The first three variations are described

in Algorithm 1 where the list of difficult tokens is defined in two different ways. The third variation is described in Algorithm 2. The following subsections provide details of these model variants.

Algorithm 1 Sampling for difficult words

Input: Difficult tokens $\mathfrak{D} = \{y_i\}_{i=1}^{D}$, monolingual corpus $\mathbb{M}$, number of required samples *N*

Output: Sampled sentences $S = \{S_i\}_{i=1}^{N}$ where each sentence S_i is sampled from $\mathbb{M}$

procedure DIFFSAMPLING ($\mathfrak{D}, \mathbb{M}, N$):
    Initialize $S = \{\}$
    repeat
        Sample S_c from $\mathbb{M}$
        for all tokens y in S_c **do** if $y \in \mathfrak{D}$
            Add S_c to S
    until $|S| = N$
    return S

5.6.1 Token frequency as a feature of difficulty

Figure 5.3 shows that the majority of tokens with high mean prediction losses have low frequencies in the training data. Additionally, the majority of decreases in prediction loss after adding synthetic sentence pairs to the training data occurs with less frequent tokens. Note that these tokens are not necessarily *rare* and some of them may have up to 1000 different occurrences in the training data. We observe in Figure 5.3 that approximately half of the tokens in the target vocabulary benefit from back-translated data.

We propose a sampling criterion based on token frequencies. Sampling new contexts from monolingual data provides context diversity proportional to the token frequencies and less frequent tokens benefit most from new contexts. Algorithm 1 presents this approach where the list of difficult tokens is defined as:

$$\mathfrak{D} = \{\forall y_i \in V_t \colon freq(y_i) < \eta\} \tag{5.2}$$

where V_t is the target vocabulary and η is the frequency threshold for deciding on the difficulty of the token.

5.6.2 Tokens with high mean prediction losses

In this approach, we use the mean losses to identify difficult-to-predict tokens. The mean prediction loss $\hat{\ell}(y)$ of token y during training is defined as follows:

$$\hat{\ell}(y) = \frac{1}{n_y} \sum_{n=1}^{N} \sum_{t=1}^{|Y^n|} -\log p(y_t^n \mid y_{<t}^n, \mathbf{s}_n) \delta(y_t^n, y) \tag{5.3}$$

where n_y is the number of times token y is observed during training, i.e., the token frequency of y, N is the number of sentences in the training data, $|\mathbf{Y}^n|$ is the length of target sentence n, and $\delta(y_t^n, y)$ is the Kronecker delta function, which is 1 if $y_t^n = y$ and 0 otherwise. By specifically providing more sentences for difficult words, we improve the model's estimation and decrease its prediction uncertainty.

Algorithm 1 presents this approach where the list of difficult tokens is defined as:

$$\mathfrak{D} = \{\forall y_i \in V_t \colon \hat{\ell}(y_i) > \mu\} \tag{5.4}$$

where V_t is the vocabulary of the target language and μ is the threshold for the difficulty of the token.

5.6.3 Tokens with skewed prediction losses

By using the mean loss for target tokens as defined above, we do not discriminate between differences in prediction loss for occurrences in different contexts. This lack of discrimination can be problematic for tokens with high loss variations. For instance, there can be a token with ten occurrences, out of which two have high and eight have low prediction loss values.

We hypothesize that if a particular token is easier to predict in some contexts and harder in others, the sampling strategy should be context-sensitive, allowing to target specific contexts in which a token has a high prediction loss. In order to distinguish between tokens with a skewed and tokens with a more uniform prediction loss distribution, we use both the mean and standard deviation of the token prediction losses to identify difficult tokens. Hence, we target tokens that have both high mean prediction loss and high amount of variation in different contexts. Algorithm 1 formalizes this approach where the list of the difficult tokens is defined as:

$$\mathfrak{D} = \{\forall y_i \in V_t \colon \hat{\ell}(y_i) > \mu \wedge \sigma(\ell(y_i)) > \rho\} \tag{5.5}$$

where $\hat{\ell}(y_i)$ is the mean and $\sigma(\ell(y_i))$ is the standard deviation of prediction loss of token y_i, V_t is the vocabulary list of the target language, and μ and ρ are the thresholds for deciding the difficulty of the token.

5.6.4 Preserving sampling ratio of difficult occurrences

Above, we examined the mean of prediction loss for each token over all occurrences, in order to identify difficult-to-predict tokens. However, the uncertainty of the model in predicting a difficult token varies for different occurrences of the token: one token can be easy to predict in one context, and hard in another. While the sampling step in the previous approaches targets these tokens, it does not ensure that the distribution of sampled sentences is similar to the distribution of problematic tokens in difficult contexts.

Algorithm 2 Sampling with ratio preservation

Input: Difficult tokens and the corresponding sentences in the bitext $\mathfrak{D} = \{y_t, Y_{y_t} = [y_1, \ldots, y_t, \ldots, y_m]\}$, monolingual corpus $\mathbb{M}$, number of required samples *N*

Output: Sampled sentences $S = \{S_i\}_{i=1}^{N}$ where each sentence S_i is sampled from $\mathbb{M}$

procedure PREDLOSSRATIOSAMPLING($\mathfrak{D}, \mathbb{M}, N$):
  Initialize $S = \{\}$
  $H(y_t) = \frac{N \times |(y_t, \cdot) \in \mathfrak{D}|}{|(y_\cdot, \cdot) \in \mathfrak{D}|}$
  repeat
    Sample S_c from $\mathbb{M}$
    for all tokens y in S_c **do** if $|y \in S| < H(y)$
      Add S_c to S
  until $|S| = N$
  return S

To address this issue, we propose an approach where we consider the number of times a token occurs in difficult-to-predict contexts and sample sentences accordingly, thereby ensuring the same ratio as the distribution of difficult contexts. If token y_1 is difficult to predict in two contexts and token y_2 is difficult to predict in four contexts, the number of sampled sentences containing y_2 is double the number of sampled sentences containing y_1. Algorithm 2 formalizes this approach.

5.6.5 Results

We measure the translation quality of various models for German→English and English → German translation tasks. As baseline we compare our approach to Sennrich et al. (2016b). For all experiments we sample and back-translate sentences from WMT monolingual data, keeping a one-to-one ratio of back-translated versus original data (1:1). We set the hyperparameters μ, ρ, and η to 5, 10, and 5000, respectively. The values of the hyperparameters are chosen on a small sample of the parallel data based on the token loss distribution.

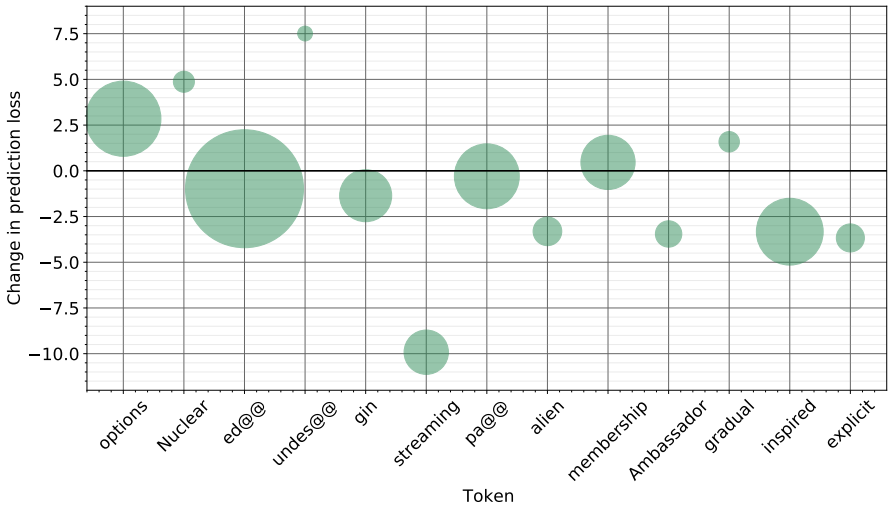


Figure 5.4: Examples of changes in average prediction loss after augmentation. Lower is better. The sizes of dots are proportional to the increase in the number of contexts for each word in the training data. Subword unit boundaries are marked with '@@'.

Our proposed augmentation method increases instances of targeted words in the training data which leads to an overall decrease in average prediction loss per token. Figure 5.4 provides examples of tokens in the training data and their changes after data augmentation. The results of the translation experiments are presented in Tables 5.4 and 5.5.

As expected using random sampling for back-translation improves the translation quality over the baseline. However, each of the proposed targeted sampling techniques outperforms random sampling. Specifically, the best performing model for German→English uses the mean of prediction loss (MPL) for the target vocabulary to frequently sample sentences including these tokens.

For the English→German experiments we obtain the best translation performance when we preserve the prediction loss ratio during sampling. We also observe that even though the model targeting tokens with skewed prediction loss distributions (MPL + SPL) improves over random selection of sentences, it does not outperform the model using only mean prediction losses. Note that frequency-based sampling, the simplest method proposed in this chapter, is very effective. We observe that the gains of other proposed approaches over frequency-based sampling are quite small. Therefore using frequency-based sampling remains a good strategy to improve translation quality over random sampling.

Table 5.4: English→German translation quality (BLEU). Experiments marked † are averaged over 3 runs. RANDOM is the standard back-translation approach with random sampling. MPL and FREQ are difficulty criteria based on mean prediction loss and token frequency, respectively. MPL + SPL is experiments with upsampling tokens with skewed prediction losses. PPLR preserves the ratio of the distribution of difficult contexts.

	En-De			
System	**WMT14**	**WMT15**	**WMT16**	**WMT17**
BASELINE†	21.2	23.3	28.0	22.4
RANDOM†	24.0	26.0	30.7	24.8
Difficulty criterion				
FREQ	24.2	27.0	31.7	25.2
MPL†	**24.7**	26.8	31.5	**25.5**
MPL + SPL	24.1	26.9	31.0	25.3
PPLR	24.5	**27.2**	**31.8**	25.5

Table 5.5: German→English translation quality (BLEU). Experiments marked † are averaged over 3 runs. RANDOM is the standard back-translation approach with random sampling. MPL and FREQ are difficulty criteria based on mean prediction loss and token frequency, respectively. MPL + SPL is experiments with upsampling tokens with skewed prediction losses. PPLR preserves the ratio of the distribution of difficult contexts.

	De-En			
System	**WMT14**	**WMT15**	**WMT16**	**WMT17**
BASELINE†	26.7	27.6	32.5	28.1
RANDOM†	28.7	29.7	36.3	30.8
Difficulty criterion				
FREQ	29.7	30.5	37.5	31.4
MPL†	29.9	**30.9**	**37.8**	**32.1**
MPL + SPL	**30.0**	30.9	37.7	31.9
PPLR	29.8	30.9	37.4	31.6

5.7 Context-Aware targeted sampling

In the previous section, we proposed methods for identifying difficult-to-predict tokens and performed targeted sampling from monolingual data. While the objective was to increase the occurrences of difficult tokens, we ignored the context of these tokens in the sampled sentences.

Arguably, if a word is difficult to predict in a given context, providing more examples of the same or similar context can aid the learning process. In this section, we focus on the context of difficult-to-predict words and aim to sample sentences that are similar to the corresponding difficult context. We first identify difficult-to-predict words and the local *context* where the prediction loss is high. Next, we sample sentences from the monolingual data that contain a difficult-to-predict word. We then compare the context of the difficult word in the sampled sentence and the initial difficult context and select the sampled sentence if the contexts are similar. Finally, we back-translate the selected sentences and augment the training data.

Algorithm 3 Sampling with context

Input: Difficult tokens and the corresponding sentences in the bitext $\mathfrak{D} = \{y_t, Y_{y_t} = [y_1, \ldots, y_t, \ldots, y_m]\}$, monolingual corpus $\mathbb{M}$, context function $context$, number of required samples *N*, similarity threshold s

Output: Sampled sentences $S = \{S_i\}_{i=1}^{N}$ where each sentence S_i is sampled from $\mathbb{M}$

1: **procedure** CONTEXTSAMPLING($\mathfrak{D}, \mathbb{M}, context, N, s$):
2: Initialize $S = \{\}$
3: **repeat**
4: Sample S_c from $\mathbb{M}$
5: **for all** tokens y_t in S_c **do** if $y_t \in \mathfrak{D}$
6: $C_m \leftarrow context(S_c, \text{index\_of}(S_c, y_t))$
7: **for all** Y_{y_t} **do**
8: $C_p \leftarrow context(Y_{y_t}, \text{index\_of}(Y_{y_t}, y_t))$
9: if $\text{Sim}(C_m, C_p) > s$: Add S_c to S
10: **until** $|S| = N$
11: **return** S

The general algorithm is described in Algorithm 3. In the following sections, we discuss different definitions of the local context ($context$ function in line 6 and line 8) and similarity measures (Sim function in line 9) in this algorithm and report the results.

5.7.1 Definition of local context

Prediction loss is a function of the source sentence and the target context. We hypothesize that one of the reasons that a token has a high prediction loss in only some contexts is because of the complexity of those contexts. This complexity can be caused by an infrequent event such as a rare sense of the word, a domain that is different from other occurrences of the word, or an idiomatic expression.

We identify *pairs* of tokens and sentences from parallel data where in each pair, the NMT model suffers a high prediction loss for the token in the given context. Note that a token can occur several times in this list since it can be considered as difficult-to-predict in different sentences.

We propose two approaches to define the local context of a difficult token:

Neighboring tokens A straightforward way is to use positional context: tokens that precede and follow the target token, typically in a window of w tokens to each side. For sentence S containing a difficult token at index i, the *context* function in Algorithm 3 is:

$$context(S, i) = [S^{i-w}, \ldots, S^{i-1}, S^{i+1}, \ldots, S^{i+w}] \tag{5.6}$$

where S^j is the token at index j in sentence S. Note that in this approach, we look at a window of fixed size and as a result, not all subwords from the same word may end up in this context window. For instance for the sentence *'a professor and a colleague at Stan|ford'*, with target word *'colleague'*, and $w = 2$, the context is [*'and', 'a', 'at', 'Stan'*]. Here, the subword *'ford'* as part of the word *'Stanford'* is not included in the context window. The symbol '|' signifies subword unit boundary.

Sibling tokens In our analysis of prediction loss during training, we observe that several tokens that are difficult to predict are indeed subword units. Current state-of-the-art NMT systems apply BPE to the training data to address large vocabulary challenges (Sennrich et al., 2016c). By using BPE, the model generalizes common subword units towards what is more frequent in the training data. This is inherently useful since it allows for better learning of less frequent words. However, a side effect of this approach is that at times the model generates subword units that are not linked to any words in the source sentence. As an example, in Table 5.6, the German source and the English reference translation highlight this problem. The word *'B|ahr'* consisting of two subword units is incorrectly translated into *'B|risk'* because of an unintended side-effect of both sharing the subword unit *'B'*.

We address the insufficiency of the context for subword units with high prediction losses by targeting these tokens in sentence sampling. Algorithm 3 formalizes this approach in sampling sentences from the monolingual data. For a sentence S containing

a difficult subword at index i, the context function is defined as:

$$context(S, i) = [S^n, \ldots, S^{i-1}, S^{i+1}, \ldots, S^m] \tag{5.7}$$

where every token S^j in the local context is a subword unit and part of the same word as S^i. Table 5.7 presents examples of sampled sentences for the difficult subword unit *'Stan'*. In this case, the difficult context for this token is *'Stan|ford'* and we use it for computation of similarity. This suggests that the subword unit *'Stan'* is difficult to predict when the context is for the word *'Stan|ford'*. This excludes other contexts where the subword unit *'Stan'* is part of another word, such as *'Stan|dard'*.

Table 5.6: An example from the synthetic data where the word *B|ahr* is incorrectly translated to *B|risk*. Subword unit boundaries are marked with '|'.

source	wer glaube, dass das Ende, sobald sie in Deutsch-land ank|ä|men, ir|re, erzählt **B|ahr**.
reference	if you think that this stops as soon as they arrive in Germany, you'd be wrong, says **B|ahr**.
NMT output	who believe that the end, as soon as they go to Germany, tells **B|risk**.

Table 5.7: Results of context-aware targeted sampling with sibling tokens for the difficult subword unit *'Stan'*. In this example, the difficult context in which the subword *'Stan'* has a high prediction loss is the complete word *'Stanford'* and we sample sentences containing this word.

*Sentence from bitext containing difficult token **'Stan'***
He attended ***Stan**|ford* University, where he double maj|ored in Spanish and History.
Sampled sentences from monolingual data
The group is headed by Aar|on K|ush|ner, a ***Stan**|ford* University gradu|ate who formerly headed a gre|eting card company.
Ford just opened a new R&D center near ***Stan**|ford* University, a hot|bed of such technological research.
Joe Grund|fest, a professor and a colleague at ***Stan**|ford* Law School, outlines four reasons why the path to the IP|O has become so steep for asp|iring companies.

