Detecting Translation Direction: A Cross-Domain Study

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

Parallel corpora are constructed by taking a document authored in one language and translating it into another language. However, the information about the authored and translated sides of the corpus is usually not preserved. When available, this information can be used to improve statistical machine translation. Existing statistical methods for translation direction detection have low accuracy when applied to the realistic out-of-domain setting, especially when the input texts are short. Our contributions in this work are threefold: 1) We develop a multi-corpus parallel dataset with translation direction labels at the sentence level, 2) we perform a comparative evaluation of previously introduced features for translation direction detection in a cross-domain setting and 3) we generalize a previously introduced type of features to outperform the best previously proposed features in detecting translation direction and achieve 0.80 precision with 0.85 recall.

1 Introduction

Translated text differs from authored text (Baker, 1993). The main differences are simplification, explicitation, normalization and interference (Volansky et al., 2013). Statistical classifiers have been trained to detect Translationese[1]. Volansky et al. (2013) state two motivations for automatic detection of Translationese: empirical validation of Translationese linguistic theories and improving statistical machine translation (Kurokawa et al., 2009). Most of the prior work focus on in-domain Translationese detection (Baroni and Bernardini, 2006; Kurokawa et al., 2009). That is, the training and test set come from the same, usually narrow, domain. Cross-domain Translationese detection serves the two stated motivations better than in-domain detection. First, automatic classification validates linguistic theories only if it works independent of the domain. Otherwise, the classifier could perform well by memorizing lexical terms unique to a specific domain without using any linguistically meaningful generalizations. Second, a Translationese classifier can improve statistical machine translation in two ways: 1) By labeling the parallel training data with translation direction[2]; 2) By labeling input sentences to a decoder at translation time and use matching models. The accuracy of the classifier is the main factor determining its impact on statistical machine translation. Most parallel or monolingual training data sources do not contain translation direction meta-data. Also, the input sentences at translation time can be from any domain. Therefore, a cross-domain setting for translation direction detection is more appropriate for improving statistical machine translation as well. We develop a cross-domain training and test data set and compare some of the linguistically motivated features from prior work (Kurokawa et al., 2009; Volansky et al., 2013) in this setting. In addition, we introduce a new bilingual feature that outperforms all prior work in both

[1] Translated text is often referred to as "Translationese" (Volansky et al., 2013).

[2] Detection of *translation direction* refers to classifying a text block pair (A and B) as A was translated to B or vice versa. In contrast, *Translationese* detection usually refers to classifying a single block of text as "Translationese" versus "Original".

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 103–109,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

in-domain and cross-domain settings.

Our work also differs from many prior works by focusing on sentence level, rather than block level classification. Although Kurokawa et al. (2009) compare sentence level versus block level detection accuracy, most other research focuses on block level detection (Baroni and Bernardini, 2006; Volansky et al., 2013). Sentence level classification serves the stated motivations above better than block level classification. For empirical validation of linguistic theories, features that are detectable at the sentence level are more linguistically meaningful than block level statistics. Sentence level detection is also more appropriate for labeling decoder input as well as some statistical machine translation training data.

In the rest of the paper, we first review prior work on sentence level and cross-domain translation direction detection. In Section 3 we motivate the selection of features used in this study. Next, we describe our cross-domain data set and the classification algorithm we use to build and evaluate models given a set of features. Experimental results are presented in Section 5.2.

2 Related Work

Volansky et al. (2013) provide a comprehensive list of monolingual features used for Translationese detection. These features include POS n-grams, character n-grams, function word frequency, punctuation frequency, mean word length, mean sentence length, word n-grams and type/token ratio. We are aware of only one prior work that presented a cross-domain evaluation. Koppel and Ordan (2011) use a logistic regression classifier with function word unigram frequencies to achieve 92.7% accuracy with ten fold cross validation on the EuroParl (Koehn, 2005) corpus and 86.3% on the IHT corpus. However testing the EuroParl trained classifier on the IHT corpus yields an accuracy of 64.8% (and the accuracy is 58.8% when the classifier is trained on IHT and tested on EuroParl). The classifiers in this study are trained and tested on text blocks of approximately 1500 tokens, and there is no comparative evaluation of models using different feature sets.

