Automatic Detection of Translated Text and its Impact on Machine Translation

David Kurokawa, Cyril Goutte and Pierre Isabelle
National Research Council, Interactive Language Technologies
283 Taché Blvd., Gatineau QC, Canada J8X 3X7
davidkurokawa@gmail.com
Cyril.Goutte,Pierre.Isabelle@nrc-cnrc.gc.ca

Abstract

We investigate the possibility of automatically detecting whether a piece of text is an original or a translation. On a large parallel English-French corpus where reference information is available, we find that this is possible with around 90% accuracy. We further study the implication this has on Machine Translation performance. After separating our corpus according to translation direction, we train direction-specific phrase-based MT systems and show that they yield improved translation performance. This suggests that taking directionality into account when training SMT systems may have a significant effect on output quality.

1 Introduction

In this paper, we address two main questions. First, is there a sufficient, detectable difference between texts originally written in a language and texts produced by translators from another language? Second, if we can reliably distinguish between original and translation, what impact can this knowledge have on a Statistical Machine Translation (SMT) system?

The differences between original and human-translated text, and the ability to detect those, have been considered for some time (cf. Baroni and Bernardini (2006) and references therein). The effects of these differences on machine translation however, have not been extensively studied.

We use a large portion of the proceedings of the Canadian parliament, a bilingual English-French parallel resource in which the original language is indicated. Using Support Vector Machine (SVM) classifiers, we show that we can reliably identify whether English or French was the original, based either on the French text, the English text or both. We obtain classification accuracies of up to 90% based on n-gram words. On single sentences, the classification accuracy reaches a lower, but still respectable 77%.

Having established the presence and detectability of differences between original and translated text, we explore how this can impact the performance of SMT systems. We find that the original language does impact translation quality. A system trained only on the data that was originally in French gets roughly the same performance as a system trained on the entire corpus, but uses only 1/5th of the training data. We exploit this in a "mixture of experts" setting, where specialized SMT systems are selected depending on whether the SVM predicts the text to be originally French or English. This yields a gain of around 0.6 BLEU (Papineni et al., 2002) points over the baseline SMT trained on the entire corpus. Moreover, we find that the SVM performance is sufficient to ensure virtually no loss compared to the reference source language information.

The next section discusses the problem and positions our work w.r.t. earlier results. Section 3 describes the data and models we used. Section 4 reports on how we learn to detect the translation direction based on either blocks of text or single sentences. Section 5 shows how the translation direction can impact the quality of SMT output.

2 Problem Setting

Many readers will have experienced that translated text often feels different from an original text, and even carries a "flavour" of the original, a property sometimes called *translationese*. The detectability and differences between original and human-translated text has been considered for example by Baroni and Bernardini (2006) and van Halteren (2008), in slightly different settings. Baroni and Bernardini use a monolingual corpus of Italian texts containing both originals and translations from a variety of language. They evaluate the performance of a SVM to detect original vs. translation and analyse various linguistic effects and features. van Halteren (2008) uses a multilingual parallel corpus (Europarl) containing 6 languages, and focuses on finding actual textual markers indicating that a text is a translation, and which language it is translated from.

Our work is closely related, and we use SVM as they do, but there are also crucial differences. One is that we work from a bilingual corpus, which constrains the problem somewhat. On the other hand, we use a much larger corpus (at least 30 times larger). This may contribute to the higher accuracy we obtained on the task of detecting translation direction (sec. 4.2). We also investigate which n-grams constitute important clues as to whether a text is an original or a translation (sec. 4.3).

On the classification task, one important contribution is that we explore how the classification performance depends on the size of the textual unit we consider. In particular, we obtain 77% accuracy and F-score when working from single sentences.

Finally, one main contribution is that we evaluate the impact of our classifiers on a specific task, namely the use of Machine Translation (sec. 5).

3 Data and Methods

We work with the multi-document transcripts of sessions of the Canadian parliaments (a.k.a. Canadian Hansard) containing much (but not all) of the 35th to 39th parliaments, spanning years 1996-2007. We use it for three main reasons. First and foremost, this bilingual English-French corpus is tagged with indications of the original language of the speaker, an essential requirement for our purposes. Second, the Hansard translations are generally considered to be top quality, hence ruling out the possibility that we are picking up on bad translation. Finally, there is no shortage of quality translation: we were able to extract around 4.5 million aligned sentence pairs (table 1), for a total of around 80 million words.

	fo	eo	mx
words (fr)	14,648K	72,054K	86,702K
words (en)	13,002K	64,899K	77,901K
sentences	902,349	3,668,389	4,570,738
blocks	40,538	42,750	83,288

Table 1: Breakdown of the raw usable data by original language (fo/eo=French/English original; mx=All data).

