Statistical Machine Translation with Automatic Identification of Translationese

Naama Twitto-Shmuel
Dept. of Computer Science
University of Haifa
Israel
naama.twitto@gmail.com

Noam Ordan
Cluster of Excellence, MMCI
Universität des Saarlandes
Germany
noam.ordan@gmail.com

Shuly Wintner
Dept. of Computer Science
University of Haifa
Israel
shuly@cs.haifa.ac.il

Abstract

Translated texts (in any language) are so markedly different from original ones that text classification techniques can be used to tease them apart. Previous work has shown that awareness to these differences can significantly improve statistical machine translation. These results, however, required meta-information on the ontological status of texts (original or translated) which is typically unavailable. In this work we show that the predictions of translationese classifiers are as good as meta-information. First, when a monolingual corpus in the target language is given, to be used for constructing a language model, predicting the translated portions of the corpus, and using only them for the language model, is as good as using the entire corpus. Second, identifying the portions of a parallel corpus that are translated in the direction of the translation task, and using only them for the translation model, is as good as using the entire corpus. We present results from several language pairs and various data sets, indicating that these results are robust and general.

1 Introduction

Research in Translation Studies suggests that translated texts are considerably different from original texts, constituting a sublanguage known as *Translationese* (Gellerstam, 1986). Awareness to translationese can significantly improve statistical machine translation (SMT). Kurokawa et al. (2009) showed that French-to-English SMT systems whose translation models were constructed from human translations from French to English yielded better translation quality than ones created from translations in the other direction. These results were corroborated by Lembersky et al. (2012a, 2013), who showed that translation models can be adapted to translationese, thereby improving the quality of SMT even further. Awareness to translationese also benefits the *language* models used in SMT: Lembersky et al. (2011, 2012b) showed that language models complied from translated texts better fit the reference sets in term of perplexity, and SMT systems constructed from such language models perform much better than those constructed from original texts.

To benefit from these results, however, one has to know whether the texts used for training SMT systems are original or translated, and previous work indeed used such meta-information. Unfortunately, annotation reflecting the status of texts, or the direction of translation, is typically unavailable. The research question we investigate in this work is whether the predictions of translationese classifiers can replace manual annotation. In a variety of evaluation scenarios, we demonstrate that this is indeed the case. When a monolingual corpus in the target language is given for constructing a language model for SMT, we show that automatically identifying the translated portions of the corpus, and using only them for the language model, is as good as using the entire corpus. Similarly, when a parallel corpus is given, we show that automatically identifying the portions of the corpus that are translated in the direction of the translation task, and using only them for training the translation model, is again as good as using the entire corpus. We present results from several language pairs and various data sets, indicating that the approach we advocate is general and robust.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 47–57,
Lisboa, Portugal, 17-18 September 2015.

The main contribution of this work is a general approach that, provided labeled data for training classifiers, can be applied to *any* corpus before it is used for constructing SMT systems, resulting in systems that are as good as (or better than) those that use the entire corpus, but that rely on significantly smaller language and translation models.

We briefly review related work in Section 2. Section 3 describes our methodology and experimental setup. Section 4 details the experiments and their results. We conclude with an analysis of the results and suggestions for future research.

2 Related work

Until recently, SMT systems were agnostic to the ontological status of a text (as original vs. translated). Several recent works, however, underscore the relevance of translationese for SMT. Kurokawa et al. (2009) were the first to show that translationese matters for SMT. They defined two translation tasks, English-to-French and French-to-English, and used a parallel corpus in which the translation direction of each text was indicated. They showed that for the English-to-French task, translation models compiled from English-translated-to-French texts were better than translation models compiled from texts translated in the reverse direction; and the same holds for the reverse translation task. These results were corroborated by Lembersky et al. (2012a, 2013), who further demonstrated that translation models can be adapted to translationese, thereby improving the quality of SMT even further.

Lembersky et al. (2011, 2012b) focused on the *language* model (LM). They built several SMT systems for several pairs of languages. For each language pair they built two systems, one in which the LM was compiled from original English text, and another in which the LM was compiled from text translated to English from each of the languages. They showed that LMs complied from translated texts better fit the reference set in term of perplexity. Moreover, SMT systems that were constructed from translationese-based LMs perform much better than those constructed from original LMs. In fact, an original corpus must be as much as ten times larger in order to yield the same translation quality as a translated corpus.

