Investigations into Machine Translation Using OpenNMT

Rebecca Rempt, 4120418

MA Thesis Vertalen, Utrecht University

Supervisor: Lette Vos, M.A.

Second Reader: Dr. Gys-Walt van Egdom

August 2019

Table of Contents

1. Introduction

In the past century, translation technology has been rapidly increasing in both quality and quantity. The development of new hardware and a better understanding of the capabilities of the tools we have at our disposal have enabled us to create continually improving systems for machine translation. Also referred to as automatic translation at times, the general definition of machine translation according to the European Association for Machine Translation is "the application of computers to the task of translating texts from one natural language to another" ("What is Machine Translation"). However, there is some discussion about what exactly falls under the header 'machine translation': some people consider computer-aided translation or interactive translation to fall under this header as well as fully automatic translation. Machine translation is often integrated into computer-aided translation tools (CAT-tools), giving the user the option to autofill the source text segments with machine translation, to see the machine translation separately, or not to include or provide it altogether.

Machine translation supports ever-faster international communication and can help improve international connections without necessitating the intervention of other people. Machine translation can be especially useful in situations where time is of the essence or in situations where understanding a foreign-language text quickly is important, but the translation does not have to be of a publishable quality. Although machine translation systems have been

rapidly increasing in quality over the last few years, it is still unlikely that machine translation will soon, if ever, be of the same quality as trained human translation. As a result, machine translation is usually accompanied by human post-editing before publication.

This thesis explores the history and development of machine translation from a theoretical and practical perspective. The theoretical side focuses mainly on the history of machine translation and the development and evaluation of machine translation systems in a broader sense rather than on technical details. For the practical side, I will use OpenNMT, an open source neural machine translation toolkit, to train a neural machine translation system using the Dutch-English Europarl parallel corpus. Through the use of the BLEU evaluation metric, I hope to track the development and improvement of this system over the training period. Using OpenNMT will likely allow me to focus on the translation output itself as well as the way in which such a translation system improves over time. Existing research on machine translation has mostly been done from the perspective of natural language processing rather than translation studies, and as such focuses mainly on the technical aspects of translation systems. My background in translation studies gives me the opportunity to examine machine translation history, development and application from another perspective.

This thesis is structured as follows: I will first discuss a few different common uses for machine translation in the 'Theoretical Background' section,

as well as different approaches to the construction of a machine translation system. Secondly, I will give an overview of the history and evolution of machine translation, touching on important projects and developments. In the 'Theoretical Background' section, I will also examine the two approaches to machine translation that are currently the most common, namely statistical and neural machine translation, in more detail. In this section, I will also discuss a number of evaluation metrics for machine translation. In the 'Methodology' section, I will discuss the corpus and tools used in this research and the processes applied in the research. Finally, I will examine the results of the project and their implications in the 'Results' section which is finally followed by the 'Conclusion and Discussion' section.

2. Theoretical Background

This section focuses on the theoretical side of machine translation. It gives a historical overview of the development of machine translation, after which it touches on the most commonly used approaches to machine translation system, with more attention to the two most common approaches: statistical and neural machine translation. Finally, it discusses several ways in which machine translation are used as well as some evaluation metrics for machine translation output.

2.1. Historical Background

According to Hutchins (1993), the development of machine translation can be divided into five basic eras: the pioneer era (1947-1954), the first generation of machine translation systems (1954-1966), a 'quiet period' (1966-1975), a revival period with second generation systems (1976-1989), and a 'new' era with likely third generation systems (1989-1993+). As Hutchins made this division in 1993, the period from 1993 onwards and its developments were not taken into consideration. However, an argument can be made for a sixth era of machine translation starting around the year 2000 that goes hand in hand with wider use of the Internet and the World Wide Web. In the following section, I will discuss several important milestone and projects from the history of machine translation systems. This is by no means an exhaustive history of machine translation systems and their development. General trends and some large projects have been singled out to focus on to present a general timeline of the course of the development of machine translation from the early 20th century to the present. For a more detailed history, see Hutchins, 2007 and Slocum, 1985.

Although the ideas of universal languages and mechanical dictionaries have been around for a long time, the idea of automatic or machine translation really started gaining ground during the first half of the 20th century. In 1933, two separate patents for mechanical translators cropped up in France and Russia. However, when the second generation of computers, also called transistor computers, was developed at the end of the 1940s, interest in machine

translation systems and research on the subject greatly increased. The smaller size, lower cost and higher reliability compared to first generation computers, combined with the use of high level programming languages, made these computers more accessible and more useful for a variety of purposes (Govindarajalu, 2010). One of these purposes was the further development of machine translation. In 1952, the first Machine Translation Conference was held at MIT and at that time people in the field were already aware that "full automation of good quality translation" was virtually impossible, and that machine translation would always need to be accompanied by pre- and post-editing (Hutchins, 2007, p. 2). At this conference, two additional observations were made, namely that "analysis of word frequency and word meaning should be conducted on a large scale, in various fields and in as many languages as possible" and that "operational analysis and syntax should be developed and be available when required; namely, for use with available large and high-speed computers." (Henisz-Dostert, Bozena, et al., 1979, p. 14). In 1954, the first edition of the journal *Machine Translation* was published. This journal was in publication until 1970. In 1955, the first book about machine translation was published: *Machine Translation of Languages* by Locke and Booth was an anthology of fourteen early essays about machine translation as well as an annotated bibliography of 46 publications on the subject, virtually all the literature that existed at the time. At the same time, in the 1950s, various machine translation research groups and projects were active around the world

in e.g. the United States (MIT, Columbia), the United Kingdom (Cambridge), Italy (Milan), Russia, China and Japan. Most of the research in the United States was specifically geared towards automatic translations of Russian, and vice versa. During this first decade of machine translation research, there were three basic approaches taken in the development of machine translation systems, all of which were rule-based: direct translation systems, transfer-based translation systems and interlingua systems.

In 1964, a committee was established by the United States government to evaluate progress in computational linguistics and machine translation, the Automatic Language Processing Advisory Committee or ALPAC. In 1966, this committee issued a report, after which there was a large drop in the amount of research done into machine translation - especially in the United States. According to the report, "[t]he Committee … believes that it is wise to press forward …, but that the motive for doing so cannot sensibly be any foreseeable improvement in practical translation." (ALPAC, 1966, p. 24). The reasons for continuing research into machine translation would be purely scientific, and as the committee did not find any "pressing need for machine translation" (ALPAC, 1966, p. 24), and did not feel continuing research at that time would have practical benefits. The committee also found machine translation in combination with post-editing to be slower and more expensive than human translation. According to the report, readers of both raw MT output and post-edited output were significantly less accurate and slower, and had a

significantly lower comprehension level than those reading human translations (ALPAC, 1966). Despite the fact that the committee found it "wise to press forward" (ALPAC, 1966, p. 24) with machine translation research, even if for purely scientific purposes, many ongoing research projects took the committee's verdict regarding the possibilities of, imminent use for, and improvement of machine translation to heart, and the amount of research done on the subject significantly dropped. Other countries, however, were less affected by the ALPAC report and continued with their research into machine translation, which produced a number of results. A particularly successful example is the Météo weather-translation system in Canada, which was developed from 1975 onwards and was officially in use starting in 1985, although it had been in operation starting in May 1977. This system translated weather bulletins from English to French and was in use until the year 2000. The Official Languages Act in Canada gave English and French equal status and meant that most information had to be available in both French and English. Due to the amount of work involved in translating these weather bulletins, the development of a dedicated translation machine was a good solution. Even though the translations produced by the Météo system still had to be post-edited by senior translators, the system led to a decrease in time spent as well as a decrease in costs, with no significant changes in the translation quality. The system was, according to Thouin (1982), unique in the respect that "translators played a crucial role in all stages of the design, development, use and refinement of the

system" (p. 43). The specialised nature of the translations is likely a big part the reason why this system was successful for so long.

Other successful specialised machine translation systems were developed during the 1970s and 1980s, such as the Smart systems at the beginning of the 1980s. These systems by the Smart corporation relied on "strict control of input [...] vocabulary and syntax" in order to produce translations of a reasonable quality with minimal revision for a variety of companies (Hutchins, 2007, p. 7). By the late 1980s, translation systems for personal computers had become available as well. The resurgence of research in the 1970s and 1980s was characterised by "the almost universal adoption of the three-stage transfer-based approach" which was largely syntax-oriented and "founded on the formalisation of lexical and grammatical rules influenced by linguistic theories of the time" (Hutchins, 2007, p. 8). Additionally, many researchers in the 1980s believed that "natural language processing research within the context of artificial intelligence" would be the most likely means of improving machine translation output quality (Hutchins, 2007, p. 9). One of the best-known systems of the second half of the 20th century was the Eurotra system, a transfer-based system that was developed by the European Communities between 1978 and 1992. It was a collaboration between various institutions in participating member states with two aims: firstly to develop a (prototype of a) machine translation system capable of translating between the 9 languages of the European Communities at the time; Dutch, English, German,

Spanish, Portuguese, Danish, Italian, Greek and French, and secondly to stimulate research on the subject of computational linguistics in the European Communities (Raw et al., 1988). At the end of the project, however, the desired end result was not reached. Although, according to Krauwer (1999), the development of a fully operational machine translation system proved to be unattainable at that time, starting points for later machine translation activities were made and the project had an enormous impact on natural language processing research in many European countries. Additionally, the project helped establish a connected network of institutions and individuals all over Europe (Krauwer, 1999, slide 16). As such, even though the original goal was not reached, the project did have several useful results in the end.

