Master Thesis

Synthetic Data for Domain Adaptation in Neural Machine Translation

Lahorka Nikolovski

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Sophie Arnoult
2nd reader: Hennie van der Vliet

Submitted: July 1, 2022

Abstract

This thesis project, executed during an internship at the language data company TAUS, focuses on evaluating the usefulness of synthetic data for domain adaptation in Neural Machine Translation. Synthetic data is generated using translation-based methods, namely forward and back-translation, from the source and target sides of natural parallel corpora that are approximating monolingual source and target-side data. Different experiments are run based on English→Dutch parallel corpora in financial, pharmaceutical and e-commerce domains, using two models and two adaptation methods: proprietary Amazon Translate and Active Custom Translation, and the open-source OPUS-MT model (Tiedemann and Thottingal, 2020) and fine-tuning. The quality of synthetic data is evaluated extrinsically, by evaluating the performance of NMT systems adapted using synthetic parallel corpora. In addition to the more traditional, string-based automatic machine translation metrics BLEU and chrF, all experiments are additionally evaluated using the pretrained, neural COMET metric (Rei et al., 2020), which has shown higher correlation with human judgment in recent research (Kocmi et al., 2021). Experiments indicate that including synthetic data obtained by back-translation into TAUS Data-Enhanced Machine Translation pipeline should result in translation models that are better adapted to the domains of interest. Synthetic data generation shows the most promise if used as a data augmentation technique in lower-resource scenarios.

Keywords: Domain Adaptation for Neural Machine Translation, Synthetic Data Generation, Back-translation, COMET score, OPUS-MT

Declaration of Authorship

I, Lahorka Nikolovski, declare that this thesis, titled *Synthetic Data for Domain Adaptation in Neural Machine Translation* and the work presented in it are my own. I confirm that:

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

Date: July 1, 2022

Signed: Lahorka Nikolovski

Acknowledgments

I would like to thank my supervisor, Sophie Arnoult, for her guidance and for providing constructive feedback at every step of the journey that writing this thesis was. Thank you also to Hennie van der Vliet, for being the second reader.

I would also like to extend my gratitude to TAUS, where I have completed my internship, and where I will embark on my next journey, that of an NLP practitioner. Thank you, Dace and Anderson, for believing in me and offering me the internship position. To Anderson, thank you also for being my company supervisor, and for always being there for me, whatever I needed. Thank you also to my colleagues from the "NLP team": Amir, András, David, Lisa and Marcell. I benefited a lot from your ideas and the feedback you provided while discussing my project. I also benefited from being able to use code written and provided by Amir, András and David, especially when it came to cleaning the data and using LASER embeddings to analyse and filter the corpora I was working with.

I am also indebted to all the wonderful professors I had the privilege of learning from during the course of the master's programme. Additionally, this last year spend studying would have been much less fun were it not for my talented colleagues. I am especially grateful to Yota for her friendship and comradery.

Lastly, I would like to thank my family and friends, who were there not just during the fun times, but also during the hard times. To my partner Goran, without whose support I would not be writing these words today, thank you for always being there for me and supporting me unconditionally. And of course, special thanks has to be extended to our wonderful dogs Nera and Deda. Without those cuddles, this journey would have been a much lonelier one.

List of Tables

4.1 Number of sentence pairs per domain before cleaning. 22
4.2 Statistics for cleaned datasets. 23
4.3 Examples of relevant sentences from the test sets for each of the domains. 23
4.4 Evaluation of unadapted en→nl models. 24
4.5 Evaluation of nl→en models that will be used for back-translation. . . . 24
4.6 Evaluation of adaptation using natural parallel data. 25
4.7 Performance of ACT using fully synthetic parallel corpora obtained by back-translation. 27
4.8 Performance of OPUS-FT models fine-tuned with fully synthetic parallel corpora obtained by back-translation. 27
4.9 Performance of ACT using synthetic parallel corpora obtained by FT and combining FT and BT. 28
4.10 Performance of OPUS-FT models fine-tuned with synthetic parallel corpora obtained by FT and combining FT and BT. 29
4.11 Performance of ACT using fully synthetic parallel corpora obtained by back-translation using the OPUS-MT model. 30
4.12 Performance of OPUS-FT models fine-tuned with fully synthetic parallel corpora obtained by back-translation using the OPUS-MT model. . . . 31
4.13 Performance of ACT adapted with 1 part synthetic to 1 part natural data. 32
4.14 Performance of OPUS-FT models fine-tuned with 1 part synthetic to 1 part natural data. 33
4.15 Performance of ACT in the pharmaceutical domain, adapted with different ratios of synthetic to natural data. 33
4.16 Performance of ACT in the financial domain, adapted with different ratios of synthetic to natural data. 34
4.17 Performance of OPUS models, pharmaceutical domain, adapted with different ratios of synthetic to natural data. 34
4.18 Performance of OPUS models, financial domain, adapted with different ratios of synthetic to natural data. 35
4.19 Performance of ACT adapted with 1 part synthetic to 1 part natural data, where some data was selected at random, and some using LASER embeddings. 36
4.20 Performance of OPUS-FT models fine-tuned with 1 part synthetic to 1 part natural data, where some data was selected at random, and some using LASER embeddings. 36

Contents

Abstract i

Declaration of Authorship iii

Acknowledgments v

List of Tables vii

1 Introduction 1

1.1 Motivation 1

1.2 Goal and Research Questions 2

1.3 Outline 3

2 Background and Related Work 5

2.1 Successes and Challenges of Neural MT 5

2.2 Domain Adaptation in Machine Translation 6

2.3 Data Augmentation and Synthetic Data for NLP 8

2.4 Synthetic Data for Neural Machine Translation 9

2.4.1 Back-translation 9

2.4.2 Forward translation 10

2.4.3 Combining back-translation and forward translation 10

2.4.4 Refining back-translation 11

2.5 Summary 12

3 Methodology 13

3.1 Models 13

3.1.1 Amazon Translate and Active Custom Translation 14

3.1.2 OPUS-MT 15

3.2 Evaluation 16

3.2.1 BLEU score 16

3.2.2 chrF score 17

3.2.3 COMET score 18

3.2.4 Statistical significance 18

3.3 Cleaning and Matching Data with LASER Embeddings 19

3.4 Summary 20

4 Experiments **21**
4.1 Datasets 21
4.2 Baseline: Performance of Unadapted Models 23
4.3 Upper Bound: Adaptation with Natural Data 25
4.4 Experimenting with Synthetic Data 26
4.4.1 Simple back-translation 26
4.4.2 Forward translation 28
4.4.3 Choosing the best translations 29
4.4.4 Experimenting with the quantity of synthetic data 31
4.4.5 Selecting data more representative of the domain 35
4.4.6 Summary 37
4.5 Qualitative Analysis 37

5 Discussion and Conclusion **41**
5.1 Answering the Research Questions 41
5.2 Discussion and Limitations 42
5.3 Conclusion and Future Work 44

A Examples selected for qualitative analysis **47**
A.1 47
A.2 49
A.3 51
A.4 53
A.5 55

Chapter 1

Introduction

Since their introduction in 2014, deep neural network-based machine translation models have taken over both the academia and the industry. Occasionally, a company or a research team even claims their system has obtained human parity (for an early example, see Hassan et al. (2018), scrutinized by Toral et al. (2018)).

Even though these Neural Machine Translation (NMT) models trained on huge amounts of data do perform exceedingly well when translating texts that are sufficiently similar to the data they have been trained on, their performance can deteriorate quickly when translating texts from domains that do not match their training data well (Koehn and Knowles, 2017). In order to address this challenge, different domain adaptation methods have been developed over the years (Chu and Wang, 2018; Saunders, 2021). Still, a bottleneck for many approaches is the fact that, in many of the world's language pairs and domains, enough high-quality parallel data that could serve as training data when adapting models to different domains of interest simply does not exist.

As one of the approaches to this low-resource problem, different ways of incorporating synthetic data have been proposed (Chu and Wang, 2018; Saunders, 2021). Synthetic data generation methods allow for obtaining additional training examples, without explicitly collecting and/or labelling new instances. In the field of Machine Translation (MT), using synthetic data, and more specifically, synthetic data produced by machine translation models, has a long tradition, and, as we shall see, has become a *de facto* standard for training models in certain contexts. As synthetic data generation approaches gained popularity in recent years, powered in part by the successes of Transformer-based language generation models, there is also renewed interest in researching how synthetic data can be used in various other Natural Language Processing (NLP) tasks, thus further fuelling the interest of researchers and companies in using synthetic data in their NLP pipelines.

1.1 Motivation

The motivation for this project comes from the language data company TAUS[1]. Recently, TAUS has introduced a new service called DEMT[2], which stands for Data-Enhanced Machine Translation (Aslan, 2022). The idea behind DEMT is to provide customers with the best possible machine translations in predefined language pairs and domains of interest. For example, if a client is interested in translating English

[1] `https://www.taus.net/`

[2] `https://datamarketplace.taus.net/enhance-mt`

texts from the financial domain into Dutch, DEMT translations will be generated using Amazon's Active Custom Translation (ACT) service[3], with customization performed using a financial-domain English→Dutch corpus curated by TAUS. At the moment, the service only uses ACT, but in the future, there are plans to explore adapting MT models provided by Google[4] and Microsoft[5], and possibly other vendors, with the goal of providing translations generated using the best-performing domain-adapted model in a certain language pair and domain, and thus really providing the customers with the best possible translations.

As a language data company, TAUS has at their disposal numerous high-quality parallel datasets in different domains. Still, there are always new language pairs and new domains, where customers will be interested in obtaining high-quality, domain-adapted translations. Collecting and curating new parallel corpora is always time-intensive and costly, and sometimes even borderline impossible[6], and thus TAUS was interested in exploring if synthetic data generation methods could be used to obtain parallel corpora in different language pairs and domains of interest.

Two translation scenarios that are especially relevant for TAUS are the "defined-domain scenario" and the "custom-translate scenario". In the first case, we are talking about the scenario currently covered by the DEMT service. Here, customers are interested in translations adapted to a specific, predefined domain (for example, the anti-money laundering domain). The "custom-translate scenario", on the other hand, would be at play when clients approach TAUS with textual examples of "domains" of interest. They might provide a number of websites, and ask for a translation engine adapted to those example sources. Although the DEMT service does not cover this scenario yet, it might be expanded in the future. From the data standpoint, again, two scenarios were singled out as relevant. In the first, there might be no parallel data in the language pair and domain of interest, thus TAUS would be interested in generating and using a fully synthetic parallel corpus for domain adaptation. In the other, some parallel data would be available, but TAUS would be interested in augmenting it with additional synthetic data in order to be able to obtain even better translations after adapting MT models. These two translation scenarios, and the data augmentation and the fully synthetic parallel corpus generation scenario, will inform the synthetic data generation methods explored and the experimental setup devised in the scope of this thesis. But first, we will narrow down the goal and the research questions that will guide us along the way, which is the topic of the next section.

1.2 Goal and Research Questions

Motivated by considerations described in the last section, the goal of this thesis was narrowed down to evaluating if synthetic data can be useful for customizing the neural machine translation models such as those presently offered through TAUS DEMT service (Amazon Translate[7]), and those considered for future use (the likes of Google

[3] https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-parallel-data.html

[4] Using Google's AutoML Translation, https://cloud.google.com/translate/automl/docs/.

[5] Using Microsoft's Custom Translator, https://docs.microsoft.com/en-us/azure/cognitive-services/translator/custom-translator/overview.

[6] In some low-resource scenarios, it can be very hard to find human translators who could provide the needed translation services, or to obtain original texts from specific domains.

[7] https://aws.amazon.com/translate/

Translate[8] and Microsoft Translator[9]). The research will be guided by the following research questions:

Q1. Can synthetic data be useful for domain adaptation in the context of TAUS DEMT?

Q2. Which method of generating synthetic data is the most useful?

To answer these questions, in this thesis we will look more closely at using synthetic data in the context of domain adaptation in neural machine translation, and narrow down the methods that seem the most promising for the TAUS context. Next, we will use them to generate synthetic parallel corpora, and evaluate their usefulness for domain adaptation.

1.3 Outline

In the following chapter, I provide an overview of relevant work in the fields of Neural Machine Translation, domain adaptation for NMT, and synthetic data generation in the context of NLP and NMT. Next, in Chapter 3, the methods that will be used in the scope of this thesis will be discussed, including neural machine translation models and evaluation methods. Afterwards, in Chapter 4, I will describe the experimental setup. First, we will explore the datasets provided by TAUS, and this will be followed by the evaluation of the baseline NMT models and the description of different domain adaptation experiments. Then, the performance of translation models adapted using synthetic data will be evaluated, followed by a brief qualitative analysis. Lastly, in Chapter 5, I will answer the research questions. The thesis will be rounded up with a discussion, conclusion, and recommendations for future work.

[8] https://cloud.google.com/translate/

[9] https://www.microsoft.com/en-us/translator/

Chapter 2

Background and Related Work

This chapter will provide an introduction to the field of NMT and discuss domain adaptation methods that can be used to obtain higher-quality translations of sentences from specific domains. Additionally, approaches to data augmentation using synthetic data in the wider context of Natural Language Processing will briefly be presented. Lastly, methods of synthetic parallel corpora generation that can be used in the context of domain adaptation for NMT will be discussed.

2.1 Successes and Challenges of Neural MT

“Efforts to build machine translation systems started almost as soon as electronic computers came into existence” (Koehn, 2020, p. 33). Warren Weaver was the first to propose using computers to aid translation (Weaver, 1952). The first functioning machine translation systems, developed in the post–World War II period, were rule-based. They depended on meticulous work by linguists who hand-crafted dictionaries and instructions on how to translate from the source to the target language. Rule-based systems continued to dominate the field of Machine Translation throughout the 20th century, with data-driven approaches being developed alongside them since the 1980s (Koehn, 2020).

One of those early data-driven methods was proposed by Nagao (1984), who envisioned an example-based system that would be built by providing many example sentences and their translations, reflecting the way humans learn a (foreign) language. Parallel corpora, i.e. (usually very many) sentence pairs each containing a sentence in the source language and its translation in the target language, are a cornerstone of all data-driven approaches, with the idea of enabling machines to extract the translation rules themselves. The first data-driven approach to gain prominence was Statistical Machine Translation (SMT), that benefited from the accelerating development of the internet in the early 21st century, bringing with it access to large scale parallel corpora, as well as the increasing speed of computation. A plethora of systems were developed, both in the academia and the industry (Koehn, 2009).

SMT systems were consistently obtaining *state-of-the-art* (SOTA) results until 2014, when Bahdanau et al. (2015) developed the first competitive Neural Machine Translation system, an encoder-decoder model with attention that could obtain results comparable to that of the best SMT models. This achievement is even more striking when we consider that the first NMT models with encoder-decoder architecture have only been proposed earlier that same year (Sutskever et al., 2014; Cho et al., 2014). Even though neural methods for MT have been researched as early as the 1980s and 1990s (Bastings, 2020), computational complexity as well as data scarcity made them unfeasible until much later. But, once those obstacles were lifted, NMT took over the field of MT in just a few years. NMT models are today ubiquitous in research, and widely used commercial models such as Google Translate, Amazon Translate and Microsoft Translator all utilize NMT architectures.

The idea behind NMT is to build a single large neural network that will, in an end-to-end fashion, take as input a sentence in the source language and output the correct translation in the target language. The whole network is trained jointly with the goal of maximizing the probability of correct translations. It consists of a source-language encoder that encodes the input sentences, and a target-language decoder that reads the encoded input and outputs the translation. Early models encoded the source sentences into fixed-length vectors (Cho et al., 2014). As a consequence, they performed poorly on long sentences, which could not be encoded appropriately. Since the introduction of the attention mechanism by Bahdanau et al. (2015), which addresses this issue, all *state-of-the-art* models integrate it into their architectures. Bahdanau et al. (2015) use recurrent neural networks for the encoder and the decoder, and later, convolutional neural network approaches are also developed (Gehring et al., 2017). The current SOTA NMT models are Transformer models with self-attention, first proposed by Vaswani et al. (2017). The Transformer model is based only on the attention mechanism, and requires significantly less training time than convolution or recurrence-based systems, while obtaining better results (Vaswani et al., 2017).

NMT also entails specific challenges. In particular, NMT models perform poorly in low-resource settings and on data that is significantly different from the data they have been trained on, and they are sensitive to noise in the training data (Koehn and Knowles, 2017; Zhang and Zong, 2020; Koehn, 2020). Even though, with ample training data, NMT systems perform better than SMT models, their performance deteriorates quicker when scaling down the size of the corpus (Koehn and Knowles, 2017). Not only are parallel corpora small or non-existent in most of the world language pairs, even for high-resource languages, there are many domains where little or no parallel data is available. NMT models are also less robust than SMT systems, meaning that higher quality data is needed to obtain satisfactory results (Koehn and Knowles, 2017). The challenges listed all come into play when we want to translate data from a specific domain, which is different from the data employed in training the model, and a usual requirement in the "real-world" scenario. The challenges of performance of MT models on data from specific domains are addressed using domain adaptation, which is the topic of the next section.