To benefit from these results, one has to know whether the texts used for training SMT systems are original or translated; such meta-information is typically unavailable. Due to the unique properties of translationese, however, this information can be determined automatically using text-classification techniques. Several works address this task, using various feature sets, and reporting excellent accuracy (Baroni and Bernardini, 2006; van Halteren, 2008; Ilisei et al., 2010; Eetemadi and Toutanova, 2014). Some of these works, however, only conduct in-domain evaluation; much evidence suggests that out-of-domain accuracy is much lower (Koppel and Ordan, 2011; Islam and Hoenen, 2013; Avner et al., Forthcoming).

A thorough investigation was conducted by Volansky et al. (2015), who focused on the features of translationese (in English) from a translation theory perspective. They defined several classifiers based on various linguistically-informed features, implementing several hypotheses of Translation Studies. We adopt some of their best-performing classifiers in this work.[1]

3 Experimental setup

The experiments we describe in Section 4 consist of three parts: 1. Training classifiers to tease apart original from translated texts. 2. Constructing SMT systems with language models compiled from the predicted translations, comparing them with similar SMT systems whose language models consist of the entire monolingual corpora. 3. Constructing SMT systems with translation models compiled from bitexts that are predicted as translated in the same direction as the direction of the SMT task, comparing them with similar SMT systems whose translation models consist of the entire parallel corpora. In this section we describe the language resources and tools required for performing these experiments.

3.1 Tools

Our first task is text classification; to ensure that the length of each text does not influence the classification, we partition the training corpus in most experiments into chunks of approximately 2000 tokens (ending on a sentence boundary). We henceforth use *chunk units* to define the size of a sub-corpus. Our major experiments involve 2,500 chunks (of approximately 2,000 tokens each, hence 5M tokens). To detect sentence

[1] Volansky et al. (2015) only identified English translationese; we extend the experimentation also to French and adapt their classifiers accordingly.

boundaries, we use the UIUC CCG tool.[2]

We use MOSES (Koehn et al., 2007) for tokenization and case normalization. Part-of-speech (POS) tagging is done with *OpenNLP*[3] for English and the *Stanford* tagger[4] for French. For classification we use *Weka* (Hall et al., 2009) with the *SMO* algorithm, a support-vector machine with a linear kernel, in its default configuration.

To construct language models and measure perplexity, we use *SRILM* (Stolcke, 2002) with interpolated modified Kneser-Ney discounting (Chen and Goodman, 1996) and with a fixed vocabulary. We limit language models to a fixed vocabulary and map out-of-vocabulary (OOV) tokens to a unique symbol to overcome sparsity and better control the OOV rates among various corpora.

We train and build the SMT systems using MOSES. For evaluation we use MultEval (Clark et al., 2011), which takes machine translation hypotheses from several runs of an optimizer and provides three popular metric scores, BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and TER (Snover et al., 2006)), as well as standard deviations and p-values.

3.2 Corpora

To construct SMT systems we need both monolingual corpora (for the language model) and bilingual ones (for the translation model). The main corpora we use are Europarl (Koehn, 2005) and the Canadian Hansard. Europarl is a multilingual corpus recording the proceedings of the European Parliament. Some portions of the corpus are annotated with the original language of the utterances, and we use the method of Lembersky et al. (2012a) to identify the source language of other segments. The Hansard is a parallel corpus consisting of transcriptions of the Canadian parliament in English and (Canadian) French from 2001-2009. We use a version that is annotated with the original language of each parallel sentence.[5] We also use the News Commentary corpus (Callison-Burch et al., 2007), a French-English corpus in the domain of politics, economics and science. The direction of translation of this corpus is not annotated.[6]

3.2.1 Language model experiments

Our main experiments focus on French translated to English (FR→EN), and we define a classifier that can identify English translationese. However, to further establish the robustness of our approach, we also experiment with German translated to English (DE→EN) and with English translated to French (EN→FR). We also conduct cross-corpus experiments in which we train translationese classifier on one corpus (Europarl) and test its contribution to SMT on another (Hansard, News). These experiments are crucial for evaluating the robustness of our approach, in light of the findings that translationese classification is much less accurate outside the training domain.

