Automatic Detection of Translation Direction

Ilia Sominsky and **Shuly Wintner**
Department of Computer Science
University of Haifa
31905 Haifa, Israel

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. 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.

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. This is an extremely difficult task for humans, in most cases: a single sentence pair often does not reveal any obvious signal of which of the two sentences is the original. It is also a highly challenging task for machines: Figure 1 depicts a few English-French examples of sentence pairs whose translation direction none of our classifiers predicted correctly.

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 (Toury, 1980, 1995; Baker, 1993). These include the tendency of translated texts to be simpler (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993; Laviosa, 1998, 2002); the tendency of translators to explicate the source text (Blum-Kulka, 1986; Baker, 1993); the different distributions of various statistical phenomena (e.g., the frequencies of function words or certain syntactic structures) between the source and the translation (Gellerstam, 1986; Blum-Kulka, 1986; Øverås, 1998; Koppel and Ordan, 2011); and *interference* of language constructions from the source to the target (Toury, 1979; Teich, 2003).

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.[1] Importantly, these features help shed light on the characteristics of translated language. (2) Furthermore, we demonstrate the robustness of our method by evaluating it on several language pairs and on three different

[1] As we explain in Section 2, a direct comparison with the state of the art is problematic as not enough detail is provided in the original publications for us to replicate existing results.

Proceedings of Recent Advances in Natural Language Processing, pages 1131–1140,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_130

English→French	Now the question is , who's going to pay for it all ?
	La question est de savoir qui va payer .
French→English	Admit it and we will understand each other .
	Dites -le moi et on va bien se comprendre .
French→English	We should at least ensure that there is no need to produce many more reports .
	Il ne faudrait tout de même pas qu' il y ait besoin d' en faire de nombreux encore .

Figure 1: Some examples of sentence pairs with their translation direction

datasets. (3) We show that detecting the translation direction can indeed be used for improving the quality of both statistical and neural machine translation systems. (4) Finally, 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.

After reviewing related work in the next section, we describe our experimental setup in Section 4, and the features we used in Section 5. The results are presented and discussed in Section 6. We conclude with suggestions for future research.

2 Related Work

The differences between original and translated texts have been a major field of investigation in Translation Studies (Toury, 1980, 1995; Baker, 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 (Baroni and Bernardini, 2006; van Halteren, 2008; Kurokawa et al., 2009; Koppel and Ordan, 2011; Ilisei et al., 2010; Volansky et al., 2015; Avner et al., 2016) and with unsupervised methods (Rabinovich and Wintner, 2015; Nisioi, 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, the classification unit used in all the above-mentioned research was larger chunks of text, typically 2,000 tokens. 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 is to improve the accuracy of translationese detection systems with much smaller text chunks, as available parallel texts are not guaranteed to be long.

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 approach was taken 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

POS	PP	VVP	TO	VV	PP
English	I	want	to	congratulate	him
French	Je	voudrais		le	feliciter
POS	PRO:per	VER:cond		PRO:per	VER:infi

Figure 2: POS-MTUs, English–French

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; 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.

As another example, the two aligned English–German sentences depicted in Figure 3 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'*.

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.

3 Motivation

This work was partly motivated by previous research that demonstrated 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 three corpora (Hansard, Europarl and UN; see below). 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 MultEval (Clark et al., 2011).

The results, listed in Table 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 1: Accuracy of NMT systems with varying configurations of the training material

4 Methodology

Task Given a sentence pair in a parallel corpus, our task is to identify the direction of translation,

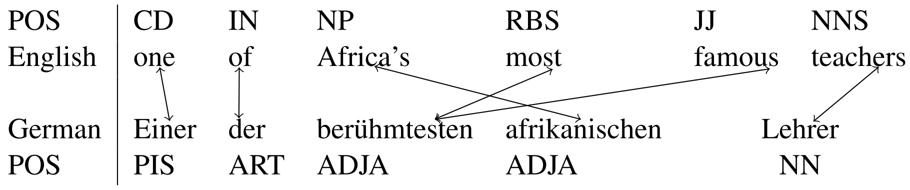

Figure 3: POS-MTUs, English–German

thereby determining the source and the target sentences. Our 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.[2] Details on the available data are presented in Table 2.