We are also aware of two prior works that investigate Translationese detection accuracy at the sentence level. First Kurokawa et al (2009) use the Hansard English-French corpus for their experiments. For sentence level translation direction detection they reach F-score of 77% using word n-grams and stay slightly below 70% F-score with POS n-grams using an SVM classifier. Second, Eetemadi and Toutanova (2014) leverage word alignment information by extracting POS tag minimal translation units (MTUs) (Quirk and Menezes, 2006) along with an online linear classifier trained on the Hansard English-French corpus to achieve 70.95% detection accuracy at the sentence level.

Label	Description
ENG.LEX	English word n-grams
FRA.LEX	French word n-grams
ENG.POS	English POS Tag n-grams
FRA.POS	French POS Tag n-grmas
ENG.BC	English Brown Cluster n-grams
FRA.BC	French Brown Cluster n-grams
POS.MTU	POS MTU n-grams
BC.MTU	Brown Cluster MTU n-grams

Table 1: Classification features and their labels.

3 Feature Sets

The goal of our study is to compare novel and previously introduced features in a cross-domain setting. Due to the volume of experiments required for comparison, for an initial study, we select a limited number of feature sets for comparison. Prior works claim POS n-gram features capture linguistic phenomena of translation and should generalize across domains (Kurokawa et al., 2009; Eetemadi and Toutanova, 2014). We chose source and target POS n-gram features for $n = 1 \ldots 5$ to test this claim. Another feature we have chosen is from the work of Eetemadi and Toutanova (2014) where they achieve higher accuracy by introducing POS MTU[3] n-gram features.

POS MTUs incorporate source and target side information in addition to word alignment. Prior work has also claimed lexical features such as word n-grams do not generalize across domains due to corpus specific vocabulary (Volansky et al., 2013). We test this hypothesis using source and target word n-gram features. Using n-grams of length 1 through 5 we run 45 (nine data matrix entries times n-gram lengths of five) experiments for each feature set mentioned above.

In addition to the features mentioned above, we

[3] Minimal Translation Units (Quirk and Menezes, 2006)

Corpus	Authored Language	Translation Language	Training Sentences	Test Sentences
EuroParl	English	French	62k	6k
EuroParl	French	English	43k	4k
Hansard	English	French	1,697k	169k
Hansard	French	English	567k	56k
Hansard-Committees	English	French	2,930k	292k
Hansard-Committees	French	English	636k	63k

Table 2: Cross-Domain Data Sets

make a small modification to the feature used to obtain the best previously reported sentence level performance (Eetemadi and Toutanova, 2014) to derive a new type of features. POS MTU n-gram features are the most linguistically informed features amongst prior work. We introduce Brown cluster (Brown et al., 1992) MTUs instead. Our use of Brown clusters is inspired by recent success on their use in statistical machine translation systems (Bhatia et al., 2014; Durrani et al., 2014). Finally, we also include source and target Brown cluster n-grams as a comparison point to better understand their effectiveness compared to POS n-grams and their contribution to the effectiveness of Brown cluster MTUs.

Given these 8 feature types summarized in Table 1, n-gram lengths of up to 5 and the 3×3 data matrix explained in the next section, we run 360 experiments for this cross-domain study.

