Automatic Detection of Translation Direction

Ilia Sominsky

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Science
Department of Computer Science

Fabruary 2019

Automatic Detection of Translation Direction

By: Ilia Sominsky

Supervised by: Prof. Shuly Wintner

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Science
Department of Computer Science

Fabruary 2019

Approved by: ______________________ Date: ______________________

(Supervisor)

Approved by: ______________________ Date: ______________________

(Chairperson of Master's studies Committee)

Acknowledgments

Contents

Automatic Detection of Translation Direction

Ilia Sominsky

Abstract

Parallel corpora are crucial resources for NLP applications, most notably for machine translation. The direction of the (human) translation of parallel corpora has been shown to have significant implications for the quality of statistical machine translation systems that are trained with such corpora. Determining the translation direction of parallel corpora is therefore an important task.

We describe a method for determining the direction of the (manual) translation of parallel corpora *at the sentence-pair level*. Using several linguistically-motivated features, coupled with a neural network model, we obtain high accuracy on several language pairs. Furthermore, we demonstrate that the accuracy is correlated with the (typological) distance between the two languages. Finally, we show that our method can be used for improving not only the quality of statistical machine translation, but also of neural machine translation.

List of Tables

List of Figures

Chapter 1

Introduction

Parallel corpora are used for various purposes, including for training and evaluation of statistical machine translation (SMT) systems (Koehn, 2010). While traditional SMT systems are agnostic with respect to the direction in which the parallel corpora they are trained on were (manually) translated, several studies have shown that taking directionality into account when training SMT systems has a significant effect on the quality of the translation (Kurokawa et al., 2009; Lembersky et al., 2012, 2013; Twitto-Shmuel et al., 2015). In this paper we show the same effect also holds for neural machine translation (NMT) systems.

We address the task of determining the direction of translation given a parallel text; this is cast as a binary classification task. To strain the classifier, we focus on retaining high accuracy when the size of text chunks to be classified is minimal: single sentence pairs. We define sets of features that reflect insights drawn from Translation Studies regarding the special properties of translated texts, and in particular the *asymmetric* nature of translation (Baker, 1993; Toury, 1980, 1995). These include the tendency of translated texts to be simpler (Baker, 1993; Blum-Kulka and Levenston, 1983; Laviosa, 1998, 2002; Vanderauwerea, 1985); the tendency of translators to explicate the source text (Baker, 1993; Blum-Kulka, 1986); the different distributions of various statistical phenomena (e.g., the frequencies of function words or certain syntactic structures) between the source and the translation (Blum-Kulka, 1986; Gellerstam, 1986; Koppel and Ordan, 2011; Øverås, 1998); and *interference* of language constructions from the source to the target (Teich, 2003; Toury, 1979).

The contribution of this paper is manifold.

(1) First and foremost, we introduce a method for accurately determining the translation direction of sentence pairs in parallel corpora; the method is based on the introduction of several new, linguistically motivated, types of features for this task. We show that the combination of these features outperforms the previous state-of-the-art in detection of translation direction.

(2) Furthermore, we demonstrate the robustness of our method by evaluating it on several language pairs and on three different datasets.

(3) From a theoretical perspective, this work corroborates the intuitive hypothesis that the translation detection task is easier when the two languages involved are typologically more distant.

(4) Finally, we demonstrate that our method can indeed be used for improving the quality of both statistical and neural machine translation systems.

After reviewing related work in the next section, we describe the experimental setup of this work in Section 3, and the features we used in Section 4. The results are presented and discussed in Section 5. We conclude with suggestions for future research.

Chapter 2

Related work

The differences between original and translated texts have been a major field of investigation in Translation Studies (Baker, 1995; Toury, 1980, 1995). Translated texts have unique characteristics that set them apart from texts originally written in the same language. These are not necessarily artifacts of poor translation; rather, they reflect different statistical distributions across the two genres. The sub-language of translated texts (in any language) was referred to as *translationese* (Gellerstam, 1986). The unique properties of translationese are attributed to various reasons, some of which are considered "universal" (e.g., translated texts tend to simplify the original message; they tend to use more standard language than originals), while others are related to *interference*, namely the "fingerprints" of the source language found in the translation product.

