Automatic Dictionary Construction and Identification of Parallel Text Pairs

Sumithra Velupillai[†] **Martin Hassel**[†] **Hercules Dalianis**[†‡]

sumithra@dsv.su.se ph: +46 8 16 11 74
xmartin@dsv.su.se ph: +46 8 674 74 14
hercules@dsv.su.se ph: +46 70 568 13 59

[†]DSV/KTH-Stockholm University
SE-164 40 Kista
Sweden

[‡]Euroling AB
Igeldammsgatan 22c
112 49 Stockholm, Sweden

Abstract

When creating dictionaries for use in for example cross-language search engines, parallel or comparable text pairs are needed. Multilingual web sites may contain parallel texts but these can be difficult to detect. For instance, a multilingual website, Hallå Norden, contains information in five languages; Swedish, Danish, Norwegian, Icelandic and Finnish. Working with these texts we discovered two main problems: the parallel corpus was very sparse, containing on average less than 80.000 words per language pair (in the final version of the corpora), and it was difficult to automatically detect parallel text pairs. We discovered that, on average, around 55 percent of the texts were not parallel. Creating dictionaries with the word aligner Uplug gave on average 213 dictionary entries. Despite the corpus sparseness the results were surprisingly good compared to other experiments with larger corpora.

Following this work, we made two sets of experiments on automatic identification of parallel text pairs. The first experiment utilized the frequency distribution of word initial letters in order to map a text in one language to a corresponding text in another in the JRC-Acquis corpus (European Council legal texts). Using English and Swedish as language pair, and running a ten-fold random pairing, the algorithm made 87 percent correct matches (baseline-random 50 percent). Attempting to map the correct text among nine randomly chosen false matches and one true yielded a success rate of 68 percent (baseline-random 10

percent). In another experiment features such as word, sentence and paragraph frequencies were extracted from a subset of the JRC-Acquis corpus and used with memory-based learning on Swedish-Danish, Swedish-Finnish and Finnish-Danish, respectively, achieving a pair-wise success rate of 93 percent. We believe methods such as these will improve, for instance, automatic bilingual dictionary construction from unstructured corpora and our experiments will be further developed and evaluated along these lines.

1 Introduction

Dictionaries are an important part of natural language processing tasks and linguistic work. Domain-specific dictionaries can for example be used in cross-language web and intranet search engines. Creating dictionaries manually is labor intensive and time consuming, and many methods to make this process automatic have been proposed. Word alignment tools are often used for the creation of bilingual word lists. Many assumptions about the characteristics of words and their translations for extracting bilingual vocabulary underlie the algorithms in such tools, and parallel or comparable corpora are needed as input. However, finding such corpora is often a difficult and arduous task, especially for small languages. The Internet is a useful resource for finding corpora in different languages, and many large corporations and organizations have abundant information in multilingual web sites. However, these text sets are often noisy, containing a lot of non-parallel parts which need to be removed in order to create useful parallel corpora.

In this paper, three experiments are described. The first, described in Section 3, is an experiment on creating parallel corpora and bilingual dictionaries from the web site Hallå Norden (Hello Scandinavia)[1]. After extracting text pairs covering all the Nordic language pairs by treating the entire set of texts on the web site as one multilingual parallel corpus, ten parallel corpora were created. These were further used as input to the word alignment tool Uplug (Tiedemann 2003) for the automatic creation of dictionaries covering the Nordic

languages.

However, in these corpora, we discovered that all text pairs were not completely parallel. Therefore, we have developed and evaluated methods for identifying parallel and non-parallel texts in corpora covering different language pairs. In Section 3, an initial experiment on deleting non-parallel texts from the ten Nordic corpora is described. This method did not prove very successful, and two more thorough experiments on alternate methods for automatically identifying non-parallel texts in bilingual corpora have been performed.

The first experiment, described in Section 4, exploits the frequency distribution of word initial letters in order to map a text in one language to a corresponding text in another. In this experiment, the JRC-Acquis corpus (European Council legal texts)[2] was used, with English and Swedish as language pair. In the second experiment, described in Section 5, a memory-based machine learning technique was used with simple frequency features such as word, sentence and paragraph frequencies. The method was evaluated on a subset of the JRC-Acquis corpus as well as the entire set of Hallå Norden texts (described above), and used on Swedish-Danish, Swedish-Finnish and Finnish-Danish, respectively.