4 Data, Preprocessing and Feature Extraction

We chose the English-French language pair for our cross-domain experiments based on prior work and availability of labeled data. Existing sentence-parallel datasets used for training machine translation systems, do not normally contain gold-standard translation direction information, and additional processing is necessary to compile a dataset with such information (labels). Kurokawa et al (2009) extract translation direction information from the English-French Hansard parallel dataset using speaker language tags. We use this dataset, and treat the two sections “main parliamentary proceedings” and “committee hearings” as two different corpora. These two corpora have slightly different domains, although they share many common topics as well. We additionally choose a third corpus, whose domain is more distinct from these two, from the EuroParl English-French corpus. Islam and Mehler (2012) provided a customized version of Europarl with translation direction labels, but this dataset only contains sentences that were authored in English and translated to French, and does not contain examples for which the original language of authoring was French. We thus prepare a new dataset from EuroParl and will make it publicly available for use. The original unprocessed version of EuroParl (Koehn, 2005) contains speaker language tags (original language of authoring) for the French and English sides of the parallel corpus. We filter out inconsistencies in the corpus. First, we filter out sections where the language tag is missing from one or both sides. We also filter out sections with conflicting language tags. Parallel sections with different number of sentences are also discarded to maintain sentence alignment. This leaves us with three data sets (two Hansard and one EuroParl) with translation direction information available, and which contain sentences authored in both languages. We hold out 10% of each data set for testing and use the rest for training. Our 3×3 corpus data matrix consists of all nine combinations of training on one corpus and testing on another (Table 2).

4.1 Preprocessing

First, we clean all data sets using the following simple techniques.

- Sentences with low alphanumeric density are discarded.
- A character n-gram based language detection tool is used to identify the language of each sentence. We discard sentences with a detected language other than their label.
- We discard sentences with invalid unicode characters or control characters.
- Sentences longer than 2000 characters are excluded.

Next, an HMM word alignment model (Vogel et al., 1996) trained on the WMT English-French corpus (Bojar et al., 2013) word-aligns sentence pairs.

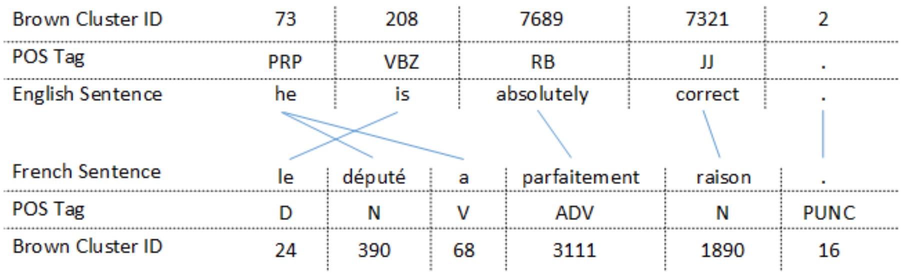

Figure 1: POS Tagged and Brown Cluster Aligned Sentence Pairs

We discard sentence pairs where the word alignment fails. We use the Stanford POS tagger (Toutanova and Manning, 2000) for English and French to tag all sentence pairs. A copy of the alignment file with words replaced with their POS tags is also generated. French and English Brown clusters are trained separately on the French and English sides of the WMT English-French corpus (Bojar et al., 2013). The produced models assign cluster IDs to words in each sentence pair. We create a copy of the alignment file with cluster IDs instead of words as well.

4.2 Feature Extraction

The classifier of our choice (Section 5) extracts n-gram features with n specified as an option. In preparation for classifier training and testing, feature extraction only needs to produce the unigram features while preserving the order (n-grams of higher length are automatically extracted by the classifier). POS, word, and Brown cluster n-gram features are generated by using the respective representation for sequences of tokens in the sentences. For POS and Brown cluster MTU features, the sequence of MTUs is defined as the left-to-right in source order sequence (due to reordering, the exact enumeration order of MTUs matters). For example, for the sentence pair in Figure 1, the sequence of Brown cluster MTUs is: 73⇒(390,68), 208⇒24, 7689⇒3111, 7321⇒1890, 2⇒16.