5.7.2 Similarity of the local contexts

In context-aware targeted sampling, we compare the context of a sentence candidate and the difficult context in the parallel data and select the sentence if they are *similar*. In the following, we propose two approaches for measuring the similarities.

Matching the local context (Exct) In this approach, we aim to sample sentences containing the difficult token matching the exact context to the problematic context. By sampling sentences that match in a local window with the problematic context and differ in the rest of the sentence, we have more instances of the difficult token for training. Algorithm 3 formalizes this approach where the similarity function is defined as:

$$\text{Sim}(C_m, C_p) = \frac{1}{c} \sum_{i=1}^{c} \delta(C_m^i, C_p^i) \tag{5.8}$$

C_m and C_p are the contexts of the sentences from monolingual and parallel data, respectively, and c is the number of tokens in the contexts. The δ function returns 1 when C_m^i and C_p^i are the same token, and 0 otherwise.

Word representations (Sem) Another approach to sampling sentences that are similar to the problematic context is to weaken the matching assumption. Allowing sentences that are similar in subject and not match the exact context words allows for lexical diversity in the training data. We use embeddings obtained by training the Skipgram model (Mikolov et al., 2013a) on monolingual data to calculate the similarity of the two contexts. For this approach we define the similarity function in Algorithm 3 as:

$$\text{Sim}(C_m, C_p) = \cos(\mathbf{v}(C_m), \mathbf{v}(C_p)) \tag{5.9}$$

where $\mathbf{v}(C_m)$ and $\mathbf{v}(C_p)$ are the averaged embeddings of the tokens in the contexts. Table 5.8 gives examples of sampled sentences for the difficult word *Rock*. In this example, the context where the word *'Rock'* has high prediction loss is about the *music genre* and not the most prominent sense of the word, *stone*. Sampling sentences that contain this word in this particular context provides an additional signal for the translation model to improve parameter estimation.

5.7.3 Results

The results of the translation experiments are given in Tables 5.9 and 5.10 for German→ English and English→German, respectively. In these experiments, we set the hyperparameters s and w to 0.75 and 4, respectively. Comparing the experiments with different similarity measures, *Exct* and *Sem*, we observe that in all test sets we achieve the best results when using word embeddings. This indicates that for targeted sampling it is more beneficial to have diversity in the context of difficult words as opposed to having

Table 5.8: Results of context-aware targeted sampling for the difficult token *'Rock'*

Sentence from bitext containing difficult word		
Bud|dy Hol|ly was part of the first group induc|ted into the **Rock** and R|oll Hall of F|ame on its formation in 1986.		
Sentences from monolingual data	*Similarity*	*Sampled*
A 2008 **Rock** and R|oll Hall of F|ame induc|t|ee, Mad|onna is ran|ked by the Gu|inn|ess Book of World Rec|ords as the top-selling recording artist of all time.	0.86	✓
The winners were chosen by 500 voters, mostly musicians and other music industry veter|ans, who belong to the **Rock** and R|oll Hall of F|ame Foundation.	0.81	✓
The **Rock** and R|oll Hall of Fam|ers gave birth to the California rock sound.	0.79	✓
After an ice cold San Miguel beer at the H|ard **Rock** Café (Ay|ala Center) just enter the Bur|gos Street and enjoy the different clubs.	0.42	✗
See a play on Broad|way, enjoy stunning views from the Top of the **Rock**, or spend the day at the Museum of Mo|dern Art, all situated nearby.	0.34	✗
The Library received the donations and endo|w|ments of prominent individuals such as John D. **Rock**|ef|eller and James B. Wil|b|ur.	0.29	✗

the exact n-grams. When using embeddings as the similarity measure, it is worth noting that with a context of size 4 the model performs very well but fails when we increase the window size to include the whole sentence. The experiments focusing on tokens from the same words (sibling tokens) achieve improvements over the baselines, however, they perform slightly worse than the experiments using neighboring tokens as context.

The best BLEU scores are obtained with the mean of prediction loss as difficulty criterion (MPL) and using word representations to identify the most similar contexts. We observe that summarizing the distribution of the prediction losses by its mean is more beneficial than using individual losses. Our results motivate further explorations of using context for targeted sampling sentences for back-translation.

5.8 Qualitative results

Finally, we review our proposed approach and further investigate individual token losses. We observed that the *individual* token loss, even after training converges, has a degree of instability and for the same word, it varies from context to context. However, in our experiments in the previous section, using local context to identify these difficult words

Table 5.9: German→English translation quality (BLEU). Experiments marked † are averaged over 3 runs. RANDOM is the standard back-translation approach with random sampling. PREDLOSS is the contextual prediction loss and MPL is the average loss. *token* and *SubUnit* are context selection definitions from neighboring tokens and subword units, respectively. Note that token includes both subword units and full words. *Sent* regards the entire sentence as the context. *Sem* is computing context similarities with token embeddings and *Exct* is comparing the context tokens.

System	Difficulty criterion						De-En			
							WMT14	WMT15	WMT16	WMT17
BASELINE †							26.7	27.6	32.5	28.1
RANDOM †							28.7	29.7	36.3	30.8
	Context			Similarity						
	Neighbor	Sibling	Sentence	Exct	Sem					
FREQ	✓				✓		30.0	30.8	37.6	31.7
PREDLOSS		✓		✓			29.1	30.1	36.9	31.0
PREDLOSS	✓			✓			29.7	30.6	37.6	31.8
PREDLOSS	✓				✓		29.9	30.8	37.7	31.9
PREDLOSS			✓		✓		24.9	25.5	30.1	26.2
MPL	✓				✓		**30.2**	**31.4**	**37.9**	**32.2**

Table 5.10: English→German translation quality (BLEU). Experiments marked † are averaged over 3 runs. RANDOM is the standard back-translation approach with random sampling. PREDLOSS is the contextual prediction loss and MPL is the average loss. *token* and *SubUnit* are context selection definitions from neighboring tokens and subword units, respectively. Note that token includes both subword units and full words. *Sent* denotes the sentence as the context. *Sem* is computing context similarities with token embeddings and *Exct* is comparing the context tokens.

System						En-De			
						WMT14	WMT15	WMT16	WMT17
BASELINE †						21.2	23.3	28.0	22.4
RANDOM †						24.0	26.0	30.7	24.8
Difficulty criterion	**Context**			**Similarity**					
	Neighbor	Sibling	Sentence	Exct	Sem				
FREQ	✓				✓	24.4	26.3	31.5	25.6
PREDLOSS		✓		✓		23.8	26.2	28.8	23.2
PREDLOSS	✓			✓		24.3	27.4	31.6	25.5
PREDLOSS	✓				✓	24.5	**27.5**	31.7	25.6
PREDLOSS			✓		✓	22.0	24.6	27.9	22.5
MPL	✓				✓	**24.4**	27.2	**31.8**	**25.6**

was not very successful.

We look at some examples from the training data where individual token loss is unstable in different contexts. Figure 5.5 illustrates several sentences from the training data containing the subword *'danger@@'* and the respective prediction losses of the trained model. The symbol '@@' signifies subword unit boundary. In all instances, this subword unit is part of the word *'dangerously'*. All of the source sentences of these examples include the same German translation, *'gefährlich'*, that corresponds to the translation of this word. In this particular example, we see no clear indication in the context of why the model's confidence for the token *'danger'* is considerably different for different contexts.

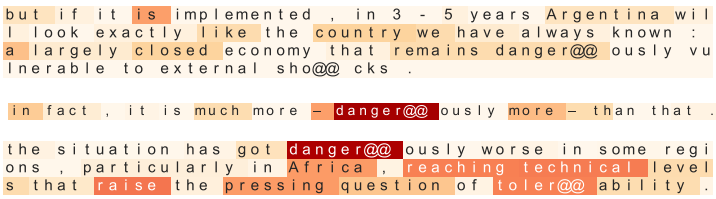


Figure 5.5: Visualization of token prediction loss (final training epoch) for the subword ***danger***@@ in three different sentences. Darker means the model has less confidence predicting the word. Subword unit boundaries are marked with '@@'

Finally, we study the importance of the position of the token in the confidence of the model. The prediction loss of the token i in the target sentence is conditioned on

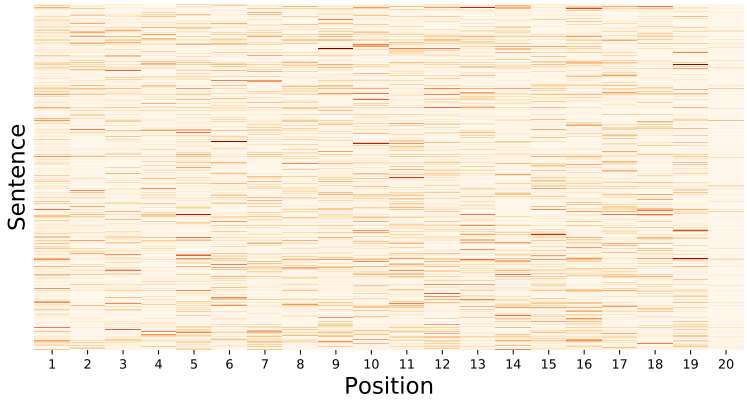


Figure 5.6: Prediction losses of 1000 randomly sampled sentences of the same length (20 tokens) from the training data. Darker means the model has less confidence predicting the word.

the source sentence and the target tokens generated up to the token in position i. As a result, words that occur later in the sentence have more contextual evidence than words appearing earlier and this could lead to better prediction. We examine whether the position of the token in the sentence is a notable factor in prediction loss values.

We randomly sample 1000 sentences of the same length (20 tokens) from the training data and observe the confidence of the model in the prediction. Figure 5.6 shows the spread of prediction loss values in each position in the sentence. The only distinct pattern we observe is that the last position has consistently low prediction loss. This is expected since the end-of-sentence symbol always follows the end of sentence markers, such as '.', '!', or '?'. We observe that the average loss values are slightly higher for the first position (12% higher). However, other positions have very similar average losses. We conclude that the position in the sentence is not a significant factor in individual prediction losses.

5.9 Conclusion

Motivated by our observations in Chapter 4 that synthesizing new context is useful for translation of rare words, we further explored in this chapter the impact of additional contexts on the translation of words that are difficult to predict by the baseline model. We asked:

RQ2.3 *Do signals from the NMT model help identify low-confidence words that could benefit from additional context?*

In this chapter, we explored different aspects of the back-translation method to gain a better understanding of its performance. Our analyses showed that the quality of the synthetic data has a small impact on the effectiveness of back-translation once there is sufficient training data available. However, the ratio of synthetic to real training data plays a more critical role. When the ratio of synthetic to real data is high, the model becomes biased towards noise in the synthetic data and the quality decreases.

Next, we examined the NMT model and found that words with high prediction losses after training benefit the most from additional back-translated data. While individual prediction losses are not a distinctive factor in identifying difficult words and same words in very similar contexts have high variance of prediction loss, by averaging these values we can successfully spot difficult words. Our findings showed that, when original model has a low confidence in predicting words, the addition of contexts for those words to the training data increases the overall accuracy of the model on the unseen test set.

Equipped with this information, we asked:

RQ2.4 *How can we successfully apply data selection of monolingual data to diversify the contexts of low-confidence words?*

As an alternative to random sampling target sentences for back-translation, we proposed targeted sampling and specifically targeted words that are difficult to predict. We found that data augmentation with the goal of increasing contexts for difficult-to-predict words improved the translation quality in German↔English. Interestingly, the proposed frequency-based sampling approach is a simple, yet effective strategy that is hard to outperform. This indicates that signals from the data distribution are on par with signals from the failures of the model.

This allows us to answer our more general question:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

In this chapter, we continued our study on the influence of having diverse contexts on translation quality. We found that translation quality improves when we diversify the context of difficult words. We investigated the effective method of back-translation for NMT and explored alternatives to the typically used random selection of target sentences that are to be back-translated into the source language.

In Chapters 4 and 5, we studied the impact of the availability of data and why models suffer from a lack of diverse contexts during training. We proposed two main data augmentation approaches with multiple variants targeting different problems in translation. Both approaches proposed in Chapters 4 and 5 lead to improvements in translation quality.

6

Translating Idiomatic Expressions

6.1 Introduction and research questions

In Chapter 3, we experimented with changing the scope of the context from sentence-level to document-level to capture the meaning of ambiguous words. In Chapters 4 and 5, we demonstrated that added context significantly help translating rare and difficult words. The next question we are interested in is which phenomena we still *fail* to capture with current approaches to using contexts. Neural machine translation has achieved substantial improvements in translation of different linguistic phenomena over traditional rule-based and phrase-based models. For instance, reordering, subject-verb agreement, double-object verbs, and overlapping subcategorization are various areas where neural models successfully overcome the limitations of phrase-based models (Isabelle et al., 2017, Bentivogli et al., 2016).

In this chapter, we examine which phenomena are not fully captured by current NMT models. NMT models use both source and target sentences as contexts to generate a target word. We are interested in cases where this scope is not sufficient for the NMT model. To shed light on this vulnerability of current NMT models, we ask:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

We study the ability of NMT models to translate fairly complex linguistic phenomena. To examine this question, in this chapter, we focus on the translation of units that possibly require cues beyond the literal context, namely idiomatic expressions. Idioms, a category of multiword expressions, are an interesting language phenomenon where the overall meaning of the expression cannot be inferred from the meanings of its parts. For the most part, idiom acquisition for humans requires additional resources such as explicit definitions of the expressions. NMT models, however, only have access to the nearby and local context of the idiomatic expression.

To further investigate why neural translation models struggle in this area, we ask:

RQ3.1 *What are the challenges of idiom translation with neural models?*

The first challenge for learning and evaluating idiom translation is the lack of dedicated data sets. In this chapter, we address this problem by creating the first large-scale data set for idiom translation. Building a hand-crafted data set for idiom translation is costly and time-consuming. In this chapter, we automatically build a new bilingual data set for idiom translation extracted from an existing general-purpose German↔English parallel corpus. The first part of our data set consists of 1,500 parallel sentences where the German side contains an idiom, while the second part consists of 1,500 parallel sentences where the English side contains an idiom. Additionally, we provide the corresponding training data sets for German→English and English→German translation where source sentences including an idiom phrase are identified.

We then study how this data set can aid in assessing the translation quality of idiomatic expressions, thus asking:

RQ3.2 *How is the translation quality of NMT influenced by idiomatic expressions?*

Having prepared the idiom translation training and test data, we investigate how to assess the translation quality of idiomatic expressions. The labels in our data are indicators of the existence of idioms in a sentence. We use these labels as an additional signal during training of the NMT model and examine whether this flag is sufficient in identifying a phrase as idiomatic and translating it correctly. Finally, we introduce several metrics to evaluate the translation quality of idiom phrases in a sentence.

Organization. This chapter is organized as follows: In Section 6.2, we provide an overview of existing work on idiom identification and translation. Next, in Section 6.3 we introduce our data collection procedure and details on the extracted training and test data. Section 6.4 describes the design of the experiments for translating idioms. Section 6.5 proposes various metrics to locally evaluate idiom translation and provides experimental results on the translation task and analyzes the performance. Finally, we discuss the conclusions and implications of this work in Section 6.6.

6.2 Idiomatic expressions

Non-compositional multiword expressions, or idioms, are lexical semantic units where the meaning is often not merely a function of the meaning of its constituent parts (Nunberg et al., 1994, Kövecses and Szabó, 1996).

The non-compositionality characteristic of idiomatic expressions exists to different degrees in a language (Nunberg et al., 1994). In English for example, for the idiom *"spill*

the beans", the word *'spill'* symbolizes *'reveal'* and *'beans'* symbolizes the *'secrets'*. For the idiomatic expression *"kick the bucket"*, on the other hand, no such analysis is possible. Automatically identifying these idiomatic expressions in a sentence is challenging. In the following section, we discuss previous works in this area.

6.2.1 Idiom identification

Expressions that potentially have idiomatic meanings can be recognized using various lexical association measures (Evert and Krenn, 2001, Evert and Kermes, 2003). However, other methods are necessary to decide whether a particular multiword expression (MWE) has an idiomatic use in a particular context. Katz and Giesbrecht (2006) use distributional semantics as a model of context similarity to examine whether the local context of an MWE can distinguish its idiomatic use from its literal use. Salehi and Cook (2013) use the translation of the components of the MWE in multiple languages to compute similarities between strings. This compositionality score illustrates the relative degree of compositionality of the MWE. Salehi et al. (2015) implement a similar approach but uses word embeddings to compute the compositionality score of an MWE.

Salton et al. (2016) use skip-thought vectors, sent2vec, first introduced by Kiros et al. (2015) for idiom classification. In this approach, they define the classes as to whether an MWE is used literally or idiomatically. More recently, Klyueva et al. (2017) propose to use an RNN that predicts the possible tags of an MWE. The system scored better in more 'syntactic' MWEs like inherently reflexive verbs, light verb constructions, and verb-particle constructions. However, they were not able to detect idioms with reasonable accuracy.

