Comparing solutions for the OOV word problem in Neural Machine Translation

Sem Meereboer

ANR 759310

Master's thesis

Communication and Information Sciences

Human Aspects of Information Technology

School of Humanities

Tilburg University, Tilburg

Supervisor: Dr. M.M. van Zaanen

Second reader: Dr. G.A. Chrupala

July 2016

Abstract

Machine Translation is an increasingly more used method to translate large corpora of text, because of its time and cost efficiency. Neural Machine Translation (NMT) is a relatively new approach to Machine Translation. A weakness in any machine translation system is its ability to translate unknown or Out-Of-Vocabulary (OOV) words, those that do not occur in the training data. Nematus is an upcoming NMT system that has a built-in model to deal with the OOV word problem. Nematus' script is slightly adjusted for different models that deal with the OOV word problem. The new models either leave the OOV word out or retain the OOV word in its source language. The test set was also manually translated by Tilburg University students. It is hypothesized that the default method of Nematus for dealing with OOV words performs better than the adjustments, but worse than manually translated texts. Participants in a survey (N = 42) confirm these hypotheses, with manually translated texts scoring significantly higher. BLEU scores have been calculated, yet do not provide confirmation to our hypotheses. The test set for this study are 1756 sentences that originate from a project called TraMOOC (Translation for Massive Open Online Courses), funded by the European Commission.

Keywords: NMT, OOV, out-of-vocabulary, Nematus, Neural Machine Translation, MOOC, TraMOOC.

Table of Contents

1. Introduction

Machine translation (MT) investigates the use of software to translate text or speech from a source language into another. The first set of proposals to machine translation was presented by Warren Weaver back in 1949 (Hutchins, 2007). It became of practical use in the Cold War during the 1950s, where human translators were scarce and hard to trust. Both the Soviet Union and the United States made use of relatively simple MT systems that translated intercepted messages of their adversaries. However, the grand expectations of that time were not fulfilled, mainly because machines did not have the required computing power yet. Another challenge was the existence of double meanings of ambiguous words. The infamous ALPAC report of 1966, that claimed there was no future for MT, seized almost all funding for further research and thereby delayed any progress for a long period of time (Hutchins, 1986). In the late 1980s, IBM's Research Center continued to investigate the possibilities of machine translation (MT).

Machine translation consists of multiple subdivisions, but in recent years Statistical Machine Translation (SMT) is one of the most profound methods in machine translation. Online machine translation services like Google Translate and Bing Translator exist because of SMT. Their strength is the ability to train models on an enormous database of text corpora, namely bilingual sources on the World Wide Web. More recently, academics have started to combine SMT with neural networks or deep learning (Deselaers, Hasan, Bender & Ney, 2009). This was the start of Neural Machine Translation (NMT), that has been gaining popularity.

Unfortunately, current Statistical and Neural Machine Translation still face numerous challenges (Arnold, 2003). One example is ambiguous translations. Think of the Dutch word "leren", which may be translated to either "to learn" or "to teach", depending on the context.

Another challenge is that word forms or phrases that do not appear in the training data cannot be translated. This problem is referred to as out-of-vocabulary (OOV) words.

This thesis contributes to a project called TraMOOC (Translation for Massive Open Online Courses), funded by the European Commission. The TraMOOC project begun in February 2015 and will seize its activities in January 2018. It unites 10 different partners in 6 European countries. Its main focus is to provide a high quality machine translation service for all types of educational textual data available. MOOCs are online educational courses in a range of subjects that one can follow. The MOOC-platform aims to enable integration of machine translation solutions into the educational domain, thereby complementing the current system used for translating MOOC data, namely crowdsourcing.

In crowdsourcing, certain tasks are outsourced to the masses. Crowdsourcing is a clever way of translating large corpora of text, although it may cost a lot of time and money (Zaidan & Callison-Burch, 2011).

The MOOC data serves as a test set for this thesis. A characteristic of MOOC data is that the courses offered uses domain-specific terminology. This leads to the use of uncommon vocabulary, resulting in a low term frequency for certain words. Words with a low term frequency are generally harder to translate for SMT and NMT systems, because there are fewer examples to train the model on. Since the training data differs from the MOOC data, the selection of this test set provides ample OOV words.

Several solutions have been presented in the past, such as the Transliteration Model and external bilingual dictionary approach. These approaches have not been tested on English ↔ Dutch machine translation however, nor have they been tested on corpora in the educational field (MOOC data).

Nematus (Sennrich, Haddow & Birch, 2015) is a NMT system that uses neural networks to solve the OOV problem. In this thesis, we have made two adjustments to the *default* method. In the first adjustment, OOV words are dropped thus creating the *drop* method. In the second adjustment, OOV words are kept in the source language, creating the *keep* method. Quality of translations are investigated by using the adequacy/fluency model, that aims to measure the faithfulness to a source language and the correctness of the translated texts. Also BLEU scores were calculated.

The research question in this thesis is: how do our methods compare regarding the OOV word problem? The following hypotheses are tested to answer this question:

(H1a) The *default* method of Nematus provides higher adequacy in translation than the *drop* and *keep* methods.

(H1b) The *default* method of Nematus provides higher fluency in translation than the *drop* and *keep* methods.

(H2a) The manual translation provides higher adequacy in translation than Nematus.

(H2b) The manual translation provides higher fluency in translation than Nematus.

This thesis is organized as follows. The next section portrays related work on the subject of machine translation, out-of-vocabulary words and its existing solutions. Section 3 describes the three methods used in order to find an effective solution to the OOV-problem. Section 4 presents the results of our methods, which will be discussed in Section 5.

2. Background

This section describes previous literature on Statistical and Neural Machine Translation, the out-of-vocabulary problem and evaluation metrics for machine translation. We will discuss MOOC data to show what is precisely meant with this subject. Important terms and concepts coined in the introduction will also be clarified in this section.