Distinguishing between original and translated texts is a classic text classification task that has been extensively addressed both with supervised machine learning (Avner et al., 2016; Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011; Kurokawa et al., 2009; van Halteren, 2008; Volansky et al., 2015) and with unsupervised methods (Nisioi, 2015; Rabinovich and Wintner, 2015; Rabinovich et al., 2016a). The main challenge, as is usually the case in text classification, lies in the choice of features with which text chunks are represented. For the task at hand, features frequently used include function words (FW), character n-grams, part-of-speech (POS) n-grams, special sets of words such as discourse markers, etc. With the right choice of features, accuracies can reach almost ceiling levels, depending on the dataset involved.

However, all of the above-mentioned works used larger chunks of text, typically 2,000 tokens, as the classification unit. This is a rather unrealistic scenario, since parallel data available online may include much fewer sentence pairs. The accuracy of identifying translationese has been shown to drop significantly when the size of the text chunk used for classification decreases (Rabinovich and Wintner, 2015). One of our goals in this work, therefore, is to improve the accuracy of translationese detection systems with much smaller text chunks.

Previous research focused on identifying translationese in monolingual texts. However, in realistic scenarios, parallel texts are available and the actual task is to determine the *direction of translation* given texts in *two* languages. For such tasks one can use features drawn from each of the two languages, as well as from the alignments between words and phrases in the two texts. This task was first defined by Eetemadi and Toutanova (2014), who used the Canadian Hansard corpus of parallel texts in English and French.

The motivation stems from the observation that linguistic structures tend to have different distributions in original and translated texts. Therefore, assessing the frequencies of syntactic structures in two parallel texts, especially for text chunks that are aligned with each other across two parallel sentences, may shed light on the direction of the translation. As base structures, Eetemadi and Toutanova (2014) used *minimal translation units (MTUs)*, defined as pairs of source and target word sets that satisfy two conditions: (i) no alignment links exist between distinct MTUs; (ii) MTUs are not decomposable into smaller MTUs without violating the previous rule. Once MTUs were identified, each word was replaced by its POS tag, thereby creating POS-MTUs. These are the structures used as features.

As an example, consider the two aligned English–French sentences in Figure 2-1; they yield the following POS-MTUs: [PP]↔[PRO:per], [VVP, TO]↔[VER:cond], [VV]↔ [VER:infi], and [PP]↔[PRO:per]. More specifically, the POS-MTU [VVP, TO]↔[VER:cond] reflects the fact that English word pairs such as ‘*want to*’ translate to French verbs in the conditional form, e.g., ‘*voudrais*’. Incidentally, this mapping is much more common, by a factor of 10, in English-to-French translations than in the reverse direction.

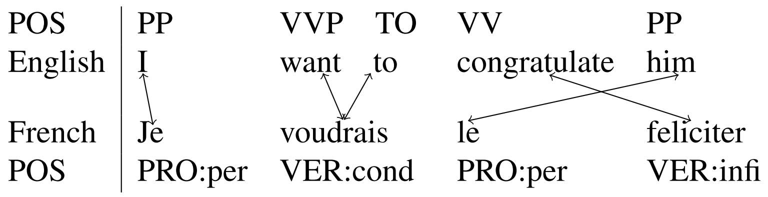

Figure 2-1: POS-MTUs, English–French

As another example, the two aligned English–German sentences depicted in Figure 2-2 yield the following POS-MTUs: [CD]↔[PIS], [IN]↔[ART], [NP]↔[ADJA], [RB, JJS]↔[ADJA], [NNS]↔[NN]. In particular, the POS-MTU [RBS, JJ]↔[ADJA] reflects the fact that English word pairs such as *'most famous'* translate to German adjectives in the superlative form, e.g., *'berühmtesten'*.