5 Experiments

We chose the Vowpal Wabbit (Langford et al., 2007) (VW) online linear classifier since it is fast, scalable and it has special (bag of words and n-gram generation) options for text classification. We found that VW was comparable in accuracy to a batch logistic regression classifier. For training and testing the classifier, we created balanced datasets with the same number of training examples in both directions. This was achieved by randomly removing sentence pairs from the English to French direction until it matches the French to English direction. For example, 636k sentence pairs are randomly chosen from the 2,930k sentence pairs in English to French Hansard-Committees corpus to match the number of examples in the French to English direction.

5.1 Evaluation Method

We are interested in comparing the performance of various feature sets in translation direction detection. Performance evaluation of different classification features objectively is challenging in the absence of a downstream task. Specifically, depending on the preferred balance between precision and recall, different features can be superior. Ideally an ROC graph (Fawcett, 2006) visualizes the tradeoff between precision and recall and can serve as an objective comparison between different classification feature sets. However, it is not practical to present ROC graphs for 360 experiments. Hence, we resort to the Area Under the ROC graph (AUC) measure as a good measure to provide an objective comparison. Theoretically, the area under the curve can be interpreted as the probability that the classifier scores a random negative example higher than a random positive example (Fawcett, 2006). As a point of reference, we also provide F-scores for experimental settings that are comparable to the prior work reviewed in Section 2.

5.2 Results

Figure 2 presents AUC points for all experiments. Rows and columns are labeled with corpus names for training and test data sets respectively. For example, the graph on the third row and first column corresponds to training on the Hansard-Committees corpus and testing on EuroParl. Within each graph we compare the AUC performance of different fea-

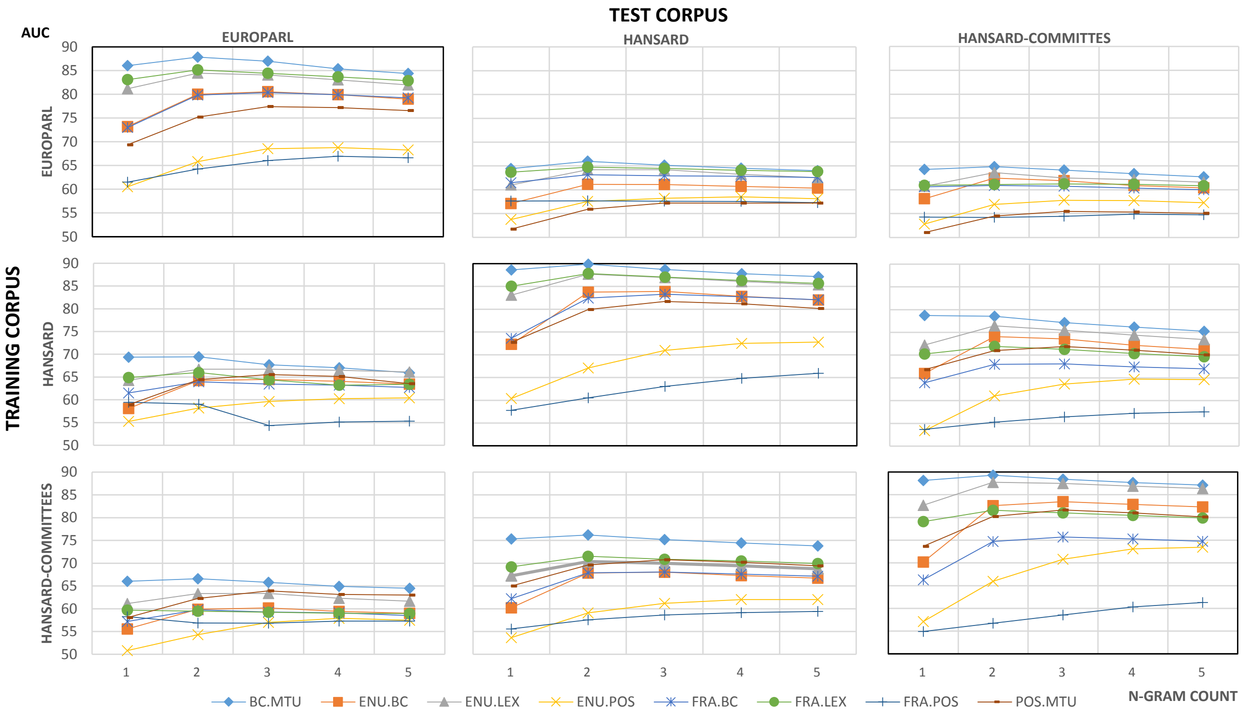

Figure 2: Comparing area under the ROC curve for the translation direction detection task when training and testing on different corpora using each of the eight feature sets. See Table 1 for experiment label description.

