Effect of translationese on machine translation quality

Stephanie Chen
Stanford University
schen751@stanford.edu

Abstract

"Translationese" refers to the unique structural characteristics of translated text compared to text originally written in a given language, a result of compromises made by translators between fidelity and fluency in source and target languages. We examine the effect of translationese on machine translation in three contexts—a phrase-based statistical system, a neural model, and an unsupervised phrase-based system—in translations from a variety of languages to English. We find that regardless of language and system, models trained on target corpora translated from the source language outperform models trained on original target-language corpora, with especially strong improvements in the case of a low-resource language.

1 Introduction

Translation theory has long held that texts in translation exhibit significant structural differences from untranslated texts, a phenomenon termed "translationese." These differences result from the translator's attempt to produce a work that reads fluently in the target language while still maintaining fidelity to the source text's style and substance. Because of this attempt at compromise, translated texts are affected by the source language itself—its grammar and norms, beyond the style of the author—in what is termed the *law of interference*, as well as by the target language's own norms, in the *law of growing standardization* (13).

With the rise of statistical machine translation (SMT) based on large text corpora, translation theory has been generally overlooked in favor of work on data collection, feature engineering, and model architectures. Recent years, however, have seen a rise in studies of translationese. Research (see Section 2) has shown that translations can be automatically distinguished from original-language texts with high accuracy (15) (11), and that language models trained on translations perform significantly better than language models trained on original-language texts in machine translation systems (9). These results suggest that the translation status of a text and the directionality of translation are important features to consider when compiling training and test corpora, yet they are largely ignored in benchmark corpora like Gigaword, News Crawl, and bilingual Europarl.

In this project, we compare translation performance between models trained on translated text and models trained on original-language text in three contexts: supervised phrase-based statistical translation, supervised neural translation, and unsupervised (monolingual) phrase-based statistical translation. We test on translation from five languages into English and find that across the three tasks, models trained on translated texts consistently perform better, even when trained on corpora smaller by up to an order of magnitude. Based on these results, we conclude that translation directionality significantly influences translation quality, and we discuss some potential practical implications, focusing on the case of low-resource languages.

2 Related work

(13) first defined the laws of interference and growing standardization, in which a translation exhibits signs of the source text's style and the source language's grammar as well as the attempt of the

translator to conform the text to the target language's norms and culture. (1) proposed several "translation universals," traits that appear in all translations, regardless of source and target languages. These include *simplification*, the tendency of translators to use simpler language (e.g. smaller or more frequent words) when describing complex concepts, and *explicitation*, in which translators tend to make explicit in the translation concepts that are implicit in the source text.

(4) and (11) find that translated texts can be automatically distinguished from original-language texts using supervised (4) and unsupervised (11) machine learning methods. (7) use machine learning to classify translations and also identify the source language given a translated text, and find that the degree of difference between translated texts is directly related to the degree of typological difference between the source languages, i.e. that translations from closely related source languages are more similar to each other than translations from distant source languages.

(15) examines 32 features of translationese in the context of the theories detailed above and compare their utility in automatically classifying texts as original-language or translated. They find that lexical variety, word rank (commonness), and features that capture source language structure like part-of-speech n-grams and positional token frequency are good indicators of translationese, while sentence length and explicitation features like naming are not. They also find that certain features perform well for only specific languages. For example, original-English texts have a much higher count of sentences beginning with "But" than English translations, likely owing to translators' adherence to the English "rule" discouraging "But" sentences in formal writing.

(9) compare the quality of language models trained on translated and original-language English texts and measure the effect of these models on phrase-based SMT systems, finding that translation-based language models more closely match reference translations and perform better in translation contexts. They find that while models trained on text translated from the source language of the task perform best, even texts translated from a mixture of other languages produce models that outperform those trained on original target-language text. They also perform an "ablation study" in which training corpora are progressively stripped of features like punctuation and named entities and eventually abstracted into part-of-speech syntax and find that results hold, confirming that translationese is a result of deep structural language features. In this project, we repeat a subset of their experiments for our phrase-based translation task.

(12) examine a prior study of Chinese-to-English neural machine translation and find that testing on English-to-Chinese corpora—that is, parallel corpora in which the Chinese text is the result of translation from original English—increases BLEU scores in the Chinese-to-English task while producing substandard translations in human evaluation. They observe that neural translation systems are sensitive to translationese, in this case unintentionally taking advantage of native-English features in the test Chinese text, and conclude that machine translation systems should be tested on corpora translated in the same direction as the given task.