From the Europarl corpus we use several portions, collected over the years 1996 to 1999 and 2001 to 2009. In all experiments, the split of the monolingual corpora to translated vs. original texts is balanced (in terms of chunks). The parallel corpora are divided to two sections according to the direction of the translation (when it is known). For example, for the French-to-English translation task, we divide the Europarl corpus to a French-original section (FR→EN) and an English-original section. We also use portions of Europarl to define reference sets for evaluating the perplexity of LMs. For this task we only use translated texts.

For constructing translation models we use parallel corpora. For the FR→EN and EN→FR tasks we use original French text, aligned with its translation to English (FR→EN). For the DE→EN translation task we use original German text, aligned with its translation to English (DE→EN). The parallel portions we use are disjoint from those used for the language model and are evenly balanced between the original text and the aligned translated text. From Europarl we use portions from the period of January to September 2000.

To tune and evaluate SMT systems we use reference sets that are extracted from a parallel, aligned corpus. These include 1000 sentence pairs for tuning and 1000 (different) sentence pairs for evaluation. The sentences are randomly extracted from another portion of the Europarl corpus, collected over the period of October to December 2000, and another portion of Hansard. All tuning and references sets are disjoint from the training materials.

3.2.2 Translation model experiments

In this set of experiments we focus again on FR→EN systems, but also experiment with

[2] http://cogcomp.cs.illinois.edu/page/tools_view/2, accessed 11.10.2013.

[3] http://opennlp.apache.org, 24.08.2012.

[4] http://nlp.stanford.edu/software/tagger.shtml, accessed 08.02.2013.

[5] We are grateful to Cyril Goutte, George Foster and Pierre Isabelle for providing us with this version of the corpus.

[6] The precise data sets we used will be made available.

DE→EN and EN→FR. We conduct in-domain experiments using the Europarl corpus, and a cross-corpus experiment in which we train on one corpus and test on another. From the Europarl corpus we use several portions, collected over the years 1996 to 1999 and 2001 to 2009.

To construct language models for the in-domain experiments we use Europarl portions from the period of January to September 2000 (this is the English/French side of the training data used for building the translation model in the language model experiments). For cross-corpus experiments we use the LM built from translated texts that we use in the Hansard language model experiments. For tuning and evaluation we use the same sets used in the language model experiments.

4 Experiments and results

4.1 Language models experiments

We build several SMT systems that use the same translation model, but differ in their language models. This involves three tasks detailed below.

4.1.1 Classification of translationese

The first task is to train a classifier to detect translationese. This has been done before, and we adapt some of the classifiers of Volansky et al. (2015). Specifically, our classifier is based on *Contextual function words*: we use counts of (contiguous) trigrams $\langle w_1, w_2, w_3\rangle$, where each element w_i is either a word or its part of speech (POS), at least two of the elements are function words, and at most one is a POS tag. An example feature is the triple $\langle$*in,the,Noun*$\rangle$. This feature set combines lexical and shallow syntactic information in a way that was proven useful for identifying translationese. We also add counts of punctuation marks, another feature that was shown accurate.[7] We evaluate the accuracy of this classifier intrinsically, using ten-fold cross-validation.

Then, we use the prediction of the classifier to determine whether test texts are original or translated. The classifier thus defines a partition of the training corpus to (predicted) originals vs. translations. Based on the classifier's prediction, we build language models from the sub-corpus determined as translated. We then evaluate the fitness of this sub-corpus to the reference set, in terms of perplexity. Specifically, we train 1-, 2-, 3-, and 4-gram LMs for this sub-corpus and measure their perplexity on the reference set. This provides an extrinsic evaluation for the quality of the classifier.