The second half of the 1980s also came with an increased interest in interlingua systems, which was partially motivated by contemporary research into artificial intelligence and cognitive linguistics. One such system was the DLT (Distributed Language Transfer) System at BSO software company in Utrecht, which used a modified form of Esperanto as the intermediary language. This project "made a significant effort in the construction of large lexical databases" and in later years proposed the construction of a bilingual knowledge bank from a corpus of human translated text (Hutchins, 2007, p. 9). Another innovative machine translation project in the 1980s was the Rosetta project, which was developed by Philips in Eindhoven. Rosetta was an interlingua system which aimed to use Montague Grammar in interlingual

representations. According to Landsbergen, Rosetta's main developer, Montague Grammar specifies two things: "(i) a set of 'basic expressions', expressions with a primitive meaning, and (ii) a set of compositional rules [...], which prescribe how larger expressions and ultimately sentences can be built from these basic expressions" (1989, p. 82). For use in a translation system, Montague Grammar rules had to be altered: basic expressions have one meaning in original Montague Grammar rules, but for practical purposes, Landsbergen found it more useful to allow them to have multiple meanings, leading to a distinction between syntactic and semantic derivation trees (Landsbergen, 1989). The Rosetta system is largely an interlingua system, but with an important difference: where interlingua systems usually depend on a universal intermediate language, the Rosetta system relied on an intermediate language that was specially defined for the specific set of languages the system worked with. A number of other projects of differing sizes and with differing levels of success also took place during the 1980s around the globe, in - among others - Japan, Korea, Russia, and North America as well as additional projects in Europe.

In the 1980s and 1990s, the use of computer aids for professional translation slowly became more common. In the 1980s, many different tools were developed for a variety of purposes such as word processing, glossary creation, and online document sharing. In the 1990s, all these features and more were combined in - at the time - revolutionary translators' workstations or

workbenches or Computer Assisted Translation Tools (CAT-Tools). These workbenches provided a dedicated environment for translators to work in, with access to terminology management software, multilingual word processing, integrated machine translation and perhaps most revolutionary, translation memories. Although Matthias Heyn, Vice President of Global Solutions at SDL, claims that at the time there was no market for these kind of systems ("Daniel Brockmann and Matthias Heyn look back on 30 years of TRADOS"), this was the time when personal computers first became available to most people, including translators, and their use increased. CAT-tools are especially useful in the translations of documents with a lot of repeated segments such as user manuals and technical documentation, as repeated segments auto-fill after the first translation, which makes it easier to produce a cohesive translation and reduces workload. Among the most commonly used CAT-tools are currently (SDL) Trados, MemoQ, Déjà Vu, Wordfast and OmegaT.

In the 1990s, a new approach to machine translation emerged. Whereas previously rule-based approaches were most common, machine translation research now turned to corpus-based approaches and at the same time returned to a statistics-based approach to machine translation (Hutchins, 2007). Corpus-based approaches to statistical machine translation have stayed common in machine translation research since then. A turning point in statistical machine translation approaches took place in 2007, with the development and release of the Moses toolkit. The Moses toolkit is a "complete

out-of-the-box translation system for academic research", which "consists of all the components needed to pre-process data, train the language models and the translation models" (Koehn et al., 2007, p. 178). Until its release, most statistical machine translation research was carried out on "proprietary and in-house research systems", which formed a barrier to a lot of research. The Moses toolkit was meant to "stimulate the development of the field" (p. 177). Although statistical machine translation has several advantages over the previously commonly used rule-base systems, such as improved semantics and improved training and translation efficiency, it still has some limitations. These include limitations in subject-verb agreement, word reordering and tense modeling, among others (Shterionov et al., 2018).

Another important development since the 1990s has been in the field of spoken language translation, which comes with "the formidable challenges of combining speech recognition and synthesis, interpretation of conversations and dialogues, semantic analysis, and sensitivity to social contexts and situations." (Hutchins, 2007, p. 13). Since the mid-1990s, the use of machine translations systems on the Internet has also increased. The output these systems produce was, and often still is, of relatively poor quality, but they serve an important demand of machine translation, namely the need for fast translation where quality is not of the essence. The first online machine translation service, launched in 1997, was Yahoo!'s Babel Fish. Although it is now defunct, it was in operation until 2012, after which it was replaced by the

Bing Microsoft Translator. Google Translate, then a statistical machine translation system, was launched at the height of statistical machine translation's rise in 2006. Although research into deep learning had been in progress since the 1990s, the first use of neural networks in machine translation appeared in 2014, and in 2016, Google started using Google Neural Machine Translation (GNMT) instead of their previously used statistical machine translation system.

2.2 Approaches to Machine Translation

The most common approaches to machine translation systems can be divided into two main categories: rule-based machine translation and corpus-based or example-based machine translation. Rule-based machine translation can be further divided into three categories: direct translation systems, transfer-based machine translation systems and interlingua systems. Corpus- or example-based machine translation can in turn be further divided into statistical machine translation and neural machine translation. At present, statistical and neural machine translation are the most commonly used approaches.

The direct translation approach involves programming rules for, as the name implies, direct word-for-word translation from one source language (SL) into one target language (TL). Because of this, this approach is sometimes also referred to as 'dictionary-based translation'. It involves minimal analysis and

syntactic reorganisation. Problems associated with homonyms and ambiguity were reduced by simplifying dictionaries and providing only one TL equivalent for a SL word, which reduced the need for analysis of contexts in the SL. The SL word order was usually maintained exactly. A problem with this approach is the simplification of dictionaries inherent to it: it is possible that the one TL equivalent that is provided semantically does not match up with the source, in which case the translation would be unsuccessful. In addition, since the SL word order is maintained exactly, this approach cannot produce translations suitable for publication except in very rare cases. This direct translation model would only be useful in highly specialised cased where there is definitely only one possible equivalent to the SL word and grammar is unimportant.

Transfer-based machine translation breaks up the translation into three stages: analysis, transfer and generation. In this approach, the grammatical structure of a ST is first analysed after which it is then transferred to a structure suitable for translation into the TL. Finally, the TT is generated according to this structure.

In the interlingua model, translation happens through two stages. The ST is analysed into an abstract meaning representation called an 'interlingua', after which the TT is generated from the interlingual representation (Jurafsky & Martin, 2009). Transfer-based machine translation operates on the same principle as the interlingua model, namely that source and target need to share an intermediate representation of 'meaning' to facilitate translation.

In statistical machine translation system, large bilingual parallel corpora are used to train a statistical translation model. When given a ST, the model produces the translation with the highest probability of being correct based on the corpus data it received. Statistical machine translation improves on rule-based machine translation in multiple ways, such as semantics and translation and training efficiency (Shterionov, 2018). Phrase-based alignment metrics and the incorporation of syntactical information can be used to improve statistical machine translation output even further. Neural machine translation is a relatively recent development that uses an end-to-end neural network to predict translation output. These networks are trained on large corpora of parallel bilingual data. Neural machine translation improves on certain weaknesses of statistical machine translation and has in certain situations been outperforming statistical translation systems.

2.3 Statistical and Neural Machine Translation

As mentioned before, the two most wide-spread approaches to machine translation at the moment are neural and statistical machine translation. Both statistical machine translation and neural machine translation depend on machine learning and use large bilingual parallel corpora, although the specific approach differs between the two approaches to machine translation. The two approaches both have their stronger and weaker points, which I will discuss in this following section.

Interest in statistical machine translation started growing at the end of the 1990s. Since then, statistical machine translation has grown to be one of the most-used forms of machine translation as it has a distinct advantages over previous approaches to machine translation. Compared to for example rule-based translation systems, statistical machine translation produces better quality output by improving on semantics as well as on training and translation efficiency (Shterionov, 2018). Statistical machine translation does, however, struggle with some aspects of translation, for instance with subject-verb agreement. Rule-based machine translation tends to produce syntactically better translations than statistical machine translation, as the statistical translation output tends to have obvious errors such as a lack of number and gender agreement (Vanmassenhove, 2016). However, rule-based systems in turn struggle with general fluency as well as lexical selection, which statistical systems improve on drastically. Word reordering and tense modeling are other aspects that statistical machine translation systems struggle with, although the first is improved by incorporating more syntactical information in the model and by focusing on phrase-based alignment over word-by-word alignment.

Jurafsky and Martin (2000) describe statistical machine translation as a way to approach translation that "focus[es] on the result, not the process" (p. 819). Brown et al. (1990) assume that "every sentence in one language is a possible translation of any sentence in the other" (p. 79). In statistical machine translation, every TL sentence is assigned a probability of being produced as a

fluent, accurate translation of a SL sentence. The statistical translation model then chooses the sentence with the highest probability as the translation. Within this model, a distinction can be made between the translation probability and the language model probability. Brown et al. (1990) explain the translation probability as ”suggesting words from the SL that might have produced the words that we observe in the target sentence” while the language model probability suggests “an order in which to place these source words” (p. 79). Parallel examples, usually a large parallel, sentence- or word-aligned corpus, are used to train such statistical models. As such, statistical machine translation relies on “statistical parameters and a set of translation and language models, among other data-driven features” (Costa-Jussà, 2012, p. 247). It is a “mathematical model in which the process of human language translation is statistically modeled” and “model parameters are automatically estimated using a corpus of translation pairs” (Yamada & Knight, 2001, “1. Introduction”, para. 1). Statistical machine translation does not generally incorporate any syntactical or semantical analysis of the ST, although there have been models that attempt to incorporate syntactical information, such as the model proposed by Yamada and Knight (2001). This model was a response to the IBM model developed by Brown et al. (1993), which “does not model structural or syntactical aspects of the language” and as such would likely not function well on language pairs with a significantly different word order (Yamada & Knight, 2001, “1. Introduction”, para. 3). Yamada and Knight’s model incorporates

structural aspects of the language by pre-processing the input sentences by a syntactic parser, which turns them into a parse tree. The model then performs three operations: reordering child nodes to simulate translation between languages with different word orders, inserting extra words such as case-marker particles where necessary, and finally, translating (Yamada & Knight, 2001). Due to the incorporation of syntactical information, this model performed better on word-alignment tasks for language pairs with different word orders than the IBM Model 5 (Yamada & Knight, 2001).

