Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

An der Universität Hamburg eingereichte Dissertation

Data Selection for Statistical Machine Translation

zur Erlangung des Doktorgrades Doctor rerum naturalium (Dr. rer. nat.)
der Fakultät für Mathematik, Informatik und Naturwissenschaften,
Fachbereich Informatik vorgelegt von

Mirela-Stefania Duma

Hamburg, 2021

Die Gutachter:
Prof. Dr.-Ing. Wolfgang Menzel
Prof. Dr. Chris Biemann

Der Vorsitz der Prüfungskommission:
Prof. Dr.-Ing. Mathias Fischer

Tag der mündlichen Prüfung:
1. Dezember 2021

Acknowledgements

I would like to praise and thank God for giving life to me and for driving me to work on this thesis and to meet wonderful people.

I want to express my deepest gratitude to the panel of supervisors that conducted my thesis: Dr. Cristina Vertan, Prof. Dr. Walther v. Hahn and Prof. Dr.-Ing. Wolfgang Menzel. I hold high esteem for all of them and I am delighted to have worked with them.

While Dr. Vertan and Prof. v. Hahn supervised my master thesis, I was very excited that our collaboration continued through this PhD. I appreciate the supervision from Prof. v. Hahn as he possesses a vast experience in the CL field. Dr. Vertan is a wonderful supervisor that shared her great expertise in the MT field with me and encouraged me to aspire for more. I will always remember my very first step into the MT world at the MT Summit in Nice 2013 when Dr. Vertan offered me the opportunity to hold together a tutorial on MT. She is a remarkable supervisor that invests time and effort into supervising her students. I highly appreciate all of her support and I feel truly honored that she was my supervisor.

I would like to kindly thank Prof. Menzel for all the meetings, ideas and questions he raised during all this time. I am also very grateful to him for all the feedback he gave me while submitting papers to conferences. His in-depth knowledge helped me greatly. This thesis shaped very well thanks to his guidance during the PhD and detailed reviews and continuous feedback during the writing phase. I feel very lucky to have had such an amazing supervisor that offers ideas and constructive criticism and communicates so well with his students.

Very helpful feedback for this thesis was also given by Prof. Chris Biemann, who holds a vast knowledge in all NLP fields. I admired his passion for research ever since I met him at UHH some years ago. I am very thankful to him for offering me valuable insights.

I would like to also thank Prof. Dr. Matthias Fischer, the head of the examination committee, for chairing the oral defense and for the positive feedback.

This work was partially funded by the University of Hamburg through a stipend that was offered for two years. I would like to wholeheartedly thank UHH for offering me this opportunity and also for financing my conference trips. Also, I received for a short period of time funding through the ”Crossmodal Learning” project (TRR 169) and I am extremely grateful for this.

Working on a PhD goes hand in hand with writing and publishing papers to conferences. I received great feedback from anonymous reviewers for my papers and I would like to also thank them. While doing manual evaluation for this thesis, three native Spanish persons supported with annotations and/ or information about the Spanish grammar and I would like to also thank them.

A warm thank you to Ms. Anna Leffler, from the academic office, for her amazing support during the doctoral proceedings.

Thank you dear family for all your love and for being there every day. This thesis is dedicated to you.

Abstract

Machine Translation (MT) is a hot topic in the Computational Linguistics (CL) community. Given a target language and a text in a source language, an MT system provides a translation of the text into the target language in an automatic fashion. The problem of performing MT is currently tackled through statistical and neural approaches. This thesis focuses on Statistical Machine Translation (SMT), with an additional experiment in Neural Machine Translation (NMT) to assess the applicability of the developed methods to another MT architecture.

One particular topic that arises in practice when using MT systems is domain mismatch. Training an MT model on a domain and using it on another one does not yield the optimal performance due to the syntactic and semantic differences between the two domains. The attempt to solve this problem is known as domain adaptation and it is addressed through model- or corpus-driven approaches. Data selection, which is the topic of this thesis, falls in the latter category.

Given a general domain corpus and a target domain (also referred to as in-domain), each sentence from the general domain corpus is scored according to its similarity to the in-domain. In the MT community, the underlying assumption is that the general domain corpus is vast and thus, contains sentences that pertain to the in-domain. The goal of using data selection is to identify a small ratio of the general domain corpus that can be used to train an MT system that outperforms a system trained on the full general domain corpus. Moreover, MT systems obtained using ratio selection are faster to train and occupy less memory than the systems trained using the full data.

There are two challenges that arise with data selection: which method to use to determine the sentence similarity and how many of the general domain sentences to select as pertaining to the in-domain. In this work, I present data selection methods that address both challenges. I developed several scoring functions that select the general domain sentences that are most similar to an in-domain and compared them with a method I developed that automatically determines the ratio of sentences to select. The methods were also compared with a random selection of sentences to assess whether the gain in performance compared with systems trained using only the in-domain data comes from simply adding more training data or from adding more in-domain sentences. Moreover, the methods were contrasted with the most commonly used data selection method from the community, and with a baseline that uses the full general domain training data. Manual evaluation is investigated on a focus language pair and the system ranking result is compared with the automatic evaluation.

Data selection is crucial for MT systems that aim to translate domain-specific texts. The methods I developed were either on a par or surpassed a strong baseline, with the automatic ratio selection method performed particularly well in most of the experimental settings. With data selection SMT models were trained faster, had a smaller size, and performing on a par or better than the models trained using the full training data.

Zusammenfassung

Maschinelle Übersetzung (MT) ist ein aktuelles Thema in der Computerlinguistik (CL). Gegeben eine Zielsprache und einen Text in einer Ausgangssprache, liefert ein MT-System eine automatische Übersetzung des Textes in die Zielsprache. MT wird derzeit durch statistische und neuronale Ansätze realisiert. Diese Arbeit konzentriert sich auf die statistische maschinelle Übersetzung (SMT), mit einem zusätzlichen Experiment zur neuronalen maschinellen Übersetzung (NMT), um die Anwendbarkeit der entwickelten Methoden auf eine andere MT-Architektur zu evaluieren.

Ein spezielles Thema, das beim praktischen Einsatz von MT-Systemen auftritt, ist das Nichtübereinstimmen von Domänen. Ein MT-Modell auf einer Domäne zu trainieren und es auf einer anderen Domäne zu verwenden, bringt nicht das optimale Ergebnis aufgrund der syntaktischen und semantischen Unterschiede zwischen den beiden Domänen. Der Versuch, dieses Problem zu lösen, wird als Domänen-Adaption bezeichnet und kann in modell- und korpusgesteuerte Ansätze aufgeteilt werden. Datenauswahl, die das Thema dieser Arbeit ist, fällt in die letztere Kategorie. Gegeben ein Korpus einer generellen Domäne und eine Ziel-Domäne, auch als In-Domäne bezeichnet, wird jeder Satz aus dem Korpus der generellen Domäne, nach seiner Ähnlichkeit mit der In-Domäne bewertet. Die zugrundeliegende Annahme ist, dass das Korpus der allgemeinen Domäne groß genug ist, um Sätze zu enthalten, die sich auf die In-Domäne beziehen. Das Ziel der Datenauswahl ist es, einen kleinen Anteil des allgemeinen Domänenkorpus zu identifizieren, der zum Trainieren eines MT-Systems verwendet werden kann, und so ein besseres Ergebnis zu erhalten, als durch ein Training mit dem gesamten Korpus der allgemeinen Domäne. Darüber hinaus sind MT-Systeme, die mithilfe von Verhältnisauswahl trainiert werden schneller zu trainieren und benötigen weniger Speicherplatz als Systeme, die mit den gesamten Daten trainiert werden.

Es gibt zwei Herausforderungen, die bei der Datenauswahl auftreten: Welche Methode soll verwendet werden, um die Satzähnlichkeit zu bestimmen und wie viele Sätze aus der allgemeinen Domäne sollen als zur In-Domäne zugehörig ausgewählt werden. In dieser Arbeit stelle ich Methoden zur Datenauswahl vor, die beide Herausforderungen angehen. Ich habe mehrere Bewertungsfunktionen entwickelt, die die Sätze der allgemeinen Domäne auswählen, die einer In-Domäne am ähnlichsten sind, und habe sie mit einer von mir entwickelten Methode verglichen, die automatisch das Verhältnis der auszuwählenden Sätze bestimmt. Die Methoden wurden auch mit einer zufälligen Auswahl von Sätzen verglichen, um festzustellen, ob der Leistungsgewinn im Vergleich zu Systemen, die nur mit den In-Domain-Daten trainiert wurden, durch einfaches Hinzufügen von mehr Trainingsdaten oder durch Hinzufügen von mehr Sätzen aus der In-Domäne kommt. Außerdem wurden die Methoden kontrastiert mit der am häufigsten verwendeten Datenauswahlmethode aus der MT-Forschungsgemeinschaft und mit einer Baseline, die die vollständigen Trainingsdaten der allgemeinen Domäne verwendet. Eine manuelle Auswertung mit einem Fokus-Sprachpaar wurde mit dem Ergebnis der automatischen Auswertung verglichen.

Datenauswahl ist entscheidend für MT-Systeme, die domänenspezifische Texte übersetzen sollen. Die entwickelten Methoden waren entweder gleich oder besser als eine starke Baseline, wobei die automatische Verhältnisauswahl-Methode in den meisten experimentellen Einstellungen besonders gut abschnitt. Mit der Datenauswahl wurden SMT-Modelle schneller trainiert, hatten einen geringeren Platzbedarf und schnitten gleich gut oder besser ab als die Modelle, die mit den vollständigen Trainingsdaten trainiert wurden.

Contents

List of Figures

List of Tables

List of Abbreviations

ATD	Automatic Threshold Detection
BERT	Bidirectional Encoder Representations from Transformers
BLEU	BiLingual Evaluation Understudy
DA	Domain Adaptation
DE	German
DS	Data Selection
DSTF	Data Selection by Term Frequency
EN	English
ES	Spanish
hATD	hybrid Automatic Threshold Detection
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MML	Method of Moore and Lewis
MT	Machine Translation
NIST	National Institute for Standards and Technology
NMT	Neural Machine Translation
NLP	Natural Language Processing
OOV	Out of Vocabulary
SEF	Sentence Embedding Filtering
SL	Source Language
SMT	Statistical Machine Translation
TL	Target Language
TER	Translation Error Rate
WMT	Workshop on Machine Translation

Chapter 1

Introduction

Translation, whether human or automatic, supports dissemination of news or information, helps people understand what they want to read in a language they do not know and facilitates communication between different communities. While translation done by a human professional yields high quality translations, it is costly, time consuming and not reusable for other documents. On the other hand, Machine Translation (MT) obtained through free services offers the possibility to obtain fast translations for any documents, in many languages. However, MT is a problem where research is still actively ongoing since it is error prone. The difficulty stems from teaching the machine how to understand a message in a language and how to convert it into another language.

Statistical Machine Translation (SMT) is an MT approach which relies on statistical models that learn how to translate using large amounts of bilingual aligned text (corpora). Probabilities to translate phrases from a source language into a target language are learned by analyzing the bilingual corpora. When training an SMT model on a domain and evaluating it on a different one, the performance of the system usually drops due to differences in the vocabulary and a mismatch in style or genre. Therefore, domain adaptation is essential in the field of SMT. In this work, I will follow the MT community which defines a domain by the corpora that it uses, together with the notion of domain mismatch that is defined by differences in topic, genre, style, formality or linguistic mode (Koehn and Knowles, 2017). Throughout this thesis, a source domain is always a general domain, which is a collection of sentences pertaining to multiple domains. The target adaptation domain is represented by an in-domain and is usually a corpus or collection of corpora of smaller size than the corpora of the general domain.

One particular corpus-driven domain adaptation technique is data selection. Given an in-domain In and a sentence s from a general domain Gen, the task of data selection is to decide whether s could belong to In and thus become useful in training an MT system targeting In. Therefore, from a large corpus of general domain sentences, the ones that are most similar to a given in-domain need to be selected. The main purpose is to use those selected sentences along with the ones from the in-domain to train faster MT systems that perform better in comparison with a system trained using the full general domain data.

The main work-flow of data selection consists of scoring all the general domain sentences according to their similarity to the in-domain, selecting the top n most similar sentences using one or more predefined thresholds/ ratios and then using the selected sentences either for training a full SMT system or parts of it.

1.1 Motivation and Objectives

Regardless of the size of the in-domain data, one of the advantages of data selection consists in providing more in-domain data selected from large amounts of general domain data. This pseudo in-domain data (term defined by (Axelrod et al., 2011)) is used in training MT systems that outperform a system trained on all the available data, both in terms of translation quality and time and resource efficiency. The challenges that arise when performing data selection are:

- developing a scoring method that produces similarity scores for the sentences from the general domain according to their similarity to the in-domain
- choosing a threshold/ ratio that determines how many of the scored sentences to keep for later use in training the MT systems.

Standard state-of-the-art methods resolve the first difficulty by means of information retrieval, perplexity or edit distance methods. Regarding the second difficulty, there is no agreed upon standard defined parameter setting in the community (for the start-threshold and for the increment threshold).

The motivation for my work was driven by these challenges. In the following, I define my research questions together with the objectives of my work.

- **RQ1**: How to transform sentences into meaningful vector representations that can be used to compute similarity scores? → **objective**: use state-of-the-art distributed representations of sentences.
- **RQ2**: How can the sentence vectors contribute to the formulation of data selection algorithms? → **objective**: develop scoring functions based on vector similarities.
- **RQ3**: How to simplify the threshold tuning step in the data selection pipeline? → **objective**: investigate methods to automate this process and analyze if they outperform the standard incremental ratio methods.
- **RQ4**: Could a gain in performance come from simply adding more general domain training data or from adding more pseudo in-domain one? → **objective**: directly compare systems that are trained on randomly selected general domain sentences with the ones trained using the data selection methods.
- **RQ5**: Having systems trained using the developed data selection methods, is the ranking of systems consistent on different test sets, language pairs

and in-domains? → **objective**: rank systems using the most common MT evaluation metric.

- **RQ6**: How to manually evaluate the developed systems and what is the relationship between the automatic and the manual evaluation? → **objective**: use standard human evaluation procedures and compare the system rankings obtained using the two evaluation paradigms.

Focusing on the aforementioned challenges of data selection and on the defined objectives, my thesis presents the methods I developed for performing data selection for SMT. Several similarity scoring functions were developed to meet the first challenge (Duma and Menzel (2016a); Duma and Menzel (2016b); Duma and Menzel (2017b)); Duma and Menzel (2018)). As for the second challenge, I developed an automatic threshold detection method which achieved good results on a variety of language pairs (Duma and Menzel, 2017a). Additionally, a hybrid data selection method is described in this thesis.

In the following, a publication list with the papers I published related to this thesis is presented. Also, my research contributions are emphasized.

1.2 Publications and Contributions

This section presents the publication list as well as the achievements gained by participating in various competitions. A short description of the conferences and workshops where I published papers follows.

An important venue in the field of Machine Translation field is the *Conference on Machine Translation (WMT)*, which was held ten times as a workshop before 2016 when it changed its status to a conference. WMT encourages participants to evaluate their methods and algorithms on the provided task datasets, in order to obtain comparable results. A series of shared tasks is given, on a different set of language pairs and corpora. Since the focus of my thesis is on data selection (as a particular approach of domain adaptation) for statistical machine translation, I participated in the WMT 2016 shared task of domain adaptation of MT to the IT domain[1]. I also took part in the WMT 2017 and WMT 2018 shared task of translating documents from the biomedical domain.

The Semantic Evaluation (SemEval) series of workshops enables participants to evaluate and compare their systems on various tasks. One of the tasks is *Semantic Textual Similarity* and it describes a challenge to assess the degree to which the meaning of two snippets of text is related.

KONVENS (Konferenz zur Verarbeitung natürlicher Sprache) is a conference on Computational Linguistics and Natural Language Processing that is being held biennially since 1992.

Below is the list of publications focusing on the topic of my thesis:

[1] http://www.statmt.org/wmt16/it-translation-task.html

1. Mirela-Stefania Duma and Wolfgang Menzel. Data selection for IT texts using paragraph vector. In Proceedings of the First Conference on Machine Translation, WMT 2016, collocated with ACL 2016, August 11-12, Berlin, Germany, pages 428-434, 2016.

2. Mirela-Stefania Duma and Wolfgang Menzel. Paragraph vector for data selection in statistical machine translation. In Proceedings of the 13th Conference on Natural Language Processing KONVENS 2016, September 19-21, Bochum, Germany, pages 84-89, 2016.

3. Mirela-Stefania Duma and Wolfgang Menzel. Sef@UHH at Semeval-2017 task 1: Unsupervised knowledge-free semantic textual similarity via paragraph vector. In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval @ACL 2017, Vancouver, Canada, August 3-4, 2017, pages 170-174, 2017.

4. Mirela-Stefania Duma and Wolfgang Menzel. Automatic threshold detection for data selection in machine translation. In Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017, pages 483-488, 2017.

5. Mirela-Stefania Duma and Wolfgang Menzel. Translation of Biomedical Documents with Focus on Spanish-English. In Proceedings of the Third Conference on Machine Translation, WMT 2018, Brussels, Belgium, October 31 - November 1, pages 648-654, 2018.

My contributions to the field of Data Selection for SMT can be grouped into two categories: participation in international competitions and scientific results.

Using the proposed methods, I achieved encouraging results in several international competitions:

- first place on the domain adaptation task at WMT 2016, for the language pair English-German, constrained mode (i. e. no additional data was used in training the models, only the data made available by the organizers) and second place out of 13 submissions in the same task when considering constrained and non-constrained submissions.

- submitted runs for seven language pairs in the biomedical task at WMT 2017, obtaining BLEU scores in the range between 32 and 49 for six of the seven language pairs, indicating that the developed method generally achieves good translation results on a variety of language pairs.

- ranked first out of 53 runs submitted in total by all participants on the Spanish-English-WMT test set of Semeval 2017[2]. Moreover, I participated in all six sub-tracks, obtaining Pearson correlation scores above the median score for five out of seven test sets.

The scoring functions that I developed are innovative and present improvements over strong baselines. An important step forward in data selection was also achieved by my automatic ratio detection algorithm which significantly reduces the time needed to obtain a domain-adapted system. All methods were investigated for

[2] Appendix C contains details of my submission. 53 runs including one baseline.

several language pairs and in-domains, validating their applicability on different settings. Moreover, given the recent trend of the MT community moving towards Neural Machine Translation, I demonstrate the applicability of my methods on the NMT paradigm[3] (see Chapter 5, Section 5.2.3).

1.3 Outline of the Thesis

In Chapter 1, the *motivation and objectives* of my work are described. The research questions are defined in relation with the challenges that data selection poses. Additionally, my contribution to the MT community is presented along with my publications from several international conferences. Encouraging results from international competitions are given, where two of them have reached a first place. Lastly, a short outline of the thesis is given in this chapter.

Given the motivation stated in the first chapter, an important aspect in data selection, as well as other computational linguistics tasks, is text representation. Chapter 2 centers on distributed representations of words and sentences and offers an overview of word representation via *Word2Vec* (Mikolov et al., 2013a) and sentence representation via *Paragraph Vector* (Le and Mikolov, 2014). The latter one is an extension of Word2Vec. An essential aspect of these embedding models is also discussed, namely the limitations that they pose. Fortunately, the shortcomings of these models have no influence on the data selection pipeline.

Continuing with background knowledge, the core concepts of SMT are described in Chapter 3. These include word alignment, language and translation modeling, decoding and tuning of parameters. Since later on in the thesis, the applicability of my data selection methods is demonstrated on the NMT architecture, this chapter also briefly covers NMT. Finally, in-depth evaluation techniques of MT system outputs are detailed.

Domain adaptation for MT with a special emphasis on data selection is introduced in Chapter 4. Fundamental terminology and definitions are presented, followed by related work.

The following three chapters are based on the notions and architectures described before. They represent the core of this thesis and present the data selection algorithms and methods that I developed together with a vast suite of experiments on different in-domains and language pairs. An *experimental setting* is defined by the language pair and the in-domain used in the experiments.

Chapter 5 introduces data selections methods that require *scoring functions* for general domain data filtering. A data selection method that uses Paragraph Vector for sentence representation is described and applied to the IT domain for the English→German language pair. The scoring function captures the similarity between sentences by means of cosine similarity between the sentences vectors and weights that were empirically determined. In addition to the Paragraph Vector sentence representation, the traditional term frequency approach is investigated for the English→Spanish and English→Portuguese language pairs on the Biomedical

[3] Even though the main focus of my thesis is on SMT.

domain. The relative difference between word counts in the in-domain domain and in the general domain constitutes the core of this scoring formula. For all these methods, the ratio that determines how many sentences to keep from the general domain data was tuned according to common practice in the MT community. RQ1 and RQ2 are answered in this chapter.

Chapter 6 identifies the limitation that most of the data selection methods have and presents a step towards a solution to improve the current methods. Developing scoring methods is a crucial step in the data selection pipeline, but also the tuning of the threshold/ ratio selection. This chapter tackles the ratio selection challenge by means of a classifier that determines whether a sentence from the general domain is kept or discarded. This approach is investigated on the Biomedical domain, on seven language pairs. A hybrid approach is also presented that improves the initial one. It relies on any of the data selection scoring functions previously presented in Chapter 5 for obtaining negative training samples for the classifier. A comparison between the two automatic ratio approaches is given. An extended evaluation follows as the ratio tuning and automatic ratio detection methods are compared on three common experimental settings. This chapter answers RQ3, RQ4 and RQ5.

The work presented in Chapter 7 focuses on the Spanish→English language pair, for the Biomedical domain. While the previous two chapters describe automatic evaluation results, this chapter presents manual evaluation via three-way ranking. The automatic and manual evaluation results are compared as RQ6 is answered.

A summary of the data selection methods, together with an overview of the experimental results and ideas for future work are covered in Chapter 8.

Chapter 2

Distributed Representations of Words and Sentences

Since data selection for MT focuses on finding the sentences from the general domain data that could be regarded as pertaining to the in-domain, one important aspect centers on representing the sentences. Whether a sentence should be discarded or kept, depends crucially on how well the meaning of the sentence can be represented. Vector representations can be computed for words, but the idea can be generalized to sentences.

Semantic representations of words could be based on the *distributional hypothesis* (Harris, 1954) which states that words with similar meaning are used in similar context and that the meaning of a word can be derived from its context[1].

The vector representation of a word is usually termed as word embedding. Word2Vec is one of the most popular type of word embedding as it has been shown effective in many applications and is easy to train and to use. This word representation was introduced by Mikolov et al. (2013b) with the aim to provide word representations that are learned from vast amounts of data. The vectors are able to encode linguistic regularities and semantic patterns (Mikolov et al., 2013b). Their success followed up immediately as the community adopted this kind of architecture and applied it to tasks like sentiment analysis (Petrolito and Dell'Orletta, 2018), sentence similarity (Mijangos et al., 2016) or quality estimation for machine translation (Paetzold and Specia, 2016). Paragraph Vector (Le and Mikolov, 2014) is an extension of Word2Vec that considers representing sentences in a similar fashion as the algorithms behind Word2Vec. As in the case of Word2Vec, Paragraph Vector was applied to information retrieval (Ai et al., 2016), and sentiment analysis (Hu and Song, 2016; Barhoumi et al., 2017).

The successful application of Paragraph Vector in various tasks constitutes my **motivation** for using it as a means of sentence representation in my work. I employed the paragraph vectors for:

[1] Summarization taken from (Kamath et al., 2019, p. 204)

- **similarity metric development** (for data selection): given a general domain sentence, obtain its similarity to the in-domain (in the form of a score).
- **sentence classification**: given a general domain sentence, deciding whether it could be considered pertaining to the in-domain or not.
- **semantic textual similarity**: the task of determining how similar in meaning two sentences are (see Appendix C for details).

This chapter offers a description of Word2Vec and Paragraph Vector with the aim to apply the latter to data selection for machine translation.

2.1 Word2Vec

This section describes the two architectures introduced by Mikolov et al. (2013b), namely Skip-gram and Continuous Bag-of-words. Examples of word similarities and analogical reasonings applied on the Biomedical and IT domains are presented in Appendix B.

Learning algorithm

Given a sentence from the training corpus, in Mikolov et al. (2013b), the algorithm scans it using a sliding window of size $2 \cdot window + 1$, with $window$ being a Word2Vec hyperparameter. The word positioned in the middle of the sliding window is termed *center word*, while the $window$ words to the left and to the right of the center word are termed *context word*[2].

A neural network (NN) is trained using a corpus with a vocabulary of size V for making predictions (either predicting the context word given a center word or vice-versa).

Every word in the vocabulary is represented as a one-hot encoding, which is a vector of the size V with a value of 1 for the index corresponding to the word and a value of 0 for all the other indexes.

There are two weight matrices that need to be optimized by the model[3]: the input weight matrix $W_{input} = [v_1, v_2, \cdots, v_V]$ and the output weight matrix $W_{output} = [v'_1, v'_2, \cdots, v'_V]$. The word vectors are a by-product of the training procedure for the prediction model. They are extracted as row vectors from W_{input}.

The size of the word embedding can be freely chosen and is usually set to a value greater than 100 depending on the application needs.

The NN is trained using Stochastic Gradient Descent (SGD) with backpropagation. The training steps follow the basic NN architecture: predict labels, compute loss and update parameters. Given a weight w, SGD uses the partial derivative of the objective function with respect to w to update the weight. The objective

[2] Sometimes in the NLP community the short terminology of ”context” is used instead of ”context words”.

[3] For the sake of brevity, ”model” is used to denote either a Skip-gram or a CBOW model.

functions for both Skip-gram and CBOW architectures are given in the following sections.

Linear activation between the input layer and the hidden layer is used. This not only simplifies the computation, but also results in word vectors which can be linearly combined[4]. Appendix B gives examples of summation and difference using word vectors. The standard activation function, *Softmax*, is the one used by the Word2Vec architecture for connecting the hidden layer to the output layer to obtain the predictions (either context or center words, depending on the choice of architecture). This function is formally defined as:

$$s(x_w) = \frac{\exp(x_w)}{\sum_{i=1}^{V} \exp(x_i)}$$

where x_w is the input vector and $\exp$ is the standard exponential function. The outputs of Softmax transformation form a probability distribution.

Skip-gram model

A large corpus or collection of documents is used to train the NN. Given a center word as the input, a NN is trained to predict the probability for every word in the context window of being in the vicinity of the center word.

Instead of using the full vocabulary, only the context words are used. They are selected using the *window* parameter, a value that stands for how many words to consider to the left and to the right of the center word (in practice, it has a value between 1 and 20) (Mikolov et al., 2013b).

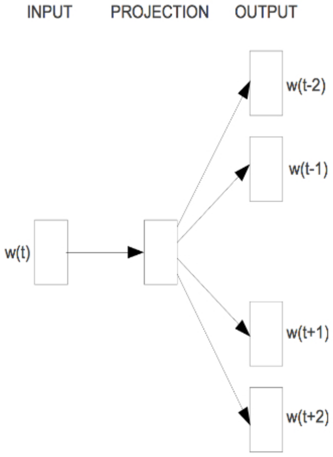


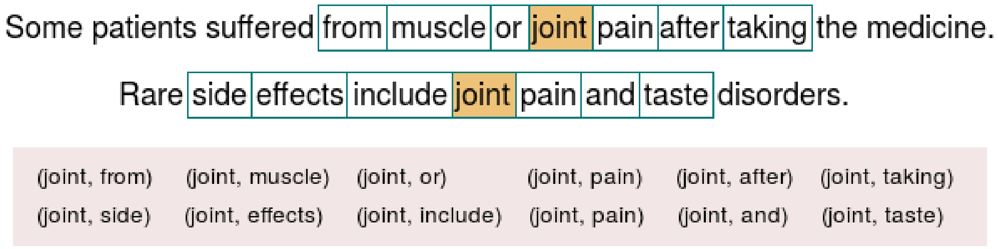
Figure 2.1: Skip-gram model (taken from (Mikolov et al., 2013a))

For every word in the vocabulary (acting as a center word) and every word in the context window, training samples (word pairs) are extracted in the form

[4] For brevity reasons, through this chapter, the general term ”word vectors” is used as means of word vectors obtained using Word2Vec.

$(center\_word, context\_word)$. Figure 2.1 (from (Mikolov et al., 2013a)) depicts the center word $w(t)$ and the output context words $w(t-2)$, $w(t-1)$, $w(t+1)$ and $w(t+2)$, given a window size of two.

Figure 2.2 shows an example with two training sentences that I have chosen from the Biomedical domain to illustrate the extraction of training samples. Given the center word *joint* and a window size of three, twelve training samples are extracted. Since the pair $(joint, pain)$ appears twice and $(joint, or)$ once, the probability for *pain* will be higher than for *or*.



Some patients suffered from muscle or joint pain after taking the medicine.

Rare side effects include joint pain and taste disorders.

(joint, from)	(joint, muscle)	(joint, or)	(joint, pain)	(joint, after)	(joint, taking)
(joint, side)	(joint, effects)	(joint, include)	(joint, pain)	(joint, and)	(joint, taste)

Figure 2.2: Example of extracting training samples for the center word *joint*

As defined in Mikolov et al. (2013b), the objective is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c}\log\ p(w_{t+j}|w_t)$$

where T is the number of training pairs and c is the window size, while $p(w_{t+j}|w_t)$ is defined using the *Softmax* function:

$$p(w_O|w_I) = \frac{\exp(v'^{\top}_{w_O}\ v_{w_I})}{\sum_{w=1}^{V}\exp(v'^{\top}_{w}\ v_{w_I})}$$

where v_w is the input vector representation of w and v'_w is the output vector representation of w (Mikolov et al., 2013b).

Continuous Bag-of-words model

This architecture is similar to the previously described one, however the task is to predict the center word based on the context words. Figure 2.3 (from (Mikolov et al., 2013a)) depicts the current word $w(t)$ being predicted from the input context words $w(t-2)$, $w(t-1)$, $w(t+1)$ and $w(t+2)$, using a window size of two.

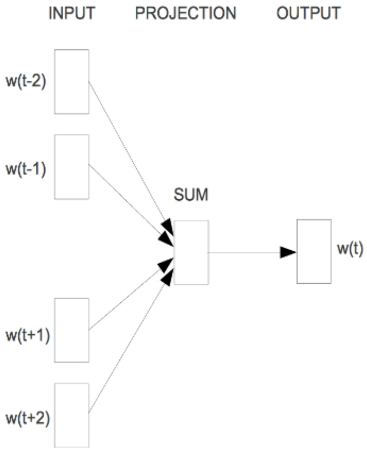


Figure 2.3: Continuous Bag-of-words model (taken from (Mikolov et al., 2013a))

Given the first sentence in Figure 2.2, if the context would be ”Rare side _____ include joint” (window size set to two), the model will give a much higher probability to the word *effects*, in contrast to words like *weather* or *car*.

As defined in Le and Mikolov (2014), the objective of this model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=c}^{T-c}\log\ p(w_t|w_{t-c},...,w_{t+c})$$

where T is the number of training pairs, c is the window size and the probability is calculated using the *Softmax* function:

$$p(w_t|w_{t-c},\cdots,w_{t+c})=\frac{\exp(y_{w_t})}{\sum_{i=1}^{V}\exp(y_i)}$$

where y_i is the log-probability for the output word i, calculated as:

$$y=b+Uh(w_{t-c},\cdots,w_{t+c};W)$$

with U and b being the softmax parameters and h is the concatenation or average of the word vectors extracted from W (Le and Mikolov, 2014).

Negative Sampling

In practice, Softmax is computationally expensive because it is calculated across many classes, which essentially represents all the words in the vocabulary. Therefore, computing $p(context_i|center_j)$ has an algorithm complexity of $O(V\ x\ V)$, given that the center word takes all values from the vocabulary ($i\in V$) and the context word also takes all values from the vocabulary (Kim, 2019). As discussed in Kim (2019), in practice the algorithm complexity is $\approx O(V)$ because the normalization factor from computing $p(context_k|center_j)$ is the same for all $k\in V$.

To overcome this still expensive computation, in the actual implementation a complexity reduction sampling technique is applied. Negative sampling is presented in Mikolov et al. (2013b) as a means of reducing computational complexity. It represents the choice I opted for my Paragraph Vector experiments.

Given a training pair made up of a center word center and a context word $\mathit{context}$ (extracted from the context window), the algorithm is adapted from a multi-label classification task, to a binary one. The probability of $\mathit{context}$, given center represents a positive sample, while a predetermined number of training samples, K, of the form $(\mathit{random\_context}, \mathit{center})$ are randomly drawn from the entire vocabulary and act like negative samples. The number of the training samples is usually very small compared to the size of the vocabulary (between 5 and 20). As a result, for each training sample, $K + 1$ weight updates are being made. The activation function Sigmoid is used instead of Softmax (Mikolov et al., 2013b; Kim, 2019).

2.2 Paragraph Vector

Based on Word2Vec, Paragraph Vector (PV) (Le and Mikolov, 2014) takes a step forward and uses word vectors in order to obtain numeric representations for texts that can be phrases, sentences or documents. Given a sentence, the model computes the vectors for all the words in it, together with a paragraph vector, that acts like a "memory" of the topic of the sentence (Le and Mikolov, 2014).

As in the case of Word2Vec, there are two architectures presented in (Le and Mikolov, 2014), namely the Distributed memory model (similar to the Continuous bag-of-words model) and the Distributed bag-of-words model (similar to the Skip-gram model), which are presented in the next sections.

The term Doc2Vec referring to the Paragraph Vector model is used (interchangeably with PV), similarly to using Word2Vec for the word vectors model. Also, the text unit of interest is the sentence, since Doc2Vec is later applied to data selection in this thesis.

Learning of Doc2Vec is very similar to the training of Word2Vec, with negative sampling being used to speed up the training (see Section 2.1). What differs between the two architectures is the addition of a matrix where every column represents the paragraph vector. As with the word vectors, the paragraph vector is randomly initialized in the beginning of the learning, and as the training converges, it will hold up the representation of the sentence. The paragraph vector is shared across all context vectors from the same sentence, but is unique for every sentence (Le and Mikolov, 2014).

Distributed memory model

The Word2Vec principle of predicting the next word in a sentence based on a given context is applied similarly by Doc2Vec: the paragraph vectors contribute together with the context vectors to the prediction task of the next word. The same notation as in the case of the word vectors is preserved, namely the matrix W_{input} contains unique vectors for every word in the vocabulary and is shared among all sentences. Additionally to the word vectors, the paragraph vector acts as the topic of the sentence (Le and Mikolov, 2014).

The paragraph vector and the word vectors are combined through concatenation or averaging. In Figure 2.4, the combination of the paragraph vector with the context vectors is used to predict the next word. The vectors are trained with stochastic gradient descent and backpropagation: at every step, a context is sampled from a random sentence and the error gradient is calculated and used in updating the parameters of the model (Le and Mikolov, 2014).

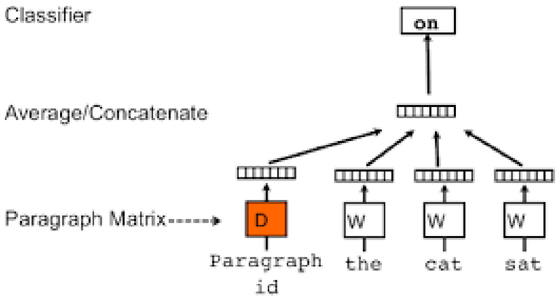


Figure 2.4: Distributed memory model (taken from (Le and Mikolov, 2014))

Distributed bag-of-words model

Similarly to the Skip-gram architecture, the Distributed bag-of-words approach predicts words randomly sampled from the sentence. The training pairs are of the form $(r, paragraph\_id)$, where r is a random word sampled from a sliding window over the sentence. As before, the neural network is trained with stochastic gradient descent. Figure 2.5 illustrates an example where the paragraph vector is trained to predict the words from the window Le and Mikolov (2014).

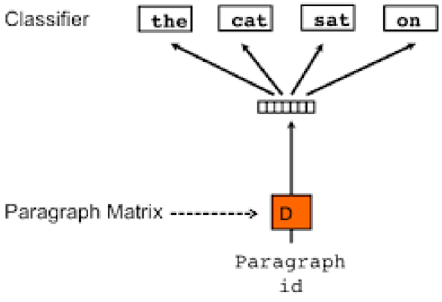


Figure 2.5: Distributed bag-of-words model (taken from (Le and Mikolov, 2014))

I used the Gensim library[5] (Řehůřek and Sojka, 2010) to implement my Para-

[5] urlhttps://radimrehurek.com/gensim/

graph Vector-based data selection methods. Appendix A presents the hyperparameters of the Doc2Vec model as offered by Gensim.

Additional approaches to sentence representation are discussed in Chapter 8, Section 8.3 where future work covers further ideas for this topic.

2.3 Limitations of Word2Vec and Doc2Vec

In this section, some of the limitations of Word2Vec and Doc2Vec are discussed with a focus on how they affect data selection. While embedding models are currently[6] widely popular and research continues on improving them, enhancements like bilingual sentence embeddings could be explored for data selection. However, the complete exploration of embeddings models is beyond the scope of my thesis and possible improvements are discussed in Chapter 8 which refers to Future Work.

According to Kamath et al. (2019), the most common drawbacks of embeddings models center on out-of-vocabulary words and antonymy.

A drawback of word embeddings models is that they fail at distinguishing antonyms (Kamath et al., 2019). This is due to the fact that usually two words a_1 and a_2 that are antonyms appear in similar contexts. However, mapping a sentence that contains a_1 close in the vector-space to a sentence that contains a_2 does not have a negative influence for data selection. On the contrary, selecting both sentences is actually required for this task. The same applies for the case of negations. Below, some examples of sentences with antonyms (*worked/ failed*) and negations (*worked/ did not work*) from the Biomedical domain are provided, where all sentences vectors are situated close in the vector-space and require the sentences to be identified as pertaining to the Biomedical domain.

The treatment worked after the first dose.
The treatment did not work after several doses.
The treatment failed using a low-dose strategy.

The vocabulary of a language is extremely vast and contains numerous words that are infrequent. For data selection and in general, domain adaptation, terminology often contains rare words that are domain-specific terms. While domain-specific terms are relevant in the setting of data selection, it is not guaranteed that all rare words from a vocabulary extracted from one or more corpora belong to that target domain. Also, running into the out-of-vocabulary problem is inevitable even when using large amounts of corpus data. Moreover, it is unpractical to train embeddings on extremely large vocabularies due to memory limitations and long training time. In the experiments presented in this thesis, I used a minimum value for word frequency defined in the Doc2Vec parameters (a default value of five) which ensures that words appearing with a frequency lower than this limit are discarded from the vocabulary. However, this ensures that the model can be loaded into memory quickly enough.

[6] At the time of writing this section.

Another point to be mentioned, is that training Word2Vec/ Doc2Vec models is self-supervised and evaluating them can only be done via the end application. In the application case of data selection for MT, the evaluation is very expensive from the resource point of view.