2.2 Domain Adaptation in Machine Translation

In machine translation, a domain is usually "defined by a corpus from a specific source, and may differ from other domains in topic, genre, style, level of formality, etc."(Koehn and Knowles, 2017, p. 29). van der Wees (2017) provides a more detailed analysis of the way the term *domain* is used in the Machine Translation field. They distinguish between provenance, topic and genre. Provenance tells us about the origin of a document and cannot be gleaned from the data itself. An example would be Europarl Corpus[1]. Topic stands for the subject of the corpus, and can be broad (e.g. politics) or narrow (e.g.

[1] https://www.statmt.org/europarl/

fishery regulation). Genre refers to non-topical text properties, such as formality or other stylistic properties (e.g. parliamentary language). Still, they emphasize that the way the term *domain* is used by many domain adaptation researchers is ambiguous, and that domain very frequently just means a "different data set" (van der Wees, 2017, p. 33).

The usage of terms *in-domain* and *out-of-domain* can also be ambiguous. For example, Koehn and Knowles (2017) mention the *out-of-domain performance* of NMT systems, meaning the performance of a system trained on one domain, and tested on another. But, more often, *out-of-domain* refers to data that is not relevant to the domain of interest, which we are trying to adapt our model to. This usage will be retained in this work. The term *in-domain* will be used to refer to the domain of the document we are interested in translating, and the domain we want to adapt our model to. The models we will be adapting are usually trained on huge amounts of data from various domains and are meant to perform well on a number of translation tasks. We will refer to this as *general domain*.

The problem encountered the most often when translating texts from different domains is that words have different meanings and frequency of use, and that stylistic features, such as sentence length or politeness level, can vary significantly. NMT models trained on general-domain data do not retain the same level of performance on domain-specific translation tasks. A model trained on news corpora or data scraped from the web might perform abysmally if we task it with translating medical texts. But even if a part of the data the model has been trained on did include medical texts, the general-domain model will usually not perform as well as it could if it was build with a more specific goal of translating medical texts in mind.

The role of domain adaptation for NMT is to address those challenges. Generally, domain adaptation approaches can be classified as either model-based or data-based (Chu and Wang, 2018; Saunders, 2021). Model-based methods focus on changing the architecture of the model in some way, to facilitate learning to perform the translation task on data from a specific domain. Data-based methods, on the other hand, focus on the data that will be used for domain adaptation.

There is no *one-fits-all* approach for domain adaptation (Koehn, 2020). As a rule of thumb, as with domain adaptation in NLP more generally, the method that can be employed will depend on the specific scenario and the resources that we have at our disposal.

For the context of this thesis, the relevant approaches to domain adaptation are data-centric. TAUS uses proprietary models and associated, proprietary methods of domain adaptation, which entails that the datasets used for domain adaptation are the only elements that can be influenced or changed.

As was already mentioned, having relevant in-domain parallel data of sufficient quantity and quality is not something we can expect in many cases. Even for high-resource language pairs, there are many domains that lack such parallel corpora. When it comes to low-resource languages, we usually cannot even obtain sufficient general-domain parallel datasets, let alone parallel corpora in a specific domain. This is where the data-centric method of generating synthetic parallel in-domain corpora (Chu and Wang, 2018; Saunders, 2021), the focus of this work, comes into play. In the next section, we will give a brief introduction to synthetic data generation in the context of NLP, after which we concentrate on approaches in the field of NMT.

2.3 Data Augmentation and Synthetic Data for NLP

High-quality data in sufficient quantity is a prerequisite for training a well-performing machine learning system. However, obtaining these data is always costly and sometimes outright impossible[2]. Data augmentation (DA) and synthetic data generation (SDG) are methods that aim to produce more data without actually collecting and labelling more natural data. The line between those two methods is not clear-cut (Nikolenko, 2021), so I will first take a moment to try and delineate them, before briefly discussing their place in NLP.

Nikolenko (2021) defines data augmentation as the "first step to synthetic data" (Nikolenko, 2021, p. 88). Data augmentation techniques increase the quantity of available data by changing natural instances in ways that result in predictable changes of associated labels. The field of computer vision pioneered these techniques and uses them extensively, which is understandable. Shifting, cropping, rotating, changing the colour of or blurring an image can result in many new training examples, with the image still depicting the same entity (for example, a cat). Synthetic data, on the other hand, would be data produced completely artificially, such as using generative models to produce images of cats based only on the label *cat* (Nikolenko, 2021).

DA and SDG are gaining popularity in recent years, primarily because of the advent of data-hungry deep learning. Additionally, as machine learning is gaining popularity, many new tasks and domains are being explored, where there is often not enough natural data available. Another important use-case for synthetic data besides limited data availability is privacy, since there are domains where enough data is available, but those data can not be shared because of privacy concerns, such as the medical domain.[3] All of those driving forces also hold for NLP, but our field poses specific challenges to utilizing them because of the discrete nature of language data. While it is easy to slightly change an image and have it still depict the same entity, removing a word from a sentence or changing word order might result in generating completely ungrammatical instances.

Recently, Feng et al. (2021) published the first survey paper on data augmentation approaches for NLP. It showcased increased interest in the approach in recent years, with many new methods and techniques being explored and applied to different tasks. Techniques range from rule-based data augmentation, for example Easy Data Augmentation (EDA, Wei and Zou (2019)), where data perturbations, such as synonym replacement, are performed on the token level, to model-based techniques, where purely synthetic data can be produced, for example using translation or generative language models. As for the tasks, the most techniques have been developed for text classification, with machine translation figuring as the second task for which the most papers have been published. In the next subsection, we dive deeper into how data augmentation and synthetic data are used in the NMT field.

[2] Here, one can think of data that could be used to counteract specific "real-world" biases, for example the fact that some professions are mostly "male" or "female". If we would like to train a model that would not reflect this "real-world" bias, we can imagine producing additional synthetic data the model will be trained on, which cannot be obtained naturally.

[3] There is an interesting application for NLP in this regard, creating synthetic clinical notes that can be published while ensuring complete anonymity for the patients. See, for example, Melamud and Shivade (2019).

2.4 Synthetic Data for Neural Machine Translation

In the last section, we mentioned that Machine Translation is one of the fields where synthetic data is often used for data augmentation. That this is so, is not completely surprising, if we consider that, by translating a sentence by a machine translation system, we already obtain a synthetic sentence.[4]

Different data augmentation techniques and synthetic data have already been used in the SMT framework (for example, Bertoldi and Federico (2009)), but here we will concentrate on NMT. The two most frequently encountered use-cases for synthetic data both have to do with the low-resource setting, either in the context of domain adaptation, where not enough parallel in-domain data is available, or low-resource languages. Other possible use cases are in training more robust models (synthetizing additional noisy data) and mitigating biases (for example, gender bias, which is very prominent in MT, can be addressed by generating additional synthetic instances that include underrepresented genders). This work is focused on the domain adaptation problem, but we note that many methods are similar or shared between domain adaptation and low-resource language translation.[5]

The most widely used methods are translation-based. Those methods allow us to leverage monolingual data that is usually a lot easier to come by than parallel data. Since training NMT models always entails having a parallel corpus, translation-based methods can be seen as fundamental, and are needed by virtually all approaches. This is why we will concentrate on them in this thesis.

2.4.1 Back-translation

One of the most widely used methods for synthetic parallel corpora generation is back-translation. The idea behind this method is simple. If there is no or not enough parallel data available, but we have ample monolingual target side data, we can use a target→source NMT system to (back-)translate target monolingual sentences into the source language. The synthetic parallel corpus obtained as a result can then be used to train the source→target model.

This method was first explored for NMT by Sennrich et al. (2016). It is primarily used for training translation models, and is today a *de-facto* standard in training high-performing NMT systems, where huge amounts of monolingual target side data are translated to source and added to natural parallel corpora (Edunov et al., 2020). Sennrich et al. (2016) also evaluated the potential of back-translation for domain adaptation. They have shown that fine-tuning a general domain source→target model using synthetic parallel in-domain data obtained by back-translating a monolingual target in-domain corpus, although not as successful as using natural in-domain parallel data for adaptation, is still effective. In their research, models adapted using synthetic data obtained an improvement over their non-adapted counterparts.

The data obtained by back-translation seems to works by strengthening the model decoder (Burlot and Yvon, 2018). Note that the data on the target side is actually natural, either a human-produced utterance or a human-produced translation. This also means that the noise obtained on the source side of the synthetic corpus, by

[4] I first completely realized this fact when my supervisor, Sophie Arnoult, mentioned it in a discussion. March 2022, personal communication.

[5] And low-resource language translation can also be posed as a domain adaptation problem (Saunders, 2021).

translating targets using imperfect MT models, should not influence the model adapted with synthetic data too much, since it should not be tasked with translating similar faulty examples.

After Sennrich et al. (2016) seminal work on back-translation for NMT, numerous other studies have been published, testing and refining their approach, among others Edunov et al. (2018), Poncelas et al. (2018) and Burlot and Yvon (2018). We will explore those works shortly, but before that, we will mention the other possibilities when using translation to generate synthetic data: forward translation or self-learning and a combination of forward and back-translation.

2.4.2 Forward translation

If we translate the source and the target side of parallel corpora into machine learning terminology of instances and labels, back-translation can be classified as *bona fide* synthetic data generation, since in the case of back-translation, we generate instances from labels. The direction of translation can also be the other way around. If we have a source→target model and source side monolingual data, we can use the model to translate the source data into target, and then use the synthetic parallel corpus obtained by pairing natural sources and synthetic targets to further train or fine-tune the same model we used for translation. This method was first explored for NMT by Zhang and Zong (2016), and is also referred to as *self-learning* or *self-training.*

While back-translation is useful because it strengthens the decoder (natural data is on the target side), forward translation should work by strengthening the encoder (natural data is on the source side, while the target side is synthetic data produced by the MT engine). While we would expect that noise in the form of incorrect translations will have a greater influence than when using back-translation, and although it seems at first sight that this method is not as useful (Bogoychev and Sennrich, 2019), Bogoychev and Sennrich (2019) have shown that the effectiveness of the translation direction (forward or back) can also depend on whether the sentences being translated were originally in the source or in the target language. Although forward translation can obtain superior results in terms of automatic evaluation in some cases when source sentences have originally been in the source language, their research has also shown that human evaluators always prefer results obtained using back-translation.

Specifically for domain adaptation, Burlot and Yvon (2018) conducted a detailed study where they compared the usefulness of forward and back-translation, and their results indicate that back-translation should be the most useful method for domain adaptation. On the other hand, Chinea-Ríos et al. (2017) have used a combination of data selection and forward-translation to generate synthetic parallel corpora for domain adaptation. Their method is attractive because it allows them to select monolingual source side data that is similar to the test set, before translating it to target.

2.4.3 Combining back-translation and forward translation

To generate a synthetic parallel corpus, a mix of forward and back-translation can also be used (Bogoychev and Sennrich, 2019). Park et al. (2017) proposed mixing equal parts of forward- and back-translated data to build a parallel corpus, and obtained slight improvements over both forward and back-translation when training their models. Burlot and Yvon (2018) evaluate the same approach for domain adaptation, but with different results. In their research, back-translated data proved to be more useful than a mixed corpus.

2.4.4 Refining back-translation

Based on research mentioned in the preceding subsections, we expect that back-translation will be the most useful translation-based method for domain adaptation. In principle, there are three considerations that could influence the quality of synthetic data obtained by back-translation, or how useful they are for adapting an NMT system: the quality of translations on the source side, the quantity of synthetic data, and the relevance of synthetic data for a particular domain. In the rest of this subsection, we explore each of those three considerations in turn.

Translation quality

Regarding the quality of translations, from the first exploration by Sennrich et al. (2016), it seems that, as long as a "good-enough" model is used for back-translation, the quality does not seem to matter much (see also Fadaee and Monz (2018)). Burlot and Yvon (2018) explored the difference between using a poor-quality and a high-quality model for back-translation, and found that better data does result in better domain-adapted models in this context.

Edunov et al. (2018) analysed the impact of different ways of generating synthetic source sentences. They found that, even though most other researchers used back-translated data generated by beam (Sennrich et al., 2016) or greedy (Lample et al., 2018) search, using data obtained by sampling is more effective. They also found that using synthetic data can sometimes match the performance of using real data when training the models.

Data selection or cleaning can be employed to obtain better-quality synthetic data. For example, Xu et al. (2019) calculated a semantic similarity score between the source (synthetic) and target sentences using bilingual word embeddings, and then used a cosine similarity between the two sentence vectors as a measure of translation quality.

Quantity of synthetic data

As was already mentioned, the quality of translations (as long as they are produced with a reasonably well-performing system), does not seem to influence the performance of back-translated data too much. The quantity of this data, however, as well as the ratio of natural to synthetic data, seem to have a much greater impact (Fadaee and Monz, 2018; Poncelas et al., 2018). In Poncelas et al. (2018) investigation, the best performance was obtained when using a ratio of two times as much synthetic, back-translated data, as natural parallel data. In the study by Fadaee and Monz (2018), they experimented with ratios as high as 1 part natural data to 10 parts synthetic data. For them, a ratio of 1 to 4 turned out to be the best performing. They also noticed how the quality of systems adapted with synthetic data does not increase linearly with the increase in the quantity of synthetic data. Systems trained using a 1 to 4 ratio performed only slightly better than systems trained with 1 part natural to 1 part synthetic data. This is in contrast to using natural parallel data for adaptation, where system performance usually scales linearly with the quantity of natural parallel corpora used for adaptation.

Fadaee and Monz (2018) also observed that there seems to be a limit to learning from synthetic data. If we continue using bigger and bigger quantities of synthetic data, at some point, the model will unlearn its parameters and performance will deteriorate. A model adapted with too large quantity of synthetic data might perform worse than an unadapted model. A similar conclusion was obtained by Burlot and Yvon (2018), who observed that using synthetic data obtained by back-translation encourages overfitting.

Relevance of synthetic data

Lastly, the relevance of synthetic data to the domain or test set of interest can also influence the performance of models adapted with this data. The idea is to obtain better performance by picking the most useful sentences, according to some criteria (Fadaee and Monz, 2018). Data selection techniques are widely used in data-based domain adaptation, and the idea is to select sentences that are more relevant to the domain adaptation problem at hand (Saunders, 2021). Poncelas et al. (2019) explored whether data selection could successfully be applied to synthetic sentences generated by back-translation, and concluded that, even though the source side sentences can be noisy, data selection seems to be useful for synthetic data. More concretely, for the domain adaptation scenario, Poncelas and Way (2019) have used the test dataset as a seed to retrieve synthetic sentences that will be used for adaptation. They have shown that, in certain scenarios, synthetic sentences can even be more useful than natural sentences.

2.5 Summary

In this chapter, we briefly introduced the field of Natural Machine Translation and discussed the challenges of translating texts from specific domains using NMT systems that haven't been explicitly built with the goal of excellent performance on data from those domains. Then, we discussed approaches to addressing this *in-domain* performance of NMT systems using domain adaptation. Since the goal of this thesis is evaluating the usefulness of synthetic data generation for domain adaptation, we briefly discussed its place in the field of Natural Language Processing, before diving deeper into specific techniques relevant for the field of Machine Translation. The most space was given to the translation-based methods of creating synthetic data, and chiefly the back-translation method, for which we expect to be the most useful when it comes to the specific context of this thesis.

Translation based methods—back-translation, forward translation, as well as their combination—and the refinements of the back-translation method, including obtaining translations of higher quality, experimenting with the quantity of synthetic data, and data selection applied to synthetic data, will guide the work presented in the rest of this thesis. First, in the next chapter, we will discuss the methodology, including NMT models and evaluation methods, before presenting experimental results in Chapter 4.

Chapter 3

Methodology

In this chapter, I will describe the NMT models that will be used in the experiments, how the quality of synthetic data will be evaluated, and how sentence embeddings will be used for cleaning the data and selecting data that is more relevant for a particular domain.

The experiments, that will be introduced in the next chapter, will employ two NMT models, one proprietary (Amazon Translate), and one open-source (OPUS-MT). *Off-the-shelf* models will be used to produce synthetic data, and adapted using natural and synthetic data. Experiments will be evaluated using two string-based metrics (BLEU and chrF score), and one neural, pretrained metric (COMET). LASER multilingual embeddings will be used for cleaning and data selection.

3.1 Models

The primary goal of this research, from the viewpoint of TAUS, was to evaluate the usefulness of synthetic data for domain adaptation in the context of models they have implemented in their production workflow.

