Asymmetric Features of Human Generated Translation

Sauleh Eetemadi
Michigan State University, East Lansing, MI
Microsoft Research, Redmond, WA
saulehe@microsoft.com

Kristina Toutanova
Microsoft Research
Redmond, WA
kristout@microsoft.com

Abstract

Distinct properties of translated text have been the subject of research in linguistics for many year (Baker, 1993). In recent years computational methods have been developed to empirically verify the linguistic theories about translated text (Baroni and Bernardini, 2006). While many characteristics of translated text are more apparent in comparison to the original text, most of the prior research has focused on monolingual features of translated and original text. The contribution of this work is introducing bilingual features that are capable of explaining differences in translation direction using localized linguistic phenomena at the phrase or sentence level, rather than using monolingual statistics at the document level. We show that these bilingual features outperform the monolingual features used in prior work (Kurokawa et al., 2009) for the task of classifying translation direction.

1 Introduction

It has been known for many years in linguistics that translated text has distinct patterns compared to original or authored text (Baker, 1993). The term "Translationese" is often used to refer to the characteristics of translated text. Patterns of Translationese can be categorized as follows (Volansky et al., 2013):

1. **Simplification**: The process of translation is often coupled with a simplification process at several levels. For example, there tends to be less lexical variety in translated text and rare words are often avoided.
2. **Explicitation**: Translators often have to be more explicit in their translations due to lack of the cultural context that speakers of the source language have. Another manifestation of this pattern is making arguments more explicit which can be observed in the heavy use of cohesive markers like "therefore" and "moreover" in translated text (Koppel and Ordan, 2011).
3. **Normalization**: Translated text often contains more formal and repeating language.
4. **Interference**: A translator is likely to produce a translation that is structurally and grammatically closer to the source text or their native language.

In Figure 1 the size of a word in the "Translated" section is proportional to the difference between the frequency of the word in original and in the translated text (Fellows, 2013). For example, it is apparent that the word "the" is over-represented in translated English as noted by other research (Volansky et al., 2013). In addition, cohesive markers are clearly more common in translated text.

In the past few years there has been work on machine learning techniques for identifying Translationese. Standard machine learning algorithms like SVMs (Baroni and Bernardini, 2006) and Bayesian Logistic Regression (Koppel and Ordan, 2011) have been employed to train classifiers for one of the following tasks:

i. Given a chunk of text in a specific language, classify it as "Original" or "Translated".
ii. Given a chunk of translated text, predict the source language of the translation.
iii. Given a text chunk pair and their languages, predict the direction of translation.

There are two stated motivations for the tasks above: first, empirical validation of linguistic theories about Translationese (Volansky et al., 2013), and second, improving statistical machine translation by leveraging the knowledge of the translation direction in training and test data (Lember-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 159–164,
October 25-29, 2014, Doha, Qatar.

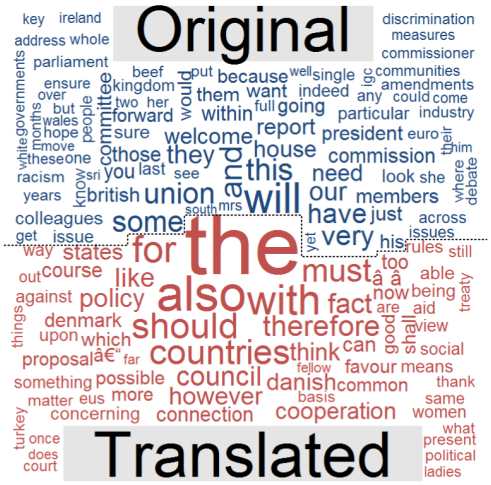


Figure 1: EuroParl Word Cloud Data Visualization (Translated vs Original) [1]

sky et al., 2012a; Lembersky et al., 2013; Lembersky et al., 2012b). Few parallel corpora including a customized version of EuroParl (Islam and Mehler, 2012) and a processed version of Hansard (Kurokawa et al., 2009) are labeled for translated versus original text. Using these limited resources, it has been shown that taking the translation direction into account when training a statistical machine translation system can improve translation quality (Lembersky et al., 2013). However, improving statistical machine translation using translation direction information has been limited by several factors.

1. **Limited Labeled Data**: The amount of labeled data is limited by language and domain and therefore by itself is not enough to make a significant improvement in statistical machine translation.
2. **Cross-Domain Scalability**: Current methods of Translationese detection do not scale across different corpora. For example, a classifier trained on EuroParl corpus (Koehn, 2005) had in-domain accuracy of 92.7% but out-of-domain accuracy of 64.8% (Koppel and Ordan, 2011).
3. **Text Chunk Size**: The reported high accuracy of Translationese detection is based on relatively large (approximately 1500 tokens) text chunks (Koppel and Ordan, 2011). When similar tasks are performed at the sentence level the accuracy drops by 15 percentage points or more (Kurokawa et al., 2009). Figure 2 shows how detection accuracy drops with the reduction of the input text chunk size. Since parallel data are often available at the sentence level or small chunks of text, existing detection methods aren't suitable for this type of data.