2.4 Summary

The motivation behind employing Paragraph Vector in my work is offered in the beginning of this chapter. An overview of word representations using Word2Vec (Mikolov et al., 2013b) was presented afterwards. Details on the two architectures, Skip-Gram and Continuous Bag of Words was given. An extension of Word2Vec, namely Paragraph Vector (Le and Mikolov, 2014) followed together with details on its corresponding architectures, Distributed Memory Model and Distributed Bag of Words. Lastly, the limitations of word embeddings were covered by discussing their impact on the data selection task.

Chapter 3

Machine Translation

Machine Translation (MT) is a field of computational linguistics that deals with the automatic translation of a text from a source language to a target language. Training of an MT system relies on learning from a bilingual corpus, which is a collection of aligned sentence pairs in the source and the target language. This chapter describes the two MT paradigms used in my experiments and presents an insight into MT evaluation.

3.1 Statistical Machine Translation

This section describes the core concepts of phrase-based Statistical Machine Translation (SMT) as presented in Koehn (2010b) together with references to the tools that implemented them.

The SMT model is built using a combination of several components (the alignment model, the phrase-based model, the reordering model and the language model). Combining all the components is done using the noisy-channel model which was developed by Shannon (1948). A message is delivered through a noisy channel to the receiver. The noise corrupts the message, thus the purpose is to reconstruct the message using knowledge about the distortions that occurred due to the channel. Adapting this scenario to MT, the assumption is that a sentence (message) e in a language got distorted via the noisy channel resulting in a sentence f in another language (Koehn, 2010b). Given the channel represented by the translation model $p(f|e)$, and the language model for the distorted sentence $p(e)$, the task of machine translation is to construct the initial sentence, e. Using Bayes rule[1], the best translation $\hat{e}$ for an input sentence f is (Koehn, 2010b):

$$\begin{aligned}\hat{e} &= argmax_e p(e|f) \\ &= argmax_e \frac{p(f|e)p(e)}{p(f)} \\ &= argmax_e p(f|e)p(e)\end{aligned}$$

[1] The denominator $p(f)$ is dropped due to being a constant for all e

The components of the channel, namely the translation and the language models, will be described after introducing word alignments.

As a first step, the parallel training corpora are preprocessed using tokenization, lowercasing and cleaning, which consists of removing too short or too long sentences.

3.1.1 Word Alignment

Given a sentence pair from a parallel corpus, the objective of a word alignment model is to map the words from the source sentence to their translations from the target sentence.

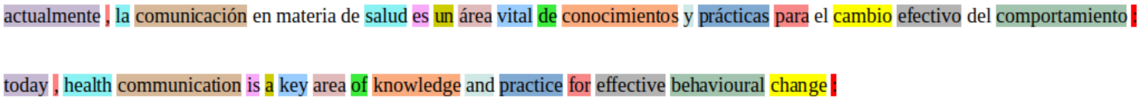
In all of my experiments, GIZA++ (Och and Ney, 2003) was employed for word alignment. As a first step, vocabularies are extracted and an integer is assigned to every word. The preprocessed parallel corpora are converted into a numeric format and the vocabularies are clustered into word classes. Two alignment files are generated, one for each language pair direction (source to target and target to source). Alignment points indicate which words from a source sentence correspond to words from its counterpart target sentence (and vice versa). A heuristics is applied in order to combine the two alignments and obtain the final one. In all of the experiments I used the default *grow-diag-final-and* which computes the intersection of the two alignments with additional alignment points for symmetrization (Och and Ney, 2003; Koehn, 2010b) .

```
First alignment file

Sentence pair (33) source length 23 target length 17
today , health communication is a key area of knowledge and practice for effective
behavioural change :
NULL ({ }) actualmente ({ 1 }) , ({ 2 }) la ({ }) comunicación ({ 4 }) en ({ })
materia ({ }) de ({ }) salud ({ 3 }) es ({ 5 }) un ({ 6 }) área ({ 8 }) vital ({
7 }) de ({ 9 }) conocimientos ({ 10 }) y ({ 11 }) prácticas ({ 12 }) para ({ 13
}) el ({ }) cambio ({ 16 }) efectivo ({ 14 }) del ({ }) comportamiento ({ 15 }) : ({ 17 })

Second alignment file

Sentence pair (33) source length 17 target length 23
actualmente , la comunicación en materia de salud es un área vital de conocimientos
y prácticas para el cambio efectivo del comportamiento :
NULL ({ 5 7 18 21 }) today ({ 1 }) , ({ 2 }) health ({ 3 8 }) communication ({ 4 6
}) is ({ 9 }) a ({ 10 }) key ({ 12 }) area ({ 11 }) of ({ 13 }) knowledge ({ 14 }) and
({ 15 }) practice ({ 16 }) for ({ 17 }) effective ({ 20 }) behavioural ({ 22 }) change
({ 19 }) : ({ 23 })
```

Figure 3.1: Example of word alignment taken from one my experiments using GIZA++

In Figure 3.1 an example from the GIZA++ alignment files is presented for a Spanish-English sentence pair[2] where for instance, the Spanish word *salud* is aligned with the third word in the English sentence. The two sentences have different lengths, therefore there are Spanish words that have no correspondent in the English sentence: in the snippet of the first alignment file, the words *la* and *materia*. However, in the second alignment file, the word *health* is aligned with two Spanish words, *la* and *salud*.

After symmetrization (Och and Ney, 2003; Koehn, 2010b), the alignment for the sentence pair presented in the example is depicted in Figure 3.2 where I marked every pair of aligned words with an individual color. The words left unaligned are unmarked. This is the sentence alignment for the presented example in 3.1.



actualmente , la comunicación en materia de salud es un área vital de conocimientos y prácticas para el cambio efectivo del comportamiento .

today , health communication is a key area of knowledge and practice for effective behavioural change .

Figure 3.2: Example aligned sentence pair visualized by means of GIZA++

3.1.2 Language Model

Intuitively, a sentence could be translated word by word by means of a dictionary. However, lexical translation has severe limitations that affects both the meaning and fluency: frequent one-to-many mappings and many-to-one mappings where one word in the source language can be translated using multiple words in the target language (and vice-versa). Moreover, a word in a source language can have different translations in a target one. The purpose of a language model is to take a sentence in a language (the target language in an MT scenario) and give the probability of that sentence being uttered by a native speaker in that language (Koehn, 2010b).

The probability distribution of a sequence of n words $W = w_1, w_2, \dots, w_n$ in a language L constitutes a statistical language model (LM). Only statistical language models are covered in this thesis since it is the type of LM used in my experiments[3]. The probability of W is estimated using the chain rule (Koehn, 2010b):

$$p(w_1, w_2, \dots, w_n) = p(w_1)p(w_2|w_1)\dots p(w_n|w_1, w_2, \dots, w_{n-1})$$

Given the assumption that the calculation of the probability for a next word w_n is only influenced by a number of previous words m, the calculation of the word probability distributions is simplified to (Koehn, 2010b):

$$p(w_n|w_1, w_2, \dots, w_{n-1}) \simeq p(w_n|w_{n-m}, \dots, w_{n-1})$$

[2] Taken from one of my experiments.

[3] Probabilistic language models are referred to with the general abbreviation LM in order to avoid specifying every time its type.

In practice, the history of words considers a small value of m, typically between three (trigrams) and five (Koehn, 2010b). For example, the trigram $p(w_3|w_1, w_2)$ is estimated by counting how many times the sequence w_1, w_2, w_3 appeared in the training corpus divided by the sum of how many times the sequence w_1, w_2 followed by any word appeared:

$$p(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3)}{\sum_w count(w_1, w_2, w)}$$

For language model estimation I used the SRILM (Stolcke, 2002) and the KENLM (Heafield, 2011) toolkits. The order of the LM was set to 5 for all of my experiments (5-gram LMs). LM interpolation between the in-domain and the general domain corpora was exploited, which empirically has been proved beneficial for domain adaptation (Koehn and Schroeder (2007); Duma and Vertan (2013)).

3.1.3 Translation Model

A phrase-based SMT model learns a phrase translation table which is built by creating a word alignment between the sentence pairs from the parallel training corpus. Afterwards, it uses this alignment model to it extract phrase pairs. For all the word-aligned sentence pairs from the training corpus, the phrase extraction algorithm extracts the consistent phrase pairs together with the sentence alignment. Following the definition from Koehn (2010b), given a word alignment $\mathcal{A}$ and a phrase pair $(\bar{f}, \bar{e})$, where $\bar{f}$ is a source phrase and $\bar{e}$ is a target phrase, the phrase pair is $\mathsf{consistent}$ if for all the words in the source phrase that have alignment points in $\mathcal{A}$ the corresponding target word from the alignment point is contained in the target phrase and vice versa. The alignment points represent constraints for extracting phrase pairs (Koehn, 2010b).

$$\begin{aligned} (\bar{f}, \bar{e}) \text{ consistent with } \mathcal{A} \Leftrightarrow & \\ & \forall e_i \in \bar{e} : (e_i, f_j) \in \mathcal{A} \Rightarrow f_j \in \bar{f} \\ \textit{AND} \quad & \forall f_j \in \bar{f} : (e_i, f_j) \in \mathcal{A} \Rightarrow e_i \in \bar{e} \\ \textit{AND} \quad & \exists e_i \in \bar{e},\ f_j \in \bar{f} : (e_i, f_j) \in \mathcal{A} \end{aligned}$$

The algorithm for phrase extraction (Koehn, 2010b) is presented on page 21 where indexes in the source (f_{start} and f_{end}) and in the target sentences (e_{start} and e_{end}) act as a sliding window over the sentence pair. The alignment points indicate how to build the phrase pairs starting with the smallest unit (a word-to-word alignment) transformed into a phrase pair and increasingly adding more alignment points that connect to the previous one, thus generating multiword phrases (Koehn, 2010b).

Figure 3.3 presents an example of phrase extraction [4] where the matrix depicts the word alignments together with a list of phrase pairs in the order they were extracted.

[4] Taken from one of my experiments and depicted using a similar fashion from (Koehn, 2010b, p. 134)

Algorithm 1 Phrase pair extraction from (Koehn, 2010b)

```
function EXTRACT($f_{start}, f_{end}, e_{start}, e_{end}$))
    for all $(e, f) \in \mathcal{A}$ do
        if $f_{start} \leq f \leq f_{end}$ and $(e < e_{start}$ or $e > e_{end})$ then return {}
    E = {}
    $f_s = f_{start}$
    repeat
        $f_e = f_{end}$
        repeat
            add phrase pair $(e_{start} \ldots e_{end}, f_s \ldots f_e)$ to set E
        until $f_e$ aligned
        $f_e = f_e - 1$
    until $f_s$ aligned
    return E

function PHRASE_EXTRACTION($e, f, \mathcal{A}$)
    BP = {}
    for $e_{start} = 1 \ldots length(e)$ do
        for $e_{end} = e_{start} \ldots length(e)$ do
            $(f_{start}, f_{end}) = (length(f), 0)$
            for all $(e, f) \in \mathcal{A}$ do
                if $e_{start} \leq e \leq e_{end}$ then
                    $f_{start} = min(f, f_{start})$
                    $f_{end} = max(f, f_{end})$
            add $EXTRACT(f_{start}, f_{end}, e_{start}, e_{end})$ to set BP
    return BP
```

The translation model consists of the phrase translation table which contains phrase translation probabilities estimated by the relative frequency:

$$\phi(\bar{f}, \bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} count(\bar{e}, \bar{f}_i)}$$

where the numerator represents how often the phrase pair appeared in all the sentence pairs and the denominator sums up the frequency of the target phrase together with all the extracted phrase pairs that contain it (Koehn, 2010b).

An example of a snippet from a phrase table is given in Table 3.1 where the first column represents Spanish phrases, the second column English ones and the third column is the probability of the Spanish phrase to be translated into the English one[5]:

[5] Snippet extracted from one of the phrase tables trained in my experiments.

	¿	cómo	se	usa	aerius	?
how	■	■				
is			■			
aerius					■	
used				■		
?						■

Extracted phrase pairs:

(how	¿ como)
(how is	¿ como se)
(how is aerius used	¿ como se usa aerius)
(how is aerius used ?	¿ como se usa aerius ?)
(is	se)
(is aerius used	usa aerius)
(is aerius used ?	se usa aerius ?)
(aerius	aerius)
(aerius used	usa aerius)
(aerius used ?	usa aerius ?)
(used	usa)
(?	?)

Figure 3.3: Example of extracted phrase pairs using Algorithm 1

Spanish phrase	**English phrase**	**p(e|f)**
al mecanismo de	mechanism of	0.931
al mecanismo de	the mechanism of	0.92
al mecanismo de	by the mode of	0.6
al mecanismo de acción de	the mechanism of action of	0.894
al mecanismo de acción de	the way	0.2

Table 3.1: Examples of phrases and their probabilities

3.1.4 Reordering Model

Often word order in the source language differs from the one in the target language. Phrase translations are able to capture such word order differences to a certain degree, but there is no guarantee that by adding the next phrase to the partial translation the result would be fluent in terms of word order. A lexicalized reordering model is used to learn the reordering preference for each phrase pair. During phrase pair extraction it gathers knowledge about three types of phrase orientations by means of alignment points (Koehn, 2010b):

- monotone: if a word alignment point exists to the top left
- swap: if a word alignment point exists to the top right
- discontinuous: neither monotone nor swap

Given the example from Figure 3.3, the three cases of orientation are depicted in Figure 3.4 [6] .

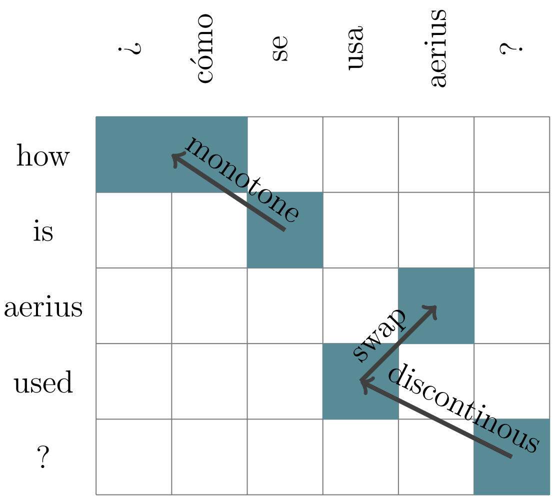


Figure 3.4: Example of orientation types

For any phrase pair $\bar{f}, \bar{e}$, the maximum likelihood principle is used in calculating the probability distribution for each orientation type $p_o(orientation|\bar{f}, \bar{e})$ and at decoding time, each orientation type is considered as a single feature function. The calculation is given below (Koehn, 2010b):

$$p_o(orientation|\bar{f}, \bar{e}) = \frac{count(orientation, \bar{e}, \bar{f})}{\sum_o count(o, \bar{e}, \bar{f})}$$

For the experiments presented in this thesis, the probabilities of bilingual phrases with the three orientations are considered for both translation directions (as presented in (Koehn, 2010b)).

[6] Depicted using a similar fashion from (Koehn, 2010b, p. 143)

3.1.5 Decoding

Given the phrase-based model, the reordering model and the language model, the best translation $\hat{e}$ for the foreign sentence f can be obtained according to (Koehn, 2010b):

$$\hat{e} = argmax_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ p_{RM} \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1})$$

The search for the best translation is referred to as decoding. Because it is computationally too expensive to search for all possible translations given an input, heuristic search methods are applied. Building a translation is done incrementally, starting from the first word in the input sentence and gradually expanding the translation by adding more words. Partial translations (hypotheses) are built, which are given partial scores computed using the probabilities obtained with the translation, reordering and language models. After all active hypotheses have been expanded, the one with the highest probability score is considered to be the best translation (Koehn, 2010b).

3.1.6 Tuning of parameters

The components can be weighted in order to give more importance to the language model, for example. Parameter tuning is an important step in finalizing an SMT model and consists of learning the feature weights using a development set (unseen data), which is considered to be a true representation of the test set, and an evaluation metric, which helps guiding the tuning algorithm towards optimal weights by measuring the errors (Koehn, 2010b, p. 264).

Given the development set, the SMT system is used to translate the set using the default values for the parameters to be adjusted. This results in a collection of the top n best translations for every sentence from the set (n-best list). The space of possible parameter settings is explored, usually using the Powell search (Powell, 1977), and the effect on the n-best lists is investigated through measuring the translation error. The Powell search explores the space of parameter values by adjusting one parameter at a time; if the translations get better then the parameter value is considered optimal and another parameter is chosen to be adjusted (Koehn, 2010b). For my experiments, tuning of the SMT systems was done with MERT (Minimum Error Training Rate)(Och, 2003) using BLEU (Papineni et al., 2002) as the evaluation metric, which is defined in Section 3.3.

The state-of-the-art SMT toolkit Moses (Koehn et al., 2007) together with the Experiment Management System (EMS) was used for my experiments. The EMS encapsulates the SMT pipeline for running an experiment into configuration files (Koehn, 2010a). The pipeline for running an SMT experiment in Moses consists of four steps: preprocessing of the training corpora, building the language and the translation models, tuning and evaluating the system.

3.2 Neural Machine Translation

This thesis centers on the research and development of data selection methods for Statistical Machine Translation. To show that my results are not restricted to this particular type of MT architecture, I also demonstrate the applicability of my data selection algorithms to Neural Machine Translation (NMT) which emerged more recently (in Chapter 5, Section 5.2.3). However, NMT is not the core MT architecture of this thesis and its further exploration is beyond the scope of my work. This section briefly describes NMT.

Neural Machine Translation (NMT) is based on an encoder-decoder architecture built with recurrent neural networks (RNN), proposed by Cho et al. (2014) and Sutskever et al. (2014). It uses word embeddings for sentence representations. In contrast to a phrase-based SMT, it does not require the training of a suite of models; instead a single sequence-to-sequence model is trained.

In the following, the components of an NMT model are shortly described: the encoder, the decoder and the attention mechanism. For a detailed description of the architecture, see Koehn (2017).

The encoder is a recurrent neural network used for encoding the source sentences. Given an input sentence, each word is represented using a one-hot encoding that is also used in learning word embeddings (see Chapter 2). The encoder network combines the word vector representation for a current word w_i with the representation of its context vector. For example, for the sentence "inflammation was treated with steroid injections", the network combines the word embedding for inflammation with the embedding of was, resulting in a representation $e(\mathit{inflammation\ was})$ which is afterwards combined with the embedding of the next word, producing the vector for $e(\mathit{inflammation\ was\ treated})$, till in the end the whole sentence is represented as a unique vector. Therefore, every word is encoded considering the left context (left-to-right RNN) and usually, also considering the a right context (right-to-left RNN), an architecture referred to as a bidirectional RNN (Bahdanau et al., 2014).

The decoder is also a recurrent neural network that uses the source sentence vector representation together with word predictions and context vectors in order to predict the next word to be translated (the word with the highest probability in the output vocabulary). When an error occurs (the predicted word is not the actual target word), the wrong prediction is kept and the network forces the next input to be the actual target word. The accuracy of the predicted translation is calculated using a loss function, which is usually the Cross-Entropy (sum of negative log likelihoods of correctly predicted words). The loss is used in updating the weights of both the encoder and the decoder.

It has been shown that the translation accuracy can be improved using an attention mechanism (Bahdanau et al., 2014). The aforementioned context vector acts as a memory built using the hidden states of the encoder. Therefore, instead of using only the final state of the encoder, all the previous states contribute to predicting the next translated word. The motivation for using this mechanism is that a closer connection between the decoder and the input words is desired (Koehn, 2017).

In order to deal with the problem of translating unknown words, word segmentation was introduced by Sennrich et al. (2016a) using the Byte Pair Encoding algorithm. The assumption is that some rare words can be translated by breaking them down into subword units. For example, the German sentence ”wurde der authentifizierungscode zurückgewiesen .” becomes ”wurde der authentifizierungs@@ code zurückge@@ wiesen .”, where authentifizierungscode was identified as a rare word and split using the @@ markup into smaller units. Also the sentence ”maximale uploadgeschwindigkeit in kib / s” becomes ”maximale up@@ load@@ geschwindigkeit in ki@@ b / s” where uploadgeschwindigkeit was marked as a rare word. For details on the algorithm, see Sennrich et al. (2016a).

Translating unseen sentences (or test sets) is done via inference. During inference, at every step one word is generated. After the input sentence is encoded, the final state of the encoder is used as an initial state for the decoder. At each time step, the states of the decoder are used as initial states for the next time step and the predicted output is used as input. The main difference between training and inference is that there are no target words to be forced on the decoder. For more details on NMT see Koehn (2017).

The Tensorflow NMT (Luong et al., 2017) implementation was used in conducting the NMT experiments[7]. Most of the default hyperparameter values were used in my experiments, with exceptions that are noted where the experiments are described. Following the recommendations from Britz et al. (2017b), I have chosen bidirectional encoders as they usually outperform unidirectional ones and 4 layers for both the encoder and the decoder.

Currently[8] the state-of-the-art for MT is based on transformers (Vaswani et al., 2017), architecture that is briefly described in in Chapter 8, Section 8.3 where future work is presented.

3.3 Machine Translation Evaluation

Currently, machine translation cannot provide perfect translations for every sentence, for every language pair. Even large amounts of training data cannot cover the whole vocabulary of a language, resulting in MT output that contains untranslated words (out-of-vocabulary words). The training data itself could be noisy, for example containing misaligned sentences, sentences in another language than the source and target languages, sentence pairs in the same language, incomplete sentences, machine translated sentences, typos or misspellings. MT presents difficulties in disambiguating word senses and correctly choosing lexical translations. Even though automatic tools could be applied to clean the data, this does not guarantee that all errors are eliminated. The MT models learn from this noisy data and inevitably produce errors when decoding. Also, if decoding is not able to explore all hypotheses due to a high computational cost (Koehn, 2010b), that might lead to some good translations being left out.

[7] Available at `https://github.com/tensorflow/nmt`.

[8] Referring to the date of writing this section, May 2021.

It is not only important to know how good an MT system performs, but also a ranking between two or more MT system variants should be determined, with the purpose of improving the MT quality. Measuring the quality of an MT output is a highly debated topic in the community, which lead to producing a collection of evaluation methods that also are far from being perfect. This chapter covers methods for the manual and the automatic evaluation of MT output, the correlation between the two evaluation types and significance tests which are useful in assessing whether an increase in quality as measured by automatic evaluation metrics is statistically significant.

3.3.1 Manual Evaluation

This section briefly presents the manual evaluation methods used so far in the MT community.

In an ideal setting, the output of an MT system should be evaluated using human evaluators (judges) that are bilingual, thus understanding both the source and the target language, are specialized in translation and ideally, are familiar with the terminology of the domain the MT is applied to. In practice, it is difficult to find fluent bilingual judges, and even more challenging, to find persons that fit the other two criteria. Translation is highly subjective and a sentence can be translated in more than one way. The example from Figure 3.5 is taken from Koehn (2010b). Ten human translations obtained from ten different persons are provided for a Chinese source sentence. More than one judge is needed in order to better evaluate a system. However, having more than one judge introduces another difficulty, namely annotator disagreement.

The consistency and reliability of the evaluators is measured through the intra- and inter-user agreement and predominantly using κ, the Cohen's Kappa coefficient (Cohen, 1960).

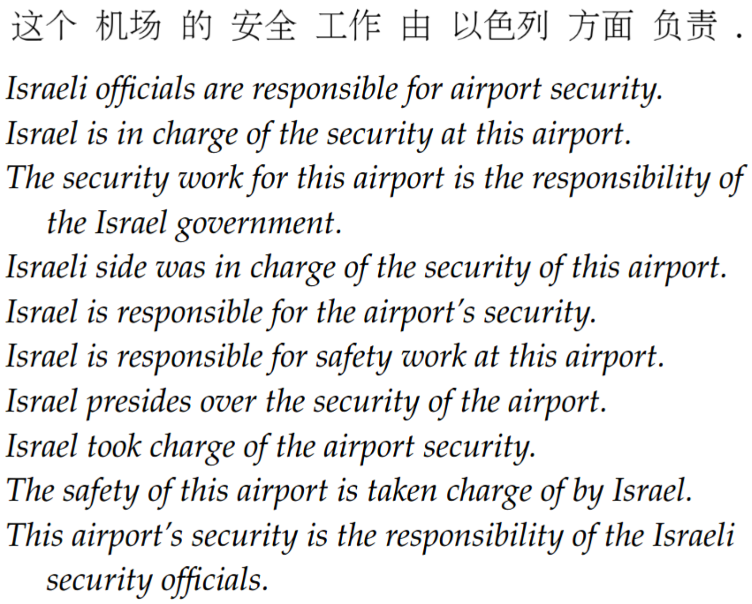
The formula for calculating the κ score is presented below:

$$\kappa = \frac{P(o) - P(e)}{1 - P(e)}$$

where $P(o)$ is the proportion of pairwise comparisons for which the two annotators agreed. Given system A and system B, the possible agreement types are:

- **A** $<$ **B** : translation **A** is better than translation **B**
- **A** $=$ **B** : translation **A** has the same quality as translation **B**
- A $>$ **B** : translation **A** is worse than translation **B**

All system comparisons that were judged twice by two annotators are investigated and the proportion of times where $A > B$, $A = B$ and $A < B$ is calculated. Similarly, for intra-agreement all pairwise comparisons annotated twice by the same annotator are considered. The proportion of times the raters would agree



这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Figure 3.5: Examples of human translations for a sentence from the 2001 NIST evaluation campaign (taken from (Koehn, 2010b)

by chance, $P(e)$, is calculated as the sum of the squared probabilities that two annotators would agree on each one of the three cases (Bojar et al., 2016a).

$$P(e) = P(\text{A<B})^2 + P(\text{A=B})^2 + P(\text{A>B})^2$$

Early methods for assessing the quality of MT output relied on the fluency of the output in terms of grammatical errors and lexical choices, and on the adequacy of the output, which focuses on how much of the meaning of the input sentence is transfered to the output (Koehn, 2010b). Both fluency and adequacy could be rated by an annotator with values ranging from 1 (no meaning preserved/ incomprehensible) to 5 (flawless/ all meaning preserved). Even though these measurements were popular and used in WMT campaigns, they have some disadvantages. As pointed out by Koehn (2010b), some annotators tend to give average scores of 4, while other average scores of 2. Normalizing all the judgments from all evaluators i.e. bringing all average scores per evaluator to the same value (by adding an adjustment value) fixes the latter problem. However, the definitions are vague and it is difficult for annotators to be consistent Koehn (2010b).

More recent methods focus on directly comparing the quality of two or more MT systems via three-way-ranking (Koehn, 2010b) (for comparing two systems) and ranking (for more systems). For ranking, the evaluator is presented with the input sentence and optionally, a reference translation, together with the MT outputs. The task is to identify which system performed better than the other ones by ranking them. The evaluation of the WMT Biomedical task (Jimeno Yepes et al., 2017; Neves et al., 2018; Barrault et al., 2019) relies on the three-way-ranking method and I will use this method when performing manual evaluation for my systems.

In the three-way-ranking procedure, an annotator is presented with three sentences: the input sentence in the source language and two sentences in the target language corresponding to two system translations for the input sentence. The first translation is named **A** and the second one **B**. The options from which the user can choose are the same as previously described: $A > B$, $A = B$ and $A < B$ (Koehn, 2010b, p. 220).

To sum up, manual evaluation is an ideal way of evaluating MT systems, but it has several limitations: the annotators can disagree, it is expensive and highly time consuming, it is subjective and not reusable on other translations.

3.3.2 Automatic Evaluation

To compensate the shortcomings of manual evaluation, automatic evaluation methods have been introduced. An MT evaluation metric always produces the same score, is cost-free, yields instant results, is objective and can be applied on other translations as many times as needed (as opposed to manual evaluation). In the following, the most frequently used evaluation metrics are described.

BLEU

Given one or more references, the most commonly used metric in MT evaluation, BLEU (Papineni et al., 2002), uses matching of words or continuous sequences of words between the translation (MT output) and the reference(s), namely n-gram matching. This metric is based on the commonly used precision metric defined as the number of words from the translation that occur in any of the references divided by the number of words in the translation). The problem that n-gram matching poses is that in case of very long translations, there is a higher probability that some words will match with words from the reference, thus leading to high recall (low precision). On the other hand, if the translated sentence is very short but containing some words that appear in the reference, it will have high precision (low recall) (Koehn, 2010b).

In order to overcome these problems, BLEU uses a modified n-gram precision which lowers the count of a correct n-gram to its maximum total count in any of the references. These modified counts are summed up for every distinct word in the candidate translation and divided by the total number of n-grams in the candidate translation (Papineni et al., 2002).

In addition to the modified precision, a brevity penalty was introduced in order to reduce the BLEU score for too short translations. Given the length of the candidate translation, lc, and the length of the reference translation, lr, the formula for calculating the brevity penalty, BP, is given below:

$$BP = \begin{cases} 1, & if\ lc > lr \\ e^{(1-\frac{lr}{lc})}, & else \end{cases}$$

Combining the n-gram modified precision and the brevity penalty, the formula for calculating the BLEU score for a given n is using the product of the brevity penalty and the geometric mean of the n-gram precisions using positive weights that sum up to 1 (Papineni et al., 2002):

$$BLEU_n = BP \cdot exp\left(\sum_{i=1}^{n} w_i \log precision_i\right)$$

Below is an example of modified precisions for unigrams using a translation from one of my NMT experiments (sentences are given in preprocessed form). Using the non-modified precision metric, the sentence precision of 7/8 is fairly high, because the word umbenennen appears three times in the translation, whereas the modified precision results in a lower value of 5/8 that better reflects the quality of the output.

Source	choose rename , the filename is highlighted in blue .
Reference	wählen sie umbenennen , wird der dateiname in blau hervorgehoben .
NMT translation	wählen sie umbenennen , umbenennen und umbenennen .

Table 3.2: Example of translation that benefits from modified precision

The default BLEU uses 4-grams and uniform weights of 1/4. The scores range from 0 (very bad translation) to 1 (translation identical to reference). Lavie (2010) indicates that BLEU scores above 0.3 reflect understandable translations, while scores over 0.5 are considered fluent translations. Punctuation tokens are considered as words.

Since BLEU only matches words, it can give a low score to a good translation that uses synonyms or other words that share the same meaning with the corresponding reference. For example, for the input sentence "Right-click the file icon", the translation "Klicken Sie mit der rechten Maustaste auf die gewünschte Datei" is perfectly fluent and equivalent in meaning with the reference "Rechtsklick auf die gewünschte Datei", however the translation receives a poor BLEU score of 0.23.

Callison-Burch et al. (2006) criticizes BLEU contesting that an actual improvement in translation quality depends on improving the BLEU score. Moreover, the paper points out the limitation of this metric which sometimes assigns the same BLEU to translations of different quality. Also, the paper provides empirical evidence that this metric does not always correlate well with human judgments.

TER

An intuitive approach to evaluate MT output would be to determine the number of changes that need to be applied to the hypothesis to obtain the reference. Snover et al. (2006) introduces this method under the name of TER (Translation Edit Rate)[9]. There are four types of edits desirable for obtaining the exact matching with the reference: word insertion, word deletion, word substitution and phrasal

[9] Also referred to as Translation Error Rate in Dorr et al. (2011)

shift. I refer to the first three as primary operations, to avoid enumerating them. The latter operation moves a contiguous sequence of words from the hypothesis to another location within the hypothesis. All the edits have an equal cost of 1, punctuation tokens are considered as words and wrong capitalization is also treated as an edit (Snover et al., 2006). As opposed to BLEU, lower scores of TER account for better translations. The calculation of TER is:

$$\mathit{TER} = \frac{\text{number of edits}}{\text{average number of reference words over all references}}$$

Given a hypothesis and one or more references, the algorithm for calculating its TER score follows two phases: the number of primary operations is calculated using dynamic programming, while the set of shifts is found by a greedy search (repeatedly select the shift that reduces the number of primary operations most, until no shifts that reduce the edit distance remain) (Snover et al., 2006). When given a set of references, the number of edits is calculated for each reference and the hypothesis. The lowest one is selected as the final number of edits.

Source	click , hold , and drag the mouse until all of the cells are highlighted , then release the mouse .
Reference	klicken , halten und ziehen sie die maus , bis alle zellen markiert sind , dann lassen sie die maus los .
NMT translation	klicken , halten und ziehen sie die maus , bis alle zellen hervorgehoben werden , dann lassen sie die maus los .
TER score	**0.0909** due to two *substitutions* (reference length of 22 → score is 2/22).

Source	select the file and click preview .
Reference	wählen sie die datei und klicken sie auf vorschau .
NMT translation	wählen sie die datei aus und klicken sie auf vorschau .
TER score	**0.1** due to one *deletion* (reference length of 10 → score is 1/10).

Source	the 19-character is much better .
Reference	das mit 19-zeichen ist viel besser .
NMT translation	das 19-zeichen ist viel besser .
TER score	**0.1428** due to one *insertion* (reference length of 7 → score is 1/7).

Source	open the file , then click file > export and select your desired conversion format .
Reference	öffnen sie die datei , klicken sie dann auf datei > exportieren und wählen sie das gewünschte konvertierungsformat .
NMT translation	öffnen sie die datei , dann klicken sie auf datei > exportieren und wählen sie das gewünschte konvertierungsformat .
TER score	**0.0526** due to one *shift* (reference length of 19 → score is 1/19).

Table 3.3: Examples of edit operations with their TER scores

Table 3.3 presents examples of the four edit operations extracted from the output of one of my NMT experiments for the IT domain.

TER shares the same limitations as BLEU as it is a purely lexical evaluation metric that does not consider semantic equivalence. In order to overcome this problem, Snover et al. (2006) introduces a human component into TER using human annotators to post-edit the hypothesis and thus, generating a new targeted reference. In this manner, the number of edits required for the hypothesis to reach the new reference is calculated as hTER (Human-Targeted Translation Edit Rate) using the same algorithm as described before. The clear advantage of this method is that translations containing synonyms or highly related phrases in meaning with the reference(s) obtain lower TER scores when the targeted reference also contains them while the initial reference does not. However, this metric is semi-automatic and therefore, highly dependent on the availability of annotators that are fluent in the target language, which as discussed previously in the Manual Evaluation Section is time consuming and not portable to other MT system outputs.

METEOR

An MT evaluation metric that attempts to overcome the limitations of the previously described ones is METEOR (Metric for Evaluation of Translation with Explicit Ordering) (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007)[10]. It correlates well with human judgments and it incorporates synonyms and stems.

Given a hypothesis sentence (translation) and a reference sentence, this metric creates a monolingual word alignment between them that is incrementally produced by a sequence of match modules (Banerjee and Lavie, 2005):

- exact: maps two words if they are exactly the same
- stem: maps two words if they have the same stem
- paraphrase: maps two words/ phrases using a paraphrase table
- synonym: maps two words if they are considered synonyms according to Wordnet (Fellbaum, 1998) synsets (only for English)

In this thesis, the sequence of modules ”exact stem paraphrase” is used for reporting METEOR scores for non-English languages and with an addition of synonym for English, together with the default parameters and weights values.

An advantage of METEOR compared to the other two metrics is the introduction of function word lists (for some languages) and a parameter used to weight content words (the words not found in the functor list). The functor lists are extracted from the corpora made available by the WMT 2011 translation task[11] filtering out the words that have a very high frequency (Denkowski and Lavie, 2011).

[10] https://www.cs.cmu.edu/ alavie/METEOR/README.html

[11] http://www.statmt.org/wmt11/translation-task.html

Another enhancement of this evaluation metric is the use of paraphrases. This ensures that hypotheses which share meaning with the reference, but use different words, contribute positively to the final score (Denkowski and Lavie, 2010). For example, the German paraphrases list used by METEOR includes the pairs ("1 bis 2", "ein bis zwei"), ("1 bis 2", "zwischen 1 und 2") and the pairs ("die soziale sicherheit", "die sozialversicherung"), ("die soziale sicherheit", "sozialschutz").

As follows, in this section, the formulas for calculating the METEOR score are taken from (Denkowski and Lavie, 2011). The weighted precision P and recall R, the parameterized harmonic mean (van Rijsbergen, 1979) of P and R, F_{mean}, the fragmentation penalty and finally, the METEOR score are presented. Given a reference, a hypothesis and their word alignment, the functor list is used to identify the content and function words in the hypothesis (h_c, h_f) and reference (r_c, r_f). The number of content and function words is counted for each of the match modules m_i with:

- the hypothesis $m_i(h_c)$ and $m_i(h_f)$
- the reference $m_i(r_c)$ and $m_i(r_f)$

The fragmentation penalty accounts for gaps and differences in word order and is calculated using a contiguous series of matches that is identically ordered in the hypothesis and the reference (chunks).

The weights used in calculating precision and recall are given for every match module. When reporting METEOR scores in this thesis, the default weight values were used: $w_{exact} = 1$, $w_{stem} = 0.8$ and $w_{paraphrase} = 0.2$. Also, the default values were used for the parameter values: the parameter needed for the parameterized harmonic mean $\alpha = 0.95$, the parameters for the fragmentation penalty $\beta = 1$ and $\gamma = 0.55$ and the content-function word parameter $\delta = 0.55$. These parameters can be tuned when human annotations are available in order to obtain a better correlation with human judgments.

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|}$$

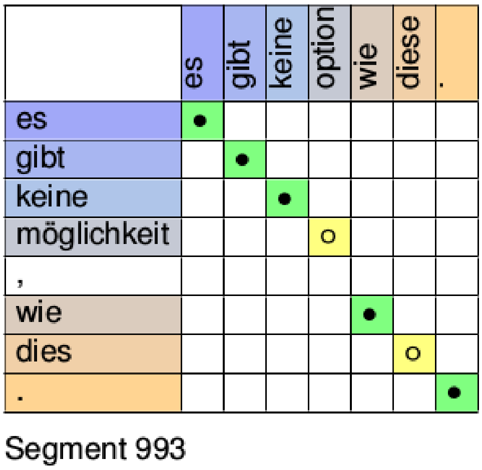
$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|}$$

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$FragmPen = \gamma \cdot \left(\frac{chunks}{matches}\right)^{\beta}$$

$$Score = (1 - FragmPen) \cdot F_{mean}$$

The figures that follow were generated with the METEOR tool and present on the top the reference, on the left the hypothesis and statistics including the precision, the recall, the fragmentation penalty and the final evaluation score. The green squares identify exact word matches, while the yellow ones stem or paraphrase matches.



P:	0.762
R:	0.853
Frag:	0.157
Score:	0.715

Figure 3.6: Example of scoring with METEOR

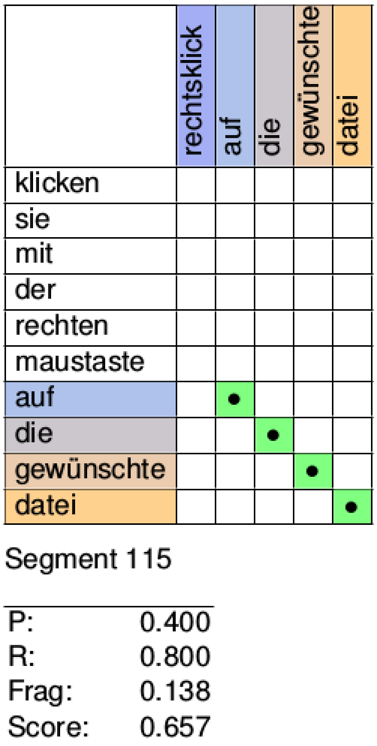
Figures 3.6 and 3.7 present examples of word alignments between the reference and the hypothesis for three sentences from the English to German test set of the WMT 2016 IT domain task[12]. The translations were obtained with one of my NMT experiments. For the sentence from the first example (Figure 3.6), all hypothesis tokens are matched except for the comma token. The yellow boxes indicate a paraphrase match between the reference word option and the hypothesis word möglichkeit and a stem match between diese and dies.