2.1 Machine translation

Machine translation is software-driven translation of text or speech of a source language into a target language. These paragraphs provide a closer look at classical approaches for machine translation. In the second paragraph, Statistical Machine Translation is elaborated on. The third paragraph will introduce Neural Machine Translation. The last paragraph will describe the machine translation system that is used in this thesis, Nematus.

2.1.1 Approaches to MT

When looking at the history of machine translation, there are three classical approaches (Hutchins, 1986; Jurafsky & Martin, 2014; Nirenburg, Carbonell, Tomita & Goodman, 1994) used by linguists. In word-based or direct translation, every source word is directly translated onto the target source text using very little intermediate structures. In transfer approaches, rules are applied to transform the source language parse structure into a target language parse structure. Parse structures describe syntactic constructions or sentences. Interlingua approaches analyze source language texts into abstract meaning representations. The target language is generated from these representations accordingly. The general concepts of MT are often visualized in the Vauquois-triangle, as seen in Figure 1.

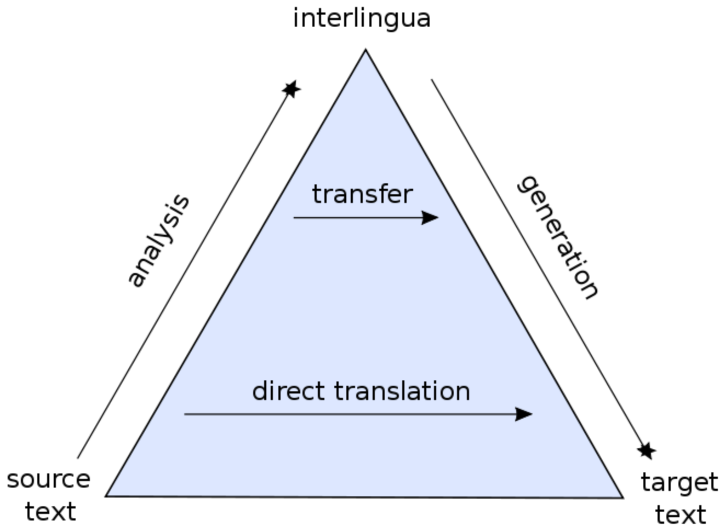

Figure 1. The Vauquois-triangle, often used to show the classical approaches.

Problems with any type of translation arise when a language uses culture-specific concepts such as metaphors or any other constructions without an exact parallel in the other language.

2.1.2 Statistical Machine Translation

In SMT, the analysis of bilingual text corpora creates a model that is able to translate a new text corpus. These models predict a translation for a new phrase or text. The quality of translations increases when the text corpora are larger. An apparent benefit of SMT is its cost and time efficiency compared to manual translation.

When statistical methods proved their use for automatic speech recognition and natural language processing, Brown et al. (1990) applied these methods in machine translation, resulting in the birth of SMT. This method requires fewer linguistic features and focuses more on creating statistical models. Through analyzing bilingual text corpora, it calculates the most probable translation outcome. If we use faithfulness to the source input and fluency in the target output as our quality metrics for probability, the translation from source language sentence *S* to target language sentence *T* can be modelled as follows (Brown et al., 1990; Jurafsky & Martin, 2014):

$$\text{best-translation } \hat{T} = \frac{\text{argmax}}{T} \text{ faithfulness}(T,S) \text{ fluency}(T)$$

All statistical translation models are based on the idea of word alignment. A word alignment is a mapping between source words and target words in a set of parallel sentences (Jurafsky & Martin, 2014). Figure 2 shows a visual representation of word alignment. Word alignment is important, because languages differ in terms of syntax structure.

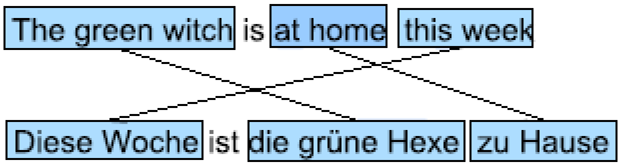

Figure 2. A statistical translation, showing phrasal reordering and word alignment.

An engine that focuses on scientific research into SMT is Moses (Koehn et al., 2007). Moses enables users to build their own statistical models that help translate large textual corpora.

2.1.3 Neural Machine Translation

A relatively new method in machine translation is the implementation of neural networks. A neural network is a system that consists of a large amount of processors operating in parallel. Usually the neural networks are patterned like the neurons of a human brain, hence the ‘neural’ adjective (Sutskever, Vinyals, & Le, 2014). A neural network is trained by feeding it large amounts of data. Neural networks can adapt, by modifying themselves as they learn from every new training run.

Neural networks are not new, but Deselaers, Hasan, Bender and Ney (2009) started using them in combination with machine translation, setting the cornerstone for NMT. The focus of NMT is to design a fully trainable model that adjusts for every component the training corpora provides, instead of relying on pre-designed features, in order to maximize its translation performance (Bahdanau, Cho & Bengio 2014). Basically, an NMT system is a large neural network that reads the entire source sentence and produces output one word at a time (Thang, Sutskever, Le, Vinyals & Zaremba (2014). NMT systems are well-suited to any translation problem because it uses minimal domain knowledge (Sutskever, Vinyals, & Le, 2014).

Nematus (Sennrich, Haddow & Birch, 2015, 2016) is the MT system used in this thesis. It is roughly based on SMT systems, but has additional features in the form of neural networks. One such feature is the BPE (byte pair encoder) package, which transliterates unknown words based on an open vocabulary. BPE (Gage, 1994) is a data compression algorithm that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. BPE is Nematus' built-in solution to the fixed vocabulary problem, which is the root of the OOV word problem.

2.2 Out-of-vocabulary words

This section will describe how OOVs are defined in the literature and how they are being differentiated. Also, this section will look at existing approaches for handling OOV words in machine translation.