The results are reported in Table 1. Replicating the results of Volansky et al. (2015), we demonstrate that the classifier is indeed excellent. Not surprisingly, good classification yields good language models. The rightmost columns of Table 1 list the perplexity of language models trained on the sub-corpus that was predicted as translated, when applied to the reference set. For comparison, we provide in Table 1 also the perplexity of language models compiled from the entire training set; from the *actual* (as opposed to predicted) translated texts; and from the actual *original* texts. Clearly, and consistently with the results of Lembersky et al. (2012b), the original texts yield the worst language models (highest perplexity), whereas the actual translated texts yield an upper bound (lowest perplexity). Still, due to the high accuracy of the classifier, its perplexity is very similar to this upper bound. The model that is built from all texts, both original and translated, is twice as large as the corpus used for the other models, hence the lower perplexity rates.

To further establish the robustness of these results, we repeat the experiments with other corpora, this time consisting of German translated to English (DE→EN), and also English translated to French (EN→FR). We only report results for the 4-gram LMs (Table 2). The accuracies of the classifiers are high, comparable to the case of FR→EN. Moreover, the perplexities of the induced language models are very close to the upper bound obtained by taking actual translated texts.

4.1.2 Language models compiled from predicted translationese

We established the fact that translated texts can be identified with high accuracy, and that language models compiled from predicted translations fit the reference sets well. Next, we construct SMT systems with these language models. Our hypothesis is that language models compiled from (predicted) translationese will perform as well as (or even better than) language models compiled from the entire corpus. We evaluate this hypothesis in several scenarios: when the corpus used for the language model is the same corpus used for training the classifiers; or a different one, but of the same type; or from a completely different domain.

We begin with a French-to-English translation task. We use the same (4-gram) language models

[7] The code for feature generation will be released.

			Perplexity			
Data set	Chunks	Acc. (%)	1-gram	2-gram	3-gram	4-gram
Predicted translations	1245	98.96	463.51	94.81	71.60	68.76
Translated texts	1255		463.58	94.59	71.24	68.37
Original texts	1258		500.56	115.48	91.14	88.31
All texts	2513		473.00	93.34	67.84	64.47

Table 1: Classification of translationese, and fitness to the reference set of FR→EN language models compiled from texts predicted as translated

		DE→EN			EN→FR	
Data set	Chunks	Acc. (%)	Ppl	Chunks	Acc. (%)	Ppl
Predicted translations	1,146	99.08	62.23	1,410	98.47	47.92
Translated texts	1,153		62.07	1,413		47.89
Original texts	1,153		76.68	1,411		59.75
All	2,306		57.48	2,824		44.49

Table 2: Accuracy of the classification, and fitness of language models compiled from texts predicted as translated to the reference set, DE→EN and EN→FR

described in Section 4.1.1, constructed from the predictions of the classifier. We also fix a single translation model, compiled from the parallel portion of the training corpus (Section 3.2). We then train a French-to-English SMT system with the (predicted) LM. As a baseline, we build an SMT system that uses the entire training corpus for its language model; we refer to this system as *All*. As an upper bound (for a system that uses only a portion of the corpus), we build a system that uses the (actual) translated texts for its LM. We also report results on a system that uses only original texts for its LM. All systems are tuned on the same tuning set of 1000 parallel sentences, and are tested on the same reference set of 1000 parallel sentences.

We evaluate the quality of each of the SMT systems using MultEval (Section 3.1). The results are presented in Table 3, reporting the BLEU, METEOR (MET), and TER evaluation measures, as well as the p-value defining the statistical significance with which the system is different from the baseline (with respect to the BLEU score only).

Replicating some of the results of Lembersky et al. (2011, 2012b), we find that using only translated texts for the language model is not inferior to using the entire corpus (although the size of the latter is double the size of the former). In terms of BLEU scores, both yield the same score, 29.1. Similarly, as reported by Lembersky et al. (2011, 2012b), using only original texts is markedly worse, with a BLEU score of 27.8. The main novelty of our current results, however, is the observation that the language model that only uses *predicted*, rather than actual translated texts, performs just as well.[8]

For completeness, we repeat the same experiments with two more language pairs: German to English and English to French. The setup is identical, and we report the same evaluation metrics. The results are presented in Table 4. The emerging pattern is identical to that of French to English.