METEOR also exhibits shortcomings due to its use of paraphrase tables and use of synonymy only for English. Figure 3.7 depicts an example where the hypothesis receives a relatively low score of 0.657 even though it is equivalent in meaning with the reference. The German word rechtsklick from the reference actually means klicken mit der rechten maustaste, the sequence of words that appears in the MT translation. Since this pair is not contained in the paraphrase table for German, the final score is affected.

Despite overcoming all the drawbacks of the human evaluation methods, the automatic ones have their own limitations (as previously presented for each of the three MT metrics) (Koehn, 2010b):

- when the human reference differs from the MT output it gives a poor score to the MT system even if the translation was good

[12] http://www.statmt.org/wmt16/it-translation-task.html



	rechtsklick	auf	die	gewünschte	datei
klicken					
sie					
mit					
der					
rechten					
maustaste					
auf		●			
die			●		
gewünschte				●	
datei					●

Segment 115

P:	0.400
R:	0.800
Frag:	0.138
Score:	0.657

Figure 3.7: Example of scoring with METEOR that fails in giving a reasonably high score

- sometimes they do not correlate well with human judgments
- they need to be evaluated as well which poses another challenge and field of research

3.3.3 Statistical Significance

In Machine Translation not only the performance of an individual system needs to be assessed, but also whether it performs better or worse than other systems. If system A obtained a higher BLEU score compared to system B , hypothesis testing is applied in order to determine if system B significantly outperforms system A in terms of BLEU. But it could be that B obtained a better score due to the composition of the test set and that an evaluation with another test set might lead to the opposite result.

The most commonly used method in the MT community for hypothesis testing is bootstrap resampling (Koehn, 2010b). Given a test set, 1000 samples are extracted and the BLEU score (or other metric) is computed for each of the samples, for both systems that should be compared. The 25 highest and the 25 lowest BLEU scores are ignored. If one system outperforms the other one for at least 950 of the samples, then it is considered to be statistically significant better at a *p-value* $\leq$ 0.05. The MTCompar-Eval tool (Klejch et al., 2015; Sudarikov et al., 2016) was used in conducting statistical significance tests for my experiments.

3.4 Summary

This chapter introduced in detail Statistical Machine Translation, namely how word alignments are generated, how the language, translation and reordering models are built. Since the current state-of-the-art MT is shifting to neural approaches, this chapter also offered a short description of Neural Machine Translation. To complete the MT flow, an overview of the state-of-the-art manual and automatic evaluation methods for MT were presented. Finally, the performance of MT systems in comparison with each other was questioned via statistical significance, which was briefly described.

Chapter 4

Data Selection

This chapter introduces domain adaptation for SMT and focuses on data selection, as a corpus level approach to domain adaption. The terminology for data selection is offered. The initial directions are presented, followed by the related work to the data selection methods I developed.

4.1 Domain Adaptation

One common premise in statistical applications is that the training data and the test data are drawn from the same distribution. However, a model can be trained on one domain and used on another one (Sankaran et al., 2012). In such a case, the test data of the application domain will be drawn from a distribution that differs from the distribution of the training data (Daumé and Marcu, 2006). To mitigate this problem, domain adaptation techniques can be applied to deal with the syntactic and semantic differences between the different data sets for training and testing.

In this thesis, the notion of domain follows the definition from Plank (2011, p. 60) where a domain is identified or defined through a corpus, similarly to the definition from Koehn and Knowles (2017) who state that "*a domain is defined by a corpus from a specific source, and may differ from other domains in topic, genre, style, level of formality, etc.*".

There is a wide range of domain adaptation techniques that fall into two categories: corpus level methods and model level methods. The corpus level methods can be further divided into data selection methods and corpus generation. The data selection methods will be covered in the next sections. The generation of parallel corpora covers data extraction from comparable corpora (Daumé III and Jagarlamudi, 2011; Irvine et al., 2013; Irvine and Callison-Burch, 2014; Chu, 2015), building synthetic corpora using information retrieval (Abdul-Rauf et al., 2016), cross-language adaptation by means of a dictionary for closely related languages (Popović and Ljubešić, 2014) or domain-focused web crawling (Lu et al., 2014; Pecina et al., 2015).

Adaptation at the model level focuses on the language model or the translation model. A discounting approach for language modeling was introduced by (Guo et al., 2014), while LM linear interpolation (weighted average of the model probabilities) was proposed by Koehn and Schroeder (2007) where more weight is given to the in-domain data. In their work, the SMT systems also benefit from factored translation models where words are represented by factors (lemma, part-of-speech, morphology). A combination technique for phrase tables, called alternative decoding paths, is used by Birch et al. (2007) where one path is the in-domain translation table and the other path is the general domain translation table. Wang et al. (2014) also use this approach in combination with a data selection technique.

For translation model interpolation, Sennrich (2012) summarizes the different approaches that the MT community adopted for choosing the weights: uniform weights (Cohn and Lapata, 2007), incremental tuning of the weight (Yasuda et al., 2008; Nakov and Ng, 2009; Axelrod et al., 2011), or weights set as a function of distance metrics (Foster and Kuhn, 2007).

Dynamic adaptation is employed by Hasler et al. (2014) where the domain of the test document is unknown. Their approach focuses on context words occurring in the same sentence. The assumption is that all phrase pairs share a topic list. Learning of the topic distributions for each phrase pair is done by representing them as documents that contain the context words from the source sentence. A topic modeling algorithm is applied using the distributional profiles of the phrase pairs. Topic modeling is also used by Su et al. (2015), Cuong et al. (2016) and Xiong et al. (2016).

Other translation adaptation techniques include weighting the phrase pairs (Matsoukas et al., 2009; Cuong and Sima'an, 2014; Mansour and Ney, 2014) by means of cross-entropy difference (Axelrod et al., 2011), or using simplification-translation-restoration (Chen et al., 2012).

A comprehensive survey of domain adaptation for Statistical Machine Translation is presented by Cuong and Sima'an (2017).

4.2 Data Selection for Statistical Machine Translation

Domain adaptation via data selection represents the core of this thesis and is presented in detail in the following sections.

This section summarizes the related work for data selection for Statistical Machine Translation. While a wide range of methods are covered, an exhaustive survey of techniques[1] on the topic is beyond the purpose of this section.

Given a large pool of sentences that belong to different domains (defined as general domain in the community) and a specific domain, data selection aims at identifying the sentences from the general domain that could be considered as belonging to the specific domain.

The terminology in data selection is defined in the MT community as follows:

[1] Refer to Eetemadi et al. (2015) for a survey on data selection for SMT.

- selection pool (D_{Gen}): a large general domain corpus (millions of sentence pairs)
- in-domain (D_{In}): a domain-specific corpus that is the target domain of adaptation
- pseudo in-domain (P_{In}): a subset of selected sentences from D_{Gen} with the property that it is highly relevant to D_{In} [2]
- ratio/ threshold: amount of sentences to select from D_{Gen} either based on a ratio (for example, 20% selection) or on a number θ (for example, all sentences that were scored higher than θ)

The task of data selection centers on determining a pseudo-in domain that helps to train an MT system tailored to output translations with a better quality as opposed to systems that are trained using the entire general domain data. Compared to the training of an MT system on the full corpus, using data selection is less memory intensive, ensures reduced usage of storage and eventually results in smaller MT system that can be ported to offline MT applications for devices like smartphones and tablets (Eetemadi et al., 2015). Moreover, higher translation quality is generally achieved.

The selection of an optimal pseudo in-domain is based on one or more scoring functions which are used in ranking all the general domain sentences. Given a sentence from the general domain, s_{Gen}, the scoring functions reflect the similarity of s_{Gen} to D_{In}. The size of P_{In} is usually determined empirically.

Unfortunately, it is extremely time and resource consuming to exhaustively evaluate all the possible pseudo in-domains. Given the size of the general domain corpus, $|D_{Gen}|$, the number of possible selections of sentences is factorial: $\frac{|D_{Gen}|}{|P_{In}|*(|D_{Gen}|-|P_{In}|)}$, for all $|P_{In}| < |D_{Gen}|$ (Eetemadi et al., 2015).

The problem of data selection can be defined formally as a constrained optimization problem (Eetemadi et al., 2015). The optimal pseudo in-domain has a maximum size, $|P_{In}|$. The translation quality needs to be maximized when using models trained on P_{In} and evaluated on an unseen test set. Below is the formulation of the problem with BLEU (Papineni et al., 2002)[3] being used as the automatic evaluation measure of the translation quality:

$$P^* = \underset{P_{In} \in D_{Gen}}{\operatorname{argmax}} BLEU(Test_{ref}, T_{P_{In}}(Test_{src}))$$

In the formulation, P^* is the optimal pseudo in-domain, $Test_{ref}$ is the reference side of the test set and $T_{P_{In}}(Test_{src})$ is the translation of the source side of the test set using an MT system trained on P_{In}.

In the following, different initial approaches to data selection for SMT are presented, as well as related work to my methods.

[2] Term introduced by (Axelrod et al., 2011)

[3] The most common evaluation measure of the MT community (Bojar et al., 2016b) used for evaluating the translation output. A description was given in Chapter 3.3.2.

4.2.1 Information Retrieval Methods

Initial work on data selection was inspired by Information Retrieval, namely the cosine TF-IDF[4] was used to assess the similarity between sentences pertaining to different domains. TF-IDF (Luhn, 1958; Sparck Jones, 1972) is formally defined as follows[5]:

$$TF(term, doc) = count(term, doc)$$

$$IDF(term, Docs) = \frac{\#Docs}{\#Docs\ containing\ term}$$

$$TF\text{-}IDF(term, doc, Docs) = TF(term, doc) \times IDF(term, Docs)$$

where given a term, *term*, a document, *doc*, and a collection of documents, *Docs*, *count* is the frequency (occurrence) of *term* in *doc*.

In the data selection pipeline using TF-IDF, each sentence from the general domain is considered a document and each sentence from the test set is considered a query (term). The sentences are represented using TF-IDF. Given a query, the cosine is applied between each document vector and the query vector, producing a ranking of the documents. The top N sentences are considered to be most similar to the in-domain and therefore, they are selected. Eck et al. (2004) pioneered this direction for data selection by applying the TF-IDF on news stories covering multiple topics for language model adaptation. They used the source side of the test set to look up similar sentences in a large news corpus and trained several adapted language models. However, language model perplexity did not significantly correlate with the translation performance, therefore the SMT experiments were used for evaluating the method.

On the same research line, Hildebrand et al. (2005) also used every sentence from the source side of the test data as an individual query for retrieving similar sentences using the cosine distance similarity. However, in contrast to Eck et al. (2004) where adapted language models are trained, Hildebrand et al. (2005) trained a translation system using the selected sentences and achieved better results compared to the baseline trained on the full data.

Offline and online data selection using TF-ID is applied by Lü et al. (2007) also following the assumption that the test data is available. For the offline adaptation, the weight of a general domain sentence is increased in accordance to the number of times it was retrieved after querying using the test sentences. Therefore, if a sentence pair appeared once in the general domain and it was retrieved by a query, then it's weight (count) increases. This weighting scheme is then applied to the input for the alignment with GIZA++ step. This results in the translation model giving higher probabilities to the adapted words. For the online scenario, the training data is split into several sub-corpora using a clustering method. Translation models are trained on these sub-corpora and additionally, one translation model

[4] Term Frequency - Inverse Document Frequency

[5] Note: there are several variants of TF; in the formula, the raw count of term occurrences in a document is given.

is trained on the full data. When querying a sentence from the test set, the top N most similar sentences are selected together with the sub-corpora information and used later on for training translation models. The resulting translation model is a linear interpolation of the sub-models and achieves similar BLEU scores to the baseline trained on full data.

The source side of the test set is translated using a baseline in Tamchyna et al. (2012). After lemmatization, the translated sentences are used for querying the general domain corpus and the most similar selected sentences are used for training an adapted language model which outperforms the baseline trained on the full general domain data.

4.2.2 Methods based on Language Model Perplexity

Another paradigm for data selection is based on the language model perplexity, which is a metric used for evaluating a language model. Before detailing related work on this data selection approach, the perplexity and cross-entropy are formally defined below (as taken from the MT community and described in (Koehn, 2010b)):

$$\text{cross-entropy} : H(W, LM) = -\frac{1}{n}\sum_{i=1}^{n} p(w_i) \log P_{LM}(w_i|w_1, ..., w_{i-1})$$

$$\text{perplexity} : PP(W, LM) = b^{H(W,LM)}$$

where w_i is the i-th word in a sentence W, $p(w_i)$ is the probability distribution of w_i, P_{LM} is the probability of a language model, LM, over a sequence of words, and b is a base, usually two. As noted by Moore and Lewis (2010), perplexity and cross-entropy are monotonically related and the lower the perplexity of a sentence, the more fluent it is.

Initial steps in the direction of language model adaptation using the perplexity is taken by Lin et al. (1997) and Gao et al. (2002) where documents are scored using the perplexity value of the document with respect to the in-domains. Moore and Lewis (2010) apply this idea to the task of data selection for SMT. The general domain sentences are ranked according to the difference between the cross-entropies computed by means of language models for the in-domain and for the general domain data[6]. Both language models are trained on the source side of the corpora.

Improvements over the cross-entropy difference approach are achieved by Axelrod et al. (2011) where language models are trained on both, the source and the target side of the corpora, leading to the following scoring formula, for a given sentence s from the general domain:

$$H(s, LM_{src}^{In}) - H(s, LM_{src}^{Gen}) + H(s, LM_{trg}^{In}) - H(s, LM_{trg}^{Gen})$$

[6] More precisely, the LM is not trained on the full-sized general domain, but on a random sample of it having the same size as the in-domain data

where In denotes the in-domain, Gen refers to the general domain, src denotes the source side of the corpus and trg the target side. Thus, LM^{In}_{src} is a language model trained on the source side of the in-domain corpus.

The methods proposed by Moore and Lewis (2010) and Axelrod et al. (2011) were widely used in the MT community as standard comparison methods as they are fast to train and achieve better results compared to baseline systems trained on the full general domain corpora. The bilingual cross-entropy difference constitutes the state-of-the-art method that I use throughout this thesis to compare my methods with. It will be referred to as *MML* (Modified Moore-Lewis) in this thesis. This method is contained in the Moses toolkit Koehn et al. (2007) and I used it in my experiments.

The cross-entropy approach suffers from the same limitation as the other data selection methods - the threshold that is used as the cutoff for forming the pseudo in-domain needs to be empirically determined, which is time and resource expensive. This problem is tackled in detail in Chapter 6.

4.2.3 Related work to domain adaptation

In order to get a better overview of related work to domain adaptation, the methods that were concurrent with mine are described. While these methods do not cover the full spectrum of domain adaptation techniques that were applied until present, they represent the methods that were compared to most of my data selection methods during WMT campaigns, thus they are highly relevant for the purpose of this section. Moreover, the developed approaches can be directly compared since most of the MT systems were trained under the same constrained condition, namely training using only the corpora made available by the task campaign.

Avramidis et al. (2016) presented a word-sense-disambiguated factored SMT with two decoding paths. In the basic path, the nouns from the English side of the training corpus are annotated with senses, while the alternative path allows for decoding when no senses have been found. A commercial rule-based system is also used for translating. These MT systems are used in a selection procedure where the MT outputs are ranked by aggregating pairwise decisions obtained using a binary classifier trained on test sets from previous WMT years campaigns. The data selection approach of Pahari et al. (2016) is based on the method introduced by Axelrod et al. (2011). Instead of the bilingual cross-entropy difference, the cross-perplexity is used (power of two). After the most relevant sentences have been selected, the MT system benefits from the interpolation of the language and translation models. A direct comparison of BLEU results between one of my data selection methods, Avramidis et al. (2016) and Pahari et al. (2016) is presented in Bojar et al. (2016a, p. 152).

Rosa et al. (2016) introduced a dictionary-based approach for domain adaptation that does not require retraining of the SMT system. They make use of a feature from Moses that allows forced translation of words/ phrases based on a markup file that suggests translations. This is achieved by means of a dictionary made available by the WMT organizers for this IT task. Then an already trained

SMT system is used for translating the enriched input file.

Among the factored models, Gaudio et al. (2016) investigated the use of lemmas for a lemma-based word alignment. They trained a hybrid MT system (rule-based and statistical modules), namely TectoMT (Popel et al., 2016). It distinguishes between surface dependency trees (a-tree) that contain all tokens from the sentences as nodes and deep dependency trees (t-tree) that contain as nodes only content words. The approach is based on the assumption that the deep structures for the source and target languages should be similar. The nodes are enriched with information like the lemma, the functor or tense. The pipeline starts with a dependency analysis where the source sentence is tokenized, tagged, parsed and named entities are detected. Hand-written rules are used for converting the dependency parse of the source sentence into a structure where only the content words matter. Transfer on the deep layer follows where each node's lemma is translated. The last step transforms the deep structure of the target sentence into an a-tree [7]. According to the results from Gaudio et al. (2016), TectoMT outperforms Moses in terms of BLEU on several language pairs.

Moses baseline systems for several language pairs were trained by Xu et al. (2017). A combination of three data selection methods is applied by Wolk and Marasek (2017), namely the perplexity approach introduced by Axelrod et al. (2011), the edit distance and the cosine TF-IDF. The three pseudo in-domain corpora are joined and used to train SMT systems for various language pairs.

Automatic evaluation results, as well as a manual validation of the systems[8] can be found in the WMT 2017 Biomedical task findings (Jimeno Yepes et al., 2017). My submissions were in the constrained setting (only used in-domain data made available by the organizers).

Grozea (2018) trains various models that differ on the one hand in the training corpora, and on the other hand in the type of word segmentation[9]. The best translation is selected automatically from the set of candidates produced by the different models by means of heuristics. The paper focuses on one language pair, English to Romanian, and uses the Tensor2Tensor (Vaswani et al., 2018) implementation for training NMT systems. The Tensor2Tensor implementation is also adopted by Huck et al. (2018) who demonstrate that compared to their submission from the previous year, (Huck et al., 2017), which used Nematus (Sennrich et al., 2017), the Tensor2Tensor approach gives better BLEU results.

Domain adaptation through transfer learning is adopted by Khan et al. (2018) where different Biomedical in-domains are used to train a series of models. The parameters of the previous training model are used in the initialization step for the next one. Using this technique BLEU scores significantly increased over the baseline. Soares and Becker (2018) used the OpenNMT toolkit (Klein et al., 2017) and the Moses toolkit. Their baseline systems trained on four language pairs put SMT and NMT on a par in terms of BLEU.

[7] Available at `https://ufal.mff.cuni.cz/tectomt/translation-example`. For more details refer to Popel et al. (2016)

[8] Including my submission.

[9] See Chapter 3, Section 3.2 for details on word segmentation for NMT.

4.3 Summary

A brief introduction to domain adaptation techniques for SMT was given in this chapter. The main focus was on one particular domain adaptation method, namely data selection. The terminology defined by the MT community for data selection was presented since I adopted it for the thesis too.

The problem of data selection was formally defined, followed by approaches to data selection for SMT. Methods based on information retrieval were presented as they pioneered data selection. Additionally, language model perplexity methods were also described, with focus on one of the most frequently used comparison method which is based on the difference between the bilingual cross-entropies. This approach constituted the method I have chosen to compare my data selection methods with. Finally, related work to domain adaptation was given.

Chapter 5

Scoring Functions for Data Selection

This chapter describes the data selection methods I developed and their application on several language pairs. The core of the methods are scoring functions that are used for filtering the general domain data. A data selection method based on Paragraph Vector (Le and Mikolov, 2014) is presented, followed by a data selection technique that uses term frequency. The methods developed using these text representations were published in Duma and Menzel (2016a) and Duma and Menzel (2018).

Whether a general domain sentence should be selected as being relevant to an in-domain or not depends highly on the underlying representation of text that can be used for assessing similarities between sentences. Therefore, one crucial step in developing a data selection method is to transform the sentences into numerical representations that become the input for the algorithm.

Considering the research questions defined in Chapter 1, this chapter addresses the first two ones. Different approaches for sentence representation are used in the data selection pipeline (**RQ1**). As a result, scoring functions are developed with the purpose of selecting pseudo in-domain sentences (**RQ2**). With the emergence of neural approaches to MT, a follow-up research question arises: can the developed data selection methods be applied on NMT?

This chapter firstly introduces the data selection method with text representation based on Term Frequency (TF). Another method which uses Paragraph Vector (PV) for representing text is afterwards presented. The data and resources used are described, together with an analysis of the experimental results, for both methods.

5.1 Term Frequency Based Method

This section introduces a data selection method based on Term Frequency. In my work, a document profile is a bag-of-words model where a document (text) is represented by its vocabulary and the frequency of its respective words. I developed an algorithm that is based on differences between a general domain profile and an in-domain profile. This model is limited because it ignores lexical semantics. However, the model is simple, no models need to be trained, the method is unsupervised, fast to apply and is language- and domain-independent. Due to the fact that in my setting there is only one training document, I did not use TF-IDF, but simply term frequency. I also did not use the cosine to compute text similarity, but developed a new sentence scoring method.

5.1.1 Algorithm

The scoring algorithm builds on top of the simple TF representation and introduces a scoring formula with a new weighting scheme. This method is referred to as *DSTF (Data Selection via Term Frequency)*.

In the initial phase, a profile consisting of word counts is built for each domain, either for the source language or for the target language side of the domains. In order to build the profile for a corpus, all of its sentences are preprocessed using tokenization, lowercasing, removal of stop words and lemmatization or stemming in the case a lemmatizer is not available for a language. Numbers or punctuation marks are ignored and only words contribute to the scoring (procedure *Preprocess_Corpus* in Algorithm 2)[1].

For every sentence from the preprocessed general domain data, the algorithm iterates through all of the words from a sentence. Given a sentence s and the word w, the relative difference[3] between the frequency of w in the in-domain profile, $TF(w, \mathcal{I}N_{side})$, and the frequency of w in the general domain profile, $TF(w, \mathcal{G}EN_{side})$, is squared. By squaring, higher differences have more impact than the lower ones. The same relative difference formula as in Keselj et al. (2003)[4] is used where the difference is divided by the arithmetic mean of the frequencies. Each relative difference is multiplied with an empirically determined weight that represents the impact that w made in the two profiles[5]. While Keselj uses the relative difference of frequencies of character n-grams, I use the relative difference of a word frequency in two domains and introduce a weighting for the word to account for its impact on the overall sentence score.

5.1.2 Data and Resources

This section presents the resources used when applying DSTF on the Biomedical domain on English→Spanish and English→Portuguese.

[1] For word count I used the script *ngram-count* from SRILM[2] (Stolcke, 2002)

[3] Given the numbers a and b, the relative difference is defined as the difference between a and b divided by a function of a and b, usually the maximum, minimum or average of the two numbers.

[4] Keselj uses character n-grams and profiles built using the most frequent character n-grams for authorship attribution

[5] Division by zero is never reached for neither calculating the weight, nor the score for a word because the algorithm iterates through all words from a sentence from the general domain, thus the word exists and has at least a frequency of one.

```
Algorithm 2 DSTF Filtering (Duma and Menzel, 2018)
  procedure PREPROCESS_CORPUS($\mathcal{C}$)
      $tokenize(\mathcal{C})$
      $lowercase(\mathcal{C})$
      $removeStopWords(\mathcal{C})$
      $lemmatize(\mathcal{C})$                                  ▷ or stem if unavailable
      $keepWords(\mathcal{C})$
      $wordCount(\mathcal{C})$
  procedure FILTER($\mathcal{G}EN_{side}, \mathcal{I}N_{side}$)        ▷ side refers to either source or target
      Preprocess_Corpus($\mathcal{G}EN_{side}$)
      Preprocess_Corpus($\mathcal{I}N_{side}$)
      for each sentence $s \in \mathcal{G}EN_{side}$ do
          for each word $w \in s$ do
              $weight = TF(w, \mathcal{I}N_{side}) / TF(w, \mathcal{G}EN_{side})$
              $score_w = \left(\frac{2 \cdot (TF(w,\mathcal{I}N_{side}) - TF(w,\mathcal{G}EN_{side}))}{TF(w,\mathcal{I}N_{side}) + TF(w,\mathcal{G}EN_{side})}\right)^2 \cdot weight$
          $score_s \mathrel{+}= score_w$   ▷ all intermediate scores contribute to the final score
```

For the general domain training data, the Commoncrawl[6] corpora and the Wikipedia (Wolk and Marasek, 2014) one were concatenated for English→Spanish, and Paracrawl[7] and Wikipedia for English→Portuguese. For the in-domain, I used the EMEA (Tiedemann, 2012) and the Scielo corpora (health and biological)(Neves et al., 2016) for both test language pairs.

Different sets belonging to multiple corpora were used for tuning the systems, depending on the availability of a tuning set for a certain language pair. In the case of English→Spanish, I aimed at diversity in the medical data, thus I concatenated two medical development sets: the Khreshmoi development set from the Medical Task of WMT 2014[8] consisting of 500 sentence pairs and the ECDC corpus made available from UFAL[9]. Using the full size of the ECDC corpus (2357 sentence pairs for English↔Spanish) would have made the tuning of the SMT systems very time and memory intensive. Therefore, I limited the size of the sentences to a minimum of 20 words and a maximum of 80 words which downsized the ECDC corpus to 850 sentences. The resulting development set for English↔Spanish consisted of 1350 sentences. Tuning of the systems for English→Portuguese used a sample of 1000 sentences from the Scielo development set[10].

Statistics that include the number of sentences, the number of tokens and vocabulary size after text preprocessing is given in the tables below for every corpus used in the training of the SMT systems. The number of documents is reported

[6] http://commoncrawl.org/

[7] https://paracrawl.eu/index.html

[8] http://www.statmt.org/wmt14/medical-task/

[9] http://ufal.mff.cuni.cz/ufal_medical_corpus

[10] http://www.statmt.org/wmt16/biomedical-translation-task.html

for the test sets.[11]

Corpora/ Dataset	Sent. / Docs	Tokens		Vocabulary	
		English	Spanish	English	Spanish
Commoncrawl	1.8M	46.5M	47.8M	459K	566K
Wikipedia	1.6M	42.9M	42.1M	634K	719K
EMEA	678K	13.0M	14.2M	71K	86K
Scielo-gma	166K	4.7M	5.1M	102K	118K
Development set	1350	36K	41K	5221	6239
Test set	100	8033	7795	2164	2241

Table 5.1: Corpora statistics for English→Spanish after preprocessing

Corpora/ Dataset	Sent. / Docs	Tokens		Vocabulary	
		English	Portuguese	English	Portuguese
Paracrawl	2.1M	58.9M	59.0M	286K	381K
Wikipedia	1.6M	44.1M	42.3M	588K	667K
EMEA	1.08M	14.7M	15.8M	103K	117K
Scielo-gma	613K	17.1M	17.5M	114K	136K
Development set	1000	40K	42K	5495	6349
Test set	92	8274	8357	2200	2461

Table 5.2: Corpora statistics for English→Portuguese after preprocessing

While the general domain corpus[12] is more than four times bigger than the in-domain corpus for English→Spanish, for the other language pair it is two times bigger than the in-domain corpus, even though the size of the two general domain corpora for the two language pairs is similar. The in-domain corpus size for English→Portuguese is two times bigger than the English→Spanish corpus. The vocabulary size per corpus is comparable between English, Spanish and Portuguese, with the latter two exhibiting slightly larger numbers than the English corpus.

The *nltk* toolkit(Bird et al., 2009) was used for text processing, as well as the WordNet (Fellbaum, 1998) lemmatizer for English and the Snowball stemmer (F. Porter, 2001) for Spanish and Portuguese.

5.1.3 Experimental Results on the Biomedical domain

Since DSTF can be applied on both source and target languages, this section presents three of its variants: the first one only considers the scores obtained using the English side of the training corpora (DSTF-src), the second variant made use of only the non-English side of the training corpora (DSTF-trg), and the third one

[11] In this thesis, tables that present corpora statistics or parameters are depicted with a basic formatting, while tables that include experimental results present a blue header. This formatting distinction was made for better visualization.

[12] Referring only to concatenated corpora for both general and in-domain.

sums up the scores obtained using DSTF applied for the source language and the scores for the target language (DSTF-bi).

Considering that one of the aims of data selection is to use small selections of pseudo in-domain sentences for training MT systems, this section presents experimental results for a selection of 10%.

Table 5.3 presents the number of sentence pairs that were selected, as well as the total number of sentence pairs that were used in the training of the SMT systems (concatenation of in-domain and pseudo in-domain corpora).

Language pair	**English→Spanish**	**English→Portuguese**
10% of Gen	350K	378K
total training data	1.62M	2.07M

Table 5.3: Number of selected sentences from the General domain data for the two language pairs

Table 5.4 presents the BLEU scores for the two language pairs. DSTF-src performs on a par with DSTF-bi for English→Spanish, with small BLEU differences to DSTF-trg (not statistical significant). On the other language pair, DSTF-src significantly outperforms DSTF-trg. The English→Spanish BLEU scores are smaller than the English→Portuguese scores with more than three BLEU points. This result was expected since the in-domain corpus for English→Portuguese is two times bigger than the English→Spanish corpus. The number of unknown words was similar among the variants, this together with the small BLEU differences indicates that any DSTF variant can be used.

Language Pair	**English→Spanish**		**English→Portuguese**	
Evaluation	BLEU	OOV	BLEU	OOV
DSTF-src	31.32	3.8	34.92	1.4
DSTF-trg	31.05	3.9	34.19	1.5
DSTF-bi	31.33	3.8	34.49	1.4

Table 5.4: BLEU scores together with OOV rates

5.2 Paragraph Vector Based Method

As emphasized in Chapter 2 which describes Paragraph Vector (PV), sentence representation using PV has gained popularity in fields like sentiment analysis, and semantic textual similarity. For example, Hu and Song (2016) experimented with sentiment analysis using microblogs where the sentences were very short. Duma and Menzel (2017b) assessed the semantic textual similarity between monolingual and cross-lingual sentence pairs.

Data selection is based on the notion of similarity. PV has been successfully applied in other similarity-based approaches. Therefore, it is interesting to investigate the application of PV to data selection.

This section presents the algorithm and scoring functions that I developed for my PV-based data selection methods.

5.2.1 Algorithm

The algorithm consists of three parts. It receives as input the in-domain corpora ($\mathcal{I}n$), the general domain corpora ($\mathcal{G}en$), the number of most similar sentences to a given one ($\mathcal{N}$) and the amount of pseudo in-domain sentences to select ($\mathcal{P}$). Given a general domain sentence s, the top $\mathcal{N}$ most similar sentences to it are retrieved based on the cosine similarity between s and the retrieved sentences.

Initially, all the available parallel corpora are concatenated and labeled using the corpus tag together with an unique numeric index. The algorithm can be applied on either the source or the target language (in Algorithm 3 the source side is used) [13].

Algorithm 3 Doc2vec Filtering

```
procedure DS_D2V($\mathcal{I}n, \mathcal{G}en, \mathcal{N}, \mathcal{P}$)
    $\mathcal{C} \leftarrow \mathcal{G}en + \mathcal{I}n$
    for each sentence $s_i \in \mathcal{C}_{source}$ do
        tag $s_i$ with its domain and the line number $i$
    train doc2vec model $\mathcal{M}$ using tagged $\mathcal{C}_{source}$
    for each sentence pair $(s_i, t_i) \in \mathcal{G}en$ do
        $\mathcal{R}_i = top(\mathcal{N}, most\_similar(\mathcal{M}, s_i))$
        $Sim_{s_i} = \{(tag, score) \in \mathcal{R}_i | \ tag \in \mathcal{C}_{tags}, score \in (0,1)\}$
        $sentence\_scores_i = \mathcal{S}core(\mathcal{N}, (s_i, t_i), Sim_{s_i})$
    $sort \uparrow sentence\_scores$
    add top $\mathcal{P}$ sentences to $FilteredCorpus_{\mathcal{P}}$
```

After the sentence tagging step, the doc2vec model is trained[14]. The Gensim doc2vec built-in function $most\_similar$ was used to obtain the top-$\mathcal{N}$ most similar docvecs (corresponding to sentences from the training data) for each sentence pair from the general domain. It uses the cosine similarity between the projection weight vectors of the given sentence and all the other sentences from the training data (general domain concatenated with in-domain). In Algorithm 3, $\mathcal{R}_i$ represents the set of most similar sentences for the given sentence s_i and it is retrieved as a set of $(tag, score)$ pairs where the similarity score ranges between 0 and 1. For every sentence s_i belonging to the general domain, its selection score is calculated by applying a scoring function based on its sentence similarity set $\mathcal{R}_i$.

[13] Training a PV model for a particular language can make use of arbitrary bilingual corpora as long as one of the languages involved is the desired one.

[14] Doc2vec hyperparameters are presented in Appendix A

$\mathcal{N}$ influences the number of similar sentences that will be considered in calculating the final score of a sentence. A too small value assigned to $\mathcal{N}$ could result in missing out pseudo in-domain sentences that are medium or marginally close to the in-domain. However, the only disadvantage to assigning a too high value to $\mathcal{N}$ could result in a more expensive computation of final sentence scores.

After all the general domain sentences get a selection score, they are sorted in descending order reflecting their proximity to the in-domain. The top $\mathcal{P}$ sentences constitute the pseudo in-domain, which will later be used in the training of MT systems together with the in-domain corpus.

```
function SCORE($\mathcal{N}, (s, t), Sim_s$)
    $sentence\_score = 0$
    for $(tag_j, sim\_score_j) \in Sim_s$ do
        if $s_{tag_j} \in \mathcal{I}n$ then
            $sim\_score_j = sim\_score_j * (\mathcal{N} - j + 1)^2$
        else
            $sim\_score_j = 0$
        $sentence\_score = sentence\_score + sim\_score_j$
    return $sentence\_score$
```

Figure 5.1: Scoring function *SEF*

The scoring function 5.1, SEF (Sentence Embedding Filtering), assigns a continuously valued score to each sentence from the selection pool. Given a sentence s from the general domain data, the sentence score is determined by the sum of all similarity scores between s and sentences that belong to the in-domain multiplied with a weight. The weighting factor was determined using the observation that given a similarity score between s and a sentence belonging to the in-domain, $s_{in-domain}$, the rank of $s_{in-domain}$ in Sim_s (which represents the top $\mathcal{N}$ most similar sentences to s), correlates well with how closely related s is to the in-domain.[15].

Additional PV-based scoring functions that I implemented are SEF* (presented in Duma and Menzel (2016a)) and SEFp (described in Duma and Menzel (2016b)). Both of them are based on Algorithm 3. The key differences between SEF and SEFp are the weighting attributed to the cosine similarity and the sentence weight. SEF*, on the other hand, considers a general domain sentence to be part of the pseudo in-domain, if it contains at least one in-domain sentence in its Top $\mathcal{N}$ list. Preliminary experiments indicated that SEF outperforms the other two scoring functions, thus it is used as the PV-based scoring method for the experiments and for the evaluation presented in this thesis.

[15] $\mathcal{N} - j + 1$ was used instead of j as higher weights need to be attributed to lower ranks of in-domain sentences in Sim_s

5.2.2 Data and Resources

The experimental setup and results are described for SEF being applied to the IT domain. Training of the SMT systems was carried on for the English→German language pair, using the Commoncrawl corpus[16], constituting the general domain data. For the in-domain, the IT corpus provided through WMT 2016 (Bojar et al., 2016a) was used. The systems have been tuned using a concatenation of two sets (Batch1a and Batch2a) from the QTLeap corpus[17], resulting in 2000 sentence pairs. The test set consists of 1000 sentences also from the QTLeap corpus. Statistics for the training data used are indicated in Table 5.5.

Corpora/ Dataset	**Sentences**	**Tokens**		**Vocabulary**	
		English	German	English	German
Commoncrawl	2.34M	59.13M	55.16M	709K	1.54M
IT corpora	210K	2.27M	2.26M	104K	125K
Development set	2000	53K	55K	3493	4820
Test set	1000	23K	24K	2334	2926

Table 5.5: Corpora statistics after preprocessing

While the in-domain corpus size is rather small (only 210K sentence pairs), the general domain corpus is more than eleven times bigger than the in-domain corpus. The vocabulary size for German is twice as big as the one for English. This is not surprising given the fact that German is highly inflectional and presents compound words that are merged together in single words (as opposed to English).

Data preprocessing included tokenization, cleaning (restriction to a maximum sentence length of 80 words), lowercasing and removal of sentence pairs that did not belong to the English-German language pair[18].

5.2.3 Application to Neural Machine Translation

Since data selection is a domain adaptation technique that focuses on creating domain-specific training data, all of the data selection methods presented in this thesis can be applied on both SMT and NMT frameworks. In this section I explore the SEF method applied to the IT domain, for the English→German language pair.

According to Britz et al. (2017a), one of the main drawbacks of NMT is that training is highly time-expensive, requiring even weeks of GPU time to converge. This creates a strong motivation for employing data selection for NMT. My main goal is to validate the applicability of one of my data selection methods to the neural MT architecture. An exhaustive hyperparameter search for obtaining the best BLEU score has not been performed. Also, since not all the hyperparameters were tuned, this section does not engage into comparing SMT and NMT.

[16] `http://commoncrawl.org/`

[17] Available at `http://metashare.metanet4u.eu/go2/`

[18] Using the jlangdetect library: `https://github.com/melix/jlangdetect`

The Tensorflow NMT (Luong et al., 2017) implementation was used in conducting the experiments. It builds sequence-to-sequence models for MT briefly described in Chapter 3, Section 3.2. The same preprocessed data as in the SMT experiments was used. Byte Pair Encoding (BPE) (Sennrich et al., 2016b) using 32,000 merge operations was employed for learning shared subword units[19]. I used the attention mechanism as described in Bahdanau et al. (2014).

After training, the best BLEU score obtained was 33.6[20]. However, as in the case of SMT systems, the output sometimes differs greatly from the reference due to a different word choice for the same term (for example, in the case synonyms are available). This leads to a low BLEU score since this metric is only a syntactic one. For example, given the input sentence *Check the volume control on the taskbar or in the control panel.* and the reference *Überprüfen Sie die Lautstärkeregelung in der Taskleiste oder in der Systemsteuerung.*, the translation obtained with my NMT system is *Kontrollieren Sie die Lautstärkesteuerung auf der Taskleiste oder im Kontrollpanel.*. Even though this translation is not perfect, it is understandable and the pairs of words *(überprufen, kontrollieren)*, *(lautstärkeregelung, lautstärkesteuerung)* and *(systemsteuerung, kontrollpanel)* can be used interchangeably in this sentence.