2.2.1 Different types of OOV

Whether or not a word is out of vocabulary or unknown to an MT model, depends on the training data that is used for a model. It may be assumed that the more training data is used, fewer words are unknown to the MT model. OOV words are usually named entities, technical terms, compounds, misspelled words or foreign words that cannot be translated to the target (Durrani, Sajjad, Hoang and Koehn, 2014).

Researchers (Aminian, Ghoneim & Diab, 2014; Arora et al., 2008) argue that the problem of OOV words is twofold and make a clear distinction between two types of OOV: Completely Out-of-Vocabulary (COOV) and Sense Out-of-Vocabulary (SOOV). COOV words are completely absent in the training data. SOOV words are observed in the training data but with a different usage or sense, different from that of the test data occurrence. Aminian, Ghoneim and Diab (2015) show that the problem of SOOVs can be tackled using a word sense disambiguator (WSD) in a pre-processing phase of machine translation.

MOOC data are in-depth courses on different subjects that deviate with respect to content and possibly the training data in vocabulary. For this reason, it can be assumed that MOOC data may contain words that are absent in the training data. Exploratory analysis revealed this assumption to be true. Therefore, we mainly address COOVs in this thesis.

2.2.2 Approaches for handling OOV words

There have been numerous efforts in dealing with OOV words in Statistical Machine Translation, somewhat less in Neural Machine Translation.

For SMT, Paul and Sumita (2008) show that a lexical approximation method (LA) and phrase-table extension method (PTE) can be successfully combined to handle OOV words. In the LA method, OOV words are determined by translating the source text into a target text. Then, all OOV words in the source text are replaced with appropriate word variants that are found in the training corpus, thus reducing the amount of OOV words in the input. The SMT model is extended by adding new phrase translations for all source language words that are completely out of vocabulary (COOV) in the original phrase-table, but only appear in the context of larger phrases. These methods were tested for translations of Hindi ↔ Japanese, languages that both have a rich morphology (Arora, Paul & Sumita, 2008).

Similar to Arora, Paul and Sumita, Huang et al. (2011) argue that the problem of OOV could be better handled if a pre-processing model recognized and translated the constituents of an OOV word. The model they propose constrains the choices of sublexical translations and eliminates unlikely ones by analyzing a collection of monolingual and bilingual lexical databases. The translation candidates are ranked and returned as the output of the model, which are examined by human translators directly or passed on to MT decoders. The model returns a reasonable-sized set of translation candidates that contains suitable translations for the OOV word. When the model is finished, a phrase-based SMT system is used to translate the corpora.

A tool specifically created to deal with OOV words, is REMOOV (Habash, 2009). REMOOV utilizes four different techniques to solve the OOV problem in machine translation: morphological expansion, spelling expansion, dictionary word expansion and proper name transliteration. Thus this tool addresses several types of OOV, such as the problem of named entities and spelling errors. The research by Habash was done in Arabic ↔ English MT, using a MT system called Pharaoh (Koehn, 2004), the predecessor of Moses.

Webmining

In the past decade a different practice has begun to emerge in academic research. The extraction of translations for OOV words from external knowledge sources such as online dictionaries and encyclopedia is becoming a commonly used method (Eck et al., 2008; Vilar et al., 2007; Zhang, Huang & Vogel, 2005). The latter authors propose that translations for OOV words can be mined from the web through cross-lingual query expansion. When a query is sent to Google, snippets containing the query and possible a translation are returned. The translation is extracted from the top-N returned snippets and implemented into the translated target corpus.

Neural network approaches

Similar to SMT, NMT is in need of methods to address the OOV word problem. NMT uses a fixed vocabulary with 30,000 to 50,000 words, but translation is an open-vocabulary problem (Thang, Sutskever, Le, Vinyals & Zaremba, 2014; Sennrich, Haddow & Birch, 2015).

Thang, Sutskever, Le, Vinyals and Zaremba (2014) have proposed a technique specifically designed for NMT. Their approach annotates the training corpus with explicit information that enables the NMT system to emit, for each OOV word, a pointer to its corresponding word in the source sentence. The information is later used in post-processing to translate the OOV words using a dictionary.

Sennrich, Haddow and Birch (2015) continued to work on the previous method and have implemented it in Nematus. The BPE package is a new implementation to the previous work. Since Nematus is used in this study, thus the BPE package serves as the default method to test the OOV problem in the experiments. The authors have showed that the algorithm improves models over a back-off dictionary baseline.

2.3 Evaluation of machine translation

The effectiveness of machine translation can be measured in different ways. The following paragraph will elaborate on manual and automatic evaluation of MT, both have their merits.

2.3.1 Human evaluation

An obvious way of evaluating MT models is the use of human participants. One can provide different translations to a human judge and ask them to rank the translations from best to worst. Another way is the *adequacy*/*fluency* model, the most common used human evaluation model (Koehn & Monz, 2006). Linguistic Data Consortium defines the *adequacy* and *fluency* concepts as follows: “how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?” and “to what extent the translation is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker.” These concepts resemble faithfulness and fluency (Brown et al., 1990), which were introduced in Section 2.1.2.

For this thesis, the adequacy/fluency evaluation with TAUS industry guidelines is used in order to perform human evaluation of the MT models.

2.3.2 Automatic evaluation

The use of human evaluators on smaller samples of translated texts is reliable, although time consuming. Automated methods exist and are used by computational linguists to test their statistical models. The strength in these methods resides in the fact that single sentence errors are averaged out by the use of large corpora. Automated evaluation techniques essentially compare a machine translated text to one or more gold standards or references of the target text. Notable mentions are BLEU, NIST, Word error rate and METEOR.

The BLEU method is developed by Papineni, Roukos, Ward and Wei-Jing Zhu (2002). This automatic evaluation technique is language independent and correlates highly with human evaluation. BLEU calculates n-gram precision of translated texts and adds equal weights to each translated segment based on its length. The use of n-grams also means that using the BLEU metric is less reliable for languages with no specific grammatical word order. Its metric ranges from 0 to 100. Generally, BLEU scores are considered above average when the metric is larger than 50 (Papineni, Roukos, Ward & Wei-Jing Zhu, 2002).