6.2.2 Idiom translation

Automatic translation of idiomatic phrases is a long-established problem in NLP (Schenk, 1986). As we illustrated in previous chapters, NMT models battle with translating rare words. In a way, idioms are similar to this problem. While the occurrence of the expression might not be rare, the idiomatic meaning of the expression in a particular context is often uncommon (Salton et al., 2014a, Isabelle et al., 2017, Agrawal et al., 2018). The challenge of translating idiomatic phrases in NMT is partly due to the underlying complexity of identifying a phrase as idiomatic and generating its correct non-literal translation, and partly due to the fact that idioms are rarely encountered in the standard data sets used for training NMT systems.

As an example, in Table 6.1, we provide an idiomatic expression in German and the literal and idiomatic translations in English. We note that the literal translation of an idiom is not the correct translation; neither does it capture part of the meaning. To illustrate the problem of idiom translation we also provide the output of three NMT systems for this sentence: GoogleNMT (Wu et al., 2016), DeepL[1], and the OpenNMT

[1] www.deepl.com/translator

Table 6.1: Example of an idiomatic phrase in German and its translation. We compare the output of state-of-the-art commercial models (DeepL and GoogleNMT), as well as our trained model (based on OpenNMT). In translating a sentence containing this idiomatic phrase, we notice that none capture the idiom translation correctly.

German phrase	*eine weiße Weste haben*
Literal translation	to have a white vest
Idiomatic translation	to have clean slate
Sentence	Coca-Cola und Nestlé gehören zu den Unterzeichnern. Beide **haben** nicht gerade **eine weiße Weste**.
Reference translation	Coca Cola and Nestlé are two signatories with *"spotty" track records*.
DeepL	Coca-Cola and Nestlé are among the signatories. Neither of them is **exactly the same**.
GoogleNMT	Coca-Cola and Nestlé are among the signatories. Both do not **have just a white vest**.
OpenNMT	Coca-Cola and Nestlé are among the signatories. Both don't **have a white essence**.

implementation (Klein et al., 2017) based on Bahdanau et al. (2015) and Luong et al. (2015a) trained on WMT17 parallel corpora. All systems fail to generate the proper translation of the idiomatic expression. This problem is particularly pronounced when the source idiom is very different from its equivalent in the target language, as is the case here.

Although there are a number of monolingual data sets available for identifying idiomatic expressions (Muzny and Zettlemoyer, 2013, Markantonatou et al., 2017), there is limited work on building a parallel corpus annotated with idioms, which is necessary to investigate this problem more systematically. Salton et al. (2014b) selected a small subset of 17 English idioms, collected 10 sentence examples for each idiom from the Internet, and manually translated them into Brazilian-Portuguese to use for translation. Isabelle et al. (2017) built a challenge set of 108 short sentences that each focus on one difficult phenomenon of the language. Their manual assessment of the eight sentences containing an idiomatic phrase showed that NMT systems struggle with the translation of these phrases.

Shao et al. (2018) introduced a new evaluation metric for detecting literal translation errors in Chinese→English translation, and concluded that idiom translation remains an open problem in MT. Moussallem et al. (2018) released a multilingual resource on idioms currently containing five languages: English, German, Italian, Portuguese, and Russian. In this work, the authors built the data set by crawling various sources and then have them manually evaluated by native speakers.

While these approaches are valuable for studying the problem of idiom translation,

they each require manual efforts to identify and label idioms. To further research in idiom translation, we still need *large-scale* training and testing resources which are hard to obtain with manual labeling.

6.3 Data collection

In this section, we introduce our proposed data collection procedure for building a training and test set. We focus on German↔English translation of idioms. This is an established language pair commonly used in machine translation literature. Automatically identifying idiomatic phrases in a parallel corpus requires a gold standard data set annotated manually by linguists. We use an online dictionary containing idiomatic and colloquial phrases[2], which is built manually, as our gold standard for extracting idiom phrase pairs.

Examining the WMT German↔English test sets from 2008 to 2016 (Bojar et al., 2017), we observe very few sentence pairs containing an idiomatic expression. The standard parallel corpora available for training however contain a sizeable number of such sentence pairs. Therefore, we automatically select sentence pairs from the training corpora where the source sentence contains an idiom phrase to build the new test set. Note that we only focus on idioms on the source side and we have two separate lists of idioms for German and English. Hence, we independently build two test sets (for German idiom translation and English idiom translation) with different sentence pairs selected from the parallel corpora.

Table 6.2: Two examples displaying different constraints of matching an idiom phrase with occurrences in the sentence.

German idiom	*alles über einen kamm scheren*
English equivalent	to measure everything by the same yardstick
Matching German sentence	Aber man kann eben nicht **alle** Inseln **über einen Kamm scheren**.
English translation	But we cannot measure all islands by the same standards.
German idiom	*in den kinderschuhen stecken*
English equivalent	to be in the fledgling stage
Matching German sentence	Es **steckt** immer noch **in den Kinderschuhen**.
English translation	It is still in its infancy.

Depending on the language, the words making up an idiomatic phrase are not always contiguous in a sentence. For instance, in German, the subject can appear between the

[2] `www.dict.cc`

verb and the prepositional phrase making up the idiom. German also allows for several re-orderings of the phrase.

In order to generalize the process of identifying idiom occurrences, we lemmatize the phrases and consider different re-orderings of the words in the phrase as an acceptable match. We also allow for a fixed maximum number of words to occur in between the words of an idiomatic phrase. Table 6.2 shows two examples of idiom occurrences that match these criteria. Following this set of rules, we extract sentence pairs containing idiomatic phrases and create a set of sentence pairs for each unique idiom phrase.

There are various ways of combining regular and idiomatic sentences and building training and test data. We know that the NMT model is capable of translating a word correctly at test time if it has observed it at training time. In the previous chapters, we showed that the frequency of occurrences in the training data and the quality of the contexts are important factors in helping the model learn to translate. Motivated by this, we distribute sentences with idiomatic phrases between training and test sets so that there are no idioms in the test set that we have not seen during training. We also make sure that there is no overlap between the training and test sets.

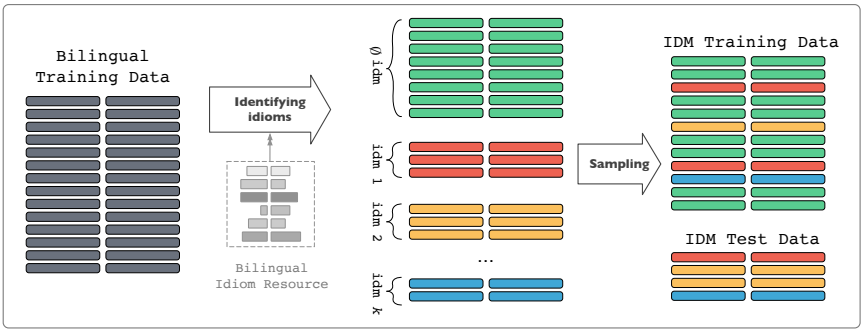


Figure 6.1: The process of data collection and construction of the test set containing only sentence pairs with idiomatic phrases.

Considering these principles, we build the training and test data as follows: First, we sample without replacement from WMT data sets and select individual sentence pairs to build the idiom test set. To build the new training data, we use the remaining sentence pairs in each idiom set as well as the sentence pairs from the original parallel corpora that did not include any idiomatic phrases. In this process, we ensure that for each idiomatic expression there is at least one occurrence in both training and test data and that no sentence pair is included in both training and test data.

Figure 6.1 visualizes the process of constructing the new training and test sets. As a result of this construction, for each language direction, we obtain a targeted test set for idiom translation and the corresponding training corpus representing a natural distribution of sentences with and without idioms. We annotate each sentence pair with the canonical form of its source-side idiom phrase and its equivalent in the target

Table 6.3: Statistics of the constructed German and English idiom translation data sets.

German idiom translation data set	
Number of unique idioms	103
Training size	4.5M
Idiomatic sentences in training data	1848
Test size	1500
English idiom translation data set	
Number of unique idioms	132
Training size	4.5M
Idiomatic sentences in training data	1998
Test size	1500

language.

Table 6.3 provides some statistics of the two data sets. For each unique idiom in the test set, we also provide the frequency of the respective idiom in the training data. Note that this is based on the lemmatized idiom phrase under the constraints mentioned in Section 6.3 and is not necessarily an exact match of the phrase. Table 6.4 shows several examples from the data set for German idiom translation. We observe that for some idioms the literal translation in the target language is close to the actual meaning, while for others it is not the case.

Note that multiword expressions that at times have an idiomatic meaning can also be translated literally depending on the context (e.g., *"spill the beans"* to literally describe the act of *spilling the beans*). This data set represents this additional difficulty: Models cannot just memorize fixed translation of idioms but also have to consider the specific context in which they are used.

6.4 Translation experiments

While the main focus of this chapter is to generate data sets for training and evaluating idiom translation, we also perform a few NMT experiments using our data set to measure the problem of idiom translation on large-scale data.

In the first experiment, following the conventional settings, we do not use any labels indicating whether a particular phrase is used idiomatically or not in the training data. In the second experiment, we use the labels in the training data as an additional feature to investigate the effect of informing the model of the existence of an idiomatic phrase in a sentence during training. We perform a German→English experiment by providing the model with additional input flags. This approached is similar to the work by Sennrich et al. (2016a), where they control the honorifics produced at test time by adding a side

Table 6.4: Examples from the German idiom translation test set.

German idiom	*in den kinderschuhen stecken*
English equivalent	to be in the fledgling stage
German sentence	Eine Bemerkung, Gentoo/FreeBSD **steckt** noch **in den Kinderschuhen** und ist kein auf Sicherheit achtendes System.
English sentence	Note that Gentoo/FreeBSD is still **in its infancy** and is not a security supported platform.
German idiom	*den kreis schließen*
English equivalent	to bring sth. full circle
German sentence	Die europäische Krise **schließt den Kreis**.
English sentence	The European crisis is **coming full circle**.
German idiom	*auf biegen und brechen*
English equivalent	by hook or crook
German sentence	Nehmen wir zum Beispiel die Währungsunion: Sie soll **auf Biegen und Brechen** eingeführt werden.
English sentence	Take, for example, the introduction -**come what may**- of the single currency.
German idiom	*sie haben das wort*
English equivalent	the floor is yours
German sentence	Berichterstatterin. - (FR) Herr Präsident! Danke, dass **Sie** mir **das Wort** erteilt **haben**.
English sentence	rapporteur. - (FR) Mr President, thank you for **giving** me **the floor**.

constraint to the source side.

The additional flag indicates whether a source sentence contains an idiom and are implemented as a special extra token <idm> that is prepended to each source sentence containing an idiom both in the training and test data. This a simple approach that can be applied to any sequence-to-sequence architecture.

We use a 4-layer attention-based encoder-decoder model as described in Section 2.4 trained with hidden dimension size of 1000, and batch size of 80 for 20 epochs. In all experiments, the NMT vocabulary is limited to the most common 30K words in both languages and we preprocess source and target language data with Byte Pair Encoding (BPE) (Sennrich et al., 2016c) using 30K merge operations. We also use a phrase-based translation system similar to Moses (Koehn et al., 2007) as baseline to measure PBMT performance for idiom translation. Several examples of our idiom translation test set and the output translations of the PBMT and NMT models are illustrated in Table 6.5.

6.5 Idiom translation evaluation

Ideally, idiom translation should be evaluated manually, but this is a very costly process. Automatic metrics, on the other hand, can be used on large data sets at no cost and have the advantage of replicability (Section 2.6). We use three metrics to evaluate the translation quality with a specific focus on idiom translation accuracy: BLEU, Modified Unigram Precision, and Word-level Idiom Accuracy. We describe each metric below.

6.5.1 BLEU

The traditional BLEU score (Papineni et al., 2002), discussed in Section 2.6, is a good measure to determine the overall quality of the translations. However, this measure considers the precision of *all* n-grams in a sentence and by itself does not focus on the translation quality of the idiomatic expressions.

6.5.2 Modified unigram precision

To specifically concentrate on the quality of the translation of idiomatic expressions, we also look at the *localized* precision. In this approach, we translate the idiomatic expression in the context of a sentence and only evaluate the translation quality of the idiom phrase.

To isolate the idiom translation in the sentence, we look at the word-level alignments between the idiomatic expression in the source sentence and the generated translation in the target sentence. We use `fast-align` (Dyer et al., 2013) to extract word alignments. Since idiomatic phrases and the respective translations are not contiguous in many cases, we only compare the unigrams of the two phrases. We compute unigram matches between the reference translation and the translation output, the $candidates$

Table 6.5: Examples from the resulting test set of sentence pairs containing idiomatic expressions. NMT and PBMT translations of sentences are provided, highlighting the challenge of idiom translation.

SRC	Seitdem aber begannen sich zwischen Polívka und Harabiš die Streitigkeiten zu häufen, die in der Absetzung "König Boleslavs I. gewählt **bis zum Sankt-Nimmerleins-Tag**" gipfelten.
REF	From then on , quarrels begin to accumulate between Polívka and Harabiš , which culminated in the dethronement of "the king Boleslav I elected **forever and ever**".
PBMT	Since then, however, the disputes between Polívka and Harabiš began to accumulate, culminating in the departure of King Boleslavs I.
NMT	Since then, but began between Polívka and Harabiš disputes to accumulate in the removal "king Boleslavs I. elected by the Sankt-Nimmerleins-Tag culminated".
SRC	Sie wurde **vor Ort** notärztlich behandelt und von Rettungskräften in ein Krankenhaus gebracht.
REF	She was treated **at the site** by an emergency doctor and taken to hospital by ambulance.
PBMT	It was treated on site in the field, and it was brought to a hospital from the rescue forces.
NMT	she was on the ground and notärztlich treated by rescue workers in a hospital.
SRC	Janson ist selbst ein **alter Hase** in seinem Metier, der Schauspielkunst.
REF	Janson is an **old hand** himself when it comes to his profession, the art of acting.
PBMT	Janon himself is an old hase in his painter, the artistic art.
NMT	Janson itself is an old hand in his subjects, the schauspielkunst.
SRC	Mit unserer Mitteilung vom letzten Sommer haben wir **den Stein ins Rollen** gebracht und demonstriert, dass Europa an der Erarbeitung eines internationalen Instruments beteiligt ist.
REF	Our communication of last summer enabled us to **get things up and running** and to demonstrate that Europe was participating in the drawing up of an international instrument.
PBMT	With our communication of last summer we have the ball rolling and demonstrated that Europe in the drafting of an international instrument is involved.
NMT	With our communication last summer, we launched the stone and demonstrated that Europe is involved in the development of an international instrument.

set, as follows:

$$UniPrec = \frac{\sum_{\mathbf{C} \in \{Candidates\}} \sum_{\text{unigram} \in \mathbf{C}} \text{count}_{clip}(\text{unigram})}{\sum_{\mathbf{C}' \in \{candidates\}} \sum_{\text{unigram}' \in \mathbf{C}'} \text{count}(\text{unigram}')} \tag{6.1}$$

where $\text{count}_{clip} = \min(\text{count}, \textit{max\_ref\_count})$. By computing the clipped count, we truncate each word's count so that it does not exceed the largest count observed in a reference for that word. Note that for this metric we have two references: The idiom translation as an independent expression, and the human-generated idiom translation in the target sentence.

6.5.3 Word-Level idiom accuracy

We also use another metric to evaluate the word-level translation accuracy of the idiom phrase. We use word alignments between source and target sentences to determine the number of correctly translated words. We use the following equation to compute the accuracy:

$$WAcc = \frac{H - I}{N} \tag{6.2}$$

where H is the number of correctly translated words, I is the number of extra words in the idiom translation, and N is the number of words in the gold idiomatic expression.

Table 6.6: Translation performance on the German idiom translation test set. *Word-level Idiom Accuracy* and *Unigram Precision* are computed only on the idiom phrase and its corresponding translation in the sentence.

	WMT08-16	**Idiom test set**		
Model	BLEU	BLEU	UniPrec	WAcc
PBMT baseline	20.2	19.7	57.7	71.6
NMT baseline	26.9	24.8	53.2	67.8
NMT SRC flag	25.2	22.5	64.1	73.2
NMT TGT flag	17.8	16.2	54.3	64.0

6.5.4 Evaluation results

Table 6.6 presents the results for the translation task using different metrics. Looking at the overall BLEU scores, we observe that baseline performance on the idiom-specific test set is lower than on the union of the standard test sets (WMT 2008-2016). While

the scores on these two data sets are not directly comparable, this result is in line with previous findings that sentences containing idiomatic expressions are harder to translate (Isabelle et al., 2017). We can also see that the performance gap is not as pronounced for a PBMT system, suggesting that phrase-based models are capable of *memorizing* the idiomatic phrases to some extent.