5.3 Summary

This chapter presented the data selection methods that I developed which require a ratio for determining the amount of sentences to select from the general domain as being pseudo in-domain. Two research questions, RQ1 and RQ2, were tackled by transforming sentences into vector representations that can be used to compute similarity scores. The sentence vectors contributed to the formulation of data selection methods through the scoring functions that were based on vector similarities.

Two types of sentence representations were presented, namely Term Frequency (TF) and Paragraph Vector (PV). A scoring function using TF resulted in a data selection method named DSTF that was applied to two language pairs for the Biomedical domain. This scoring function used the relative difference between term frequencies in two domain corpora and a weighting based on the ratio of frequencies of a word in two domain corpora.

A scoring function which made use of PV for the initial phase of text representation was applied to one language pair in the IT domain (SEF). This scoring function used the rank of a sentence in a list of most similar sentences. While DSTF was applied to SMT, SEF was also validated with an NMT system to demonstrate that my data selection methods can be applied to both MT paradigms.

[19] I used the implementation available at `https://github.com/rsennrich/subword-nmt`

[20] The training took 43 hours on a machine with one GPU GeForce 940MX.

Chapter 6

Automatic Ratio Detection for Data Selection

The previous chapter presented a suite of methods that focus on scoring the general domain sentences with respect to their similarity to a given domain. Beyond scoring, data selection includes also the task to identify the ratio of general domain sentences to keep for training an MT system. The methods described so far did not deal with this problem. Instead, the same procedure commonly used in the research community was applied: empirically determining the optimal number of sentences to be kept based on the BLEU score of the trained system for different predetermined ratios of domain specific sentences.

There are several problems regarding this approach: there is no standard agreement in the community for the parameter setting (the minimal number and the increment size for selecting sentences). This leads to different schemes of reporting the empirical results.

For example, Axelrod et al. (2011) uses the top $N = 35K, 70K, 150K$ sentence pairs from the scored general domain pool, while Biçici and Yuret (2011) increasingly select $N \in 100, 200, 500, 1000, 2000, 3000, 5000, 10000$ instances for each test set sentences. In contrast to these absolute numbers, Kirchhoff and Bilmes (2014) selects 10% of the data till a maximum value of 40% and van der Wees et al. (2017) apply data selection on SMT and NMT selecting the top $5\%, 10\%, 20\%$ and 50% for the general domain corpus. Also for NMT, Silva et al. (2018) use different data selection sizes corresponding to a factor of $1, 2, 4$ and 8 in relation to a preprocessed general domain corpus, while Poncelas et al. (2019) evaluates NMT systems trained on pseudo in-domains of size $100K$ and $200K$.

This chapter addresses the research question **RQ3** defined in Chapter 1, which concerns the automation of the ratio detection. Such a method (Duma and Menzel, 2017a), iATD, is introduced with preliminary evaluation, followed by an improvement to it by means of a hybrid approach that makes use of the scoring step from a threshold tuning method and the automatic ratio detection one. This hybrid method, namely hATD, is contrasted with the automatic ratio detection, iATD, approach and the evaluation results are discussed.

The preliminary evaluation is followed by an extended one, where hATD together with the methods presented in the previous chapter, the state-of-the-art method, and several baselines are evaluated on common experimental settings. This was essential, since the methods have been previously applied on different in-domains and language pairs, thus MT outputs and results were not comparable across mixed training corpora and test sets.

Two research questions will be answered in the extended evaluation: **RQ4** and **RQ5**. The objective of RQ4 is to investigate whether a gain in translation performance comes from adding more training data, or from adding more pseudo in-domain sentences. Since all data selection methods will be applied in several experimental settings, RQ5 aims at finding out whether the systems ranking is consistent across them.

The chapter is structured as follows: firstly, iATD is introduced with its preliminary evaluation, followed by an improvement to it, namely hATD. While the preliminary evaluation contrasts iATD and hATD and aims at answering RQ3, the extended evaluation brings together the incremental ratio methods based on scoring functions and hATD under a common evaluation framework. The extended evaluation is more detailed than the preliminary one since all data selection methods are contrasted against each other on the same in-domains and the same language pairs, aiming at answering the RQ4 and RQ5.

6.1 Algorithm

This section describes the mechanism of the automatic threshold detection methods, thus answers **RQ3**. Firstly, iATD is detailed, followed by an improvement to it via a hybrid approach, hATD, that combines iATD with the previously developed data selection methods (presented in Chapter 5).

Considering the data selection problem as a classification problem was an important step towards solving the automatic ratio detection of sentences. I used a MultiLayer Perceptron (MLP) classifier, also known as a Feed-forward Neural Network classifier, to obtain a model that is able to make a binary decision: to keep or to discard a sentence. A diagram representing a vanilla MLP is depicted in Figure 6.1[1] where the input features are a set of $X = (x_1, x_2, ..., x_n)$ neurons that together with a bias are propagated through the network.

In order to avoid overfitting, I used the dropout technique which randomly drops neurons (units) during the training of the neural network. A dropout value of 0.5 was used, which was selected in accordance with the findings from Srivastava et al. (2014).

The input to the network consists of Paragraph Vectors with the positive samples being randomly selected from the in-domain pool, while the negative samples are randomly selected from the general domain pool. The assumption is that the general domain pool is large enough, therefore the probability of selecting false

[1] Figure taken from `https://scikit-learn.org/stable/modules/neural_networks_supervised.html`

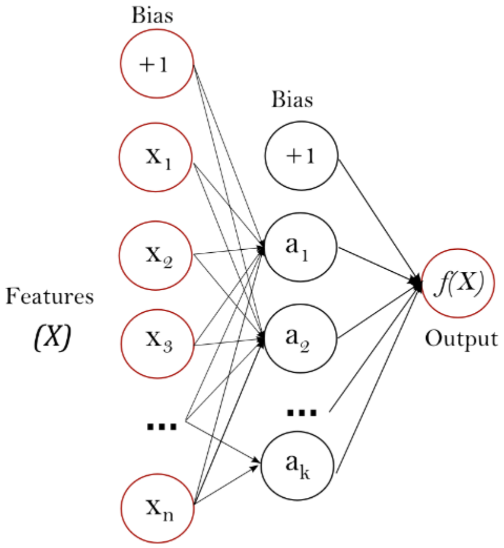


Figure 6.1: Vanilla MLP (Pedregosa et al., 2011)

negatives (which actually are pseudo[2] in-domain sentences) is small.

An equal number of positive and negatives samples was randomly selected and a Doc2Vec model was trained on all available data (using tags to identify the domain for each sentence). I used the scikit-learn (Pedregosa et al., 2011) and scikit-neuralnetwork[3] libraries for training the MLP classifier.

The presented method, *iATD* comes with a drawback: the negative samples are randomly drawn from the general domain data. This assumes that the probability to select biomedical sentences from it (false negatives) is small.

Actually, this assumption does not hold. The following sample inspection shows that the rate of false negatives among randomly sampled sentences is high indeed. Given a general domain consisting of 3.5 million sentences for Spanish→English, the ratio of sentences that belong to the Biomedical domain is unknown[4]. I inspected the randomly selected sentences by automatically searching for all the terms (multi-words) from a biomedical terminology set[5].

A terminology set was preferred instead of a dictionary which contains words that could be polysemous. The risk is, however, that the search terms are either too complex or too specific and biomedical sentences are missed out because they don't contain them. Examples of terms from the terminology set are, for instance, adrenal glands, administered subcutaneously, x-ray of right wrist and whiteness. As it can be noted, some of the terms are very specific (x-ray of right wrist) and rare to find in corpora. On the other hand, too general terms (whiteness) might cause non-biomedical sentences to be retrieved. The search resulted in a ratio of 4.3%

[2] As previously noted, sentences selected from the general domain are called pseudo in-domain ones in this thesis to differentiate them from the in-domain sentences.

[3] Available at: `https://github.com/aigamedev/scikit-neuralnetwork`

[4] Detailed information on the corpora and data selection task is given in Section 6.2.1.

[5] The terminology set was available through the WMT 2019 Biomedical task.

Algorithm 4 Hybrid Automatic Threshold Detection

```
procedure hATD($\mathcal{I}n, \mathcal{G}en$, MLPClassifier)
    $PosSamples \leftarrow$ RandomSelection($\mathcal{I}n, sampleSize$)
    score $\mathcal{G}en$ and sort $\downarrow$          ▷ using any data selection scoring functions
    $NegSamples \leftarrow$ Top($scored\mathcal{G}en, sampleSize$)
    split $PosSamples$ and $NegSamples$ into train and test data
    sources $\leftarrow \{PosTrainSamples, NegTrainSamples,$
                $PosTestSamples, NegTestSamples, GenTrain\}$
    for each sentence $s_i \in$sources do
        tag $s_i$ with its domain and the line number $i$
    train doc2vec model $\mathcal{M}$ using tagged sources
    best_estimator $\leftarrow$ GridSearch(MLPClassifier)
    predict labels for $GenTrain$ with best_estimator
    $PseudoInDomain \leftarrow$ positive_labeled($GenTrain$)
```

medical sentences (duplicate sentences were removed as multiple search terms can retrieve the same sentence).

The problem of selecting false negatives is tackled by identifying the most dissimilar sentences to the in-domain and treating them as negative samples. I have chosen the *DSTF* method, presented in the previous chapter, to score the general domain sentences. Instead of using the top most similar sentences, the sentences with the lowest scores were selected. I applied again the search for biomedical terminology using the same data set and observed a reduction of the false negatives rate from 4.3% to 0.02%.

The positive samples were selected with the previously described iATD method and the same classifier was applied. The term for the improved negative sampling method is *hATD*, since it is a hybrid between a data selection scoring function and an automated data selection method.

The procedure for automatically detecting the data selection ratio is given in Algorithm 4. The positive instances for training the Doc2Vec model are randomly sampled from the in-domain data, while any data selection scoring function can be applied to score the general domain sentences and sampling the ones with the lowest scores as negative instances. The instances are split into training and testing samples with a test set size that amounts to 5% of the training set size. After the Doc2Vec model is trained, grid search is used to determine the best estimator and finally, this is used to predict for each sentence from the general domain whether it belongs to the in-domain or not.

6.2 Preliminary Evaluation

Firstly, results using iATD obtained for the Biomedical domain for several language pairs are presented. Afterwards, the improvement to the method is explored through the hybrid approach, hATD, revealing significant improvements over iATD.

6.2.1 Data and Resources

This section presents the data, resources and data preparation used to apply iATD to the Biomedical domain.

The SMT systems were developed using the Moses toolkit. The preprocessing of the data consisted in tokenization, cleaning with a cutoff of 6-80, lowercasing and normalizing punctuation. For the tuning of the systems the data provided by the WMT 2016 Biomedical task[6] (Bojar et al., 2016a) was used.

As general domain data, Commoncrawl[7] and Wikipedia (Wolk and Marasek, 2014) were exploited for all language pairs except for English-Portuguese where no Commoncrawl data was provided by WMT. As in-domain corpora, EMEA (Tiedemann, 2012) was used for all language pairs. In addition to that, Pubmed[8] and other medical corpora from the UFAL Medical Corpus[9] (ECDC, Muchmore, PatTR Medical) were used depending on their availability for each language pair. The Scielo corpus provided by the previous Biomedical task from 2016 was also used. The table below reports statistics for every corpus used to train the SMT systems, including the number of sentences, the number of tokens and the vocabulary size after preprocessing.

For some language pairs, several test sets were available (Jimeno Yepes et al., 2017) covering a variety of topics: Scielo (ecosystem studies, descriptions of clinical cases), Cochrane (medicine descriptions, experimental studies), EDP (findings of health articles) and NHS (health recommendations, medical advice on addictions).

Corpora/ Dataset	**Sent. / Docs**	**Tokens**		**Vocabulary**	
		English	French	English	French
Commoncrawl	3.1M	81M	781K	97.9M	886K
Wikipedia	770K	20.7M	19.8M	417K	422K
EMEA	672K	12.9M	16.6M	70K	81K
ECDC	2043	46K	62K	5229	6040
Scielo-gma 2016	17K	489K	680K	17K	22K
Development set	1516	26K	36K	4410	5090
Test set EN→FR EDP	750	17K	20K	3691	4151
Test set EN→FR Cochrane	467	10K	13K	1762	2093
Test set EN→FR NHS	1044	15K	20K	2509	3105
Test set FR→EN EDP	699	16K	18K	3706	3862

Table 6.1: Corpora statistics for English↔French after preprocessing

[6] Available at: `https://www.statmt.org/wmt16/biomedical-translation-task.html`

[7] Available at: `http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz`

[8] Available through WMT 2016 at: `https://www.ncbi.nlm.nih.gov/pubmed`

[9] Available at: `http://ufal.mff.cuni.cz/ufal_medical_corpus`

Corpora/ Dataset	**Sent. / Docs**	**Tokens**		**Vocabulary**	
		English	Portuguese	English	Portuguese
Wikipedia	1.6M	44.1M	42.3M	588K	667K
EMEA	1.08M	14.7M	15.8M	103K	117K
Scielo-gma 2016	613K	17.1M	17.5M	114K	136K
Pubmed	67K	1.2M	965K	36K	54K
Development set	7000	203K	212K	14K	16K
Test set EN→PT Scielo	1806	48K	50K	5997	7200
Test set PT→EN Scielo	1897	50K	51K	6015	7139

Table 6.2: Corpora statistics for English↔Portuguese after preprocessing

Corpora/ Dataset	**Sent. / Docs**	**Tokens**		**Vocabulary**	
		English	German	English	German
Commoncrawl	2.34M	59.13M	55.16M	709K	1.54M
Wikipedia	2.2M	54.8M	47.4M	803K	1.1M
EMEA	646K	12.3M	12.2M	70K	112K
ECDC	1931	43K	44K	5030	7107
Muchmore	28K	717K	614K	32K	78K
PatTR Medical	1.4M	46.2M	43.1M	223K	714K
Development set	1960	35K	35K	4006	5195
Test set EN→DE Cochrane	467	10K	10K	1762	2349
Test set EN→DE NHS	1044	15K	15K	2509	3224

Table 6.3: Corpora statistics for English↔German after preprocessing

6.2.2 Experimental Results

iATD was validated through the 2017 WMT evaluation campaign on seven language pairs (English→German and both directions for English→Spanish, English-Portuguese and English-French) and several test sets.

The classifiers were trained on 200K sentences with an equal number of positive and negative samples. The default parameters values for training the Doc2Vec models were used. The MLP classifier was trained with the `tanh` activation function and the `momentum` learning rule.

Since the algorithm can be applied on either the source or the target language, I exploited both directions for each language pair and each test set, as follows:

- iATD-src: SMT system trained on the selected sentences obtained using the classifier trained on the source language data
- iATD-trg: SMT system trained on the selected sentences obtained using the classifier trained on the target language data

Corpora/ Dataset	**Sent. / Docs**	**Tokens**		**Vocabulary**	
		English	Spanish	English	Spanish
Commoncrawl	1.8M	46.5M	47.8M	459K	566K
Wikipedia	1.6M	42.9M	42.1M	634K	719K
EMEA	678K	13.0M	14.2M	71K	86K
ECDC	1769	37K	43K	4654	5566
Scielo-gma 2016	166K	4.7M	5.1M	102K	118K
Pubmed	250K	3.2M	3.3M	75K	108K
Development set	3933	119K	128K	12K	14K
Test set EN→ES Scielo	1082	31K	33K	4612	6076
Test set EN→ES Cochrane	467	10K	11K	1762	2050
Test set EN→ES NHS	1044	15K	16K	2509	2950
Test set ES→EN Scielo	1180	31K	34K	4602	5527

Table 6.4: Corpora statistics for English↔Spanish after preprocessing

- iATD-bi: SMT system trained on the union of the selected sentences proposed by the two classifiers

Table 6.5 presents all the BLEU scores for my submissions as reported by the WMT competition. In the following, I will discuss the results for each test separately. The Scielo dataset consisted of titles and abstracts from scientific publications retrieved from the Scielo database[10] . My participation was the only one from all teams that submitted system runs for this dataset. For the English→Portuguese and the English→Spanish language pairs, all my submissions improved over the baseline provided by the organizers with almost 10 BLEU points. For the other directions, Portuguese→English and Spanish→English, my experiments achieved a significant improvement over the baseline with almost 7 BLEU points.

Language pair	**English→German**		**English→Spanish**			**Spanish→English**	**English→French**			**French→English**	**English→Portuguese**	**Portuguese→English**
Test set	Cochrane	NHS	Scielo	Cochrane	NHS	Scielo	EDP	Cochrane	NHS	EDP	Scielo	Scielo
iATD-src	22.03	18.71	36.08	**48.99**	40.97	37.14	22.43	32.46	31.79	22.64	39.14	43.84
iATD-trg	22.37	**19.80**	35.93	48.45	41.20	37.47	22.25	32.59	31.89	22.37	**39.38**	**43.93**
iATD-bi	**22.63**	19.66	**36.23**	48.70	**41.22**	**37.49**	**22.79**	**33.16**	**33.36**	**23.41**	39.21	43.88

Table 6.5: BLEU results for all language pairs as reported in Jimeno Yepes et al. (2017)

The EDP dataset was made up of a collection of titles and abstracts from five journals from the *Health*, *Life* and *Environmental Sciences* fields (Jimeno Yepes et al., 2017) with a reported misalignment rate of 6% on a corpus sample, with 20% of the sentence pairs containing additional content in one of the languages. It was made available only for the language pairs English→French and French→English. I obtained again significant improvements over the baseline, with an increase in 6 BLEU points for French→English and 10 BLEU points for English→French.

[10] `http://www.scielo.org/`

Among the three variants, the one which was trained on the union of the selected sentences gave the best results.

The Cochrane and NHS test sets consist of health related documents obtained during the KConnect[11] and HimL[12] projects. There were no baselines results provided by the organizers for these test sets. I obtained very high BLEU scores for English→Spanish (almost BLEU of 49 for Cochrane and BLEU of 41 for NHS).

For English→French I obtained BLEU scores revolving around 33 and for English→German close to 23 for Cochrane and 20 for NHS. The differences in BLEU scores among language pairs vary with the amount of training data used, as well as with the size of the in-domain corpora. For English→German, there was no Scielo corpus available. Very high BLEU scores have been reached on Portuguese→English and Spanish→English (and vice-versa) due to a relatively large amount of Scielo data available (compared to English→French where the Scielo corpus size was 10 times smaller) and the additional use of the Pubmed corpora.

In a comparison among all participating teams, my submission ranked first for English→French for the Cochrane and NHS datasets, second on English→French and French→English on the EDP datasets, but only last on English→German for the Cochrane and NHS datasets. Moreover, it was the only one submitting for Scielo (Portuguese→English, English→Portuguese, Spanish→English, English→Spanish) as well as for Cochrane and NHS.

Lavie (2010) notes that BLEU scores above 30 reflect understandable translations, while scores over 50 are considered good and fluent translations. Within my 36 submissions, 24 obtained BLEU scores between ≈32 and ≈49, for six language pairs. Thus, the iATD method offers generally good translation results on a variety of language pairs.

The ratio of selected general domain data using iATD ranged between 3.1% and 9.35% using either the source or the target language and it ranged between 5.6% and 12.1% using the union of the selected sentences trained on both languages from both classifiers (see Table 6.6). In general, a small selection ratio like this is preferred as in large scale applications the general domain pool can consist even of billions of sentences. Therefore, iATD presents an advantage over the methods discussed in Chapter 5 since it not only avoids the need to train several MT systems, but also offers the possibility to choose a small selection ratio.

In the following, I will focus on the hybrid approach, hATD, a data selection method that I developed in order to improve iATD. I evaluated it in comparison with iATD for the Spanish→English language pair (the focus language pair in this thesis). Since the BLEU results are close to each other for the three variants, results are reported only for the classifiers trained on the English side of the corpora.

[11] http://k-connect.org/

[12] http://www.himl.eu/

Language pair	# selected src. sent.	# selected trg. sent.	Union
English→German	148K (3.1%)	188K (4.0%)	263K (5.6%)
English↔Spanish	327K (9.35%)	257K (7.36%)	425K (12.1%)
English↔French	223K (5.6%)	225K (5.7%)	345K (8.7%)
English↔Portuguese	78K (4.7%)	89K (5.3%)	123K (7.4%)

Table 6.6: Number of selected sentences and ratio selection of General domain (Duma and Menzel, 2017a)

The same amount of samples as for iATD was used to train the classifiers for hATD. The positive samples were randomly drawn from three medical corpora from which short sentences have been removed: EMEA, Scielo (2016) and Medline abstracts (made available by the WMT 2019 Biomedical task[13]). The negative samples were obtained by applying the *DSTF* method and selecting sentences with the lowest scores. This ensured that the number of false negatives could be kept low. As before, Doc2Vec was used for sentence representation: the optimal vector size was determined using the values 200, 300 and 400. Two hyperparameters of the classifier were tuned: the activation function (*Tanh*, *ExpLin* (Clevert et al., 2016) or *ReLU* (Nair and Hinton, 2010)) and the gradient descent optimization algorithm (*sgd*, *momentum* (Polyak, 1964) and *nesterov* (Nesterov, 1983)).

Table 6.7 presents the best classifier results obtained for both approaches. In particular, the best accuracy was obtained using *Tanh* and *momentum* for hATD and *ExpLin* and *nesterov* for iATD (Doc2Vec vector size of 400 for both). The following measures are reported: true negatives (TN), false positives (FP), false negatives (FN), true positives (TP), accuracy, precision, recall (Manning et al., 2008) and F1-score (van Rijsbergen, 1979). The formulae for these measures are presented below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 - score = \frac{2TP}{2TP+FP+FN}$$

For the accuracy reported on the test set, the same set is used for both methods.

The classifier for iATD selected 188K sentences and the classifier for hATD selected 526K sentences. The recall of the iATD classifier is worse than the recall of the hATD classifier showing that the hybrid approach wins over the initial method. Since the negative samples for training the iATD approach were randomly selected from the general domain, the 4.3% negatives that were actually positive samples resulted in misclassifying 183 sentences that were supposed to be identified as pseudo in-domain (2.7 times more false negatives than the hybrid method). Also,

[13] Available at: http://www.statmt.org/wmt19/biomedical-translation-task.html

the hybrid approach performed much better at identifying the pseudo in-domain sentences. The initial method identified a total of 5181 negatives, while the hybrid approach a total of 5065. As a consequence, iATD predicted less Biomedical domain sentences than the hybrid method as it predicts too many false negatives. The iATD classifier selected fewer sentences from the whole general domain pool as pseudo in-domain due to the negative training samples that contained positives. As a result, this intrinsic evaluation reveals that the hybrid approach, which used high-quality negative samples, outperforms the initial approach in terms of precision, recall and accuracy.

System	**hATD**	**iATD**
Accuracy test	99.31	98.15
True negatives	4998	4998
False positives	2	2
False negatives	67	183
True positives	4933	4817
Precision	1.00	1.00
Recall	0.986	0.963
F1-score	0.993	0.981

Table 6.7: Comparison between the hATD and iATD classifiers

Evaluating the approaches in the data selection pipeline, SMT systems were trained using the pseudo in-domain sentences selected using both classifiers. The obtained BLEU and OOV-rate results are given in Table 6.8.

System	**iATD**	**iATD-526k**	**hATD**	**hATD-188k**
BLEU	36.63	36.68	37.34	36.85
OOV	4.3	3.8	3.7	4.3

Table 6.8: Averaged BLEU scores and OOV rates for iATD and hATD

In order to ensure a fair comparison between the two approaches and to show that the gain of hATD over iATD is not caused by simply having more training data, I downgraded the hATD classifier results by subsampling the pseudo in-domain to 188K sentences and also upgraded the iATD classifier results by adding to its pseudo in-domain a sample of 338K (526K - 188K) sentences that were randomly selected from the general domain pool. The BLEU results are in accordance with the intrinsic evaluation: the hybrid approach outperforms the initial one in terms of BLEU (37.34 versus 36.63) and the OOV rate is also lower (3.7 versus 4.3).

Adding more training data to the pseudo in-domain selected using iATD does not improve the BLEU score (36.68 versus 36.63), but only reduces the OOV rate.

Subsampling the pseudo in-domain selected by means of the hybrid approach, hATD-188K, outperforms both iATD and iATD-526K.

In order to assess whether the difference in BLEU scores between the systems is statistically significant, I applied paired bootstrap resampling. The results are presented in Table 6.9. All trained systems were compared with each other. One of the results shows that hATD produces better translations than iATD in terms of BLEU (statistical significant result for the non-modified versions of the pseudo in-domains, p-value <0.001). Another result is that even though I upgraded the initial approach by adding more training data (iATD-526K), it still did not outperform iATD (they performed on a par), nor hATD, which performs better than iATD-526K (statistical significant). In conclusion, the gain that hATD achieves over the initial approach is statistical significant and it is due to the influence of the high-quality negative training samples. Moreover, this gain is not caused by the larger amount of training data as the two experiments, namely iATD-526K and hATD-188K, support this result.

Method	**p-value**	**Significance**
hATD vs iATD	<0.001	*better: Yes*
hATD vs hATD-188k	0.002	*better: Yes*
hATD-188k vs iATD	0.224	better: No
hATD vs iATD-526k	<0.001	*better: Yes*
iATD vs iATD-526k	0.327	better: No
hATD-188k vs iATD-526k	0.158	better: No

Table 6.9: Paired bootstrap resampling p-values for iATD and hATD

A high overlap of 96% ($\approx 180K$ sentences) between the sentences selected with iATD and hATD was observed (Figure 6.2). This does not come as a surprise, since essentially the same method is used, only the negative training samples differ. This result emphasizes the importance of using high-quality negative samples which help to identify a higher share of pseudo in-domain sentences. Since the intersection between iATD and hATD amounts to 96%, the sentences selected using iATD can be regarded as a subsample of the ones selected with hATD. Not surprisingly, hATD-188K performs on par with iATD, as both of them use 180K pseudo in-domain sentences subsampled from the same data (hATD pseudo in-domain of 536K). Moreover, the overlap between hATD-188K and iATD is only 6%, so the two subsamples share little common pseudo in-domain data. However, hATD-188K represents a downgrade of hATD and the full effect of the hybrid approach is highlighted by the high BLEU score.

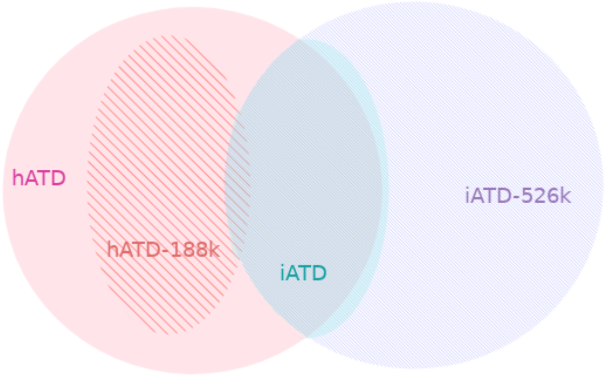


Figure 6.2: Overlap between iATD, hATD, iATD-526K and hATD-188K

6.3 Extended Evaluation

In order to evaluate all of my methods and compare them with the chosen state-of-the-art method, standard automatic evaluation metrics were applied. Moreover, the out-of-vocabulary rates were used to assess which methods produce the best vocabulary coverage. Since the methods described in this thesis were applied on different in-domains, test sets and language pairs, in this section the focus is on a common setting: the Biomedical domain, the Spanish→English language pair and two test sets.

In addition to the aforementioned empirical settings, automatic evaluation is explored on the Biomedical domain for the English→Spanish language pair and on the IT domain for English→German with the purpose of analyzing if the system rankings are independent of the in-domain and the language pair.

This section presents the empirical results obtained using standard automatic evaluation metrics (BLEU, METEOR and TER). Additionally, the out-of-vocabulary rates for the three sets of experiments are presented.

6.3.1 Data and Resources

A total of 32 SMT experiments were carried out on the same general domain and in-domain corpora as in Chapter 5, Section 5.1.2. The evaluation was on two test sets pertaining to the Biomedical domain (Khresmoi and WMT 2018(Neves et al., 2018) competitions test sets) for Spanish→English translations.

The development set for the Biomedical domain consisted of a concatenation of the Khresmoi development set (500 sentence pairs) and a cleaned up version of the ECDC data set (850 sentence pairs)[14]. Additionally to using the Khresmoi test set, all systems are evaluated on the WMT 2018 test set[15]. The two test sets differ in one essential aspect, their average sentence length, which is is 24 words for the

[14] More details can be read in Chapter 5, Section 5.1.2

[15] The latest WMT Biomedical test at the time of writing this chapter.

Khresmoi test and 30 words for WMT 2018.

A baseline using only the in-domain data (BS-IN) was trained. Also a baseline using the concatenation of the in-domain and general domain data using the same interpolated language model as the one in the data selection experiments was trained (BS-strong). The hATD method, resulted in a selection of 15% of the general domain (only one experiment because it is an automatic selection method). For the other methods, MML (state-of-the-art, cross-entropy based), DSTF (weighted term frequencies difference), SEF (PV-based scoring function) and RND (random selection), a selection in the range of [1, 5] percent[16] was used to determine the gain when dealing with very small pseudo in-domains and also ratio selections of 10 and 15 to study the impact of larger selections of general domain sentences. The ratio selection of 15 was chosen to compare these methods with the automatic data selection method in a fair manner. Experiments using random selection from the general domain were important in assessing whether the performance gain is due to simply using more training data or actually due to adding more pseudo in-domain data. This experimental setup has also been chosen for the investigations with the manual evaluation in Chapter 7.

Two other experimental settings were designed to assess whether the systems ranking provided by the automatic evaluation is stable across language pairs and in-domains: the Biomedical domain for English→Spanish translations and the IT domain for English→German translations. The motivation for the second setting is immediate: to explore the same domain (Biomedical) for the opposite translation direction. The third experimental setting was chosen due to the successful participation in the WMT 2016 IT domain task where my system ranked first on the constrained task for the English→German language pair. A straight comparison between this winning system and my other data selection methods represented the motivation for choosing the third setting. A subset of the selection steps was applied for these two settings: 1%, 5% and the selection ratio identified by the hATD method. In the case of the Biomedical English→Spanish setting the hATD method selected approximately 15% pseudo in-domain sentences, as for the IT English→German setting, the ratio was approximately 23%.

While the same training data was used for the Biomedical English→Spanish language pair (interchanged source and target sides of corpora), for the IT domain, the general domain corpora consisted of concatenating the Wikipedia (Wolk and Marasek, 2014) and the Commoncrawl corpora for English→German. The IT domain corpus, as well as the development and test set[17], are the same used in Chapter 5, Section 5.2.2.

6.3.2 Experimental Results

Following recommendations from the MT community, the averaged BLEU, METEOR and TER results over five system tunings is reported. Bootstrap resampling

[16] 1% constituted ≈ $35K$ sentence pairs.

[17] The only data difference compared with Chapter 5, Section 5.2.2 is that here the Wikipedia corpus was used additionally.

(Klejch et al., 2015) reveals if the difference between systems is statistically significant. To complete the evaluation, the OOV rates are analyzed.

Spanish→English for the Biomedical domain

The averaged BLEU scores for all the data selection methods, the random selection and the baselines are presented in Table 6.10 and visually depicted in Figure 6.3 for the Khresmoi test set and in Figure 6.4 for WMT18. For visual clarity, the standard deviation for the BLEU scores was not included in the graphics containing all methods, but in the four subplots (depicted using error bars colored in green). All baselines are represented using horizontal lines instead of dots for better visualization. Similarly, instead of representing the automatic ratio detection method using a dot at 15% selection step, an horizontal line was used, also for better visualization.

Method	**RND**	**MML**	**DSTF**	**SEF**	**hATD**	**RND**	**MML**	**DSTF**	**SEF**	**hATD**
Test set			Khresmoi					WMT 18		
1%	39.5	39.91	40.08	40.14		33.25	33.46	33.77	33.49	
2%	39.81	40.37	40.39	40.58		33.46	33.63	33.79	33.78	
3%	40.04	40.77	40.8	40.92		33.63	33.98	34.58	33.88	
4%	40.14	41.08	40.94	41.16		33.66	34.07	34.86	34.45	
5%	40.37	41.24	41.11	41.36		33.72	34.14	34.74	34.39	
10%	40.72	41.78	41.91	41.94		34.22	34.53	35.26	34.75	
15%	40.89	42.27	42.1	42.37	**42.47**	34.34	35	**35.42**	34.85	35.33
					Baselines					
BS-IN			38.35					32.44		
BS-GEN			38.56					33.51		
BS-strong			42.41					35.2		

Table 6.10: BLEU results for the Biomedical test sets for Spanish→English

The BLEU scores for the Khresmoi test set are higher than WMT18 as part of the development set was similar to the test set, both belonging to the Khresmoi data.

The averaged BLEU scores already indicate a ranking of the systems. However, in order to assess if the results are statistical significant, bootstrap resampling was applied with focus on evaluating systems per ratio selection.

On both test sets, the lowest performance in terms of BLEU is obtained with the baseline trained using only the in-domain (BS-IN). The baseline trained on the general domain (BS-GEN) performs almost on a par with BS-IN on the Khresmoi test set and it achieves significantly better results (p-value = 0.024) than BS-IN on the WMT 18 test set with almost 1 BLEU point difference. The strong baseline (BS-strong) significantly outperforms the other two baselines (p-value < 0.001) on both test sets. This is due to using the concatenation of the in-domain and general domain corpora for training and interpolating the language models for the two corpora.

RND produces the lowest BLEU scores, for all selection steps, on both test sets. Applying significance tests to the Khresmoi test set reveals that RND performs significantly worse than any of the data selection methods, on every step. Investigating the BLEU scores per step, by randomly adding more training data indeed improves the translation quality, however, the results using RND are much lower than any of the data selection methods. The difference in BLEU scores increases with the ratio of the selection. On the other hand, on the WMT18 test set, the random selection is significantly outperformed by all selection ratios only when comparing RND with DSTF and hATD[18]. The p-values when comparing the other two data selection methods with RND were quite small for selection ratios > 3, however the results were not statistically significant. While the statistical tests did not reveal significance across all system combinations for both test sets, it could be concluded that, in general, random selection performs worse than pseudo in-domain selection. These results indicate that the better BLEU performance is not a result of generally adding more training data (random selection), but effected by adding more pseudo in-domain training data (via data selection). Thus, the research question **RQ4** stated in the introduction can be answered by concluding that the experiments using random selected sentences out of the general domain are generally performing worse than the data selection methods.

Considering the BLEU performance of all ratio-tuning data selection methods on the Khresmoi test set, the SEF method outperforms the state-of-the-art method, MML, and the term-frequency method, DSTF, on all selection steps, however, the BLEU differences are not statistical significant on all steps. The automatic ratio detection method, hATD, is superior to all methods on the 15% selection of pseudo in-domain and to the strong baseline (not statistically significant). By using 15% of pseudo in-domain data obtained with any of the data selection methods, the same translation quality is achieved as with the full selection pool. This emphasizes the benefit of applying data selection to reduce the size of the training data, which directly results in faster training of MT systems and obtaining smaller models which are easier to store and load into memory at translation time.

Considering the WMT 2018 test set, when comparing the BLEU scores of the state-of-the-art method, MML, with the term frequency method, DSTF, the latter significantly outperforms MML and on 3% selection (p-value = 0.034), on 4% selection (p-value = 0.015) and on 10% selection (p-value = 0.002). When comparing MML with SEF, the only ratio where SEF outperforms MML is 2% (p-value = 0.05). It is important to note that MML does not significantly outperform any of my selection methods on any ratio step. At least for this particular test set, DSTF achieves a much better translation quality than MML and SEF performs slightly better than MML. The hATD method produced very good results, better than MML and SEF, but worse than DSTF (not statistically significant). Similar to the other test set, using 15% of pseudo in-domain data for the data selection methods performs on a par with the strong baseline.