NIST (Doddlington, 2002) is an adaptation of BLEU, but also calculates how informative a particular n-gram is. Word error rate (Klakow and Peters, 2002) has proven to be a valuable metric for speech recognition and machine translation. Finally, METEOR is a metric proposed by Banerjee and Lavie (2005) and emphasizes n-gram recall over n-gram precision, although it does not seek correlation on a corpus level. The BLEU metric is used in this thesis, since to our knowledge there are no other metrics yet that outperform BLEU with respect to correlation between MT and human judgment.

3. Method

This section describes the tools and approaches used in this research. To test the default method for handling OOV words in Nematus, two adjustments have been made. Adding the crowdsourced manual translations, this sums up to four different methods. First, the tools and resources used will be explained.

3.1 Tools and resources

Nematus

A Neural Machine Translation (Sennrich, Haddow & Birch, 2015, 2016) system that participated in the WMT ’16 shared translation task by building NMT systems for four language pairs: English ↔ Czech, English ↔ German, English ↔ Romanian and English ↔ Russian. All reported methods gave substantial improvements, with BLEU scores improving by 4.3 to 11.2 points. In human evaluation, the NMT system was tied for best system in 7 out of 8 translation tasks (Sennrich, Haddow & Birch, 2016).

The Nematus script runs several packages, as can be observed in Figure 3. When a text is entered for translation, it will be prepared by adjusting tokens and uppercase or lowercase letters. The BPE package transliterates unknown words using open vocabularies. After translating, the translated sentence is brought back to its original form by returning the tokens and lowercases or uppercases.

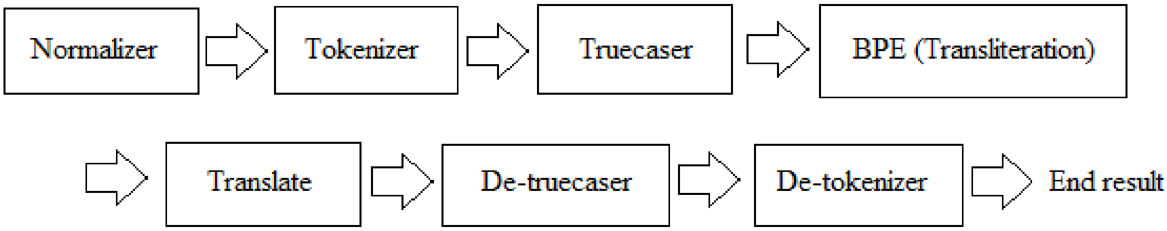

Figure 3. A schematic representation of the different steps that Nematus goes through.

BLEU-score

A metric for automatic evaluation of machine translation. It correlates highly with human evaluation and has little marginal cost per run (Papineni, Roukos, Ward & Wei-Jing Zhu, 2002). This metric is used to evaluate the different machine translation methods.

OPUS corpora

Translated texts used as a training set for Nematus. The dataset is free for public use and contains 157,669 documents of different sources such as books, subtitles, TED talks, Wikipedia pages and European Parliament proceedings. Source: http://opus.lingfil.uu.se/.

MOOC corpora

Translated texts used as a test set. The dataset is translated through the use of crowdsourcing. The total amount of sentences in this dataset is 1756. The MOOC corpora are of a slightly different domain than the OPUS corpora, which is beneficial in order to encounter more OOV words.

3.2 Methods

Nematus is a relatively new tool for machine translation, as shown in the WMT ’16 shared translation task. The system has its own way of dealing with OOV words, which is a transliteration model. To test the strength of this method, we have added two different methods by editing the Nematus script. These are the *drop* and *keep* methods. Thus for every sentence of the MOOC test data, there are different translations available based on the method. Lastly, because the MOOC test data was also translated by Tilburg University students using crowdsourcing, we have manual translations to our disposal.

Method A: Default

For this method a regular model using Nematus is built using OPUS corpora as a training set and MOOC corpora as a test set for EN ↔ NL machine translations. The default way of handling OOV words by Nematus is by using transliteration, or the BPE package as can be observed in Figure 4.

Method B: Drop

The Nematus script is adjusted by editing the BPE package. The BPE package finds OOV words, but removes them from the translation. This means OOV words are not dealt with in this method. In the target language text, the words will not reoccur and the sentence will logically appear incomplete.

Method C: Keep

The Nematus script is adjusted by replacing the BPE package with a Python script, that aims to retain the OOV word in the source language. When an OOV word is found, it will be moved to a wordlist. After translation, the source OOV words will return into the translated sentence. Copying OOV words into the target text is a reasonable strategy for names (Jean et al., 2015; Luong et al., 2015), however an untranslated word may result in lower adequacy for other types of OOV.

Method D: Manual

All 1756 sentences were translated using a crowdsourcing tool called Crowdflower. We are interested in seeing the interaction between machine and manual translations and its evaluation. This is the closest variant of a gold standard of the MOOC courses.

3.3 Evaluation

As discussed in the previous section, two forms of evaluation will be applied to our machine translation models. A human evaluation is carried out by having participants fill out a survey. The participant are asked to judge different translations of every method for 20 different sentences. Also, an automatic evaluation is conducted by using the BLEU metrics.

Procedure

The *default*, *drop* and *keep* methods yielded different results on the MT task. Of the 1756 sentences, 18 sentences differed in the target language in all methods, due to the OOV words. The hypotheses were tested by conducting a survey. The source text is written in English while the translations are in Dutch. There are 12 survey questions which have a different translation for every method. Six survey questions have translations of the *default* and *manual* method, while two questions only have the *manual* method. This is done because only 12 sentences were different in all four methods and useable for this survey. Furthermore, the six survey questions where only the *default* and *manual* method are represented, are implemented as to see which method scored highest; a translation made by a living person or Nematus. The survey amounted a total of 20 questions. All questions are drafted using the adequacy/fluency model according to TAUS guidelines.