Many early statistical translation models, especially in the 20th century, used word-to-word alignment, where often each source word would correspond to exactly one target word. However, this fails to take into account dependencies between groups of words and makes it difficult to properly handle the translation of compound nouns or deal with differences in word order between source and TLs (Och et al., 1999). Around the year 2000, statistical machine translation researchers started experimenting with phrase-based alignment instead. Och et al. (1999) created a model that uses two types of alignment: "a phrase level alignment between phrases and a word level alignment between single words within these phrases" (p. 20). Their so-called "alignment template" approach led to better translation results compared to a word-to-word alignment approach (Och et al., 1999, p. 27). Koehn et al. (2003) created a statistical machine translation framework consisting of a translation model and a decoder in order to "evaluate and compare various phrase

translation methods" (p. 54). Based on this model, they conclude that phrase-based statistical machine translation leads to better results than traditional word-based methods (Koehn et al., 2003). Overall, phrase-based statistical machine significantly improves translation results compared to translation models using word-by-word alignment. Chiang (2005) notes that phrases "can be any substring and not necessarily phrases in any syntactic theory" and that the use of phrase-based alignment allows translation models to learn "local reorderings, translation of short idioms, or insertions and deletions that are sensitive to local context" (Chiang, 2005, p. 263). Considering the fact that phrase-based translation models are good at reordering of words during translation, Chiang (2005) assumes that such models can also be used for the reordering of phrases. To test this, they developed a hierarchical phrase-based model that makes use of hierarchical phrases that contain both words and sub-phrases. This model is able to learn hierarchical phrase pairs without syntactically annotated training data and translation accuracy is significantly improved using this model compared to state-of-the-art phrase-based systems (Chiang, 2005). According to Chiang, incorporating syntactical information in statistical machine translation could potentially improve efficiency and accuracy further.

Neural machine translation systems that produce high-quality output are a relatively recent development. Interest in neural machine translation has been sharply rising since around 2013, and it has been outperforming statistical

machine translation for multiple language pairs and translation tasks since 2015 (Sterionov et al., 2018). Neural models used in machine translation "involve building an end-to-end neural network that maps aligned bilingual texts which, given an input sentence X to be translated, is normally trained to maximise the probability of a target sequence Y without additional external linguistic information" (Castilho et al., 2017, p. 110). Deep neural networks have been employed to build end-to-end encoder-decoder models for machine translation since around 2013, which is also the reason for the rising interest in neural machine translation around that time. More recent developments in the field of neural machine translation include improving attention mechanisms trained to "attend to the relevant source-language words as it generates each word of the target sentence", as well as including linguistic information in the models and incorporating more languages into the translation models (Castilho et al., 2017, p. 110). Several large MT vendors, including Google, KantanMT and Systran have started offering neural machine translation as part of their services since then. For example, Google has switched their translation service from a statistical machine translation system to their Google Neural Machine Translation system in 2016. Neural machine translation's strength lies in "its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text" (Wu et al., 2016, p. 1). Wu et al. mention that neural machine translation "sidesteps many brittle design choices in traditional

phrase-based machine translation", although they do not mention what exactly these "brittle design choices" are (2016, p. 1).

One of the main drawbacks of neural machine translation is its computational and financial cost compared to phrase-based machine translation systems. Additionally, while neural machine translation output often looks very fluent, it sometimes lacks adequacy or is even simply wrong (Shterionov, 2018). Wu et al. (2016) identify three main weaknesses in neural machine translation: its training and inference speed are slower than those of phrase-based statistical machine translation and other machine translation systems, it is ineffective when dealing with rare words, and it occasionally fails to translate all the words in a source sentence. Despite these drawbacks, neural machine translation is still a relatively young development in machine translation and is already outperforming statistical machine translation in many areas when measured with automatic evaluation methods. However, comparing neural machine translation and statistical machine translation output using automatic and human evaluation, Castilho et al. (2018) found that based on human evaluation, neural machine translation had at the time of their research not fully reached the same quality as statistical machine translation. There have since, however, been claims of neural machine translation reaching human parity (Hassan et al., 2018). Toral et al. (2018), however, reassessed this claim and found several variables in the original study that would likely have led to better evaluation scores than otherwise expected. The study by Hassan et al. focused on the

translation of news texts from Chinese to English, half of which were originally written in English and translated to Chinese prior to the study. Toral et al. argue that this makes the translation back to English more likely to result in higher evaluation scores due to the influence of 'translationese', and in particular due to simplification in the original translation to Chinese. Both in Toral et al.'s evaluation and in the original evaluation by Hassan et al., the texts originally written in Chinese did not reach human parity in evaluation (Toral et al., 2018). In addition, Toral et al. found issues in the human reference translations that indicated that the translations were "conducted by non-experts and possibly post-edited MT output", which makes it more likely for the evaluation results to be untrustworthy (Toral et al., 2018, p. 121). (Neural) machine translation output has been improving dramatically in recent years, but is not consistently at a human level yet and "human translators will continue to find gainful employment for many years to come" (Toral et al., 2018, p. 122).

2.4 Applications of Machine Translation

According to Christiane Nord, "the prime principle determining any translation process is the purpose (*Skopos*) of the overall translational action" (Nord, 2014, p. 27). Although this statement is made with a human translator in mind, it can be interpreted to apply to different levels of machine translation quality as well. Hutchins (2007) distinguishes three types of demands for the application of

machine translation, all of which benefit from different levels of translation quality. In all of these cases, the purpose of the translation is what determines what level of quality is necessary, which in turn influences the decisions made when choosing the type of machine translation and the level of post-editing.

The first type of demand for machine translation is "the traditional need for translations of 'publishable' quality" (Hutchins, 2007, p. 1). Translations of publishable quality, however, are difficult to achieve through machine translation alone. Thus, this demand is often filled by HAMT, or human-aided machine translation: machine-translated texts are post-edited to a publishable standard by human translators and editors. The use of machine translation in this case can save costs and time, as the post-editing of high-quality machine translation output can "increase the productivity of professional translators compared to manual translation 'from scratch'" (Kaponen, 2016, p. 132). Post-editing of poor quality machine translation output is, however, less productive than translation from scratch (Kaponen, 2016). According to Hutchins (2007), this application for machine translation is also called "machine translation for dissemination" (p. 1). The second type of demand is not for texts of publishable quality but for "something that can be produced quickly … conveying the essence of the original text" (Hutchins, 2007, p. 1) without having to be grammatically, lexically or stylistically perfect, or even - in many cases - particularly good. As long as the information provided in the text can be understood in the translation, the translation can be considered acceptable. This

application uses machine translation with no further human interference or interaction and is called "machine translation for assimilation" (Hutchins, 2007, p. 1). The third type of demand is a relatively recent development since the rise of the Internet. Quality is also not of the essence for this application of machine translation, which is called "machine translation for communication" (Hutchins, 2007, p. 1). As the name suggests, in this case machine translation is used in social situations to be able to communicate quickly with people in different languages than your own. Email correspondence or chatroom interactions, for instance, are translated with a machine translation system upon receipt. The only requirement for the quality of the translation in this situation is usually that the output is understandable, while fluency is not a factor. Google Translate is often used for this purpose nowadays, as it is accessible and offers a wide selection of language pairs.

2.5 Machine Translation Evaluation

In machine translation research, a distinction can be made between two types of evaluation: automatic and human evaluation. As the term suggests, automatic evaluation is the evaluation of a translation using computerised means. Usually, this is done with the aid of a reference text. The evaluation program compares the produced translation with the provided reference and scores it based on how much it differs from the reference. Because of the way in which translations are evaluation in this sort of system, a translation could be a

grammatically and semantically correct representation of the ST and still earn a low score if different synonyms are used or the word order differs from the reference. However, automatic evaluation is more efficient, easier to use on large amounts of text, and more cost-efficient than human translation.

Using automatic evaluation also implies that there is no chance of a human evaluator's subjective opinion influencing the evaluation. Automatic evaluation is always impartial and uses the same parameters for every translation, which makes it useful for quantifiable evaluation data. Automatic evaluation for machine translation has to meet several criteria to be effective and useful. It has to have a high correlation with human judgement of the quality of the machine translation and it should be "as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system" (Banerjee & Lavie, 2005, p. 66). Furthermore, it should be consistent, reliable and general: evaluating similar outputs on the same machine translation system should also produce similar scores, systems that score similarly should perform similarly, and it has to be "applicable to different MT tasks in a wide range of domains and scenarios" (Banerjee & Lavie, 2005, p. 66). Automatic evaluation metrics are generally employed for two purposes: to compare two or more machine translation systems with each other to determine which of the systems generates better translations, and to "automatically optimize or tune the parameters of a system" (Snover et al., 2009, p. 116).

Some of the most common automatic evaluation metrics are BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee & Lavie, 2005) and TER (Translation Edit Rate or Translation Error Rate) (Snover et al., 2006). The BLEU evaluation system compares a machine translation's output and a correct, usually human reference translation to compute the precision of the system (Shterionov, 2017). Specifically, it compares the n-grams, or sequences of *n* number of words, of the output translation with the n-grams of the reference translation and counts the number of matches (Papineni et al., 2002). There are three further relevant factors in calculating BLEU scores: translation length, translated words, and word order. A higher scoring translation matches the reference translation in length, the words used in the translation match those in the reference translation, and the order of the words is the same in the translation and in the reference. BLEU scores range from 0 to 1, but are often represented as a value between 0 and 100 instead. A higher score points to a 'better' translation, or at least one that matches the reference(s) more closely. BLEU assumes that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002, p. 311). In their experiments, Papineni et al. (2002) have found BLEU to correlate heavily with human judgement of translations. However, Callison-Burch et al. (2006) have shown that BLEU scores may not correlate with human judgment as much as previously thought and note that "an improved Bleu score is not sufficient to

reflect a genuine improvement in translation quality" (p. 255). While Callison-Burch et al. (2006) note that this means that BLEU is not a useful tool for the comparison of translation systems that employ different strategies, they do state that it is still useful in tracking improvements within a single system or comparing translation systems that do employ the same strategies in translating.

In the METEOR system, translations are evaluated by "computing a score based on explicit word-to-word matches between the translation and a given reference translation" (Agarwal & Lavie, 2008, p. 115). It was developed to address some of BLEU's weaknesses, in particular BLEU's emphasis on n-gram precision which "does not appropriately measure the degree to which a machine-generated translation captures the entire content of the source sentence" (Russo-Lassner et al., 2005, p. 3). Additionally, according to Russo-Lassner et al., BLEU and similar automatic evaluation metrics "do not correlate well with human judgment at the sentence level, despite correlations over large test sets" (2005, p. 3). In addition to exact word-to-word matches, METEOR also supports matching between words that are morphological variants with an identical stem and matching between synonyms (Banerjee & Lavie, 2005). According to Banerjee & Lavie (2005), METEOR is able to reach a higher correlation with human evaluation than BLEU on both a sentence-by-sentence basis and on the system level.