Table 6.11 shows examples of segment-level BLEU scores for translations ob-

[18] With one exception on 2% selection ratio of RND and DSTF

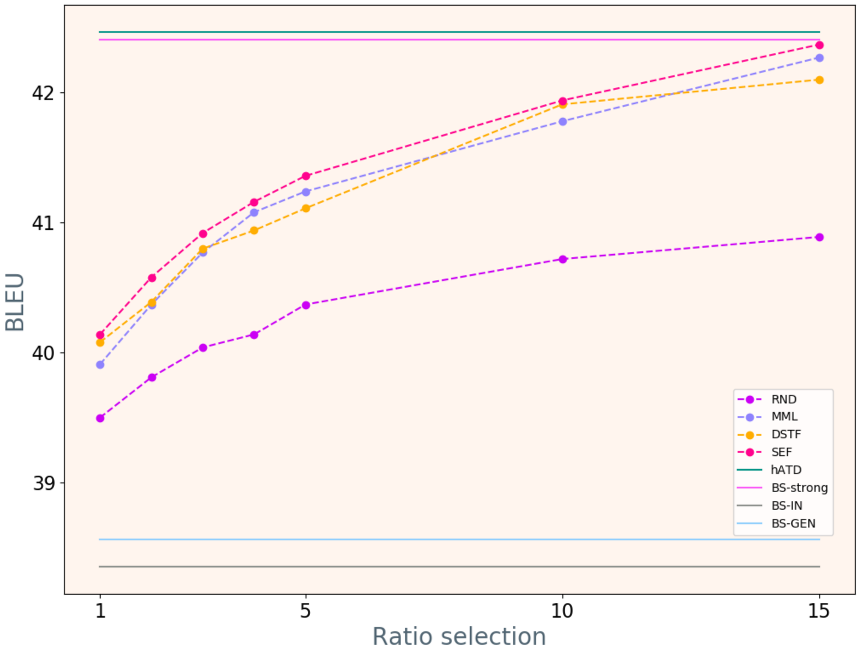


Figure 6.3: BLEU graphic of all methods for the Khresmoi test set

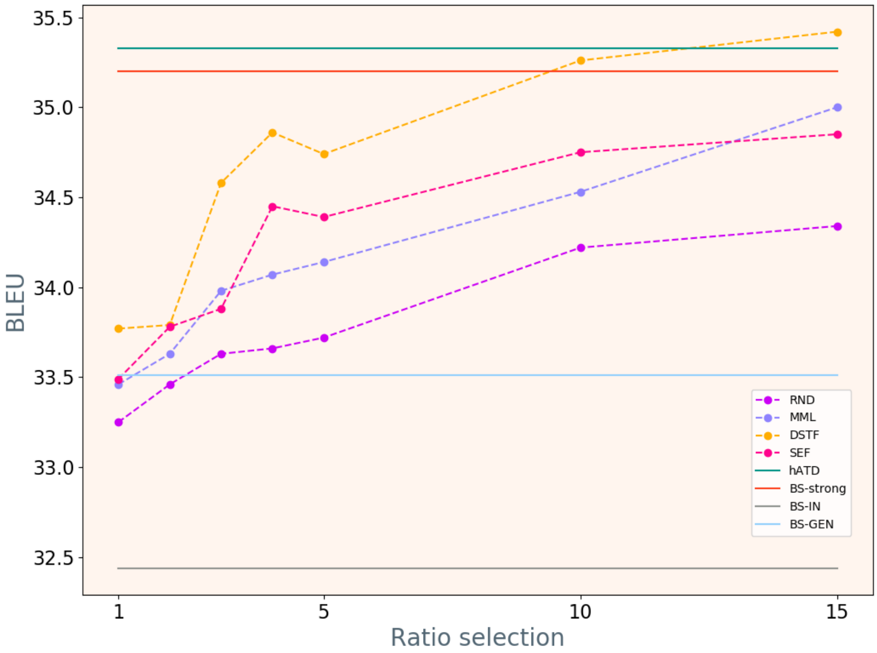


Figure 6.4: Averaged BLEU of all methods for the WMT18 test sets

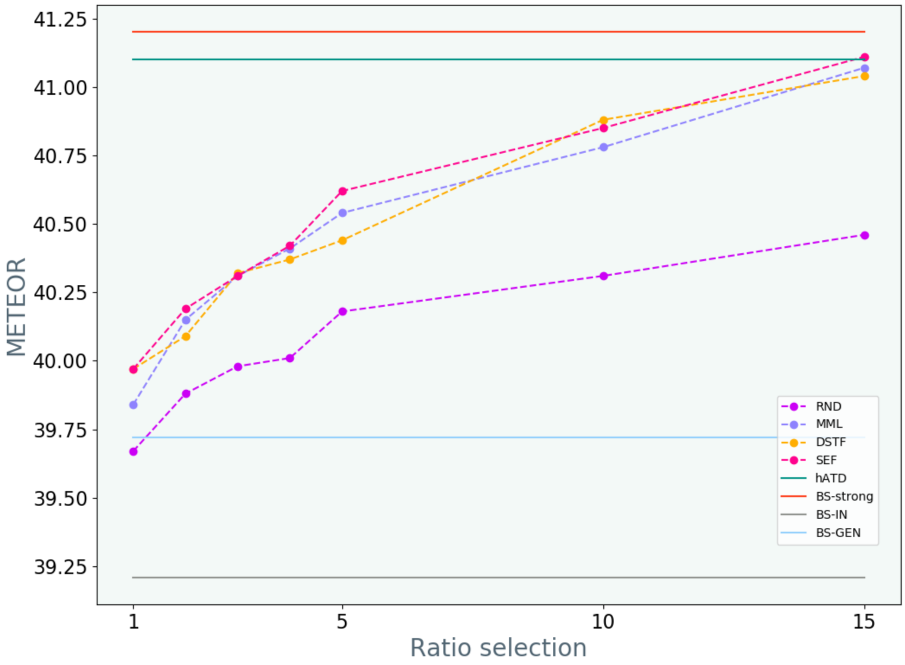


Figure 6.5: Averaged METEOR of all methods for the Khresmoi test set

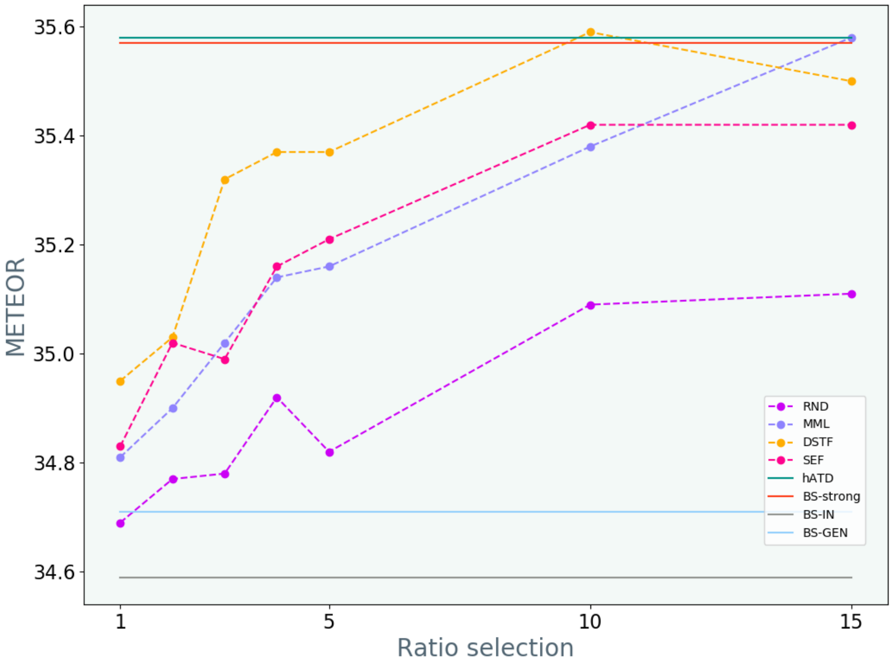


Figure 6.6: Averaged METEOR of all methods for the WMT 18 test set for Spanish→English

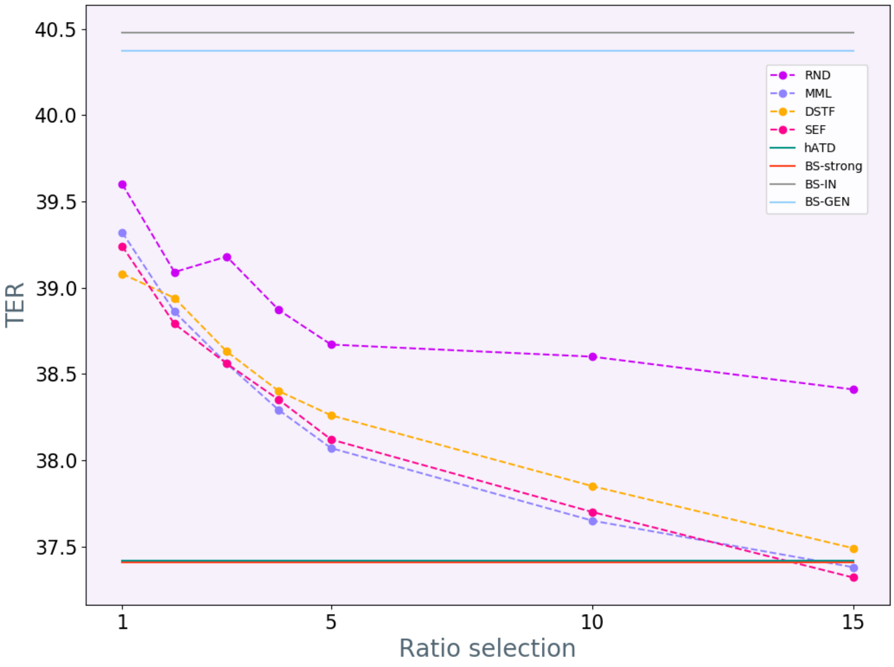


Figure 6.7: Averaged TER of all methods for the Khresmoi test set

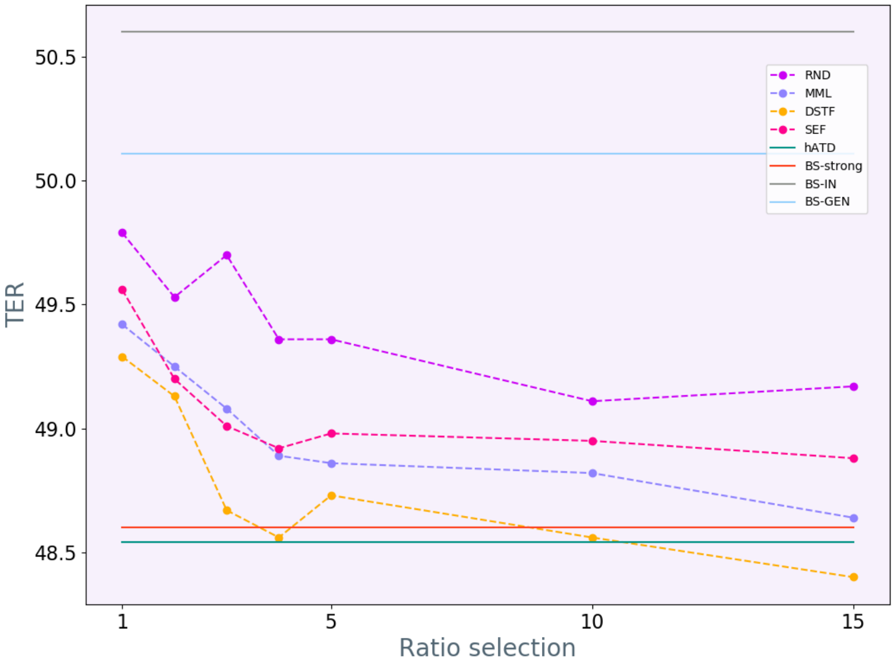


Figure 6.8: Averaged TER of all methods for the WMT 18 test set for Spanish→English

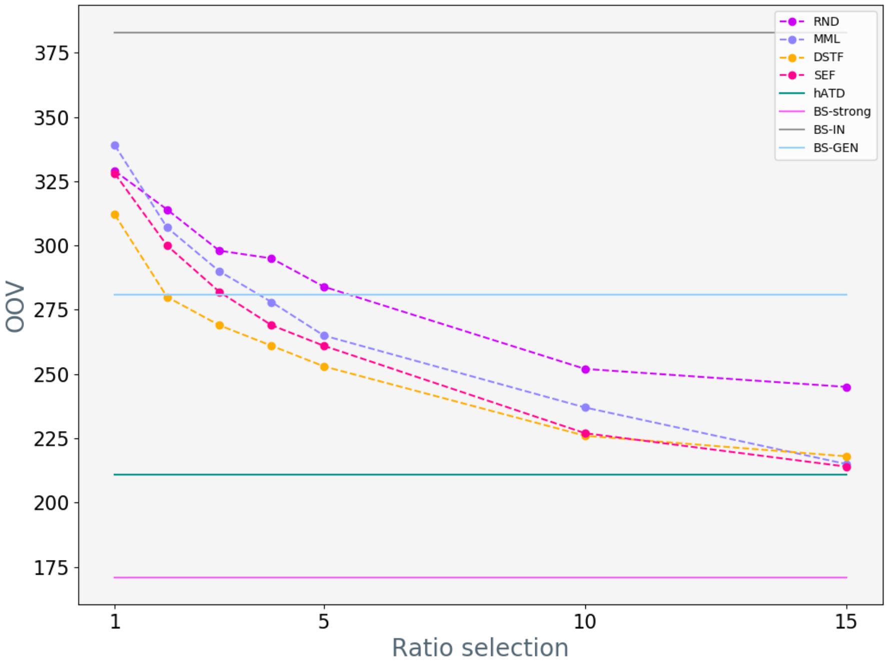


Figure 6.9: OOV graphic of all methods for the Khresmoi test set

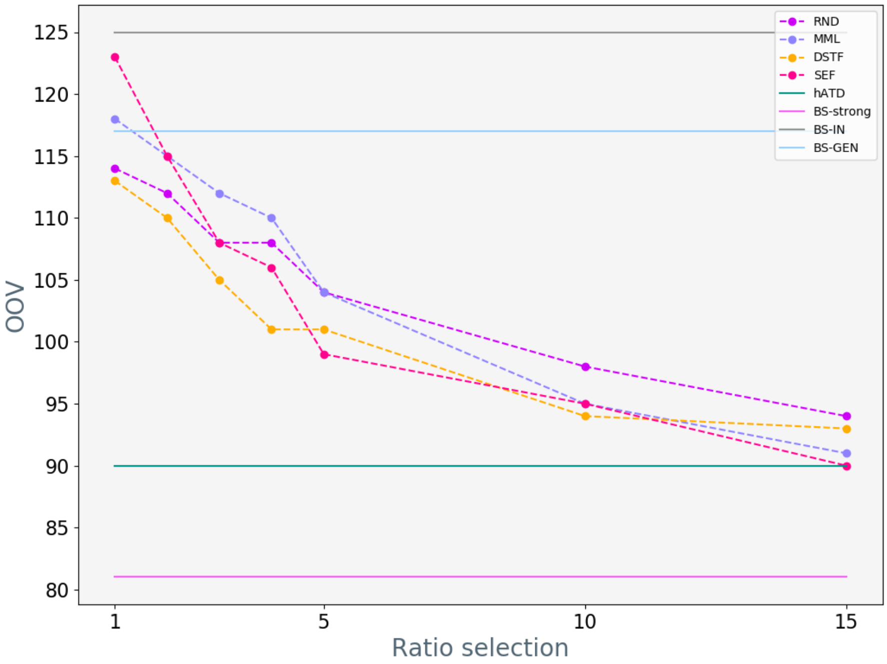


Figure 6.10: OOV rate of all methods for the WMT 2018 test set for Spanish→English

tained with all systems for the Khresmoi test set. In the first example, all systems produce poor translations of the source sentence. All of them failed to translate eritrocitaféresis (eng. reference: erythrocytapheresis) since it is an out-of-vocabulary word. The best BLEU score is obtained with MML which correctly translates ”se usará” into ”will be used”. MML is outperformed by all systems in the second example as it misses the translation of the word administración (eng. reference: administration). It is important to point out that the DSTF translation is actually identical to the reference even though it has a 0.892 score. This is due to the fact that the reference uses 3 corresponding to tres and DSTF (as well as all systems) translated it with three. The last example presents SEF offering the best translation with a maximum BLEU score. The challenge in this source sentence is translating terminology: insuficiencia renal (eng. reference: renal insufficiency) is correctly translated by SEF (renal insufficiency) and DSTF (renal failure), while hATD and MML offer a partial translation (renal).

Some examples of translations for the WMT18 test set are provided in Table 6.12. The best BLEU score is achieved by DSTF in the first example where the first part of the source sentence is correctly translated. The other methods fail at translating es preciso adoptar (eng. reference: is needed) which DSTF translated as it is necessary to adopt. Even though this part of the sentence is not completely the same as the reference counterpart, the translation produced with DSTF has the same meaning as the reference (is needed). In the second example the difficult part to translate is the enumeration Clásica, Española y Flamenco preceded by the noun grupos (eng. reference: groups (Classical, Spanish and Flamenco)). The reference encompasses the enumeration items into parentheses, even though the initial source sentence does not include them. A translation closer to the source sentence would be the Classical, Spanish and Flamenco groups, which reinforces the need to have more than one reference. Finally, the last example demonstrates the difficulty to translate very long sentences (a source sentence containing 55 tokens). Even though the reference presents the translation for ayudas diagnosticas as diagnostic facilities, the translation produced with MML is also viable (diagnostic aids).

Figures 6.5 and 6.6 depict the METEOR scores for both test sets and Figures 6.7 and 6.8 depict the TER scores. It can be observed that the METEOR and TER evaluation results are in general consistent with the BLEU scores for both test sets.

Figure 6.9 depicts the out-of-vocabulary rate of all the methods for the Khresmoi test set. All baselines are illustrated using a horizontal line since only one system is trained. The same applies for the automatic ratio detection method, hATD. The baseline trained using only the in-domain data suffers from the highest unknown words count. This emphasizes the importance of adding more training samples pertaining to the in-domain. The baseline trained only with the general domain data achieves an OOV rate on a par with the data selection methods trained using very small pseudo in-domains (1% - 3% selection).

Source	Se usará una máquina de aféresis para quitar los hematíes sólo del grupo de la eritrocitaféresis.	
Reference	An apheresis machine will be used to remove red blood cells only from the erythrocytapheresis group.	
hATD translation	We will use a multinational apheresis only for the group of eritrocitaféresis.	**0.096**
DSTF translation	We used a machine to remove red only Apheresis Group of eritrocitaféresis.	**0.164**
SEF translation	We used a machine to remove apheresis red only the eritrocitaféresis group.	**0.141**
MML translation	A apheresis machine will be used to remove red only of the eritrocitaféresis.	**0.478**

Source	Las complicaciones fueron más frecuentes inmediatamente después de la administración de los agentes quimioterapéuticos, con una declinación gradual sobre las siguientes tres semanas.	
Reference	Complications were more frequent immediately after administration of the chemotherapeutic agents, with a gradual decline over the following 3 weeks.	
hATD translation	Complications were more frequent immediately after administration of chemotherapeutic agents, with a gradual decline over the next three weeks.	**0.720**
DSTF translation	Complications were more frequent immediately after administration of the chemotherapeutic agents, with a gradual decline over the following three weeks.	**0.892**
SEF translation	Complications were more frequent immediately after administration of chemotherapeutic agents, with a gradual decline on the following three weeks.	**0.652**
MML translation	Complications were more common immediately after the chemotherapeutic agents, with a gradual decline over the next three weeks.	**0.553**

Source	En conclusión, MAHA es un indicador importante de la insuficiencia renal y la recuperación en pacientes con hipertensión maligna.	
Reference	In conclusion, MAHA is an important indicator of renal insufficiency and recovery in patients with malignant hypertension.	
hATD translation	In conclusion, MAHA is an important indicator of the renal and recovery in patients with malignant hypertension.	**0.814**
DSTF translation	In conclusion, MAHA is an important indicator of renal failure and recovery in patients with malignant hypertension.	**0.857**
SEF translation	In conclusion, MAHA is an important indicator of renal insufficiency and recovery in patients with malignant hypertension.	**1.0**
MML translation	In conclusion, MAHA is an important indicator of the renal and recovery in patients with malignant hypertension.	**0.814**

Table 6.11: Examples of segment-level BLEU scores for the Khresmoi test set for Spanish→English

Source	Una vez se confirma, es preciso adoptar un enfoque sindrómico y usar una valoración geriátrica integral para determinar sus causas y elaborar un plan de tratamiento que incluya tanto el tratamiento de los síntomas como el etiológico.	
Reference	Once confirmed, a syndromic approach is needed, based on a comprehensive geriatric assessment in order to determine its causes and prepare a treatment plan which addresses the treatment of symptoms as well as the etiology.	
hATD translation	Once confirmed an syndromic approach and using a comprehensive geriatric assessment to determine its causes and develop a treatment plan that includes both the etiological as.	**0.227**
DSTF translation	Once confirmed, it is necessary to adopt a syndromic approach and using a comprehensive geriatric assessment to determine its causes and develop a treatment plan that includes both the treatment of the symptoms such as etiological.	**0.305**
SEF translation	Once confirmed, and syndromic approach using a comprehensive geriatric assessment to determine its causes and develop a treatment plan that includes both the treatment of etiologic as.	**0.271**
MML translation	Once confirmed, syndromic approach and using a comprehensive geriatric assessment to determine its causes and develop a treatment plan that includes both the etiological as.	**0.242**
Source	Las diferencias en el porcentaje de grasa en masa entre los grupos Clásica, Española y Flamenco fueron evaluadas mediante un análisis de medidas repetidas (ANOVA).	
Reference	Differences in percent fat mass between groups (Classical, Spanish and Flamenco) were tested by using repeated measures analysis (ANOVA).	
hATD translation	The differences in the percentage of fat mass between Spanish classical and flamenco were evaluated using repeated measures (ANOVA).	**0.264**
DSTF translation	The differences in the percentage of fat mass between groups, Spanish classical and flamenco were evaluated using repeated measures (ANOVA).	**0.331**
SEF translation	The differences in the percentage of fat mass classical between groups in Spanish, and flamenco were evaluated using repeated measures (ANOVA).	**0.258**
MML translation	The differences in the percentage of fat mass between groups, and classic Spanish flamenco were evaluated using repeated measures (ANOVA).	**0.312**
Source	Este artículo es una revisión general de las herramientas diagnósticas que el médico clínico puede usar para el diagnóstico temprano de la apendicitis aguda con énfasis en la escala de Alvarado, y está destinado principalmente a los médicos generales en diferentes partes del mundo donde las ayudas diagnosticas y los recursos tecnológicos son limitados.	
Reference	This article is a general review of the diagnostic tools that the clinician can use for the early diagnosis of acute appendicitis with emphasis on the Alvarado Score, and it is aimed principally to the medical practitioners in different parts of the world where the diagnostic facilities and technological resources are limited.	
hATD translation	This article is a review of the clinical doctor diagnostic tools that can be used for the early diagnosis of acute appendicitis with emphasis on the scale of Alvarado, and is intended primarily to the general practitioners in different parts of the world where diagnostic aid and technological resources are limited.	**0.525**
DSTF translation	This article is a review of the clinical diagnostic tools that can be used for the early diagnosis of acute appendicitis with emphasis on the scale of Alvarado, and is intended primarily to the general practitioners in different parts of the world where the diagnostic helps and technological resources are limited.	**0.561**
SEF translation	This article is a review of the clinical diagnostic tools that can be used for the early diagnosis of acute appendicitis with emphasis on the scale of Alvarado, and is intended primarily to the general practitioners in different parts of the world where the diagnostic helps technological resources are limited.	**0.541**
MML translation	This article is a review of the clinical diagnostic tools that can be used for the early diagnosis of acute appendicitis with emphasis on the scale of Alvarado, and is intended primarily to the general practitioners in different parts of the world where the diagnostic aids and technological resources are limited.	**0.561**

Table 6.12: Examples of segment-level BLEU scores for the WMT18 test set for Spanish→English

By adding only 1% of pseudo in-domain data, the OOV count decreases from 383 unknown words (obtained with BS-IN) to 312 unknown words (obtained with DSTF). When comparing the 15% selection of pseudo in-domain, all methods perform on a par, with the random selection performing worse. As the graphics shows, the DSTF method achieves the lowest OOV rates across all selection ratios and the random selection produces the highest OOV rates when comparing the experiments with ratio selection (from 1% until 10%). As expected, the best OOV coverage is obtained using the full concatenation of general domain and in-domain data (BS-strong).

A similar OOV behavior is observed on the WMT 2018 test set (Figure 6.10). The lowest vocabulary coverage is obtained with the baseline trained using only the in-domain data, while the highest coverage is given by the strong baseline. This result is intuitive since more training data leads to fewer unknown words. However, minimizing the OOV rate is not the only goal when training MT systems, but also obtaining systems that perform well on target domains, that require less training data, thus decreased training time, and consequently, less disk space and faster loading into memory. With these goals in mind, data selection is the technique that allows for a compromise between low OOV rate and smaller, faster models that perform the same or even better than the models trained on the full data. Results on the WMT 2018 set are consistent with the ones obtained on the Khresmoi set. All data selection methods at 15% ratio are highly competitive with the strong baseline.

English→Spanish for the Biomedical domain

The selection steps considered for this experimental settings were 1%, 5% and 15%, with the latter representing the ratio chosen with the automatic data selection method. The same evaluation procedure as described before was employed: comparing system performance against each other in terms of BLEU, METEOR, TER and the out-of-vocabulary rate measured for different ratio selection.

Table 6.13 presents the BLEU scores obtained using all systems. When selecting only 1% of pseudo in-domain data, DSTF performs best on the Khresmoi test set, followed up with small differences by the strong baseline and by SEF. The worst performing system is MML for 1% selection, being statistically outperformed by DSTF with a p-value of 0.005. The results on the WMT18 test set indicate that all data selection methods, including the state-of-the-art, perform almost the same as BS-strong with very small BLEU differences (not statistically significant). This is a very important result because it demonstrates the benefits of data selection: using 1% versus 100% of the selection pool yields a much faster training and decreased resource requirements.

Method	MML	DSTF	SEF	hATD	MML	DSTF	SEF	hATD
Test set	Khresmoi				WMT 18			
1%	34.34	34.85	34.65		32.31	32.35	32.25	
5%	35.3	35.28	35.34		32.45	32.45	32.36	
15%	35.29	35.39	35.68	**35.89**	32.46	32.71	**33.03**	32.8
BS-strong	34.71				32.21			

Table 6.13: BLEU results for the Biomedical test sets for English→Spanish

For a selection ratio of 5%, all data selection methods achieve a similar translation quality in terms of BLEU for both test sets (very small differences with no statistical significance). When comparing them with BS-strong on the Khresmoi test set, MML and SEF significantly outperform the baseline with p-values of 0.017 and 0.001, respectively. Although for the WMT18 test set all data selection methods give better BLEU scores, bootstrap resampling indicated no statistical significance.

When considering the 15% selection, the hATD method achieves the best output quality according to BLEU on the Khresmoi test set (35.89) and SEF on the WMT18 set (33.03). Significance tests reveal that on the Khresmoi test set SEF significantly outperforms MML with a p-value of 0.012, while hATD outperforms MML with a p-value of 0.031. On the WMT18 test set, bootstrap resampling shows that SEF is significantly better than MML with a p-value of 0.016.

Method	MML	DSTF	SEF	hATD	MML	DSTF	SEF	hATD
Test set	Khresmoi				WMT 18			
1%	62.43	62.82	62.58		58.24	58.31	58.36	
5%	63.41	63.29	63.44		58.67	58.44	58.54	
15%	63.3	63.3	63.73	**63.99**	58.78	58.98	**59.09**	58.97
BS-strong	63.09				58.45			

Table 6.14: METEOR results for the Biomedical test sets for English→Spanish

The results for METEOR are given in Table 6.14 and the ones for TER in Table 6.15. Both evaluation metrics are consistent with the BLEU results for the Khresmoi test set with system rankings that place hATD on the first place, followed by SEF, DSTF, MML and finally, BS-strong. On the WMT18 test set, BLEU and METEOR place SEF first, while TER places hATD first.

Method	MML	DSTF	SEF	hATD	MML	DSTF	SEF	hATD
Test set	Khresmoi				WMT 18			
1%	46.52	46.11	46.39		54.06	53.89	54.01	
5%	45.53	45.67	45.52		53.53	53.72	53.77	
15%	45.69	45.57	45.3	**44.97**	53.54	53.56	53.28	**53.17**
BS-strong	47.03				54.63			

Table 6.15: TER results for the Biomedical test sets for English→Spanish

The number of unknown words for both test sets is given in Table 6.16. Not surprisingly, the lowest OOV is achieved on both test sets using BS-strong as the systems used full general domain training data. All data selection methods show very similar OOV numbers on the 15% selection, on both test sets.

Method	**MML**	**DSTF**	**SEF**	**hATD**	**MML**	**DSTF**	**SEF**	**hATD**
Test set	Khresmoi				WMT 18			
1%	186	179	184		94	90	97	
5%	150	145	151		84	84	85	
15%	126	127	129	126	77	77	78	80
BS-strong	**102**				**69**			

Table 6.16: Out-of-vocabulary for the Biomedical test sets for English→Spanish

Some examples of translations using all systems together with their BLEU sentence scores at the 15% selection are given in Table 6.17 for the Khresmoi test. Words highlighted in yellow represent wrong translations, while pink denotes extra words. In the first example, the best output is obtained with the hATD method which achieves almost a perfect score, with only one error by mistranslating basement with basal instead of base. The same behavior is encountered with all the other data selection methods. However, investigating the correct translation of basement membrane into Spanish using diverse linguistic tools (*DeepL Linguee*[19]) shows that the correct translation was found by all data selection systems. While the DSTF and SEF methods each introduce one extra word when translating the source sentence, the MML method fails at translating the verb show with the correct tense as it outputs the imperative form with the pronoun le. The second example presents lower-quality translations in terms of BLEU scores. All MT systems perform on a par. This example illustrates how BLEU fails to assign high scores to translations that contain synonyms to words from the reference (in this case lugar and vez). Also, the translations use the formal form of the possessive adjective in contrast with the reference that uses the informal one (su piel instead of tu piel). Another issue with this example is that the reference uses the singular form of the English noun Bacteria which has its singular form Bacterium. Therefore, all MT systems correctly translate the second part of the source sentence.

Examples of translations from the WMT18 test set are given in Table 6.18. In the first example, while all MT systems translate the verb to have in the correct tense, they all fail in the conjugation by using singular instead of the plural form (tuvo versus tuvieron). The reference uses a different verb (presentó), however a closer translation to the source sentence should use tuvieron. All MT systems have difficulties in translating non-obese patients, with SEF offering the most fluent version, while omitting to translate non, which in the end turns the whole sentence into being non-understandable (English translation: *The obese patients had a favorable clinical evolution in comparison with the group of obese patients.*). Even though the second part of the source sentence is not translated by the systems

[19] https://www.linguee.com/english-spanish/translation/basement+membrane.html

Source	The nuclei are uniform in size and shape and show normal polarity with their axes perpendicular to the basement membrane.	
Reference	Los núcleos son uniformes en tamaño y forma y muestran polaridad normal con sus ejes perpendiculares a la membrana base.	
hATD translation	Los núcleos son uniformes en tamaño y forma y muestran polaridad normal con sus ejes perpendiculares a la membrana basal.	**0.913**
DSTF translation	Los núcleos son uniformes en tamaño y forma y muestran una polaridad normal con sus ejes perpendiculares a la membrana basal.	**0.794**
SEF translation	Los núcleos son uniformes en tamaño y forma y muestran la polaridad normal con sus ejes perpendiculares a la membrana basal.	**0.794**
MML translation	Los núcleos son uniformes en tamaño y forma y muéstrele polaridad normal con sus ejes perpendiculares a la membrana basal.	**0.783**
Source	Instead of making your skin look better, tea tree oil works at the source of acne: the bacteria found on the skin's surface.	
Reference	En vez de hacer que tu piel tenga mejor apariencia, el aceite del árbol del té trabaja en la fuente del acné: la bacteria que se encuentra en la superficie de la piel.	
hATD translation	En lugar de hacer que su piel look mejor, aceite esencial de árbol de te trabaja en la fuente de acné: las bacterias que se encuentran en la superficie de la piel.	**0.362**
DSTF translation	En lugar de hacer que su piel se ven mejor, aceite esencial de árbol de té trabaja en la fuente de acné: las bacterias que se encuentran en la superficie de la piel.	**0.362**
SEF translation	En lugar de hacer que su piel luzca mejor, aceite esencial de árbol de té trabaja en la fuente de acné: las bacterias que se encuentran en la superficie de la piel.	**0.362**
MML translation	En lugar de hacer que su piel luzca mejor, aceite esencial de árbol de té trabaja en la fuente del acné: las bacterias que se encuentran en la superficie de la piel.	**0.440**

Table 6.17: Examples of segment-level BLEU scores for the Khresmoi test set English→Spanish

using the same word choices as the reference, it still has a good quality with the only flaw consisting in omitting to translate more into más. The second example highlights an out-of-vocabulary word, retinovascular, and the difficulty of the MT systems in translating therapy at follow-up. In this example, SEF produces the most fluent and closest translation to the reference.

English→German for the IT domain

In this subsection a technical domain, the IT domain, is investigated for the English→German language pair. As in the previously presented experimental setting, a subset of the selection steps was used for the experiments: 1%, 5% and

Source	Non-obese patients had a more favorable clinical course compared to the group of obese patients.	
Reference	El grupo de pacientes no obesos presentó una evolución clínica más favorable comparado con el grupo de pacientes con obesidad.	
hATD translation	No - pacientes obesos tuvo una evolución clínica favorable en comparación con el grupo de pacientes obesos.	**0.300**
DSTF translation	Non - pacientes obesos tuvo una evolución clínica favorable en comparación con el grupo de pacientes obesos.	**0.294**
SEF translation	Los pacientes obesos tuvo una evolución clínica favorable en comparación con el grupo de pacientes obesos.	**0.291**
MML translation	No pacientes obesos tuvo una evolución clínica favorable en comparación con el grupo de pacientes obesos.	**0.297**

Source	None of 21 patients with retinovascular changes required any therapy at follow-up.	
Reference	Ninguno de los 21 pacientes con cambios en la vasculatura de la retina requirió tratamiento durante el seguimiento.	
hATD translation	Ninguno de los 21 pacientes con retinovascular cambios requiere cualquier tratamiento , a seguimiento.	**0.331**
DSTF translation	Ninguno de los 21 pacientes con retinovascular cambios requiere cualquier tratamiento , a el seguimiento.	**0.371**
SEF translation	Ninguno de los 21 pacientes con retinovascular cambios requiere cualquier tratamiento en el seguimiento.	**0.374**
MML translation	Ninguno de los 21 pacientes con retinovascular cambios requiere cualquier tratamiento con el seguimiento.	**0.367**

Table 6.18: Examples of segment-level BLEU scores for the WMT18 test set for English→Spanish

23% (with the latter being the ratio of pseudo in-domain sentences selected with hATD).

Automatic evaluation was performed for the same data selection methods previously investigated. The focus in on automatic metrics (BLEU, METEOR, TER), out-of-vocabulary rate and statistical significance tests (like in the previous section).

The BLEU results are presented in Table 6.19. The best performance is obtained with the hybrid automatic ratio method (37.3), while the system trained using 23% pseudo in-domain sentences selected with SEF follows close (37.16). The 23% selection using the state-of-the-art method, MML (36.96), and the term frequency method, DSTF (36.85), also achieve good BLEU results that outperform the strong baseline (36.74). The difference in BLEU score between hATD and BS-strong is statistical significant (p-value < 0.001), as well as with MML 23% (p-value $= 0.026$) and DSTF 23% (p-value $= 0.002$). SEF 23% is also significantly better than BS-strong (p-value $= 0.006$), while MML 23% and DSTF 23% outperform BS-strong too, but not significantly. These results confirm the effectiveness of hATD on the chosen setting (IT domain, English→German) when compared with the strong baseline: only 23% selection of the general domain data outperforms the baseline. Moreover, hATD is the winning method among all the data selection methods that were compared in this setting.

Method	MML	DSTF	SEF	hATD
1%	34.77	34.81	34.61	
5%	36.41	36.03	36.44	
23%	36.96	36.85	37.16	**37.3**
BS-strong		36.74		

Table 6.19: BLEU results for the IT test set for English→German

When inspecting the other selection steps, it can be observed that neither the 1% nor the 5% selection steps is able to outperform the strong baseline. However, this is not a negative result, since it only confirms that selecting a too small amount of data limits the chances to obtain a system that outperforms the baseline.

Method	MML	DSTF	SEF	hATD
1%	53.67	53.41	53.52	
5%	55.02	54.74	55.01	
23%	55.45	55.47	55.65	**55.74**
BS-strong		55.11		

Table 6.20: METEOR results for the IT domain test sets for English→German

The METEOR evaluation scores are presented in Table 6.20 where the best performing method is hATD, followed by SEF 23% selection, and MML 23% close to DSTF 23% . The results obtained with this evaluation metric are in concordance with the ones based on BLEU. Slightly different results are obtained using TER (see Table 6.21) where the best performing method is MML 23% closely followed by hATD.

Method	MML	DSTF	SEF	hATD
1%	45.23	44.52	44.59	
5%	42.96	44.24	43.66	
23%	**42.12**	42.93	42.45	42.29
BS-strong		43.45		

Table 6.21: TER results for the IT domain test sets for English→German

The out of vocabulary rate is presented in Table 6.22. Not surprisingly, the best vocabulary coverage is obtained with the baseline (3.3) as it uses all training data. Low OOV rates are also obtained with the 23% selection of SEF (3.7) and hATD (3.8). The unknown words lists included concatenated words that should have been separated by space in the test file (e.g., "newfolder", "andclick", "tocontrol"), misspellings (e.g., "uninstal", "pintrest", "aaccess", "downolad") and many host and domain names (e.g., "developers.google.com", "www.codecademy.com", "iforgot.apple.com"). Some of the test set problems could have been resolved with external tools like a spell checker, but I decided not to modify the test set because it impedes a comparison with other MT systems trained and tested on the original data sets.

Method	MML	DSTF	SEF	hATD
1%	274 (11.7)	180 (7.7)	220 (9.4)	
5%	165 (7.1)	124 (5.3)	120 (5.1)	
23%	93 (4.0)	93 (4.0)	87 (3.7)	90 (3.8)
BS-strong		**78 (3.3)**		

Table 6.22: Out-of-vocabulary for the IT domain test sets for English→German

Examples of translations for all systems together with their BLEU by sentence scores at 23% selection are given in Table 6.23. Words highlighted in yellow represent wrong translations (when compared to the reference) and pink denotes extra words.

In the first example, both hATD and MML produce the best translations in terms of BLEU, with the most severe errors being the wrong translation of the IT technical term Befehl Bildgrenze and being unable to translate the verb is from the last part of the input sentence. Similar errors are made by the other systems, with the exception that even though the DSTF translation has the lowest BLEU score, it is the only one that produces a translation of the verb is. However, the sentence structure is incorrect (ob die gestrichelte Linie ist) as the word gestrichelt should appear after the word Linie.

In the second example, the best BLEU score is achieved by SEF and the worst by MML. The IT term Home Tab could not be translated by any of the systems, as well as the last part of the input (and begin typing to add text). The closest attempt to translate this is done by DSTF and MML, but their output is incorrect.

The last example involves translations that received poor BLEU scores due to the mistranslation of the IT terms Preview command and Transitions tab. However, the reference represents only one of the possible translations and the German term Vorschaubefehl exists, as well as the English-borrowed word Tab[20]. The systems translated Preview command with Vorschau Befehl, which is close to the compound term Vorschaubefehl. The MML system attempted to merge the two words using a hyphen.

[20] See `https://www.linguee.com/english-german/search?query=tab`

Source	Select the image, then click the Format tab, click the Picture Border command. Select a color, weight (thickness), and whether or not the line is dashed.	
Reference	Wählen Sie das Bild, klicken Sie dann auf die Registerkarte Format, klicken Sie auf den Befehl Bildgrenze. Wählen Sie eine Farbe, Gewicht (Dicke), und ob die Linie gestrichelt sein soll.	
hATD translation	Wählen Sie das Bild, klicken Sie dann auf die Registerkarte Format, klicken Sie auf das Bild Umrandung Befehl. Wählen Sie eine Farbe, Gewicht (Dicke), und ob die gestrichelte Linie.	**0.762**
DSTF translation	Wählen Sie das Bild, klicken Sie dann auf das Format Tab, klicken Sie auf das Bild Grenze Befehl. Wählen Sie eine Farbe, Gewicht (Dicke) und ob die gestrichelte Linie ist.	**0.566**
SEF translation	Wählen Sie das Image und klicken Sie dann auf die Registerkarte Format, klicken Sie auf das Bild Grenze Befehl. Wählen Sie eine Farbe, Gewicht (Dicke), und ob die gestrichelte Linie.	**0.675**
MML translation	Wählen Sie das Bild, klicken Sie dann auf die Registerkarte Format, klicken Sie auf das Bild Umrandung Befehl. Wählen Sie eine Farbe, Gewicht (Dicke), und ob die gestrichelte Linie.	**0.762**
Source	From the Home tab click New Slide, choose the desired slide layout from the menu that appears. Click any placeholder and begin typing to add text.	
Reference	Von der Registerkarte Start, klicken Sie auf Neue Folie, wählen Sie das gewünschte Folienlayout aus dem erscheinenden Menü. Klicken Sie auf einen Platzhalter und beginnen einen Text hinzuzufügen.	
hATD translation	Aus dem Hause Tab klicken Sie auf Neue Folie, wählen Sie die gewünschte Folie Layout aus dem Menü, das erscheint. Klicken Sie auf einen Platzhalter und beginnen Sie, um Text.	**0.448**
DSTF translation	Aus dem Hause Tab klicken Sie auf Neue Folie, wählen Sie die gewünschte Folienlayout aus dem Menü, das erscheint. Klicken Sie auf jedem Platzhalter und beginnen Sie zum Hinzufügen von Text.	**0.410**
SEF translation	Aus dem Hause Tab klicken Sie auf Neue Folie, wählen Sie die gewünschte Folienlayout aus dem Menü, das erscheint. Klicken Sie auf einen Platzhalter und beginnen Sie, um Text.	**0.509**
MML translation	Aus dem Hause Tab klicken Sie neue Folie, wählen Sie die gewünschte Folienlayout aus dem Menü, das erscheint. Klicken Sie auf eine beliebige Platzhalter und beginnen Sie zum Hinzufügen von Text.	**0.325**
Source	Click the Preview command on the Transitions tab.	
Reference	Klicken Sie auf den Befehl Vorschau auf der Registerkarte Übergänge.	
hATD translation	Klicken Sie auf die Vorschau Befehl auf die Übergänge Tab.	**0.260**
DSTF translation	Klicken Sie auf die Vorschau Befehl auf die Übergänge.	**0.280**
SEF translation	Klicken Sie auf die Vorschau Befehl auf die Übergänge Tab.	**0.260**
MML translation	Klicken Sie auf die Vorschau-Befehl auf die Übergänge Tab.	**0.237**

Table 6.23: Examples of segment-level BLEU scores for IT domain test set

6.3.3 Analysis of the Results

This section presents an analysis of the similarities and differences between the evaluation results for different empirical settings explored in this chapter: English→Spanish and Spanish→English translations for the Biomedical domain and English→German translations for the IT domain. The aim is to compare how the data selection methods behave on the three language pairs and on the two in-domains and to determine whether the ranking across all settings is stable. Also, the follow-up questions defined in the beginning of the chapter are answered as the analysis unfolds.