For our unsupervised translation task, we refer to (8), a recent paper that produced state-of-the-art results in neural and phrase-based unsupervised translation. They introduce neural and phrase-based systems unsupervised systems based on three techniques: initialization from dictionaries inferred from monolingual corpora, language modeling (as a denoiser in the neural case), and iterative backtranslation between monolingual corpora. They outperform the previous state of the art across several languages by up to 10 BLEU points and demonstrate the effectiveness of their system on low-resource languages. In this project, we use their phrase-based translation system for our unsupervised task.

3 Approach

3.1 Phrase-based system

In statistical machine translation, to find the best translation t given a source sentence s, we maximize

$$p(t|s) \propto p(s|t)p(t) \tag{1}$$

where $p(s|t)$ is determined by the translation model and $p(t)$ is determined by the target language model. We use a standard phrase-based SMT setup in which the translation probability $p(s|t)$ is read from a phrase table of matching source and target training phrases and the target language probability

$p(t)$ is determined by an n-gram language model. We train our SMT system using the Moses (6) toolkit's baseline translation model, which is a 3-gram language model with modified Kneser-Ney smoothing and a translation model using standard alignment factors and lexical reordering. As in (9), to examine the effect of the language model alone on translation, we compare systems with language models trained on original text with those with language models trained on translated text but maintain the same translation model, trained on bidirectional corpora, across both systems.

3.2 Neural system

For our neural system, we use the same architecture as in Assignment 4 for this class, which is a single-layer sequence-to-sequence model with attention, using a bidirectional LSTM encoder and a unidirectional LSTM decoder. We use multiplicative attention and pass the output through one linear layer with tanh activation and dropout, and compute softmax cross-entropy loss. Because this is an end-to-end translation system, unlike in the phrase-based system, we are unable to isolate the effect of the language model; experiments on the neural model thus measure the effect of translated and original-language text on the system as a whole.

3.3 Unsupervised task

We use the unsupervised phrase-based SMT implementation from (8) for this task; experiments with their unsupervised neural system were computationally infeasible in the scope of this project. Given source and target monolingual corpora, we use separate word embeddings for each corpus and align these embeddings, resulting in a rotation matrix W between the source and target embedding spaces. We can then populate source-to-target phrase tables as follows:

$$p(t_j|s_i) = \frac{e^{\frac{1}{T}\cos(e(t_j),We(s_i))}}{\sum_k e^{\frac{1}{T}\cos(e(t_k),We(s_i))}} \tag{2}$$

where s_i and t_j are the ith and jth words, respectively, in the source and target vocabularies, $\cos$ is cosine similarity, T is a hyperparameter on the phrase probability distributions, and $e(x)$ is the embedding of a word x.

Using this dictionary and an n-gram language model for the target language—we use a 5-gram modified Kneser-Ney model—we can then bootstrap an initial source-to-target translation model. We then iteratively backtranslate the source and target corpora as in Algorithm 1.

Algorithm 1: Unsupervised phrase-based SMT

Generate fastText embeddings for S and T;
Initialize phrase table by aligning embeddings;
Learn language model LM_t for T;
Build forward model $P^{(0)}_{s\to t}$ from phrase table and LM_t;
Translate source corpus C_s with $P^{(0)}_{s\to t}$, giving $C^{(0)}_t$;
for *iteration=1 to N* **do**
    Train backward model $P^{(i)}_{t\to s}$ using C_s and $C^{(i-1)}_t$;
    Translate target corpus C_t with $P^{(i)}_{t\to s}$, giving $C^{(i)}_s$;
    Train forward model $P^{(i)}_{s\to t}$ using $C^{(i)}_s$ and and C_t;
    Translate C_s with $P^{(i)}_{s\to t}$, giving $C^{(i)}_t$;
end

4 Methods

4.1 Dataset

We use version 7 (most recent public release) of the Europarl dataset (5), which contains transcripts of European Parliament proceedings from 1996 to 2011 (2007 to 2011 for certain languages). We use the full set of Europarl data for translations between five languages—French, German, Italian, Dutch, and Romanian—and English; note that Romanian-language data is only available from 2007 to 2011.