As we mentioned in the Introduction, at the moment of starting to write this thesis, TAUS DEMT service was live using Amazon's Active Custom Translation framework, with a plan to extend the service to Google's Auto ML and Microsoft's Custom Translator. All three providers use a neural machine translation architecture for their baseline models, while the adaptation method differs between Amazon's ACT on the one side, and Google and Microsoft on the other. Since all the models are proprietary, the exact details about the architectures are, of course, unknown. But, from the descriptions available, as well as the training times, it can be concluded that customization of Google and Microsoft is a sort of fine-tuning the underlying model for a few more epochs, while Amazon deploys an "on-the-fly" customization method, described in more detail in the next section. Since the goal of this research is to evaluate the usefulness of synthetic data for data-based domain adaptation, the fact that the inner workings of a customization method are unknown should not be viewed as a deterrent from using the method. Additionally, since the thesis is executed in partnership with a company, evaluating the model the company uses is important.

On the other hand, relying on proprietary models is also problematic. For one, models can change overnight, making the research done to evaluate the model obsolete. Additionally, the fact that we cannot know the exact mechanism by which they work limits our understanding of which methods might be useful. In a way, we are destined to throwing different corpora at the model, and seeing what sticks. Thirdly, using those models is costly, thus also from the standpoint of the company, it would make sense to also evaluate an open-source method, especially when it comes to how synthetic data will be generated.

Because of those reasons, instead of using the Google's and/or Microsoft's NMT model, I chose to evaluate the open-source OPUS-MT model (Tiedemann and Thottingal, 2020), both as a candidate for generating synthetic data by translation, and as a model that can be customized to better perform on data from a particular domain.

Off-the-shelf models in the source→target direction will be used as unadapted baselines, while target→source models will be used for back-translation. Adapted models will be used to evaluate adaptation using natural and synthetic data.

3.1.1 Amazon Translate and Active Custom Translation

"Amazon Translate is a neural machine translation service that delivers fast, high-quality, affordable, and customizable language translation"[1]. *Off-the-shelf* machine translation models offered through Amazon Translate (AT) are general-domain models that offer translation between 75 languages and 5550 source→target combinations.[2]

Amazon Translate offers two kinds of customization. The first allows the users to customize translations using Custom Terminology, influencing the translation of words and phrases.[3] The other, one we will use in this thesis, is called Active Custom Translation (ACT). Using ACT is "similar to using a custom translation model"[4] trained with users' example translations. Unlike training a custom model, parallel data in the form source example→target translation is used at runtime to adapt translations to "reflect the style, tone, and word choices" that are found in the parallel data submitted by the user.[5] ACT is marketed as an alternative to training custom models, with Amazon claiming that in this way, costs in terms of time and money needed to train custom models are avoided.

The exact mechanism of working for ACT is not clear. It does not seem that it simply takes phrases from the source sentences and their translations to the target language, since this is the mechanism of working for Custom Terminology. "When a custom terminology is used as part of the translation request, the engine scans the terminology file before returning the final result. When the engine identifies an exact match between a terminology entry and a string in the source text, it locates the appropriate string in the proposed translation and replaces it with the terminology entry"[6]. Translations that use custom terminology are priced the same as using *off-the-shelf* models, while translations using ACT are four times more expensive. Also, translation jobs using ACT take substantially more time than using AT, but still a lot less time than it takes to train Google's or Microsoft's customized models.[7]

[1] `https://aws.amazon.com/translate/`, accessed 05-06-2022.

[2] `https://aws.amazon.com/translate/details/`, accessed 05-06-2022.

[3] `https://docs.aws.amazon.com/translate/latest/dg/how-custom-terminology.html`, accessed 05-06-2022.

[4] `https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-parallel-data.html`, accessed 05-06-2022.

[5] `https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-parallel-data.html`, accessed 05-06-2022.

[6] `https://docs.aws.amazon.com/translate/latest/dg/how-custom-terminology.html`, accessed 05-06-2022.

[7] ACT jobs take around 30 minutes, independent of the size of parallel corpora used for adaptation,

3.1.2 OPUS-MT

In contrast to the AT models described in the previous section, when it comes to the open-source OPUS-MT models (Tiedemann and Thottingal, 2020), we know both the data they were trained on and their architecture. They are named after the corpus they have been trained on: the OPUS[8] or Open Parallel corpUS (Tiedemann, 2012). Models are based on *state-of-the-art* transformer architecture, with 6 self-attention layers in the encoder and the decoder, each layer consisting of 8 attention heads (Tiedemann and Thottingal, 2020).[9]

In this work, we use the Hugging Face[10] implementation of OPUS-MT models, more concretely an English→Dutch[11] and Dutch→English[12] model. The latter model will be used for back-translation, while the former will be used for forward translation and fine-tuned with natural and synthetic data. The models are pretrained and can be used *off-the-shelf*, and Hugging Face also provides instructions for adapting the models by fine-tuning them on in-domain data. The fine-tuning procedure is described in the next subsection.

Fine-tuning procedure

Fine-tuning, also known as continued training, first proposed for Neural Machine Translation by Luong and Manning (2015) and further analysed by Freitag and Al-Onaizan (2016), entails extending the training of a model for a few more epochs, during which the model is trained exclusively on in-domain data, with the goal of adapting the model to a specific domain. Fine-tuning is usually fast and performs well, even though some problems may arise, such as overfitting to a dataset that is too small or too noisy (Saunders, 2021). Another potential problem of fine-tuning is catastrophic forgetting, where the model overfits to in-domain data and has worse performance on data from the general domain, data that it previously translated well (Freitag and Al-Onaizan, 2016; Saunders, 2021). As such, catastrophic forgetting should not be a problem for this project, since models are only intended to be used to translate in-domain data.

In order to fine-tune the *off-the-shelf* OPUS-MT en→nl model, we follow the procedure described by Hugging Face.[13] The model is tuned for 3 epochs, with learning rate set to 0.00002, and using the AdamW optimizer. The batch size is set to 8. Fine-tuned model parameters are saved separately after every epoch, which allows evaluating models trained for a shorter time, since it will be shown that, in some cases, training for 2 epochs is better since after 3 epochs, the model overfits to training data.

Fine-tuning is performed using one GPU provided through Google Colaboratory Pro subscription[14]. It takes from around 15 minutes to more than 2 hours, depending on the size of parallel data used for adapting the model.

while training a Google or Microsoft custom model takes a few hours, and this duration does depend on the size of the parallel training dataset used.

[8] `https://opus.nlpl.eu/`

[9] `https://github.com/Helsinki-NLP/Opus-MT-train`

[10] `https://huggingface.co/`

[11] `https://huggingface.co/Helsinki-NLP/opus-mt-en-nl`

[12] `https://huggingface.co/Helsinki-NLP/opus-mt-nl-en`

[13] The code is adapted from `https://huggingface.co/course/chapter7/4?fw=pt`.

[14] `https://colab.research.google.com/`

Decoding method

In the context of this thesis, both the AT and OPUS-MT models will be used to generate the synthetic data. While, for the AT model, it is not possible to set any parameters, and we always obtain just one translation, with the OPUS-MT model we have the freedom of choosing different decoding methods, and also the possibility to generate more than one translation for each input sentence.

By default, the OPUS-MT models as implemented by Hugging Face use greedy search as the decoding algorithm. Greedy decoding is faster than the other most commonly used decoding algorithm, beam search (Stahlberg, 2020), since greedy search selects only one hypothesis - the most probable one - for every possible next word in the output, looking at the probability of this word being the correct translation given all the previous words. However, we will also experiment with another decoding method, namely sampling. This is described in more detail in the next chapter.

3.2 Evaluation

The goal of this thesis is to evaluate if synthetic data can successfully be used for domain adaptation. This entails that the synthetic data will be evaluated extrinsically, looking at how well the models adapted using this data perform. This brings us to the territory of machine translation evaluation. Usually, this process consists of comparing each translation obtained as an output of an MT system to one or more reference translations, produced by professional human translators. Metrics that take into account a reference translation are called reference-based metrics. There are also ways to evaluate the output without needing a reference translation, using a referenceless metric and comparing the translation to the source test sentence.

In this thesis, I will use reference-based metrics, that will compare machine generated outputs to reference translations. Two string-based metrics, BLEU and chrF score, will be used, as well as a pretrained COMET metric. Both the metrics themselves, and the reasons for using each of them, are described in more detail in the following subsections.

Important considerations that were taken into account when deciding which metrics will be used were the fact that we will be evaluating outputs of high-performing systems, and that our systems will be trained using synthetic data. Mathur et al. (2020) showed that automatic metrics can perform differently when used to evaluate high-performing systems, as compared to models that perform less well. Edunov et al. (2020), on the other hand, studied how well automatic metrics evaluate systems trained with back-translated data, and concluded that there are important differences that need to be taken into account when evaluating systems trained using synthetic data, as compared to models trained exclusively with natural parallel data.

3.2.1 BLEU score

The most widely used metric for automatically evaluating machine translation outputs, first proposed 20 years ago, is BLEU score (Papineni et al., 2002), which stands for Bilingual Language Evaluation Understudy. BLEU score is a string-based metric (Kocmi et al., 2021) that considers n-gram overlaps between the machine generated output and one or more reference translations[15]. BLEU score looks only at precision, and thus needs to employ a brevity penalty that penalizes translations that are too short[16]. The score is between 0 and 1 or 0 and 100, with the higher result better (1 or 100 would entail that the reference and translation are identical, while a score of 0 means that there are no overlaps between them).

Even though BLEU has been criticized from the start, it remains a *de facto* standard both for reporting experimental results, and in development, to compare the performance of different models. That this is so, probably comes down to the ease and speed with which this metric can be calculated, as well as a long history of using it in the context of automatic MT evaluation.

To offer an example of why BLEU is so problematic: since it measures n-gram overlaps between a machine-translated sentence and a reference, it will assign the same penalty for using a synonym or an antonym. Additionally, changing word order entails a substantial penalty for a translation that, for all intents and purposes, might be considered equivalent to the reference. In recent years, BLEU has predominately been criticized because it does not seem to correlate well with human judgment, especially when it comes to evaluating high performing systems (Mathur et al., 2020). Edunov et al. (2020) have additionally shown that BLEU scores cannot sufficiently discriminate between systems trained with natural and with synthetic data. Mathur et al. (2020) also point out that BLEU preforms badly when it comes to judging which of the two or more models performs better, as compared to human judgement. In their study, another string based metric, chrF (Popović, 2015), performs better than BLEU, and this is why we will use it as an additional metric, as described in the next subsection.

Still, because results in terms of BLEU scores are reported in virtually all research papers, we keep using BLEU as an evaluation metric in this thesis. Concretely, we use the SacreBLEU implementation (Post, 2018), that standardizes the metric parameters and makes it possible to compare results across different research papers[17].

3.2.2 chrF score

ChrF, or character n-gram F-score, was first proposed by Popović (2015). Basically, chrF takes into consideration the percentage of character n-gram overlaps between the machine-translated sentence and the reference. In contrast to BLEU score, chrF takes into account both precision and recall. Precision looks at how many n-grams in the hypothesis are also present in the reference, while recall calculates the number of n-grams present in the reference that can also be found in the hypothesis. As with BLEU, scores can be between 0 and 1 or 0 and 100, with a higher score implying a translation that better agrees with the reference. To calculate the chrF score, we again use SacreBLEU[18].

Recent studies by Mathur et al. (2020) and Kocmi et al. (2021), which recommend discontinuing the use of the BLEU score, both find that chrF is the string-based metric that correlates the best with human judgment. It should work especially well for

[15] Ideally, BLEU would always be used with more than one reference translation, but in practice this almost never happens.

[16] Otherwise, a one-word translation that contains one of the words in a ten-word reference would obtain a perfect score

[17] `https://pypi.org/project/sacrebleu/`, signature:
BLEU nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0.

[18] Signature: chrF2 nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0.

morphologically more complex languages (Mathur et al., 2020). Mathur et al. (2020) recommend using it instead of BLEU, while Kocmi et al. (2021) go one step further, and recommend that a pretrained metric, COMET score, should be the metric of first choice, with chrF being used to evaluate translation into languages not covered by the COMET score. We describe this metric in the next subsection.

3.2.3 COMET score

The extensive study by Kocmi et al. (2021) was conducted with the goal of finding an automatic metric that is most suitable to be used when developing MT systems, in order to decide which of the two candidate models performs better. They recommend using a pretrained metric, specifically COMET (Rei et al., 2020). COMET (Crosslingual Optimized Metric for Evaluation of Translation) is actually not a single metric, but rather a framework for training deep neural network-based evaluation models that can be used as metrics. Evaluation models are built on pretrained, multilingual language models such as XLM-RoBERTA (Conneau et al., 2020), and take into consideration not just the reference and the machine-translated output, but also the source sentence. The objective of the evaluation models it to learn to model human judgment, assigning higher scores to translations that are deemed better by human evaluators.

We use the Unbabel implementation[19] and the default, reference-based *wmt20-comet-da* model, to calculate the COMET scores. COMET models are trained with z-scores, and thus the score can be lower than 0 or higher than 1.[20]. For a *state-of-the-art* system, the score obtained when using *wmt20-comet-da* model is expected to be between 0.6 and 1.[21] The downside of using COMET is that it is quite computationally expensive, requiring the use of a GPU and taking much more time to calculate than lightning-fast BLEU and chrF scores.[22]

3.2.4 Statistical significance

To ascertain the statistical significance of results, we use the bootstrap method (Koehn, 2004). For two or more systems whose results we want to compare, a paired bootstrap resampling test is run, as implemented by SacreBLEU and COMET libraries. This test allows us to estimate how probable it is that a difference in the mean results obtained by a pair of systems is a result of chance. For example, if we were to test both systems 100 times, how many times would one system perform better than the other? If system A outperforms system B 95 times, we can say that, with a p-value of 0.05, we reject the null-hypothesis, saying that the difference between the two systems is insignificant, and conclude that the difference in performance of the two systems is indeed statistically significant. This does not allow us to be absolutely certain that system A is really better than system B, but it does imply that it is quite improbable that the observed difference in performance is accidental.

[19] `https://github.com/Unbabel/COMET`

[20] `https://unbabel.github.io/COMET/html/faqs.html`, accessed 23-06-2022.

[21] `https://github.com/Unbabel/COMET/issues/14`

[22] The good news is that Unbabel team is already working on creating faster and less computationally expensive versions of COMET. See Rei et al. (2022), which received the Best paper award at EAMT 2022.

3.3 Cleaning and Matching Data with LASER Embeddings

Since the seminal work by Mikolov et al. (2013), embeddings have been successfully used as a measure of word and sentence similarity. The idea behind word and sentence embeddings is to encode a word or a sentence into a fixed length vector. Vectors of different words or sentences can then be positioned into a shared space, and the distance between them can be used as a measure of their semantic similarity.

While first approaches to word embeddings worked with monolingual data, soon multilingual and cross-lingual embeddings gained prominence (Ruder et al., 2019). The idea behind them is to project embeddings in two or more languages into a joint vector space, where similarity can again be modelled.

One of the widely used pretrained multilingual sentence embeddings is LASER or Language-Agnostic SEntence Representations (Artetxe and Schwenk, 2019). The idea behind LASER embeddings is to generate embeddings that will be language and task independent. A single encoder processes sentences in many languages, and as a result, embedded sentences in different languages that are semantically similar should end up being close in the vector space. Relevant for this work, this means that, by employing a distance measure, such as cosine similarity, we can obtain a similarity score between a pair of sentences. Then, we can use this score to filter our noisy sentences or sentences that are closer to a particular domain or test set.

LASER embeddings have been selected since they are already widely used in TAUS. For example, one of the preprocessing steps when compiling parallel datasets provided by TAUS for this research, as will be described in the next chapter, was calculating cosine similarity of LASER embeddings between pairs of sentences. Then, the obtained score was used to filter out lower-quality translations.

In this work, LASER embedding will be used for three different tasks:

- data cleaning, in order to obtain higher-quality translations;
- data selection, finding sentences that are more similar to the test set;
- exploratory data analysis, to obtain sentences that are the most relevant for the domain of each dataset.

Using sentence embeddings and cosine similarity to clean and filter parallel sentences is well-supported by research. For example, Schwenk (2018) proposed using cosine similarity to filter noisy sentences and to mine for possible translations. They also mentioned that the same approach could be used to filter back-translated data.

Cosine similarity can also be used to find data that is more similar to an entire dataset. For example, we can use the test set as a seed, and, calculating cosine similarity with sentences from the training set, find those sentences that are more similar to the test set.

Calculating cosine similarity matrices can also be used in order to rank sentences according to their similarity to all the other sentences in a certain corpus. This allows one to rank sentences so that the most relevant sentences rank the highest. This will be used in the next chapter, when exploring the datasets we will be working with.