Evaluation metrics results per language pairs

Comparing the performance of the MT systems in terms of BLEU, METEOR and TER, per language pair that includes Spanish either as source or as target language, the BLEU scores are higher for Spanish→English systems translations than for the reverse language pair (about 6 points for the Khresmoi test set and about 2 points for WMT18). The same behavior is observed with the TER scores for these language pairs. These differences in BLEU and TER scores can be explained by the fact that translating into English is easier than translating into Spanish. English is not as morphologically rich as Spanish is. The opposite effect is encountered with METEOR which assigns much higher scores to English→Spanish translations than the reverse language pair (about 23 points for the Khresmoi test set and about 22 points for WMT18).

Investigating detailed alignments and scores from METEOR for translations from both test sets revealed that even though when evaluating English an extra module is used (Wordnet synonyms), the total number of modules usage is much lower for English than for Spanish. Particularly, the use of the paraphrase module reveals mostly the same number of times for both test sets. Instead, it is the stemming module that makes a difference between the two languages: for the Khresmoi test set, it was used for 966 Spanish words and for 299 English words. On the WMT18 test set, 297 words were stemmed by the scorer for Spanish and 92 for English. Therefore, it is not surprising that the METEOR scores significantly differ between the two language pairs. The difference between the test sets is explained by their size.

Examples of translations and scores with METEOR for both language pairs involving Spanish, using the hATD translation system, are given in Table 6.24 where exact word matches are not highlighted, yellow denotes stemming was applied, pink indicates that a paraphrase was used and green highlights synonyms for the English translations[21]. The sentences are given in the tokenized and lowercased form as used by the evaluation metric. The input is omitted for both language pairs as the reference for English→Spanish represents the input for Spanish→English and vice-versa. The dominance of the stemming module in the English→Spanish sentences (yellow) can be observed and how it affected the METEOR score compared to the Spanish→English case.

[21] See Chapter 3, Section 3.3.2 for more details on how METEOR produces scores

English→Spanish Reference	me parece que la única área de contención es si necesitas bloqueos de nervios periféricos o puedes usar inyecciones periarticulares .	
hATD translation	creo que la única área de contención es si necesita algún nervio periférico bloques o puede que utilice periarticular inyecciones .	**0.609**
Spanish→English Reference	i believe the only area of contention is whether you need peripheral nerve blocks or can you use periarticular injections .	
hATD translation	it seems that the only area is concerning the need of containment blocks peripheral nerves or you can use periarticular injections .	**0.354**
English→Spanish Reference	la formación de las células sanguíneas comienza con una célula especial localizada en la medula ósea llamada célula madre hematopoyética .	
hATD translation	la formación de células de la sangre comienza con un recuento de células especiales localizados en la médula ósea llamada un transplante de células madre hematopoyéticas .	**0.650**
Spanish→English Reference	blood cell formation begins with a special cell located in the bone marrow called a hematopoietic stem cell	
hATD translation	the formation of blood cells begins with a special cell located in the bone marrow called haematopoietic stem cell .	**0.512**

Table 6.24: Examples of segment-level METEOR alignments and scores for ES-EN and English→Spanish

When considering the English→German language pair, the closest language pair and test set from the Biomedical settings that could be considered for comparison is English→Spanish as for both settings the source language is English and both target languages Spanish as well as German have comparable inflection. The Khresmoi test set was chosen for comparison as the IT domain test set also consisted of 1000 sentence pairs. With respect to these two experimental settings, the results for BLEU differ by ≈ 1 points, for METEOR by ≈ 8 points, and for TER by ≈ 3 points. While BLEU shows stability across these two settings (translating a test set of size 1000 from English into a morphologically rich language), METEOR shows the opposite. This effect can be explained by the semantic nature of this metric where target language paraphrases and stemmers make it difficult to compare METEOR scores across different language pairs. Moreover, the in-domain also plays a role because the IT domain is highly technical and, therefore, the paraphrase module was not used as often in the METEOR computation of the scores. Count statistics of the stemming and paraphrase modules for the English→German hATD translation reveal that METEOR applied the stemming module 361 times and the paraphrase module 557 times, summing up to 918. In contrast, the English→Spanish hATD translation requested the stemming module 966 times and the paraphrase module 964 times, summing up to 1930. This is more than twice the module count of English→German (918 versus 1930). Since the modules were used much more frequently for the English→Spanish translations, the METEOR scores were higher than those for English→German (the aforementioned ≈ 8 points).

Out-of-vocabulary analysis per in-domains

Analyzing the unknown words lists for the Biomedical domain, five types of OOV words were identified:

- *NE - named entities* (for example, "marshfield")
- *LINK - hostnames/ domain names* (for example, "randyamy.com", "www.diabetesaustralia.com.au")
- *DT - domain terminology* (for example, procedures: "chromoendoscopy", "videofluoroscopy"; medicines: "ct327"; names of genes: "col9a2", "sec23b"; biology terms: "erythroblasts" ; abbreviations/ acronyms: "opmd", "nve", "s.o.b.")
- *TYPO - typos* (for example, "bestbets")
- *GEN - general domain words* (for example, "attender", "nonrandomized")

Count statistics for the identified OOV types are reported for the hybrid automatic ratio detection method and the strong baseline, for the English→Spanish Biomedical domain (the Khresmoi test set) and for the English→German IT domain. Both test sets have the same size (1000 sentences). The purpose is to compare a data selection method with the baseline (15% or 23% ratio selection versus full use of general domain data). Another aim is to determine the degree of difficulty of translating text pertaining to the two in-domains. Intuitively, given a test set from an in-domain, the ratio of unknown words that are of type *domain terminology* is a strong indicator of the difficulty of translating that test set.

Manual analysis of the two OOV lists (hATD and BS-strong) from the Biomedical domain revealed that out of the 126 unknown words for the hATD system translation, 85 entries pertained to domain terminology (31 terms were abbreviations or acronyms). On the other hand, out of the 102 unknown words for the strong baseline system translation, 81 entries belonged to domain terminology (16 terms were abbreviations or acronyms). The full distribution of OOV types is given in Table 6.25 where the percentages state how much of the total number of unknown words the given types count for (percentages were rounded up). Where a type could not be identified, the unknown word was tagged with *other* (for example, "fig3a"). Inspecting the relative frequencies of unknown words, hATD performs worse than the baseline at translating terminology (67% versus 79%) but has a better coverage of named entities (20% versus 9%). The relative frequencies of the other OOV types are very similar for both systems.

	Biomedical		IT	
OOV type	**hATD**	**BS-strong**	**hATD**	**BS-strong**
DT	85 (67%)	81 (79%)	1 (1%)	1 (1%)
NE	25 (20%)	9 (9%)	14 (14%)	12 (15%)
GEN	9 (7%)	6 (6%)	6 (6%)	1 (1%)
LINK	2 (2%)	2 (2%)	29 (29%)	24 (31%)
TYPO	2 (2%)	4 (4%)	41 (41%)	34 (44%)
other	3 (2%)	2 (2%)	9 (9%)	6 (8%)
Total OOV	126	102	90	78

Table 6.25: Count statistics of OOV types for the English→Spanish Biomedical test set and for the English→German IT test set

The same procedure was applied to the IT domain where manual analysis of the unknown words list using the same five types of OOV as for the Biomedical domain. The count statistics of the OOV types are given in Table 6.25. As opposed to the Biomedical domain, the terminology (DT) is almost fully translated by both systems. For the general domain (GEN), the baseline has a higher coverage than hATD (1% versus 6%).

Vocabulary coverage for the Biomedical test set is lower than for the IT set. The systems perform much better at translating domain-specific terms (DT) for the IT domain than for the Biomedical domain (1% versus 79% for the baselines). This is a strong indicator that the Biomedical domain is a more difficult domain for translation. Not surprisingly, the amount of OOV type LINK is much higher in the IT domain (31% versus 2% for the baselines). Also, the amount of typos (TYPO) in the IT test set is much larger than in the Biomedical set, with 44% unknown words entries out of the full OOV list accounting for typing errors in the IT test set (for the baseline) and only 4% in the Biomedical set.

Performance comparison of data selection methods

In order to answer the research question **RQ5** stated in the Introduction, an analysis of the system ranking follows.

The ranking based on BLEU scores for the IT domain positions hATD on the first place, followed by SEF, MML, DSTF (23% selection for all of them) and the strong baseline on the last place. Also for English→Spanish in the Biomedical domain, the Khresmoi test set reveals a similar ranking: hATD, followed by SEF, DSTF, MML and the baseline. On the other test set, WMT 2018, SEF outperforms hATD, followed by DSTF, MML and the strong baseline. The direction Spanish→English, reveals the same rankings for the data selection methods as in the IT domain, for the Khresmoi test set: hATD, followed by SEF, MML and DSTF (15% selection for all of them). However, the strong baseline performs slightly worse than hATD, thus its rank is two. A different ranking order is produced by the other test set where DSTF achieves first place, followed by hATD, the strong baseline, MML and finally, SEF.

These five different rankings based on BLEU scores make it impossible to clearly state that one data selection method is better than another one given any language pair, or any in-domain. A generalization of ranking of the systems cannot be applied which was also the case in all the WMT system rankings from all years where the participating systems produced different ranking depending on the language pair[22]. However, in most of the cases, hATD or SEF outperformed the other methods and the strong baseline never ranked first.

Bootstrap resampling did not show any difference in the BLEU scores between the systems for Biomedical Spanish→English, for both test sets. Some statistical significant results can be observed on the Biomedical test sets for the English→Spanish language pair. SEF outperforms DSTF and both SEF and hATD outperform MML on the Khresmoi test set. On the WMT18 test set, SEF also outperforms MML. On the IT domain, there are also some statistical significant results as hATD outperforms both DSTF and MML, while SEF performs better than DSTF. On the intersection of all the comparisons that yield significance, there are two results that are found twice: SEF is better than DSTF and hATD better than MML.

Therefore, full ranking generalization does not apply for the bootstrap resampling results in the three investigated experimental settings. However, partial ranking generalization can be achieved across three test sets. Some results are statistical significant and indicate that the automatic ratio detection method outperforms the state-of-the-art method, and that one of the methods that uses PV for text representation is better than the method that uses TF.

6.4 Summary

Automatic ratio detection was tackled in this chapter where a data selection method based on an MLP classifier was introduced (iATD). Given sentences represented by means of PV, the algorithm learned from positive training samples selected from the in-domain corpus and negative samples randomly selected from the selection pool, whether a general domain sentence should be labeled as pseudo in-domain or not. However, the negative training samples could contain pseudo-in domain sentences that harm the prediction accuracy of the model by introducing false negative samples in the training data. An improvement to iATD, hATD, considers using another data selection method for scoring all the general domain sentences and selecting the sentences with the worst scores as being false training samples for the classifier. This approach was compared with the initial one and the results indicated that hATD outperforms iATD.

To answer research questions RQ4 and RQ5 in a conclusive manner, a consolidated evaluation became necessary as before all data selection methods were introduced and evaluated on different in-domains and for different language pairs. All

[22] For example, see Table 11 from Barrault et al. (2019, page 24)

methods were applied within three common experimental settings that consisted of training SMT systems on three language pairs pertaining to two in-domains. While the BLEU evaluation scores provides a system ranking, not all of the system comparisons turned out to be statistically significant. Moreover, the system rankings did not fully generalize across all experimental settings. However, a major result is that hATD generally outperforms other data selection methods.

Chapter 7

Manual Evaluation of Data Selection Methods

7.1 Introduction

This chapter presents an analysis of human evaluation comprising of a three-way ranking procedure applied to the Spanish→English system output and an error analysis of the MT output. The WMT Biomedical campaigns also use the three-way ranking procedure for human evaluation (Bojar et al., 2016a; Jimeno Yepes et al., 2017; Neves et al., 2018). The research question **RQ6** that concerns the comparison between manual and automatic evaluation results is answered in this chapter.

A total of six system comparisons were evaluated using the three-way ranking procedure: DSTF versus SEF, hATD versus SEF, DSTF versus hATD, MML versus DSTF, MML versus SEF and MML versus hATD. Moreover, for the three-way ranking evaluation, the intra- and inter-annotator agreement was calculated in order to attest the reliability of the human judges. Error analysis was also conducted with the scope of identifying and analyzing the types of errors each system produced.

The Appraise evaluation system (Federmann, 2012) was used with a small improvement to the error analysis task where I considered necessary an additional field where a human annotator can include justifications. Two bilingual human annotators, native Spanish speakers, with background in linguistics and translation, were chosen and given two user accounts on the Appraise platform. The annotators were paid for their work. After reading an instruction material that I created, the annotators were asked to perform a training for the three-way ranking task in order to get familiarized with the procedure.

When several MT systems produced the same translation for a given input, the duplicates were eliminated from the three-way ranking task, considering the systems to perform on a par for those sentences. This step reduced significantly the work load with a mean of 34% considering all tasks.

Due to time and cost considerations, human evaluation was applied only on the

Spanish → English language pair.

In the following, the three-way ranking procedure is explained with examples. The results of the two human annotators and the ranking of the systems are also given. Afterwards, the annotator agreement is conducted, and finally, and error analysis is presented that gives on overview of the types of errors that were most commonly found in the output of the MT systems.

7.2 Three-way Ranking

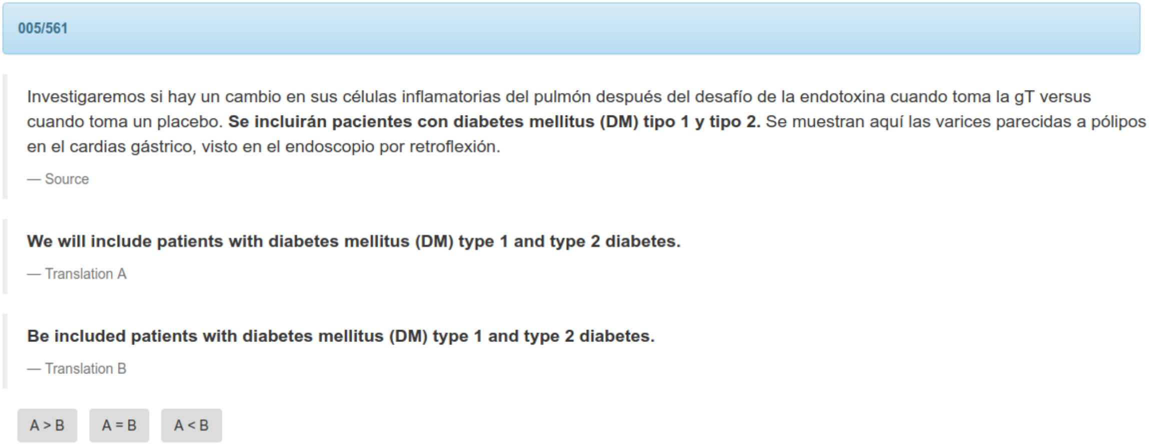
This section presents the three-way ranking procedure for the Spanish→English language pair. As the out-of-vocabulary analysis from Section 6.3.3 revealed, the Biomedical domain is harder to translate than the IT domain, due to the complex terminology. Therefore, the three-way ranking evaluation was applied to the Biomedical domain. I chose the Khresmoi test set because it contains 1000 sentence pairs and it is well aligned (in contrast to the 275 sentences from the WMT18 set which contains alignments between empty lines and sentences).

I opted not to reveal the reference to the evaluators in order to not bias them. When judging the quality of the translations, the annotators were asked to consider the amount of errors a translation contains, how well the meaning of the source is preserved by the translation, the amount of missing words and misspellings, the word order, whether poor lexical choices were produced, the fluency of the English translations, whether extra words were inserted, the morphology errors and the punctuation errors.

The annotators were given an instruction material that consisted of ranking examples that I produced. Some of them are presented below. The sample sentences were also extracted from the Khresmoi test set (source side) and the explanations offer also the Khresmoi reference. The reference offered in the explanation is mentioned as being ”one possible correct translation” due to the complex nature of human translation where one sentence can have many possible translations (see Example 3.5 from Section 3.3.1). The examples were given in order to help and guide the annotators in their initial phase. After being familiarized with the procedure, the annotators were encouraged to construct their own judgments and ranking schemes in order to make the translations comparisons.

In the first example, both translations contain an extra word inserted at the end of the sentence: diabetes. However, translation **A** is better than **B** because it is fluent in English and it uses the correct verb tense (we will include versus be included). One possible correct translation for the sentence is: *Patients with type 1 and type 2 diabetes mellitus (DM) will be included.* This is an example where it is evident that translation **A** is better than translation **B**.

In the second example, **A** and **B** preserve some meaning of the Spanish sentence, however the same errors are present in both: wrong word order, wrong translation of cardias gástrico and wrong lexical choice for retroflexión (**A** simply uses the Spanish word, not being able to translate it, and **B** uses a wrong translation; the correct translation is retroflexion). One possible correct translation for the sentence



Investigaremos si hay un cambio en sus células inflamatorias del pulmón después del desafío de la endotoxina cuando toma la gT versus cuando toma un placebo. **Se incluirán pacientes con diabetes mellitus (DM) tipo 1 y tipo 2.** Se muestran aquí las varices parecidas a pólipos en el cardias gástrico, visto en el endoscopio por retroflexión.

— Source

We will include patients with diabetes mellitus (DM) type 1 and type 2 diabetes.

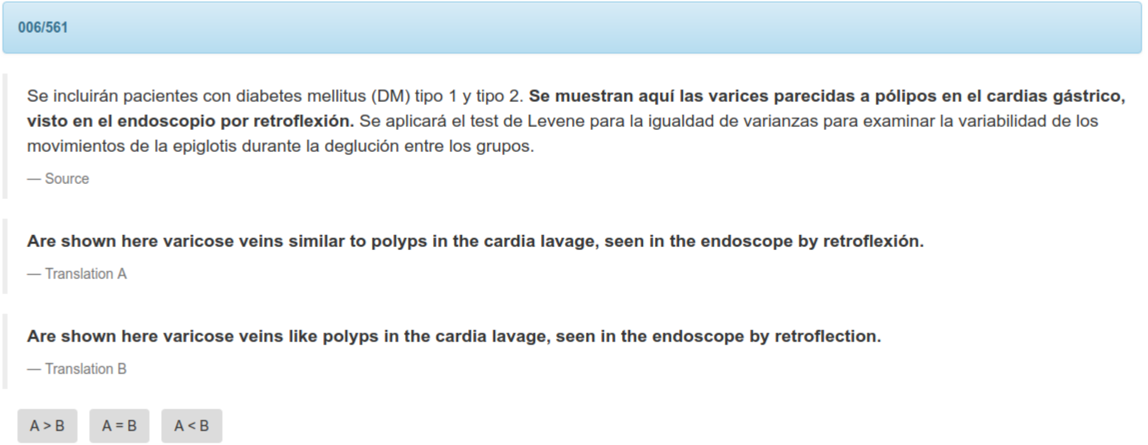
— Translation A

Be included patients with diabetes mellitus (DM) type 1 and type 2 diabetes.

— Translation B

Figure 7.1: Example 1 for the three-way ranking procedure

is: *Polyp-like varices are shown here in the gastric cardia, seen on retroflexion of the endoscope*. Judging by the fact that the translations are similar and contain the same type and number of errors it can be concluded that **A** = **B**. Judging by the fact that **A** failed to provide a translation for retroflexión, while **B** translated it using a wrong English term resulting in a more fluent sentence, it can be concluded that **A** < **B**. This is a difficult case and it is up to the human annotator to make a decision.



Se incluirán pacientes con diabetes mellitus (DM) tipo 1 y tipo 2. **Se muestran aquí las varices parecidas a pólipos en el cardias gástrico, visto en el endoscopio por retroflexión.** Se aplicará el test de Levene para la igualdad de varianzas para examinar la variabilidad de los movimientos de la epiglotis durante la deglución entre los grupos.

— Source

Are shown here varicose veins similar to polyps in the cardia lavage, seen in the endoscope by retroflexión.

— Translation A

Are shown here varicose veins like polyps in the cardia lavage, seen in the endoscope by retroflection.

— Translation B

Figure 7.2: Example 2 for the three-way ranking procedure

Two systems rankings based on the three-way ranking results are presented: the ranking procedure that the WMT Biomedical task employed in all years and the ranking that the WMT News task used from 2014 until 2017. The advantages and disadvantages of each method will be discussed.

The WMT Biomedical task follows the same manual evaluation procedure every year: a sample of 100 sentences is randomly extracted from the test set and the systems translations, the three-way ranking procedure is applied using one annotator[1], and finally the systems ranking is produced by applying the max function between every two systems comparisons. In the end, an overall ranking is produced by combining all the results[2].

For my experiments, I made two changes to the approach: using two annotators instead of one and using the full test set instead of a 10% sample. Given the Khresmoi test set of 1000 sentences, there are 2000 comparisons for system A with system B, in contrast with the sample of 100. The goal of these two changes was to obtain a more valid ranking since it considers the full test set, thus eliminates the randomness factor which could favor one system over the other (when randomly sampling 10% of the test set). Moreover, by using two annotators, the ranking for my systems is double-checked. Applying the WMT Biomedical methodology for my human comparisons results, the following conclusions can be drawn:

MML > *DSTF*, *SEF* > *MML*, *hATD* > *MML*, *SEF* > *DSTF*, *hATD* > *SEF* and *hATD* > *DSTF*.

Which leads to the following ranking:

hATD > SEF > MML > DSTF

Answering **RQ6**, this ranking is in line with the ranking obtained via BLEU for the Spanish→English translations of Biomedical data (the Khresmoi test set), and also with the ranking obtained for the IT domain for English→German. The other three test sets (the Spanish→English WMT 2018 Biomedical and the two English→Spanish Biomedical sets) produce a different system ranking. However, hATD is never placed last.

While this ranking approach is easy to apply and provides an intuitive, fast ranking of the systems, it has two important disadvantages:

- it does not handle the cases where two systems have relatively close counts
- it does not handle the *infinite test set* scenario: how does the ranking change in case more sentences are added to the test set? As Sakaguchi et al. (2014) pointed out, there is a possibility that a system is lucky in the number of comparisons that favor it and thus, generate the conclusion that it outperforms the system it is being compared to.

In order to tackle these limitations, I employed the TrueSkill[3] algorithm (Herbrich et al., 2007; Sakaguchi et al., 2014). This is the official method used in comparing systems used by the WMT organizers for ranking the systems participating in the MT evaluation News campaigns (Bojar et al., 2016a, 2017). As in

[1] One exception: for the 2018 manual evaluation, the WMT Biomedical task used two annotators only for the language pair English→Romanian.

[2] For example, if $A > B$ and $B > C$, then the ranking is $A > B > C$

[3] I used the TrueSkill version for WMT available at `https://github.com/keisks/wmt-trueskill`.

the WMT campaign, I ran the TrueSkill algorithm for 1000 bootstrap resampled (with replacement[4]) datasets over all the data.

As follows, I use Sakaguchi et al. (2014) in explaining the TrueSkill algorithm. For each system S_j, the approach assumes that its skill level follows a normal distribution where the mean μ_{S_j} reflects the current estimate of the system's ability and the variance $\sigma^2_{S_j}$ represents the algorithm's uncertainty about its estimate of each mean. The initial value for μ is 0 and for σ is 0.5.

Given an outcome (S_1, S_2, rel) where S_1 and S_2 represent the two compared systems and rel is $<$ (win), $>$ (loss) or $=$ (draw), **the update equations for the systems means** are defined as Sakaguchi et al. (2014):

$$\mu_{S_1} = \mu_{S_1} + \frac{\sigma^2_{S_1}}{c} \cdot v\left(\frac{t}{c}, \frac{\epsilon}{c}\right)$$
$$\mu_{S_2} = \mu_{S_2} - \frac{\sigma^2_{S_2}}{c} \cdot v\left(\frac{t}{c}, \frac{\epsilon}{c}\right)$$

In the formulas, c denotes the confidence of the algorithm and is calculated as $c^2 = 2\beta + \sigma^2_{S_1} + \sigma^2_{S_2}$, where β is a free parameter ($\beta = 0.025 \cdot J \cdot \sigma^2$ with J quantifying as the total number of human judgments). The v function captures how surprising the outcome was: if a translation obtained with S_1 has a high μ_{S_1} and was judged as being better than the translation obtained using S_2 which has a much lower μ_{S_2}, then it is not surprising (low update for both systems means). However, if the result is unexpected, then the updates for the means will be larger. Given t as the difference in means and ϵ a fixed parameter, the v function is calculated using the normal distribution, $\mathcal{N}$, and the cumulative distribution function, Φ Sakaguchi et al. (2014):

$$v_{win}(t, \epsilon) = \frac{\mathcal{N}(-\epsilon + t)}{\Phi(-\epsilon + t)}$$
$$v_{tie}(t, \epsilon) = \frac{\mathcal{N}(-\epsilon - t) - \mathcal{N}(\epsilon - t)}{\Phi(\epsilon - t) - \Phi(-\epsilon - t)}$$

In addition to updating the systems abilities, **the update equations for the systems confidences** are defined below Sakaguchi et al. (2014):

$$\sigma^2_{S_1} = \sigma^2_{S_1} \cdot \left[1 - \frac{\sigma^2_{S_1}}{c^2} \cdot w\left(\frac{t}{c}, \frac{\epsilon}{c}\right)\right]$$
$$\sigma^2_{S_2} = \sigma^2_{S_2} \cdot \left[1 - \frac{\sigma^2_{S_2}}{c^2} \cdot w\left(\frac{t}{c}, \frac{\epsilon}{c}\right)\right]$$

The role of the function w, defined below, is to express how surprising the outcome is by using the previously defined function v Sakaguchi et al. (2014):

$$w_{win}(t, \epsilon) = v_{win} \cdot (v_{win} + t - \epsilon)$$

[4] Given N outcomes, the algorithm randomly samples N outcomes where an outcome can occur 0, 1 or multiple times.

$$w_{tie}(t,\epsilon) = v_{tie} + \frac{(\epsilon - t)\cdot\mathcal{N}(\epsilon - t) + (\epsilon + t)\cdot\mathcal{N}(\epsilon + t)}{\Phi(\epsilon - t) - \Phi(-\epsilon - t)}$$

After running the TrueSkill algorithm over each of the 1000 samples, a rank range is computed for each system in each iteration by removing the top and bottom 2.5% (confidence level of 95%) and clustering the systems into equivalence classes. The system scores are represented by their system means Sakaguchi et al. (2014).

Table 7.1 presents the ranking results obtained using TrueSkill (system score along with ranking range). The systems in the same range are considered tied, therefore all of the data selection methods perform on a par for the Khresmoi test set (the state-of-the-art method performing slightly worse). With respect to answering **RQ6**, this result is consistent with the BLEU results for this particular test set.

System	**Score**	**Rank range**
hATD	0.156	(1.0, 4.0)
DSTF	0.067	(1.0, 4.0)
SEF	0.014	(1.0, 4.0)
MML	-0.237	(2.0, 4.0)

Table 7.1: Ranking produced with TrueSkill using the human judgments

7.3 Annotator Agreement

Several measures to determine the reliability of human judgments have been proposed in the literature. As defined in Hollnagel (1993, page. 16), reliability is "the probability that a person will perform according to the requirements of the task for a specified period of time". Taking into account a series of observations which contain repeated measurements, reliability was investigated through the agreement between the two human judges annotating the same translation pairs (inter-agreement) and the agreement for each judge with himself annotating the same translation pairs twice (intra-agreement). As pointed out in Krippendorff (2004, chapter. 11), reproducibility is "the degree to which a process can be replicated by different analysts working under varying conditions, at different locations" and is measured through the inter-agreement, while stability is "the degree to which a process is unchanging over time" and is reflected through the intra-agreement. Krippendorff presents accuracy, defined as "the degree to which a process conforms to its specifications and yields what is is designed to yield", as the strongest type of reliability. However, since no gold-standard exists that could certainly state which translation is better for each of the translation pairs (observations), accuracy could not be evaluated for the three-way ranking tasks.

The reliability of the human annotators was measured using κ, the Cohen's Kappa coefficient (Cohen, 1960). The formula for calculating the κ score was presented in Section 3.3.1. As in Machávcek and Bojar (2015), 5% of each three-way-ranking task was randomly sampled and used for the intra- and inter-annotator agreements. The random sampling was performed on the duplicate-free set of sentence pairs, as duplicate sentence pairs were also not included in the full three-way ranking tasks (see Section 7.2).

For computing the κ score I used the script made available by the WMT organizers[5]. The average per tasks intra-annotator κ was **0.608** and the inter-annotator κ was **0.470**. The interpretation according to Landis and Koch (1977) and also used by the WMT campaigns. is that a κ score between 0 and 0.20 means slight agreement, 0.21 − 0.40 is fair, 0.41 − 0.60 is moderate, 0.61 − 0.80 is substantial and 0.81 − 1.00 is almost perfect agreement. Thus, both results reflect moderate agreement. They are also consistent with the κ results reported by WMT. Therefore, both manual and automatic evaluation present the same difficulty in assessing the quality of an MT translation. Intuitively, a degree of consistence between measures performed by the same person twice is higher than the degree of consistence between measures performed by two persons. Similar results have also been reported in the findings of the WMT 2016 (see Tables 4 and 5 from Bojar et al. (2016a)) where the agreement scores are almost in the same range as the ones obtained in this thesis.

7.4 Error Analysis

This section presents the types of errors encountered in the MT systems outputs. For the language pair in the primary focus, Spanish→English, the same two native Spanish speakers that ranked the systems in the previous section were asked to classify and annotate the errors from 10% of the MT outputs for the Khresmoi test set for each system.

As follows, the classification of errors is presented, as taken from the Appraise tool, along with the examples I prepared for the two annotators. The results for the language pair in the primary focus are discussed.

For every sentence presented for inspection, each word can be annotated with one or more of the following errors (according to the Appraise tool):

- Terminology
- Lexical choice
- Syntax (ordering)
- Insertion (extra word)
- Morphology

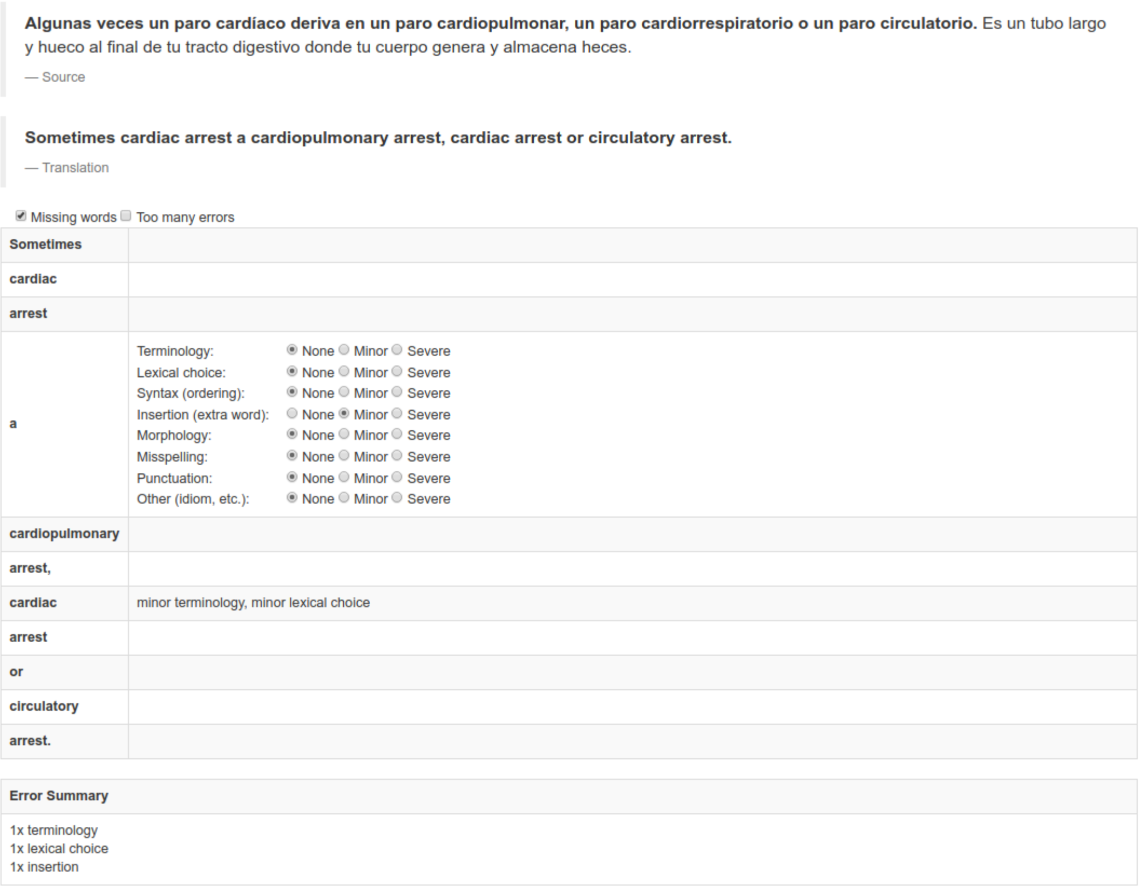
[5] https://github.com/cfedermann/wmt16/blob/master/scripts/

- Misspelling
- Punctuation
- Other

For every error, the user can mark it either as minor or severe. The instruction offered to support the evaluators in taking these decisions:

- minor: it affects in a way the meaning/ fluency of the sentence
- severe: it highly affects the meaning/ fluency of the sentence.

In addition to individual errors, the user can mark the translation as having missing words and/ or having too many errors (according to the Appraise tool). The annotators were encouraged to use any online and offline lexical resources for the medical terms they were not familiar with. Two examples that I annotated are presented below which were included in the instruction material.



> **Algunas veces un paro cardíaco deriva en un paro cardiopulmonar, un paro cardiorrespiratorio o un paro circulatorio.** Es un tubo largo y hueco al final de tu tracto digestivo donde tu cuerpo genera y almacena heces.
>
> — Source

> **Sometimes cardiac arrest a cardiopulmonary arrest, cardiac arrest or circulatory arrest.**
>
> — Translation

☑ Missing words □ Too many errors

Sometimes	
cardiac	
arrest	
a	Terminology: ● None ○ Minor ○ Severe Lexical choice: ● None ○ Minor ○ Severe Syntax (ordering): ● None ○ Minor ○ Severe Insertion (extra word): ○ None ● Minor ○ Severe Morphology: ● None ○ Minor ○ Severe Misspelling: ● None ○ Minor ○ Severe Punctuation: ● None ○ Minor ○ Severe Other (idiom, etc.): ● None ○ Minor ○ Severe
cardiopulmonary	
arrest,	
cardiac	minor terminology, minor lexical choice
arrest	
or	
circulatory	
arrest.	

Error Summary
1x terminology 1x lexical choice 1x insertion

Figure 7.3: Example 1 for the error classification procedure

In the first example, the translation contains missing words, thus the checkbox for ”missing words” was checked. Having in mind one possible translation (the reference from the test set) such as "Cardiac arrests are sometimes referred to as cardiopulmonary arrest, cardiorespiratory arrest, or circulatory arrest", the indefinite article "a" does not represent a translation of any of the Spanish words. Thus, it was marked in the list at "insertion (extra word)". However, the annotator might have a different translation in mind when reading the Spanish sentence which might include the indefinite article. Therefore, when classifying errors it is important that the annotator translates the sentence in their mind. The word "cardiac" was marked with "minor terminology" and "minor lexical choice" because the correct translation would have been "cardiorespiratory". At the end of the list there is an "Error Summary" which updates automatically as the annotator makes changes to the error list. In this example, the MT system output was annotated with three types of errors which appear once and with missing words.

In the second example, three types of errors can be identified: missing words ("It", the translation of "Es" as in "It is a long ..."), syntax ("and" is placed in a wrong order after "tube" instead of before it) and morphology (wrong verb tense for "almacena": "stored" instead of "stores").