Because Europarl's bilingual subcorpora do not specify the original language of each utterance, we extract unidirectional subcorpora between source and target language pairs using the EuroparlExtract (14) toolkit, which assigns utterances to a source language based on the language tags and speaker IDs in the raw Europarl data. We then hold out a test set (Oct.–Dec. 2000; Nov.–Dec. 2009 for Romanian) and a development set (May–Jul. 2005; Nov.–Dec. 2010 for Romanian) for tuning and evaluation.

4.1.1 Phrase-based system

For the phrase-based SMT case, the rest of the data is used to train the language model, and a smaller subset (1999, Jan.–Sep. 2000, 2004, and 2011; all data for Romanian) is used to train the translation model. Since we distinguish between language models trained on original-language corpora and translated corpora, we train target language models on exclusively original (O-L) or translated (T-L) data, and to isolate these effects from the translation model, as in (9), we train the translation model on a a concatenated corpus of approximately equal numbers of sentences translated in each direction. For example, in French-to-English translation, our T-L system would have a language model trained on English translated from French, our O-L system would have a language model trained on original-English sentences, and both would have a translation model trained on a parallel set of both French-to-English and English-to-French sentences.

4.1.2 Neural system

The neural models are trained on the same segments of data as the phrase-based language models. In French-to-English translation, our T-L neural model would be trained on a parallel corpus of original-French sentences and their English translations, and our O-L model would be trained on a parallel corpus of original-English sentences and their French translations.

4.1.3 Unsupervised task

For unsupervised translation, we split the data by alternating years into two non-overlapping monolingual corpora, with M-A containing all data from even years (less the test set) and M-B containing all data from odd years (less the dev set). We arbitrarily select M-A as the default "source" corpus, and focus on the effects of the translation status of the target-language corpus. In French-to-English unsupervised translation, our T-L model would be trained on original French sentences from M-A and unrelated English sentences translated from French in M-B, and our O-L model would be trained on original French sentences from M-A and original English sentences from M-B. Our corpora are smaller by about an order of magnitude than the standard sets used for monolingual translation, but those datasets do not indicate directionality of translation.

All testing is done on parallel corpora in the direction of the translation task; i.e. in French-to-English, the test set consists of original French sentences and their English translations.

Statistics for each corpus by language are given in Table 1. Note that the original-English corpora for each of the first four languages are roughly the same (i.e. all English utterances in Parliament in the specified years); we specify statistics for the original-English corpora in the Romanian-to-English task because of the shorter date range of the dataset.

4.2 Evaluation

We use BLEU scores for evaluation. Evaluation during development is performed on the tuning set, and evaluation of final models (for all reported results below) is performed on the test set.

Table 1: Corpus statistics (sentence count)

Corpus	fr-en	de-en	it-en	nl-en	ro-en
Parallel	133336	143496	101475	115922	94728
T-L LM	199402	233657	83654	120938	12530
O-L LM*	371374	-	-	-	82198
Mono	111469	121771	40180	64258	6020
Mono T-L en	109032	133891	48922	68165	8786
Mono O-L en*	200240	-	-	-	65289
Tuning/Dev	4648	3579	1443	1685	1553
Test	9438	7808	2305	4758	723

4.3 Experiments

All experiments were run on two standard NV6 machines on Microsoft Azure, each with 6 vCPUs, 80GB memory, and an NVIDIA Tesla M60 GPU.

4.3.1 Phrase-based system

We use Moses scripts to tokenize and truecase all corpora (including tuning and test sets), then, for each translation task, train two 3-gram language models using KenLM (included with Moses): one on original English text, and one on English text translated from the source language. We then train the Moses baseline translation model with the `grow-diag-final-and` alignment algorithm and the `msd-bidirectional-fe` lexical reordering model; these are also the options used in the WMT baselines.

4.3.2 Neural system

We reused code from Assignment 4 for this system; the model was written in PyTorch, with BLEU scores calculated with ntlk. All corpora were again tokenized and truecased with Moses before training. All models ran until the early stop criterion was met (usually around 15 epochs), with the exception of the Italian-to-English and Dutch-to-English original-English models, which both ran into memory issues around epoch 6.

4.3.3 Unsupervised task

We used (8)'s implementation of the unsupervised system. All corpora were tokenized and truecased with Moses; note that while (8) requires that the two monolingual corpora be the same length, we do not enforce this due to the already-small size of our corpora. We use pretrained fastText embeddings for all languages and align those embeddings with MUSE (3). Each 5-gram language model is trained with KenLM, and the translation systems are trained with Moses using the same alignment and lexical reordering as in the phrase-based SMT task. We run each translation task for three iterations of backtranslation.