The NMT model using a special input flag to indicate the presence of an idiom in the source sentence performs better than PBMT but slightly worse than the NMT baseline in terms of BLEU. Despite this drop in BLEU performance, by examining the *unigram precision* and *word-level idiom accuracy* scores, we observe that this model generates more accurate idiom translations. When comparing this to having the idiom flag on the source or target side, we observe a significant difference: The experiment with the target flag performs the worst, partially because during inference, only the source sentence is available and hence there is no contextual signal to aid the model.

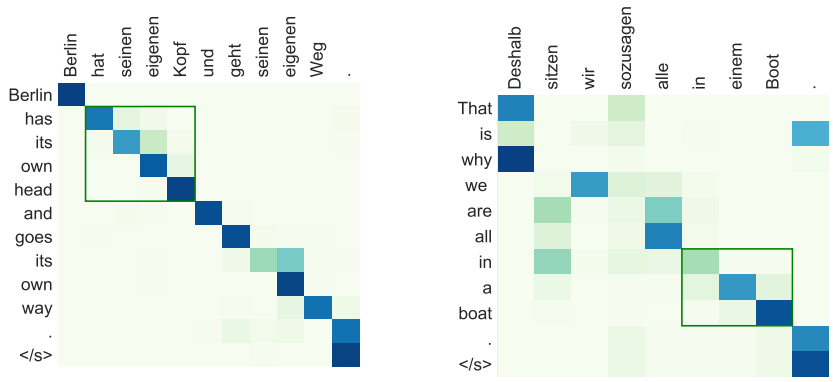


Figure 6.2: Attention visualization of the translation of two sampled German sentences. Darker color means higher weight. The blocked area marks the idiomatic expression and its generated translation. The reference translations are: (left) *"Berlin has a mind of its own and is doing its own thing."* and (right) *"We are therefore all in the same boat, so to speak."*

Figure 6.2 illustrates the attention distribution of the NMT model during translation of an example German sentence. We expected that in order to translate each word in the idiomatic expression correctly, the model would pay a noticeable degree of attention to the other words in the expression. However, we see that it does not happen and the model essentially translates the sentence word by word, i.e., literally. These preliminary experiments reiterate the problem of idiom translation with neural models, and in addition show that with a labeled data set, we can devise simple models to address this problem to some extent.

6.6 Conclusion

Motivated by our observations in Chapters 4 and 5 that illustrate the importance of local context on learning difficult words, we further explored in this chapter some shortcomings of current models. Concretely, we were interested in finding cases where NMT models are unsuccessful. One case of a complex language phenomenon is idiom translation where it is very difficult to infer the meaning of the phrase without explanation and only from the observed context.

To investigate the behaviour of NMT models in translating idiomatic expressions, we asked:

RQ3.1 *What are the challenges of idiom translation with neural models?*

> We identified two main challenges when translating idiom phrases, namely lack of dedicated data sets and lack of targeted evaluation metrics. To address this problem, we have extracted a parallel data set for training and testing idiom translation for German→English and English→German. In the test sets, we included sentences with at least one idiom on the source side. In the training set, we included a mixture of idiomatic and non-idiomatic sentences with labels to distinguish between the two. We release our new data sets which can be used to further investigate and improve NMT performance of idiom translation. Using our new resources, we performed preliminary translation experiments to evaluate the quality of idiom translation. Experiments on this test set showed PBMT models scored higher than NMT models on our metrics which explicitly measure idiom translation quality.

RQ3.2 *How is the translation quality of NMT influenced by idiomatic expressions?*

> We observed that even though the NMT model achieved a higher overall BLEU score, it performed worse on idiom translation metrics in comparison with PBMT model. Next, we studied whether a flag in the training data can help to distinguish between when a phrase is to be translated literally and when it should be translated idiomatically. Our experiments showed that adding a side flag during training improves the quality of idiom translation. However, we found that the BLEU score on standard test sets declined. Our experiments suggest that there is no correlation between overall BLEU scores and the localized precision of idiomatic phrase translations.

It allows us to return to our more general research question:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

To answer this question, we specifically examined non-compositional multiword expressions. Since the literal meaning of the components is different from the idiomatic

meaning of the entire expression, the model needs to know in which context to translate it literally and in which idiomatically. We showed that NMT models perform poorly on idiom translation despite their overall strong advantage over previous MT paradigms. We conclude that further research on idiom translation can benefit from having a dedicated data set.
In the next chapter, we continue investigating this question by examining cases where there are no complex linguistic phenomena, such as non-compositional phrases, in the observed context.

7

Volatilities of Neural Models

7.1 Introduction and research questions

In the previous chapters, we first investigated how to enhance the use of context to address some of the shortcomings of neural translation models. Then in Chapter 6, we showed that the translation of idiomatic phrases is challenging for the current NMT models. We saw that the scope of the observed context was not sufficient to infer the meaning of idioms. Based on these findings, in this chapter, we continue examining the following question:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

While the lack of suitable context exposes shortcomings in current models, we extend our research in this chapter to situations where appropriate data *is* available. We first look into the robustness of current translation models. Namely, we investigate what is the effect of small perturbations of the source sentence on the translation. Observing that in some cases translations change unexpectedly with these small perturbations, we study whether and to what extent it can be replicated and quantified with automatically modified test data. Concretely we ask:

RQ3.3 *How can contextual modifications during testing reveal a lack of robustness of translation models and affect the translation quality?*

To answer this question, we locally modify sentence pairs in the test set and identify examples where a trivial modification in the source sentence causes an 'unexpected change' in the translation. These modifications are generated conservatively to avoid insertion of any noise or rare words in the data (Section 7.4). Our goal is not to *fool* the NMT models, but instead, to identify common cases where the models exhibit unexpected behaviour and in the worst cases result in incorrect translations. We identify these unexpected and erroneous changes in the translation output as a sign of an underlying *volatility* of NMT models.

RQ3.4 *To what extent is a lack of robustness an indicator of a generalization problem in neural machine translation models?*

We investigate to what extent two current state-of-the-art NMT models are robust against changes in the input during inference. We observe that our modifications expose volatilities of both RNN and Transformer translation models in 26% and 19% of sentence variations, respectively. Our findings show how vulnerable current NMT models are to trivial linguistic variations, putting into question the generalization abilities of these models.

Organization. The chapter is organized as follows: Section 7.2 discusses prior works on the impact of noise on the performance of machine translation. In Section 7.3, we provide an example of unexpected behaviour of NMT models and discuss how it is different from the unexpected behaviour when encountering noise in the input text. In Section 7.4, we introduce our sentence variation generation approach and provide details on the experimental settings. Section 7.5 proposes various metrics to identify and quantify these unexpected changes and provides experimental results on a translation task. Finally, we discuss the conclusions and implications of this work in Section 7.6.

7.2 Noisy text translation

Recently, several approaches investigated NMT models when encountering noisy input and how *worst-case examples* of noisy input can 'break' state-of-the-art NMT models (Michel and Neubig, 2018). Noisy input text can cause mistranslations in most translation systems, and there has been growing research interest in studying the behaviour of translation systems when encountering noisy input (Li et al., 2019).

Belinkov and Bisk (2018) show that character-level noise in the input leads to poor translation performance. They propose to swap or randomize letters in a word in the input sentence. For instance, they change the word '*noise*' in the source sentence into '*iones*'. Lee et al. (2018) randomly insert words in different positions in the source sentence and observe that in some cases the translations are completely unrelated to the input. Michel and Neubig (2018) propose a benchmark data set for translation of noisy input sentences, consisting of noisy, user-generated comments on Reddit. The types of noisy input text they observe include spelling or typographical errors, word omission/insertion/repetition, and grammatical errors.

7.3 Volatility in machine translation

In the discussed works in Section 7.2, the focus of the research is on studying how the translation systems are not robust when handling noisy input text. In these approaches, the input sentences are semantically or syntactically incorrect which leads

to mistranslations. However, in this chapter, our focus is on input text that does *not* contain any types of noise. We modify input sentences in a way that the outcomes are still syntactically and semantically correct. We investigate how translation systems exhibit volatile behaviour in translating sentences that are extremely similar and only differ in one word without any noise injection. While it is to some extent expected that the performance of NMT models that are trained on predominantly clean but tested on noisy data deteriorates, other changes are more unexpected.

Table 7.1: Insertion of the German word *'sehr'* (English: *'very'*) in different positions in the source sentence results in substantially different translations. Note that all source sentences are syntactically correct and semantically plausible. We use a Transformer model trained on WMT data with 6 encoder and decoder layers and 8 attention heads. † indicates the original sentence from WMT 2017.

Source: *Ich bin* ___ [1] ***erleichtert*** *und* ___ [2] *bescheiden.*

[1]	[2]	NMT output
ϕ	ϕ	I am easier and modest.
ϕ	*sehr* †	I am **relieved** and very modest.
sehr	ϕ	I am very much easier and modest.
sehr	*sehr*	I am very easy and very modest.
		Reference
ϕ	ϕ	*I am relieved and humble.*
sehr	*sehr*	*I am very relieved and very humble.*

In this chapter, we explore unexpected and erroneous changes in the output of NMT models. Consider the simple example in Table 7.1 where the Transformer model is used to translate very similar sentences. Surprisingly, we observe that by simply altering one word in the source sentence—inserting the German word *'sehr'* (English: *'very'*)—an unrelated change occurs in the translation. In principle, an NMT model that generates the translation of the word *'erleichtert'* (English: *'relieved'*) in one context, should also be able to generalize and translate it correctly in a very similar context. Note that there are no infrequent words in the source sentence and after each modification, the input is still syntactically correct and semantically plausible. We call a model *volatile* if it displays inconsistent behaviour across similar input sentences during inference.

7.4 Variation generation

While there are various ways to automatically modify sentences, we are interested in simple semantic and syntactic modifications. These trivial linguistic variations should

have almost no effect on the translation of the rest of the sentence.

We define a set of rules to slightly modify the source and target sentences in the test data and keep the sentences syntactically correct and semantically plausible.

- **Delete**: A conservative approach to modifying a sentence automatically without breaking its grammaticality is to remove adverbs. We identify a list of the 50 most frequent adverbs in English and their translations in German.[1] For every sentence in the WMT test sets, if we find a sentence pair containing both a word and its translation from this list, we remove both words and create a new sentence pair.

- **Insert**: Randomly inserting words in a sentence has a high chance of producing a syntactically incorrect sentence. To ensure that sentences remain grammatical and semantically plausible after modification, we define a bidirectional n-gram probability for inserting new words as follows:

$$P(w_3 \mid w_1 w_2 w_4 w_5) = \frac{C(w_1 w_2 w_3 w_4 w_5)}{\sum_j C(w_1 w_2 w_j w_4 w_5)} \tag{7.1}$$

  w_3 is inserted in the middle of the phrase $w_1 w_2 w_4 w_5$, if the conditional probability is greater than a predefined threshold. The probabilities are computed on the WMT data. This simple approach, instead of using a more complex language model, serves our purposes since we are interested in inserting very common words that are already captured by the n-grams in the training data.

- **Substitute number**: Another simple yet effective approach to safely modifying sentences is to substitute numbers with other numbers. In this approach, we select every sentence pair from the test sets that contain a number and substitute the number i in both source and target sentences with $i + k$ where $1 \leqslant k \leqslant 5$. We choose a small range for change so that the sentences are still semantically correct for the most part and result in a few implausible sentences.

- **Substitute gender**: Finally, a local modification is to change the gender of the pronoun in the sentences. The goal of this modification is to investigate the existence and severity of gender bias in our models. This is inspired by recent approaches that have shown that NMT models learn social stereotypes such as gender bias from training data (Escudé Font and Costa-jussà, 2019, Stanovsky et al., 2019).

Note that in a minority of cases, these procedures can lead to semantically incorrect sentences. For instance, by substituting numbers we can potentially generate sentences such as *"She was born on October 34th"*. While this can cause problems for a reasoning

[1] Here, we use the `dict.cc` online dictionary.

task, it barely affects the translation task, as long as the modifications are consistent on the source and target side.

Table 7.2 shows examples of generated variations. We emphasize that only modifications with local consequences have been selected and we intentionally ignore cases such as *negation* which can result in wider structural changes in the translation of the sentence.

Table 7.2: Examples of different variations from WMT. $[w_i \backslash w_j]$ indicates that w_i in the original sentence is replaced by w_j. ϕ is the empty string.

Modification	**Sentence variations**
Delete	Some 500 years after the Reformation, Rome **[now\ϕ]** has a Martin Luther Square.
Insert	I loved Amy and she is **[ϕ\also]** the only person who ever loved me.
Subs number	I'm very pleased for it to have happened at Newmarket because this is where I landed **[30\31]** years ago.
Subs gender	**[He\She]** received considerable appreciation and praise for this.

We generate 10K sentence variations by applying these modifications to all sentence pairs in WMT test sets 2013–2018 (Bojar et al., 2018). We use RNN and Transformer models to translate sentences and their variations.

7.4.1 Experimental setup

In the translation experiments, we use the standard English↔German WMT-2017 training data (Bojar et al., 2018). We perform NMT experiments with two different architectures as described in Sections 2.4 and 2.5: RNN (Luong et al., 2015a) and Transformer (Vaswani et al., 2017). We preprocess the training data with Byte-Pair Encoding (BPE) using 32K merge operations (Sennrich et al., 2016c). Table 7.3 shows the case-sensitive BLEU scores as calculated by `multi-bleu.perl`.

Table 7.3: BLEU scores of different baseline models on the WMT news data for translation of German↔English.

	De-En			**En-De**		
	WMT16	WMT17	WMT18	WMT16	WMT17	WMT18
RNN	32.5	28.2	35.2	28.1	22.4	34.6
Transformer	36.2	32.1	40.1	33.4	27.9	39.8

RNN We use a 2-layer bidirectional attention-based LSTM model implemented in OpenNMT (Klein et al., 2017) trained with an embedding size of 512, hidden dimension size of 1024, and batch size of 64 sentences. We use Adam (Kingma and Ba, 2015) for optimization.

Transformer We also experiment with the Transformer model (Vaswani et al., 2017) implemented in OpenNMT. We train a model with 6 layers, the hidden size is set to 512, and the filter size set to 2048. The multi-head attention has 8 attention heads. We use Adam (Kingma and Ba, 2015) for optimization. All parameters are set based on the suggestions by Klein et al. (2017) to replicate the results of the original paper.

7.5 Unexpected and erroneous changes

The modifications described above generate sentences that are extremely similar and hence are expected to have a very similar difficulty of translation. First, we evaluate the NMT models on how robust and consistent they are in translating these sentence variations rather than their absolute quality. Next, we perform manual evaluation on the translation outputs to assess the impact of unexpected changes on translation quality.

7.5.1 Deviations from original translation

The variations are aimed to have minimal effect on changing the meaning of the sentences. Hence, major changes in the translations of these variations can be an indication of volatility in the model. To assess whether the proposed sentence variations result in major changes in the translations, we measure changes in the translations of sentence variations with Levenshtein distance (Levenshtein, 1966). We also use the first and last positions of change in the translations, which represents the span of changes.

Ideally, with our simple modifications, we expect a value of zero for the span of change and a value of at most 2 for the Levenshtein distance for a translation pair. This indicates that there is only one token difference between the translation of the original sentence and the modified sentence.

We define two types of changes based on these measures: *minor* and *major*. We choose the threshold to distinguish between minor and major changes conservatively to allow for more variations in the translations. The change in translations is empirically considered *major* if both metrics are greater than 10, and *minor* if both are less than 10. Note that edit distances and spans are based on BPE subword units.

With two very similar source sentences, we expect the Levenshtein distance and span of change between translations of these sentences to be small. Figure 7.1 shows the results for the RNN and Transformer model. While the majority of sentence variations have minor changes, a substantial number of sentences, 18% of RNN and 13% of Transformer translations, result in translations with major differences. This is a

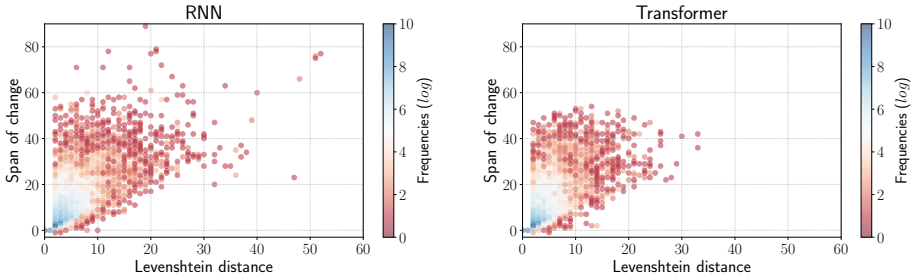


Figure 7.1: *Levenshtein distance* and *span of change* between translations of sentence variations for RNN and Transformer. The majority of sentence variations fall into the category of *minor* changes between translations (blue area). However, a surprising number of cases have significant changes (red area). RNN exhibits a slightly more unstable pattern i.e., sentence variations with large edit differences and large spans of change.

surprising indication of volatility since these trivial modifications, in principle, should only result in minor and local changes in the translations.