TER works quite differently from both BLEU and METEOR in that it does not focus on explicit word matches. Instead, TER scores are calculated by measuring "the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation," or in other words, the number of edits it would take to make the output semantically match a correct translation (Snover et al., 2006, p. 223). Snover et al. claim that this "measure of 'goodness' of MT output" is more intuitive than previous automatic evaluation metrics (2006, p. 223). Edits that are counted in calculating TER include the "insertion, deletion, and substitution of single words as well as shifts of word sequences" (Snover et al., 2006, p. 225). Incorrect capitalisation and punctuation errors are also counted as edits. A lower score is preferable in the TER metric, seeing as that points to a lower number of required edits. TER also provides an alignment between a translation and a reference in addition to providing a score for the translation, which makes it useful for things beyond translation evaluation (Snover et al., 2009). In 2009, an extension to the TER metric called TERp or TER-Plus was developed (Snover et al., 2009). TERp was designed to address several flaws of the original TER metric, in particular the fact that TER only "considers exact matches when measuring the similarity of the hypothesis and the reference, and it can only compute this measure of similarity against a single reference" (Snover et al., 2009, p. 118). In addition to aligning a translation and reference on the basis of exact word matches, TERp considers synonyms and uses stemming to allow more matches. It creates

alignment between phrases in the translation and the reference as well, through the use of "probabilistic phrasal substitutions" (Snover et al., 2009, p. 118). While TER counts mis-capitalisation as an error or needed edit, TERp does not, as doing so was shown to decrease the correlation with human judgement of the translation quality (Snover et al., 2009). According to Snover et al., TERp "achieves significant gains in correlation with human judgments over other MT evaluation metrics" (2009, p. 126).

3. Methodology

The last chapter discussed several different options for machine translation systems, as well as their advantages and disadvantages. It also examined several evaluation metrics and programmes for machine translation quality. In this chapter, I will discuss which systems and metrics are used in this research and why.

3.1 Corpus and Tools

The Europarl corpus is a parallel corpus developed by Philip Koehn, containing text in twenty-one European languages extracted from the proceedings of the European Parliament between 1996 and 2011. The Dutch-English portion of the corpus contains nearly two million sentences with fifty million words in each language, spanning April 1996 to November 2011. It is free to use and, due to the public nature of the European Parliament, contains no copyrighted or classified information. The goal of the corpus was to "generate sentence aligned text for statistical machine translation systems" (Koehn, 2005), but due to its availability and size is often used in other research that calls for large or multilingual parallel corpora as well. The files used in this research have been downloaded from the OPUS website, which is a "growing resource of freely accessible parallel corpora" (Tiedeman, 2012, p. 2214). The Europarl corpus is available on the OPUS website in various formats, including as a pre-aligned translation memory file (.tmx) and tokenised monolingual files. The version

used in this research is the 'Moses format' plain text version consisting of two aligned .txt documents containing the entire corpus.

OpenNMT is "an open source (MIT) initiative for neural machine translation and neural sequence modeling" (Klein et al., 2018). The system was designed to be "simple to use and easy to extend, while maintaining efficiency and state-of-the-art accuracy" (Klein et al., 2018, p. 177). It prioritises efficiency and modularity and according to Klein et al. was developed with three main aims in mind: "to prioritize fast training and test efficiency", to "maintain model modularity and readability", and to support significant research extensibility, as well as "providing code for core translation tasks" (Klein et al., 2018, p. 177). The primary of these aims is system efficiency in both training and testing, which is achieved in three ways. The first is memory-sharing and sharding, which enables large training datasets to be broken into shards that are shared and loaded during training to save on memory (GPU) usage. Memory size restrictions directly impact the time needed for training neural machine translation models, which is significantly improved using this method. Besides this, training and testing efficiency is achieved through options for multi-GPU training, as well as different translation implementations designed for various runtime environments. The secondary of these aims is "a desire for code readability and extensibility" (Klein et al., 2018, p. 180). This is achieved through "explicitly separating training, optimization and different components of the model, and by including tutorial documentation within the code" (Klein

et al., 2018, p. 180). Users are provided with " simple interfaces pre-process, train and translate, which only require source/target files as input" while advanced customisation of models is still available if desired. A schematic overview of the OpenNMT system is shown in figure 1 below.

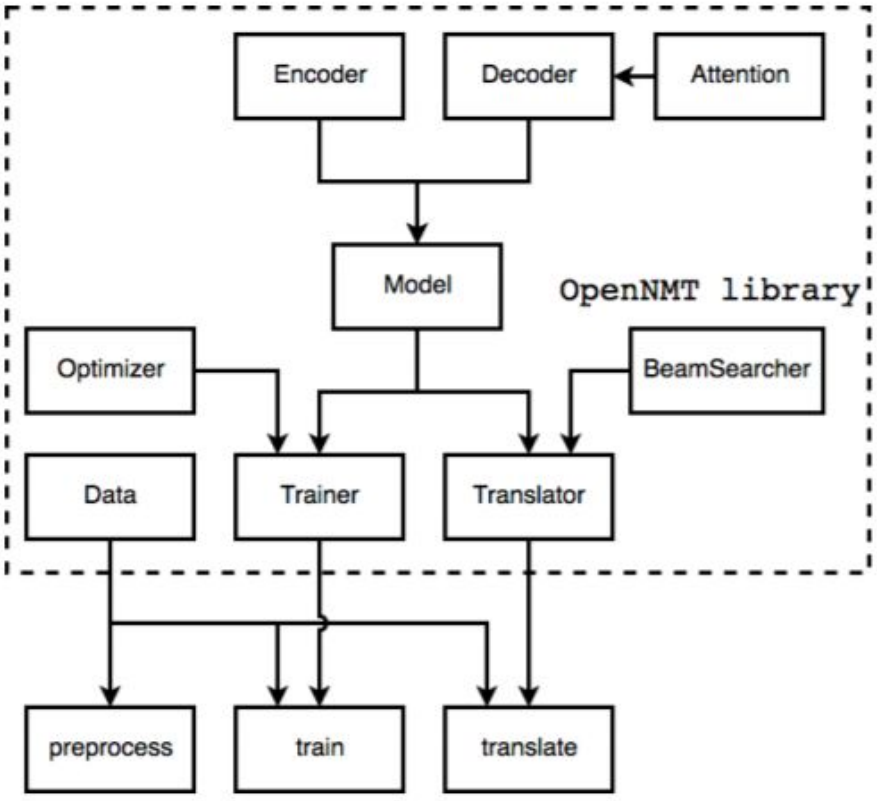

Figure 1. Schematic overview of the OpenNMT-py system (Klein et al., 2018).

As well as the source code, pre-trained models are also publicly available for download, some of which have been trained using the Europarl corpus (although none using the Dutch-English corpus). Tutorials and documentation are available on the website (opennmt.net) as well. The proven compatibility

with the Europarl corpus and the clear documentation were what ultimately made me decide to use OpenNMT in my research. Using the open source OpenNMT toolkit enables me to retain full control over the training corpus and parameters without requiring the construction of a machine translation system from the ground up. In addition, the OpenNMT toolkit has been constructed by machine translation and natural language processing experts with far more experience than me, which would make the system function better than anything I could construct in the limited time and with the limited resources available to me. Using the OpenNMT toolkit will enable me to focus more on the results of the research as well as on the theoretical background, rather than having to devote all available time to the construction process itself. The OpenNMT toolkit offers a perl script that can be used to determine BLEU scores, which will be used to evaluate the progress and improvements to the system made over training.

3.2 Method

3.2.1 Set-up

OpenNMT setup is relatively simple and is well-documented in the documentation. OpenNMT uses either the PyTorch or Tensorflow framework. As PyTorch is recommended over Tensorflow for people who are getting started with deep learning and for academic research, since it is easier to learn and use, that is the option I chose. OpenNMT includes a set of demo files that

can be used to train a rudimentary translation model. I used this demo training set to test whether the setup was successful and to examine the pre-processing and training speed. While pre-processing took no time at all, training turned out to be unfeasibly slow. The training demo consisted of 100,000 steps that were completed in sets of 50 steps. Based on the time taken to complete the first 2,000 steps, completing the whole demo on CPU would take roughly 127 hours. As the demo dataset is rather small - too small to provide a working translation model - some adjustments have to be made to facilitate training of the Europarl model in a reasonable timeframe. The best option to improve training times is to use GPU for training rather than CPU. However, training on GPU with OpenNMT relies on the CUDA toolkit, which is developed by NVIDIA and only works on NVIDIA graphics cards, which I do not have. However, an alternative to training on native GPU is using Google Colab (https://colab.research.google.com/): a "free Jupyter notebook environment that requires no setup and runs entirely in the cloud" ("Welcome to Colaboratory!", n.d.), that enables users to write and execute code in their web browser and offers 12 hours of consecutive use of an external GPU.

3.2.2 Training and Evaluation

Training a neural model requires at minimum parallel source and target data and validation files. I split both languages of the English-Dutch Europarl corpus into three parts: a test file of 400 lines, a validation file of 4944 lines, and a main

training file that consists of the remainder of the corpus which is 2,020,806 lines. The total word count for either language is approximately 50 million words according to the Europarl corpus website (Koehn, 2005). The Dutch files have a slightly higher word count, but there is no significant difference between the two languages. The vocabulary size is very nearly the same for both languages: the Dutch vocabulary size is 50,004 words while the English vocabulary size is 50,002 words. All of these files are in saved in plain text (.txt) format and uploaded to Google Drive to be accessed through Google Colab. While it is possible to upload files directly to the Google Colab environment, these files are no longer accessible after the runtime is disconnected. Training output such as the pre-processed files and the model checkpoints are also saved to Google Drive instead of in Google Colab for this reason.