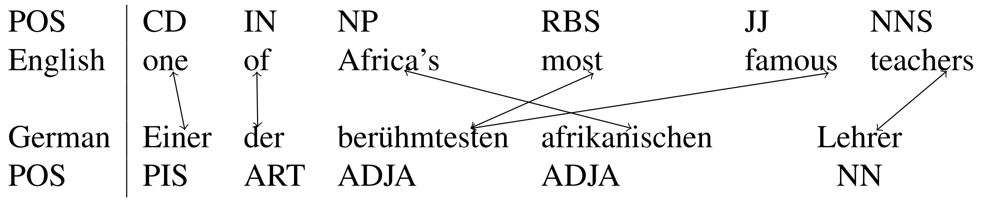

Figure 2-2: POS-MTUs, English–German

Eetemadi and Toutanova (2014) do not provide sufficient details that would enable replication of their results, but they report 71% accuracy with these features.

In a subsequent work, Eetemadi and Toutanova (2015) used Brown clusters (Brown et al., 1992), a method of clustering words according to syntactic and semantic relatedness, instead of POS tags. With *Brown cluster MTUs* as features, they reached 80% precision and 85% recall on the Hansard corpus. This is the present state of the art for this task.

Chapter 3

Methodology

Task Given a sentence pair in a parallel corpus, our task is to identify the direction of translation, thereby determining the source and the target sentences. We approach this task as a corpus-based text classification task based on supervised machine learning; the main challenge is to define a set of features that will yield the best accuracy.

Datasets We used sentence-aligned parallel corpora from three resources: the Canadian parliamentary proceedings (Hansard), with English–French sentence pairs; Europarl (Koehn, 2005), the proceedings of the European Parliament, where English is aligned with French and German; and the UN parallel corpora (Ziemski et al., 2016), in which English is aligned with Arabic, French, German, Russian and Spanish. We used subsets of these corpora in which the direction of translation has been accurately annotated (Kurokawa et al., 2009; Rabinovich et al., 2016b; Tolochinsky et al., 2018). We cleaned the data by removing editor's comments and sentences with fewer than 5 tokens. We then down-sampled the corpora and extracted equally-sized subsets with 50,000 sentence-pairs in each language pair, distributed evenly across translation direction. These are the data we used in all the experiments described below. Details on the available data are presented in Table 3.1.

Prepossessing We preprocessed the data as follows. First, all words in the two languages were tagged for part of speech using FARASA (Abdelali et al., 2016) for Arabic and TreeTagger (Schmid, 1995) for the other languages. Second, all the sentence pairs were word aligned using FastAlign (Dyer et al., 2013). With the word alignments we were able to extract the

	Europarl		**UN**				**Hansard**
	EN-FR	EN-DE	EN-FR	EN-ES	EN-RU	EN-AR	EN-FR
EN original	217	225	8100	6100	3600	4087	3377
EN original, cleaned	215	222	6600	5100	2800	3338	2981
EN translated	130	155	773	447	107	88	744
EN translated, cleaned	128	153	683	381	91	65	678

Table 3.1: Dataset sizes (in thousands of sentence-pairs)

features that will be explained in the next section.

Classification For the task of identifying the translation direction, we implemented various feature sets and used them for training a Logistic Regression classifier (with the implementation of Pedregosa et al. (2011)), mainly because it is faster yet no less accurate than SVM. We performed ten-fold cross-validation for evaluation and report accuracy in %. As our datasets are balanced, the trivial baseline is 50%.

Chapter 4

Features

We defined several novel features motivated by various insights from Translation Studies. We motivate and explain these feature in this section.

Baseline As a baseline, we implemented some of the features that were suggested by Volansky et al. (2015), including:

POS trigrams We used the frequencies of the 2000 most frequent POS trigrams for each language.

Function words Function words for many languages are available online. We used the frequencies of all the function words in each language (between 160 in Arabic and 600 in German).

Positional token frequency In different languages, the choice of words with which sentences begin is rather different, and is more constrained and formulaic than elsewhere in the sentence (Volansky et al., 2015). A clear example is greetings: parliament speakers may choose to begin their speeches by *'Ladies and gentlemen'*, but this turns out to be much more common in French than in English. We used the frequencies of words that occur in the first, second, penultimate and last positions in the sentences.

MTUs Finally, to compare with the state of the art, we also computed POS-MTUs and Brown Cluster MTUs (Eetemadi and Toutanova, 2015).