3.4 Summary

In this chapter, we described the NMT models that will be used in the experiments and the adaptation methods that will be employed to obtain domain-customized translations. We dedicated the bulk of the chapter to describing the evaluation metrics we will use to ascertain which synthetic data generation method produces corpora that are the most useful for domain adaptation. Lastly, we briefly described LASER embeddings, that will be used for three specific purposes in the reminder of the thesis: data cleaning, data selection, and exploratory data analysis. In the next chapter, we start by describing the datasets that we are going to use, and then dive into describing the experimental setup and analysing experimental results.

Chapter 4

Experiments

For the experiments, a similar setup was devised as presented by Burlot and Yvon (2018), who investigated back-translation, forward translation and a mix of the two methods for domain adaptation. As in their work, as a baseline, general-domain, *off-the-shelf* unadapted models will be used, and as an upper bound, the result we do not expect to beat, we will use adaptation with natural data. Then, in each of the experiments, different portions of target or source natural datasets will be translated to construct synthetic parallel corpora.

Before diving into the experiments, though, parallel datasets we will be working with will first be presented. Next, we will evaluate the baseline and upper bound systems. Lastly, the setup of each experiment will be described, followed by reporting and discussing experimental results. We will round up the chapter with a brief qualitative analysis.

4.1 Datasets

For this research, TAUS provided English→Dutch parallel corpora in three domains: Financial Services (Fin), Pharmaceuticals & Biotechnology (Pharma) and Retail & Wholesale Distribution / E-Commerce (E-Comm). These are the same corpora that are used in their DEMT pipeline, to provide customized translations in the English→Dutch translation direction in those domains[1].

I was provided with 3 parallel datasets (used as customization data by DEMT), and 3 test datasets of 2000 instances each, which were selected at random from the initial corpora curated from a large repository of translations by applying different selection methods, and used to evaluate DEMT performance. As I mentioned in the last chapter, cosine similarity was used as a cleaning method when curating the datasets, and only segments with LASER embeddings cosine similarity between 0.9 and 0.99 were selected, as this was taken to mean they are good translations. Most instances should contain a natural English sentence (source), and a human translation to Dutch (target), but this is not guaranteed nor is there a way, in the scope of this project, to check whether certain sentences comprise a direct or a reverse portion (direct meaning they are source natural sentences translated into target, and reverse referring to the reverse scenario).

The datasets did not contain any duplicate source→target pairs, but they did contain some duplicate source sentences, and some duplicate targets. That this is so, is

[1] `https://datamarketplace.taus.net/enhance-mt`; also available on AWS marketplace: `https://aws.amazon.com/marketplace/seller-profile?id=5e008837-9f31-46ca-9797-74ceea721e4d`.

not surprising or problematic *per se*, since the same source sentences can have different translations, and different source sentences can have the same translation. Still, in the context of this project, I wanted to clean all the duplicate sources and all the duplicate targets. Duplicate sources were cleaned because Active Custom Translation does not use them anyway. If the user provides a dataset that contains duplicate sources, ACT will only make use of the last source, or the source with the most recent date[2], with other duplicate sources filtered out from the parallel corpus. Google's AutoML does not use duplicate sources either[3], and thus this seems to be the default in the industry. Cleaning the data ourselves, instead of letting providers clean the data for us, should ensure reproducibility of the experiments. Duplicate targets were cleaned with an eye out for the method we will use most frequently to obtain synthetic parallel corpora, namely back-translation. Even though having multiple (non-duplicate) sources for the same target might actually be beneficial when using synthetic corpora generated by back-translation, as reported by Imamura et al. (2018), if we went ahead and translated duplicate targets using the *off-the-shelf* AT or OPUS-MT model, we would each time obtain the same source sentence, which we would again have to remove before using this synthetic parallel corpus.

domain	en-nl sentence pairs
Financial Services	174025
Pharmaceuticals & Biotechnology	84862
Retail & Wholesale Distribution / E-Commerce	37986

Table 4.1: Number of sentence pairs per domain before cleaning.

As a first step in preparing the data, I combined the original training and test datasets that were provided in each of the domains. Statistics about those combined corpora are given in Table 4.1. Next, I used a cleaner provided by TAUS[4] to clean the data. The cleaner both fixes the text according to prespecified rules (for example, it fixes the quotation marks and content extracted between HTML tags), and flags problematic instances, such as duplicate sources and targets and instances where one of the sentence pairs is much shorter than the other, indicating a likely wrong translation or a misaligned sentence. pair. It also flags the sentences that are very long (longer than 100 tokens). After removing all the data that was flagged by the cleaner, including all duplicate sources and duplicate targets, from each of the datasets, I randomly selected 2000 sentence pairs to serve as a test set, and 2000 sentence pairs that will be used as a development set when fine-tuning the OPUS-MT model. TAUS *data-language-cleaner* library also provides the token counts per sentence (not counting the punctuation tokens). Basic statistics about the datasets in the three domains after cleaning are provided in Table 4.2.

In Table 4.3, I present sentences from the test set that are very relevant for each of the domains. Those sentences were obtained by using LASER embeddings and calculating a cosine similarity matrix across each of the test datasets.[5]

[2] `https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-parallel-data-input-files.html`

[3] `https://cloud.google.com/translate/automl/docs/`

[4] `https://github.com/TAUSBV/data-language-cleaner`, presently only available for internal use.

[5] For this, I used another library developed by TAUS, called CosineSimilarityMatrix: `https://github.com/TAUSBV/CosineSimilarityMatrix`. Presently only available for internal use.

domain	# train	# dev	# test	# avg tok source	# avg tok target
Fin	160395	2000	2000	23	23
Pharma	79029			22	23
E-Comm	32467			25	25

Table 4.2: Statistics for cleaned datasets.

Financial Services
Contracting authorities should introduce appropriate contractual safeguards into their supply agreements to the effect that the amount and delivery schedule of ordered euro banknotes may be changed within the limits established by the ECB. *Aanbestedende diensten dienen passende contractuele waarborgen in hun leveringsovereenkomsten op te nemen inhoudende dat bedrag en leveringsschema van bestelde eurobankbiljetten binnen de door de ECB vastgestelde grenzen kunnen worden gewijzigd.*
Pharmaceuticals & Biotechnology
In a study in subjects with varying degrees of renal impairment, mild to moderate renal disease had no influence on plasma concentration of rosuvastatin or the N-desmethyl metabolite. *In een onderzoek bij patiënten met verschillende gradaties van nierinsufficiëntie had milde tot matige nierinsufficiëntie geen invloed op de plasmaconcentratie van rosuvastatine of de N- desmethylmetaboliet.*
Retail & Wholesale Distribution / E-Commerce
Standard features include a locked-down internal USB port, chassis intrusion switch, locking bezels and a built-in Trusted Platform Module (TPM) which enables system authentication, assists with encryption and helps prevent tampering. *Tot de standaardvoorzieningen behoren een vergrendelde interne USB-poort, een schakelaar die het openen van het chassis detecteert, vergrendelbare randen en een ingebouwde Trusted Platform Module (TPM) die systeemverificatie mogelijk maakt, assisteert bij versleuteling en sabotagepogingen helpt voorkomen.*

Table 4.3: Examples of relevant sentences from the test sets for each of the domains.

4.2 Baseline: Performance of Unadapted Models

As a baseline for all the other experiments, Amazon Translate and OPUS-MT English to Dutch (en→nl) *off-the-shelf* models were used to translate the source (English) portion of the test set into Dutch. The performance in terms of BLEU, chrF2 and COMET scores is reported in Table 4.4. In this table and all the following tables, best results are marked in bold, and * implies that the difference is statistically significant at p=0.05 or less. For baseline and upper bound systems, we compare the two models, while when reporting experimental results, we always compare the systems to the baseline.

As we can see from the table, BLEU and chrF2 scores indicate that *off-the-shelf* models perform the best when translating data from the financial domain. The performance is worst for the e-commerce domain, while results for the pharmaceutical

domain & model	BLEU	chrF2	COMET
Fin			
AT	49.95	72.80	0.8092
OPUS	**51.85***	**74.00***	**0.8223***
Pharma			
AT	45.72	71.37	0.8152
OPUS	**48.61***	**72.90***	**0.8298***
E-Comm			
AT	**41.11***	**68.28***	**0.7464***
OPUS	39.49	66.72	0.7105

Table 4.4: Evaluation of unadapted en→nl models.

domain are positioned in the middle. COMET scores differ in that they assign the best performance when translating sentences from the pharmaceutical domain, followed by those from the financial and e-commerce domains. All evaluation metrics agree that OPUS-MT models perform better on data from the financial and pharma domains, while Amazon Translate model performs better when it comes to e-commerce. An important fact to notice when looking at the baseline results is that they are already pretty high, especially if we compare them to the results *then-state-of-the-art* NMT systems were obtaining just a few years ago, as in the papers we referenced when describing related work on translation-based synthetic data generation such as Burlot and Yvon (2018), Poncelas et al. (2018) and Edunov et al. (2018). Because of this, it will be very interesting to see if we obtain comparable results, since we will be working with systems that perform much better before domain adaptation.

We also evaluate the other translation directions, nl→en, by taking the target side portion of the test set (Dutch) and translating it into English. This will be the direction for producing synthetic back-translated data. The results in terms of BLEU, chrF2 and COMET scores are reported in Table 4.5.

domain & model	BLEU	chrF2	COMET
Fin			
AT	54.13	74.10	0.765
OPUS	**55.12**	**75.63***	**0.7962***
Pharma			
AT	52.06	74.07	0.8274
OPUS	**56.62***	**76.78***	**0.8544***
E-Comm			
AT	**45.85**	**69.66**	0.7564
OPUS	45.66	69.71	**0.7643***

Table 4.5: Evaluation of nl→en models that will be used for back-translation.

As we can see, when translating in the nl→en direction, all metrics agree that best translations are produced in the pharmaceutical domain, with the financial domain a close second. The e-commerce domain lags behind, although different metrics disagree about the scale of difference in performance. The fact that we obtain better translation in this translation direction, is probably, at least in part, attributable to the fact that, when translating Dutch to English, we are translating sentences that are themselves translations of the source (English) original sentences. As shown by numerous studies, discussed by Edunov et al. (2020), sentences that are themselves a translation (translationese sentences) are easier to translate than sentences that are original text (source original sentences).

4.3 Upper Bound: Adaptation with Natural Data

Following Burlot and Yvon (2018), we use adaptation with natural data to define a *topline* system or an "upper bound of translation performance" (Burlot and Yvon, 2018, p. 145). In this scenario, natural parallel training datasets, as described at the start of this chapter, were used to adapt the models. In the case of ACT, they were added as parallel data to AWS, and used to customize translations of the source portion of the test data (English) into Dutch. OPUS-MT *off-the-shelf* model, on the other hand, was fine-tuned with this natural parallel corpora, using the fine-tuning procedure described in the last chapter, to obtain the fine-tuned OPUS-FT models. Results are reported in Table 4.6.

domain & model	BLEU	chrF2	COMET
Fin			
ACT	**55.05**	**75.61***	**0.8245**
OPUS-FT	54.36	75.13	0.8216
Pharma			
ACT	**51.55***	**74.33***	**0.8342**
OPUS-FT	50.33	73.81	0.8340
E-Comm			
ACT	**46.34***	**71.05***	**0.7701***
OPUS-FT	44.28	69.83	0.7531

Table 4.6: Evaluation of adaptation using natural parallel data.

The metrics again disagree, with BLEU and chrF2 painting a picture of adapted models that perform significantly better than the baselines, with an average increase in performance of more than 5 BLEU points or almost 3 chrF2 points for ACT, and 3 BLEU points or almost 2 chrF2 points for the OPUS-FT models. As for COMET evaluation, it shows significant improvements of almost 0.02 points on average for ACT. OPUS-FT models adapted with data from the financial and pharmaceutical domains, on the other hand, do not perform significantly different from the baseline models in terms of COMET scores. The model adapted with e-commerce data, though, shows the highest improvement over the baseline of any model in terms of COMET score. Granted, the baseline OPUS-MT model, when translating data from the e-commerce domain, is also the lowest performing of all baselines.

In the reminder of this chapter, when reporting experimental results, we will always show both the baseline results, as presented in the last section, as well as the upper bound results. Note that the performance of models adapted with synthetic data will always be compared to the performance of the baseline models, while upper bound results will be shown for the convenience of the reader, so that it would not be necessary to look them up in the tables presented in this section.

4.4 Experimenting with Synthetic Data

After introducing our datasets, evaluating the baseline models, and defining upper bound performance obtained when adapting models with natural data, in this section we will present five experiments that were executed in the course of this research. The experiments roughly follow the organization of the section on Synthetic Data for Neural Machine Translation in Chapter 2. The first two experiments evaluate the performance of using different translation directions to obtain synthetic parallel corpora: back-translation, forward translation, and the combination of the two methods. After that, we conduct three experiments that seek to examine back-translation in more detail. The first is meant to evaluate the influence translation quality has on back-translated parallel corpora. Next, we look at using different ratios of synthetic to natural data. Lastly, we experiment with data selection in order to filter out the data more similar to the test sets.

4.4.1 Simple back-translation

To conduct the first experiment, AT and OPUS-MT *off-the-shelf* nl→en models, evaluated in the Baseline section, were used to back-translate the target portion of the natural parallel dataset (Dutch data) into English. Synthetic parallel corpora that were obtained by combining the translated source side and the target side of the original datasets were then used as parallel data for ACT, and to adapt the OPUS-MT model.

After generating synthetic parallel corpora and before using them for adaptation, the same cleaning method that was used for cleaning natural parallel datasets was employed since, after generating the source side data, we obtained some new duplicate sources that needed to be removed.[6] The same will be repeated in all the following experiments if there is a need to clean duplicate source sentences.

Table 4.7 presents results of adapting the AT model. The only disagreement between the metrics is that COMET again, as for the baseline and upper bound models, assigns a higher score for the models adapted to the pharmaceutical domain, as compared to the financial domain. All adapted models obtain higher scores than the baselines, with the difference being statistically significant. Models adapted with synthetic data obtained by back-translation using OPUS-MT *off-the-shelf* nl→en model perform better than those translated with the AT model in all cases, although the difference is not statistically significant in most of them. As we have seen, OPUS-MT model did also obtain better results in almost all cases in the nl→en translation direction, so this is not surprising. More interesting is the fact that the worst performing baseline, e-commerce, gained the smallest improvements in terms of all the scores. While for the financial and pharmaceutical domain, performance of models adapted with purely synthetic data obtained by back-translation is somewhere in the middle between the performance of the unadapted model and the upper bound models adapted with natural data, for the e-commerce domain, the performance is much closer to that of the baseline model. Note that this is also the domain where the smallest quantity of data is available, less than half of data available for the pharmaceutical or quarter of data available for the financial domain. I am not sure if I should speculate that this disparity in data quantity could be the reason for the different performance, though. Note also

[6] This implies that a number of sentences in the original dataset were very similar, thus when translating the target side Dutch sentences we obtained duplicate English sources.

domain & model	BLEU	chrF2	COMET
Fin			
AT baseline	49.95	72.80	0.8092
ACT ATbt	52.05*	74.03*	0.8123
ACT OPUSbt	**52.27***	**74.11***	**0.8133***
ACT upper bound	55.05	75.61	0.8245
Pharma			
AT baseline	45.72	71.37	0.8152
ACT ATbt	47.68*	72.35*	0.8204*
ACT OPUSbt	**48.41***	**72.79***	**0.8229***
ACT upper bound	51.55	74.33	0.8342
E-Comm			
AT baseline	41.11	68.28	0.7464
ACT ATbt	42.62*	69.05*	0.7538*
ACT OPUSbt	**42.80***	**69.16***	**0.7542***
ACT upper bound	46.34	71.05	0.7701

Table 4.7: Performance of ACT using fully synthetic parallel corpora obtained by back-translation.

that this disparity did not seem to influence the performance of upper bound models, where all systems have shown comparable increases in performance (with the lowest performing e-commerce model actually showing comparatively the biggest increase in performance when compared to the baseline in terms of the COMET score).

domain & model	BLEU	chrF2	COMET
Fin			
OPUS baseline	51.85	74.00	**0.8223**
OPUS-FT ATbt	**51.91**	**74.22**	0.8059*
OPUS-FT OPUSbt	50.98	73.87	0.8006*
OPUS-FT upper bound	54.36	75.13	0.8216
Pharma			
OPUS baseline	**48.61**	72.90	**0.8298**
OPUS-FT ATbt	48.03*	72.58*	0.8113*
OPUS-FT OPUSbt	48.55	**72.95**	0.8134*
OPUS-FT upper bound	50.33	73.81	0.8340
E-Comm			
OPUS baseline	39.49	66.72	0.7105
OPUS-FT ATbt	**42.17***	**68.81***	**0.7346***
OPUS-FT OPUSbt	40.68*	68.01*	0.7207*
OPUS-FT upper bound	44.28	69.83	0.7531

Table 4.8: Performance of OPUS-FT models fine-tuned with fully synthetic parallel corpora obtained by back-translation.