The model is trained using the standard OpenNMT train.py script using the following command:

```
!python train.py -data '/content/drive/My Drive/Master/Scriptie/OpenNMT/opennmt/'data -save_model '/content/drive/My Drive/Master/Scriptie/OpenNMT/opennmt/model/'one -gpu_ranks 0 -save_checkpoint_steps 10000 -train_steps 1000000 -pool_factor=10
```

This divides the training in 1,000,000 steps and saves the model every 10,000 steps, which means that at the end of training there are a hundred checkpoint models available. Every 100,000th step will be used in evaluation in order to measure the progression of the model in relation to the training time. In the case

of training being interrupted, the model can be loaded from a previously saved step to resume training by adding the following to the end of the previous command:

-train-from '/content/drive/My Drive/Master/Scriptie/OpenNMT/opennmt/model/'one_step_[number].pt

One important thing to keep in mind about using Google Colab is that the runtime only stays connected for either up to 12 hours or until 12 GB of RAM is used. After this, the runtime is automatically disconnected and will have to be reconfigured for use when reconnecting. This means that after every instance that the runtime is disconnected, drive access has to be granted again and OpenNMT has to be reinstalled. In addition, training then has to be resumed from the last saved checkpoint, which can prove a setback of an hour or more depending on the training speed before the runtime was disconnected and how much progress was made since the last checkpoint was saved. In addition, the GPU used is shared between users, meaning that sole use of the GPU is not guaranteed. This can also impact training speed and efficiency, and as a result the total training time. Based on the time taken to complete the first 2,000 steps, completing the training takes about 142 hours or roughly 6 days and nights. Due to the factors mentioned above, training is expected to take slightly longer. If, for example, the runtime unexpectedly disconnects at night, it will not be reconnected until the following morning, which means losing several hours of potential training productivity.

When the training is completed, there are a hundred saved models at every 10,000 training steps. Every tenth model (after 100,000, 200,000, 300,000 steps etc.) will be evaluated with the use of the BLEU evaluation metric. Although, as seen in section 2.4, BLEU does not necessarily correlate well with human judgement and may not be useful in comparing different translation systems, it is still a useful system for tracking improvements within a single system. As such, the BLEU scores will be used to determine the rate of improvement compared to the completed training stages. The test file used to acquire these scores consists of 400 lines, or 10,123 (English) source words, of the Europarl corpus that are not included in the training corpus or the validation file. The test file is translated by each tenth checkpoint model, leading to ten translation predictions, after which the predicted translations are compared to the reference translation and scored through the BLEU metric. OpenNMT provides a perl script that can easily be run on saved prediction .txt files with the following command:

```
!perl /tools/multi-bleu.perl [selected model] < [prediction text]
```

Due to the size of the test file and consequently of the predicted translations, it is not feasible to completely evaluate them through human evaluation as well, especially taking into account the fact that ten translations would have to be evaluated. Instead, a section of approximately 200 words will be selected at random to evaluate in detail. A score will be given based on the number of translation errors in this section, which will be recorded using an evaluation model developed for this rese based on the LISA QA model. A few of the

categories from the LISA QA metric are used, since errors due to human interpretation differences and cultural differences are not relevant in a machine-produced translation. The error categories used are mistranslation/not translated, terminology, language, and omission. Mistranslation is considered a semantically incorrect translation, i.e. if "table" is translated as "chair", while 'not translated' indicates the TT is left in English instead of being rendered in Dutch. Terminology errors are mistranslations specific to specialised terminology. These errors are expected to be relatively rare, seeing as the model training and the to-be-translated text is domain-specific. Language errors include all grammar, style, punctuation and capitalisation errors. Preferential style changes that do not influence the translation semantically will not be counted as errors. Omissions are self-explanatory; any section where a segment is not included in the TT falls under this error type. All errors are weighed the same and no distinction will be made between minor, major or critical errors.

4. Results

4.1 Results and Evaluation

At the end of the training period, the test file was translated by ten models that were saved after each 100,000th training step. The resulting ten translation predictions were used to calculate BLEU scores for each of the ten checkpoint models using the provided script. The scores are calculated by comparing the translation prediction to the reference text. A perfect match would score 1 (or 100), while a translation with no matches at all would score 0. Penalties are applied for sentences that are longer or shorter than the reference, even if they contain the same tokens, as well as for differing word orders. Although BLEU scores are usually calculated on the basis of multiple reference translations, I only had one available. However, as the produced BLEU scores were not used to compare the completed translation model to other translation systems, but only to compare the improvement within a single system, this is not expected to be an issue. Although the entire English-Dutch Europarl corpus of approximately 50 million source and target words was used in training the model, the resulting BLEU scores are not very high. The scores are given in table 1 below on a scale of 1-100. BLEU scores are also visually represented in figure 2 below to illustrate the progression of the model compared to the training time.

Model checkpoint	BLEU score
100,000	20.52
200,000	20.83
300,000	20.93
400,000	20.93
500,000	20.93
600,000	20.93
700,000	20.93
800,000	20.93
900,000	20.93
1,000,000	20.93

Table 1: BLEU scores per checkpoint model.

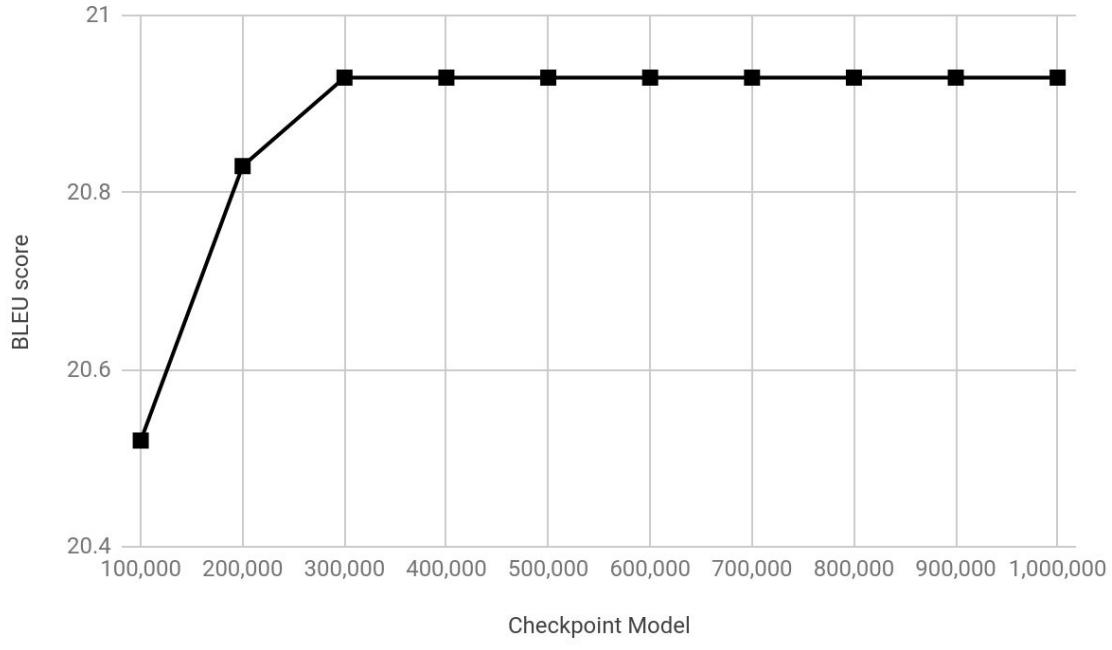

Figure 2: BLEU scores per checkpoint model.

As can be seen in the above table and cart, BLEU scores increased for the first three training stages. However, after this the BLEU scores stayed exactly the same, despite the earlier increase trend. Comparing the different output predictions also showed that the translations were exactly the same for all of these models. There are a few possible explanations for this: the first possibility is that it is simply a limitation of the neural network, and that with the training script used, it is not able to improve any further even with additional data. Another possibility is that there was an unexpected issue during the training process between two checkpoints at 300,000 and 400,000 steps that led to the model stagnating at that point. However, looking back on the training records, there is nothing that would indicate any issues at this point. A more likely explanation is that the training process used is a relatively simple one. Although a reasonably large corpus was used, the model was only trained once with default training options and the model was not fine-tuned during or after training, which would be required to improve the model after a certain point. As such, although the translation quality of the resulting model is examined in this thesis, adjusting the model to improve it further once training was complete was not a priority, especially given the time needed to completely train the model. Klein et al. mention that "OpenNMT's default setting is useful for experiments, but not optimal for large-scale NMT", which would confirm this reasoning (2018, p. 182).

The BLEU scores obtained from the first three checkpoint models should indicate an increase in output quality from each checkpoint to the next. Supposedly, this would mean that when comparing the translation predictions, there should be a clear visible improvement. Because all checkpoint models from 300,000 to 1,000,000 turned out to be exactly the same, only the first three will be evaluated in further detail. A section of 200 words has been chosen at random from the test file to be compared between the three different translation predictions, as well as to the reference test text and the ST. See Appendix A for the complete ST, reference text and the three translation predictions of these 200 words. Table 2 below gives an overview of the number of different translation errors for each of the predictions, while figure 3 gives the same overview visually. For a detailed overview of the different translation errors found in the texts, see appendix B.

	Model 100,000	Model 200,000	Model 300,000	**Total**
(1) Mistranslation or not translated	6	5	4	15
(2) Terminology	1	1	1	3
(3) Language	6	6	7	19
(4) Omission	4	3	3	10
Total	17	15	15	

Table 2: Translation errors per type and model.