Table 7.4: An example of the generated variations of an English sentence and different sentence-level metrics for the translation of each variation. We compute the oscillation range for each sentence in the test data.

Source:	*Mr Ivanov took up the post in December ♠.*
Reference:	*Ivanov nahm den Posten im Dezember ♠ an.*
♠	NMT output
2012	Herr Ivanov hat den Beitrag im Dezember 2012 übernommen.
	bleu=22.78 meteor=60.54 ter=36.36 LengthRatio=109.09
2013	Herr Ivanov nahm den Beitrag im Dezember 2013 auf.
	bleu=43.67, meteor=66.89, ter=27.27, LengthRatio=109.09
2014	Herr Ivanov nahm den Beitrag im Dezember 2014 auf.
	bleu=43.67, meteor=66.89, ter=27.27, LengthRatio=109.09
Oscillation range:	*bleu=20.9, meteor=6.4, ter=9.1, LengthRatio=0*

7.5.2 Oscillations of variation in translations

In this section, we look into various sentence-level metrics to further analyze the observed behaviour. In particular, we focus on the *substitute numbers* modification since with this modification, we can easily generate numerous variations of the same sentence. Having a high number of variations for each sentence gives us the opportunity

of observing oscillations of various string matching metrics.

We use sentence-level BLEU, METEOR (Denkowski and Lavie, 2011), TER (Snover et al., 2006), and LengthRatio to quantify changes in the translations. LengthRatio represents the translation length over reference length as a percentage. For a given source sentence, we define the *oscillation range* as changes in the sentence-level metric for the translations of all variations of the sentence (see Table 7.4 for an example).

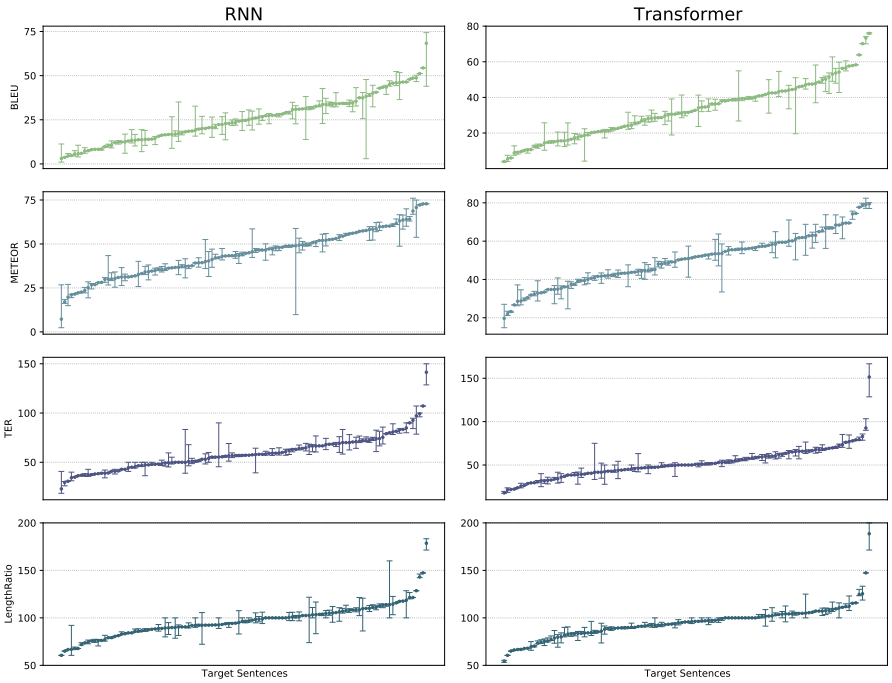


Figure 7.2: Oscillations of various sentence-level attributes for randomly sampled sentences from our test data and their *substitute number* variations. The data points are the mean values for all variations of each sentence, and the error bars indicate the range of oscillation of the metrics. The x-axis represents test sentence instances, sorted based on the corresponding metric. Ideally, each data point should have zero oscillation.

While sentence-level metrics are not reliable indicators of translation quality, they do capture fluctuations in translations. With the variations we introduced, in theory, there should be no fluctuations in the translations. Figure 7.2 and Table 7.5 provide the results. We observe that even though these sentence variations differ by only one number, there are many cases where an insignificant change in the sentence results in unexpectedly large oscillations. Both RNN and Transformer exhibit this behaviour to a certain extent.

Table 7.5: Mean oscillations for *substitute number* variations. In theory, the variations should result in zero oscillations for every metric.

	BLEU	METEOR	TER	LengthRatio
RNN	4.0	3.8	5.2	5.3
Transformer	3.8	3.3	4.2	3.4

7.5.3 The effect of volatility on translation quality

While edit distances, spans of change, and oscillation in variations provide some indication of volatility, they do not capture all aspects of this unexpected behaviour. It is also not entirely clear what effect these unexpected changes have on translation quality. To further investigate this, we also perform two manual evaluations by eight fellow PhD students working on information and language processing systems. The native language of the annotators consists of English, Dutch, and Chinese. All non-native annotators use English as a second language. Our manual evaluation does not require familiarity with the German language.

In the first evaluation, we provide annotators with a pair of sentence variations and their corresponding translations and ask them to identify the differences between the two sentence pairs. In the second evaluation, we additionally provide the source sentences and reference translations and ask the annotators to rank the sentence variations based on the translation quality similar to Bojar et al. (2016). In total, the annotators evaluated 400 randomly selected sentence quadruplets.

Table 7.6: Definitions of different labels of changes and examples for each category. The annotators identified these differences between the translations.

Label	Definition	Example
`Word form`	One or more words are different in form but belong to the same lexeme.	*observe* → *observation*
`Reordered`	One or more words are reordered in the translation sentence.	*he said go* → *go, he said*
`Paraphrased`	One word is replaced with a synonym or a section of the sentence is paraphrased.	*first six months* → *first half of the year*
`Add/Remove`	One or more words are added or dropped from the translation sentence.	*will participate* → *will also participate*
`Other`	Other changes in the translation sentence.	*were torn through* → *have been bypassed*

Table 7.6 shows the identified categories of changes from annotators' labels. The

Table 7.7: A random sample of sentences from the WMT test sets and our proposed variations shown with 'unexpected change' annotations ($\Delta Translation$). The cases where the unexpected change leads to a change in translation quality are marked in column $\Delta Quality$. $[w_i \backslash w_j]$ indicates that w_i in the original sentence is replaced by w_j. S is the original and modified source sentence, R is the original and modified reference translation, T is the translation of the original sentence, and T_m is the translation of the modified sentence. Differences in translations related to annotations are underlined.

S	Coes letztes Buch "Chop Suey" handelte von der chinesischen Küche in den USA, während Ziegelman in ihrem Buch "**[97\101]** Orchard" über das Leben in einem Wohnhaus an der Lower East Side aus der Lebensmittelperspektive erzählt.
R	Mr. Coe's last book, "Chop Suey," was about Chinese cuisine in America, while Ms. Ziegelman told the story of life in a Lower East Side tenement through food in her book "**[97\101]** Orchard."
T	Coes's last book, "Chop Suey," was about Chinese cuisine in the US, while Ziegelman, in her book "97 Orchard" talks about living in a lower East Side.
T_m	Coes last book "Chop Suey" was about Chinese cuisine in the United States, while Ziegelman writes in her book "101 Orchard" about living in a lower East Side.
$\Delta Translation$:	`[reordered] [paraphrased]` | $\Delta Quality$: `No`
S	Man hält **[bereits\ϕ]** Ausschau nach Parkbank, Hund und Fußball spielenden Jungs und Mädels.
R	You are **[already\ϕ]** on the lookout for a park bench, a dog, and boys and girls playing football.
T	We are already looking for Parkbank, dog and football playing boys and girls.
T_m	Look for Parkbank, dog and football playing boys and girls.
$\Delta Translation$:	`[word form] [add/remove]` | $\Delta Quality$: `Yes`
S	Bei einem Unfall eines Reisebusses mit **[43\45]** Senioren als Fahrgästen sind am Donnerstag in Krummhörn (Landkreis Aurich) acht Menschen verletzt worden.
R	On Thursday, an accident involving a coach carrying **[43\45]** elderly people in Krummhörn (district of Aurich) led to eight people being injured.
T	In the event of an accident involving a coach with 43 senior citizens as passengers, eight people were injured on Thursday in Krummaudin (County Aurich).
T_m	In the event of an accident involving a 45-year-old coach as a passenger, eight people were injured on Thursday in the district of Aurich.
$\Delta Translation$:	`[word form] [add/remove] [other]` | $\Delta Quality$: `Yes`
S	Es ist ein anstrengendes Pensum, aber die Dorfmusiker helfen **[normalerweise\ϕ]**, das Team motiviert zu halten.
R	It's a backbreaking pace, but village musicians **[usually\ϕ]** help keep the team motivated.
T	It's a demanding child, but the village musicians usually help keep the team motivated.
T_m	It is a hard-to-use, but the village musician helps to keep the team motivated.
$\Delta Translation$:	`[word form] [other]` | $\Delta Quality$: `Yes`

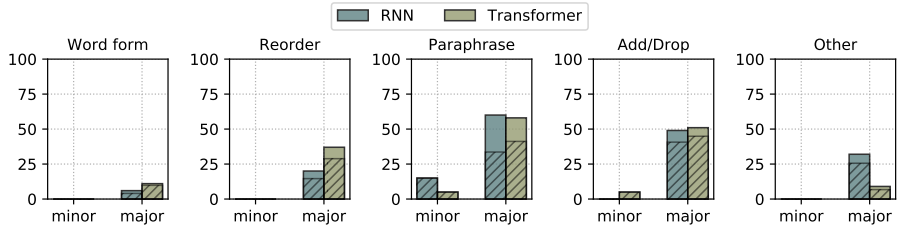


Figure 7.3: Categories of unexpected changes in the translation of sentence variations as provided by annotators. The percentage of sentence variations with *minor* and *major* edit differences, as defined in 7.5.1, are shown separately. The hatched pattern indicates the ratio of sentence variations for which the translation quality changes.

main types of unexpected changes identified by the annotators are a 'change of word form', e.g., verb tense, 'reordering of phrases', 'paraphrasing' parts of the sentence, and an 'other' category, e.g., preposition. A sentence pair can have multiple labels based on the types of changes. Table 7.7 provides examples from the test data.

Statistics for each category of unexpected change are shown in Figure 7.3. Our first observation is that, as to be expected, there are very few 'unexpected changes' when two variations lead to translations with *minor* differences. Interestingly, the vast majority of changes are due to paraphrasing and adding or dropping of words. Comparing the performance of the RNN and Transformer model, we see that both RNN and Transformer display inconsistent translation behaviour. From the annotators' assessments, we find that in 26% and 19% of sentence variations, the modification results in a change in translation quality for the RNN and Transformer model, respectively. From the manual evaluations, we conclude that the oscillations in translation outputs captured by our metrics indeed point to harmful changes in translation quality. This behaviour is not exposed by standard test sets and evaluation metrics.

7.5.4 Generalization and compositionality

Because of their ability to generalize beyond their training data, deep learning models achieve exceptional performances in numerous tasks. The generalization ability allows translation systems to generate long sentences not seen before. Recently there has been some interest in understanding whether this performance depends on recognizing shallow patterns, or whether the networks are indeed capturing and generalizing linguistic rules (Linzen et al., 2016, Chowdhury and Zamparelli, 2018).

The capability of generalization of current deep learning models can be interpreted as whether compositionality arises in learning problems where the compositional structure has not been explicitly declared. The principle of compositionality (Frege, 1892) has been extremely influential throughout the history of formal semantics and cognitive science with many arguments for and against it (Montague, 1974, Pelletier, 1994,

Janssen, 2001).

In simple terms, compositionality can be defined as the ability to construct larger linguistic expressions by combining simpler parts. For instance, if a model applies the correct compositional rules to understand *'John loves Mary'*, it must also understand *'Mary loves John'* (Fodor and LePore, 2002).

Investigating the compositional behaviour of neural networks in real-world natural language problems is a challenging task. Recently, several approaches have studied deep learning models' understanding of compositionality in natural language by using synthetic and simplified languages (Andreas, 2019, Chevalier-Boisvert et al., 2019). Hupkes et al. (2020) designed theoretically grounded tests based on different interpretations of compositionality. Their experiments showed that the current state-of-the-arts neural network architectures struggle to capture different aspects of compositionality in language and there is a need for a more extensive set of evaluation criteria to evaluate these models. Lake and Baroni (2018) introduced the SCAN data set consisting of simplified natural language commands and their translations into sequences of actions. They showed that, when there is a systematic difference between training and test data, neural models fail to generalize because they lack the ability to extract systematic rules from the training data. Baroni (2019) observed that current models seem to be able to generalize without any compositional rules. He argued that to a certain extent neural networks can be productive without being compositional.

Although we do not specifically look into the compositional potential of translation systems, we are inspired by compositionality in defining our modifications. We argue that the observed volatile behaviour of the translation systems in this chapter is a side effect of current models not being compositional. If a translation system has a good understanding of the underlying structures of the sentences *'Mary is 10 years old'* and *'Mary is 11 years old'*, it must also translate them very similarly regardless of the accuracy of the translation. While current evaluation metrics capture the accuracy of the NMT models, these volatilities go unnoticed.

Current neural models are successful in generalizing without learning any explicit compositional rules, however, our findings show that they still lack robustness. We highlighted this lack of robustness in this chapter and suspect that it is associated with these models' lack of understanding of the compositional nature of language.

7.6 Conclusion

Motivated by our findings in the previous chapters, we continued investigating the circumstances where current models do not perform as expected. Specifically, we are interested in cases where the semantic and syntactic complexity of the sentence remains unchanged with minor modifications to the observed context. Hence we investigated:

RQ3.3 *How can contextual modifications during testing reveal a lack of robustness of translation models and affect the translation quality?*

We studied an unexpected and erroneous behaviour in current NMT models by examining various metrics to quantify oscillations in translations of very similar sentences. We show that even with minor modifications preserving the grammaticality and plausibility of the sentence, we can effectively identify a surprising number of cases where the translations of extremely similar sentences are unexpectedly different. Our experiments on our test set showed that current NMT models are not completely robust and they expose their weaknesses when probed with specific test cases.

RQ3.4 *To what extent is a lack of robustness an indicator of a generalization problem in neural machine translation models?*

Models that have compositional understanding of the world are capable of generalizing to unseen composite cases (Lake and Baroni, 2018, Cogswell et al., 2019). We proposed an approach to examine the compositional understanding and measure the generalization capability of NMT models. We did so by introducing a specific test set and various evaluation metrics. If a model is capable of compositional understanding, it should have hardly any oscillations in the translation outputs on this test set. However, once we probed these models with extensive test cases, we observed that they do in fact exhibit unexpected changes in the translation outputs. Our analyses showed that both RNN and Transformer models exhibit volatile behaviour with changes in translation quality for 26% and 19% of sentence variations, respectively. Our experiments highlighted the need for a comprehensive evaluation setup for deeper analyses of current neural models.

This concludes the main research question of this chapter as follows:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

To answer this question, we examined the vulnerability of current NMT models to small changes to the observed context. Primarily, we focused on modifications that do not introduce new linguistic complexities for the translation model. We proposed a simple approach to modifying standard test sentences without introducing noise. By creating this data set, we can automatically measure if NMT models lack robustness and exhibit volatile behaviour. We observed that neural models, even with high performances on standard test sets, struggle in showing compositional understanding and suffer from a lack generalization.

8 Conclusions

In this thesis, we explored the role of context in machine translation as well as lexical modeling, using deep learning frameworks. With the increase of computational power of machines and the availability of data, progress in deep learning has focused on developing more advanced models. In fact, with the increasing amount of training data, the performance of many computer vision and language understanding tasks has increased significantly. In this thesis, we are motivated by *what* these models learn from the available data and *how* we can use this information to resolve linguistic challenges that arise from statistical learning from data. Specifically, we investigated the influence of contextual cues in understanding different words and proposed several approaches that improve how these models learn from context.

Firstly, we looked into ambiguous words and studied how document-level context assists in distinguishing different meanings of a word. Next, we focused on the influence of context in the bilingual setting of machine translation. We examined how neural translation models use context to learn and transfer meaning and showed that by diversifying data for difficult words, we can improve translation quality. In order to identify difficult words, we first looked at the data distribution and specifically targeted rare words. Since there is a lack of diverse contexts in the training data for rare words, the translation of them is challenging. Next, we investigated the failures of a trained model to identify difficult words. These difficult words are words that the model has low confidence in predicting after training on a sizable amount of data. We identified difficult words for an NMT model and performed data augmentation targeting these words. By creating new contexts for difficult words, we improved the generation capability of the NMT model and the translation quality.