Word rank The *simplification hypothesis* conjectures that translated texts tend to be simpler than originals. As one realization of this hypothesis, we assume that translations would use more common, frequent words than originals. In order to determine how common each word is, we use pre-trained frequency lists in all languages.[1]

Comparing the actual (frequency-based) ranks of word forms across languages is rather problematic, especially when the morphologies of the languages differ. (e.g., when one language has many more inflected forms per lexeme than the other). Therefore, we split the word frequency lists to seven *bins* that group together words by their frequency, and compared the bins rather than the actual ranks.[2] The first bin includes words whose accumulated frequency is up to 0.25; it includes the most frequent words in each language. The other bins include words with accumulated frequency up to 0.5, 0.7, 0.8, 0.88, 0.95 and all the rest. This facilitates comparison of words in the same frequency brackets across two different languages. This feature defines 14 bins (7 for each language); its actual value is number of words in each bin.

Additionally, we compared the (frequency-based) ranks of aligned word pairs. Given a pair of aligned sentences, consider the difference in rank between each pair of aligned words. We hypothesize that such differences would depend on the translation direction (as rarer words tend to be translated to more common ones). For example, we expect the English *'however'* (ranked 236th) to be typically translated to French *'mais'* (ranked 33rd), but French *'mais'* to be more often translated to English *'but'* (ranked 23rd).

To implement this observation, we defined a histogram representing the values of the differences in rank between pairs of aligned words in each sentence pair. For example, if the English word *'however'* is ranked 236th and its aligned French word *'mais'* is ranked 33rd, we used the value $236 - 33 = 203$. We computed these values for all the aligned words in a sentence-pair; we then used the highest and lowest values as the boundaries of a histogram and split it to 12 bins. For example, if the defined limits of the histogram are: [-100000, -50000, -25000, -8000, -4000, -300, 300, 4000, 8000, 25000, 50000, 100000] and the values of differences between the words in a sentence pair are -10953, -511, 402, -3159, 4099, 11267, 10535, 80, 4280, 345; then the resulting histogram is: [0, 0, 0, 1, 0, 2, 1, 2, 2, 2, 0 ,0]. The values of this feature for a given pair of sentences are the values of each bin in the resulting histogram.

Lexically-Anchored-POS-MTUs While POS-MTUs identify meaningful linguistic structures, they are too general and may lose important nuances of the correspondences between

[1] We used Michel et al. (2010) for all the languages in our dataset.

[2] The number of bins and their frequency ranges were determined empirically.

constructions in the two languages. For example, consider the POS-MTUs [IN]↔[ART] in Figure 2-2: clearly it is not the case that prepositions in English translate to determiners in German. However, it is reasonable to assume that the English genitive preposition *'of'* will be aligned to a German genitive article such as *'der'*. To reflect this notion, and define finer, subtler cross-language correspondences, we propose *Lexically-Anchored-POS-MTUs* (LA-POS-MTUs): we only replace *content* words by their POS tag, leaving *function* words intact. The values of these features are the actual counts of each LA-POS-MTU in the sentences. Similarly to POS-MTUs, they are distributed differently in each of the translation directions.

As an example, consider the LA-POS-MTUs in Figure 4-1: [one]↔[einer], [of]↔[der], [NP]↔[ADJA], [most, JJ]↔[ADJA], [NNS]↔[NN]. In particular, the LA-POS-MTU [most, JJ]↔[ADJA] reflects the fact that in English, some superlative adjectives can come with the adverb *'most'* or with *'est'* as a suffix, while in German there is only one form: adding a suffix to the adjective. Indeed, the LA-POS-MTU [most, JJ]↔[ADJA] is much more frequent in English to German than in the reverse direction. This is presumably an instance of *interference* of German on the English translation product. While in English there are two ways to form the superlative, and sometimes both are valid (e.g., *'most clever'* and *'cleverest'*), German has only one possible form. When a superlative adjective is translated from German to English, the translator may tend to keep it with the suffix (if possible), rather than splitting it into two words. Hence, this LA-POS-MTU is more frequent in the English to German direction.