Table 4.8 presents results of fine-tuning the OPUS-MT model with fully synthetic data obtained by back-translation. We can immediately notice that the results of adapting this model are very different to what we observed when evaluating ACT. The worst performing baseline, e-commerce, is the only one to obtain significantly better results after fine-tuning the model with synthetic data. All metric further agree that adapting the model using back-translated e-commerce data produced by AT is significantly better than using data produced by the OPUS-MT baseline model. A possible explanation for this is the fact that, as will be discusses further at a later point, OPUS-MT model produces noisier translations, to which an NMT model should be more susceptible. As for the other two domains, when looking at BLEU and chrF2 evaluation, we can notice that the adapted models perform similarly to the baseline, and in one case, even significantly worse than the baseline, as is the case when using back-translations obtained using AT model to translate data from the pharmaceutical domain. The AT model did obtain much worse results when translating the pharmaceutical data test set in the nl→en direction as compared to the OPUS-MT model, which might explain this difference. Remember also that in the case of BLEU and chrF2, models adapted using natural data obtained significantly better results in those two domains as compared to the baseline, while COMET scores showed no or insignificant improvements. After adaptation with purely synthetic data, on the other hand, COMET scores show significantly worse results. Here, the fact that NMT models have a tendency to overfit to noisy synthetic data, as shown by Burlot and Yvon (2018) and Fadaee and Monz (2018), might be at play.

4.4.2 Forward translation

In this experiment, the usefulness of forward translation and mixing equal parts of forward and back-translated data to obtain the synthetic corpora was evaluated. To obtain forward translated corpora, all source side data in English was translated to Dutch using the *off-the-shelf* en→nl AT model. For the combination of forward and back-translation, half of the source original corpora was randomly sampled and then translated into target, while the other half of target natural corpora was back-translated into source, a procedure also used by Burlot and Yvon (2018) and Park et al. (2017).

Since I did not expect to obtain such good results as with back-translation, I only executed this experiment in one domain, e-commerce, and only using AT as the translation engine to obtain the translations used when building parallel corpora. I chose the e-commerce domain since the baseline performance obtained when translating data from this domain was the lowest, thus I expected that any increases would be the easiest to obtain in this domain. Also, since this domain has the least quantity of parallel data available, those experiments, from which I didn't expect to gain much, were the quickest and least expensive to run. The results are reported in Table 4.9 for ACT, and in Table 4.10 for OPUS-FT models.

domain & model	BLEU	chrF2	COMET
E-Comm			
AT baseline	41.11	68.28	0.7464
ACT ATft	40.91*	68.14*	0.7454
ACT ATbt&ft	41.79*	68.57*	0.7484
ACT ATbt	**42.62***	**69.05***	**0.7538***
ACT upper bound	46.34	71.05	0.7701

Table 4.9: Performance of ACT using synthetic parallel corpora obtained by FT and combining FT and BT.

Performance of ACT using synthetic corpora obtained by forward translation (FT) and mixing forward and back-translated (BT) data seems to confirm the expectations. When adapting translations using forward translated data, the performance is actually (in terms of BLEU and chrF2, significantly) worse than that of the baseline model. I expect this is because of the noise introduced by the erroneous translations of the baseline model. Combining forward and back-translated data results in (again, in terms of BLEU and chrF2, significantly) better results than when using the baseline model, but the results are still not as good as when using just back-translated data.

domain & model	BLEU	chrF2	COMET
E-Comm			
OPUS baseline	39.49	66.72	0.7105
OPUS-FT ATft	40.14*	67.63*	0.7349*
OPUS-FT ATbt&ft	41.44*	68.39*	**0.7397***
OPUS-FT ATbt	**42.17***	**68.81***	0.7346*
OPUS-FT upper bound	44.28	69.83	0.7531

Table 4.10: Performance of OPUS-FT models fine-tuned with synthetic parallel corpora obtained by FT and combining FT and BT.

Performance of OPUS-FT models in terms of BLEU and chrF2 scores actually is even more in tune with what we would expect. Here, fine-tuning the model using forward translated data also results in significant gains, although less pronounced than when adding back-translated data, which is exactly what we would expect based on what is reported in the literature. COMET scores, on the other hand, paint an unexpected picture and imply that models adapted with forward translated data perform at least as good as when using back-translation (the results are actually higher for both forward translation and the best performing model adapted with a combination of forward and back-translated data, although the differences are not significant).

4.4.3 Choosing the best translations

This experiment was designed in order to check if cleaning the synthetic data differently or using a different decoding method to obtain translations could result in better performance of models adapted with those synthetic parallel corpora.

Inspired by Xu et al. (2019), who used bilingual word embeddings and cosine similarity between sentence vectors as a measure of translation quality, LASER embeddings were used to encode each source sentence (obtained by back-translation) and target sentence (natural) in turn, and then the cosine similarity between the vectors was calculated. Then, all sentence pairs where cosine similarity is lower than 0.9 were filtered out. This cut-off was selected because it was also used when selecting sentences that comprise the natural parallel corpora provided for the experiment. Thus, it should ensure that synthetic data obtained is of very similar translations quality as the natural data. Only back-translated sentences produced by OPUS-MT were used in this experiment because, considering that those translations are more noisy than translations obtained using AT, I expected higher gains could be obtained from cleaning the data.

As for exploring the utility of using a different decoding method, following Edunov et al. (2018), sampling was used instead of greedy decoding when generating back-translations using the baseline OPUS-MT model.

Results are shown in Table 4.11 for ACT, and Table 4.12 for OPUS-MT model fine-tuned with synthetic parallel data. The tables also feature results of adaptation with synthetic data obtained using greedy decoding, that was evaluated in the first experiment.

	BLEU	chrF2	COMET
Fin			
AT baseline	49.95	72.80	0.8092
ACT OPUSbt greedy	**52.27***	**74.11***	0.8133*
ACT OPUSbt greedy LASER	52.24*	74.08*	**0.8162***
ACT OPUSbt sampling	51.94*	73.89*	0.8119
ACT upper bound	55.05	75.61	0.8245
Pharma			
AT baseline	45.72	71.37	0.8152
ACT OPUSbt greedy	48.41*	**72.79***	**0.8229***
ACT OPUSbt greedy LASER	48.22*	72.63*	0.8214*
ACT OPUSbt sampling	**48.43***	**72.79***	0.8227*
ACT upper bound	51.55	74.33	0.8342
E-Comm			
AT baseline	41.11	68.28	0.7464
ACT OPUSbt greedy	42.80*	**69.16***	0.7542*
ACT OPUSbt greedy LASER	42.67*	69.12*	0.7547*
ACT OPUSbt sampling	**42.81***	69.15*	**0.7551***
ACT upper bound	46.34	71.05	0.7701

Table 4.11: Performance of ACT using fully synthetic parallel corpora obtained by back-translation using the OPUS-MT model.

As we can see in the tables, there is hardly any difference obtained when cleaning the data with LASER embeddings or generating translations via sampling, as opposed to using greedy decoding and the default cleaning method. Different metrics also do not point into the same direction when it comes to which method could be considered better than the others. When it comes to ACT, all of them lead to significant improvements over the baselines. OPUS-MT models, on the other hand, only improve over the worst-performing e-commerce baseline. In the other two domains, results are *on par* with those obtained using the *off-the-shelf*, baseline model, or even significantly worse (when it comes to COMET evaluation).

Based on related work, I did not expect that the quality of translation will influence the performance of the models to a considerable extent, since it has been shown that using a reasonably good model for translation is usually enough (Fadaee and Monz, 2018). Still, I did expect some improvements would be obtained by filtering out translation of lesser quality when constructing synthetic parallel corpora, or using a different decoding method to generate translations. Still, experimental results seem to indicate that those procedures do not result in better performance when compared to using greedy decoding and the default cleaning method.

	BLEU	chrF2	COMET
Fin			
OPUS baseline	**51.85**	74.00	**0.8223**
OPUS-FT OPUSbt greedy	50.98	73.87	0.8006*
OPUS-FT OPUSbt greedy LASER	51.19	**74.01**	0.8018*
OPUS-FT OPUSbt sampling	51.10	73.96	0.8067*
OPUS-FT upper bound	54.36	75.13	0.8216
Pharma			
OPUS baseline	**48.61**	72.90	**0.8298**
OPUS-FT OPUSbt greedy	48.55	**72.95**	0.8134*
OPUS-FT OPUSbt greedy LASER	48.47	72.75	0.8109*
OPUS-FT OPUSbt sampling	48.53	72.81	0.8114*
OPUS-FT upper bound	50.33	73.81	0.8340
E-Comm			
OPUS baseline	39.49	66.72	0.7105
OPUS-FT OPUSbt greedy	40.68*	**68.01***	0.7207*
OPUS-FT OPUSbt greedy LASER	40.67*	67.93*	**0.7210***
OPUS-FT OPUSbt sampling	**40.91***	67.88*	0.7188
OPUS-FT upper bound	44.28	69.83	0.7531

Table 4.12: Performance of OPUS-FT models fine-tuned with fully synthetic parallel corpora obtained by back-translation using the OPUS-MT model.

4.4.4 Experimenting with the quantity of synthetic data

The last experiment seems to imply that, when it comes to synthetic data of sufficient quality, further tweaks meant to ensure translation are a bit better do not bring improvements in the overall performance of adapted models. What is expected to have a bigger impact, though, is the quantity of synthetic data used, and even more crucially, the ratio of synthetic to natural data (as shown by Fadaee and Monz (2018), Poncelas et al. (2019) and Burlot and Yvon (2018)).

In the previous three experiments, the models were adapted using completely synthetic data, either produced by back-translation (experiments 1 and 3) or by forward translation and combining forward and back-translation (experiment 2). This models a scenario in which there is no natural parallel data available in the domains of interest. A scenario one might expect to encounter more frequently in a "real-world" setting, though, is the one where some natural parallel data is available, and this natural data is augmented using synthetic parallel data. The remaining experiments model this scenario.

The present experiment is set up as follows. A certain percentage of the original natural parallel dataset is selected at random (1/2, 1/5 or 1/11, depending on the experiment and the domain). Then, the remaining target natural sentences are back-translated to source using the *off-the-shelf* AT nl→en model.

Tables 4.13 and 4.14 show results for adapting AT and OPUS-MT models to each of the three domains, each time taking a random half of original datasets as natural data, and adding the second half as back-translated parallel data. Since different datasets are of different sizes, each table in this subsection includes information on how many natural sentences were randomly selected for the natural data portion of the parallel corpora.

Since for the e-commerce domain, the least amount of parallel data was available, only this first experiment was run in that domain. For the other two domains, however, there was more data, and thus more room to experiment with different ratios of synthetic to parallel data in a meaningful way. For the pharmaceutical domain, I also randomly selected 1/5 of the natural data as the natural portion, and tried adding synthetic data in ratios of 1:1 and 1:4 to this data. For the financial domain, where the most data is available, I additionally experimented with selecting just 1/11 of natural data, and adding ratios as high as 1 part natural to 10 parts synthetic data. Results for ACT in those two domains are given in tables 4.15 and 4.16, while result for fine-tuned OPUS-FT model are given in tables 4.17 and 4.18.

Selecting just a part of the natural parallel corpora as natural data in this way had an additional benefit: each time the least amount of natural data was selected in a certain domain (1/2 for e-commerce, 1/5 for pharmaceutical, and 1/11 for financial), the resulting natural parallel corpus was of roughly the same size (around 15000 sentences). This also allows for evaluating the effect different sizes of natural corpora have on adapting models to the three domains.

domain & model	BLEU	chrF2	COMET
E-Comm	16175 natural sentences		
AT baseline	41.11	68.28	0.7464
ACT natural	44.12*	69.84*	0.7595*
ACT 1 natural : 1 synthetic	**44.99***	**70.31***	**0.7615***
ACT upper bound	46.34	71.05	0.7701
Pharma	39362 natural sentences		
AT baseline	45.72	71.37	0.8152
ACT natural	49.15*	73.16*	0.8253*
ACT 1 natural : 1 synthetic	**50.15***	**73.59***	**0.8303***
ACT upper bound	51.55	74.33	0.8342
Fin	79653 natural sentences		
AT baseline	49.95	72.80	0.8092
ACT natural	52.74*	74.34*	0.8149*
ACT 1 natural : 1 synthetic	**53.53***	**74.81***	**0.8192***
ACT upper bound	55.05	75.61	0.8245

Table 4.13: Performance of ACT adapted with 1 part synthetic to 1 part natural data.

As with the previous experiments, the biggest difference is in the performance of the two models, so we will look at each of them in turn.

When it comes to ACT, every time a translation is adapted as part of this experiment we obtain a statistically significant improvement over the unadapted baseline, except a few evaluations with COMET scores (the translations adapted to pharmaceutical domain using just 1/5 of the original natural data and a number of models from the financial domain adapted with 1/11 and 1/5 of original natural data and with synthetic data added to those portions of the original data). The best performance is obtained when using a half of the original dataset as the natural portion, and adding equal quantity of synthetic parallel data (Table 4.13). This is not surprising, since this is the scenario where we are using the biggest quantity of natural data, which results

domain & model	BLEU	chrF2	COMET
E-Comm	16175 natural sentences		
OPUS baseline	39.49	66.72	0.7105
OPUS-FT natural	43.18*	69.07*	0.7474*
OPUS-FT 1 natural : 1 synthetic	**43.85***	**69.67***	**0.7491***
OPUS-FT upper bound	44.28	69.83	0.7531
Pharma	39362 natural sentences		
OPUS baseline	48.61	72.90	0.8298
OPUS-FT natural	49.35*	73.29*	**0.8343**
OPUS-FT 1 natural : 1 synthetic	**49.94***	**73.64***	0.8327
OPUS-FT upper bound	50.33	73.81	0.8340
Fin	79653 natural sentences		
OPUS baseline	51.85	74.00	0.8223
OPUS-FT natural	53.46*	74.86*	0.8184
OPUS-FT 1 natural : 1 synthetic	**54.24***	**75.20***	**0.8259**
OPUS-FT upper bound	54.36	75.13	0.8216

Table 4.14: Performance of OPUS-FT models fine-tuned with 1 part synthetic to 1 part natural data.

in the highest performance of the adapted models.

When experimenting with adding different ratios of synthetic to natural data in the pharmaceutical (Table 4.15) and the financial (Table 4.16) domain, we observe that, each time we add additional synthetic data, we obtain additional improvement. The best performing models in those two domains also come quite close to the results of using natural data for adaptation (the difference in performance even becomes statistically insignificant when looking at the COMET score of the best performing model in the pharmaceutical domain). As far as it concerns ACT, it seems that adding more synthetic data is always helpful.

model	BLEU	chrF2	COMET
AT baseline	45.72	71.37	0.8152
ACT natural, 15745 sentences	47.48*	72.27*	0.8199
1 natural : 1 synthetic	47.83*	72.43*	0.8215*
1 natural : 4 synthetic	48.59*	72.85*	0.8220*
ACT natural, 39362 sentences	49.15*	73.16*	0.8253*
1 natural : 1 synthetic	**50.15***	**73.59***	**0.8303***
ACT upper bound	51.55	74.33	0.8342

Table 4.15: Performance of ACT in the pharmaceutical domain, adapted with different ratios of synthetic to natural data.

When it comes to the OPUS-MT model, though, a slightly different picture emerges. As for the best performing models—and these are again models adapted with a half of the original natural parallel corpus, while the other half is synthetic corpus obtained by back-translation, see Table 4.14—they perform almost as good as the models fine-tuned with natural data, and in some cases even better (although this difference is not statistically significant). Only the models in pharmaceutical and the ecommerce

model	BLEU	chrF2	COMET
AT baseline	49.95	72.80	0.8092
ACT natural, 14482 sentences	50.64*	73.14*	0.8077
1 natural : 1 synthetic	50.94*	73.32*	0.8088
1 natural : 4 synthetic	51.65*	73.78*	0.8127
1 natural : 10 synthetic	52.23*	74.11*	0.8132
ACT natural, 31861 sentences	51.34*	73.58*	0.8095
1 natural : 1 synthetic	51.80*	73.86*	0.8124
1 natural : 4 synthetic	52.71*	74.34*	0.8162*
ACT natural, 79653 sentences	52.74*	74.34*	0.8149*
1 natural : 1 synthetic	**53.53***	**74.81***	**0.8192***
ACT upper bound	55.05	75.61	0.8245

Table 4.16: Performance of ACT in the financial domain, adapted with different ratios of synthetic to natural data.

domain, and only when evaluated with BLEU score, perform significantly worse than models adapted with natural data.