The experiments described in this paper show very promising results. However, further development and evaluation is needed. Language-independent methods for creating language resources, especially for small languages, are still scarce but important. Some concluding remarks and thoughts on future work are described in the final section, with the intent of raising some directions for further studies in this intriguing and important research area.

2 Related Work

Bilingual parallel corpora are useful for many natural language processing tasks, such as machine translation systems. For the automatic creation of dictionaries, word alignment

systems are often used. Such systems need to make some assumptions regarding translated texts (Somers 2001):

- Words have one sense per corpus
- Words have a single translation per corpus
- There are no missing translations in the target document
- The frequencies of words and their translations are comparable
- The positions of words and their translations are comparable

These assumptions affect word alignment algorithms and, as can be seen, for the systems to work optimally, parallel or comparable corpora are needed.

The distinction between a parallel and a comparable corpus has been discussed in several research articles. In Somers (2001), it is pointed out that a "comparable" corpus has been used both interchangeably with "parallel" corpus, and as a term describing a corpus with similar but not necessarily equivalent texts. A more detailed discussion on the distinctions between how the terms parallel, comparable and non-parallel corpora are used can be found in Fung & Cheung (2004) for instance.

Freely available multilingual resources are often noisy and non-parallel sections need to be removed. Many methods for identifying such sections automatically have been proposed. Maximum entropy (ME) classification is used in Munteanu & Marcu (2005) in order to improve machine translation performance. From large Chinese, Arabic and English non-parallel newspaper corpora, parallel data was extracted. For this method, a bilingual dictionary and a small amount of parallel data for the ME classifier is needed. By selecting pairs of similar documents from two monolingual corpora, all possible sentence pairs are passed through a word-overlap based filter and then sent to the ME classifier. The results were evaluated in different ways, one evaluation was made by testing the system on the news test corpus used for the NIST 2003 MT evaluation[3], using the BLEU score, reporting significant improvements over the baseline (the highest score for Arabic-English was 47.97

and for Chinese-English 30.03).

In Fung & Cheung (2004) a method for extracting parallel sentences through bootstrapping and Expectation Maximization (EM) learning methods is presented. An iterative bootstrapping framework is presented, based on the idea that documents, even those with a low similarity score, containing one pair of parallel sentences must contain others. In particular, the proposed method works well for corpora with very disparate contents. The approach achieves 65.7 percent accuracy and a 50 percent relative improvement over their baseline.

Latent Semantic Indexing (LSI) has been experimented with in Katsnelson & Nicholas (2001) in order to identify parallel sequences in corpora. In this work, the hypothesis that LSI reveals similarities between parallel texts not apparent in non-parallel texts is presented and evaluated. Corpora from digital libraries were used with the language combinations English-French, English-Russian, French-Russian and English-Russian-Italian. Applying correlation coefficient analysis, a threshold of 0.75 was reported to successfully hold as a lower bound for identifying parallel text pairs. Non-parallel text pairs did not, in these experiments, exceed a correlation coefficient value of 0.70.

Unfortunately, most work has been performed on different types of corpora and on different language pairs. Moreover, they have been evaluated differently depending on available resources and the nature of the experiments, which makes them difficult to compare. However, the different approaches show the need for these types of methods.

3 Automatic Construction of Domain-specific Dictionaries on Sparse Corpora in the Nordic Languages

In an experiment described in Velupillai & Dalianis (2008), dictionaries covering the Nordic languages using corpora obtained from the web site Hallå Norden (Hello Scandinavia) were automatically created. Hallå Norden contains information regarding mobility between the

Nordic countries in five languages; Swedish, Danish, Norwegian, Icelandic and Finnish. Treating the entire set of texts on the web site as one multilingual parallel corpus, ten parallel corpora for each Nordic language pair were extracted and used for the creation of ten different dictionaries. The creation of the corpora was semi-automatic. The texts on the web site were structured in a site map which was exploited to automatically find parallel text pair candidates. However, after manual inspection of these candidates, we discovered that only around 45 percent of the initial corpora from the web site contained parallel text pairs. The remaining texts were either single texts with no matching translated text, texts in the wrong language, or just empty pages. We removed almost all such texts manually.