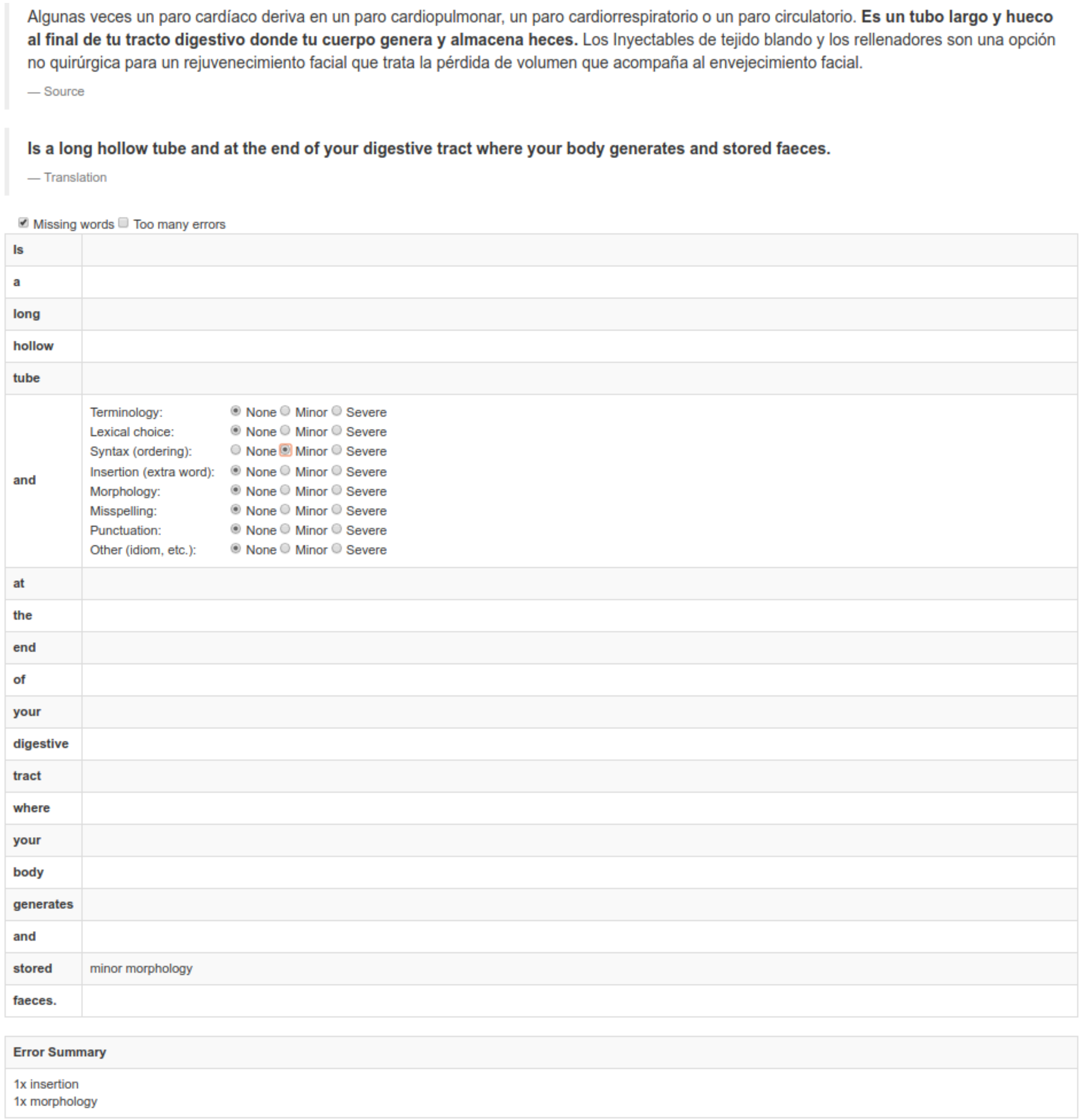
The same two native Spanish speakers that ranked the systems annotated 150 sentences from the output of all the MT systems. One annotator marked errors for 50 sentences, while the other one (a translation specialist) marked errors for 100 sentences (the same 50 sentences as the other annotator and 50 additionally ones). The motivation for overlapping 50 sentences was to observe the inter-annotator agreement on the types of errors. Since the task of annotating sentences is time consuming and also costly, only the translation specialist was asked to annotate more sentences. The same MT output sentences were investigated for all systems in order for the results to be comparable.

For each MT system, counts of each type of errors are depicted in Figure 7.5 where each bar is associated with one system. From the cumulative counts of errors it can be observed that the most frequent type of error is *syntax*, followed by wrong *lexical choice* and then *missing words*. According to this error classification, the SEF system produced the most syntax errors, with the other three systems performing almost on a par.

The SEF system exhibits the least number of missing words, almost on a par with hATD. The hATD system not only produced the smallest number of wrong lexical choice errors, but also came up with the smallest number of other error types, like insertion of words that should not be in the translation, morphology errors and terminology problems. The other types of errors (punctuation, misspelling and other) have only a small number of occurrences in the translations.

The punctuation errors are mainly wrong word capitalizations that are produced by the recaser. Since the recaser is not part of the data selection pipeline, the errors marked as punctuation do not offer much insight into the performance of one system compared to another (in terms of data selection). The purpose of the recaser is to simply transform the lowercased machine translation output into a recased sentence that is easier for humans to read.

Examples of error classification along with the justification that the annotators provided are given in Table 7.2. The judgments attributed by the first annotator



> Algunas veces un paro cardíaco deriva en un paro cardiopulmonar, un paro cardiorrespiratorio o un paro circulatorio. **Es un tubo largo y hueco al final de tu tracto digestivo donde tu cuerpo genera y almacena heces.** Los Inyectables de tejido blando y los rellenadores son una opción no quirúrgica para un rejuvenecimiento facial que trata la pérdida de volumen que acompaña al envejecimiento facial.
>
> — Source

> **Is a long hollow tube and at the end of your digestive tract where your body generates and stored faeces.**
>
> — Translation

☑ Missing words ☐ Too many errors

Is	
a	
long	
hollow	
tube	
and	Terminology: ● None ○ Minor ○ Severe Lexical choice: ● None ○ Minor ○ Severe Syntax (ordering): ○ None ● Minor ○ Severe Insertion (extra word): ● None ○ Minor ○ Severe Morphology: ● None ○ Minor ○ Severe Misspelling: ● None ○ Minor ○ Severe Punctuation: ● None ○ Minor ○ Severe Other (idiom, etc.): ● None ○ Minor ○ Severe
at	
the	
end	
of	
your	
digestive	
tract	
where	
your	
body	
generates	
and	
stored	minor morphology
faeces.	

Error Summary
1x insertion 1x morphology

Figure 7.4: Example 2 for the error classification procedure

are visually enhanced with the blue color, while for the second annotator the orange color is used.

The translation produced by the hybrid automatic ratio detection system, hATD, wrongly attributes the adjective unknown to ion channels. The first annotator identified this error and marked it as being severe as it changes the meaning of the sentence. While the other annotator did not catch this error, both of them correctly marked the word desalineado as being a wrong lexical choice. The first annotator penalizes the error with the attribute severe consequently on all the MT system translations which fail to translate the Spanish word. There is also one exception when for the DSTF translation, the same annotator does not mark the Spanish word. An agreement is observed across both users for the error produced by the indefinite article a followed by the adjective unknown where both of them mark the error as a minor morphological one.

A case of consistent error treatment is observed also for the second annotator which marks the word desalineado as being a minor wrong lexical choice and also a minor terminology error. This user also makes small exceptions for hATD where it only marks it as being a wrong lexical choice and for MML where, in addition, the word is marked as a syntax error too. Given the Spanish word desalineado, which was not translated by any of the MT systems, it can be observed that the notion of domain-specific terminology is ambiguous to humans as for one annotator it did not trigger a flag that this word pertains to the Biomedical domain, but it did trigger the notion of domain-specific vocabulary for the other one. Intuitively, if the word is correctly translated into misaligned, in the context of the sentence, it could be regarded as medical terminology.

Intuitively, the best error classification annotators for the sentences pertaining to this domain are native Spanish speakers, that are specialized in linguistics and that also have a background in Biomedicine. However, such expertise is difficult to find and most probably, also very costly and comes with the disadvantages discussed in Chapter 3, Section 3.3.1.

This example demonstrates the difficulty of maintaining consistency in judgments for error classification. Both annotators were trained linguists, native Spanish speakers, however their annotations differ in some places. The task of identifying errors and grading them according to their influence on the whole translation is not only laborious, but also subjective.

In addition to subjective differences between annotators, intra-annotator disagreement poses another challenge for manual annotation. Even though the same translation was produced with two different MT systems, the same annotator marked the word submerged with the error *lexical choice* for the hATD system, while the same word was marked with the error *terminology* for the SEF and MML systems (see Table 7.3). The translation produced with the DSTF system differs in two places (insertion of the definite article the and the lexical choice of using the synonym extirpated of excised). For this sentence, the same annotator marked again the word submerged as an error, but with both lexical_choice and terminology. There is however an agreement in the severity of these errors as they were all marked with minor. Interestingly, the other annotator also annotated the word submerged as an error, but for all systems consistency was preserved (only lexical_choice). However, the other annotator was inconsistent in the severity of the error, marking it sometimes with minor and other times with severe. From my point of view, the word submerged is both a wrong lexical choice and a terminology error because the correct English translation of descalcificado is decalcified, which is a term belonging to the medical domain, as the Merriam-Webster online dictionary explains decalcification as *the removal or loss of calcium or calcium compounds (as from bones or soil)*[6]. Also, using instead the word submerges is indeed a wrong lexical choice.

[6] https://www.merriam-webster.com/dictionary/decalcification

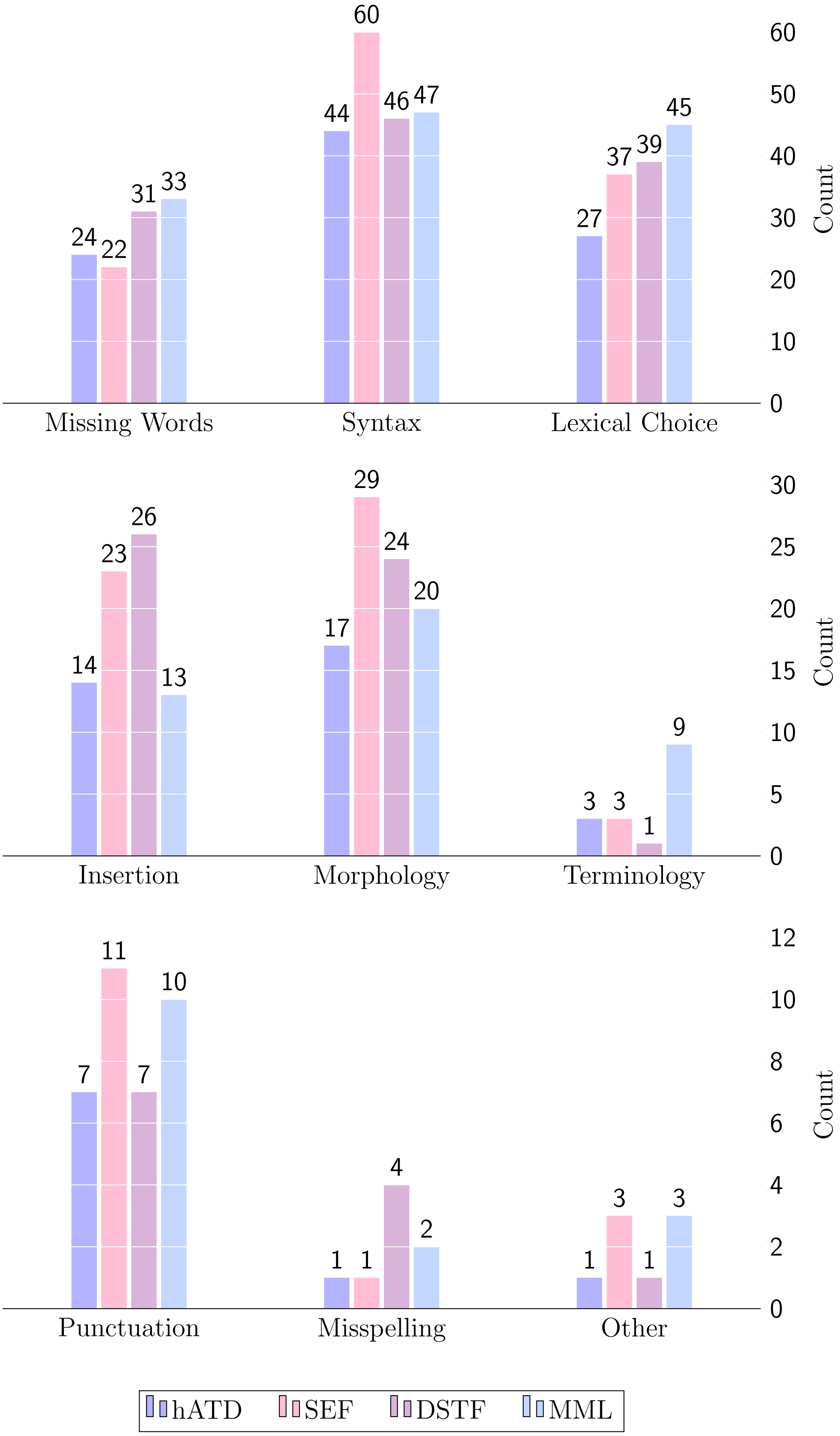


Figure 7.5: Cumulative Counts of Errors per Type

Source

Una nueva investigación reveló una conexión previa desconocida entre dos canales de iones, lo cual, cuando desalineado, puede causar los más extraños síntomas de la condición.

Reference

New research revealed a previously unknown connection between two ion channels, which, when misaligned, can cause the many bizarre symptoms that characterize the condition.

hATD Translation

A new research revealed a previous connection between two unknown ion channels, which, when desalineado, may cause the strangest symptoms of the condition.

1 → unknown=**syntax**:SEVERE; desalineado=**lexical_choice**:SEVERE; Justification: 'unknown' is an adjective to 'previous connection';'desalineado' is in Spanish.

2 → desalineado=**lexical_choice**:MINOR, **misspelling**:MINOR; Justification: 'misaligned' instead of 'desalineado'

SEF Translation

New research revealed a unknown prior connection between two channels, which, when desalineado, may cause the most strange symptoms of the condition.

1 → MISSING_WORDS a=**morphology**:MINOR desalineado=**lexical_choice** :SEVERE; Justification: it should be 'an' since the next word starts with a vowel; 'of ions' is missing; it did not translate a word.

2 → a=**morphology**:MINOR desalineado=**lexical_choice**:MINOR, **terminology**:MINOR Justification: none

DSTF Translation

New research revealed a connection between two unknown prior ion channels, which, when desalineado, may cause the strangest symptoms of the condition.

1 → prior=**lexical_choice**:MINOR, **syntax**:SEVERE unknown=**syntax**:SEVERE Justification: New research revealed a previously unknown connection between two ion channels.

2 → **MISSING_WORDS** prior=**syntax**:MINOR desalineado=**lexical_choice**: MINOR, **terminology**:MINOR unknown=**syntax**:MINOR Justification: before 'new' the word 'a' is missing; 'unknown prior' should go before 'connection'; 'misaligned' instead of 'desalineado'

MML Translation

New research revealed a unknown prior connection between two ion channels, which, when desalineado, may cause the most bizarre symptoms of the condition.

1 → a=**morphology**:MINOR prior=**other**:MINOR desalineado=**lexical_choi-ce**:SEVERE

2 → a=**morphology**:MINOR desalineado=**lexical_choice**:MINOR, **terminology**:MINOR, **syntax**:MINOR; Justification: 'an' instead of 'a'; 'misaligned' instead of 'desalineado'

Table 7.2: Examples of annotations from the two annotators for all the MT systems output

Source

Las válvulas aórticas son extirpadas y el anillo de la aorta (bases sobre las cuales las válvulas encajan) es limpiado y descalcificado si es necesario.

Reference

Aortic valves are excised and the ring of the aorta (basis on which the valves fit) is cleaned and decalcified if necessary.

hATD Translation

Aortic valves are excised and the ring of the aorta (basis on which the valves fit) is cleaned and submerged if necessary. → submerged=**lexical_choice**:MINOR

SEF Translation

Aortic valves are excised and the ring of the aorta (basis on which the valves fit) is cleaned and submerged if necessary. → submerged=**terminology**:MINOR

DSTF Translation

The aortic valves are extirpated and the ring of the aorta (basis on which the valves fit) is cleaned and submerged if necessary. → submerged=**lexical_choice**:MINOR, **terminology**:MINOR

MML Translation

Aortic valves are excised and the ring of the aorta (basis on which the valves fit) is cleaned and submerged if necessary. → submerged=**terminology**:MINOR

Table 7.3: Examples of annotations from the same annotator for all the MT systems output

The inconsistencies at an intra-annotator level show how difficult it is even for a trained human translator to agree with himself when judging the same sentences at different points in time.

The purpose of this error analysis is not to rank systems by counts of errors, but to gain an insight into the types of errors produced by the MT system. For example, given Figure 7.5 with the cumulative error types for all systems, the DSTF and MML systems commit the most missing words errors. The data selection methods used by these two systems perform worse than the other two methods (hATD and SEF) because the pseudo in-domain sentences selected with DSTF and MML have a lower lexical coverage over the Biomedical domain. Investigating the number of wrong lexical choices for each system, the difference of almost 20 errors between hATD and MML indicates that hATD selected pseudo in-domain sentences that are closer to the Biomedical domain than MML.

7.5 Summary

This chapter brought together all the data selection methods presented in the previous chapters through manual evaluation.

Human evaluation was carried out for one language pair and the manual results lead to a similar ranking as the one produced with the automatic evaluation results. The best performing data selection method among the three that I developed as well as the state-of-the-art approach is the hybrid automatic ratio detection method. It constitutes an essential result towards automatically obtaining pseudo in-domain data with considerable advantages: it is fast to train (due to a small corpus size and no need to tune the ratio of selected sentences). At the same time it produces high-quality translations, for several language pairs and in-domains. A comparison between different annotation paradigms showed a high degree of agreement between system rankings produced via automatic and manual procedures.

Chapter 8

Conclusions

This chapter summarizes the data selection methods that I developed and presents an overview of the experimental results with a focus on the system ranking obtained using automatic and manual evaluation methods. Finally, ideas for future work are presented.

8.1 Summary of Data Selection Methods

An initial step in developing a standard data selection method is text representation. In this thesis two approaches were investigated: Paragraph Vector (PV) and Term Frequency (TF). After this initial step of representing the sentences, the standard data selection pipeline consists of scoring the sentences from the selection pool according to their similarity to an in-domain, sorting them by their scores and selecting the top most similar sentences to build the pseudo in-domain which is later used together with the in-domain for training MT systems.

Given a sentence from the general domain, s_{Gen}, scoring produces a list of sentences most similar to s_{Gen}. The similarity is calculated through the use of the cosine between two vectors. Each sentence from the list of the most similar ones is weighed according to its rank in the list, which assigns higher importance to the sentences that are most similar to s_{Gen}. Whether the sentences from the most similar list come from the in-domain, or from the general one influences the function that determines if s_{Gen} should belong to the pseudo in-domain or not. This data selection method was termed **SEF**.

Sentence representation using Term Frequency (**DSTF**) was employed aiming at comparing this simple representation with the one based on PV. Using such a simple approach yields fast results since it is based on word frequency. The first step was to create a frequency distribution for the in-domain and for the general domain. Each sentence from the general domain got scored through the relative differences between its word frequencies in the in-domain and in the general domain (for each word of the sentence). A weight was applied in the sentence scoring calculation in order to account for the domain-specificity of the word in both domains.

There is no consensus in the community on how many of the scored general domain sentences should be selected to build the pseudo in-domain. This results in different schemes of reporting empirical outcomes. Experiments using several ratios or thresholds are carried on and the best ratio is determined through the highest BLEU score that an experiment using a certain ratio had. An attempt towards solving this problem is achieved through the use of a feed-forward neural network classifier. The input features were paragraph vectors with the positive samples randomly selected from the in-domain pool and the negative ones randomly selected from the general domain pool. The assumption was that the general domain was large enough, thus the probability to select false negatives is small. Another approach to select negatives was to use one of the previously developed data selection methods to score the general domain and to use the sentences with the lowest similarity to the in-domain as negative samples for the classifier. This method was called **hATD**. In the end, the classifier was used to predict for each sentence from the general domain whether it should be kept or discarded.

8.2 Overview of Experimental Results

The data selection methods (SEF, DSTF, hATD) were evaluated on a common setting: same language pairs and same in-domains. Additionally, the state-of-the-art method, MML, described in Chapter 4, Section 4.2.2 was included into the evaluation. All methods were automatically evaluated using BLEU and manually by means of a three-way ranking for the language pair Spanish→English with sentences from the Biomedical in-domain. For this setting, also a random selection of pseudo in-domain sentences, RND, was evaluated with the purpose of assessing if the gain in translation quality is due to the data selection methods, or due to the additional training data.

In order to verify the system ranking in other settings, the methods were also automatically evaluated on the Biomedical in-domain, for English→Spanish, and on the IT domain, for English→German. For the Biomedical domain, the automatic evaluation was performed on two test sets, Khresmoi and WMT 2018.

The automatic evaluation metrics results produced the following system rankings, in terms of BLEU:

- Biomedical, ES→EN, Khresmoi: *hATD* > *SEF* > *MML* > *DSTF* > *RND*
- Biomedical, ES→EN, WMT: *DSTF* > *hATD* > *MML* > *SEF* > *RND*
- Biomedical, EN→ES, Khresmoi: *hATD* > *SEF* > *DSTF* > *MML*
- Biomedical, EN→ES, WMT: *SEF* > *hATD* > *DSTF* > *MML*
- IT, EN→DE: *hATD* > *SEF* > *MML* > *DSTF*

The systems trained using the random selection of general domain sentences perform poorly on both test sets of the Biomedical domain. This result confirms that the gain in translation performance is not due to adding more general domain training data, but can be attributed to adding more domain-specific data through the data selection methods.

Given the rankings obtained using BLEU on the five experimental settings, a general system ranking cannot be determined since the rankings are not consistent between different choices of language pairs and in-domains. In three out of five settings, hATD outperforms all methods and in the other two ones, it ranks second. This result indicates that the automatic ratio detection approach generally performs better than the standard methods. However, DSTF and SEF also rank first in one setting each. The state-of-the-art method, MML, ranks third on three cases and last for the other two cases. This result shows that the data selection methods that I developed outperform MML. The term frequency approach, DSTF, ranks generally on the last places, with one exception where it ranks first (with a very small BLEU difference to hATD). When considering how statistical significant the results are, there are two partial rankings that are found twice to be statistical significant: SEF better than DSTF and hATD better than MML. Thus, full ranking generalization does not apply for all the results for the experimental settings when using bootstrap resampling. However, partial rankings that are significant can be observed across three test sets[1].

Human evaluation through the tree-way ranking method was applied in the Biomedical domain, on the Khresmoi test set, for the Spanish→English language pair. Six system pairs have been compared: DSTF versus SEF, hATD versus SEF, DSTF versus hATD, MML versus DSTF, MML versus SEF and MML versus hATD. Two Spanish annotators with linguistics background assessed whether a system output is better than another one. Given the test set size of 1000 sentences, there were 2000 human judgments per systems pair. How often a system outperformed another one at sentence level lead to the following results:

MML > *DSTF*, *SEF* > *MML*, *hATD* > *MML*, *SEF* > *DSTF*, *hATD* > *SEF* and *hATD* > *DSTF*.

The following ranking can be derived from the results:

hATD > SEF > MML > DSTF

The system rankings obtained using automatic evaluation placed hATD first as well on this particular experimental setting. Thus, the manual results are consistent with the particular result from the automatic evaluation.

Given two systems A and B, one of the limitations of the three-way ranking procedure is its inability to deal with the cases where the frequency of system A outperforming system B is close the the frequency of B outperforming A. In this case, the higher frequency dominates the ranking. Moreover, three-way ranking does not handle the infinite test set scenario, as pointed out by (Sakaguchi et al., 2014). In order to overcome these drawbacks, I employed the TrueSkill algorithm which computes ranking ranges instead of a system ranking. The result was that the hATD method obtained the highest score, however it was in the same rank

[1] Details on the statistical significant tests can be found in Chapter 6, Section 6.3.3.

range (1, 4) as the other data selection methods that I developed. Since all my methods are in the same rank range, they are considered to be tied. Another outcome was that the MML method had the lowest score and the rank range (2, 4) revealed that it cannot rank first.

The reliability of human judgments was measured through intra- and inter-annotator agreement using the Kappa coefficient and the result was moderate agreement: an intra-annotator score of 0.608 and an inter-annotator score of 0.470.

8.3 Future Work

A list of possible ideas to improve or extend the presented data selection methods is given below:

1. A direction to further investigate, according to the findings from Appendix C, is to use the Bray-Curtis metric in order to assess the similarity between two paragraph vectors instead of the cosine one.

2. Regarding text representation, bilingual word/ sentence embeddings (Shi et al., 2019) is worth investigating, as well as contextualized embeddings, such as Sentence-BERT (Reimers and Gurevych, 2019).

   Word2Vec provides context-free models which generate static embeddings. Therefore, homonyms get the same vector representation irrespective of the context they appeared in. The NMT model presented in Chapter 3, Section 3.2 is based on RNN which has been shown to perform poorly on capturing very long-term dependency between words (Wang et al., 2019). A model that tackles both of these challenges is the transformer, introduced in Vaswani et al. (2017).

   The transformer architecture binds together a stack of $\mathcal{N}$ encoders and $\mathcal{N}$ decoders. The encoder blocks are identical and consist of a multi-head attention layer and a feedforward NN. In contrast to RNN where a sentence is fed word by word, with the transformer all the words are fed in parallel which results in decreased training time and support in learning very long-term dependencies. In order to retain the word order in the sentence, positional encoding is used. The input and the output of the layers are connected through the add and norm component. The decoder blocks also consist of the same layers as the encoder, but with an additional layer for masked multi-head attention. The transformer is trained using cross-entropy as the loss function and the Adam optimizer (Ravichandiran, 2021).

   Bidirectional Encoder Representation from Transformer (BERT) (Devlin et al., 2018) is based on the encoder component of the transformer. Pre-trained BERT models are available which differ in the training data being used and in the configuration (number of encoders, number of attention heads). Sentence-BERT (Reimers and Gurevych, 2019) is used for computing sentence representations and also for assessing if a given sentence pair is similar or not, and for finding most similar sentences given a sentence using cosine. Thus, BERT is highly relevant for the data selection task since it can be used in the initial phase of text representation and for obtaining most similar sentences. The data selection methods that I presented in this thesis can be applied making use of Sentence-BERT. Other pre-trained BERT models that could be taken to future work are ClinicalBERT (Huang et al., 2019) and BioBERT (Lee et al., 2019), which are trained on biomedical data and could be used in the data selection pipeline for the Biomedical in-domain.

3. Since much smaller in size models can be achieved by training MT models using data selection (versus using the full selection pool), it would be interesting to develop an offline MT application (free to download). Offline applications are very useful in the case an Internet connection is not available when traveling, for example. Integrating optical character recognition would make the application even more useful. Applying my data selection methods for language pairs that include Asian languages and the Biomedical domain would benefit travelers that are in need or want to prepare for any medical situation that could arise. For such mobile applications, the size of the model has a great importance.

4. Newly developed measures could also be evaluated with YiSi (Lo, 2019), which is an MT evaluation metric based on BERT that recently showed very good correlation with human judgments.

With NMT being the current[2] state-of-the-art for MT, data selection and domain adaptation for NMT are still hot topics that are actively researched by the MT community. For example, Aharoni and Goldberg (2020) employ BERT with data selection, while Dou et al. (2020) exploits the MML method for NMT using iterative back-translation.

[2] Current refers to the date of writing this section, June 2021.

Appendix A

Doc2Vec Hyper-Parameters

All the Paragraph Vectors were trained using Doc2Vec from the *Gensim* library. The table below gives give a description of the hyper-parameters used in training my models, along with the default values. Due to brevity, only a subset of the parameters are enumerated here [1]. The hyper-parameters that I tuned were the dimension of the vectors and the Doc2Vec algorithm.

Hyper-Parameter	**Default value**	**Explanation**
dm	1	the algorithm used for training the Doc2Vec model. Default value is 1, encoding 'distributed memory' (PV-DM), whereas 0 encodes 'distributed bag of words' (PV-DBOW).
documents		the corpora used for training the model (each document is a TaggedDocument object).
size	100	the dimension of the vectors. Default value is 100.
window	5	the number of words used as left and right context.
min count	5	a threshold indicating that words that have a lower frequency than min count should be ignored.
sample	0.001	a threshold indicating which words with high-frequency should be downsampled (stop-words can be removed with this hyper-parameter)
negative	5	negative sampling will be used if the value set is higher than 0. Introduced in (Mikolov et al., 2013b)
dbow words	0	faster training if set to default value as it trains only doc2vec vectors. If set to 1 it will train word vectors using the skip-gram algorithm along with DBOW training of doc2vec vectors.
dm concat	0	if set to 1 it uses concatenation of context vectors.
iter	5	the number of epochs used for training the model.
workers	3	the number of threads used for training (on a multicore machine it results in faster training).

Table A.1: Hyper-Parameters explanation for Doc2Vec

[1] The complete list can be found on the Gensim webpage `https://radimrehurek.com/gensim/models/doc2vec.html`.

Appendix B

Word2Vec for the Biomedical and the IT domains

This appendix presents examples of operations using Word2Vec for two domains and visualizations of word vectors.

In Mikolov et al. (2013b) word embeddings are presented as being highly capable of encoding certain linguistic regularities and patterns. Examples are given regarding linear combinations of vectors or other algebraic operations, like the frequently-cited examples:

(1) $\mathit{vector}(''\mathit{king}'') - \mathit{vector}(''\mathit{man}'') + \mathit{vector}(''\mathit{woman}'')$ results in a vector that is closest to the vector representation of the word "*queen*" (Mikolov et al., 2013a)

(2) $\mathit{vector}(''\mathit{madrid}'') - \mathit{vector}(''\mathit{spain}'') + \mathit{vector}(''\mathit{france}'')$ results in a vector that is closest to the vector representation of the word "*paris*" (Mikolov et al., 2013b)

Considering the two in-domains that are explored in this thesis, I tested similar operations with corresponding word embedding models[1], as an initial step to determine whether these models are suitable for my architecture and experiments. For the medical in-domain I selected the terms presented below (underlined):

- **Inflammation** in **joints** is a common sign of <u>arthritis</u>.
- **Thyroid** hormone **deficiency** leads to <u>hypothyroidism</u>.

Vectors for the words emphasized in bold were given as a list of words that contribute positively in finding out the most similar word in terms of cosine similarity, formally presented below[2]. The similarity score is presented together with the most similar word.

$$\mathcal{M}(\mathit{positive} = ['\mathit{inflammation}', '\mathit{joints}']) \rightarrow ('\mathit{arthritis}', 0.72)$$

$$\mathcal{M}(\mathit{positive} = ['\mathit{deficiency}', '\mathit{thyroid}']) \rightarrow ('\mathit{hypothyroidism}', 0.83)$$

[1] A model trained on a Wikipedia dump and made available on `https://github.com/jhlau/`, as well as a model I trained on the general domain data

[2] The similarity scores were trimmed down to two digits for layout reasons; $\mathcal{M}$ stands for "model" and the *most_similar* method is applied using it

In these examples, the words from the list of positives contributed to successfully selecting the most similar word to them (the sum of the word vectors for the words in the positive list results in a word vector that is most similar to the sum).

An analogical reasoning query (similar to examples (1) and (2)) was also defined using medical terms and achieved an expected result. The reasoning I used is based on the assumption that vectors for diseases should be situated in close proximity in the vector space (similar to 'man' and 'woman' from (1)) and that the vectors for the underlying aspect of diseases could be found by a similar offset (similar to 'king' and 'queen' in (1)). The diseases chosen for the query were **arthritis**[3] and **ulcer**[4] and the corresponding chosen aspects were **joints** and **stomach**. Arthritis affects the joints and ulcer affects the stomach. Below the cosine distance between the pairs of vectors is given, as well as the analogical reasoning query result:

$$\mathcal{M}(\mathsf{\textit{positive}} = [\mathsf{\textit{'arthritis','stomach'}}], \mathsf{\textit{negative}} = [\mathsf{\textit{'joints'}}]) \rightarrow (\mathsf{\textit{'ulcer'}}, 0.67)$$

The same testing procedure was repeated in the IT domain with the following three terms (underlined):

- A <u>Malware</u> is a **malicious software**.
- <u>HTML</u> is a **markup language** for creating **web** pages.

Similarly to the previous domain, vectors for the words emphasized in bold were summed up resulting in the vectors of the underlined words, formally presented below:

$$\mathcal{M}(\mathsf{\textit{positive}} = [\mathsf{\textit{'malicious','software'}}]) \rightarrow (\mathsf{\textit{'malware'}}, 0.74)$$

$$\mathcal{M}(\mathsf{\textit{positive}} = [\mathsf{\textit{'markup','language','web'}}]) \rightarrow (\mathsf{\textit{'html'}}, 0.74)$$

In addition to performing algebraic operations on word vectors, they can also be visualized in a 2D space using Principal Component Analysis (PCA) for dimensionality reduction (Jolliffe, 1986).

Plotting all the words from the vocabulary leads to a visual overlap which cannot be interpreted, therefore I picked sixty representative words for each in-domain. The English side of the medical corpora that I used in the Biomedical Task of WMT 2017 was employed to obtain a vocabulary and then automatically extracting the top 200 most frequent words that appeared in the whole corpus. Since these most frequent words were not pertaining only to the medical domain, I manually inspected the list and selected sixty medical terms. The same procedure was applied to the IT domain where the IT corpus used in the domain adaptation task of WMT 2016 was employed as the corpus for obtaining a vocabulary.

[3] A common condition that causes pain and inflammation in the joints (Definition from `https://www.nhs.uk/conditions/arthritis`)

[4] Open sores that develop on the lining of the stomach (Definition from `https://www.nhs.uk/conditions/stomach-ulcer`)

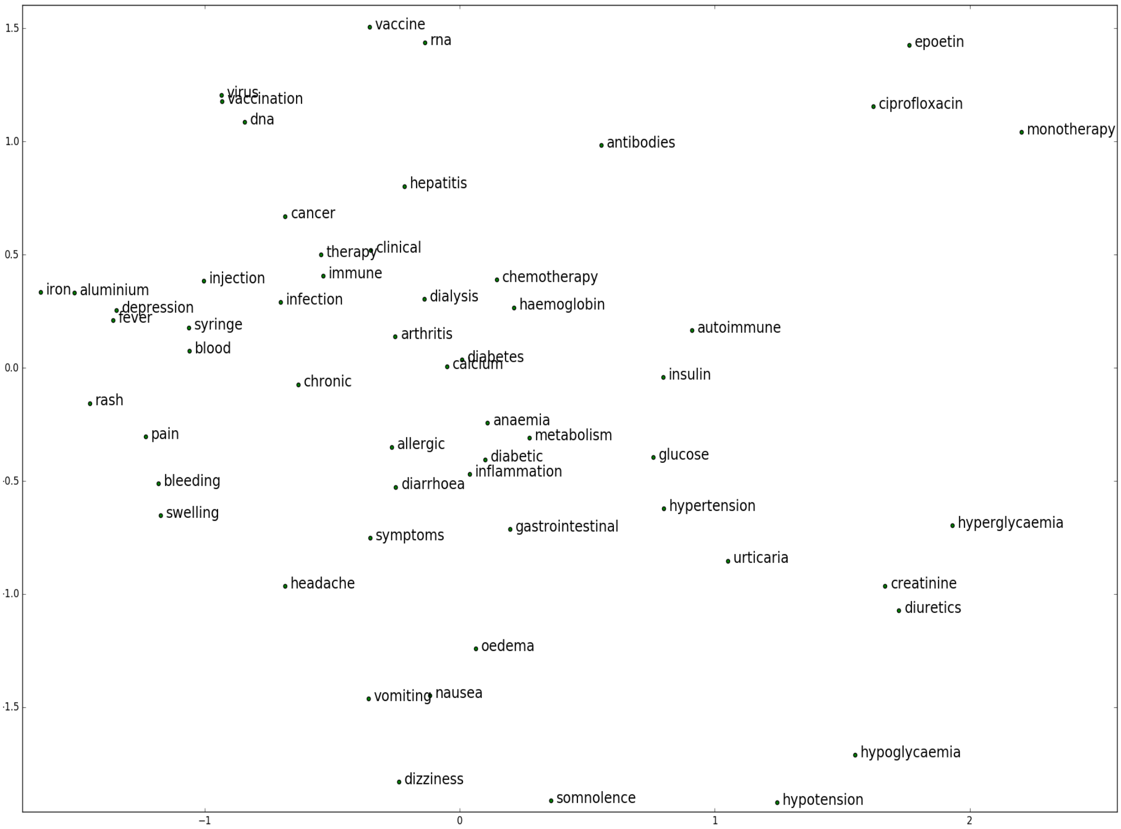


Figure B.1: Visualization of frequent medical terms using two-dimensional PCA projection of their corresponding word embeddings.

In Figure B.1 the vectors for a set of ≈ 60 frequently used medical terms (obtained using the vocabulary extracted from the Biomedical Task from WMT 2017) are depicted according to their two-dimensional PCA projection. The model is able to group together symptoms/ side-effects[5] like *headache*, *vomiting*, *nausea*, *dizziness* and *somnolence*. It can also be observed that *iron* and *aluminium* are situated very close to each other, as well as *diabetes* and *insulin*(a hormone usually injected to the patients suffering of diabetes).

Figure B.2 shows the projection of the vector space for a set of ≈ 60 frequently used IT terms (according to the vocabulary extracted from the corpora provided by WMT 2016 domain adaptation task). The model groups words close in meaning and usage and is observed with pairs like $(\mathit{database}, \mathit{oracle})$ or $(\mathit{menu}, \mathit{click})$[6].

Figure B depicts the visualization of frequent, domain-specific terms for the medical and IT domain using two-dimensional PCA projection of their correspond-

[5] NHS (National Health Service) defines side effects as unwanted symptoms triggered by medical treatment(`https://www.nhs.uk/common-health-questions/medicines/what-are-side-effects`)

[6] A database is a collection of data organized in a structured way and Oracle is a database management system. In the IT domain, a menu refers to a list of items that trigger a command usually when performing a click event (user clicking on the item)

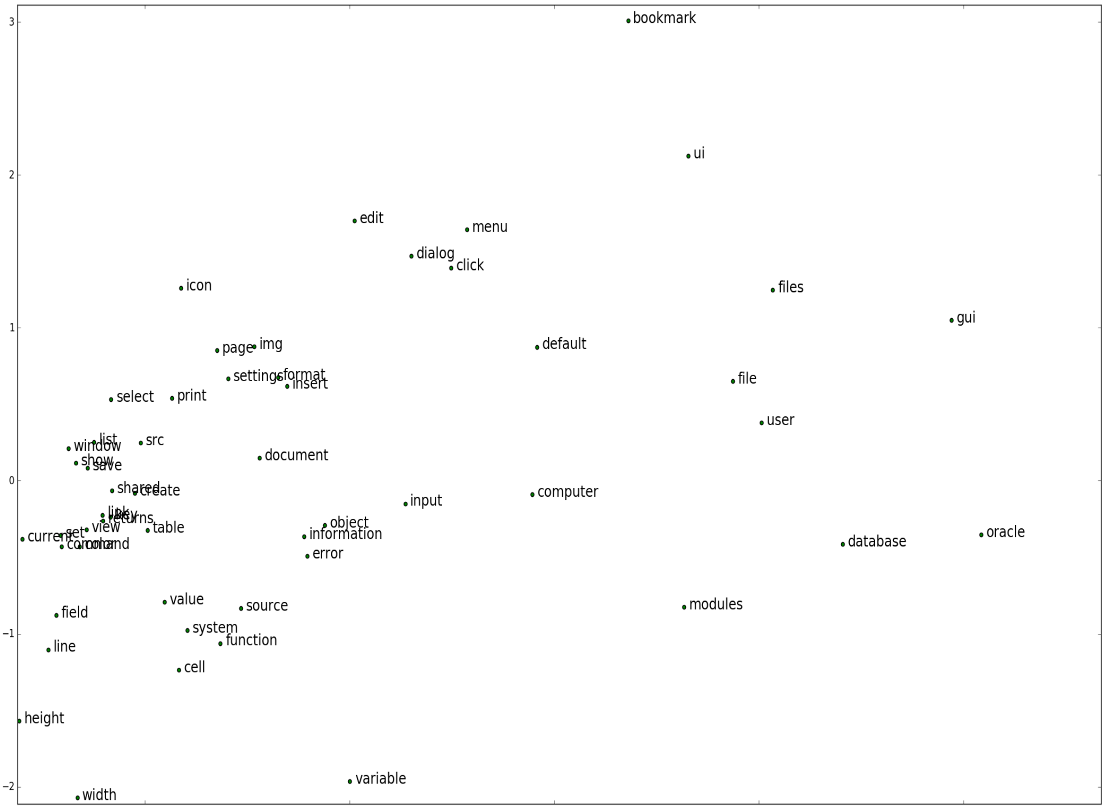


Figure B.2: Visualization of frequent IT terms using two-dimensional PCA projection of their corresponding word embeddings.

ing word embeddings[7]. The set of fourteen words situated at the intersection between the two in-domains are highlighted in a blue rectangle[8].