The results of all the experiments confirm our hypothesis; SMT systems built from *predicted* translationese language models perform as well as SMT systems built from (actual) translated language models, and similarly to (twice as large) mixed language models.

4.1.3 Cross-corpus experiments

The experiments discussed above all use the same type of corpus both for training the translationese classifiers and for training the SMT systems (the actual portions differ, but all are taken from the same corpus). In a typical translation scenario, a monolingual corpus is available for constructing a language model, but the status of its texts (original or translated) is unknown, and has to be predicted by a classifier that was trained on a potentially dif-

[8] In Table 3 and henceforth we highlight in boldface entries that correspond to classifiers whose performance is better than, or not significantly worse than, the performance of the *All* classifier, which is considered the baseline against which all other systems are compared.

Data set	BLEU↑	MET↑	TER↓	*p*
Predicted translations	**28.9**	**33.2**	**53.8**	0.16
Translated texts	**29.1**	**33.3**	**53.6**	0.58
Original texts	27.8	32.9	54.7	0.00
All	**29.1**	**33.3**	**53.8**	

Table 3: Evaluation of the FR→EN SMT system built from LMs compiled from predicted translationese

	DE→EN				EN→FR			
Data set	BLEU↑	MET↑	TER↓	*p*	BLEU↑	MET↑	TER↓	*p*
Predicted translations	**21.9**	**28.6**	**63.8**	0.87	**26.3**	47.8	**58.3**	0.47
Translated texts	**21.8**	**28.6**	**63.9**	0.37	26.1	47.7	**58.5**	0.03
Original texts	21.0	28.4	64.6	0.00	25.1	47.0	59.5	0.00
All	**21.9**	**28.6**	**63.7**		**26.3**	**48.0**	**58.7**	

Table 4: Evaluation of the DE→EN and EN→FR SMT systems built from LMs compiled from predicted translationese

ferent domain. The question we investigate here, then, is whether a classifier trained on texts in one domain is useful for predicting translationese in a different domain.

As a first experiment, we use an (English) translationese classifier that is trained on the Europarl training data, but use the Hansard training data for constructing the SMT system. In this experiment, we do not use the meta-information of the Hansard corpus, but instead use the predictions of the classifier. Based on these predictions, we define a partition of the Hansard training corpus to (predicted) originals vs. translations and use the text chunks that were classified as translated to build 4-grams language models.

Again, as in the in-domain experiment, we construct a single, fixed translation model from the parallel portion of the (Hansard) corpus. We then train a French-to-English SMT system with the (predicted) LM. As a baseline, we build an SMT system that uses the entire Hansard training corpus for its language model (*All*). As an upper bound, we build a system that uses the (real) translated texts for its LM. We also report results on a system that uses only original texts for its LM. All systems are tuned and tested on the same tuning and evaluation reference set.

The results (Table 5) are consistent with the findings of the in-domain experiments. Although the classifier only performs at 78% accuracy, its predictions are sufficient for defining a language model whose BLEU score (37.8) is statistically indistinguishable with the score (38.0) of LMs based on real translations or the entire corpus.

We repeat the cross-corpus experiments with the News Commentary corpus, a French-English parallel corpus for which the direction of translation is not annotated; we only use its English side. Presumably, most of the texts in this corpus consist of original English, but we hypothesize that the classifier may be able to select chunks with translationese-like features and consequently provide a better SMT system. Additionally, as the News Commentary corpus is a collection of editorials, we partition the corpus into (not necessarily equal-length) articles, rather than to 2000-token chunks, to maintain the coherence of chunks.

The results (Table 6) reveal the same pattern: the predicted-translationese system yields a BLEU score of 27.0, statistically insignificant difference compared with the *All* system that uses the entire corpus (27.2). This is obtained with much smaller corpora, only 1,470 chunks (58% of the entire corpus of 2,527 chunks).

4.2 Translation model experiments

We now move to experiments that address the translation model. We build SMT systems that use a fixed language model but differ in their translation model training data. For all systems we use fixed tuning and evaluation sets.