More synthetic data, on the other hand, does not equal better performance as it did in the case of ACT. This is best illustrated by the performance of the model adapted to the financial domain (Table 4.18), where ratios as high as 1:10 natural to synthetic data were used. When 10 times as much synthetic data as natural data is used, the model seems to unlearn its parameters, and performs worse than the baseline. The model seems to overfit to noisy training data, with noise stemming from synthetic data. I experimented and found out that this phenomenon can be partially ameliorated by fine-tuning the model for one epoch less (thus for 2 instead of 3 epochs). Results that are lower than they would be after fine-tuning for 2 epochs are market with an exclamation mark in tables 4.17 and 4.18. Nevertheless, those models would still perform worse than models fine-tuned with less synthetic data. When fine-tuning OPUS-MT models, thus, there definitely seems to be a limit to learning from synthetic data, as postulated by Fadaee and Monz (2018).

model	BLEU	chrF2	COMET
OPUS baseline	48.61	72.90	0.8298
OPUS-FT natural, 15745 sentences	48.49	72.88	0.8286
1 natural : 1 synthetic	48.64	72.95	0.8233*
1 natural : 4 synthetic	48.67!	73.16	0.8216!*
OPUS-FT natural, 39362 sentences	49.35*	73.29*	**0.8343**
1 natural : 1 synthetic	**49.94***	**73.64***	0.8327
OPUS-FT upper bound	50.33	73.81	0.8340

Table 4.17: Performance of OPUS models, pharmaceutical domain, adapted with different ratios of synthetic to natural data.

In the scope of this experiment, natural and synthetic data that was used for each experiment was selected at random. In the following, last experiment, LASER embeddings will be used to select data that is more similar to the test set and thus more representative of the domain of interest.

model	BLEU	chrF2	COMET
OPUS baseline	51.85	74.00	0.8223
OPUS-FT natural, 14482 sentences	52.48	74.23	0.8164
1 natural : 1 synthetic	52.74*	74.50*	0.8156
1 natural : 4 synthetic	52.69*	74.54*	0.8113*
1 natural : 10 synthetic	45.46!	74.09!	0.8097!*
OPUS-FT natural, 31861 sentences	52.38	74.31	0.8174
1 natural : 1 synthetic	53.10*	74.67*	0.8136
1 natural : 4 synthetic	53.47*	74.90*	0.8189
OPUS-FT natural, 79653 sentences	53.46*	74.86*	0.8184
1 natural : 1 synthetic	**54.24***	**75.20***	**0.8259**
OPUS-FT upper bound	54.36	75.13	0.8216

Table 4.18: Performance of OPUS models, financial domain, adapted with different ratios of synthetic to natural data.

4.4.5 Selecting data more representative of the domain

The goal of this last experiment was to evaluate whether selecting adaptation data that is more representative of the domain would be beneficial when adapting models with synthetic data. Data selection has a long tradition among data-based methods for domain adaptation for machine translation, and some works propose that the same method be used when selecting synthetic data (see, for example, Poncelas and Way (2019), as mentioned in Chapter 2). To obtain data that is more representative of the domain, the source portion of the test set (English data) was used as the seed to retrieve relevant sentences that are more similar to the test set. Similarity was again calculated using LASER embeddings and cosine similarity as the similarity measure, this time comparing a candidate sentence from the training corpus to all the sentences from the test set.

The setup of this experiment builds on one of the experimental setups from the last subsection, where 1/5 of the natural data in the pharmaceutical and financial domains was selected, and the same quantity of synthetic data was added to the natural dataset. To compare the data selection method to those experiments, in this one, natural data was also selected based on similarity to the test set, with the goal of evaluating whether the method works when it comes to selecting natural data. Next, the same random portion of natural data was used as in the previous experiment, but synthetic data was selected based on similarity to the test set. This time, similarity was calculated between the synthetic, back-translated source portion of synthetic data, and the natural test set. Results are reported on in Table 4.19 for ACT, and 4.20 for OPUS-MT models.

When it comes to ACT, the experiments in the pharmaceutical domain seem to corroborate the hypothesis that data selection (DS) should help both when choosing natural and when selecting synthetic data, even though the differences between the performance of different models are slight. When adapting the model to the financial domain, though, natural data selection shows more significant increases in performance over selecting data at random, while selecting synthetic data more similar to the test set seems to work less well than random selection, which is a bit surprising (the differences are really minimal, though). We must note that these experiments are really not extensive enough to warrant any strong conclusions.

domain & model	BLEU	chrF2	COMET
Pharma	15745 natural sentences		
AT baseline	45.72	71.37	0.8152
ACT natural, random	47.48*	72.27*	0.8199
ACT natural, DS	47.72*	72.39*	0.8200*
ACT 1 natural, random : 1 synthetic, random	47.83*	72.43*	0.8215*
ACT 1 natural, random : 1 synthetic, DS	**48.00***	**72.56***	**0.8219***
ACT upper bound	51.55	74.33	0.8342
Fin	31861 natural sentences		
AT baseline	49.95	72.80	0.8092
ACT natural, random	51.34*	73.58*	0.8095
ACT natural, DS	**52.25***	**74.02***	**0.8137***
ACT 1 natural, random : 1 synthetic, random	51.80*	73.86*	0.8124
ACT 1 natural, random : 1 synthetic, DS	51.72*	73.78*	0.8117
ACT upper bound	55.05	75.61	0.8245

Table 4.19: Performance of ACT adapted with 1 part synthetic to 1 part natural data, where some data was selected at random, and some using LASER embeddings.

domain & model	BLEU	chrF2	COMET
Pharma	39362 natural sentences		
OPUS baseline	48.61	72.90	**0.8298**
OPUS-FT natural, random	48.49	72.88	0.8286
OPUS-FT natural, DS	48.00*	72.38*	0.8191*
OPUS-FT 1 natural, random : 1 synthetic, random	48.64	72.95	0.8233*
OPUS-FT 1 natural, random : 1 synthetic, DS	**48.84**	**72.98**	0.8223*
OPUS-FT upper bound	50.33	73.81	0.8340
Fin	79653 natural sentences		
OPUS baseline	51.85	74.00	**0.8223**
OPUS-FT natural, random	52.38	74.31	0.8174
OPUS-FT natural, DS	52.38	74.25	0.8094*
OPUS-FT 1 natural, random : 1 synthetic, random	**53.10***	**74.67***	0.8136
OPUS-FT 1 natural, random : 1 synthetic, DS	53.07*	74.74*	0.8153
OPUS-FT upper bound	54.36	75.13	0.8216

Table 4.20: Performance of OPUS-FT models fine-tuned with 1 part synthetic to 1 part natural data, where some data was selected at random, and some using LASER embeddings.

Fine-tuning OPUS-MT models with data selected based on similarity to the test set, on the other hand, seems to hurt models' performance, even when it comes to natural data selection (again, results are pretty close, not warranting any strong conclusions). We hypothesize that OPUS-MT models were not as good suited to evaluate this experiment because of the differences in the adaptation method. When it comes to these models, namely, as will be further discussed later, adaptation data is shuffled at the start of each epoch, and quite different results can be obtained by changing the random seed that is used in the process. Thus, results obtained in this experiment, that are very close, can not really tell us much.

4.4.6 Summary

In this section, the five experiments design to evaluate the usefulness of translation-based methods for generating synthetic data for domain adaptation were presented. The first three experiments modelled the scenario where no in-domain parallel data is available in the language pair and domain of interest, thus a fully synthetic parallel corpus was used. Fully synthetic corpora generated by back-translating target side data proved to be the most useful, and it seemed that data obtained by the default method of translation generation and cleaned with the default cleaning method could not be improved by using a different decoding algorithm or a different cleaning approach. The last two experiments modelled a scenario where synthetic data is used to augment natural parallel corpora. The experimental results indicated that when using a ratio of 1 part natural to 1 part synthetic data to construct a parallel corpus, similar results could be obtained as when using a natural corpus for adaptation. This was also the best performing method of using synthetic data for adaptation.

Furthermore, it was shown that the performance of the two models (ACT and OPUS-FT) differs significantly when it comes to the quantity of synthetic data used. While when it came to ACT, there seemed to be no limit to learning from synthetic data, and each experiment obtained improvements over the baseline, OPUS-FT models were quicker to overfit to noisy synthetic data, using which turned out to be harmful to model performance in more than one scenario. Additionally, the choice of an automatic metric used for evaluation turned out to be crucial, since the string-based metrics BLUE and chrF2 did not agree with the COMET score in many instances.

In the next, final chapter, we will discuss those issues in more detail, but first, in the remainder of this one, a brief qualitative analysis will be presented, with the goal of allowing us to go past evaluation using automatic metrics.

4.5 Qualitative Analysis

Recent research on automatic evaluation was mentioned when corroborating the choice of using the COMET metric for evaluation (Kocmi et al., 2021; Mathur et al., 2020). But, all that research also suggested that using human evaluation should be the best way to evaluate machine translation results. Still, in a project like this one, where so many systems are build and need to be evaluated, it is pretty hard to actually design a framework that makes sense. Since I do feel that looking at the actual data is very important, and not just relying on automatic metrics, I wanted to at least attempt something resembling qualitative analysis. After consideration, I decided to select a number of sentences from the test set, and show what kinds of outputs were produced by different systems evaluated in the scope of this research. Inspired in part by the work of Koot (2022), who calculated segment BLEU score differences in order to analyse which sentences gained the most from domain adaptation, I decided to select my sentences based on COMET score differences between sentences translated by *off-the-shelf* AT model, and ACT customized with natural data from the e-commerce domain. I chose this domain since all the experiments, including the ones using forward translation, have been conducted with e-commerce data. When presenting the datasets in Chapter 3, I mentioned using LASER embeddings to sort test sentences in order to obtain the ones most relevant for the domain. Now, I went through them one by one, and looked at the difference in COMET scores obtained by AT and ACT translations. I selected the ones where translations adapted with natural data were at least 0.2 COMET scores better than the baseline. Those were selected because, since there was a substantial difference in performance between the unadapted and adapted models, I expected there to be enough “room for improvement” with regard to adaptation. Hopefully, when looking at those sentences, we will be able to observe meaningful differences between the various models.

All sentences selected for the analysis, as well as the outputs of the different systems, are presented in the Appendix. I selected just five sentences in order to make the analysis manageable. After selecting the source sentences, it was just a matter of retrieving translations produced by each of the systems selected for analysis. In addition to the outputs of the baseline and upper-bound models, I retrieved translations produced by 10 different systems using synthetic data for adaptation: models adapted with completely synthetic data obtained by back-translation, using both the AT and the *off-the-shelf* OPUS-MT model to obtain synthetic sentences, then the models adapted using forward translated data and a mix of forward and back-translation, and lastly the models adapted using 1 part natural to 1 part synthetic data. In addition to presenting the translations, I also calculated the COMET score for each segment, taking into account both the source sentence and the target reference. This was done in the hopes of getting a bit of a feeling for what different COMET scores mean in terms of different translations obtained by various systems.

In the case of the first sentence (A.1), the biggest difference in translation is in the use of the more formal *dienen te worden gebruikt* instead of *moeten worden gebruikt* as a translation of *are to be used* and *zoals beschreven* instead of *zoals aanbevolen* for *as recommended.* Both systems adapted with natural data reflect this wording, that is also present in the reference, and so do the models adapted with 1 part natural to 1 part synthetic data, thus the best performing models in our experiments. Some models adapted with purely synthetic back-translated data, such as ACT ATbt and ACT OPUSbt, use *dienen te worden gebruikt* and *zoals aanbevolen*, thus showing a performance that can be thought of as the middle ground between the best and the worst performing models. Additionally, the baseline AT models and some systems adapted with forward translated data use untranslated *cartridges* instead of *patronen.* (For this example and the following ones, all the pertinent differences that we mention here are marked in bold in the Appendix.)

When looking at the second example (A.2), models adapted with natural data use the term *verdovend* for *incapacitating*, and *chemische stoffen* for *chemical agents.* Unadapted models, on the other hand, use a more literal translation of *chemische agentia* for *chemical agents*, while imprecise terms like *niet-bekwaam* (*uncompetent*), *schadelijk* (*harmful*) or *ontplofbaar* (*explosive*) are used instead of the very precise *verdovend.* As for the systems adapted using synthetic data, about half of them use *chemische stoffen*, while only one (OPUS-FT 1 natural : 1 synthetic) uses the adjective *verdovend* for *incapacitating.*

As for our third example (A.3), the best translation in terms of the COMET score is actually obtained by the unadapted OPUS-MT model that uses *gerecycled materiaal* as a translation for *post-consumer recycled materials*, which is closest to the reference translation (*gerecyclede materialen*). Other systems use incorrect terms such as *gerecyclede materialen uit de consumentenkringloop* and literal translation such as *gerecyclede materialen na de consument.* A more correct term would be *gerecyclede materialen na consumptie*, which is used by the OPUS model fine-tuned with natural data and with 1 part natural and 1 part synthetic data. Other systems fine-tuned with synthetic data do not obtain meaningful improvements over the baselines.

The fourth sentence selected for analysis (A.4), features *Cyan, Magenta and Yellow cartridges*, which are not translated at all, or the difference is only minimal (using *Cyaan* instead of *Cyan*), in case of the unadapted models. The reference features *cyaan, magenta en gele cartridges*, which is echoed by models adapted using natural data, and also some models adapted with synthetic data, most notably again, both models adapted with a mix of natural and synthetic data, which in this case produced translations of the same quality as models adapted with natural data.

When it comes to our last, fifth example (A.5), all the systems obtain results that are very close to the reference, except for systems adapted using forward translated data or a combination of forward and back-translation. The best performing systems use the term *patroon* instead of untranslated *cartridge*, and a more precise *klontjes* instead of *klonten* as a translation of *clumps.*

In conclusion, this limited qualitative analysis seems to confirm the conclusions we got from automatic evaluation metrics, that the best performance can be expected by models adapted with natural data, followed closely by models adapted using a ratio of 1 part natural to 1 part synthetic data obtained by back-translation. Models using forward translation and a mix of forward and back-translation seem to perform the worst.

Chapter 5

Discussion and Conclusion

5.1 Answering the Research Questions

At the start of this thesis, I posed the following research questions:

Q1. Can synthetic data be useful for domain adaptation in the context of TAUS DEMT?

Q2. Which method of generating synthetic data is the most useful?

To answer the first question, I will first have to rephrase it slightly. As has been shown in the last chapter, the results obtained in the experiments primarily differed in respect to two criteria: the model that was adapted, and the metric that was used to evaluate the results. As at the start of writing this thesis, TAUS used Amazon's Active Custom Translation for their Data-Enhanced Machine Translation service, and BLEU scores to evaluate its performance, the answer to the first research question, if posed for ACT and BLUE as a method of evaluation, would have to be overwhelmingly positive. Of all the experiments, the only scenario in which we did not obtain BLEU score improvements when using synthetic data to customize the *off-the-shelf* AT model was using only forward translated data for adaptation. Moreover, the experiment that evaluated how different quantities of synthetic data influence the performance of adapted models seemed to imply that adding more synthetic data was always helpful. Even using 10 times more synthetic than natural data resulted in gains in ACT translation performance, which is something that was not expected based on related work.

Another question that can quickly be pondered, even though it was not posed as a research question, is: should synthetic data be used for domain adaptation? Well, as we have seen, using synthetic data never performed as good as using natural data. Thus, a recommendation would be to always prioritize collecting natural data for adaptation, if at all feasible. Still, it is important to note that, in the experiments executed in this thesis, a high-resource scenario was evaluated, namely adapting English to Dutch (both high-resource languages) models in three pretty general domains. The baseline models already obtained very high results, and we were still able to obtain significant improvements by using only back-translated synthetic parallel corpora for adaptation (modelling a low-resource scenario, where no parallel in-domain data is available), and even better improvements when synthetic data was used to augment natural data. In a truly low-resource scenario, where the baseline models would not perform as well, I would expect even better results might be obtained by using synthetic data.

As for the second research question, it can be answered for both models and substantiated by results obtained by all three automatic metrics: the best results were obtained by using synthetic parallel corpora generated by back-translation as augmentation to natural parallel corpora. In some cases of fine-tuning the open-source OPUS-MT model, the results obtained were even better than adaptation with natural data (although the difference was not statistically significant). As for the quantities of natural and synthetic data that should be used, this should probably be evaluated separately for each model, and possibly also for each domain of interest, since experimental results varied quite considerably, especially when it came to evaluating OPUS-FT models and contrasting BLEU and chrF22 scores on the one hand, and COMET evaluation on the other.