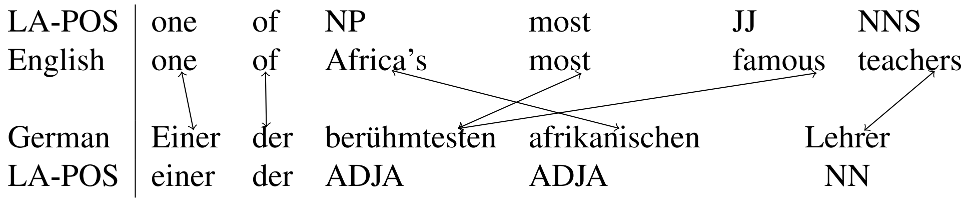

Figure 4-1: LA-POS-MTUs

Syntactic structure The simplification hypothesis implies that the structure of translated sentences tends to be simpler than that of originals. We therefore parsed the corpus with universal dependencies (Straka and Straková, 2017) and defined several measures that supposedly reflect sentence complexity: the height of the dependency tree; its depth; and the average number of dependents per word. In addition, we used dependency tag trigrams as features, similarly to POS-trigrams.

Back translation Translated texts carry a unique signal; the challenge is to identify this signal at the sentence-pair level, where it may be extremely subtle. The motivation for the back translation feature is to amplify this signal.

To do so, we use machine translation (Google Translate) to translate the sentences again. Given a sentence pair $\langle e_1, f_1 \rangle$, we machine-translate both sentences, yielding the pair $\langle f_2, e_2 \rangle$, where $f_2 = MT(e_1)$ and $e_2 = MT(f_1)$, MT indicating machine translation. Now assume, without loss of generality, that e_1 is the original; hence f_1 is its manual translation, namely $f_1 = HT(e_1)$, where HT indicates human translation. Therefore, $e_2 = MT(f_1) = MT(HT(e_1))$. In other words, e_2 is "twice removed" from e_1, being translated once by a human and once automatically. In contrast, $f_1 = HT(e_1)$ and $f_2 = MT(e_1)$; both f_1 and f_2 are only "once removed" from e_1: f_1 was translated manually and f_2 automatically, but only once. Therefore, we expect f_1 and f_2 (two French sentences) to be closer to each other than e_1 and e_2 (two English sentences) are. This is only the case if f_1 is the translation of e_1; if the translation direction is reversed, we would expect e_1 and e_2 to be closer to each other than f_1 and f_2 are.

To measure the similarity between the two sentences we used three metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and Levenshtein distance (Levenshtein, 1965). Each metric results in two scores: one for the distance between the two English sentences and one for the two French sentences. These six scores were used as features for the classifier.

Neural network In addition to the classifiers described above, we also approached the task of determining translation direction with a neural network. The input of the network is the two sentences, where the words are mapped to pre-trained word embedding vectors of 50 dimensions (we used Pennington et al. (2014) for English and Bojanowski et al. (2017) for the other languages.) The network consists of one Bi-directional Long Short-Term Memory (BiLSTM) layer with 100 units, followed by a fully connected layer with a single output; the loss is defined as binary cross-entropy (the network was implemented with Keras.)

Chapter 5

Results

Table 5.1 depicts the accuracy of 10-fold cross validation evaluation of classifiers reflecting the various features. The "All" column indicates a dataset constructed from the French–English sentence pairs in all the three different corpora; it is therefore a heterogenous dataset, which makes the task much more challenging (Rabinovich and Wintner, 2015). Indeed, the results on this dataset are worst, lower than each individual dataset in isolation. Still, even for this challenging experimental scenario, our best classifier achieves over 72% accuracy.