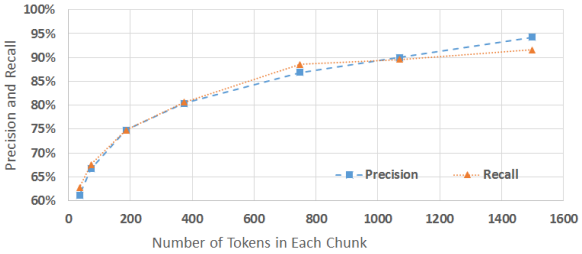


Figure 2: Effects of Chunk Size on Translationese Detection Accuracy[2]

Motivated by these limitations, in this work we focus on improving sentence-level classification accuracy by using non-domain-specific bilingual features at the sentence level. In addition to improving accuracy, these fine-grained features may be better able to confirm existing theories or discover new linguistic phenomena that occur in the translation process. We use a fast linear classifier trained with online learning, Vowpal Wabbit (Langford et al., 2007). The Hansard French-English dataset (Kurokawa et al., 2009) is used for training and test data in all experiments.

2 Related Work

While distinct patterns of Translationese have been studied widely in the past, the work of Baroni and Bernardini (2006) is the first to introduce a computational method for detecting Translationese with high accuracy. Prior work has shown in-domain accuracy can be very high at the chunk-level if fully lexicalized features are used (Volansky et al., 2013), but then the phenomena learned are clearly not generalizable across domains. For example, in Figure 1, it can be observed that content words like "commission", "council" or "union" can be used effectively for classification while they do not capture any general linguistic phenomena and are unlikely to scale

[1]This word cloud was created using the *wordcloud* and *tm* **R** packages (Fellows, 2013) from EuroParl parallel data annotated for translation direction (Islam and Mehler, 2012) obtained from http://www.hucompute.org/ressourcen/corpora/56.

[2]This is a reproduction of the results of Koppel and Ordan (2011) using function word frequencies as features for a logistic regression classifier. Based on the description of how text chunks were created, the results of the paper (92.7% accuracy) are based on text chunk sizes of approximately 1500 tokens.

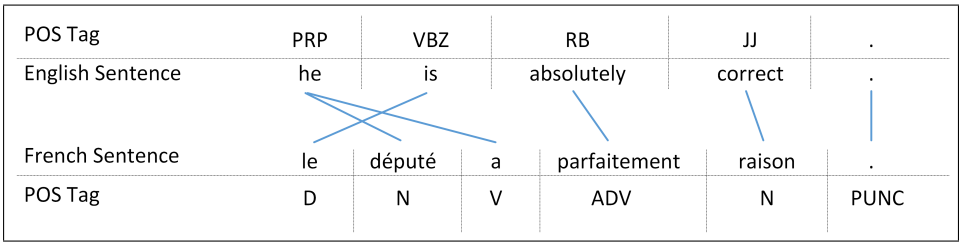


Figure 3: POS Tagged Aligned Sentence Pairs

to other corpora. This is also confirmed by an average human performance of 72.7% precision with 82.1% recall on a similar task where the test subjects were not familiar with the domain and were not able to use domain-specific lexical features (Baroni and Bernardini, 2006). A more general feature set still with high in-domain accuracy is POS tags with lexicalization of function words (Baroni and Bernardini, 2006; Kurokawa et al., 2009). We build on this feature set and explore bilingual features.

The only work to consider features of the two parallel chunks (one original, one translated) is the work of Kurokawa et al. (2009). They simply used the union of the n-gram mixed-POS[3] features of the two sides; these are monolingual features of the original and translated text and do not look at translation phenomena directly. Their work is also the only work to look at sentence level detection accuracy and report 15 percentage points drop in accuracy when going from chunk level to sentence level classification.

3 Bilingual Features for Translation Direction Classification

We are interested in learning common localized linguistic phenomena that occur during the translation process when translating in one direction but not the other.

3.1 POS Tag MTUs

Minimal translation units (MTUs) for a sentence pair are defined as pairs of source and target word sets that satisfy the following conditions (Quirk and Menezes, 2006).

1. No alignment links between distinct MTUs.
2. MTUs are not decomposable into smaller MTUs without violating the previous rule.