tures with n-gram lengths of 1 through 5.

Graphs on the diagonal correspond to in-domain detection and demonstrate higher performance compared to off diagonal graphs. This confirms the basic assumption that cross-domain translation direction detection is a more difficult task. The overall performance is also higher when trained on the Hansard corpus and tested on Hansard-Commitee and vice versa. This is because the Hansard corpus is more similar to the Hansard-Committees corpus compared to the EuroParl corpus. It is also observable that the variation in performance of different features diminishes as the training and test corpora become more dissimilar. For instance, this phenomenon can be observed on the second row of graphs where the features are most spread out when tested on the Hansard corpus. They are less spread out when tested on the Hansard-Committees corpus, and compressed together when tested on the EuroParl corpus. The same phenomenon can be observed for classifiers trained on other corpora.

For different feature types, different n-gram order of the features is best, depending on the feature granularity. To make it easier to observe patterns in the performance of different feature types, Figure 3 shows the performance for each feature type and each train-test corpus combination as a single point, by using the best n-gram order for that feature/data combination. Each of the 9 train/test data combinations is shown as a curve over feature types.

We can see that MTU features (which look at both languages at the same time) outperform individual source or target features (POS or Brown cluster) for all datasets. Brown clusters are unsupervised and can provide different levels of granularity. On the other hand, POS tags usually provide a fixed granularity and require lexicons or labeled data to train. We see that Brown clusters outperform corresponding POS tags across data settings. As an example, when training and testing on the Hansard corpus FRA.BC outperforms FRA.POS by close to 20 AUC points.

Lexical features outperform monolingual POS and Brown cluster features in most settings although their advantages diminish as the training and test corpus become more dissimilar. This is somewhat contrary to prior claims that lexical features will not generalize well across domains – we see that lexical features do capture important generalizations across domains and models that use only POS tag features have lower performance, both in and out-of-domain.

Figure 4 shows the rank of each feature amongst

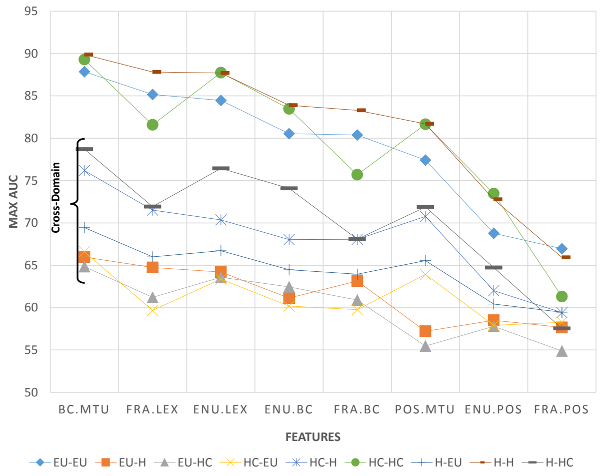

Figure 3: Translation detection performance matrix for training and testing on three different corpora - We ran experiments for n-grams of up to length five for each feature (See Table 1 for feature label descriptions). Unlike Figure 2 where we report AUC values for all n-gram lengths, in this graph we only present the highest AUC number for each feature. Each marker type indicates a training and test set combination. The format of experiment labels in the legend is [TrainingSet]-[TestSet] and **EU**: EuroParl, **H**: Hansard, **HC**: Hansard Committees. For example, EU-HC means training on EuroParl corpus and testing on Hansard Committees corpus.