Next, we addressed the shortfalls of relying only on the contexts observed in the training data to learn the meanings of words. We examined under which conditions context is not enough for capturing various linguistic phenomena. In particular, we studied the interesting case of idiom translation and showed that current NMT models often fail to capture such nuances. Neural networks optimize the learning process on the available data and the lack of data for complex linguistic phenomena such as idiom

translation is an obstacle for developing stronger models.

Finally, we raised more general questions about the learning capabilities of current state-of-the-art translation models. We analyzed how these models fail unexpectedly even in cases where there are no evident complex linguistic phenomena. By introducing simple contextual modifications, we identified an underlying generalization problem of state-of-the-art translation models.

In the following section, we revisit our research questions and summarize the main findings of this thesis. We then propose a number of questions that remain open for further exploration.

8.1 Main findings

In Chapter 3, we started with a preparatory question on the importance of context when learning word representations for difficult words. Going past local neighboring words, we looked at document-level contexts by asking:

RQ1.1 *To what extent can distributions over word senses be approximated by distributions over topics of documents without assuming these concepts to be identical?*

To answer this question, we investigated whether document-level information is an adequate contextual cue to help distinguish between different senses of a word. We experimented with a hierarchical Dirichlet process for modeling document topics which generated two sets of distributions that we used in our methods: distributions over topics for words in the vocabulary and distributions over topics for documents in the corpus. We observed that these distributions distinguish between senses of words. Next, we examined how we can leverage this information to learn word representations by asking:

RQ1.2 *How can we exploit document-level topics to distinguish between different meanings of a word and learn the corresponding representations?*

We found that the distribution over topics is different for different senses of an ambiguous word. This motivated us to combine this distribution with the Skipgram model to provide information on word senses to the embeddings. To achieve this, we devised three model variations that learned multiple representations per word based on the assigned topic in different contexts. We then evaluated these embeddings by asking:

RQ1.3 *What are the advantages of using document-level topics in learning multiple representations per word?*

We evaluated word embeddings on the word similarity task and observed slight improvements under different settings. However, there was no clear winner across

all data sets. Since word similarity data sets consider individual words in isolation and do not provide any contexts, we then evaluated the embeddings in a more context-aware setting. Our evaluation on the lexical substitution task showed that topic distributions capture word senses to a large extent. Moreover, we obtained statistically significant improvements in a lexical substitution task without using any syntactic information. The best results were achieved by our HTLE model which learns topic-sensitive representations by hard-labeling topics to target words and not using generic representations.

These three sub-questions together allowed us to answer our first main research question:

Research Question 1: *Can document-level topic distribution help infer the meaning of a word?*

Our experiments showed that we can use document-level topic distribution to improve word representation learning. To summarize, we introduced an approach in Chapter 3 to learn topic-sensitive word representations that exploits the document-level context of words and does not require annotated data or linguistic resources. Additionally, we also learned representations for topics and our qualitative analyses showed that words belonging to the same topics also tend to be clustered together.

Having observed the effectiveness of wider context in capturing polysemy in word embeddings, we investigated in Chapter 4 the impact of context on the translation of difficult words. We first asked:

RQ2.1 *How can we successfully augment the training data with diverse contexts for rare words?*

By leveraging language models trained on large amounts of monolingual data, we generated new sentence pairs containing rare words in new contexts. We first confirmed that the translation performance is primarily affected by low-frequency and out-of-vocabulary words. Our analysis in Section 4.6 further showed that the poor translation quality of rare words is a result of a lack of diverse training examples. To address this problem, we proposed a method to automatically generate new contexts for these words. Next, we used this data to augment the parallel corpus used to train the translation model and re-trained the entire system. Our results showed that this approach improves the representations of rare words learned by the model and consequently increases the number of times the model generates these words correctly.

We observed substantial improvements in simulated low-resource English→German and German→English settings.

A natural follow-up question is whether we can perform augmentation during test time as well. So we asked:

RQ2.2 *Do rare words benefit from augmentation via paraphrasing during test time?*

Our experiments showed that augmentation at test time reduces the number of `unks` in the output and results in more fluent sentences. We cannot modify a source sentence resulting in a change of meaning without modifying the reference translation as well. Since we do not have access to the reference translations during inference, any alteration we made to the source sentence must keep the meaning of the sentence unchanged. In Section 4.7, we introduced a substitution via paraphrasing method to replace rare and out-of-vocabulary words in source sentences. We used different paraphrase knowledge resources to do this: WordNet, PPDB, GermaNet, CBOW, and our embedding approach proposed in Chapter 3. We gained improvements in BLEU scores while significantly reducing the number of `unk` generated in the target output.

In Chapter 5, we continued addressing our second research question by further analyzing the effectiveness of additional context for learning the meaning of a word. Rather than looking at the distribution of the training data, we investigated the behaviour of a neural MT system during training by asking:

RQ2.3 *Do signals from the NMT model help identify low-confidence words that could benefit from additional context?*

We found that signals from failures of the model can be used to identify where the model is not learning satisfactorily. To investigate this question we first explored different aspects of other influential augmentation methods, in particular back-translation, in Section 5.4. Our analyses showed that the quality of the synthetic data generated with a reasonably good model has a small impact on the effectiveness of back-translation, but that the ratio of synthetic to real training data plays a more important role. With a higher ratio, the model becomes biased towards noise in the synthetic data and unlearns the parameters. Next, we looked into which words benefit most from additional back-translated data. We observed that words with high prediction losses in the original model undergo the most changes after training with synthetic data. Our findings showed that with the addition of contexts for words with high prediction loss, we can increase the overall accuracy of the model.

Equipped with this information, we addressed the following question:

RQ2.4 *How can we successfully apply data selection of monolingual data to diversify the contexts of low-confidence words?*

In Section 5.6, we proposed our sampling approach targeting words that are difficult to predict. These words benefit the most from a more diverse context after augmentation. Our approach included several variants of using the prediction loss for identifying relevant sentences to back-translate. We also used the contexts

of difficult words by incorporating context similarities as a feature to sample sentences for back-translation. We discovered that using the prediction loss to identify weaknesses of the translation model and providing additional synthetic data targeting these shortcomings improved the translation quality of German→English and English→German translations.

Having discussed our specific sub-questions, we return to our more general question:

Research Question 2: *How is the translation quality of a word influenced by the availability of diverse contexts?*

In Chapters 4 and 5, we investigated the effect of the availability of data, where *the lack of* diverse contexts during training causes difficulties. Rare words, by definition, suffer from a lack of diverse context. Our studies showed that both translating and generating rare words is a challenging task with NMT models. We then continued with the impact of diverse contexts on translation quality in NMT. In particular, we focus on the back-translation method and the synthetic contexts that are generated with a reverse trained NMT model. We investigated this method and explored alternatives to select the monolingual data in the target language that is to be back-translated into the source language to improve translation quality. Both data augmentation approaches proposed in Chapters 4 and 5 lead to improvement of translation quality by generating diverse contexts for training.

Continuing our research on the impact of context on the quality of NMT models, we subsequently investigated some of its limitations. In Chapter 6, we looked into idiom translation with neural models. Since the literal meaning of the components is different than the idiomatic meaning of the entire expression, the model needs to know in which context to translate it literally and in which idiomatically. Neural MT, in particular, has been shown to perform poorly on idiom translation despite its overall strong advantage over previous MT paradigms (Isabelle et al., 2017).

We began by asking:

RQ3.1 *What are the challenges of idiom translation with neural models?*

One of the main challenges of studying idiom translation is the lack of dedicated and labeled data for evaluation and analysis. As an essential step towards answering this question, we required a test set explicitly tailored to evaluating idiom translation quality. To this end, we harvested a parallel data set for training and testing idiom translation for German→English and English→German. The test sets included sentences with at least one idiom on the source side while the training data is a mixture of idiomatic and non-idiomatic sentences with labels to distinguish between the two. Using our new resources, we performed preliminary

translation experiments and proposed different metrics to evaluate the quality of idiom translation.

We then evaluated the translation quality of neural models on our idiom translation test set:

RQ3.2 *How is the translation quality of NMT influenced by idiomatic expressions?*

We observed that the NMT model achieved a higher overall BLEU score but scored lower in idiom translation metrics. This is in agreement with previous works on investigating idioms as one of the weak points of neural models (Shao et al., 2018). Since there are no explicit signals in the sentence to identify when a phrase is to be translated literally and when it is to be translated idiomatically, we examined whether such a signal would help. Our experiments in Section 6.5 showed that adding a flag during training to indicate idiomatic use improves the quality of idiom translation in general. However, the overall BLEU score declined slightly. We concluded that there is little correlation between overall BLEU scores and the localized precision of idiomatic phrase translation. Our experiments showed that idiom translation can benefit from having a tailored development and test set and more specific metrics for evaluation.

Our next research question focused on other cases where NMT models fail to generate a correct translation given the observed context. In Chapter 7, we first examined how to expose this shortcoming in translation models by asking:

RQ3.3 *How can contextual modifications during testing reveal a lack of robustness of translation models and affect the translation quality?*

We studied the behaviour of NMT models and observed an unexpected but recurring pattern: A model that translates a given phrase correctly in a sentence fails to translate it correctly in another sentence which is very similar to the first. To explain these observations, we introduced new quantitative metrics measuring such unexpected changes. These metrics measured oscillations in translations of very similar sentences. Our experiments further showed that even with minor modifications preserving the grammaticality and plausibility of the sentence, we can effectively identify a surprising number of cases where the translations of extremely similar sentences are very different. By further manual inspection, we observed that these differences included ‘changes of word forms’, ‘reorderings of phrases’, ‘changes by paraphrasing’, ‘adding or dropping words from the translations’, and ‘semantically different translations’. We concluded that with contextual modifications during testing, we can reveal a lack of robustness of translation models.

Knowing this shortcoming of NMT models, we then asked:

RQ3.4 *To what extent is a lack of robustness an indicator of a generalization problem in neural machine translation models?*

We created a test set from our the sentence variation method proposed in Section 7.4. Next, we examined the robustness of current NMT models with this new test data and various evaluation metrics. Our analyses showed that both RNN and Transformer models exhibit volatile behaviour with changes in translation quality. We observed that these models fall short of capturing the compositional nature of the language, which confirms previous findings on the lack of compositional behaviour of NMT systems (Lake and Baroni, 2018). Additionally, we found that current evaluation sets do not spot the unexpected patterns we identified with our test data.

These answers allow us to return to our more general question:

Research Question 3: *To what extent are neural translation models vulnerable as a result of relying on the observed context in the training data to infer meaning?*

To study the influence of the observed context, we investigated how NMT models handle translating *non-compositional* and *compositional* events. Our findings in Chapters 6 and 7 showed that even well-performing models with high translation quality still suffer from a number of problems in both cases.

First, we looked into non-compositional multiword expressions or idioms. Idiom translation is one of the more difficult challenges of machine translation. To study this, in Section 6.3, we created a data set for training and testing idiom translation. We found that with the observed context, current NMT models struggle with translating non-compositional expressions. PBMT models on the other hand, despite underperforming on general-purpose data sets, achieve better idiom translation quality.

Next, we investigated the impact of observed context on translating compositional expressions. To achieve this, we defined a test set and evaluation metrics and investigated the NMT model's behaviour in this particular setting. In creating this test set, we focused on modifications that do *not* explicitly introduce new challenges for the translation model. Large oscillations in translations in the test set are an indication that the models do not capture composition in a systematic way, but often rely on memorized patterns to translate new sentences. In Section 7.4, we proposed a simple approach to modify standard test sentences without introducing noise and hence generating semantically and syntactically correct variations. Our findings showed that even well-performing models with high translation quality are prone to this problem and more extensive evaluations are necessary for assessing a system. We believe that our insights will be useful for developing more robust NMT models.

8.2 Future work

In this thesis, we studied how context is used by neural models to learn and generate words in a language and proposed methods to improve it. While this work highlighted the potentials of neural translation models in learning from data and studied some cases where they fall short, there are still many questions left to explore. Here, we discuss a few of these questions:

Are evaluation metrics for generation tasks still adequate? Since manual evaluation is very expensive, several automatic metrics have been designed to evaluate generation tasks as described in Section 2.6, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006).

These metrics compare an automatically generated candidate with a reference that is created manually. Matching the words in the candidate and reference gives us useful and reproducible results on the performance of a system and has helped advance tasks such as machine translation. However, with the improvements in generation models, the errors in the candidates are becoming increasingly subtle and idiosyncratic and existing metrics are not fully capable of highlighting them. While this issue has recently been addressed by Chen et al. (2019) and Ribeiro et al. (2020), further research on approaches and metrics that highlight deeper problems in generation models and go beyond n-gram based matching are necessary.

How can we learn complex nuances and structures of language? Many neural models still struggle with complex language structures, such as idiomatic expressions, in their respective tasks. One reason is that many interesting phenomena in language do not occur frequently. As a result, exclusively data-driven models fail to capture these nuances. One way of addressing this issue requires constructing data sets that are both adequately large and of high quality. With the availability of data sets that target specific linguistic phenomena, the process of learning them will be measurable and developing models targeting these challenges will be more accessible.

How compositional are sequence-to-sequence models? Compositionality is the ability to construct larger linguistic expressions by combining simpler parts (Frege, 1892, Fodor and Lepore, 1992). Investigating the compositional behaviour of neural networks in real-world natural language problems is a challenging task.

Current NMT models deliver high average translation quality provided enough training data and a good training-test domain match. It is not entirely clear, though, how much of this success stems from learning the underlying compositional structure of the sentence. In general, many traits of neural models are still a black box which hinders advancements to some extent. Recently, a few studies have focused on studying the

level of compositionality in neural sequence-to-sequence models using toy data sets (Lake and Baroni, 2017, Hupkes et al., 2020). Creating evaluation paradigms to further analyze this aspect can potentially lead to a better understanding of the inner workings of these influential models.

Bibliography

I. Abdulmumin, B. S. Galadanci, and A. Isa. Using self-training to improve back-translation in low resource neural machine translation, 2020. URL https://arxiv.org/abs/2006.02876. (Cited on page 56.)

A. Agarwal and A. Lavie. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W08-0312. (Cited on page 26.)

R. Agrawal, V. Chenthil Kumar, V. Muralidharan, and D. Sharma. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1048. (Cited on page 101.)

R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1388. (Cited on pages 23 and 57.)

J. Andreas. Measuring compositionality in representation learning. *CoRR*, abs/1902.07181, 2019. URL http://arxiv.org/abs/1902.07181. (Cited on page 124.)

M. Artetxe, G. Labaka, and E. Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, Oct. 2018a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1399. (Cited on pages 29 and 57.)

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, Apr. 2018b. URL https://openreview.net/pdf?id=Sy2ogebAW. (Cited on page 57.)

M. Artetxe, G. Labaka, and E. Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1019. (Cited on page 57.)

A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2011. URL http://www.aclweb.org/anthology/D11-1033. (Cited on pages 75 and 77.)

A. Axelrod, Y. Vyas, M. Martindale, M. Carpuat, and J. Hopkins. Class-based n-gram language difference models for data selection. In *IWSLT (International Workshop on Spoken Language Translation)*, pages 180–187, 2015. URL http://workshop2015.iwslt.org/downloads/IWSLT%5F2015%5FRP%5F17.pdf. (Cited on page 15.)

L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450. (Cited on page 25.)

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.0473. (Cited on pages 16, 19, 21, 22, 56, and 102.)

S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0909. (Cited on pages 26 and 134.)

M. Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *CoRR*, abs/1904.00157, 2019. URL http://arxiv.org/abs/1904.00157. (Cited on page 124.)

M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010. URL https://www.aclweb.org/anthology/J10-4006. (Cited on page 16.)

M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P14-1023. (Cited on page 16.)

L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. 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, Aug. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-5301. (Cited on page 14.)

Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL http://people.csail.mit.edu/belinkov/assets/pdf/iclr2018.pdf. (Cited on pages 7 and 114.)

Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-1080. (Cited on page 21.)

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, Mar. 2003. URL http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. (Cited on page 29.)

L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1025. (Cited on page 99.)

G. Blackwood, M. Ballesteros, and T. Ward. 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, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1263. (Cited on page 57.)

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2301. (Cited on pages 62 and 121.)

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4717. (Cited on pages 78 and 103.)

O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-6401. (Cited on page 117.)

J. Boyd-Graber, D. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, Prague, Czech Republic, June 2007.

Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D07-1109. (Cited on page 3.)

T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D07-1090. (Cited on pages 14 and 15.)

D. Britz, A. Goldie, M.-T. Luong, and Q. Le. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1151. (Cited on page 21.)