We use POS tags to capture linguistic structures and MTUs to map linguistic structures of the two languages. To obtain POS MTUs from a parallel corpus, first, the parallel corpus is word aligned. Next, the source and target side of the corpus are tagged independently. Finally, words are replaced with their corresponding POS tag in word-aligned sentence pairs. MTUs were extracted from the POS tagged word-aligned sentence pairs from left to right and listed in source order. Unigram, bi-gram, and higher order n-gram features were built over this sequence of POS MTUs. For example, for the sentence pair in Figure 3, the following POS MTUs will be extracted: `VBZ⇒D, PRP⇒(N,V), RB⇒ADV, JJ⇒N, .⇒PUNC`.

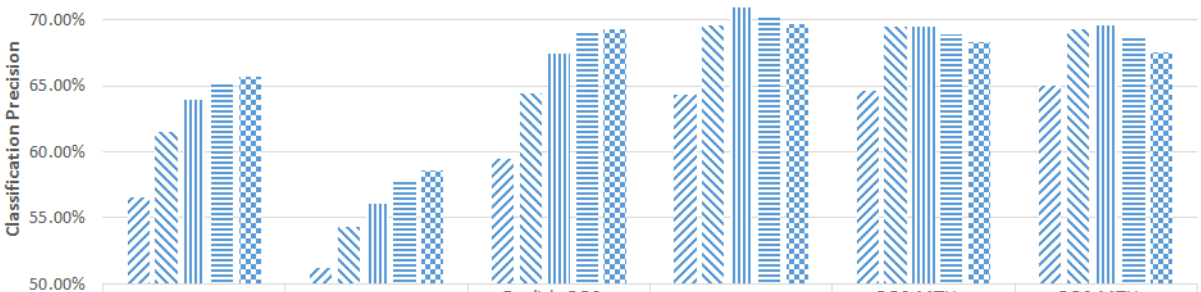
3.2 Distortion

In addition to the mapping of linguistic structures, another interesting phenomenon is the reordering of linguistic structures during translation. One hypothesis is that when translating from a fixed-order to a free-order language, the order of the target will be very influenced by the source (almost monotone translation), but when translating into a fixed order language, more re-ordering is required to ensure grammaticality of the target. To capture this pattern we add distortion to POS Tag MTU features. We experiment with absolute distortion (word position difference between source and target of a link) as well as HMM distortion (word position difference between the target of a link and the target of the previous link). We bin the distortions into three bins: "$= 0$", "> 0" and "< 0", to reduce sparsity.

4 Experimental Setup

For the translation direction detection task explained in section 1, we use a fast linear classifier trained with online learning, Vowpal Wabbit (Langford et al., 2007). Training data and classification features are explained in section 4.1 and 4.2.

[3]Only replacing content words with their POS tags while leaving function words as is.



	English-POS	French-POS	English-POS + French-POS	POS-MTU	POS-MTU + Absolute Distortion	POS-MTU + HMM Distorition
1-gram	56.59%	51.20%	59.46%	64.37%	64.58%	65.01%
2-gram	61.48%	54.27%	64.43%	69.60%	69.47%	69.29%
3-gram	63.93%	56.04%	67.43%	70.95%	69.45%	69.57%
4-gram	65.28%	57.72%	69.01%	70.29%	68.91%	68.58%
5-gram	65.71%	58.61%	69.32%	69.63%	68.36%	67.55%

Figure 4: Sentence level translation direction detection precision using different features with n-gram lengths of 1 through 5.

4.1 Data

For this task we require a parallel corpus with sentence pairs available in both directions (sentences authored in language A and then translated to language B and vice versa). While the customized version of EuroParl (Islam and Mehler, 2012) contains sentence pairs for many language pairs, none of the language pairs have sentence pairs available in both directions (e.g., it does contain sentences authored in English and translated into French but not vice versa). The Canadian Hansard corpus on the other hand fits the requirement as it has 742,408 sentence pairs translated from French to English and 2,203,504 sentences pairs that were translated from English to French (Kurokawa et al., 2009). We use the Hansard data for training classifiers. For training the HMM word alignment model used to define features, we use a larger set of ten billion words of parallel text from the WMT English-French corpus.

4.2 Preprocessing and Feature Extraction

We used a language filter[4], deduplication filter[5] and length ratio filter to clean the data. After filtering we were left with 1,890,603 English-French sentence pairs and 640,117 French-English sentence pairs. The Stanford POS tagger (Toutanova and Manning, 2000) was used to tag the English and the French sides of the corpus. The HMM alignment model (Vogel et al., 1996) trained on WMT data was used to word-align the Hansard corpus while replacing words with their corresponding POS tags. Due to differences in word breaking between the POS tagger tool and our word alignment tool there were some mismatches. For simplicity we dropped the entire sentence pair whenever a token mismatch occurred. This left us with 401,569 POS tag aligned sentence pairs in the French to English direction and 1,184,702 pairs in the other direction. We chose to create a balanced dataset and reduced the number of English-French sentences to 401,679 with 20,000 sentence pairs held out for testing in each direction.