The survey is enclosed in Appendix A.

Participants

There are 42 participants (18 male and 24 female), all of whom were of Dutch origin. The average age is 26.24 (SD = 7.83), the youngest being 19 and oldest being 63 years old. The average age of male participants was 26.94 (SD = 9.21), while the average age of female participants was 25.71 (SD = 6.96). Two participants were removed due to their answers being exceptionally low, below 2 SDs of the means of adequacy and fluency. It is possible they have misunderstood the explanations of adequacy and fluency. Furthermore, the proficiency of English of the participants is unknown to us.

Measurements

The participants are asked to rate several items on *Adequacy* and *Fluency*. The *Adequacy* scale consisted of 4 items (α = .71), the *Fluency* scale also consisted of 4 items (α = .71). The Cronbach's alphas of .71 indicate high internal consistency within the variables. The average *Adequacy* is 2.60 (SD = 0.96), the average *Fluency* is 2.62 (SD = 1.03). It is assumed that all observations are independent.

Automatic evaluation

A BLEU script is used to check the BLEU scores of the three machine translated methods for the complete test set. The reference text for the scoring is the *manual* method as described in the previous paragraph. This method is a combined set of manually translated sentences obtained through crowdsourcing. Ideally, a BLEU test is done by using a *gold standard* translation of the source text. The crowdsourced translations serves as the *gold standard* in this study.

4. Results

This section describes the results of both the human and automatic evaluation.

4.1 Human evaluation

Before starting the main analysis, the descriptions of the variables are given. Next, we tested whether the outcome variables are normally distributed. Lastly, the correlations between the variables are investigated. Gender did not seem to affect any of the correlations. All observations were independent of each other.

Two datasets were rendered from the survey results. The first dataset is rendered from the 12 survey questions that compare all 4 methods, testing hypotheses 1a and 1b. The second dataset is rendered from the 6 survey questions that compare manual translation with Nematus, testing hypotheses 2a and 2b.

For the variable *Adequacy*, (M = 2.60, SD = 0.96), scores seemed normally distributed with a skewness of -0.18 (*SE* = .08) and kurtosis of -.97 (*SE* = .15). For the variable *Fluency*, (M = 2.62, SD = 1.03), scores seemed normally distributed with a skewness of -0.09 (*SE* = .08) and kurtosis of -1.14 (*SE* = .15).

After this, the variables were further explored upon, by investigating the means and plotting the results in two tables. The first table describes the first dataset of 12 survey questions.

Table 1. *Mean proportions, standard deviations and confidence intervals of percent correct recognition for Adequacy and Fluency when comparing all methods (Dataset 1).*

Method	Mean	*SD*	95% CI
Adequacy			
A. Default	2.93**	0.93	[2.81, 3.04]
B. Drop	2.13*	0.76	[2.04, 2.23]
C. Keep	2.04*	0.71	[1.95, 2.12]
D. Manual	3.31**	0.76	[3.22, 3.41]
Fluency			
A. Default	2.71**	1.03	[2.58, 2.84]
B. Drop	2.26*	0.95	[2.14, 2.38]
C. Keep	2.26*	0.97	[2.14, 2.38]
D. Manual	3.25**	0.81	[3.15, 3.35]

**The mean difference is significant at the .05 level with all other methods.

*The mean difference is significant at the .05 level with Method A and D.

The hypotheses for *Adequacy* in this dataset were tested with a one-way analysis of variance (*F*(3,1004) = 152.952, *p* < .001). Levene's test indicated unequal variances (*F* = 10.887, *p* < .001). Because the assumption of homogeneity of variance is violated while scores seemed normally distributed, Games-Howell post hoc test was used to reveal the direction. The Games-Howell test showed that the adequacy of the *manual* method was significantly higher (*p* < .001) than the other methods. The *default* method was significantly lower than *manual*, but higher than *drop* and *keep* (*p* < .001). *Drop* and *keep* were significantly lower than *default* and *manual*, but did not differ from each other (p = .43). These results suggest that participants thought *manual* was the most adequate, followed by *default*, and then *drop* and *keep*.

The hypotheses for *Fluency* in this study were tested with a one-way analysis of variance (*F*(3,1004) = 63.841, *p* < .001). Levene's test indicated unequal variances (*F* = 7.025, *p* < .001). Because the assumption of homogeneity of variance is violated while scores seemed normally distributed, Games-Howell post hoc test was used to reveal the direction. The Games-Howell test showed that the adequacy of the *manual* method was significantly higher (*p* < .001) than the other methods. The *default* method was significantly lower than *manual*, but higher than *drop* and *keep* (*p* < .001). *Drop* and *keep* were significantly lower than *default* and *manual*, but did not differ from each other (*p* = 1). These results suggest that participants thought *manual* was the most fluent, followed by *default*, and then *drop* and *keep*.

Table 2 describes the second dataset of 6 survey questions wherein only *default* and *manual* were taken into account.

Table 2. *Mean proportions, standard deviations and confidence intervals of percent correct recognition for Adequacy and Fluency when comparing only manual and NMT (Dataset 2).*

Method	Mean	*SD*	95% CI
Adequacy			
A. Default	3.00**	0.85	[2.80, 3.12]
D. Manual	3.42**	0.77	[3.30, 3.62]
Fluency			
A. Default	2.94**	0.86	[2.80, 3.17]
D. Manual	3.26**	0.90	[3.12, 3.50]

**The mean difference is significant at the .05 level with the other method.

The hypotheses for *Adequacy* in this dataset were tested with an independent-samples t-test. There was a significant difference in the scores for *default* (M = 3.00, SD = 0.85) and *manual* (M = 3.42, SD = 0.77), t (212) = 3.81, p < .001. These results suggest that participants found manual translations more adequate and true to its original meaning.