The corpus is not without its share of problems however. One is the (infrequent) inconsistency of the source language tags, i.e. when both sides claim to be the original. Although we simply did not use such sections, they put into question the reliability of the tags. There also appeared to be missing portions of text, causing perfectly usable English or French sentences to be aligned with blank lines. It is uncertain whether other problems exist and what, if any, effect such problems would have on the results. Finally, and perhaps less of an issue and more of an idiosyncrasy of the corpus, is the imbalance between the number of English original and French original sentences (4:1 ratio).

3.1 Preprocessing

As an initial pre-processing step, we first lowercased, tokenized and sentence-aligned the corpus using NRC's tools. The tokenizer is standard, and the sentence aligner implements the well-known Gale and Church algorithm (Church and Gale, 1991). As an additional step, alignments with null sentences on either side were removed.

We considered the data at two levels of granularity. For the simpler, *sentence level* variant, each line of the aligned corpus, is considered a basic *textual unit*, also called a fragment. We also used larger units containing consecutive sequences of sentences with the same original language, henceforth called *blocks*. This yields our *block level* results. Note that by definition, blocks are of very different length, containing 3 to several thousand words each. In addition, as one block of French-original (fo) text is usually followed by a block of English-original (eo)

text, there are by construction nearly equal numbers of blocks, despite the fact that there are 4 times as many eo sentences, cf. table 1. This obviously indicates that fo blocks are on average considerably shorter than their eo counterparts.

The corpus was tagged with part-of-speech (POS) using the freely available tree-tagger (Schmid, 1994). We produced 4 versions of the corpus depending on the representation of each token: 1) word, 2) lemma, 3) POS and 4) mixed. The mixed representation replaces content words (nouns, verbs, etc.) with the corresponding POS, while the grammatical words are kept in their original surface form. The idea, inspired by Baroni and Bernardini (2006), is to abstract from contextual or lexical clues and focus on the linguistic description of the text.

3.2 Statistical Classifier

We used Support Vector Machines (SVM) with a linear kernel. This choice is motivated in part by the state-of-the-art performance of SVM on many binary text classification tasks, and in part by the fact that Baroni and Bernardini (2006) and van Halteren (2008) reported good performance on similar tasks with these models.

An SVM with a linear kernel produces a score that is a linear combination of the feature values, on which the classification decision is made. One attractive aspect of SVMs is that there are a number of theoretical results (Cristianini and Shawe-Taylor, 2000) that ensure that the classifier has good generalization performance over unseen examples. Traditional SVM training algorithms based on constrained optimization scale badly with the number of training data. However, recent research has shown how to speed up training considerably (Joachims, 2006; Bottou et al., 2007). In our experiments, we use the publicly available implementation of SVM-Perf[1] in order to train SVMs on 85K examples (block-level) to 1.8M examples (sentence-level) in manageable time.

In the experiments presented below, the features described above were just put in the correct input format for SVM-Perf, which we used in a totally straightforward manner, without any additional *ad hoc* tuning. We use a cross-validation setting, i.e. we split the training data into 10 folds, and estimate a model on each subset of 9 folds, testing on the remaining, left-out fold. This produces unbiased predictions on each example, albeit each obtained on 90% of the training data. These are the predictions analysed in the results section below.

3.3 Statistical Machine Translation

For our Machine Translation experiment, we use the National Research Council of Canada (NRC)'s phrase-based SMT system Portage (Ueffing et al., 2007).[2] This is a fairly straightforward PBMT system based on a log-linear model using four main components: a phrase table obtained from the word-aligned corpus using HMM, a target-language model, a distortion model and a sentence-length feature. Once the model is trained from a large aligned bilingual corpus, test translations are produced as the output of a beam-search decoder. While Portage features an optional rescoring component, no rescoring stage was used. The idea is to use Portage as a "black box" SMT model as much as possible.

4 Detection of Translation vs. Original

We first turn our attention to the problem of detecting translation direction, that is, whether a given text fragment is an original or a translation.