4.2.1 Translation models compiled from predicted translationese

We first train a classifier to detect the direction of the translation (FR→EN vs. EN→FR). We classify the English side of the parallel corpus; for the

Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	p
Predicted translations	1,321	78.22	**37.8**	**37.7**	**45.9**	0.11
Translated texts	2001		**38.0**	**37.8**	**45.7**	0.86
Original texts	2001		37.5	37.6	46.1	0.00
All	4002		**38.0**	**37.7**	**45.8**	

Table 5: Cross-corpus evaluation: Hansard-based SMT system, Europarl-based classification

Data set	Chunks	BLEU↑	MET↑	TER↓	p
Predicted translations	1,470	27.0	**33.0**	**55.2**	0.02
All	2,527	**27.2**	**33.0**	**55.2**	

Table 6: Cross-corpus evaluation: News Commentary corpus

FR→EN and DE→EN tasks, chunks predicted as *translated* are assumed to be translated in the right direction ($S \rightarrow T$). For the EN→FR task, chunks predicted as *original* are assumed to be translated in the right direction. Then, we use the prediction of the classifier to construct translation models: we only use the chunks predicted as translated in the right direction. For each partition, we match the English with the aligned French (or German) sentences, thereby defining the SMT training data.

We hypothesize that translation models built from such training data are better for SMT. To explore this hypothesis we fix a single language model (Section 3.2), and train an SMT system with the (predicted) partitions and their aligned sentences. As a baseline, we build an SMT system, *All*, that uses the entire training corpus for its translation model. As an upper bound, we build a system that uses for its translation model the portion of the parallel corpus that was indeed translated in the right direction ($S \rightarrow T$). We also report results on a system that uses only the portion of the parallel corpus that was translated in the opposite direction ($T \rightarrow S$) for its translation model. All systems are tuned on the same tuning set and are tested on the same reference set.

The results are presented in Table 7. They are consistent with previous works that showed that SMT systems trained on $S \rightarrow T$ parallel texts outperformed systems trained on $T \rightarrow S$ texts (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013). Indeed, the best-performing systems use either (actual) $S \rightarrow T$ texts (BLEU score of 31.3), or the entire corpus (31.3); the worst system uses (actual) $T \rightarrow S$ texts (28.4). What we add to previous results is the corroboration of the hypothesis that a predicted-translationese system performs just as well as the actual ones.

As in the language model experiments, we repeat the same experiments with two more translation tasks: German to English and English to French. The setup is identical, and we report the same evaluation metrics. The emerging pattern (Table 7) confirms our hypothesis: SMT systems built from *predicted* $S \rightarrow T$ systems perform as well as SMT systems built from the entire corpus.

4.2.2 Cross-corpus experiments

The above results are not very surprising given the high accuracy of the translationese classifier. The question we investigate in this section is whether a classifier trained on texts in one domain is useful for predicting translationese in a different domain.

We train an (English) translationese classifier on the Europarl training data, but use the Hansard corpus for the translation model. We apply the classifier to the English side of the Hansard corpus, and based on its predictions, define a partition of the Hansard training corpus to use for the translation model. As in the in-domain experiment, we construct a single, fixed language model from a portion of the (Hansard) corpus. We then train a French-to-English SMT system with the (predicted) translation model, comparing it to systems that use the entire Hansard training corpus, the (actual) $S \rightarrow T$ texts and the actual $T \rightarrow S$ texts.

Table 8 reports the results. The best-performing systems use either actual $S \rightarrow T$ texts or the entire corpus (BLEU score of 37.3). The classifier performs worse, at 36.3, but still much better than the system that is based on $T \rightarrow S$ texts. This should be attributed to the very small number of chunks predicted by the classifier as $S \rightarrow T$.