Corpus	**Europarl**		**UN corpus**				**Hansard**	**All**
Feature set	EN-FR	EN-DE	EN-FR	EN-ES	EN-RU	EN-AR	EN-FR	EN-FR
POS-MTUs	64.4	63.1	63.4	62.6	69.2	76.2	62.7	58.1
LA-POS-MTUs	65.6	66.2	63.4	64.0	68.4	75.2	64.8	59.9
Brwn Clstr MTUs	73.0	67.1	66.4	68.3	71.9	79.0	64.8	60.3
Rank	63.5	64.8	58.0	59.0	60.8	65.2	56.6	56.0
POS-trigrams	65.0	65.7	64.0	63.2	67.0	74.3	64.1	59.6
Function words	65.6	68.0	66.3	66.1	72.3	69.0	66.5	56.6
Pos. token freq.	62.0	64.7	65.9	66.7	76.0	80.8	64.2	61.0
Syntactic structure	64.0	62.0	65.0	63.3	68.6	67.0	61.4	58.8
Back translation	61.2	58.5						
All	**81.0**	**78.1**	**75.6**	**78.0**	**84.5**	**90.1**	**75.1**	**67.9**
Neural network	81.0	80.9	79.8	84.8	90.8	89.0	78.4	74.6
Stacking	**83.0**	**82.3**	**80.3**	**84.9**	**91.1**	**90.0**	**76.5**	**72.1**

Table 5.1: Results: accuracy (%)

The "All" row indicates the concatenation of all features into one feature vector. Since these features encode different aspects of the relations between the two languages, we believe that they are at least partially independent. Indeed, the results of feature combination support this assumption.

The signal of translationese is indeed subtle, but the results show that many of our basic classifiers are able to detect it, albeit to a small extent. For most language pairs and datasets, each of the feature sets we defined yielded accuracy of over 60%, sometimes over 70%, and reaching 80% in a few cases. Brown cluster MTUs, which were used by the state of the art (Eetemadi and Toutanova, 2015), are indeed a good feature set. MTUs based on Brown clusters turned out to be better than LA-POS-MTUs; presumably, Brown clusters encode lexical semantic information that is helpful for the task. However, they are outdone in more than half of the cases by simpler features such as function words or positional token frequencies.

Back translation turned out to be a less beneficial feature than we have expected on Europarl. As it is a computation-intensive feature, we refrained from computing it on the other datasets.

Combining features together yielded a sizable boost in accuracy, advancing the state of the art to the area of 80-90% accuracy in all cases.

The features that we defined are obviously not mutually independent; it therefore makes sense to try some dimensionality reduction method to remove redundant features. In order to obtain the best accuracy, we tried several dimensionality reduction methods, with various dimensionalities. Principal Component Analysis (PCA) (Jolliffe, 2002) and Singular Value Decomposition (SVD) (Deerwester et al., 1990) do not use the labels of the sentences for dimensionality reduction. They both produce similar results with some advantage for SVD since it is more suitable for sparse matrices (specifically, MTUs result in very sparse feature vectors). Finally, we used an algorithm that selects the k features with the highest scores. This method produced better results in cross-validation only when using smaller data; when training on the full dataset this method did not improve the results.

Evidently (and frustratingly), the accuracy of the neural network is higher than feature combination in all cases but one; yet we suspect that the features capture phenomena that are not reflected by the neural network. To test that, we used *stacking*. We defined three different classifiers: one with features computed from the English texts only (rank, POS trigrams, function words, positional tokens, and syntactic structure); another with the same features computed from the other language; and a third from the alignment features computed from both languages (the three MTU feature types). We additionally trained the neural network. We then used all four classifiers to predict the direction of translation, and used their confidence scores as features for a stacked classifier, whose prediction is the class we use. The results are listed in Figure 5.1 under “Stacking”, and show a small but consistent improvement for all language pairs.

Still, the BiLSTM turned out to be better for the Hansard corpus and for the mixed dataset. We do not have a clear explanation for this outcome. We used paired t-test to determine the statistical significance of the improvement in results between using all the features (“All”) and the best results obtained by Stacking. The test yielded p-values $<$0.001 for all language pairs except English–Arabic. Similarly, comparing the neural network with Stacking in the same way, the test yielded p-values $<$0.001 in all language pairs except English–Spanish. We thus conclude that the generalizations of the neural network are, at least to some extent, different from the features we defined.

Finally, observe that the results clearly support our theoretical hypothesis: the accuracy of the classification improves when the two languages involved are more typologically distant. The task is particularly hard for English-French and English-German, and easiest for English-Arabic and English-Russian. We tentatively conclude, therefore, that translationese is more pronounced, and interference is more powerful, when the two languages are more distant. This chimes in with recent results that show the relationships between interference and language typology (Rabinovich et al., 2017).