4.1 Feature space

The first step is to build the feature vectors from the 4 text representations described above (word, lemma, POS and mixed). We consider n-gram representations, with $n \in \{1, 2, 3, 4, 5\}$. Because of sparsity, we consider 4-grams and 5-grams for the POS and mixed representations only. As there are many infrequent n-grams in most feature spaces, we only consider n-grams that appear in at least 10 fragments (blocks or sentences). Table 2 contains a breakdown of the number of n-grams before and after thresholding in each representation. Not surprisingly, the word-based representations are sparsest and have more n-grams filtered out, with the POS representations at the opposite end of the spectrum.

In the block-level experiments, we also considered tf-idf in order to downweight common features.

[1] http://svmlight.joachims.org/svm_perf.html

[2] PORTAGE is made available by the NRC to Canadian universities for research and education purposes.

		English		
n	token	used	total	% used
1	word	37,598	156,767	23.98
1	lemma	20,745	44,944	46.16
1	mixed	262	289	90.66
1	POS	58	58	100.00
2	word	456,187	4030,175	11.32
2	lemma	361,667	2511,102	14.40
2	mixed	14,189	24,289	58.42
2	POS	2249	2731	82.35
3	word	887,787	19,593,088	4.53
3	lemma	858,944	15,239,737	5.64
3	mixed	138,641	420,496	32.97
3	POS	31,678	56,358	56.21
4	mixed	528,087	2865,923	18.43
4	POS	189,769	540,714	35.10
5	mixed	1000,100	10,717,832	9.33
5	POS	566,806	2823,960	20.07

Table 2: Total number of distinct n-grams before ("total") and after ("used") thresholding, when building the SVM feature vectors for English, on the block-level problem. Figures for French, bilingual and sentence level give a similar picture.

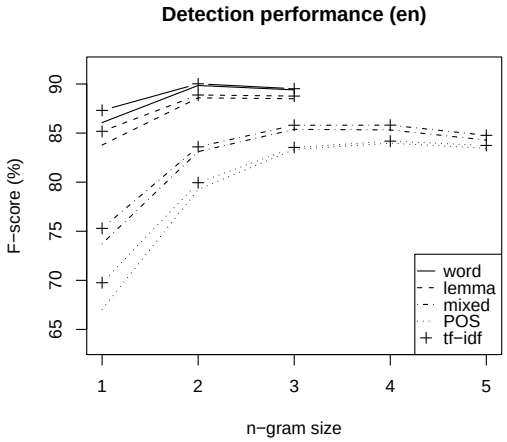

Figure 1: SVM performance on the English monolingual text (en) for various features (word, lemmas, POS and mixed) and n-gram length, with and without tf-idf

As its effect was found to be small but consistent, we only ran the sentence-level experiments with tf-idf turned on. In all experiments, the feature vectors were normalised w.r.t. the Euclidean norm. This was done on either the French side of the corpus, the English side, or on both (bilingual processing).

4.2 Results

Our experiments explored the effect of four key parameters: 1) language used (French/English/bilingual), 2) length of n-grams (i.e. n), 3) representation used as feature space (word, lemma, POS, or mixed) and 4) use of tf-idf.

Figure 1 shows results obtained on the English side of the corpus. Performance on French is very similar, only very slightly lower, while the use of both languages is slightly better (as expected), by 1-2%. The overall picture is identical, however. Note though that the bilingual case corresponds to a different use case where both sides are available. In all feature spaces, there seems to be an optimal n-gram length: bigram for words and lemmas, trigrams for POS and mixed. We attribute this to the fact that the sparsity introduced by using longer n-grams tends to offset the potential increase in performance of a richer representation. Also, note that we do not mix different n-gram length in the feature space.

Globally , the relationship between the feature representations is clear: word $>$ lemma $>$ mixed $>$ POS. The word bigram representations reach around 90% F-score (92% on the bilingual data). On this balanced dataset, accuracy is similar, meaning that the detection is correct for 9 out of 10 blocks. Although the performance of word and lemma representations may be helped by contextual or lexical cues (more on that later), we also notice that the POS and mixed representations, which focus solely on linguistic patterns, still reach around 85% F-score and accuracy (88% on the bilingual data).[3] This certainly shows that there are detectable differences in translated and original documents at the general, linguistic level.