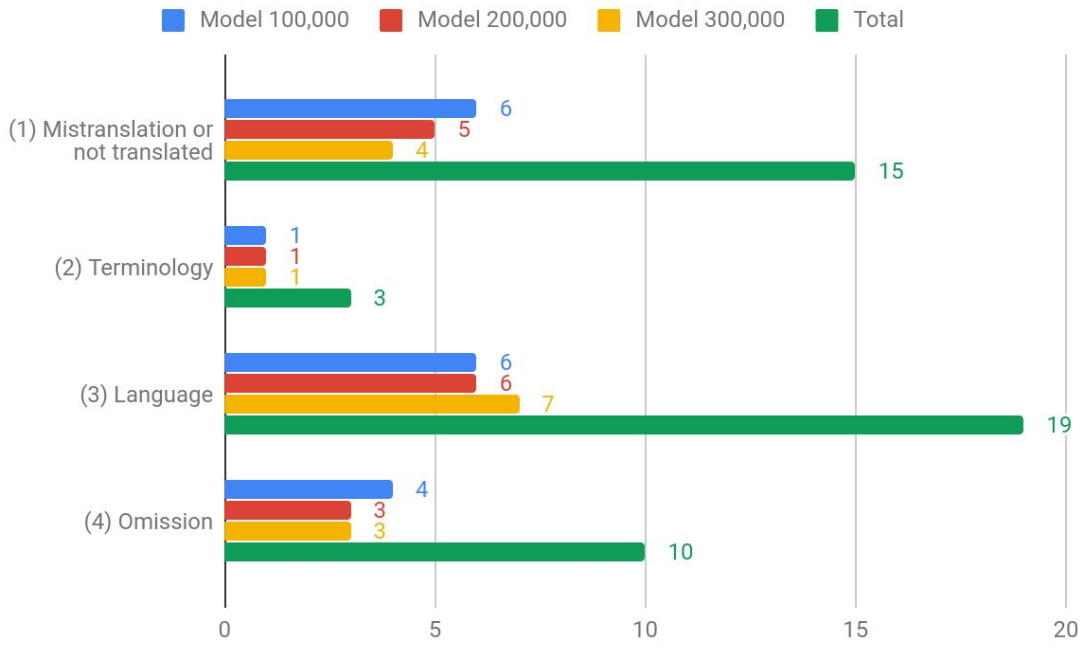

Figure 3: Translation errors per type and model.

As can be seen in these statistics, language-type errors are the most commonly occurring errors in the evaluated passage, likely due to the fact that this is the broadest category and also includes style errors. The number of errors does not decrease significantly from one model to the next, although a slight decrease can be observed. It is possible that detailed evaluation over a longer passage would show a significant change, but that is not certain from these results. As can be seen in figure 3 above, errors to do with mistranslations, untranslated segments, and omissions decrease in frequency as the model is trained further. However, language errors increase in frequency, while the number of terminology errors stays the same. A number of the segments in the evaluated passage are identical across the three models, but the segments that do evolve across the different models are mostly the ones including mistranslations or

omissions. When the model acquires a broader vocabulary and thus learns to translate mistranslated or omitted segments over training, these are in some cases translated with language errors, which explains the fact that the number of language errors increases in the later models. An example of this is the following segment in the ST: “Parliament and the Council have turned Kosovo into an affair of honour over the financial ceiling of category 4.”. In the first evaluated model, this is translated as “ Het Parlement en de Raad hebben Kosovo in een zaak van de honour van rubriek 4 turned”. In this segment, both ‘honour’ and ‘turned’ are not translated, both of which which are counted as ‘mistranslation/not translated’ errors. In the second model’s translation, “Het Parlement en de Raad hebben Kosovo in een zaak van de financiële bovengrens van categorie 4 turned”, ‘honour’ is completely left out, replacing the ‘mistranslation’ error with an ‘omission’ error. ‘Turned’ is still not translated in this version. However, in the third model’s translation ‘honour’ is translated , leading to the the following sentence: “Het Parlement en de Raad hebben Kosovo in een kwestie van eer laten zien over het financiële plafond van categorie 4.” Although this is stylistically still not the best possible translation of ‘affair of honour’, it is technically correct and in this case not counted as an error. In the third model’s output, ‘have turned Kosovo into’ is translated as ‘hebben Kosovo laten zien over’, which as it is translated, is no longer counted as a ‘mistranslation/not translated’ error but is a ‘language’ type error, as it is a grammatically and semantically incorrect representation of the ST.

There are several instances where the translation by the second model is an improvement on the first, but the third model's translation is reverted to the first instead of keeping the second, better translation. Take for example, the ST phrase "is reason enough to vote against it", which the first model translates as follows: "is daarom voldoende om tegen te stemmen". This is a clear omission of 'reason' in the translation. The second model translates this phrase as "en dat is een reden om tegen te stemmen", which in turn omits 'enough', but this is not counted as an error. Even with the omission this translation is acceptable and contains no errors according to the used evaluation system, although a stricter system could count the omission of 'enough' as an error. The third translation, however, is the same as the first: "en is daarom voldoende om tegen te stemmen", which again omits the 'reason' and is marked as an omission error. There are also cases in which the translation output provided by the first model is the best quality output, and deteriorates over the following models. An example of this is the following sentence (which is not in the evaluated passage):

Source: "One of the hunger strikers, after 44 days of fasting, has today been transferred to the Hôtel de Dieu Hospital in Paris in extremely poor health after losing 24 kilos.

Model 1: "Een van de hunger na 44 dagen van fasting, is vandaag overgedragen naar de Hôtel de Dieu in Parijs in zeer slechte gezondheid na het verlies van 24 kilos."

Model 2: “Een van de hunger na 44 dagen van fasting, is vandaag overgedragen naar de Hôtel de Dieu in Parijs in een zeer slechte gezondheid na 24 kilos.”

Model 3: “Een van de hunger na 44 dagen van fasting, is vandaag overgedragen naar de Hôtel de Dieu in Parijs in de Hospital health in de health van 24 kilos.”

Although there are several errors in the first translation (two mistranslation errors: ‘hunger’ and ‘fasting’ are not translated; two language errors: ‘has been transferred’ is translated as ‘is overgedragen naar’ when it should be ‘is overgedragen aan’ or ‘is overgebracht naar’, and ‘24 kilos’ should be ‘24 kilo’ in Dutch, and finally an omission error where ‘Hospital’ is not translated) it is clearly the best of the three options. The second and third models produce progressively lower quality output. The second translation, in addition to the same errors in the first sentence, has an additional omission error where ‘losing’ is not translated. The third translation has additional mistranslation and language errors in the following phrase, which is of such poor quality that it is very difficult to identify specific errors rather than counting the whole phrase as a critical error: “in de Hospital health in de health”.

Idioms are “linguistic expressions which are typical for a language and specific to a single culture” which makes them a very difficult aspect of (human) translation (Adelnia & Dastjerdi, 2011, p. 879) The translation of idioms relies on an understanding of the cultural and social norms that underlie

its meaning, and on the ability to transfer that cultural impact from one culture to another. Machine translation systems lack this understanding and are in most cases only able to translate idioms literally, in which case they usually lose their meaning. In an evaluation of rule-based machine translation systems, phrase-based statistical machine translation systems and neural machine translation systems by Burchardt et al. (2017), only the phrase-based systems were able to correctly learn to translate certain idioms. As such, it comes as no surprise that the translation of idioms by the translation model in this thesis produces very poor quality output. In the evaluated passage, the source sentence "But this is not in the offing as yet." is translated as "Dat is echter nog niet in de offing" by the first translation model. The second model translated it as ""Maar dit is nog niet in de offing", while the output of the third model is the same as that of the first. All of these are counted as mistranslation errors due to the fact that 'offing' has not been translated. Another example where an idiom is not translated, is the following sentence, which is not in the evaluated passage: "Robbing Peter to pay Paul has never been an indication of great political imagination.". It is represented as "Robbing Peter to Paul" by all three models. In the reference text, the corresponding sentence is "aan de ene kant weghalen wat je aan de andere kant uitgeeft is nog nooit een teken van grote politieke creativiteit geweest". While it is not unexpected that the idiom in this sentence has not been translated correctly or in this case at all, I would have expected the rest of the sentence to have been translated word-for-word as that

is often how machine translation systems handel idioms. According to Wu et al. (2016) neural machine translation models "sometimes produce output sentences that do not translate all parts of the input sentence" (p. 2), evidence of which can also be seen in the earlier examples and in the detailed error evaluation in Appendix B.

There are many more possible examples to be found in the translation predictions, but the ones discussed in this sections are singled out to give a good overview of some different types of errors in the model translations as well as illustrate different possible issues with neural machine translation. Although most of this section has focused on translation errors, it has to be said that there are also many passages in the model output where the translation is of a perfectly acceptable quality and would not require further post-editing to be suitable for publication. These are likely sections that are very similar to sections included in the training data, which would give the system an advantage during their translation. Despite several sections of very poor quality, the output quality is overall a better quality than expected based on the limited training parameters.

4.2 Technological and Methodological Problems

As expected, training the corpus took slightly longer than the 142 hours projected due to the limitations of Google Colab and the number of times training was interrupted by being disconnected from their GPU. However,

training was completed in approximately 160 to 170 hours. Before starting training on the English-Dutch corpus, I trained a demo using the English-German Europarl demo dataset provided in the OpenNMT toolkit to test training speed and identify potential issues in advance of training the final model. The demo finished training in a roughly 5 hours. Due to the small size of the dataset, translation predictions were expected to be of a low quality. However, when testing the system it turned out that the model had somehow not been trained properly. Because of this, every sentence in the test demo had been translated as only "Die ist", while no other words in the sentences had been translated. Although it is still unclear what exactly caused this issue, there seemed to be a training loop that reset every 150 steps. This was causing the training to start over from scratch after every 150 completed steps and made it impossible to advance. Despite not knowing the exact cause, research into similar issues led to the advice to add 'pool_ factor 10' to the training command to stop this training loop from occurring. Testing this out by training a new demo model proved to work. The output produced by the new demo model is still not of a very good quality, but this was to be expected due to the small dataset. Training the demo prior to training the real model turned out to be very helpful in this case, seeing as otherwise the actual translation model would have gone through several days of training with no end result.