5 Results

The results of our experiments on the translation direction detection task are listed in Table 4. We would like to point out several results from the table. First, when using only unigram features, the highest accuracy is achieved by the "POS-MTU + HMM Distortion" feature, which uses POS minimal translation units together with distortion. The highest accuracy overall if obtained by a "POS-MTU" trigram model, showing the advantage of bilingual features over prior work using only a union of monolingual features (reproduced by the "English-POS + French-POS" configuration). While higher order features generally show better in-domain accuracy, the advantage of low-order bilingual features might be even higher in cross-domain classification.

[4]A character n-gram language model is used to detect the language of source and target side text and filter them out if they do not match their annotated language.

[5]Duplicate sentences pairs are filtered out.

[6]For description of English POS tags see (Marcus et al., 1993) and (Abeillé et al., 2003) for French

	POS MTU (E⇒F)	FE#	EF#	Example
1	NNPS⇒(N,C)	336	12	quebecers(NNPS) ⇒ québécoises(N) et(C) des québécois
2	IN⇒(CL,V)	69	1027	a few days ago(IN) ⇒ il y(CL) a(V) quelques
3	PRP⇒(N,V)	18	663	he(PRP) is ⇒ le député(N) à(V)
4	(NNP,POS)⇒A	155	28	quebec(NNP) 's(POS) history ⇒ histoire québécoises(A)
5	(FW,FW)⇒ADV	7	195	pro(FW) bono(FW) work ⇒ bénévolement(ADV) travailler
6	(RB,MD)⇒V	2	112	money alone(RB) could(MD) solve ⇒ argent suffirait(V) à résoudre

Table 1: POS MTU features with highest weight. FE# indicates the number of times this feature appeared when translating from French to English.[6]

6 Analysis

An interesting aspect of this work is that it is able to extract features that can be linguistically interpreted. Although linguistic analysis of these features is outside the scope of this work, we list POS MTU features with highest positive or negative weights in Table 1. Although the top feature, NNPS⇒(N,C)[7], in this context is originating from a common phrase used by French speaking members of the Canadian Parliament, québécoises et des québécois, it does highlight an underlying linguistic phenomenon that is not specific to the Canadian Parliament. When translating a plural noun from English to French it is likely that only the masculine form of the noun appears, while if it was authored in French with both forms of the nouns, a single plural noun would appear in English as English doesn't have masculine and feminine forms of the word. A more complete form of this feature would have been NNPS⇒(N,C,N), but since word alignment models, in general, discourage one-to-many alignments, the extracted MTU only covers the first noun and conjunction.

7 Conclusion and Future Work

In this work we introduce new features for translation direction detection that leverage word alignment, source POS and target POS in the form of POS MTUs. POS MTUs are a powerful tool for capturing linguistic interactions between languages during the translation process. Since POS MTUs are not lexical features they are more likely to scale across corpora and domains compared to lexicalized features. Although most of the high weight POS MTU features used in classification (Table 1) are not corpus specific, unfortunately, due to lack of training data in multiple domains, experiments were not run to validate this claim. In future work, we intend to obtain training data from multiple domains that enables us to verify cross-domain scalability of POS-MTUs. In addition, observing linguistic phenomena that occur in one translation direction but not the other can be very informative in improving statistical machine translation quality. Another future direction for this work is leveraging sentence level translation direction detection to improve statistical machine translation output quality. Finally, further investigation of the linguistic interpretation of individual feature that are most discriminating between opposite translation directions can lead to discovery of new linguistic phenomena that occur during the translation process.

[7] NNPS: Plural Noun, N: Noun, C:Conjunction

Acknowledgement

The authors would like to thank Lee Schwartz for analyzing classification features and providing linguistic insight for them. We would like to also acknowledge the thoughtful comments and detailed feedback of the reviewers which helped us improve the paper.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In Anne Abeillé, editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 165–187. Springer Netherlands.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: in honour of John Sinclair*, 233:250.

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.

Ian Fellows, 2013. *wordcloud: Word Clouds*. R package version 2.4.

Zahurul Islam and Alexander Mehler. 2012. Customization of the europarl corpus for translation studies. In *LREC*, page 2505–2510.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

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-Volume 1*, page 1318–1326. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. *Proceedings. MT Summit XII, The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas*.

J Langford, L Li, and A Strehl, 2007. *Vowpal wabbit online learning project*.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012a. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 255–265. Association for Computational Linguistics.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012b. 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.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Chris Quirk and Arul Menezes. 2006. Do we need phrases?: Challenging the conventional wisdom in statistical machine translation. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Literary and Linguistic Computing*, page fqt031.