The hypotheses for *Fluency* in this dataset were tested with an independent-samples t-test. There was a significant difference in the scores for *default* (M = 2.94, SD = 0.86) and *manual* (M = 3.26, SD = 0.90), t (212) = 2.64, p = .009. These results suggest that participants found manual translations more fluent.

Levene’s test of equality indicated that error variance of the dependent variables *Adequacy* (*F* = .314, *p* = .56) and *Fluency* (*F* = 3.620, *p* = .06) were equal across groups.

4.2 Automatic evaluation

In the table below the different BLEU scores can be observed.

Table 3. *BLEU Scores of translation methods A-D*

Method	*BLEU score*	*BP**	*Ratio*	*hyp_len*	*ref_len*
A. Default	27.73	1.000	1.027	20666	20116
B. Drop	27.56	1.000	1.011	20334	20116
C. Keep	27.35	1.000	1.024	20592	20116

*Brevity penalty

Note how seemingly the default method of Nematus is performing strongest. Though the values are close to one another, they are lower than 30. A possible explanation for this are the different domains for test and training corpora. The MOOC data are a collection of online courses, while the OPUS data range from European Parliament proceedings, subtitles of TV shows and books.

5. Conclusion and discussion

This thesis aims to compare different methods to handle the OOV word problem in NMT. This is done by answering the research question: how do our methods compare regarding the OOV word problem? We hypothesized that *adequacy* and *fluency* in translation would be rated higher in the *default* method than *drop* and *keep* (H1). Also, we hypothesized that *adequacy* and *fluency* would be rated higher in the *manual* method than the *default* method of Neural Machine Translation (H2).

The current *default* method of Nematus performs higher at machine translation than the adjustments made for this thesis, as was expected in H1. Participants scored the *drop* and *keep* methods significantly lower than *default*. The *drop* and *keep* methods were expected to score lower, because they did not react to the OOV problem. Leaving out words or keeping the OOV in source language is obviously not beneficial to a correct translation of a source text. One has to consider that, of all 1756 sentences, a deliberate selection of 12 sentences was made wherein all methods differed. By zooming in on this portion, significant differences were revealed. We assume the p-value of the ANOVA would increase if more sentences from the dataset were used, or randomization was added. The reason other sentences in the dataset were not affected by the *drop* and *keep* methods, because the sentences contained no OOV words. This is because the training data for Nematus is substantial.

BLEU scores were relatively close to one another. *Default* scored highest, followed by *drop* and *keep*. This suggests there is more proof that supports the first hypothesis in this thesis.

Manual translation scores significantly higher than the *default* method of Nematus, as was expected in H2. These results can be derived from both the ANOVA on the first dataset, as well as the independent t-test from the second dataset. One has to take into account that of the manual crowdsourced translations, a selection was made by the researcher for the most representative translation of the source text. In conclusion, NMT still has its limitations when compared to human translation.

A limitation in this research is the lack of a true gold standard for the MOOC corpora. The MOOC corpora were translated by multiple human translators, never were their translations measured on translation quality.

In the future, it would also be interesting whether to see external dictionaries like Google's top-N snippets could enhance the solution to the OOV word problem (Eck et al., 2008; Vilar et al., 2007; Zhang, Huang & Vogel, 2005).

Also, for this study a small portion of the test set was used to manually evaluate. The portion contained all of the OOV words that were encountered by Nematus. This has resulted in significant differences. In the future, it would be interesting to see if the effect holds when the portion that is evaluated is randomized. Though this study does suggest that manual translation will outperform Nematus nonetheless, when looking at the effect of the six sentences in the second dataset.

References

Aminian, M., Ghoneim, M., & Diab, M. (2014). Handling OOV words in dialectal Arabic to English machine translation. *LT4CloseLang 2014*, 99-108.

Aminian, M., Ghoneim, M., & Diab, M. (2015). Unsupervised False Friend Disambiguation Using Contextual Word Clusters and Parallel Word Alignments. *Syntax, Semantics and Structure in Statistical Translation*, 39-48.

Arnold, D. (2003). *Why translation is difficult for computers*. H. Somers (Ed.) Computers and Translation: A translator's guide, Amsterdam and Philadelphia: John Benjamins, 119-42.

Arora, K., Paul, M. & Sumita, E. (2008). Translation of unknown words in phrase-based statistical machine translation for languages of rich morphology. *SLTU*, 70-75.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S., & Lavie, A. (2005). 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*, 65-72.

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

Deselaers, T., Hasan, S., Bender, O., & Ney, H. (2009). A deep learning approach to machine transliteration. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 233-241. Association for Computational Linguistics.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* 138-145. Morgan Kaufmann Publishers Inc..

Durrani, N., Sajjad, H., Hoang, H., & Koehn, P. (2014). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *EACL*, 148-153.

Eck, M., Vogel, S., & Waibel, A. (2008). Communicating Unknown Words in Machine Translation. *In Proceedings of the international conference on Language Resources and Evaluation*, 1-6.

Gage, P. (1994). A New Algorithm for Data Compression. *C Users J., 12(2),* 23 – 38.

Habash, N. (2009). REMOOV: A tool for online handling of out-of-vocabulary words in machine translation. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*.

Huang, C., Yen, H., Yang, P., Huang, S., & Chang, J. (2011). Using sublexical translations to handle the OOV problem in machine translation. *ACM Transactions on Asian Language Information Processing (TALIP), 10*(3), 16.

Hutchins, W. (1986). *Machine translation: past, present, future*. Chichester: Ellis Horwood.

Hutchins, J. (2007). Machine translation: A concise history. *Journal of Translation Studies, 13*, 29-70.

Jean, S., Firat, O., Cho, K., Memisevic, R., & Bengio, Y. (2015). Montreal neural machine translation systems for WMT15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 134-140.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. Pearson.