5. Conclusion and Discussion

This thesis offers a theoretical overview of the history and development of machine translation systems, discusses several possible applications of machine translation and examines the most commonly used types of systems for machine translation, with special attention to statistical machine translation systems and neural machine translation systems. It also discusses several automatic evaluation metrics for machine translation research. The knowledge gained in the theoretical overview is then put into practice though the training of a basic, domain-specific neural machine translation model using the English-Dutch Europarl corpus. Training parameters for the model were very basic, which led to relatively poor quality output and a low rate of improvement, as shown by BLEU scores and detailed human evaluation of a randomly selected section of the test file. Of course, human evaluation is subjective and the output was only evaluated by one person. To get a more objective judgement of the quality of this system, it would have to be scored by multiple evaluators on the basis of a larger amount of output. This research has pointed out several recurring errors in the output produced by the translation models. As expected, idioms and figurative language prove to be impossible to translate properly with the sort of system trained in this thesis. The system also proved to at times revert back to output produced by earlier checkpoints at later stages in training, despite the fact that this output is less accurate than output produced by the previous checkpoint model. This points to a need to constantly

evaluate the translation model during training to be able to adjust training parameters to make the model choose the better translation options. At several points not all words in a sentence are translated. This is, however, a common issue in neural machine translation (Wu et al., 2016). Output produced by this model is of a rather poor quality in general and significant post-editing would be necessary for it to reach a publishable quality, although there are sections that are already of an acceptable quality. Machine translation is often thought of as an alternative to human translation. However, as this research shows, developing more than a very basic machine translation system requires a lot of resources, time and expertise and a basic system is not able to produce translations that come close to expert human translation in terms of quality. Although machine translation system are rapidly advancing in recent years, they are not likely to reach the same quality as human translation in the near future. Rather than replacing the need for human translation, it is more likely that the use of machine translation will create an additional need for post-editing and related activities in the translation sphere. In addition, different types of text are more or less suited for machine translation. It is highly unlikely that poetry, for example, will be a good candidate for machine translation unless MT systems improve drastically in several areas, such as the translation of culture-specific elements and handling different word orders. A further exploration of the relationship between human translation and machine translation and its possible future development, as well as translators' current

attitudes towards machine translation would be an interesting avenue for future research.

There are several other interesting avenues for possible future research based on this project. Seeing as the model training capped out at 300,000 steps, it would be interesting to see if this can be related back to corpus size. In a 50 million word corpus, 3/10th corresponds to 15 million words. In additional research, it would be interesting to use 15 million words of the Europarl corpus to train a neural machine translation model, using the same parameters used to train the model in this thesis, to see whether it would reach the same translation quality as the model using the complete 50 million word corpus reached. This translation quality could be measured by translating the same test text with the two models and calculating BLEU scores for both. Undoubtedly, a 15 million word corpus would have a smaller unique vocabulary than a 50 million word corpus, which would most likely influence translation quality. Seeing as this is only a small amount of research and it is closely connected to the current project, it would have been a good addition to this research. Due to the time involved in training a neural network, it was unfortunately not feasible at this time. In a broader sense, this opens up questions about the relation between corpus size and the quality of neural machine translation systems. Traditionally, neural translation models need large amounts of data to be able to produce good quality output and are expected to increase in quality as they undergo more extensive training. However, research into neural machine

translation for language pairs with access to small amounts of bilingual data is devising ways to substantially improve neural translation output with smaller amounts of data. One way is described in research done by Zoph et al. (2016), which uses a transfer learning method to "train a high-resource language pair [...] then transfer some of the learned parameters to the low-resource pair" (par. 1).

Another interesting and currently relevant avenue of further research, would be to compare the neural translation model trained in this research with a statistical model trained using the same data. Although BLEU scores have been shown not to be of much use when comparing fundamentally different machine translation models, given enough time it is possible to compare the translation output of both systems in detail through human evaluation. As such, comparing a neural and statistical machine translation system in such a manner would be better suited to an independent research project. Neural machine translation is a rather recent development and is currently starting to compete with state-of-the-art statistical machine translation systems. This has led to a large amount of interest in comparing the two in various ways. Some examples of research in this direction are studies by Bentivogli et al. (2016), Castilho et al. (2017) and Shterionov et al. (2019). Although several studies have already been done on the comparison of neural and statistical machine translation, the results are not conclusive: which system outperforms the other does not always correspond between studies and is often based on the exact

parameters and data used in the research. In addition, research on this subject often compares the two types of systems using BLEU scores, which have been shown not to be useful for the comparison of different types of translation systems (Callison-Burch et al., 2006). As such, more research on this subject can be useful for further development of neural machine translation systems and to gain a better understanding of the specific situations in which neural or statistical machine translation might function better than the other.

References

Adelnia, A., & Dastjerdi, H. V. (2011). Translation of idioms: a hard task for the translator. *Theory and Practice in Language Studies, 1*(7), 879-883.

Agarwal, A., & Lavie, A. (2008, June). Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation* (pp. 115-118). Association for Computational Linguistics.

ALPAC. (1966). *Language and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee.* Washington, DC: National Academy of Sciences.

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631.*

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics, 16*(2), 79-85.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J. T., & Williams, P. (2017). A linguistic evaluation of rule-based, phrase-based, and neural

MT engines. *The Prague Bulletin of Mathematical Linguistics, 108*(1), 159-170.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. 249-255.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art?. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109-120.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 263-270). Association for Computational Linguistics.

Costa-Jussà, M. R., Farrús, M., Mariño, J. B., & Fonollosa, J. A. (2012). Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and informatics, 31*(2), 245-270.

CUDA Zone. (2019, April 30). Retrieved from https://developer.nvidia.com/cuda-zone

Daniel Brockmann and Matthias Heyn look back on 30 years of TRADOS. (n.d.) [Video File]. Retrieved from https://www.sdltrados.com/about/history.html

Google Colaboratory. (n.d.). *Welcome to Colaboratory!* Retrieved from https://colab.research.google.com/

Govindarajalu, B. (2010). C*omputer Architecture and Organisation, 2E*. Tata McGraw-Hill Education.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Liu, S. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Hutchins, J. (1993). Latest developments in machine translation technology: beginning a new era in MT research. *MT Summit (1993)*, 11-34.

---. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice, 13*, 29-70.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River: Pearson/Prentice Hall.

---. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River: Pearson/Prentice Hall.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2018). Opennmt: Open-source toolkit for neural machine translation. *Proceedings of AMTA 2018, vol. 1: MT Research Track*. 177-184.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Compeutational Linguistics on Human Language*

Technology-Volume 1 (pp. 48-54). Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit*, *5*, 79-86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177-180).

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, *25*, 131-148.

Krauwer, S. (1999). EUROTRA(UMA?)The First Crusade against the Multilinguality Problem (1978-1993) [PowerPoint Slides]. Retrieved from http://www-sk.let.uu.nl/ond/eurotrauma-en_files/v3_document.htm

Landsbergen, J. (1989). The Rosetta Project. *Proceedings of MT Summit II*. Munich, 82-87.

Nord, C. (2014). *Translating as a purposeful activity: Functionalist approaches explained*. Routledge.

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on*

Empirical Methods in Natural Language Processing and Very Large Corpora. 20-28.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Raw, A., Vandecapelle, B., & Van Eynde, F. (1988). Eurotra: an overview. *Interface, 3*(1), 5-32.

Russo-Lassner, G., Lin, J., & Resnik, P. (2005). *A paraphrase-based approach to machine translation evaluation* (No. LAMP-TR-125). MARYLAND UNIV. COLLEGE PARK INST. FOR ADVANCED COMPUTER STUDIES.

Shterionov, D., Casanellas, P. N. L., Superbo, R., & O'Dowd, T. (2017). Empirical evaluation of NMT and PBSMT quality for large-scale translation production. *20th Annual Conference of the European Association for Machine Translation*. n.p.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation, 32*(3), 217-235.

Slocum, J. (1985). A survey of machine translation: its history, current status, and future prospects. *Computational linguistics, 11*(1), 1-17.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, *200*(6), 223-231.

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation, 23*(2-3), 117-127.

Thouin, B. (1982). The METEO system. *Practical experience of machine translation*, 39-44.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Lrec* (Vol. 2012, pp. 2214-2218).

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432.*

Vanmassenhove, E., Du, J., & Way, A. (2016). Improving subject-verb agreement in SMT. *ADAPT Centre/School of Computing, Dublin City University*, 1-13.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 523-530).

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201.*

Appendix A: 200-word abstract for detailed evaluation

Source
This situation is all the more disgusting in that, at the same time, we are paying EUR 45 million to ACP banana producers, some of which are American multinationals established in Cameroon and the Ivory Coast and which are being paid EUR 45 million under this European budget.
This, then, is stupid spending on top of customer spending on top of ideological spending.
It all adds up to a lot, it adds up to much too much and that is reason enough to vote against it.
Mr President, if there is one issue on which the Council and Parliament should follow the same line of action, then surely it is the reconstruction of Kosovo.
But this is not in the offing as yet.
Parliament and the Council have turned Kosovo into an affair of honour over the financial ceiling of category 4.
This is taken to such an extreme that the rapporteur even threatened to include only EUR 115 million into the budget instead of the EUR 500 million which is required.
If this were to happen, then the Kosovars would be the big losers.
The only lifeline which then remains is a supplementary and amending budget later on in the year.

Reference
Die situatie is des te weerzinwekkender omdat wij ondertussen 45 miljoen uitkeren aan de producenten van ACS-bananen, waarvan sommigen Amerikaanse multinationals zijn die zich in Kameroen en in Ivoorkust gevestigd hebben en die nu 45 miljoen ontvangen uit die Europese begroting.
Dit zijn gewoonweg dwaze uitgaven die bovenop uitgaven voor vriendjes komen, die bovenop ideologische uitgaven komen.
Dat is veel, veel te veel van het goede en voldoende reden om tegen te stemmen.
Voorzitter, als er één onderwerp is waarbij Raad en Parlement eensgezind zouden moeten optreden is het wel de wederopbouw van Kosovo.
Maar tot nog toe is daar geen sprake van.
Parlement en Raad hebben Kosovo onderwerp gemaakt van een prestigestrijd over het financiële plafond van categorie 4.
Die gaat zo ver dat de rapporteur zelfs dreigde om slechts 115 miljoen euro in de begroting op te nemen terwijl 500 miljoen euro nodig is.
Als dat zou gebeuren, zijn de Kosovaren de grote verliezers.

De enige strohalm die dan overblijft is een aanvullende en gewijzigde begroting, later in het jaar.