4.4 Results

Results for each task are shown in Tables 2, 3, 4, and 5. For both the phrase-based and neural translation tasks, models trained on English text translated from the source language consistently outperform models trained on original-English text across all five languages tested, with the exception of the Romanian-to-English neural translation. This discrepancy might be due to corpus size effects on the neural model; given the small T-L corpus (nearly seven times smaller than the next-smallest corpus), the T-L neural model could be overfitting to the training data.

We can see, however, that in all other cases, the T-L model performs significantly better than the O-L model, as expected. Improvements range from 0.73 to 3.34 BLEU points with the phrase-based model and from 1.01 to 4.43 BLEU points with the neural model. These improvements occur despite the fact that the O-L English training corpora are significantly larger (from 85% to over 300% larger across the first four languages, and nearly seven times larger in the case of Romanian) than the T-L corpora.

Table 2: BLEU scores, phrase-based statistical model

Task	Model	BLEU
fr→en	T-L LM	32.23
	O-L LM	31.37
de→en	T-L LM	24.93
	O-L LM	24.20
it→en	T-L LM	28.32
	O-L LM	27.51
nl→en	T-L LM	29.47
	O-L LM	28.49
ro→en	T-L LM	38.61
	O-L LM	35.27

Table 3: BLEU scores, sequence-to-sequence neural model

Task	Model	BLEU
fr→en	T-L	33.15
	O-L	28.72
de→en	T-L	26.85
	O-L	24.15
it→en	T-L	27.78
	O-L	25.07*
nl→en	T-L	29.42
	O-L	25.60*
ro→en	T-L	31.81
	O-L	32.80

Also in line with previous work, we find that the neural models generally outperform their corresponding phrase-based models for languages like French and German with large datasets, but that the phrase-based models perform better for low-resource languages like Romanian.

Tables 4 and 5 contain BLEU scores on the test set (only output to measure model quality, not used to adjust model parameters) for each stage of backtranslation, with the best final scores bolded for each task. We can see that the effects of translationese persist even in the case of unsupervised translation: the T-L models perform better across the board, with improvements of 0.90 to 7.55 BLEU points. Note that even the backtranslation models perform slightly better in the T-L case than the O-L case when tested on the backwards test corpus. As with the phrase-based and neural tasks, this improvement occurs despite a significant corpus size difference between the T-L and O-L target English corpora.

5 Analysis

As expected, the T-L models outperform the O-L models consistently across all tasks and all translation systems. Based on the scores observed, we make a few general observations.

In both phrase-based translation experiments (supervised and unsupervised), the difference in BLEU score between T-L and O-L models is generally larger for language tasks with smaller corpora. In the supervised case, for example, we see a much larger improvement in Romanian (3.34 BLEU points) than in French or German (about 1 BLEU point). We see an even greater improvement in the unsupervised case, where the Romanian T-L model performs over 7 BLEU points better than the O-L model, while the French and German models see improvements of 1-2 BLEU points. This could be due to the small size of the Romanian corpora, or the scores could be affected by the large difference in size between the Romanian O-L and T-L corpora; more work is needed here.

Table 4: BLEU scores, unsupervised PBSMT, T-L

	en-fr	fr-en	en-de	de-en	en-it	it-en	en-nl	nl-en	en-ro	ro-en
UPT	-	11.21	-	8.05	-	9.33	-	12.74	-	14.34
Iter. 1	14.33	18.36	6.88	11.22	9.57	15.22	11.92	15.19	5.76	17.68
Iter. 2	17.48	**19.69**	9.09	12.70	12.32	16.68	12.81	15.72	7.87	20.81
Iter. 3	17.98	19.63	9.57	**12.90**	12.59	**16.70**	12.88	**16.05**	8.70	**21.84**

Table 5: BLEU scores, unsupervised PBSMT, O-L

	en-fr	fr-en	en-de	de-en	en-it	it-en	en-nl	nl-en	en-ro	ro-en
UPT	-	10.77	-	7.91	-	8.96	-	12.47	-	7.01
Iter. 1	14.05	16.92	6.82	10.58	9.46	14.29	12.08	14.28	5.46	9.26
Iter. 2	17.21	**17.81**	8.79	11.70	12.15	14.83	12.69	14.79	6.90	11.58
Iter. 3	17.20	17.41	9.26	**12.00**	12.38	**14.92**	12.98	**14.85**	7.86	**14.29**

If translationese-based models can be shown to significantly outperform original-language models in the case of low-resource languages regardless of source language, there are some significant potential practical implications. In general, there is an inverse relationship between the amount of text published in a language and the rate of translations into that language (10); i.e. the number of original-English texts vastly outnumbers the number of texts translated into English, but in lower-resource languages like Romanian, there are potentially many more translations into Romanian than original-Romanian texts. This poses challenges for translation *out* of low-resource languages. However, if there are qualities of translationese that can be shown to improve translation regardless of source language, then translated texts from related source languages to the same target language could be used to bolster performance.