To turn to those score differences: the experiments have demonstrated that it is critically important to select an appropriate automatic metric for evaluation. Luckily, the evaluation of the performance of different automatic metrics, especially when it comes to model development and comparing performance of different NMT models, has been at the forefront of many excellent recent research papers, and it does seem that the field will be able to move past evaluation using the BLEU score, which has dominated it since its inception 20 years ago. Even though in the scope of this thesis, the string-based chrF22 has been used as a second metric, following research that recommended its use instead of the BLUE score, most experimental results where the metrics did not agree have shown similar BLEU and chrF2 score differences between different systems, with the pretrained neutral COMET metric painting a dissimilar picture. This was especially noticeable when it came to OPUS-FT models. Not only did COMET scores indicate that using forward translated data and a combination of forward and back-translations could be much more useful than I expected, they also implied that, with baseline models of sufficiently high performance, using synthetic data for domain adaptation was possibly harmful and definitely not useful. In financial and pharmaceutical domains, COMET scores obtained using the baseline OPUS-MT model were already very similar to those obtained using ACT customized with natural parallel data, thus the upper bound for AT. Those scores did not improve when models were adapted with natural data, and, perhaps not surprisingly, seeing that using natural data for adaptation did not increase the performance of fine-tuned models, using synthetic data harmed those already high-performing baselines. On the other hand, when fine-tuned to the e-commerce domain, the OPUS-MT model, which when used *off-the-shelf* resulted in the lowest baseline, obtained the biggest improvement, with adaptation with synthetic data helpful not only when it came to back-translation, but also forward translation. This seems to indicate that there might be a certain threshold of model performance, where the models already perform very well on data from a certain domain, and cannot be adapted any further. In the financial and pharmaceutical domains, those COMET scores were above 0.8, which should indicate a very well performing SOTA model. On the other hand, results obtained in the e-commerce domain substantiate the expectation that in a truly low-resource scenario, synthetic data should be even more useful for domain adaptation.

5.2 Discussion and Limitations

The first and biggest limitation of my research that I would like to address in this section is the fact that, when using back-translation to generate synthetic corpora, a method used in all experiments, the target side Dutch sentences that were translated to source (English) were actually themselves translations. This is not a setup we expect to encounter in a “real-world” scenario, which is ultimately the scenario that we care about. The reason experiments were set up in this way was because it was deemed important not only to compare adaptation with synthetic data to an unadapted baseline, but also to an upper bound system obtained by adapting a model using natural parallel corpora, to get a sense of where on a scale from no adaptation to adaptation with natural data, adaptation using synthetic data would lie. The impacts of this decision are two-fold. First, were we to use real target side monolingual data to obtain a synthetic parallel corpus by back-translation, we expect that (back-)translations obtained using the *off-the-shelf* model would be of lesser quality than those we obtained in our experiments. This is because, as we already mentioned, translationese sentences are easier to translate than original data (Edunov et al., 2020). Secondly, we expect that using real monolingual target side data would impact our automatic evaluation to a greater extent than using translationese data did, as discussed by Edunov et al. (2020). Thus, we effectively evaluated a scenario that we will not encounter in practice, which is definitely a limiting factor. I could add to that something that I already discussed briefly in the last chapter, and that is the fact that we evaluated the method in a high-resource scenario, while we will probably use it in lower-resource scenarios where sufficient quantity of natural parallel data for adaptation is unattainable. Still, this is not such a limiting factor, since we just expect results at least as good as we obtained in the high-resource scenario. The performance of models used for back-translation in lower-resource scenarios, though, might prove problematic, since we won’t be able to obtain synthetic data of equal quality.

An additional limitation can be found when we look at the test sets that were used. Usually in the machine translation field, the standard is to use translations produced by human translators as references. Ideally, not even a single reference, but multiple references would be used when calculating metrics such as BLEU score, even though this is seldom encountered in practice. Our test sets, however, were picked at random from parallel training datasets that were provided by TAUS. These datasets have been cleaned and filtered using automatic metrics such as LASER embeddings similarities, and we do expect they were of sufficient quality, but the additional layer of human evaluation was still missing. Evaluating the quality of test sets was unfortunately out of scope in this research, but in future work, it would be recommended to use test sets that went through the additional step of vetting by human translators.

The last limitation I would like to discuss is the fact that the research presented in this thesis is not expected to be replicable. Of course, the corpora used cannot be shared publicly, because they are a property of TAUS, but the problem of replicability actually runs much deeper. For one, the first model that was used and evaluated, AT, is a proprietary model. For all I know, this model and/or the customization method used by ATC could have changed already, and even using the same data, the results obtained today might be very different from what was reported in this thesis. Even though the other model I used, OPUS-MT, is open-source, and was picked as such precisely because of reproducibility considerations, obtaining the same results might still be challenging. This stems from the inherent non-determinism of some of the processes that were run in the course of fine-tuning this model. Even though all precautions were taken, such as using different types of seeds to control non-deterministic behaviour, complete replicability still cannot be guaranteed across different platforms and even different

GPUs.[1] Note that, even though this research is probably not replicable, meaning that even if one used the same proprietary data, results obtained today might differ from what was reported in this thesis, I do expect that the conclusions drawn after reproducing the experiments would be very similar. Additionally, since we mentioned random seeds that were used to try and ensure reproducibility, I can briefly discuss another challenge that I encountered, and that influenced results obtained by OPUS-MT models. Namely, since training data was shuffled differently before feeding it to the model at the start of every fine-tuning epoch, when experimenting with different seeds, I sometimes obtained results that were quite different. Definitely the results that implicate that a certain OPUS-MT fine-tuned model is better than another, certainly if differences are not statistically significant, need to be taken with some reservation, since were I to use a different seed, those results could probably suggest a different ordering of the systems.

5.3 Conclusion and Future Work

This thesis presented an evaluation of the utility of using synthetic parallel corpora produced by translation for domain adaptation in Neural Machine Translation. In the first part of the thesis, I situated the topic by discussing successes and challenges of deep learning approaches to machine translation, as well as presented related work from the areas of domain adaptation for NMT and synthetic data generation for NLP. Then, the question this thesis tries to answer—could TAUS, the company that the research was performed at, successfully use synthetic data in their domain adaptation pipeline—was narrowed down to a method that was deemed as probably the most useful: back-translation. A series of experiments was designed, testing adaptation with synthetic data using two models: Amazon Translate, a model currently used by TAUS in their DEMT pipeline, as well as an open-source model, OPUS-MT. A big theme of this thesis was also automatic evaluation for machine translation, since I needed a reliable evaluation method in order to conclude which of the models performs the best. Luckily, the field of MT has seen a lot of high quality research into automatic metrics in recent years that I was able to benefit from. Following the recommendations by Kocmi et al. (2021), I added the neural, pretrained metric COMET to the more traditionally used string-based BLEU and chrF2 metrics. Even though it was at times confusing to try and analyse what the results in terms of different metrics could mean, as there were many experiments where BLEU and chrF2 results told one story, and COMET another, incorporating COMET into the evaluation arsenal was an important decision, since research has repeatedly shown that it correlates better with human judgment and especially when it comes to high-performing systems, which the systems evaluated in the scope of this thesis were.

As for the experimental results, they have unquestionably shown that there is potential for using synthetic data in domain adaptation in the scope of TAUS's DEMT, even in the context of high-performing baseline models. Since results were compared not only to the unadapted baselines, but also to the upper bound of adaptation with natural data, the experimental results also urge us to stress that there is no way around using natural data, since models adapted with it tended to show the best performance in all circumstances. Still, adaptation with synthetic data showed considerable poten-

[1] For more information, as well as procedures that control sources of randomness, see `https://pytorch.org/docs/stable/notes/randomness.html`.

tial, especially in the scenario of data augmentation, where synthetic data was added to natural parallel corpora. Customizing the proprietary AT model using synthetic data always showed at least some improvement over the baseline, and there seemed to be no point where synthetic data would start to be harmful, leading to catastrophic forgetting, since using as much as 10 times more synthetic than natural data proved to be useful. Fine-tuning OPUS-MT models with synthetic data, on the other hand, proved to be more nuanced. While in the two of the three domains that were evaluated, and where unadapted models already performed *on par* with customized AT models, improvements were not obtained even when using natural data for adaptation, in the third domain, where the unadapted model performed the worst of the six baselines, we were actually able to obtain significant improvements over the baseline using fully synthetic parallel corpora.

When it comes to recommendations for future work, firstly, the method for adapting models using synthetic data presented in this thesis needs to be evaluated in the "real-world" scenario, where target side data that will be back-translated to source to construct a synthetic parallel corpus will be natural data, and not translations. I was already able to secure some data in a domain not explored in this thesis, but unfortunately, do to time constraints, I was unable to run an additional experiment in the scope of this work.

Secondly, there are promising methods that refine back-translation even more, including iterative back-translation, where models used to generate back-translations are adapted in steps, which is especially relevant when it comes to low-resource scenarios (Wei et al., 2020). Additionally, one method that I did explore, but very briefly and quite superficially, is selecting synthetic data that is more in-domain or closer to the test set. I only evaluated the method of using LASER embeddings to compute semantic similarity, and calculated it exclusively on the source side of the parallel corpora, but there are other methods that proved promising in previous research that I would like to explore further, such as using transductive data selection algorithms (Poncelas et al., 2019; Poncelas and Way, 2019). Furthermore, this thesis concentrated only on evaluating synthetic data generated by translation, but there are many other promising methods for SDG in the context of NMT, for example, targeting domain-specific words that are the hardest to translate (Fadaee et al., 2017).

Lastly, and connected to all the points above, I would like to evaluate this experimental setup in a true low-resource scenario, for which the methods of synthetic data generation are primarily designed.

Appendix A

Examples selected for qualitative analysis

Below, the five examples from the e-commerce domain selected for qualitative evaluation can be found, as described in the Qualitative Analysis section. For each example, the source (S) and reference (R) sentences from the test set are listed, followed by translations obtained by the baseline and upper-bound systems, and then each of the models selected for qualitative evaluation. Next to each translation, the COMET score this translation obtained is provided.

A.1

S
Insulin pen ***The cartridges are to be used*** *in conjunction with an insulin pen such as OptiPen and other pens suitable for Insulin Human Winthrop cartridges and* ***as recommended*** *in the information provided by the device manufacturer.*

R
Insulinepen ***De patronen dienen te worden gebruikt*** *met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen, en* ***zoals beschreven*** *in de informatie verstrekt door de fabrikant van de pen.*

AT
Insulinepen ***De cartridges moeten worden gebruikt*** *in combinatie met een insulinepen zoals OptiPen en andere pennen die geschikt zijn voor Insulin Human Winthrop cartridges en* ***zoals aanbevolen*** *in de informatie die door de fabrikant van het apparaat wordt verstrekt.* 0.6783

OPUS-MT
Insulinepen ***De patronen moeten gebruikt worden*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals aanbevolen*** *in de informatie verstrekt door de fabrikant van het hulpmiddel.* 0.7433

ACT natural
Insulinepen ***De patronen dienen te worden gebruikt*** *met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals beschreven*** *in de informatie verstrekt door de fabrikant van de pen.* 0.9290

OPUS-FT natural
Insulinepen ***De patronen dienen te worden gebruikt*** *met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals beschreven*** *in de informatie verstrekt door de fabrikant van het hulpmiddel.* 0.8044

ACT ATbt
Insulinepen ***De patronen dienen te worden gebruikt*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals aanbevolen*** *in de informatie verstrekt door de fabrikant van het apparaat.* 0.7593

ACT OPUSbt
Insulinepen ***De patronen dienen te worden gebruikt*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals aanbevolen*** *in de informatie verstrekt door de fabrikant van het apparaat.* 0.7593

OPUS-FT ATbt
Insulinepen ***De patronen moeten worden gebruikt*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals aanbevolen*** *in de informatie verstrekt door de fabrikant van het hulpmiddel.* 0.7418

OPUS-FT OPUSbt
Insulinepen ***De patronen dienen te worden gebruikt*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en zoals* ***aanbevolen*** *in de informatie verstrekt door de fabrikant van het hulpmiddel.* 0.7494

ACT ATft
Insulinepen ***De cartridges moeten worden gebruikt*** *in combinatie met een insulinepen zoals OptiPen en andere pennen die geschikt zijn voor Insulin Human Winthrop-cartridges en* ***zoals aanbevolen*** *in de informatie die door de fabrikant van het apparaat wordt verstrekt.* 0.667

ACT ATbt&ft
Insulinepen ***De cartridges moeten worden gebruikt*** *in combinatie met een insulinepen zoals OptiPen andere pennen die geschikt zijn voor Insulin Human Winthrop-cartridges en* ***zoals aanbevolen*** *in de door de fabrikant van het apparaat verstrekte informatie.* 0.615

OPUS-FT ATft
Insulinepen ***De cartridges moeten worden gebruikt*** *in combinatie met een insulinepen zoals OptiPen en andere pennen die geschikt zijn voor Insulin Human Winthrop-cartridges en* ***zoals aanbevolen*** *in de door de fabrikant van het apparaat verstrekte informatie.* 0.6649

OPUS-FT ATbt&ft
Insulinepen ***De patronen moeten worden gebruikt*** *in combinatie met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals aanbevolen*** *in de informatie verstrekt door de fabrikant van het apparaat.* 0.7540

ACT 1 natural : 1 synthetic
Insulinepen ***De patronen dienen te worden gebruikt*** *met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals beschreven*** *in de informatie verstrekt door de fabrikant van de pen.* 0.9289

OPUS-FT 1 natural : 1 synthetic
Insulinepen ***De patronen dienen te worden gebruikt*** *met een insulinepen zoals de OptiPen en andere pennen geschikt voor Insulin Human Winthrop patronen en* ***zoals beschreven*** *in de informatie verstrekt door de fabrikant van het hulpmiddel.* 0.8043

A.2

S
Fixed equipment for the dissemination of ***incapacitating*** *or irritating chemical substances, which can be attached to a wall or to a ceiling inside a building, comprises a canister of irritating or* ***incapacitating chemical agents*** *and is activated using a remote control system*

R
Vaste apparatuur voor de verspreiding van ***verdovende*** *of irriterende chemische stoffen die kan worden vastgemaakt aan een muur of een plafond in een gebouw, die een bus bevat met irriterende of* ***verdovende chemische stoffen*** *en die met afstandsbediening wordt geactiveerd*

AT
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.4603

OPUS-MT
Vaste apparatuur voor de verspreiding van ***schadelijke*** *of irriterende chemische stoffen, die aan een wand of aan een plafond binnen een gebouw kan worden bevestigd, bestaat uit een busje van irriterende of* ***arbeidsongeschikte chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.4995

ACT natural
Vaste apparatuur voor de verspreiding van ***verdovende*** *of irriterende chemische stoffen, die aan een muur of plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische stoffen*** *en wordt geactiveerd door middel van een afstandsbedieningssysteem.* 0.6679

OPUS-FT natural
Vaste apparatuur voor de verspreiding van ***verdovende*** *of irriterende chemische stoffen, die aan een wand of aan een plafond in een gebouw kan worden bevestigd, bestaat uit een bus van irriterende of* ***verlammende chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.7297

ACT ATbt
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.4607

ACT OPUSbt
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem.* 0.5882

OPUS-FT ATbt
Vaste uitrusting voor de verspreiding van ***ontplofbare*** *of irriterende chemische stoffen, die aan een wand of aan een plafond binnen een gebouw kan worden bevestigd, bestaat uit een bus van irriterende of* ***ontplofbare chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.4039

OPUS-FT OPUSbt
Vaste apparatuur voor de verspreiding van ***ontplofbare*** *of irriterende chemische stoffen, die aan een wand of aan een plafond binnen een gebouw kan worden bevestigd, bestaat uit een bus van irriterende of* ***ontplofbare chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.5125

ACT ATft
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische stoffen*** *en wordt geactiveerd door middel van een afstandsbediening* 0.6071

ACT ATbt&ft
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.5848

OPUS-FT ATft
Vaste apparatuur voor de verspreiding van ***ongeschikte*** *of irriterende chemische stoffen, die aan een muur of aan een plafond in een gebouw kunnen worden bevestigd, bestaat uit een busje irriterende of* ***ontregelende chemische agentia*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.5065

OPUS-FT ATbt&ft
Vaste apparatuur voor de verspreiding van ***inactieve*** *of irriterende chemische stoffen, die aan een muur of aan een plafond binnen een gebouw kan worden bevestigd, bestaat uit een busje irriterende of* ***actieve chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.6438

ACT 1 natural : 1 synthetic
Vaste apparatuur voor de verspreiding van ***niet-bekwame*** *of irriterende chemische stoffen, die aan een muur of plafond in een gebouw kan worden bevestigd, bestaat uit een bus met irriterende of* ***niet-bekwame chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.5865

OPUS-FT 1 natural : 1 synthetic
Vaste apparatuur voor de verspreiding van ***verdovende*** *of irriterende chemische stoffen, die aan een wand of aan een plafond binnen een gebouw kan worden bevestigd, bestaat uit een bus van irriterende of* ***verstorende chemische stoffen*** *en wordt geactiveerd met behulp van een afstandsbedieningssysteem* 0.7148

A.3

S
This means arsenic-free glass, a mercury-free panel, halogen-free laminates in its circuit boards and a chassis that features more than 25% ***post-consumer recycled materials****.*