Task	Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	p
FR→EN	Predicted $S \to T$	1,678	98.93	**31.1**	**34.7**	**52.1**	0.13
	$S \to T$	1,690		**31.3**	**34.8**	**51.7**	0.94
	$T \to S$	1,689		28.4	33.3	54.4	0.00
	All	3,379		**31.3**	**34.7**	**51.9**	
DE→EN	Predicted $S \to T$	1,607	99.44	23.7	30.3	61.6	0.00
	$S \to T$	1,613		**24.0**	30.4	**61.3**	0.05
	$T \to S$	1,612		21.7	29.0	63.9	0.00
	All	3,225		**24.2**	**30.5**	**61.1**	
EN→FR	Predicted $S \to T$	1,678	98.93	**29.4**	**50.7**	**55.3**	0.11
	$S \to T$	1,689		**29.3**	**50.8**	**56.1**	0.18
	$T \to S$	1,690		26.7	48.2	58.2	0.00
	All	3,379		**29.1**	**50.6**	**56.0**	

Table 7: Accuracy of the classification and evaluation of SMT systems built from translation models compiled from predicted translationese

Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	p
Predicted $S \to T$	1,840	79.36	36.3	36.9	46.6	0.00
$S \to T$	3,000		**37.3**	**37.3**	**46.2**	0.94
$T \to S$	3,000		34.1	35.8	48.9	0.00
All	6,000		**37.3**	**37.4**	**46.0**	

Table 8: Cross-corpus evaluation: Hansard-based SMT system, Europarl-based classification

5 Conclusion

Two fundamental insights, motivated by research in Translation Studies, drive our work:

1. *Direction matters.* When constructing translation models from parallel texts it is important to identify which side of the bitext is the source and which is the target. Translation from the source of the SMT task to its target is always better than the reverse option. In fact, direction itself was utilized as features for classification of translationese by selecting alignment patterns from O to T and vice versa (Eetemadi and Toutanova, 2014, 2015).
2. *Translationese matters.* When constructing language models, translated texts (especially from the source language, but not only) are preferable to texts written originally in the target language of the task at hand.

Our main hypothesis was that these benefits to SMT still hold when meta-information on the status of the texts is unavailable, and has to be predicted, especially in light of the deterioration in the accuracy of translationese classifiers in the face of out-of-domain texts. We trained classifiers to identify translationese, and then used their predictions to construct language- and translation-models for SMT, demonstrating that attention to translationese can yield state-of-the-art translation quality with only a fraction of the corpora. We find that one can generally rely on classifiers that identify at least half of the data as translated for both the language model and the translation model.

In future work we would like to improve our classifiers such that smaller chunks of text suffice for accurate identification of translationese. We also believe that combining various feature sets is a key to improving the accuracy, and especially the robustness, of translationese classifiers. In this work we combined two complementary feature sets; more work should be done in this direction. In particular, there is ample evidence that features should be sensitive to language *family*, as translations from similar languages look more similar than translations from unrelated languages (Pym and Chrupała, 2005; Koppel and Ordan, 2011). To further improve the generality and domain-independence, we currently experiment with *unsupervised* classification of translationese, with very encouraging preliminary results (Rabinovich and Wintner, 2015).

Finally, we mainly experimented with English and French in this work, but we are confident that many language pairs can benefit from the methodology we propose.

Acknowledgments

This research was supported by a grant from the Israeli Ministry of Science and Technology. The second author was supported by Cluster of Excellence MMCI at Saarland University. We are grateful to Gennadi Lembersky for his continuous help.

References

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, Forthcoming. doi: http://dx.doi.org/10.1093/llc/fqu047. URL http://dx.doi.org/10.1093/llc/fqu047.

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-0718.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/981863.981904. URL http://dx.doi.org/10.3115/981863.981904.

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*, pages 176–181, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-2031.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91. Association for Computational Linguistics, July 2011. URL http://www.aclweb.org/anthology/W11-2107.

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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL http://dx.doi.org/10.1145/1656274.1656278.

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.

Zahurul Islam and Armin Hoenen. Source and translation classification using most frequent words. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1299–1305, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I13-1185`.

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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-2045`.

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.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1034`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France, April 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1026`.

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

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.

Anthony Pym and Grzegorz Chrupała. The quantitative analysis of translation flows in the age of an international language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*, pages 27–38. John Benjamins, Amsterdam, 2005.

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`.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006. URL `http://www.cs.umd.edu/~snover/tercom/`.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Proced-*

ings of International Conference on Spoken Language Processing, pages 901–904, 2002. URL `citeseer.ist.psu.edu/stolcke02srilm.html`.

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`.

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