S. Brody and M. Lapata. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 103–111, Athens, Greece, Mar. 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E09-1013. (Cited on pages 32 and 33.)

E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P12-1015. (Cited on pages 38 and 40.)

F. Burlot and F. Yvon. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6315. (Cited on page 14.)

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, Apr. 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E06-1032. (Cited on page 26.)

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-0218. (Cited on page 14.)

D. S. Chaplot and R. Salakhutdinov. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5062–5069, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17415/16787. (Cited on page 3.)

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. URL https://arxiv.org/abs/1405.3531. (Cited on page 53.)

A. Chen, G. Stanovsky, S. Singh, and M. Gardner. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-5817. (Cited on page 134.)

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996. URL https://www.aclweb.org/anthology/P96-1041. (Cited on page 82.)

C. Cherry, G. Foster, A. Bapna, O. Firat, and W. Macherey. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1461. (Cited

on page 57.)

M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJeXCo0cYX. (Cited on page 124.)

K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014. URL https://arxiv.org/abs/1409.1259. (Cited on pages 16 and 19.)

S. A. Chowdhury and R. Zamparelli. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1012. (Cited on page 123.)

M. Cogswell, J. Lu, S. Lee, D. Parikh, and D. Batra. Emergence of compositional language with deep generational transmission, 2019. URL https://openreview.net/forum?id=r1gzoaNtvr. (Cited on page 125.)

R. Collobert, A. Hannun, and G. Synnaeve. A fully differentiable beam search decoder. *International Conference on Machine Learning*, pages 1341–1350, 2019. URL https://arxiv.org/abs/1902.06022. (Cited on page 22.)

M. R. Costa-jussà and J. A. R. Fonollosa. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-2058. (Cited on page 57.)

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. URL https://arxiv.org/abs/1805.09501. (Cited on page 55.)

A. Currey and K. Heafield. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong, Nov. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-5610. (Cited on page 14.)

A. Currey, A. V. M. Barone, and K. Heafield. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, 2017. URL https://www.aclweb.org/anthology/W17-4715. (Cited on pages 14 and 56.)

M. Denkowski and A. Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2107. (Cited on pages 26 and 120.)

M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W14-3348. (Cited on page 26.)

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423. (Cited on pages 18, 29, 40, and 50.)

T. Domhan and F. Hieber. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1158. (Cited on page 14.)

L. Duong, T. Cohn, S. Bird, and P. Cook. A neural network model for low-resource universal dependency

parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1040. (Cited on page 54.)

C. Dyer, V. Chahuneau, and N. A. Smith. 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, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-1073. (Cited on pages 59 and 107.)

S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1045. (Cited on pages 1, 15, 22, 23, and 79.)

J. Escudé Font and M. R. Costa-jussà. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-3821. (Cited on page 116.)

S. Evert and H. Kermes. Experiments on candidate data for collocation extraction. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, Apr. 2003. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E03-1080. (Cited on page 101.)

S. Evert and B. Krenn. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France, July 2001. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P01-1025. (Cited on page 101.)

M. Fadaee and C. Monz. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1040. (Cited on page 11.)

M. Fadaee and C. Monz. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.ngt-1.10. (Cited on page 11.)

M. Fadaee, A. Bisazza, and C. Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July 2017a. Association for Computational Linguistics. URL http://aclweb.org/anthology/P17-2090. (Cited on page 11.)

M. Fadaee, A. Bisazza, and C. Monz. Learning topic-sensitive word representations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 441–447, Vancouver, Canada, July 2017b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-2070. (Cited on page 11.)

M. Fadaee, A. Bisazza, and C. Monz. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association, 2018. URL http://aclweb.org/anthology/L18-1148. (Cited on page 11.)

M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016. URL http://arxiv.org/abs/1605.02276v1. (Cited on pages 30 and 40.)

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA, 2001. ACM. URL http://doi.acm.org/10.1145/371920.372094. (Cited on page 38.)

O. Firat, K. Cho, and Y. Bengio. 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, June 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N16-1101. (Cited on page 57.)

J. R. Firth. A synopsis of linguistic theory 1930-55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, volume 1952-59, pages 1–32, Oxford, 1957. The Philological Society. URL https://ci.nii.ac.jp/naid/10020680394/. (Cited on pages 16 and 29.)

J. Fodor and E. Lepore. *Holism: A Shopper's Guide*. Wiley, 1992. URL https://books.google.nl/books?id=fjoji3YAwoUC. (Cited on page 134.)

J. Fodor and E. LePore. *The Compositionality Papers*. Clarendon Press, 2002. URL https://books.google.nl/books?id=qcI6IVquMdgC. (Cited on page 124.)

G. Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892. URL https://pure.mpg.de/rest/items/item%5F2320751/component/file%5F2555768/content. (Cited on pages 123 and 134.)

M. Freitag and Y. Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver, Aug. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-3207. (Cited on page 21.)

P. Gage. A new algorithm for data compression. *The C Users Journal archive*, 12(2):23–38, Feb. 1994. URL https://www.derczynski.com/papers/archive/BPE%5FGage.pdf. (Cited on page 16.)

W. A. Gale, K. W. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439, 1992. URL https://link.springer.com/article/10.1007/BF00136984. (Cited on page 38.)

J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. 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, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1092. (Cited on page 69.)

M. García-Martínez, O. Caglayan, W. Aransa, A. Bardet, F. Bougares, and L. Barrault. Lium machine translation systems for WMT17 news translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 288–295, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4726. (Cited on page 75.)

S. Garg, T. Vu, and A. Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/abs/1911.04118. (Cited on pages 18 and 19.)

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/gehring17a.html. (Cited on page 25.)

H. Ghader and C. Monz. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1004. (Cited on page 22.)

H. Ghader and C. Monz. An intrinsic nearest neighbor analysis of neural machine translation architectures. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 107–117, Dublin, Ireland, Aug. 2019. European Association for Machine Translation. URL https://www.aclweb.org/anthology/W19-6611. (Cited on page 21.)

W. Gharbieh, V. C. Bhavsar, and P. Cook. A word embedding approach to identifying verb–noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions MWE 2016*, page 112, 2016. URL https://www.aclweb.org/anthology/W/W16/W16-1817. (Cited on page 29.)

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.

org/abs/1412.6572. (Cited on page 6.)

A. Graves, G. Wayne, and I. Danihelka. Neural Turing machines. *ArXiv*, abs/1410.5401, 2014. URL https://arxiv.org/abs/1410.5401. (Cited on page 22.)

J. Gu, H. Hassan, J. Devlin, and V. O. Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018a. URL https://arxiv.org/abs/1802.05368. (Cited on page 57.)

J. Gu, Y. Wang, Y. Chen, V. O. K. Li, and K. Cho. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium, Oct. 2018b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1398. (Cited on page 53.)

C. Gulcehre, O. Firat, K. Xu, K. Cho, and Y. Bengio. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148, 2017. URL https://www.sciencedirect.com/science/article/abs/pii/S0885230816301395. (Cited on page 15.)

T.-L. Ha, J. Niehues, and A. Waibel. Effective strategies in zero-shot neural machine translation. *ArXiv e-prints*, Nov. 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv171107893H/abstract. (Cited on page 75.)

A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24 (2):8–12, Mar. 2009. URL http://dx.doi.org/10.1109/MIS.2009.36. (Cited on pages 54 and 55.)

B. Hamp and H. Feldweg. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997. URL https://www.aclweb.org/anthology/W97-0802. (Cited on page 70.)

S. Hassan and R. Mihalcea. Semantic relatedness using salient semantic analysis. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2011. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3616/3972. (Cited on page 39.)

D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc., 2016a. URL http://papers.nips.cc/paper/6469-dual-learning-for-machine-translation. (Cited on page 15.)

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016b. URL https://arxiv.org/abs/1512.03385. (Cited on page 25.)

V. Henrich and E. Hinrichs. GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/264%5FPaper.pdf. (Cited on page 70.)

F. Hill, R. Reichart, and A. Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, Dec. 2015. URL https://www.aclweb.org/anthology/J15-4004. (Cited on pages 38 and 40.)

S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 6(2):107–116, Apr. 1998. URL https://doi.org/10.1142/S0218488598000094. (Cited on page 19.)

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. URL https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735. (Cited on page 19.)

E. Huang, R. Socher, C. Manning, and A. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P12-1092. (Cited on pages 32, 40, and 42.)

Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu,

and z. Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems 32*, pages 103–112. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8305-gpipe-efficient-training-of-giant-neural-networks-using-pipeline-parallelism. (Cited on page 55.)

D. Hupkes, V. Dankers, M. Mul, and E. Bruni. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5065–5069. International Joint Conferences on Artificial Intelligence Organization, 7 2020. URL https://doi.org/10.24963/ijcai.2020/708. (Cited on pages 124 and 135.)

M. Hurtado Bodell, M. Arvidsson, and M. Magnusson. Interpretable word embeddings via informative priors. 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 6323–6329, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1661. (Cited on page 51.)

N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998. URL https://www.aclweb.org/anthology/J98-1001. (Cited on page 31.)

H. Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018. URL https://arxiv.org/abs/1801.02929. (Cited on page 56.)

P. Isabelle, C. Cherry, and G. Foster. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1263. (Cited on pages 99, 101, 102, 110, and 131.)

C. Jacquemin. A temporal connectionist approach to natural language. *SIGART Bull.*, 5(3):12–22, July 1994. URL https://doi.org/10.1145/181911.181913. (Cited on page 19.)

T. M. V. Janssen. Frege, contextuality and compositionality. *Journal of Logic, Language and Information*, 10 (1):115–136, 2001. URL https://doi.org/10.1023/A:1026542332224. (Cited on page 124.)

S. Jean, K. Cho, R. Memisevic, and Y. Bengio. 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, July 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P15-1001. (Cited on page 16.)

F. Jelinek. *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press, Jan. 1998. URL http://www.amazon.fr/exec/obidos/ASIN/0262100665/citeulike04-21. (Cited on page 21.)

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. URL https://www.aclweb.org/anthology/Q17-1024. (Cited on page 57.)

M. I. Jordan. The era of big data. *The official bulletin of the International Society for Bayesian Analysis*, 18 (2), 2011. URL http://jfsowa.com/ikl/Jordan11.pdf. (Cited on page 32.)

M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 2020. URL https://arxiv.org/abs/1907.10529. (Cited on page 18.)

L. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/pdf?id=SJTQLdqlg. (Cited on page 1.)

T. Kajiwara and M. Komachi. 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*, pages 1147–1158, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1109. (Cited on page 14.)

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. URL https://arxiv.org/abs/1812.04948. (Cited on page 55.)

G. Katz and E. Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-1203. (Cited on page 101.)

H. Khayrallah and P. Koehn. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-2709. (Cited on page 81.)

A. Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1997. URL https://doi.org/10.1023/A:1000583911091. (Cited on page 32.)

D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980. (Cited on page 118.)

R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5950-skip-thought-vectors. (Cited on page 101.)

K. Kishida. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005. URL https://www.nii.ac.jp/TechReports/public%5Fhtml/05-014E.pdf. (Cited on page 43.)

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics, 2017. URL http://www.aclweb.org/anthology/P17-4012. (Cited on pages 78, 102, and 118.)

N. Klyueva, A. Doucet, and M. Straka. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-1707. (Cited on page 101.)

A. Koç, I. Utlu, L. K. Senel, and H. M. Özaktas. Imparting interpretability to word embeddings. *CoRR*, abs/1807.07279, 2018. URL http://arxiv.org/abs/1807.07279. (Cited on page 51.)

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86. Citeseer, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.5497&rep=rep1&type=pdf. (Cited on page 13.)

P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Aug. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-3204. (Cited on pages 1, 4, 22, 53, and 56.)

P. Koehn, F. J. Och, and D. Marcu. 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, 2003. URL https://www.aclweb.org/anthology/N03-1017. (Cited on pages 13, 14, and 21.)

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 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, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-2045. (Cited on pages 14, 59, and 107.)

Z. Kövecses and P. Szabó. Idioms: A view from cognitive semantics. *Applied Linguistics*, 17(3):326–355, 09 1996. URL https://doi.org/10.1093/applin/17.3.326. (Cited on page 100.)

G. Kremer, K. Erk, S. Padó, and S. Thater. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E14-1057. (Cited on page 43.)

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks. (Cited on pages 53 and 55.)

T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3927–3936. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8648-improved-precision-and-recall-metric-for-assessing-generative-models. (Cited on page 55.)

B. M. Lake and M. Baroni. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *CoRR*, abs/1711.00350, 2017. URL http://arxiv.org/abs/1711.00350. (Cited on page 135.)

B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. Dy and A. Krause, editors, *35th International Conference on Machine Learning, ICML 2018*, 35th International Conference on Machine Learning, ICML 2018, pages 4487–4499. International Machine Learning Society (IMLS), Jan. 2018. URL https://arxiv.org/abs/1711.00350. 35th International Conference on Machine Learning, ICML 2018 ; Conference date: 10-07-2018 Through 15-07-2018. (Cited on pages 124, 125, and 133.)

S. M. Lakew, M. Cettolo, and M. Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1054. (Cited on page 23.)

P. Lambert, H. Schwenk, C. Servan, and S. Abdul-Rauf. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT11, pages 284–293, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2132960.2132997. (Cited on pages 75, 80, and 81.)

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 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*, 2018a. URL https://openreview.net/forum?id=rkYTTf-AZ. (Cited on page 57.)

G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, Oct. 2018b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1549. (Cited on page 57.)

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS. (Cited on page 18.)

J. H. Lau, P. Cook, D. McCarthy, S. Gella, and T. Baldwin. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1025. (Cited on page 32.)

J. Lee, K. Cho, and T. Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. URL https://www.

aclweb.org/anthology/Q17-1026. (Cited on page 57.)

K. Lee, O. Firat, A. Agarwal, C. Fannjiang, and D. Sussillo. Hallucinations in neural machine translation. In *Neural Information Processing Systems (NeurIPS) Workshop on Interpretability and Robustness for Audio, Speech, and Language*. NeurIPS, 2018. URL https://openreview.net/pdf?id=SkxJ-309FQ. (Cited on page 114.)

G. Lembersky, N. Ordan, and S. Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1034. (Cited on page 14.)

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966. URL https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf. (Cited on page 118.)

O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2050. (Cited on page 18.)

O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. URL https://transacl.org/ojs/index.php/tacl/article/view/570. (Cited on page 38.)

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 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, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.703. (Cited on page 18.)

G. Li, L. Liu, C. Zhu, T. Zhao, and S. Shi. Detecting and understanding generalization barriers for neural machine translation. *CoRR*, abs/2004.02181, 2020. URL https://arxiv.org/abs/2004.02181. (Cited on page 2.)

J. Li and D. Jurafsky. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1200. (Cited on pages 32 and 45.)

L. Li, B. Roth, and C. Sporleder. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-1116. (Cited on page 3.)

X. Li, P. Michel, A. Anastasopoulos, Y. Belinkov, N. Durrani, O. Firat, P. Koehn, G. Neubig, J. Pino, and H. Sajjad. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-5303. (Cited on page 114.)

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013. (Cited on page 134.)

T. Linzen, E. Dupoux, and Y. Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. URL https://www.aclweb.org/anthology/Q16-1037. (Cited on page 123.)

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46448-0%5F2. (Cited on page 55.)

T. Luong, R. Socher, and C. Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https:

//www.aclweb.org/anthology/W13-3512. (Cited on page 38.)

T. Luong, H. Pham, and C. D. Manning. 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, Sept. 2015a. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1166. (Cited on pages 19, 21, 22, 102, and 117.)

T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in 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 11–19, Beijing, China, July 2015b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P15-1002. (Cited on pages 4 and 16.)

S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze, editors. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-1700. (Cited on page 102.)

Y. Marton, C. Callison-Burch, and P. Resnik. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore, Aug. 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D09/D09-1040. (Cited on page 54.)

D. McCarthy and R. Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S/S07/S07-1009. (Cited on pages 31, 43, and 50.)

I. D. Melamed. *Manual annotation of translational equivalence: The Blinker project*. University of Pennsylvania, 1998. URL https://arxiv.org/abs/cmp-lg/9805005. (Cited on page 53.)

O. Melamud, O. Levy, and I. Dagan. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-1501. (Cited on pages 43 and 44.)

P. Michel and G. Neubig. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1050. (Cited on pages 7 and 114.)

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a. URL https://arxiv.org/abs/1301.3781. (Cited on pages 3, 17, 29, 32, 33, 36, 38, 42, and 92.)

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013b. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality. (Cited on pages 37 and 70.)

T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090. (Cited on page 18.)

G. A. Miller. Dictionaries of the mind. In *23rd Annual Meeting of the Association for Computational Linguistics*, pages 305–314, Chicago, Illinois, USA, July 1985. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P85-1038. (Cited on page 1.)