The domains form two well-defined clusters, indicating that the word embeddings are powerful in encoding relations between words. A remark on the intersection of the two domains is that the marked words are obviously belonging to the medical domain but not so clearly to the IT domain. One possible reason behind this lies in the fact that the transfer of bio-inspired patterns, algorithms, ideas into the IT domain is much more dominant than the opposite direction. Also, an intersection between the two domains is inevitable since both domains are integrated into a parent-domain: science.

[7] Obtained using a model trained on a Wikipedia dump and available at `https://github.com/jhlau/doc2vec`

[8] The words are: *DNA*, *Virus*, *Cell*, *Vaccination*, *Cancer*, *Therapy*, *Injection*, *Syringe*, *Blood*, *Fever*, *Pain*, *Rash*, *Bleeding* and *Swelling*

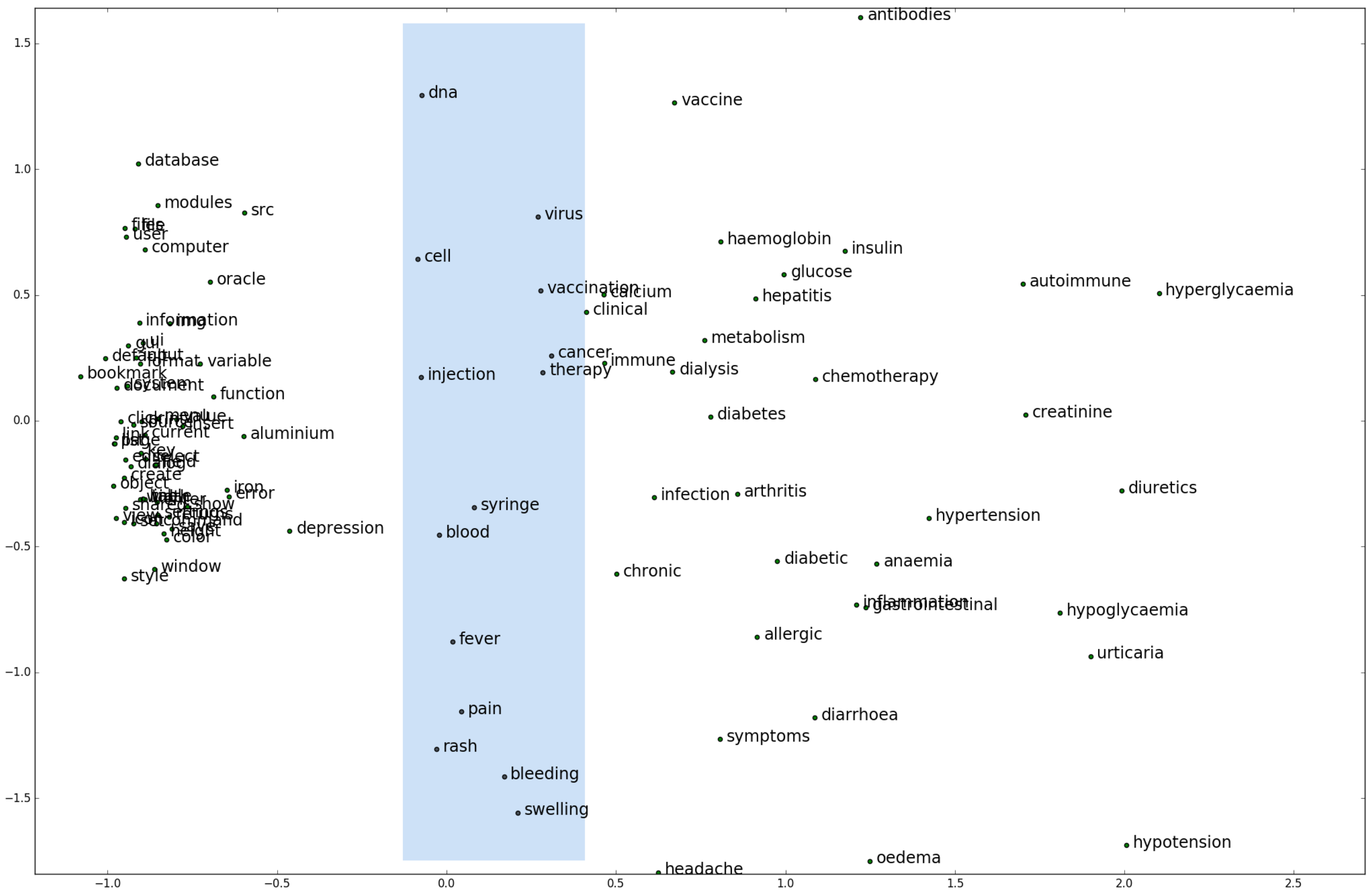


Figure B.3: Visualization of frequent medical and IT terms using two-dimensional PCA projection of their corresponding word embeddings. The blue rectangle emphasizes the set of words situated at the intersection between the two in-domains.

Appendix C

Doc2Vec for Semantic Textual Similarity

This appendix describes the task of Semantic Textual Similarity (STS) situated in connection with data selection for MT. Moreover, my participation in a STS task from an international competition is presented (Duma and Menzel, 2017b). The goal was to evaluate Paragraph Vector (PV) for STS since it is a related task to data selection.

Semantic Textual Similarity (STS) assesses the degree in which two snippets of text are similar in meaning to each other. For my experiments, a snippet of text consists of a single sentence. STS can be applied in machine translation (Zou et al., 2013; Castillo and Estrella, 2012; Duma and Menzel, 2016a,b), information retrieval (Kim et al., 2017), text summarization (Al-Khassawneh et al., 2016; Verma and Verma, 2020) and question answering (Park et al., 2014; Özyurt et al., 2020). Some difficulties in STS are determining the degree of meaning overlap (Cer et al., 2017), capturing the relationship between the words contained in the sentences and choosing a method to represent the sentences in order to develop a way that lends itself to STS.

The connection between data selection (for MT) and STS is that both tasks aim at identifying the similarity between two snippets of text. Given an in-domain and a general domain corpus, the goal of data selection is to extract the sentences from the general domain that are most similar to the in-domain. Therefore, developing semantic textual similarity methods could be applied to data selection. Even when considering cross-lingual STS, it could be considered equivalent with the task of bilingual-focused data selection since both tasks act on the two language pairs involved using a combination of results produced using both languages.

In my data selection experiments, sentences are represented as PVs for both the general domain and the in-domain corpora. In order to obtain the similarity between two sentences, I used the default cosine similarity. Another similarity metric is explored in this appendix with the purpose of evaluating an alternative to the standard cosine similarity on the STS task.

Substantial research has been conducted for STS and SemEval constitutes an annual workshop and competition that offers a track on STS (Cer et al., 2017).

Given two sentences in the same language, the task is to assign a similarity score ranging from 0 to 5, with 0 indicating that the semantics of the sentences are completely independent and 5 signifying semantic equivalence (Cer et al., 2017). In 2017, a cross-lingual version of STS was introduced which is similar to the initial task, but differs in the input sentences which come from two languages.

The 2017 shared task featured six sub-tasks: Arabic-Arabic, Arabic-English, Spanish-Spanish, Spanish-English (two test sets), English-English and a surprise task for which no annotated data was offered- Turkish-English. I participated in all sub-tasks, submitting three runs per sub-task.

The evaluation data sets are extracted from the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). SNLI was used for all language pairs. For English-Spanish, however, two sub-tracks were given: one extracted from SNLI and one extracted from WMT quality estimation (QE) data. For the cross-lingual sub-tasks, the sentences were translated by humans. The Spanish side of the WMT QE sub-track was constructed by translating the English sentences using various MT systems, annotated with the Human-targeted Translation Error Rate (HTER) (Snover et al., 2006). HTER represents the minimum number of edits needed to correct a translation divided by its length post-editing by humans. Every test set was comprised of 250 sentence pairs.

The English-Spanish WMT QE sub-task presents more challenges compared to the other sub-tasks: the average sentence length is 19.4 words (compared to the English-English sub-task with an average sentence length of 8.7) and capturing the meaning differences introduced by MT errors presents more difficulties (Cer et al., 2017).

Table C.1 presents examples that I extracted from the cross-lingual English-Spanish track from SemEval 2017 test set with explanations of scores (according to Agirre et al. (2013)). For each example I identified the differences between the two sentences in order to better sustain the score explanation. The gold standard scores were provided by human annotators[1] and were prepared by SemEval using Amazon Mechanical Turk[2].

In the following, details on the methods I used for the participation to the STS SemEval competition as well as additional experiments are presented.

The sentences were represented using Paragraph Vector and the semantic similarity score between two sentences was computed using standard similarity metrics. The most traditional similarity metric used in the STS community is the Cosine and it is also the metric implemented into *Gensim*, the library I used for implementation.

My approach for STS has the advantages of being unsupervised and knowledge-free. The empirical setting included various Doc2Vec sizes (200, 300 and 400) and standard similarity metrics as Cosine and Bray-Curtis (Bray and Curtis, 1957).

[1] Without any formal background in linguistics.

[2] `https://www.mturk.com/`

Score	**Sentences**	**Explanation**
0	Un camarógrafo sale a comer con su familia. (translation: A cameraman goes out to eat with his family.) A man is waiting for his friend to catch up.	The two sentences are completely not similar: there is no overlap in meaning (”cameraman” $\neq$ ”man”, ”goes out” $\neq$ ”waiting”, ”to eat” $\neq$ ”to catch up”, ”his family” $\neq$ ”his friend”).
1	Cuatro chicos en un barco juntos. (translation: Four boys on a boat together.) Three boys put together a sailboat.	The two sentences are not equivalent, but share the same topic: boys with a boat is the topic of the two sentences and the differences are the number of boys, the action they do and the type of boat.
2	El hombre baila en la calle mientras otro mira. (translation: The man dances in the street while another one is looking.) A woman and a man are dancing in the street.	The two sentences are not equivalent, but share some details: the action of dancing in the street is shared by the two sentences.
3	Las bicicletas estàn en una carretera. (translation: The bicycles are on a road.) Bicycles are laying on the side of the road.	The two sentences are roughly equivalent, but some important information differs/missing: specific detail on where the bicycles are situated.
4	Un grupo de personas posando fuera de un edificio para una sesión de fotos de una revista. (translation: A group of people posing outside a building for a photo shoot for a magazine). A group of people pose for a photograph.	The two sentences are mostly equivalent, but some unimportant details differ: the Spanish sentence indicates in addition the place where the group of people are and that the photographs are for a magazine.
5	Una niña peqeã corriendo en un campo. A little girl is running in a field.	The two sentences are completely equivalent, as they mean the same thing: each sentence could be a translation of the other one.

Table C.1: Score explanation for sentences from SemEval EN-ES test set

The metrics are defined as:

$$Cosine : 1 - \frac{u \cdot v}{||u||_2||v||_2}$$

$$Bray - Curtis : \frac{\sum_i |u_i - v_i|}{\sum_i |u_i + v_i|}$$

where u_i and v_i are the vector representations of the two sentences, $\bar{u}$ and $\bar{v}$ denote the mean value of the elements of u and v, and $x \cdot y$ is the dot product of x and y.

The Cosine metric is directly implemented by *Gensim* and the Bray-Curtis implementation was obtained from the *spatial* library of *scipy*[3].

For any monolingual sub-task, I trained the Doc2Vec model on the corpora available for the respective task. For any cross-lingual sub-task, Doc2Vec models were trained for both languages.

The Stanford Arabic Segmenter (Monroe et al., 2014) was used for the Arabic-Arabic and Arabic-English sub-tasks aiming to reduce lexical sparsity. For all the other sub-tasks, text normalization, tokenization and lowercasing using the scripts available in the Moses Machine Translation Toolkit (Koehn et al., 2007) was applied to the corpora.

SemEval uses the Pearson Correlation coefficient of systems scores with gold standard human judgments to measure the performance of the submitted systems (Cer et al., 2017). According to the SemEval evaluation ranking procedure, $System_A$ is considered to outperform $System_B$, if the Pearson Correlation of $System_A$ is higher than the Pearson Correlation of $System_B$.

Several parameters were explored, including the size of the Doc2Vec vectors (200, 300 and 400), considering both sides of the bilingual corpora for the cross-lingual tasks and applying Cosine and Bray-Curtis metrics. Table C.2 shows the Pearson Correlation scores obtained by experimenting with all combinations of the parameters values.

Given the seven test sets, the best performing similarity metric in terms of Pearson correlation[4] is Bray-Curtis outperforming Cosine with a tie with Cosine on the EN-EN test set. This result could be due to the fact that Bray-Curtis is sensitive to the differences between every element of the two vector representations of the sentences. This actually represents the core of identifying the similarity between two vectors as from a logical point of view *differences* lead to *dissimilarity.*

It is not possible to indicate the best vector size for the PV since all three investigated sizes lead to best results for at least two sub-tasks. Regarding the choice of computing vector similarity, Bray-Curtis outperforms Cosine. The combination of the two scores obtained by averaging the scores for both languages leads to

[3] `https://docs.scipy.org/doc/scipy-0.18.1/reference/spatial.html`

[4] The Pearson correlation does not offer an indication of how close/ distant the metric scores are to the gold standard scores. Its formula relies on computing the normalized covariance of the gold standard and the metric scores, thus how well the two variables move together.

Task		Cosine			Bray-Curtis	
AR-AR						
200		0.5587			0.5790	
300		0.5825			**0.5984**	
400		0.5773			0.5943	
AR-EN	AR	EN	Mean	AR	EN	Mean
200	0.4789	0.4971	0.5221	0.4755	0.503	0.5268
300	0.4963	0.5141	0.5429	0.502	0.5085	0.5432
400	0.4813	0.5266	0.5381	0.4949	0.5288	**0.5469**
ES-ES						
200		**0.7455**			0.7423	
300		0.7002			0.7054	
400		0.6979			0.7072	
ES-EN-a	ES	EN	Mean	ES	EN	Mean
200	0.5738	0.6021	0.6212	0.5852	0.6208	**0.6353**
300	0.5676	0.6162	0.6219	0.5793	0.6253	0.6299
400	0.566	0.6092	0.6187	0.5767	0.6162	0.6253
ES-EN-b	ES	EN	Mean	ES	EN	Mean
200	0.3069	0.1933	0.3111	0.306	0.1686	0.2953
300	0.3234	0.1784	0.3193	0.3187	0.1685	0.3099
400	0.3407	0.1873	0.3303	**0.3436**	0.1575	0.3113
EN-EN						
200		**0.7880**			**0.7880**	
300		0.7237			0.7396	
400		0.7185			0.7264	
TR-EN	TR	EN	Mean	TR	EN	Mean
200	0.4990	0.5554	0.5804	0.5080	0.5577	0.5846
300	0.4919	0.5718	0.5792	0.4869	**0.6001**	0.5879
400	0.4878	0.5832	0.5775	0.5024	0.6000	0.5930

Table C.2: Pearson Correlation results for various parameters

the best results for two out of four cross-language sub-tasks. This outcome suggests that the straightforward average of scores is worth investigating for other cross-lingual tasks.

SemEval introduced in 2017 the STS Benchmark[5] (Cer et al., 2017) where I was asked to contribute with evaluation results, since my method achieved the best result on one sub-track: first place out of 53 total submissions from all participants (including one baseline) on the Spanish-English-WMT test set. The purpose of the STS Benchmark is to establish state-of-the-art approaches and state their results on standard data sets.

The benchmark collected a selection of previous data sets from 2012 until 2017 for the EN-EN sub-task, comprised of 8628 sentence pairs. I did not use any

[5] `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`

annotated training dataset as my method is knowledge-free and unsupervised. The Pearson correlation scores obtained when evaluating my method were 0.6158 for the development set and 0.5922 for the test set.

In conclusion, a wide range of experiments were conducted indicating that the Bray-Curtis metric could be considered as a replacement for the traditional Cosine approach since it achieved better results on five out of seven language pairs. Regarding the size of the paragraph vectors, no conclusion can be drawn from the results. Using the Bray-Curtis metric for the similarity calculation in the data selection pipeline is an encouraging idea for future work.

Bibliography

Sadaf Abdul-Rauf, Holger Schwenk, Patrick Lambert, and Mohammad Nawaz. 2016. Empirical Use of Information Retrieval to Build Synthetic Data for SMT Domain Adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(4):745–754. https://doi.org/10.1109/TASLP.2016.2517318.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 32–43. https://www.aclweb.org/anthology/S13-1004.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 7747–7763. https://aclanthology.org/2020.acl-main.692.

Qingyao Ai, Liu Yang, Jiafeng Guo, and W. Bruce Croft. 2016. Analysis of the Paragraph Vector Model for Information Retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM, New York, USA, ICTIR '16, pages 133–142. http://doi.acm.org/10.1145/2970398.2970409.

Yazan Al-Khassawneh, Naomie Salim, and Adekunle Obasae. 2016. Sentence Similarity Techniques for Automatic Text Summarization. *Journal of Soft Computing and Decision Support Systems* 3:35–41.

Eleftherios Avramidis, Aljoscha Burchardt, Vivien Macketanz, and Ankit Srivastava. 2016. DFKI's system for WMT16 IT-domain task, including analysis of systematic errors. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 415–422. http://www.aclweb.org/anthology/W/W16/W16-2329.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, EMNLP '11, pages 355–362. http://dl.acm.org/citation.cfm?id=2145432.2145474.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473.

Satajeev Banerjee and Alon Lavie. 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*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 65–72. https://www.aclweb.org/anthology/W05-0909.

Amira Barhoumi, Yannick Estève, Chafik Aloulou, and Lamia Belguith. 2017. Document embeddings for Arabic Sentiment Analysis. In *Conference on Language Processing and Knowledge Management, LPKM 2017*. Sfax, Tunisia. https://hal.archives-ouvertes.fr/hal-02042060.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, pages 1–61. http://www.aclweb.org/anthology/W19-5301.

Ergun Biçici and Deniz Yuret. 2011. Instance Selection for Machine Translation Using Feature Decay Algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, WMT '11, pages 272–283. http://dl.acm.org/citation.cfm?id=2132960.2132996.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, USA, StatMT '07, pages 9–16.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. In *O'Reilly Media Inc.*

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin M. Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016b. Ten Years of WMT Evaluation Campaigns:

Lessons Learnt. In *LREC 2016 Workshop Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*. pages 27–34. https://publications.rwth-aachen.de/record/668803.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 169–214. http://www.aclweb.org/anthology/W17-4717.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

J Roger Bray and John T Curtis. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological monographs* 27(4):325–349.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017a. Massive Exploration of Neural Machine Translation Architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1442–1451. https://www.aclweb.org/anthology/D17-1151.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V. Le. 2017b. Massive Exploration of Neural Machine Translation Architectures. *CoRR* abs/1703.03906. http://arxiv.org/abs/1703.03906.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Trento, Italy. https://www.aclweb.org/anthology/E06-1032.

Julio Castillo and Paula Estrella. 2012. Semantic Textual Similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada, pages 52–58. https://www.aclweb.org/anthology/W12-3103.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada, pages 1–14. https://www.aclweb.org/anthology/S17-2001.pdf.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 545–560. https://www.aclweb.org/anthology/C12-1034.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. ACL, pages 1724–1734. https://www.aclweb.org/anthology/D14-1179.

Chenhui Chu. 2015. Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. http://arxiv.org/abs/1511.07289.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, volume 20(1):37 - 46.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 728–735. https://www.aclweb.org/anthology/P07-1092.

Hoang Cuong and Khalil Sima'an. 2014. Latent Domain Phrase-based Models for Adaptation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 566–576. https://doi.org/10.3115/v1/D14-1062.

Hoang Cuong and Khalil Sima'an. 2017. A Survey of Domain Adaptation for Statistical Machine Translation. *Machine Translation* 31(4):187–224. https://doi.org/10.1007/s10590-018-9216-8.

Hoang Cuong, Khalil Sima'an, and Ivan Titov. 2016. Adapting to All Domains at Once: Rewarding Domain Invariance in SMT. *Transactions of the Association for Computational Linguistics* 4:99–112. https://www.aclweb.org/anthology/Q16-1008.

Hal Daumé and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research* 26(1):101–126.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 407–412. https://www.aclweb.org/anthology/P11-2071.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Uppsala, Sweden, pages 339–342. https://www.aclweb.org/anthology/W10-1751.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 85–91. https://www.aclweb.org/anthology/W11-2107.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805. http://arxiv.org/abs/1810.04805.

Bonnie J. Dorr, Matt Snover, and Nitin Madnani. 2011. *Chapter 5: Machine Translation Evaluation*. Springer Publishing Company, Incorporated, 1st edition.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic Data Selection and Weighting for Iterative Back-Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 5894–5904. https://aclanthology.org/2020.emnlp-main.475.

Mirela-Stefania Duma and Wolfgang Menzel. 2016a. Data Selection for IT Texts using Paragraph Vector. In *Proceedings of the First Conference on Machine Translation, August 11-12, Berlin, Germany*. pages 428–434. http://aclweb.org/anthology/W/W16/W16-2331.pdf.

Mirela-Stefania Duma and Wolfgang Menzel. 2016b. Paragraph Vector for Data Selection in Statistical Machine Translation. In *Proceedings of the 13th Conference on Natural Language Processing KONVENS 2016, September 19-21, Bochum, Germany*. pages 84–89. https://www.linguistics.rub.de/konvens16/pub/11_konvensproc.pdf.

Mirela-Stefania Duma and Wolfgang Menzel. 2017a. Automatic Threshold Detection for Data Selection in Machine Translation. In *Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, September 7-8, 2017*. pages 483–488. http://aclanthology.info/papers/W17-4754/w17-4754.

Mirela-Stefania Duma and Wolfgang Menzel. 2017b. SEF@UHH at SemEval-2017 Task 1: Unsupervised Knowledge-Free Semantic Textual Similarity via Paragraph Vector. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. pages 170–174. https://doi.org/10.18653/v1/S17-2024.

Mirela-Stefania Duma and Wolfgang Menzel. 2018. Translation of Biomedical Documents with Focus on Spanish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*. pages 637–643. https://aclanthology.info/papers/W18-6444/w18-6444.

Mirela-Stefania Duma and Cristina Vertan. 2013. Integration of Machine Translation in On-line Multilingual Applications : Domain Adaptation. *Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe* https://www.blogs.uni-mainz.de/fb06-tc3/files/2015/11/33-147-1-PB.pdf.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Language Model Adaptation for Statistical Machine Translation Based on Information Retrieval. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/374.pdf.

Sauleh Eetemadi, William Lewis, Kristina Toutanova, and Hayder Radha. 2015. Survey of Data-selection Methods in Statistical Machine Translation. *Machine Translation* 29(3-4):189–223. https://doi.org/10.1007/s10590-015-9176-1.

M F. Porter. 2001. Snowball: A language for Stemming Algorithms. In *Retrieved March*. volume 1.

Christian Federmann. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics* 98:25–35. https://ufal.mff.cuni.cz/pbml/98/art-federmann.pdf.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. In *Cambridge, MA: MIT Press*. https://wordnet.princeton.edu/.

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, pages 128–135. https://www.aclweb.org/anthology/W07-0717.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a Unified Approach to Statistical Language Modeling for Chinese. *ACM Transactions on Asian Language Information Processing* 1(1):3–33. https://doi.org/10.1145/595576.595578.

Rosa Gaudio, Gorka Labaka, Eneko Agirre, Petya Osenova, Kiril Simov, Martin Popel, Dieke Oele, Gertjan van Noord, Luís Gomes, João António Rodrigues, Steven Neale, João Silva, Andreia Querido, Nuno Rendeiro, and António Branco. 2016. SMT and Hybrid systems of the QTLeap project in the WMT16 IT-task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 435–441. http://www.aclweb.org/anthology/W/W16/W16-2332.

Cristian Grozea. 2018. Ensemble of Translators with Automatic Selection of the Best Translation – the Submission of FOKUS to the WMT 18 Biomedical Translation Task –. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 655–658. http://www.aclweb.org/anthology/W18-6446.

Junfei Guo, Juan Liu, Qi Han, and Andreas Maletti. 2014. A Tunable Language Model for Statistical Machine Translation. Proceedings of AMTA 2014, Vancouver, BC, pages 356–368. http://www.informatik.uni-leipzig.de/alg/pub/guoliuhanmal14.pdf.

Zellig Harris. 1954. Distributional Structure. *Word* 10:146–162. https://link.springer.com/chapter/10.1007/978-94-009-8467-7_1.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014. Dynamic Topic Adaptation for SMT using Distributional Profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 445–456. http://www.aclweb.org/anthology/W/W14/W14-3358.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *WMT@EMNLP*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 20*. MIT Press, pages 569–576. https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of EAMT*. pages 133–142.

Erik Hollnagel. 1993. *Human Reliability Analysis: Context and Control*, volume Cognition. Academic Press.

Chengcheng Hu and Xuliang Song. 2016. Microblog Sentiment Analysis Based on Paragraph Vectors. *Journal of Computers* 11:83–90. https://doi.org/10.17706/jcp.11.1.83-90.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *CoRR* abs/1904.05342. http://arxiv.org/abs/1904.05342.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 315–322. http://www.aclweb.org/anthology/W17-4730.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. LMU Munich's Neural Machine Translation Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 659–665. http://www.aclweb.org/anthology/W18-6447.

Ann Irvine and Chris Callison-Burch. 2014. Using Comparable Corpora to Adapt MT Models to New Domains. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 437–444. http://www.aclweb.org/anthology/W/W14/W14-3357.

Ann Irvine, Chris Quirk, and Hal Daumé III. 2013. Monolingual Marginal Matching for Translation Model Adaptation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1077–1088. https://www.aclweb.org/anthology/D13-1109.

Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 234–247. http://www.aclweb.org/anthology/W17-4719.

Ian T. Jolliffe. 1986. *Principal Component Analysis*. Springer Verlag.

Uday Kamath, John Liu, and James Whitaker. 2019. *Deep Learning for NLP and Speech Recognition*. Springer. https://doi.org/10.1007/978-3-030-14596-5.

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-Gram-Based Author Profiles For Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING 2003*.

Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter NMT System for WMT18 Biomedical Translation Task: Transfer Learning in Neural

Machine Translation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 666–672. http://www.aclweb.org/anthology/W18-6448.

Eric Kim. 2019. Demystifying Neural Network in Skip-Gram Language Modeling. Accessed on 2020-09-30. https://aegis4048.github.io/demystifying_neural_network_in_skip_gram_language_modeling.

Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. 2017. Bridging the Gap: Incorporating a Semantic Similarity Measure for effectively mapping PubMed Queries to Documents. *J. Biomed. Informatics* 75:122–127. http://dblp.uni-trier.de/db/journals/jbi/jbi75.html.

Katrin Kirchhoff and Jeff A. Bilmes. 2014. Submodularity for Data Selection in Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 131–141. http://aclweb.org/anthology/D/D14/D14-1014.pdf.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, pages 67–72. https://www.aclweb.org/anthology/P17-4012.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval : Graphical evaluation interface for Machine Translation development. In *The Prague Bulletin of Mathematical Linguistics, Number 104*. pages 63–74.

Philipp Koehn. 2010a. An Experimental Management System. In *Proceedings of the Machine Translation Marathon 2010*. The Prague Bulletin of Mathematical Linguistics, volume 94, pages 86–96.

Philipp Koehn. 2010b. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Philipp Koehn. 2017. Neural Machine Translation. *CoRR* abs/1709.07809. http://arxiv.org/abs/1709.07809.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, pages 28–39. https://doi.org/10.18653/v1/W17-3204.

Philipp Koehn and Josh Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation* pages 224–227. https://doi.org/10.3115/1626355.1626388.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to its Methodology*. Sage Publications.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33 1:159–74.

Alon Lavie. 2010. Evaluating the Output of Machine Translation Systems. In *AMTA*. Denver, Colorado, USA.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '07, pages 228–231. http://dl.acm.org/citation.cfm?id=1626355.1626389.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. JMLR.org, ICML'14, pages II–1188–II–1196. http://dl.acm.org/citation.cfm?id=3044805.3045025.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *CoRR* abs/1901.08746. http://arxiv.org/abs/1901.08746.

Sung-Chien Lin, Chi-Lung Tsai, L. Chien, K. Chen, and L. Lee. 1997. Chinese Language Model Adaptation based on Document Classification and Multiple Domain-specific Language Models. In *EUROSPEECH*.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. pages 507–513. https://doi.org/10.18653/v1/W19-5358.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, pages 343–350. https://www.aclweb.org/anthology/D07-1036.

Yi Lu, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Yiming Wang. 2014. Domain Adaptation for Medical Text Translation using Web Resources. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 233–238. http://www.aclweb.org/anthology/W/W14/W14-3328.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.* 2(2):159–165. https://doi.org/10.1147/rd.22.0159.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural Machine Translation (seq2seq) Tutorial. *https://github.com/tensorflow/nmt* .

Matouvs Machávcek and Ondřej Bojar. 2015. Evaluating Machine Translation Quality Using Short Segments Annotations. *The Prague Bulletin of Mathematical Linguistics* 103:85–110. https://doi.org/10.1515/pralin-2015-0005.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Saab Mansour and Hermann Ney. 2014. Unsupervised Adaptation for Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 457–465. http://www.aclweb.org/anthology/W/W14/W14-3359.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 708–717. https://www.aclweb.org/anthology/D09-1074.

Victor Mijangos, Gerardo Sierra, and Abel Herrera. 2016. A Word Embeddings Model for Sentence Similarity. *Research in Computing Science* 117:63–74.

Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. pages 1–12. https://www.researchgate.net/publication/319770439_Efficient_Estimation_of_Word_Representations_in_Vector_Space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Will Monroe, Spence Green, and Christopher D. Manning. 2014. Word Segmentation of Informal Arabic with Domain Adaptation. In *Proceedings of* *the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics (ACL). https://www.aclweb.org/anthology/P14-2034.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, pages 220–224. https://www.aclweb.org/anthology/P10-2041.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML 2010*. pages 807–814.

Preslav Nakov and Hwee Tou Ng. 2009. Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 1358–1367. https://www.aclweb.org/anthology/D09-1141.

Yurii Nesterov. 1983. A Method for Unconstrained Convex Minimization Problem with the Rate of Convergence o$(1/k^2)$. *Doklady Akademii Nauk SSSR* 269:543–547.

Mariana Neves, Antonio Jimeno Yepes, Aurelie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 Biomedical Translation Shared Task. In *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Brussels, Belgium.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pages 160–167. http://dx.doi.org/10.3115/1075096.1075117.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19–51.

Ibrahim Burak Özyurt, Anita E. Bandrowski, and Jeffrey S. Grethe. 2020. Bio-AnswerFinder: a System to find Answers to Questions from Biomedical Texts. *Database J. Biol. Databases Curation* 2020. https://doi.org/10.1093/database/baz137.

Gustavo Paetzold and Lucia Specia. 2016. SimpleNets: Quality Estimation with Resource-Light Neural Networks. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*. Berlin, Germany, WMT, pages 802–808. http://www.aclweb.org/anthology/W/W16/W16-2387.

Koushik Pahari, Alapan Kuila, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay, and Josef van Genabith. 2016. JU-USAAR: A Domain Adaptive MT System. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 442–448. http://www.aclweb.org/anthology/W/W16/W16-2333.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Seonyeong Park, Hyosup Shim, and Gary Geunbae Lee. 2014. ISOFT at QALD-4: Semantic Similarity-based Question Answering System over Linked Data. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*. CEUR-WS.org, volume 1180 of *CEUR Workshop Proceedings*, pages 1236–1248. http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-ParkEt2014.pdf.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Ales Tamchyna, Andy Way, and Josef van Genabith. 2015. Domain Adaptation of Statistical Machine Translation with Domain-focused Web Crawling. *Lang. Resour. Evaluation* 49(1):147–193. https://doi.org/10.1007/s10579-014-9282-3.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830. https://scikit-learn.org/stable/index.html.

Ruggero Petrolito and Felice Dell'Orletta. 2018. *Word Embeddings in Sentiment Analysis*, pages 330–334. https://doi.org/10.4000/books.aaccademia.3589.

Barbara Plank. 2011. *Domain Adaptation for Parsing*. Groningen Dissertations in Linguistics 96.

Boris T. Polyak. 1964. Some Methods of Speeding up the Convergence of Iteration Methods. *USSR Computational Mathematics and Mathematical Physics* 4(5):1–17. http://www.sciencedirect.com/science/article/pii/0041555364901375.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of Machine Translation Models with Back-translated Data using Transductive Data Selection Methods. *CoRR* abs/1906.07808. http://arxiv.org/abs/1906.07808.

Martin Popel, Roman Sudarikov, Ondřej Bojar, Rudolf Rosa, and Jan Hajič. 2016. TectoMT – a Deep Linguistic Core of the Combined Chimera MT System. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*. Baltic Journal of Modern Computing, Riga, Latvia. https://www.aclweb.org/anthology/2016.eamt-2.1.

Maja Popović and Nikola Ljubešić. 2014. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*. Association for Computational Linguistics, Doha, Qatar, pages 76–84. https://doi.org/10.3115/v1/W14-4210.

Michael James David Powell. 1977. Restart Procedures for the Conjugate Gradient Method. *Mathematical programming* 12(1):241–254.

Sudharsan Ravichandiran. 2021. *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. `http://is.muni.cz/publication/884893/en`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pages 3980–3990. https://doi.org/10.18653/v1/D19-1410.

Rudolf Rosa, Roman Sudarikov, Michal Novák, Martin Popel, and Ondřej Bojar. 2016. Dictionary-based Domain Adaptation of MT Systems without Retraining. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 449–455. http://www.aclweb.org/anthology/W/W16/W16-2334.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 1–11. http://www.aclweb.org/anthology/W14-3301.

Baskaran Sankaran, Majid Razmara, Atefeh Farzindar, Wael Khreich, Fred Popowich, and Anoop Sarkar. 2012. Domain adaptation techniques for machine translation and their evaluation in a real-world setting. In *Canadian Conference on AI*. Springer, volume 7310 of *Lecture Notes in Computer Science*, pages 158–169.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, pages 539–549. https://www.aclweb.org/anthology/E12-1055.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Valerio Miceli Antonio Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* pages 65–68.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. https://www.aclweb.org/anthology/P16-1162.

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27(3):379–423. https://ieeexplore.ieee.org/document/6773024/.

Weijia Shi, Muhao Chen, Yingtao Tian, and Kai-Wei Chang. 2019. Learning Bilingual Word Embeddings Using Lexical Definitions. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Association for Computational Linguistics, Florence, Italy, pages 142–147. https://doi.org/10.18653/v1/W19-4316.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting In-domain Training Corpora for Neural Machine Translation Using Data Selection Methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 224–231. https://www.aclweb.org/anthology/W18-6323.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of Translation Edit Rate with Targeted Human Annotation. In *In Proceedings of Association for Machine Translation in the Americas*. pages 223–231.

Felipe Soares and Karin Becker. 2018. UFRGS Participation on the WMT Biomedical Translation Shared Task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 673–677. http://www.aclweb.org/anthology/W18-6449.

Karen Sparck Jones. 1972. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, Taylor Graham Publishing, pages 132–142.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15(1):1929–1958. http://dl.acm.org/citation.cfm?id=2627435.2670313.

Andreas Stolcke. 2002. SRILM-an Extensible Language Modeling Toolkit. In *Interspeech*. volume 2002.

Jinsong Su, Deyi Xiong, Yang Liu, Xianpei Han, Hongyu Lin, Junfeng Yao, and Min Zhang. 2015. A Context-Aware Topic Model for Statistical Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 229–238. https://doi.org/10.3115/v1/P15-1023.

Roman Sudarikov, Martin Popel, Ondrej Bojar, Aljoscha Burchardt, and Ondrej Klejch. 2016. Using MT-ComparEval. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173.

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting Data for English-to-Czech Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '12, pages 374–381. http://dl.acm.org/citation.cfm?id=2393015.2393068.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1400–1410.

Cornelis Joost van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Association for Machine Translation in the Americas, Boston, MA, pages 193–199. https://www.aclweb.org/anthology/W18-1819.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., volume 30, pages 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Pradeepika Verma and Anshul Verma. 2020. A Review on Text Summarization Techniques. *Journal of Scientific Research* 64:251–257. https://doi.org/10.37398/JSR.2020.640148.

Longyue Wang, Yi Lu, Derek F. Wong, Lidia S. Chao, Yiming Wang, and Francisco Oliveira. 2014. Combining Domain Adaptation Approaches for Medical Text Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 254–259. http://www.aclweb.org/anthology/W/W14/W14-3331.

Zhiwei Wang, Yao Ma, Zitao Liu, and Jiliang Tang. 2019. R-Transformer: Recurrent Neural Network Enhanced Transformer. *CoRR* abs/1907.05572. http://arxiv.org/abs/1907.05572.

Krzysztof Wolk and Krzysztof Marasek. 2014. Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs. In *Procedia Technology, 18*. Elsevier, pages 126–132. https://10.1016/j.protcy.2014.11.024.

Krzysztof Wolk and Krzysztof Marasek. 2017. PJIIT's systems for WMT 2017 Conference. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 416–421. http://www.aclweb.org/anthology/W17-4743.

Deyi Xiong, Fandong Meng, and Qun Liu. 2016. Topic-based Term Translation Models for Statistical Machine Translation. *Artificial Intelligence* 232:54–75. http://www.sciencedirect.com/science/article/pii/S0004370215001782.

Jia Xu, Yi Zong Kuang, Shondell Baijoo, Jacob Hyun Lee, Uman Shahzad, Mir Ahmed, Meredith Lancaster, and Chris Carlan. 2017. Hunter MT: A Course for Young Researchers in WMT17. In *Proceedings of the Second*

Conference on Machine Translation, Volume 2: Shared Task Papers. Association for Computational Linguistics, Copenhagen, Denmark, pages 422–427. http://www.aclweb.org/anthology/W17-4744.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of Selecting Training Data to build a Compact and Efficient Translation Model. IJCNLP, pages 655–660. https://www.aclweb.org/anthology/I08-2088.pdf.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1393–1398. https://www.aclweb.org/anthology/D13-1141.

Eidesstattliche Versicherung
Declaration on oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare, on oath, that I have written the present dissertation independently and have not used further resources and aids than those stated.

Hamburg, 2021