R
Dat betekent arsenicumvrij glas, een kwikvrij beeldscherm, halogeenvrije platen in de printplaten en een chassis dat voor meer dan 25% uit ***gerecyclede materialen*** *bestaat.*

AT
Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3097

OPUS-MT
Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaat in de printplaten en een chassis met meer dan 25% ***gerecycled materiaal****.* 0.8109

ACT natural

Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in printplaten en een chassis met meer dan 25% ***gerecyclede materialen uit de consumentenkringloop****.* 0.7564

OPUS-FT natural

Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na consumptie****.* 0.7309

ACT ATbt

Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

ACT OPUSbt

Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

OPUS-FT ATbt

Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in zijn printplaten en een chassis dat voorzien is van meer dan 25% ***postconsumer gerecyclede*** *materialen.* 0.6513

OPUS-FT OPUSbt

Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in zijn printplaten en een chassis dat voorzien is van meer dan 25% ***postconsumer gerecyclede materialen****.* 0.6513

ACT ATft

Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

ACT ATbt&ft

Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

OPUS-FT ATft

Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

OPUS-FT ATbt&ft
Dit betekent arseenvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na de consument****.* 0.3104

ACT 1 natural : 1 synthetic
Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in printplaten en een chassis met meer dan 25% ***gerecyclede materialen uit de consumentenkringloop****.* 0.7562

OPUS-FT 1 natural : 1 synthetic
Dit betekent arsenicumvrij glas, een kwikvrij paneel, halogeenvrije laminaten in de printplaten en een chassis met meer dan 25% ***gerecyclede materialen na consumptie****.* 0.7306

A.4

S
This cartridge, when used in conjunction with ***the Cyan, Magenta and Yellow cartridges****, can produce high resolution color printouts with sharp images and text.*

R
Wanneer deze cartridge in combinatie met ***de cyaan, magenta en gele cartridges*** *wordt gebruikt, kunt u in kleur en met hoge resolutie afdrukken waarbij tekst en afbeeldingen haarscherp worden weergegeven.*

AT
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de Cyaan, Magenta en Yellow cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7514

OPUS-MT
Deze cartridge, bij gebruik in combinatie met ***de Cyan, Magenta en Yellow cartridges****, kan hoge resolutie kleurenprints met scherpe afbeeldingen en tekst produceren.* 0.7512

ACT natural
Wanneer deze cartridge in combinatie met ***de cyaan, magenta en gele cartridges*** *wordt gebruikt, kunt u in kleur en met hoge resolutie afdrukken waarbij tekst en afbeeldingen haarscherp worden weergegeven.* 1.0235

OPUS-FT natural
Wanneer deze cartridge wordt gebruikt in combinatie met ***de cartridges cyaan, magenta en geel****, kan deze kleurafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7532

ACT ATbt
Deze cartridge kan, indien gebruikt in combinatie met ***de cyaan, magenta en gele cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7658

ACT OPUSbt
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de Cyaan, Magenta en Yellow cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7516

OPUS-FT ATbt
Deze cartridge, bij gebruik in combinatie met ***de Cyan-, Magenta- en Yellowcartridges****, kan kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7113

OPUS-FT OPUSbt
Deze cartridge, bij gebruik in combinatie met ***de Cyan-, Magenta- en Yellowcartridges****, kan kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7113

ACT ATft
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de cyaan, magenta en gele cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7657

ACT ATbt&ft
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de cyaan, magenta en gele cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7657

OPUS-FT ATft
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de cyaan-, magenta-en gele cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7662

OPUS-FT ATbt&ft
Deze cartridge kan, wanneer deze wordt gebruikt in combinatie met ***de cyaan-, magenta-en gele cartridges****, kleurenafdrukken met hoge resolutie produceren met scherpe afbeeldingen en tekst.* 0.7662

ACT 1 natural : 1 synthetic
Wanneer deze cartridge in combinatie met ***de cyaan, magenta en gele cartridges*** *wordt gebruikt, kunt u in kleur en met hoge resolutie afdrukken waarbij tekst en afbeeldingen haarscherp worden weergegeven.* 1.0234

OPUS-FT 1 natural : 1 synthetic
*Wanneer deze cartridge wordt gebruikt in combinatie met **de cartridges cyaan, magenta en geel**, kan deze kleurafdrukken met hoge resolutie met scherpe afbeeldingen en tekst produceren.* 0.756

A.5

S
*Do not use Humalog Mix50, if **clumps** of material are present or if solid white particles stick to the bottom or wall of **the cartridge**, giving it a frosted appearance.*

R
*Gebruik Humalog Mix50 niet wanneer er **klontjes** of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van **de patroon** plakken, en een bevroren indruk geven.*

AT
*Gebruik Humalog Mix50 niet als er **klonten** materiaal aanwezig zijn of als vaste witte deeltjes aan de onderkant of wand van **de cartridge** kleven, waardoor het een mat uiterlijk krijgt.* 0.6586

OPUS-MT
*Gebruik Humalog Mix50 niet als er **klontjes** of deeltjes aanwezig zijn of als er witte vaste deeltjes aan de bodem of wand van **de patroon** plakken, wat een bevroren indruk geeft.* 0.9133

ACT natural
*Gebruik Humalog Mix50 niet wanneer er **klontjes** of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van **de patroon** plakken, en een bevroren indruk geven.* 1.0232

OPUS-FT natural
*Gebruik Humalog Mix50 niet wanneer er **klontjes** of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van **de patroon** plakken, en een bevroren indruk geven.* 1.0236

ACT ATbt
*Gebruik Humalog Mix50 niet, als er **klontjes** materiaal aanwezig zijn of als er vaste witte deeltjes aan de bodem of wand van **de patroon** plakken, waardoor deze er mat uitziet.* 0.7220

ACT OPUSbt
*Gebruik Humalog Mix50 niet wanneer er **klontjes** materiaal aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van **de patroon** plakken, en een bevroren indruk geven.* 0.9023

OPUS-FT ATbt
Gebruik Humalog Mix50 niet, als er ***klontjes*** *materiaal aanwezig zijn of als er vaste witte deeltjes aan de bodem of wand van* ***de patroon*** *plakken, wat een bevroren uiterlijk geeft.* 0.8104

OPUS-FT OPUSbt
Gebruik Humalog Mix50 niet, wanneer er ***klontjes*** *of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van* ***de patroon*** *plakken, en een bevroren indruk geven.* 0.9736

ACT ATft
Gebruik Humalog Mix50 niet als er ***klonten*** *materiaal aanwezig zijn of als vaste witte deeltjes aan de onderkant of wand van* ***de cartridge*** *kleven, waardoor het een mat uiterlijk krijgt.* 0.6587

ACT ATbt&ft
Gebruik Humalog Mix50 niet als er ***klonten*** *materiaal aanwezig zijn of als vaste witte deeltjes aan de onderkant of wand van* ***de cartridge*** *kleven, waardoor het een mat uiterlijk krijgt.* 0.6587

OPUS-FT ATft
Gebruik Humalog Mix50 niet als er ***klonten*** *materiaal aanwezig zijn of als vaste witte deeltjes aan de onderkant of wand van* ***de cartridge*** *kleven, waardoor het een mat uiterlijk krijgt.* 0.6584

OPUS-FT ATbt&ft
Gebruik Humalog Mix50 niet als er ***klonten*** *materiaal aanwezig zijn of als vaste witte deeltjes aan de onderkant of wand van* ***de patroon*** *kleven, waardoor het een mat uiterlijk krijgt.* 0.5999

ACT 1 natural : 1 synthetic
Gebruik Humalog Mix50 niet wanneer er ***klontjes*** *of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van* ***de cartridge*** *plakken, en een bevroren indruk geven.* 0.8376

OPUS-FT 1 natural : 1 synthetic
Gebruik Humalog Mix50 niet wanneer er ***klontjes*** *of deeltjes aanwezig zijn of wanneer er witte vaste deeltjes aan de bodem of wand van* ***de patroon*** *plakken, en een bevroren indruk geven.* 1.0233

Bibliography

M. Artetxe and H. Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL `https://aclanthology.org/Q19-1038`. Place: Cambridge, MA Publisher: MIT Press.

S. Aslan. TAUS Launches Data-Enhanced Machine Translation - TAUS - The Language Data Network, 2022. URL `https://www.taus.net/resources/blog/taus-launches-data-enhanced-machine-translation`.

D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

J. Bastings. *A tale of two sequences: interpretable and linguistically-informed deep learning for natural language processing.* PhD thesis, Universiteit van Amsterdam, 2020. URL `https://eprints.illc.uva.nl/id/eprint/2178/1/DS-2020-09.text.pdf`. ISBN: 9789083091211 OCLC: 1199123834.

N. Bertoldi and M. Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09*, page 182, Athens, Greece, 2009. Association for Computational Linguistics. doi: 10.3115/1626431.1626468. URL `http://portal.acm.org/citation.cfm?doid=1626431.1626468`.

N. Bogoychev and R. Sennrich. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. 2019. doi: 10.48550/ARXIV.1911.03362. URL `https://arxiv.org/abs/1911.03362`.

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. doi: 10.18653/v1/W18-6315. URL `https://aclanthology.org/W18-6315`.

M. Chinea-Ríos, Á. Peris, and F. Casacuberta. Adapting Neural Machine Translation with Parallel Synthetic Data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4714. URL `https://aclanthology.org/W17-4714`.

K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL `https://aclanthology.org/W14-4012`.

C. Chu and R. Wang. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1111`.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

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.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL `https://aclanthology.org/D18-1045`.

S. Edunov, M. Ott, M. Ranzato, and M. Auli. On The Evaluation of Machine Translation Systems Trained With Back-Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.253. URL `https://aclanthology.org/2020.acl-main.253`.

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. doi: 10.18653/v1/D18-1040. URL `https://aclanthology.org/D18-1040`.

M. Fadaee, A. Bisazza, and C. Monz. Data Augmentation for Low-Resource Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, 2017. doi: 10.18653/v1/P17-2090. URL `http://arxiv.org/abs/1705.00440`. arXiv: 1705.00440.

S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.84. URL `https://aclanthology.org/2021.findings-acl.84`.

M. Freitag and Y. Al-Onaizan. Fast Domain Adaptation for Neural Machine Translation. *arXiv:1612.06897 [cs]*, Dec. 2016. URL `http://arxiv.org/abs/1612.06897`. arXiv: 1612.06897.

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. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/gehring17a.html`.

H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. Achieving Human Parity on Automatic Chinese to English News Translation, 2018. URL `https://arxiv.org/abs/1803.05567`.

K. Imamura, A. Fujita, and E. Sumita. Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2707. URL `https://aclanthology.org/W18-2707`.

T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, Nov. 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.57`.

P. Koehn. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-3250`.

P. Koehn. *Statistical Machine Translation.* Cambridge University Press, 2009. doi: 10.1017/CBO9780511815829.

P. Koehn. *Neural Machine Translation.* Cambridge University Press, 2020. doi: 10.1017/9781108608480.

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. doi: 10.18653/v1/W17-3204. URL `https://aclanthology.org/W17-3204`.

D. Koot. Understanding BLEU Scores in Customized Machine Translation - TAUS - The Language Data Network, 2022. URL `https://www.taus.net/resources/blog/understanding-bleu-scores-in-customized-machine-translation`.

G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv:1804.07755 [cs]*, Aug. 2018. URL `http://arxiv.org/abs/1804.07755`. arXiv: 1804.07755.

M.-T. Luong and C. Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam, Dec. 3-4 2015. URL `https://aclanthology.org/2015.iwslt-evaluation.11`.

N. Mathur, T. Baldwin, and T. Cohn. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL `https://aclanthology.org/2020.acl-main.448`.

O. Melamud and C. Shivade. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1905. URL `https://aclanthology.org/W19-1905`.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013. URL `https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf`.

M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence: Edited Review Papers Presented at the International NATO Symposium on Artificial and Human Intelligence*, 1984. URL `https://aclanthology.org/www.mt-archive.info/70/Nagao-1984.pdf`.

S. I. Nikolenko. *Synthetic Data for Deep Learning*, volume 174 of *Springer Optimization and Its Applications*. Springer International Publishing, 2021. ISBN 978-3-030-75177-7 978-3-030-75178-4. doi: 10.1007/978-3-030-75178-4. URL `https://link.springer.com/10.1007/978-3-030-75178-4`.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

J. Park, J. Song, and S. Yoon. Building a Neural Machine Translation System Using Only Synthetic Parallel Data. *arXiv:1704.00253 [cs]*, Sept. 2017. URL `http://arxiv.org/abs/1704.00253`. arXiv: 1704.00253.

A. Poncelas and A. Way. Selecting Artificially-Generated Sentences for Fine-Tuning Neural Machine Translation. *arXiv:1909.12016 [cs]*, Sept. 2019. URL `http://arxiv.org/abs/1909.12016`. arXiv: 1909.12016.

A. Poncelas, D. Shterionov, A. Way, G. M. d. B. Wenniger, and P. Passban. Investigating Backtranslation in Neural Machine Translation. *arXiv:1804.06189 [cs]*, Apr. 2018. URL `http://arxiv.org/abs/1804.06189`. arXiv: 1804.06189 version: 1.

A. Poncelas, G. M. d. B. Wenniger, and A. Way. Adaptation of Machine Translation Models with Back-translated Data using Transductive Data Selection Methods. *arXiv:1906.07808 [cs]*, June 2019. URL `http://arxiv.org/abs/1906.07808`. arXiv: 1906.07808.

M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `http://aclweb.org/anthology/W15-3049`.

M. Post. A Call for Clarity in Reporting BLEU Scores. *arXiv:1804.08771 [cs]*, Sept. 2018. URL `http://arxiv.org/abs/1804.08771`. arXiv: 1804.08771.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A Neural Framework for MT Evaluation. *arXiv:2009.09025 [cs]*, Oct. 2020. URL `http://arxiv.org/abs/2009.09025`. arXiv: 2009.09025.

R. Rei, A. C. Farinha, J. G. de Souza, P. G. Ramos, A. F. Martins, L. Coheur, and A. Lavie. Searching for COMETINHO: The Little Metric That Could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium, June 2022. European Association for Machine Translation. URL `https://aclanthology.org/2022.eamt-1.9`.

S. Ruder, I. Vulić, and A. Søgaard. A Survey Of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631, Aug. 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL `http://arxiv.org/abs/1706.04902`. arXiv:1706.04902 [cs].

D. Saunders. Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. 2021. doi: 10.48550/ARXIV.2104.06951. URL `https://arxiv.org/abs/2104.06951`.

H. Schwenk. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037. URL `https://aclanthology.org/P18-2037`.

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. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL `https://aclanthology.org/P16-1009`.

F. Stahlberg. Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, 69:343–418, Oct. 2020. ISSN 1076-9757. doi: 10.1613/jair.1.12007. URL `https://jair.org/index.php/jair/article/view/12007`.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. Technical Report arXiv:1409.3215, arXiv, Dec. 2014. URL `http://arxiv.org/abs/1409.3215`. arXiv:1409.3215 [cs] type: article.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *LREC*, 2012. URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf`.

J. Tiedemann and S. Thottingal. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for*

Machine Translation, pages 479–480, Lisboa, Portugal, Nov. 2020. European Association for Machine Translation. URL `https://aclanthology.org/2020.eamt-1.61`.

A. Toral, S. Castilho, K. Hu, and A. Way. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6312. URL `https://aclanthology.org/W18-6312`.

M. van der Wees. *What's in a Domain? Towards Fine-Grained Adaptation for Machine Translation*. PhD thesis, Universiteit van Amsterdam, 2017. URL `https://pure.uva.nl/ws/files/19726462/Thesis.pdf`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. URL `https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

W. Weaver. Translation. In *Proceedings of the Conference on Mechanical Translation*, Massachusetts Institute of Technology, 17-20 June 1952. URL `https://aclanthology.org/1952.earlymt-1.1`.

H.-R. Wei, Z. Zhang, B. Chen, and W. Luo. Iterative Domain-Repaired Back-Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5884–5893, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.474. URL `https://aclanthology.org/2020.emnlp-main.474`.

J. Wei and K. Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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 6382–6388, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670. URL `https://aclanthology.org/D19-1670`.

G. Xu, Y. Ko, and J. Seo. Improving Neural Machine Translation by Filtering Synthetic Parallel Data. *Entropy*, 21(12):1213, Dec. 2019. ISSN 1099-4300. doi: 10.3390/e21121213. URL `https://www.mdpi.com/1099-4300/21/12/1213`. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

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. doi: 10.18653/v1/D16-1160. URL `https://aclanthology.org/D16-1160`.

J. Zhang and C. Zong. Neural Machine Translation: Challenges, Progress and Future. *Science China Technological Sciences*, 63(10):2028–2050, 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1632-x. URL `https://doi.org/10.1007/s11431-020-1632-x`.