5.1 Adversarial learning

Our in-domain testing yielded high accuracy results, but unfortunately, the accuracy when testing out of domain decreased significantly. For example, using all of the features when training on Hansard and testing on Europarl, the model accuracy is only 56.1, while testing in-domain the accuracy is 75.1. The deterioration in accuracy was observed both in the feature-based classification and in the neural network model. We attribute this outcome to the fact that the languages of the different domains have different styles; for example the English used in the European Parliament is different from the one used in the Canadian parliament.

To circumvent this problem, we designed an adversarial neural network that learns weights in such way so as to maximize the accuracy on the main task but at the same time minimize the effect of domain-specific features. We changed the neural network model described above only in the definition of the final output layer: instead of a single output, the network will output two values. The first output will predict the direction of translation, while the second will predict the domain. As the objectives of the two outputs is different, we defined different loss functions for the two outputs. In the learning process, the weights of the network are expected to change according to these loss functions. In the default case, the objective of the network is to minimize both of the loss functions, whereas in this case the network will learn to classify correctly both the direction and the domain. Since we want the network to learn to correctly predict only the direction of translation, we defined for the final network the following loss function, which is the sum:

$$Loss = L_{direction} + (-L_{domain})$$

The objective of the network is to minimize the direction loss (i.e., the error in learning the correct translation direction) and to maximize the domain loss (i.e., refrain from learning the domain class). In that way, if a feature that is specific to some domain contributes to predicting the direction classification, the network should assign it a lower weight.

We tested the adversarial network on our three data sets, focusing on English-to-French translations. The results are shown in table 5.2. Our conclusion from this experiment is that some differences between original and translated texts are specific to each domain. These differences can be observed and learned, but they do not completely transfer to other domains. There are some mutual differences across the domains but they are not powerful enough to detect translations with sufficiently high accuracy.

Train		**Test**	
	UN	Europarl	Hansard
UN+EU	75.3	72.2	**59.1**
UN+Hansard	74.6	**61.7**	68.6
EU+Hansard	**56**	72.7	68.7

Table 5.2: Results: Adversarial learning

Chapter 6

Applications for machine translation

This work was partly motivated by previous research that suggested that *statistical* machine translation can be improved by training on source-translated-to-target corpora rather than target-translated-to-source texts (Kurokawa et al., 2009; Lembersky et al., 2013; Twitto-Shmuel et al., 2015). In this section we verify that such benefits hold also for *neural* machine translation (NMT). We used French–English data from the three corpora (Hansard, Europarl and UN). The total data that was available to us consisted of 1.6 million sentences annotated as French original, and 11.7 million sentences annotated as English original. Focusing on translating French to English, we trained three different NMT systems using Marian (Junczys-Dowmunt et al., 2018). In one system (FO), the training material consisted only of French original sentence pairs; in the other (EO), we only used English original sentence pairs; and in the third (MIX), we mixed equal portions of both. In all three cases we used an equal number of sentence pairs (1.6 million). We tested the three NMT systems on a reference set of 10,000 sentences taken from French original data, following the methodology of Lembersky et al. (2013). We evaluated the quality of the resulting NMT systems by comparing BLEU, METEOR and TER scores using MaltEval (Clark et al., 2011).

The results, listed in table 6.1, clearly corroborate our hypothesis: for the task of French to English translation, training data that were manually translated from French to English yield much better NMT systems than training data that were translated in the reverse direction.

Train Data	BLEU↑	METEOR↑	TER↓
FO	41.0	38.4	46.1
MIX	38.2	36.7	48.5
EO	34.4	35.0	52.8

Table 6.1: Accuracy of NMT systems

Chapter 7

Conclusion

We have shown that linguistically-motivated features, based on Translation Studies insights pertaining to the asymmetry of the translation process, can yield high, state-of-the-art accuracy on the task of translation direction detection. We introduced several novel features and used stacking to produce highly accurate sentence-pair-level classifiers for five language pairs. We also confirmed the hypothesis that this task is harder the more closely-related the two languages involved are. Finally, we showed that these results can be used to improve the accuracy of neural machine translation systems.