G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995. URL https://dl.acm.org/doi/abs/10.1145/219717.219748. (Cited on pages 31, 37, and 69.)

R. Montague. Universal grammar. In R. H. Thomason, editor, *Formal Philosophy: Selected Papers of Richard Montague*, number 222–247 in Theoria, New Haven, London, 1974. Yale University Press. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-

2567.1970.tb00434.x. (Cited on page 123.)

R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-2041. (Cited on page 77.)

D. Moussallem, M. A. Sherif, D. Esteves, M. Zampieri, and A.-C. Ngonga Ngomo. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1392. (Cited on page 102.)

G. Muzny and L. Zettlemoyer. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1145. (Cited on page 102.)

R. Navigli. *SOFSEM 2012: Theory and Practice of Computer Science: 38th Conference on Current Trends in Theory and Practice of Computer Science, Špindlerův Mlýn, Czech Republic, January 21-27, 2012. Proceedings*, chapter A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches, pages 115–129. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. URL https://link.springer.com/book/10.1007%2F978-3-642-27660-6. (Cited on page 32.)

R. Navigli and S. P. Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1023. (Cited on page 31.)

A. Neelakantan, J. Shankar, A. Passos, and A. McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1113. (Cited on pages 32, 33, 40, 41, 42, 45, 47, 49, and 50.)

T.-V. Ngo, T.-L. Ha, P.-T. Nguyen, and L.-M. Nguyen. Overcoming the rare word problem for low-resource language pairs in neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 207–214, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-5228. (Cited on pages 4, 53, and 69.)

G. Nunberg, I. A. Sag, and T. Wasow. Idioms. *Language*, 70(3):491–538, 1994. URL http://www.jstor.org/stable/416483. (Cited on page 100.)

R. Östling and J. Tiedemann. Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*, 2017. URL https://arxiv.org/abs/1708.05729. (Cited on pages 53 and 57.)

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2015. URL https://ieeexplore.ieee.org/abstract/document/6802355. (Cited on page 32.)

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P02-1040. (Cited on pages 26, 62, 78, 107, and 134.)

F. J. Pelletier. The principle of semantic compositionality. *Topoi*, 13:11–24, 1994. URL https://link.springer.com/article/10.1007%2FBF00763644. (Cited on page 123.)

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1162. (Cited on pages 17, 29, 32, 36, and 37.)

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1202. (Cited on pages 18, 29, 40, and 50.)

N.-Q. Pham, J. Niehues, T.-L. Ha, E. Cho, M. Sperber, and A. Waibel. The Karlsruhe Institute of Technology systems for the news translation task in WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, pages 366–373, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-4736. (Cited on page 15.)

A. Poncelas, D. Shterionov, A. Way, G. Maillette de Buy Wenniger, and P. Passban. Investigating backtranslation in neural machine translation. *ArXiv e-prints*, Apr. 2018. URL http://adsabs.harvard.edu/abs/2018arXiv180406189P. (Cited on page 79.)

A. Poncelas, G. Maillette de Buy Wenniger, and A. Way. Adaptation of machine translation models with back-translated data using transductive data selection methods. *arXiv e-prints*, art. arXiv:1906.07808, June 2019a. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190607808P. (Cited on page 77.)

A. Poncelas, M. Popovic, D. Shterionov, G. M. de Buy Wenniger, and A. Way. Combining SMT and NMT back-translated data for efficient NMT. *CoRR*, abs/1909.03750, 2019b. URL http://arxiv.org/abs/1909.03750. (Cited on page 77.)

Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig. When and why are pre-trained word embeddings useful 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 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-2084. (Cited on page 18.)

L. Qiu, K. Tu, and Y. Yu. Context-dependent sense embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 183–191, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1018. (Cited on page 32.)

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *Technical report*, 2019. URL https://www.ceid.upatras.gr/webpages/faculty/zaro/teaching/alg-ds/PRESENTATIONS/PAPERS/2019-Radford-et-al%5FLanguage-Models-Are-Unsupervised-Multitask-%20Learners.pdf. (Cited on page 50.)

K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM. URL http://doi.acm.org/10.1145/1963405.1963455. (Cited on pages 38 and 39.)

M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732. (Cited on page 22.)

R. Rapp. The back-translation score: Automatic MT evaluation at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 133–136, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1667583.1667625. (Cited on page 14.)

J. Reisinger and R. J. Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N10-1013. (Cited on page 29.)

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.442. (Cited on page 134.)

A. Rios, M. Müller, and R. Sennrich. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels, Oct.

2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6437. (Cited on page 1.)

H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, Oct. 1965. URL http://doi.acm.org/10.1145/365628.365657. (Cited on page 38.)

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Representations by Back-Propagating Errors*, pages 696–699. MIT Press, Cambridge, MA, USA, 1988. URL https://www.nature.com/articles/323533a0. (Cited on page 19.)

B. Salehi and P. Cook. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (∗SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S13-1039. (Cited on page 101.)

B. Salehi, P. Cook, and T. Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N15-1099. (Cited on pages 29 and 101.)

G. Salton, R. Ross, and J. Kelleher. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden, Apr. 2014a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W14-0806. (Cited on page 101.)

G. Salton, R. Ross, and J. Kelleher. An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden, Apr. 2014b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1007. (Cited on page 102.)

G. Salton, R. Ross, and J. Kelleher. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-1019. (Cited on page 101.)

A. Schenk. Idioms in the Rosetta machine translation system. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, pages 319–324, Stroudsburg, PA, USA, 1986. Association for Computational Linguistics. URL https://doi.org/10.3115/991365.991458. (Cited on page 101.)

J. Schmidhuber. *Habilitation thesis*. Institut für Informatik, Technische Universität München, 1993. URL ftp://ftp.idsia.ch/pub/juergen/habilitation.pdf. (Cited on page 19.)

H. Schwenk. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, 2008. URL https://www.isca-speech.org/archive/iwslt%5F08/papers/slt8%5F182.pdf. (Cited on page 14.)

R. Sennrich and B. Zhang. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1021. (Cited on page 56.)

R. Sennrich, B. Haddow, and A. Birch. 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, June 2016a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N16-1005. (Cited on pages 6 and 105.)

R. Sennrich, B. Haddow, and A. Birch. 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, Aug. 2016b. Association for Computational Linguistics.

URL http://www.aclweb.org/anthology/P16-1009. (Cited on pages 1, 5, 15, 53, 56, 62, 63, 75, 78, 79, 81, 82, and 86.)

R. Sennrich, B. Haddow, and A. Birch. 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, Aug. 2016c. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1162. (Cited on pages 16, 54, 57, 62, 69, 78, 90, 107, and 117.)

R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. Miceli Barone, and P. Williams. The university of edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4739. (Cited on page 75.)

Y. Shao, R. Sennrich, B. Webber, and F. Fancellu. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1005. (Cited on pages 102 and 132.)

C. C. Silva, C.-H. Liu, A. Poncelas, and A. Way. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6323. (Cited on page 77.)

B. Singh, M. Najibi, and L. S. Davis. Sniper: Efficient multi-scale training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9310–9320. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8143-sniper-efficient-multi-scale-training. (Cited on page 55.)

S. Small, G. Cottrell, and M. Tanenhaus. *Lexical Ambiguity Resolution: Perspective from Psycholinguistics, Neuropsychology and Artificial Intelligence*. Elsevier Science, 2013. URL https://books.google.nl/books?id=-J-fAgAAQBAJ. (Cited on page 1.)

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006. URL http://mt-archive.info/AMTA-2006-Snover.pdf. (Cited on pages 26, 120, and 134.)

P. Sprent and N. C. Smeeton. *Applied nonparametric statistical methods*. CRC Press, 2016. URL https://link.springer.com/book/10.1007/978-94-009-1223-6. (Cited on page 45.)

G. Stanovsky, N. A. Smith, and L. Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1164. (Cited on page 116.)

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*, 2006. URL https://arxiv.org/abs/cs/0609058. (Cited on page 13.)

H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, and T. Zhao. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.324. (Cited on page 58.)

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks. (Cited on pages 16, 19, 21, and 56.)

G. Szarvas, R. Busa-Fekete, and E. Hüllermeier. Learning to rank lexical substitutions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1932, Seattle,

Washington, USA, Oct. 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1198. (Cited on page 43.)

D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1146. (Cited on page 29.)

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1385–1392. MIT Press, 2005. URL http://papers.nips.cc/paper/2698-sharing-clusters-among-related-groups-hierarchical-dirichlet-processes. (Cited on page 33.)

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. URL https://www.seas.harvard.edu/courses/cs281/papers/teh-jordan-beal-blei-2005.pdf. (Cited on pages 32 and 38.)

J. Tiedemann, F. Cap, J. Kanerva, F. Ginter, S. Stymne, R. Östling, and M. Weller-Di Marco. Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 391–398, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W16-2326. (Cited on page 15.)

F. Torabi Asr, R. Zinkov, and M. Jones. Querying word embeddings for similarity and relatedness. 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 675–684, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1062. (Cited on page 40.)

K. Tran, A. Bisazza, and C. Monz. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1503. (Cited on page 25.)

N. Ueffing, G. Haffari, and A. Sarkar. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2):77–94, June 2007. URL https://doi.org/10.1007/s10590-008-9036-3. (Cited on page 15.)

L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605, 2008. URL http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf. Pagination: 27. (Cited on page 49.)

M. van der Wees, A. Bisazza, and C. Monz. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1147. (Cited on pages 75 and 77.)

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need. (Cited on pages 22, 23, 24, 25, 117, and 118.)

L. Wang, D. F. Wong, L. S. Chao, J. Xing, Y. Lu, and I. Trancoso. Edit distance: A new data selection criterion for domain adaptation in SMT. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 727–732, Hissar, Bulgaria, Sept. 2013. INCOMA Ltd. Shoumen, BULGARIA. URL https://www.aclweb.org/anthology/R13-1094. (Cited on page 77.)

W. Wang, I. Caswell, and C. Chelba. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292, Florence, Italy, July 2019a. Association for

Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1123. (Cited on page 77.)

X. Wang, H. Pham, Z. Dai, and G. Neubig. SwitchOut: An efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1100. (Cited on page 53.)

Y. Wang, L. Cui, and Y. Zhang. Using dynamic embeddings to improve static embeddings. *ArXiv*, abs/1911.02929, 2019b. URL https://ui.adsabs.harvard.edu/abs/2019arXiv191102929W/abstract. (Cited on page 16.)

Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7029–7039. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7278-learning-to-model-the-tail. (Cited on page 1.)

S. Wiseman and A. M. Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D16-1137. (Cited on page 22.)

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144. (Cited on pages 18, 21, 22, and 101.)

S. Wubben, A. van den Bosch, and E. Krahmer. 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, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P12-1107. (Cited on page 14.)

W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423, 2017. (Cited on page 19.)

Z. Yang, W. Chen, F. Wang, and B. Xu. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P18-1005. (Cited on page 57.)

X. Yao and B. Van Durme. Nonparametric Bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14, Portland, Oregon, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-1102. (Cited on pages 30 and 32.)

W.-t. Yih and V. Qazvinian. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 616–620, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N12-1077. (Cited on page 39.)

L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1056. (Cited on page 29.)

Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation, 2019. URL https://arxiv.org/abs/1909.11065. (Cited on page 19.)

J. Zhang and C. Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://www.aclweb.org/

anthology/D16-1160. (Cited on page 80.)

X. Zhang and M. Lapata. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1062. (Cited on page 14.)

Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou. Semantics-aware BERT for language understanding, 2019. URL https://arxiv.org/abs/1909.02209. (Cited on page 18.)

Z. Zhang, J. jie Yang, and H. Zhao. Retrospective reader for machine reading comprehension. *ArXiv*, abs/2001.09694, 2020. URL https://arxiv.org/abs/2001.09694. (Cited on page 19.)

B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1163. (Cited on pages 53 and 57.)

W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1141. (Cited on page 29.)

Summary

Neural networks learn patterns from data to solve complex problems. To understand and infer meaning in language, neural models have to learn complicated nuances. Discovering distinctive linguistic phenomena from data is not an easy task. For instance, lexical ambiguity is a fundamental feature of language which is challenging to learn. Even more prominently, inferring the meaning of rare and unseen lexical units is difficult with neural networks. Meaning is often determined from *context*. With context, languages allow meaning to be conveyed even when the specific words used are not known by the reader. To model this learning process, a system has to learn from a few instances in context and be able to generalize well to unseen cases. Neural models use a sizable amount of data that often consists of contextual instances to learn patterns. The learning process is hindered when training data is scarce for a task. Even with sufficient data, learning patterns for the long tail of the lexical distribution is challenging.

In this thesis, we focus on understanding certain potentials of contexts in neural models and design augmentation models to benefit from them. We focus on machine translation as an important instance of the more general language understanding problem. To translate from a source language to a target language, a neural model has to understand the meaning of constituents in the provided context and generate constituents with the same meanings in the target language. This task accentuates the value of capturing nuances of language and the necessity of generalization from few observations. The main problem we study in this thesis is what neural machine translation models learn from data and how we can devise more focused contexts to enhance this learning. First, we study how document-level contexts aid in distinguishing different meanings of a word. Second, we investigate how translation models exploit context to learn and transfer meaning and show that different and diverse contexts resolve various obstacles of translation. Third, we examine under which conditions the observed context in the data is not enough for inferring meaning and capturing various linguistic phenomena.

Looking more in-depth into the role of context and the impact of data on learning models is essential to advance the Natural Language Processing (NLP) field. Understanding the importance of data in the learning process and how neural network models interact with and benefit from data can help develop more accurate NLP systems. Moreover, it helps highlight the vulnerabilities of current neural networks and provides insights into designing more robust models.

Samenvatting

Neurale netwerken zijn computermodellen die patronen leren uit data, om zo complexe problemen op te lossen. Om betekenis in taal te begrijpen en af te leiden, moeten neurale netwerken ingewikkelde nuances leren. Voorbeelden van taalkundige fenomenen die niet gemakkelijk zijn voor een neuraal netwerk, zijn *lexicale ambiguïteit* (woorden met meerdere betekenissen) en zeldzame of nieuwe woorden die het neurale model niet of slechts enkele keren heeft gezien in de trainingsdata. De betekenis van zeldzame of ambigue woorden hangt af van de context waarin ze voorkomen. Door middel van context kunnen talen betekenis overbrengen, zelfs als de lezer bepaalde gebruikte woorden niet kent. Om dit leerproces te modelleren, moet een systeem leren van slechts enkele voorbeelden in een bepaalde context en moet het goed kunnen generaliseren naar ongeziene gevallen. Neurale modellen gebruiken een aanzienlijke hoeveelheid data, vaak bestaande uit voorbeelden in context die gebruikt worden om patronen te leren. Als er maar weinig trainingsdata voor een bepaalde taak is, wordt het leerproces gehinderd, maar ook als er genoeg data beschikbaar is, blijft het leren van zeldzame woorden een uitdaging.

De focus van dit proefschrift ligt op het begrijpen van de mogelijkheden die context biedt voor neurale modellen om hier vervolgens van te kunnen profiteren. We richten ons op machinaal vertalen als een belangrijk voorbeeld van het meer algemene probleem van taalbegrip. Om te vertalen van een brontaal naar een doeltaal, moet een neuraal model de betekenis van woorden of zinnen in de gegeven context begrijpen en woorden of zinnen met dezelfde betekenis genereren in de doeltaal. Deze taak benadrukt het belang van het vastleggen van nuances in taal en de noodzaak om te kunnen generaliseren vanuit slechts een beperkt aantal observaties. Het belangrijkste probleem dat we in dit proefschrift bestuderen is wat neurale machinale vertaalmodellen leren van data en hoe we meer gefocuste contexten kunnen bedenken om dit leerproces te verbeteren. Allereerst bestuderen we hoe context op documentniveau kan helpen bij het onderscheiden van verschillende betekenissen van een woord. Daarnaast onderzoeken we hoe vertaalmodellen context gebruiken om betekenis te leren en over te dragen en laten we zien dat het gebruik van verschillende contexten een aantal obstakels tijdens het vertalen kan overwinnen. Ten slotte onderzoeken we onder welke voorwaarden de waargenomen context in de data onvoldoende is om betekenis af te kunnen leiden en bepaalde taalkundige fenomenen te vatten.

Door dieper in te gaan op op de rol die context speelt en de impact van data op het leren van modellen, draagt het werk in dit proefschrift bij aan de vooruitgang van Natural Language Processing (NLP). Het is belangrijk dat we goed begrijpen hoe neurale modellen interageren met en profiteren van data, om zo accuratere NLP-systemen te ontwikkelen. Bovendien helpt het onderzoek in dit proefschrift om de kwetsbaarheden van huidige neurale netwerken te benadrukken en geeft het inzicht in hoe robuustere modellen ontworpen kunnen worden.