For the comparatively expensive sentence-level problem, we reduced the number of experiments significantly by not testing 1-gram configurations, bigram POS and mixed, trigram mixed, and limiting ourselves to tf-idf. In addition, to counter the imbalance between the two classes at the sentence level, we sub-sampled the eo data in order to have roughly equal classes. Figure 2 shows that the performance drops dramatically on the sentence-level task. Again, this is not unexpected considering that the amount of information in sentences is much lower than in blocks. Yet, the F-score and ac-

[3] On balanced block-level data, random yields $\sim$50%

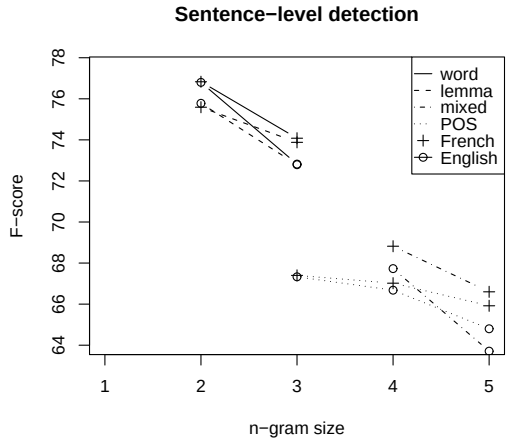

Figure 2: Performance on the sentence-level problem, for the combinations of experimental factors considered.

curacy reach around 77% using word bigrams. On this task, this is equivalent to predicting every other sentence correctly, while the other is guessed at random. This is no small achievement considering the limited amount of information in one sentence. Using the mixed representation, we stay slightly below 70% accuracy (and F-score).

4.3 Differences in Original and Translated Text

SVM classifiers provide good performance on the detection task, but do not provide clues about the actual differences between the two classes. In order to investigate that, we rely on the n-gram frequencies. We calculate the contribution of each n-gram to the symmetrised Kullback-Leibler divergence between the n-gram distributions for each class (Dobrokhotov et al., 2003), and use that to assess which n-grams were most characteristic of each original language. We report here on the results of our examination of the English side of the corpus for signs of eo or fo text.

Table 3 lists the most important bigram words according to this metric. One obvious pattern clue is the political parties. Speakers are identified by parties, and MPs from the Canadian Alliance or CPC (Conservatives) are more prominent in the western, English-speaking part of Canada, while MPs from the Bloc Quebecois (BQ), representing a French-speaking minority, tend to use French. More subtle, the expressions "a couple of" (combining bigrams #1 and #3) is highly indicative of eo text. We

eo	fo
couple_of	of_the
alliance_)	mr_.
a_couple	,_the
do_that	in_the
,_canadian	to_the
the_record	,_i
forward_to	._the
,_cpc	)_:
cpc_)	speaker_,
of_us	._i
this_country	:_mr
this_particular	,_and
many_of	._speaker
canadian_alliance	bq_)

Table 3: Bigram words most characteristic of each original language, on the English side of the corpus. Heavy use of articles and prepositions seems to be indicative of text translated from French; political parties (Canadian alliance, CPC, BQ) also provide important clues.

assume that because it has no direct equivalent in French, it is rarely produced as a translation, but is relatively frequently used by English speakers. We also notice the higher presence of the definite article *the* and prepositions in text translated from French. We conjecture that this is pure *translationese*: it is well-known that French uses more articles and prepositions than English. Apparently, translators have a tendency to carry this over, to a certain extent, in their translations. The method of address may be more culturally significant. French speakers appear to have a propensity to prefix titles such as "Mr." and "Honorary" to the names of their addressees.

Some of these considerations are backed up by a similar analysis of important POS and mixed 4-grams, which we do not have space to present here.

5 Impact on SMT

We now look at a practical situation where detecting the translation direction may be of use. In standard practice, when creating a MT system, the identity of the source language is not taken into consideration: all training data is used to create the model (in either direction) and thus all testing data is translated via that model. We conjecture that if there is a detectable difference between original and translated text, this may have an influence on the translation quality, i.e. it would be best to translate original

	fo			eo			mx		
	#sent	# K words		#sent	# K words		#sent	# K words	
		fr	en		fr	en		fr	en
Train	897,430	16,760	14,753	3,645,506	81,547	72,599	4,544,936	98,346	87,387
Train-b	897,430	16,760	14,753	897,430	20,084	17,878	1,796,860	36,886	32,668
Test	2 749	55	48	20,114	472	421	22,863	528	468

Table 4: Corpus statistics for the French-original (fo), English-original (eo) and mixed (mx) corpora used below.

French text into English using a MT system that was trained on pairs of original French + translation.

We first test that hypothesis using the reference information in the corpus. In all the experiments described below, we randomly sampled the corpus (at the document level) until we amassed at least 20,000 aligned sentence pairs. This resulted in a test set of 22,863 sentence pairs, 2749 of which have a French original (Table 4). Note that in that test set, only 12% of the sentences are originally French–considerably lower than the 20% proportion in the training data. We will come back to this when analysing the results.