In the three systems studied here, the structure of each imposes different constraints on the influence of the target-language corpus over the system's learned translation probabilities, but the improvements in BLEU score of T-L models over O-L models doesn't seem significantly different between systems. In the supervised phrase-based model, we intentionally train the translation model on a balanced bidirectional parallel corpus and only change the training of the language model between T-L and O-L models, while in the neural and unsupervised models, the target-language corpus directly influences the learned translation probabilities. Yet we still see significant improvement in T-L models in the supervised phrase-based systems, suggesting that the choice of target-language corpus has the most impact in the language modeling aspect of translation. In neural translation, language models can be used as denoisers in the decoding stage as in (8); future work could investigate if a T-L language model added at this stage could further improve translation quality.

Note that in general, our results are not comparable with (9), (8), or other prior work given the datasets used; we expect our BLEU scores to generally be lower due to the comparatively small corpora used here. Because standard benchmark datasets (like those put up by WMT) do not indicate translation direction, most work dealing with translationese also cannot be compared against other prior papers; given that translationese has a clear effect on translation quality, this should indicate a need for more labeling of translation direction in standard corpora.

6 Conclusions

We examine the effects of translationese on three machine translation systems: supervised phrase-based SMT, neural translation, and unsupervised phrase-based SMT. We test translation of five languages of varying resource levels into English and find that models trained on translated text almost always outperform models trained on original target-language text, regardless of language, system, and resource level, with significant potential implications for translation out of low-resource languages. However, our results are difficult to compare to existing work in the field due to the lack of corpora with labeled translation direction. Further research into and better understanding of the effects of translationese, with better-labeled corpora, is needed to continue advancing the state of machine translation.

7 Additional info

Project mentor: Michael Hahn

References

[1] Mona Baker, Gill Francis, & Elena Tognini-Bonelli. Corpus Linguistics and Translation Studies: Implications and Applications. 1993. *Text and Technology: In Honour of John Sinclair*.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, & Tomas Mikolov. Enriching Word Vectors with Subword Information. 2017. *Transactions of the Association for Computational Linguistics* 5: 135-146.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, & Hervé Jégou. Word translation without parallel data. 2018. ICLR 2018.

[4] Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, & Ruslan Mitkov. Identification of Translationese: A Machine Learning Approach. 2010. *Computational Linguistics and Intelligent Text Processing* 2010: 503-511.

[5] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. 2005. MT Summit 2005.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, & Evan Herbst. Moses: Open source toolkit for statistical machine translation. 2007. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*: 177-180.

[7] Moshe Koppel & Noam Ordan. Translationese and Its Dialects. 2011. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: 1318-1326.

[8] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, & Marc'Aurelio Ranzato. Phrase-Based & Neural Unsupervised Machine Translation. 2018. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*: 5039–5049.

[9] Gennadi Lembersky, Noam Ordan, & Shuly Wintner. Language Models for Machine Translation: Original vs. Translated Texts. 2011. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*: 363-374.

[10] Anthony Pym & Grzegorz Chrupala. The quantitative analysis of translation flows in the age of an international language. 2005. *Less Translated Languages*: 27-38.

[11] Ella Rabinovich & Shuly Wintner. Unsupervised Identification of Translationese. 2015. *Transactions of the Association for Computational Linguistics* 3: 419–432.

[12] Antonio Toral, Sheila Castilho, Ke Hu, & Andy Way. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. 2018. *Proceedings of the Third Conference on Machine Translation: Research Papers*: 113-123.

[13] Gideon Toury. In search of a theory of translation. 1980.

[14] Michael Ustaszewski. Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. 2018. *Perspectives: Studies in Translation Theory and Practice* 27 (1): 107-23.

[15] Vered Volansky, Noam Ordan, & Shuly Wintner. On the features of translationese. 2015. *Literary and Linguistic Computing* 30(1): 98-118.