Model 100,000
Deze situatie is des te disgusting terwijl we tegelijkertijd 45 miljoen euro betalen voor de bananenproducenten in de ACS-landen, waarvan sommige Amerikaanse multinationals zijn gevestigd in Kameroen en Ivoorkust en die 45 miljoen euro onder deze Europese begroting worden betaald.
Dit is dus dom voor de uitgaven van de klant boven de ideological
Het voegt veel toe aan een veel te veel toegevoegde waarde en is daarom voldoende om tegen te stemmen.
Mijnheer de Voorzitter, als er één kwestie is waarover de Raad en het Parlement dezelfde lijn moeten volgen, dan is het toch de wederopbouw van Kosovo.
Dat is echter nog niet in de offing
Het Parlement en de Raad hebben Kosovo in een zaak van de honour van rubriek 4 turned
Dit is een zo extreem feit dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting heeft opgenomen in plaats van de 500 miljoen euro die nodig is.
Als dit zou gebeuren, zouden de Kosovaren de grote verliezers zijn.
De enige lifeline blijft een aanvullende en aanvullende begroting later in het jaar.

Model 200,000
Deze situatie is des te disgusting terwijl we tegelijkertijd 45 miljoen euro betalen voor de bananenproducenten in de ACS-landen, waarvan sommige Amerikaanse multinationals zijn gevestigd in Kameroen en Ivoorkust en die 45 miljoen euro onder deze Europese begroting worden betaald.
Dit is dus dom voor de uitgaven van de klant boven de ideological
Het voegt veel te veel toe aan een veel te veel en dat is een reden om tegen te stemmen.
Mijnheer de Voorzitter, als er één kwestie is waarover de Raad en het Parlement dezelfde lijn moeten volgen, dan is het toch de wederopbouw van Kosovo.
Maar dit is nog niet in de offing
Het Parlement en de Raad hebben Kosovo in een zaak van de financiële bovengrens van categorie 4 turned
Dit is een zo extreem feit dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting heeft opgenomen in plaats van de 500 miljoen euro die nodig is.
Als dit zou gebeuren, zouden de Kosovaren de grote verliezers zijn.
De enige lifeline blijft een aanvullende en aanvullende begroting later in het jaar.

Model 300,000
Deze situatie is des te disgusting terwijl we tegelijkertijd 45 miljoen euro betalen voor de bananenproducenten in de ACS-landen, waarvan sommige

Amerikaanse multinationals zijn gevestigd in Kameroen en Ivoorkust en die 45 miljoen euro onder deze Europese begroting worden betaald.
Dit is dus dom voor de uitgaven van de klant boven de ideological
Het voegt veel toe aan een veel te veel toegevoegde waarde en is daarom voldoende om tegen te stemmen.
Mijnheer de Voorzitter, als er één kwestie is waarover de Raad en het Parlement dezelfde lijn moeten volgen, dan is het toch de wederopbouw van Kosovo.
Dat is echter nog niet in de offing
Het Parlement en de Raad hebben Kosovo in een kwestie van eer laten zien over het financiële plafond van categorie 4.
Dit is een zo extreem feit dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting heeft opgenomen in plaats van de 500 miljoen euro die nodig is.
Als dit zou gebeuren, zouden de Kosovaren de grote verliezers zijn.
De enige lifeline blijft een aanvullende en aanvullende begroting later in het jaar.

Appendix B: Detailed error evaluation results

Changes in output compared to the last model are highlighted in yellow.

Reference	Source	Target 1	Type	Target 2	Type	Target 3	Type
Die situatie is **des te weerzinwekkender** omdat wij ondertussen 45 miljoen uitkeren	This situation is **all the more disgusting** in that, at the same time, we are paying EUR 45 million	Deze situatie is **des te disgusting** terwijl we tegelijkertijd 45 miljoen euro betalen	1	Deze situatie is **des te disgusting** terwijl we tegelijkertijd 45 miljoen euro betalen	1	Deze situatie is **des te disgusting** terwijl we tegelijkertijd 45 miljoen euro betalen	1
Die situatie is des te weerzinwekken der **omdat wij ondertussen** 45 miljoen uitkeren	This situation is all the more disgusting **in that, at the same time**, we are paying EUR 45 million	Deze situatie is des te disgusting **terwijl we tegelijkertijd** 45 miljoen euro betalen	3	Deze situatie is des te disgusting **terwijl we tegelijkertijd** 45 miljoen euro betalen	3	Deze situatie is des te disgusting **terwijl we tegelijkertijd** 45 miljoen euro betalen	3
Dit zijn gewoonweg **dwaze uitgaven** die bovenop uitgaven voor vriendjes komen, die bovenop ideologische uitgaven komen.	This, then, is **stupid spending** on top of customer spending on top of ideological spending.	Dit is dus **dom voor de uitgaven** van de klant boven de ideological	3	Dit is dus **dom voor de uitgaven** van de klant boven de ideological	3	Dit is dus **dom voor de uitgaven** van de klant boven de ideological	3
Dit zijn gewoonweg dwaze uitgaven die **bovenop** uitgaven voor vriendjes komen, die bovenop ideologische uitgaven komen.	This, then, is stupid spending **on top of** customer spending on top of ideological spending.	Dit is dus dom voor de uitgaven van de klant **boven** de ideological	3	Dit is dus dom voor de uitgaven van de klant **boven** de ideological	3	Dit is dus dom voor de uitgaven van de klant **boven** de ideological	3

Dit zijn gewoonweg dwaze uitgaven die bovenop **uitgaven voor vriendjes** komen, die bovenop ideologische uitgaven komen.	This, then, is stupid spending on top of **customer spending** on top of ideological spending.	Dit is dus dom voor de uitgaven **van de klant** boven de ideological	4	Dit is dus dom voor de uitgaven **van de klant** boven de ideological	4	Dit is dus dom voor de uitgaven **van de klant** boven de ideological	4
Dit zijn gewoonweg dwaze uitgaven die bovenop uitgaven voor vriendjes komen, die bovenop **ideologische uitgaven** komen.	This, then, is stupid spending on top of customer spending on top of **ideological spending**.	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	1	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	1	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	1
Dit zijn gewoonweg dwaze uitgaven die bovenop uitgaven voor vriendjes komen, die bovenop **ideologische uitgaven** komen.	This, then, is stupid spending on top of customer spending on top of **ideological spending**.	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	4	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	4	Dit is dus dom voor de uitgaven van de klant boven **de ideological**	4
Dat is veel, veel te veel van het goede	It all adds up to a lot, it adds up to much too much	Het voegt veel toe aan een veel te veel toegevoegde waarde	3	Het voegt veel te veel toe aan een veel te veel	3	Het voegt veel toe aan een veel te veel toegevoegde waarde	3
voldoende **reden** om tegen te stemmen.	is **reason** enough to vote against it.	is daarom **voldoende** om tegen te stemmen.	4	en dat is een **reden** om tegen te stemmen	n/a	en is daarom **voldoende** om tegen te stemmen.	4
eensgezind zouden moeten optreden	should follow the same line of action	dezelfde lijn moeten volgen	3	dezelfde lijn moeten volgen	3	dezelfde lijn moeten volgen	3

Maar tot nog toe is daar geen sprake van.	But this is not in the **offing** as yet.	Dat is echter nog niet in de **offing**	1	Maar dit is nog niet in de **offing**	1	Dat is echter nog niet in de **offing**	1
Parlement en Raad hebben Kosovo onderwerp gemaakt van een **prestigestrijd** over het financiële plafond van categorie 4.	Parliament and the Council have turned Kosovo into an **affair of honour** over the financial ceiling of category 4.	Het Parlement en de Raad hebben Kosovo in een **zaak van de honour** van rubriek 4 turned	1	Het Parlement en de Raad hebben Kosovo in **een zaak** van de financiële bovengrens van categorie 4 turned	4	Het Parlement en de Raad hebben Kosovo **in een kwestie van eer** laten zien over het financiële plafond van categorie 4.	n/a
Parlement en Raad **hebben Kosovo onderwerp gemaak**t van een prestigestrijd over het financiële plafond van categorie 4.	Parliament and the Council **have turned Kosovo into** an affair of honour over the financial ceiling of category 4.	Het Parlement en de Raad **hebben Kosovo** in een zaak van de honour van rubriek 4 **turned**	1	Het Parlement en de Raad **hebben Kosovo** in een zaak van de financiële bovengrens van categorie 4 **turned**	1	Het Parlement en de Raad **hebben Kosovo** in een kwestie van eer **laten zien** over het financiële plafond van categorie 4.	3
Parlement en Raad hebben Kosovo onderwerp gemaakt van een prestigestrijd over het **financiële plafond van categorie 4.**	Parliament and the Council have turned Kosovo into an affair of honour over **the financial ceiling of category 4**.	Het Parlement en de Raad hebben Kosovo in een zaak van de honour van **rubriek 4** turned	4	Het Parlement en de Raad hebben Kosovo in een zaak van **de financiële bovengrens van categorie 4** turned	n/a	Het Parlement en de Raad hebben Kosovo in een kwestie van eer laten zien over **het financiële plafond van categorie 4.**	n/a
dat de rapporteur zelfs **dreigde** om slechts 115 miljoen euro in de begroting **op te nemen**	that the rapporteur even **threatened to include** only EUR 115 million into the budget	dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting **heeft opgenomen**	3	dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting **heeft opgenomen**	3	dat de rapporteur zelfs nog maar 115 miljoen euro in de begroting **heeft opgenomen**	3

De enige **strohalm** die dan overblijft is een aanvullende en gewijzigde begroting, later in het jaar.	The only **lifeline** which then remains is a supplementary and amending budget later on in the year	De enige **lifeline** blijft een aanvullende en aanvullende begroting later in het jaar.	1	De enige **lifeline** blijft een aanvullende en aanvullende begroting later in het jaar.	1	De enige **lifeline** blijft een aanvullende en aanvullende begroting later in het jaar.	1
De enige strohalm die dan overblijft is een **aanvullende en gewijzigde** begroting, later in het jaar.	The only lifeline which then remains is a **supplementary and amending budget** later on in the year	De enige lifeline blijft een **aanvullende en aanvullende begroting** later in het jaar.	2	De enige lifeline blijft een **aanvullende en aanvullende begroting** later in het jaar.	2	De enige lifeline blijft een **aanvullende en aanvullende begroting** later in het jaar.	2