Because we are given the source language for each sentence pair, we constructed three training sets. The first, using all the available data, which we refer to as mx, corresponds to standard SMT practice and is oblivious to the original language. The other two are a split of the corpus according to the original language: eo for the $\sim$ 80% English-original data, and fo for the French original text. In order to investigate possible effects of this imbalance, we also subsampled the eo side in order to obtain as many sentence pairs as the fo corpus. This "balanced" corpus is shown as "Train-b" in table 4 ("Train-b").

We then trained Portage on each training set and each direction. Note that for the eo and fo splits, this means that one of the models will learn the "correct" translation direction (i.e. from original to translation) while the other will learn to "untranslate" (i.e. from translation to "original").

5.1 Impact of Original Language

We first check whether the difference between original and translation is large enough to have an impact on SMT quality. Table 5 shows the performance observed on the test set and on its fo and eo splits.

The first row shows that the system trained on the mixed data (i.e. entire training set) performs quite uniformly over the fo and eo data. The last row shows the test results of the MT model trained only on eo data. There is a clear effect due to the data mismatch: performance improves on the eo test set, by up to 1 BLEU point over the baseline mx model. In contrast, it degrades significantly on the fo test set. This suggests that the difference in language between original and translation is enough to produce a statistically significant difference in performance. Note also that this model was actually trained on *less* data than the mx model, but still manages to produce a better performance on eo and even on mx.[4]

The middle row is especially interesting. The BLEU score of the fo model on the full test set is 5-6 points lower than the baseline, which is not surprising as it was trained using five times less data. The performance is even a bit worse on the eo testing data, which again confirms the significant practical difference between eo and fo data. Note however, that on the fo test set, the fo model is quite close to the baseline. This means that a model trained on five times less data yields a similar performance, simply because it is trained on the *right kind* of data. Performance on the balanced training sets (Train-b, not indicated here for lack of space) is qualitatively slightly different, but confirms the impact of the original language on translation performance.

Note that it is well-known that SMT is very sensitive to *topic* or *genre* differences. However, in our case the corpus is fairly homogeneous overall. It is therefore likely that the difference is solely due to whether a test sentence was an original or a translation.

[4]This is due to the mismatched proportions of eo and fo data in the training and testing sets. The eo model does better on eo data, which is proportionally over-represented in the test set (88% vs. 80%).

Train	mx test set		fo test set		eo test set	
	fr▷en	en▷fr	fr▷en	en▷fr	fr▷en	en▷fr
mx	36.2	37.1	36.1	37.3	36.1	36.9
fo	31.2	30.8	36.2	36.5	30.5	30.1
eo	36.6	37.8	33.7	36.0	36.8	38.0

Table 5: BLEU score of MT systems trained and tested on fo, eo, and mx data.

	Full test set	
	fr→en	en→fr
mx	36.86	37.78
fo	32.00	31.85
eo	37.20	38.23
SVM	37.44	38.35
ref	37.46	38.35

Table 6: Performance (BLEU) of three SMT models (mx, fo, eo) and of their combination either using SVM prediction (SVM) or reference labels (ref).

5.2 Impact of Automatic Detection

We now check whether the automatic detection of translation direction provided by the sentence-level SVM classifier is good enough to produce an effect on the translation quality. Instead of translating the entire test set with the same model, we use the SVM prediction to select the appropriate SMT system. For example, when evaluating the French to English translation, we first apply the SVM classifier trained on the French monolingual data,[5] and depending on its decision, use either the fo or eo model for fr→en (and similarly for English to French translations). Note that this is more of a research experiment than a realistic use case, as in practice, translators tend to only translate originals from a well-identified language. As a *gold standard*, we use the reference source language information instead of the SVM predictions to select the SMT system. For fr→en, we used the fo model for sentences that are known to be originals, and the eo model otherwise.

Table 6 displays the final BLEU scores. It shows that the SVM prediction yields essentially the same performance as the reference data. Note that this is not unusually surprising, as all incorrect classifications are likely to be on translations that are very similar to original text (or reverse). The table also shows that both combinations of fo and eo models (last two lines) outperform the three models fo, eo and mx taken individually. Compared to the mx model, which essentially corresponds to standard practice in SMT research, the gain is about 0.6 BLEU points. This difference is unlikely to produce, on actual translations, an impact that has *practical* significance, but: 1) It is *statistically* significant, meaning that observed differences are not due only to stochastic fluctuations; and 2) It provides strong evidence that detecting whether a text is an original or a translation and using the appropriate fo or eo models actually makes sense in terms of performance.