In future work, we intend to provide a deeper analysis of the results, focusing on the constructions whose frequencies differ most across the two languages. We would also like to evaluate our systems cross-domain, as it has been shown (Rabinovich and Wintner, 2015) that the signal of translationese is subtle, and can be overshadowed by signals of the datasets used for training and testing. Finally, and depending on the availability of datasets, we would like to extend the experiments described herein to more language pairs.

Bibliography

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-3003.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, 31(1):30–54, April 2016. URL http://dx.doi.org/10.1093/llc/fqu047.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 1993.

Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, September 1995.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W05/W05-0909.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September 2006. URL http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies*, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*, pages 119–139. Longman, 1983.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL http://aclweb.org/anthology/Q17-1010.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based N-gram models of natural language. *Computational Linguistics*, 18(4): 467–479, 1992. ISSN 0891-2017.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 176–181, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL http://dl.acm.org/citation.cfm?id=2002736.2002774.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990. URL citeseer.ist.psu.edu/deerwester90indexing.html.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-1073.

Sauleh Eetemadi and Kristina Toutanova. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 159–164. Association for Computational Linguistics, October 2014. URL http://www.aclweb.org/anthology/D14-1018.

Sauleh Eetemadi and Kristina Toutanova. Detecting translation direction: A cross-domain study. In *NAACL Student Research Workshop*. ACL – Association for Computational Linguistics, June 2015. URL http://research.microsoft.com/apps/pubs/default.aspx?id=249114.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL http://dx.doi.org/10.1007/978-3-642-12116-6.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2nd edition, 2002.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-4020.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86. AAMT, 2005. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 2010. ISBN 0521874157, 9780521874151.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1132.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88, 2009.

Sara Laviosa. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta*, 43(4):557–570, December 1998.

Sara Laviosa. *Corpus-based translation studies: theory, findings, applications*. Approaches to translation studies. Rodopi, 2002. ISBN 9789042014879.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825, December 2012. URL http://dx.doi.org/10.1162/COLI_a_00111.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4): 999–1023, December 2013. URL http://dx.doi.org/10.1162/COLI_a_00159.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Holberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 2010. URL http://www.sciencemag.org/content/331/6014/176.full.

Sergiu Nisioi. Unsupervised classification of translated texts. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane, and Elisabeth Métais, editors, *Natural Language*

Processing and Information Systems: Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB, volume 9103 of *Lecture Notes in Computer Science*, pages 323–334. Springer, June 2015. URL http://dx.doi.org/10.1007/978-3-319-19581-0_29.

Lin Øverås. In search of the third code: An investigation of norms in literary translation. *Meta*, 43(4):557–570, 1998.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Ella Rabinovich and Shuly Wintner. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432, 2015. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/618.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1870–1881, August 2016a. URL http://aclweb.org/anthology/P/P16/P16-1176.pdf.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. A parallel corpus of translationese. In *Proccedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, April 2016b. URL http://arxiv.org/abs/1509.03611.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics, July 2017. URL http://aclweb.org/anthology/P17-1049.

Helmut Schmid. Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995.

Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Elke Teich. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, 2003.

Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. The UN parallel corpus annotated for translation direction. arXiv:1805.07697 [cs.CL], 2018. URL http://arxiv.org/abs/1805.07697.

Gideon Toury. Interlanguage and its manifestations in translation. *Meta*, 24(2):223–231, 1979.

Gideon Toury. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

Gideon Toury. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam / Philadelphia, 1995.

Naama Twitto-Shmuel, Noam Ordan, and Shuly Wintner. Statistical machine translation with automatic identification of translationese. In *Proceedings of WMT-2015*, 2015.

Hans van Halteren. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 937–944, 2008. ISBN 978-1-905593-44-6. URL http://www.aclweb.org/anthology/C08-1118.

Ria Vanderauwerea. *Dutch novels translated into English: the transformation of a 'minority' literature*. Rodopi, Amsterdam, 1985.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April 2015.

Michałl Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.