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication, 38*(1), 19-28.

Koehn, P. (2004). Pharaoh: a beam search decoder for phrase-based Statistical Machine Translation models. In *Machine translation: From real users to research*, 115-124. Springer Berlin Heidelberg.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, 102-121. Association for Computational Linguistics.

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 ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

Luong, M., Sutskever, I., Le, Q., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Luong, M. T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Nirenburg, S., Carbonell, J., Tomita, M., & Goodman, K. (1994). *Machine translation: A knowledge-based approach*. Morgan Kaufmann Publishers Inc.

Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics, 29*(1), 19-51.

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*, 311-318. Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural Machine Translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. *arXiv preprint arXiv:1606.02891*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 3104-3112.

Vilar, D., Peter, J. T., & Ney, H. (2007). Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, 33-39. Association for Computational Linguistics.

Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1,* 1220-1229. Association for Computational Linguistics.

Zhang, Y., Huang, F., & Vogel, S. (2005). Mining translations of OOV terms from the web through cross-lingual query expansion. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 669-670). ACM.

Appendices

Appendix A

This appendix shows the content and nature of the survey, which was distributed through Qualtrics.

Beste deelnemer,

Bedankt dat u mee wilt werken aan dit onderzoek. We zijn benieuwd naar uw evaluatie van automatisch vertaalde teksten. Het invullen van de vragenlijst duurt ongeveer 10 minuten. Zorg dat u deze tijd daadwerkelijk hebt en laat u gelieve zo min mogelijk afleiden.

Met vriendelijke groet,

Sem Meereboer

Voor een Europees project genaamd TraMOOC, zijn we op zoek naar de optimale manier om online cursussen in het Engels te vertalen naar andere talen. Voor dit experiment zijn Engelse zinnen vertaald naar Nederlands, met behulp van een vertaalmachine-systeem (Nematus).We zijn benieuwd naar uw oordeel over de kwaliteit van de vertaalde zinnen. De vertaalkwaliteit wordt gemeten met behulp van het Adequacy/Fluency model. U mag deze begrippen als volgt interpreteren: Source: de oorspronkelijke tekstbron Adequacy: hoeveel van de Source-tekst komt terug in de vertaalde tekst? Fluency: hoe sterk scoort de vertaalde tekst op gebied van grammatica, spelling en zinsstructuur? Hierna volgt de eerste vraag. U mag de antwoorden scoren door te kijken naar de relatie tussen de antwoorden.

After the small introduction, the participants were inquired for their age and gender. Consquently, every participant had to answer 10 questions that were randomly picked of a set of 20 questions. The 20 questions can be observed below. First, the English source sentence was provided to the participant. Next, the participants were either shown four, two or one different answers and were asked to rate them on a four-point scale for adequacy and fluency.