all 8 different features for each entry in the cross-corpus data matrix (Similar to Figure 3 the highest performing n-gram length has been chosen for each feature). Brown cluster MTUs outperform all other features with rank one in all dataset combinations. Source and target POS tag features are the lowest performing features in 8 out of 9 data set combinations. The POS.MTU has its lowest ranks (7 and 8) when it is trained on the EuroParl corpus and its highest ranks (2 and 3) when trained on the Hansard-Committees corpus. High number of features in POS.MTU requires a large data set for training. The variation in performance for POS.MTU can be explained by the significant difference in training data size between EuroParl and Hansard-Committees. Finally, while FRA.LEX and ENG.LEX are mostly in rank 2 and 3 (after BC.MTU) they have their lowest ranks (6 and 4) in cross-corpus settings (HC-EU and HC-H).

Finally, we report precision and recall numbers to enable comparison between our experiments and previous work reported in Section 2. When training and testing on the Hansard corpus, BC.MTU achieves 0.80 precision with 0.85 recall. In comparison, ENG.POS achieves 0.65 precision with 0.64 recall and POS.MTU achieves 0.73 precision and 0.74 recall. These are the highest performance of each feature with n-grams of up to length 5.

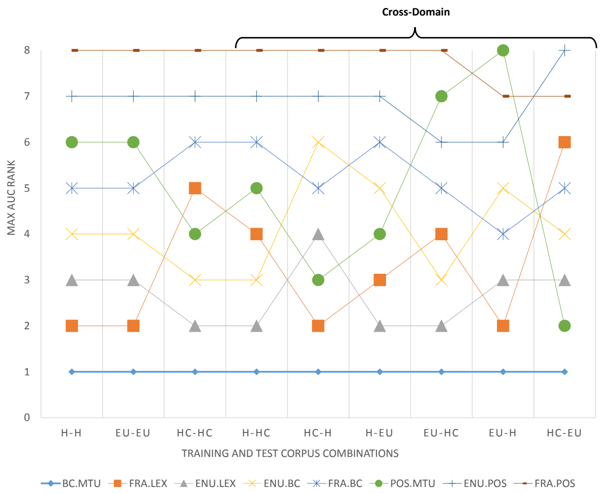

Figure 4: Translation direction detection AUC performance rank for each training and test set combination. For corpus combination abbreviations see description of Figure 3. For feature label descriptions see Table 1.

6 Conclusion and Future Work

From among eight studied sets of features, Brown cluster MTUs were the most effective at identifying translation direction at the sentence level. They were superior in both in-domain and cross-domain settings. Although English-Lexical features did not perform as well as Brown cluster MTUs, they performed better than most other methods. In future work, we plan to investigate lexical MTUs and to consider feature sets containing any subset of the eight or more basic feature types we have considered here. With these experiments we hope to gain further insight into the performance of feature sets in in out out-of-domain settings and to improve the state-of-the-art in realistic translation direction detection tasks. Additionally, we plan to use this classifier to extend the work of Twitto-Shmuel (2013) by building a more accurate and larger parallel corpus labeled for translation direction to further improve SMT quality.

References

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.

Austin Matthews Waleed Ammar Archna Bhatia, Weston Feely, Greg Hanneman Eva Schlinger Swabha Swayamdipta, Yulia Tsvetkov, and Alon Lavie Chris Dyer. 2014. The cmu machine translation systems at wmt 2014. *ACL 2014*, page 142.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44.

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.

Nadir Durrani, Philipp Koehn, Helmut Schmid, and Alexander Fraser. 2014. Investigating the usefulness of generalized word representations in smt. In *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING), Dublin, Ireland*, pages 421–432.

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.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

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*.

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.

Naama Twitto-Shmuel. 2013. *Improving Statistical Machine Translation by Automatic Identification of Translationese*. Ph.D. thesis, University of Haifa.

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 31.