[5]For these experiments, the bigram word SVM was re-estimated without using the documents from the test set.

6 Discussion

We have demonstrated that on the Canadian Hansard, it is possible to automatically determine the direction of the translation with high accuracy. We have also shown that SMT systems tend to perform better when the translation direction is the same in the training and test sets. At this point, our work could be refined and extended in several different directions.

From the point of view of the text classification methods, we faithfully stuck to SVMs, but there are many other classification techniques and kernel machines that could be applied, each with a gamut of possible parameter and pre-processing configurations, e.g. normalization, thresholding or term weighting.

With respect to the classification task itself, we may wonder how general our methods and resulting models will prove to be. Would models developed on the Hansard perform reasonably well on other English-French corpora? It seems clear that our SVM is able to figure out at least some of the rather strong correlations that can arise between a particular source language and some specific topics. This happens for example when a particular party is strongly associated with some particular topics while the constituency of that party is strongly associated with one of the two languages. On the other hand, the fact that good classification accuracy was obtained even when texts were reduced to part-of-speech sequences (with or without the addition of function words) clearly indicates that us-

able clues run deeper than simple lexical associations. But then, to what extent are such clues dependent on highly specific speaker and/or translator communities? Further investigation of such factors and their potential effects on classification quality, and thereafter SMT quality would be enlightening. One problem is that there do not appear to be many other sources of direction-annotated translation data available for pushing our experiments further; however the English-French subset of the Europarl corpus would certainly be suitable for a second set of experiments.

Also, would the same approach work on a different language pair? It may well be more difficult to detect translation direction for languages that are closer together such as two romance languages. Again, further experiments and suitable data would be required to evaluate that. The Europarl used by van Halteren (2008) is one obvious possibility. It may also be interesting to analyse a monolingual corpus as in Baroni and Bernardini (2006), where translations come from different languages.

Another possibility would be to use the translation detection to split the *training set* according to the original languages, for corpora where this information is not available.

At a more abstract level, a quintessential process of this work is the idea of selecting a translation model depending on the input. This can be applied to many practical categorizations. For instance, the mother tongue of the original speaker/writer is likely to manifest itself in some form (especially if it differs from the language spoken/written). Thus, creating translation models for various first languages may also be beneficial.

7 Conclusion

We considered two problems in the context of the English-French Canadian Hansard corpus. First, can we tell the difference between an original and translated document, and to what level of accuracy? Second, is the knowledge of the translation direction useful for machine translation, and if so, is the classification performance sufficient?

Using various textual representations, we found that we could detect original text vs. translation using SVMs with high accuracy: 90+% using word bigrams and 85% using POS or mixed representations. We also uncovered various patterns that are indicative of original vs. translation in English and French. The success in classification did impact Machine Translation quality. Using the SVM to select the appropriate MT system yielded a 0.6 BLEU increase in test performance w.r.t. to a single model, and was practically indistinguishable from the gold standard, implying that improvements to the classifier may not further improve translation quality.

We point out that we are at an early stage in research on detecting and exploiting translation direction in bilingual corpora and we hope that further work will explore these issues further.

References

Baroni, M. and Bernardini, S. (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.

Bottou, L., Chapelle, O., DeCoste, D., and Weston, J., editors (2007). *Large-Scale Kernel Learning*. MIT Press.

Church, K. W. and Gale, W. A. (1991). A program for aligning sentences in bilingual corpora. In *Proc. ACL-91*.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Dobrokhotov, P. B., Goutte, C., Veuthey, A.-L., and Gaussier, E. (2003). Combining NLP and probabilistic categorisation for document and term selection for swiss-prot medical annotation. *Bioinformatics*, 19(Suppl. 1):i91–i94.

Joachims, T. (2006). Training linear svms in linear time. In *Proc. KDD-06*, pages 217–226.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL-02*, pages 311–318.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. Int'l Conf. on New Methods in Language Processing*.

Ueffing, N., Simard, M., Larkin, S., and Johnson, J. H. (2007). NRC's PORTAGE system for WMT 2007. In *ACL-2007 Second Workshop on SMT*, pages 185–188.

van Halteren, H. (2008). Source language markers in europarl translations. In *Proc. Coling-2008*, pages 937–944.