#	Engels	Crowdsourced	Default	Drop	Keep
1	I will add that I have taken philophy classes that made me ponder on the philosphy of democracy.	Ik zal toevoegen dat ik filosofie lessen genomen heb die me deden nadenken over de filosofie van democratie	Ik voeg er aan toe dat ik de Latijnse les heb gevolgd waardoor ik over de filosofie van de democratie heb nagedacht.	Ik voeg hieraan toe dat ik lessen heb gevolgd waardoor ik over de democratie heb nagedacht.	Ik voeg hieraan toe dat ik de lessen heb gevolgd waarin ik over de democratie heb nagedacht.
2	50% was my first estimate but while coming up with that number i didn't think of 'small' every day decissions like tying shoelaces.	50% was mijn eerste gok maar terwijl ik dat getal bedacht, dacht ik niet aan dagelijkse kleine beslissingen zoals het strikken van veters.	50% was mijn eerste schatting, maar toen ik eraan kwam met dat nummer... dacht ik niet aan kleine 'kleine beslissingen zoals veters strikken'.	50% was mijn eerste schatting, maar toen ik met dat nummer kwam, dacht ik niet dat ik schoenveters zou strikken.	50% was mijn eerste schatting, maar toen ik eraan kwam met dat nummer... dacht ik niet aan 'klein' elke dag alsof ik schoenveters zou strikken.
3	And as you can see in this slide, the orangeish yellow area shows statistically significant activation in the, in recalling resonant leader, moments with resonant leaders, versus moments with dissonant leaders.	Zoals je kan zien in de slide, laat het oranjegele gebied zien dat er een statistische significante actie plaatsvindt in de momenten met leiders	En zoals u ziet in deze glijbaan, is de orangeish gele zone, statistisch significante activering in de, in het terugroeping van resonant leider, momenten met resonante leiders, versus momenten met slechte leiders.	En zoals u ziet in deze glijbaan, is het gele gebied statistisch significante activering in de, in het terugroeping van resonant leider, momenten met resonante leiders, versus momenten met slechte leiders.	En zoals u ziet in deze glijbaan, blijkt het gele gele gebied statistisch significante activering in de, in het terugroeping van resonant leider, momenten met resonante leiders, versus momenten met slechte leiders.
4	Does anyone know where I can find historical data of EMBIG and NEXGEM indexes?	Weet iemand waar ik historische data van EMBIG en NEXGEM indexen kan vinden?	Weet iemand waar ik historische data van EMBIG en NEXGEM-indexen kan vinden?	Weet iemand waar ik historische data van en indexxen kan vinden?	Weet iemand waar ik historische data gevonden heb?
5	My name is Alexandra Maratchi, and I'm going to be your interlocutor for the next few weeks in which this MOOC will run its course.	Mijn naam is Alexandra Maratchi, en ik zal je gesprekspartner zijn voor de komende paar weken in welke deze MOOC cursus zal lopen.	Mijn naam is Alexandra Maratchi, en ik ga uw gesprekspartner zijn voor de komende weken waarin deze MOOC zijn gang zal gaan.	Mijn naam is Alexandra, en ik zal jouw gesprekspartner zijn voor de komende weken waarin dit zijn beloop zal hebben.	Mijn naam is Alexandra... en ik zal jouw gesprekspartner zijn voor de komende weken waarin deze kwestie zijn gang zal gaan.
6	I'm from Barcelona, I studied in Paris and worked for several years at FCMG in Milano, and I came back to my home sweet home about two years ago to co-found HOMUORK with a friend of my childhood.	Ik kom uit Barcelona, ik studeeder in Parijs en werkte enkele jaren bij FCMG in Milaan, en ik kwam terug naar huis ongeveer twee jaar geleden om HOMUORK te co-founden met een vriend van mijn kindertijd.	Ik ben uit Barcelona, ik studeerde in Parijs en werkte voor verscheidene jaren bij FCMG in Milaan, en ik kwam terug naar mijn thuis Sweet home, ongeveer twee jaar geleden om de HOMUORK te vinden met een vriend uit mijn kindertijd.	Ik ben uit Barcelona, ik studeerde in Parijs en werkte voor verscheidene jaren in Milaan, en ik kwam terug naar mijn thuis Sweet home, ongeveer twee jaar geleden om samen te werken met een vriend uit mijn kindertijd.	Ik ben uit Barcelona, ik heb in Parijs gestudeerd en gewerkt voor verscheidene jaren in Milaan, en ik kwam terug naar mijn thuis Sweet home ongeveer twee jaar geleden om samen te werken met een vriend uit mijn kindertijd.
7	Homuork is an edtech start-up made to produce MOOC's for firms.	Homuork is een edtech start-up gemaakt om MOOC's voor bedrijven te produceren.	Homuork is een edtech start-up voor MOOC 's voor bedrijven.	Is een start-up voor bedrijven.	Voor firma 's is een begin gemaakt met een start-up.
8	But we mainly make SPOC's through our LMS, and that's why we decided to undertake this adventure along with Iversity!	Maar we maken vooral SPOC's door onze LMS, en dat is waarom we besloten dit avontuur te ondergaan met Iversity!	Maar we maken vooral de SPOC in onze LMS, en daarom besloten we om dit avontuur samen te doen met Iversiteit!	Maar we maken vooral de LMS door. Daarom hebben we besloten om dit avontuur samen te doen.	Maar we maken vooral de lijn door onze LMS, en daarom besloten we om dit avontuur samen te doen met...
9	For example, the changes in dress codes, such as wearing a short sleeve shirt at work to save energy, were already in place in Japan before the tsunmai.	Bijvoorbeeld, de veranderingen in de dress code, zoals het dragen van een shirt met korte mouwen op het werk om energie te besparen, werden al toegepast in Japan voor de tsunami.	Bijvoorbeeld, de veranderingen in de kledingvoorschriften, zoals het dragen van een overhemd met korte mouwen op het werk om energie te besparen, waren al in Japan vóór de tsunami.	Bijvoorbeeld, de veranderingen in de kledingvoorschriften, zoals het dragen van een overhemd met korte mouwen op het werk om energie te besparen, waren er al in Japan voor.	Bijvoorbeeld, de veranderingen in de kledingvoorschriften, zoals het dragen van een overhemd met korte mouwen op het werk om energie te besparen, waren al in Japan voor de opstand.
10	Welcome to our MOOC on the future of storytelling	welkom bij onze MOOC, de toekomst van verhalen vertellen.	Welkom bij ons MOOC op de toekomst van verhalen.	Welkom in de toekomst van verhalen.	Welkom in ons verhaal over de toekomst van verhalen.
11	Second finding, I have to go back and, and explain something that Professor Tony Jack here at Case Western Reserve University has just published an article in NeuroImage	Tweede bevinding, Ik moet terug gaan en, en iets uitleggen dat Professor Tony Jack hier op Case Western Reserve University een artikel gepubliceerd heeft in	Tweede gevonden, ik moet terug en, en leg iets uit dat Professor Tony Jack hier heeft op de zaak Western Reserve University heeft net een artikel gepubliceerd waaruit blijkt	Tweede vondst, ik moet teruggaan en uitleggen dat Professor Tony Jack hier op de zaak Western Reserve University net een artikel heeft gepubliceerd waaruit blijkt dat	Tweede vondst, ik moet terug en, en leg iets uit dat Professor Tony Jack hier heeft in de zaak Western Reserve University heeft zojuist een artikel gepubliceerd in een artikel, waaruit blijkt

	showing that when we aregiven an analytic task, we're given	Neurolmage dat laat zien wanneer we een analystische taak worden toegewezen, we zijn gegeven.	dat als we een analytische taak krijgen, we krijgen	als we een analytische taak hebben, we krijgen	dat als we een analytische taak hebben, we krijgen
12	To have the real experience of a MOOC!	om een echte MOOC ervaring te hebben	Om de echte ervaring van een MOOC te hebben.	Om de echte ervaring van een.	Om de echte ervaring te hebben.
13	Thanks also for interesting link.	Ook bedankt voor de interessante link.	Bedankt voor de interessante verbinding.		
14	Since it would be inconsistent I guess I'd better itemize them.	Sinds het inconsistent is, is het beter om te splitsen	Omdat het niet consequent zou zijn, zou ik het beter doen.		
15	Or maybe there's a sink inside the OR, too.	Of misschien is er een ook gootsteen in de OK	Of er zit ook een gootsteen in de OK.		
16	Again, no other possibility.	Nogmaals, geen andere mogelijkheid.	Nog een keer, geen andere mogelijkheid.		
17	Hello, again.	Hallo, nogmaals	Hallo, nog een keer.		
18	Socrates is a man, therefore Socrates is mortal.	Socrates is een man, dus hij is sterfelijk.	SOCRATES is een man, daarom is Socrates sterfelijk.		
19	Yes that's right.	Ja dat klopt			
20	Now, the fact is that we need both.	Nu, het is feit is dat we het beide nodig hebben.			