Preprossessing 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 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 tenfold cross-validation for evaluation and report accuracy in %. As our datasets are balanced, the trivial baseline is 50%.

Neural network In addition to the classifiers described below, we also approached the task of determining translation direction with a neural network. Our main goal here was to guarantee best performance, even the cost of interpretability. We used a network consisting 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.) The input of the network is the two sentences, where the words are mapped to pre-trained GloVe word embedding vectors of 50 dimensions (we used Pennington et al. (2014) for English and Bojanowski et al. (2017) for the other languages.)

5 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

[2]The only other parallel corpora that we are aware of where the direction of translation is marked are the Dutch Parallel Corpus (Macken et al., 2011), aligning Dutch with English abd French, and EuroParl-UdS (Karakanta et al., 2018), which largely overlaps with our dataset.

	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 2: Dataset sizes (in thousands of sentence-pairs)

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, as defined by Eetemadi and Toutanova (2014, 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 used pre-trained frequency lists in all languages (Michel et al., 2010).

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.[3] 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 resulting value from the 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 constructions in the two languages. For example, consider the POS-MTUs [IN]↔[ART] in Figure 3: 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

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

LA-POS	one	of	NP	most	JJ	NNS
English	one	of	Africa’s	most	famous	teachers
German	Einer	der	berühmtesten	afrikanischen		Lehrer
LA-POS	einer	der	ADJA	ADJA		NN

Figure 4: LA-POS-MTUs

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: [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.

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.

6 Results

Table 3 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. For the Europarl and UN

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**
BiLSTM	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 3: Results: accuracy (%) of predicting the translation direction

datasets, however, our results range between 80% and over 90% accuracy; given the difficulty of the task (refer back to Figure 1), we view this as a significant contribution.

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. We tried several dimensionality reduction methods, with various dimensionalities, but none yielded better results (on the full datasets) than using all features.

As could be expected, the accuracy of the BiLSTM 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 Table 3 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. In future work, we intend to consider new ways for incorporating linguistically-motivated features in neural network architecture, e.g., along the lines of Strubell et al. (2018).

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).

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 when the two languages involved are more closely related.

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.

Acknowledgements

This research was supported by Grant No. 2017699 from the United States-Israel Binational Science Foundation (BSF), by Grant No. 1813153 from the United States National Science Foundation (NSF), and by grant No. LU 856/13-1 from the Deutsche Forschungsgemeinschaft. We are grateful to Noam Ordan and Ella Rabinovich for helpful discussions and creative ideas. We thank the anonymous RANLP reviewers for their constructive comments. All remaining errors and misconceptions are, of course, our own.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. 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. Association for Computational Linguistics.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, 31(1):30–54.

Mona Baker. 1993. 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.

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

Satanjeev Banerjee and Alon Lavie. 2005. 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. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. 2006. 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.

Shoshana Blum-Kulka. 1986. 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.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. 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. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. 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. Association for Computational Linguistics.

Sauleh Eetemadi and Kristina Toutanova. 2014. 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.

Sauleh Eetemadi and Kristina Toutanova. 2015. Detecting translation direction: A cross-domain study. In *NAACL Student Research Workshop*. ACL – Association for Computational Linguistics.

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

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 937–944.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *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.

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. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. EuroParl-UdS: Preserving and extending metadata in parliamentary debates. In *Proceedings of ParlaCLARIN*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86. AAMT.

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

Moshe Koppel and Noam Ordan. 2011. 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. Association for Computational Linguistics.

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

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

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

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023.

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

Lieve Macken, Orphée De Clercq, and Hans Paulussen. 2011. Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta*, 56(2):374–390.

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. 2010. Quantitative analysis of culture using millions of digitized books. *Science*.

Sergiu Nisioi. 2015. Unsupervised classification of translated texts. In *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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. Association for Computational Linguistics.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016a. 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.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. 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.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2016b. A parallel corpus of translationese. In *Proccedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*.

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

Milan Straka and Jana Straková. 2017. 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. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics.

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

Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. 2018. The UN parallel corpus annotated for translation direction. ArXiv:1805.07697 [cs.CL].

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

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

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

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

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

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

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