Creating parallel corpora from multilingual web sites often involves analyzing the contents and structures, as well as removing a lot of noise. For instance, on a Scandinavian bank corporation web site with information in Swedish, Danish and Finnish, more than 50 percent of the texts were non-parallel. However, although a lot of texts may be removed, the final size of the created parallel corpora will naturally depend on the types of texts. The Hallå Norden texts, for example, are in general very short, while other types of texts available on other web sites, annual reports for instance, may be much longer.

The final version of the created Hallå Norden corpora contained on average less than 80.000 words per language pair, which is considered as a sparse corpora. For the creation of the dictionaries we used the word alignment system Uplug, since it is a non-commercial system which does not need a pre-trained model and is easy to use. It is also updated continuously and incorporates other alignment models, such as GIZA++ (Och & Ney 2003).

The produced dictionaries gave on average 213 dictionary entries (frequency > 3). Combinations with Finnish, which belongs to a different language family, had a higher error rate, 33 percent, whereas the combinations of the Scandinavian languages only yielded on average 9 percent errors. Despite the corpus sparseness the results were surprisingly good

compared to other experiments with larger corpora.

However, we discovered that the created corpora were to some extent non-parallel containing some extra non-aligned paragraphs. We believed that these text pairs affected the results negatively, and made a small experiment on trying to automatically delete texts pairs that were not parallel.

We used a simpler algorithm than in for instance Munteanu & Marcu (2006). The total number of paragraphs and sentences in each parallel text pair were counted. If the total number for each language in some language pair differed more than 20 percent either in the total number of paragraphs, sentences, or both, these texts were automatically deleted. On

	Initial		Deleting non-parallel	
Language pair	No. dictionary words	Erroneous translations, %	No. dictionary words	Erroneous translations, %
sw-da	322	7.1	305	7.2
sw-no	269	6.3	235	9.4
sw-fi	138	29.0	133	34.6
sw-ice	151	18.5	173	16.2
da-no	322	3.7	304	4.3
da-fi	169	34.3	244	33.2
da-ice	206	6.8	226	10.2
no-fi	185	27.6	174	30.0
no-ice	159	14.5	181	14.4
Average	213	16.4	219	16.1

Table 1: Produced dictionary words and error rate for the initial and the refined corpora, from Velupillai & Dalianis (2008).

average 5 percent of the manually processed corpora were detected as being non-parallel using this algorithm. The refined corpora were re-aligned with Uplug and evaluated, but unfortunately about the same error rate as before deleting the non-parallel texts was obtained, although with some differences in the produced word pairs (see Table 1). Perhaps our simple algorithm was too coarse for these corpora, especially since they were so sparse. The texts were in general very short and simple frequency information on paragraph and sentence amounts might not have captured non-parallel fragments on such texts. A more detailed discussion on the results of this experiment can be found in Velupillai & Dalianis (2008).

More elaborate and efficient methods for identifying parallel and non-parallel texts in bilingual corpora are described in the following sections.

4 Identifying Parallel and Non-parallel Texts in Bilingual Corpora using Fingerprints

When comparing documents for content similarity it is common practice to produce some form of document signatures, or "fingerprints". These fingerprints represent the content in some way, often as a vector of features, which are used as the basis for such comparison. One common method when comparing the likeness of two documents is to utilize the so-called Vector Space model (Salton 1971, 1983). In this model the documents' fingerprints are represented as feature vectors consisting of the words that occur within the documents, with weights attached to each word denoting its importance for the document. We can, for example, for each feature (in this example, a word) record the number of times it occurs within each document. This gives us what is commonly called a document-by-term matrix where the rows represent the documents in the document collection and the columns each represent a specific term existing in any of the documents (a weight can thus be zero). We can now, somewhat simplified, compare the documents' fingerprints by looking at how many times each feature occurs in each document, taking the cosine angle between the vectors, and pair the two most similar together. One obvious drawback of the basic use of this model is that when comparing texts written in different languages we do not necessarily know which feature in one language corresponds to which feature in another.

Another drawback when building a word vector space representing more than one language is that the vocabulary, i.e. the number of features in the feature vectors, grows alarmingly (this is in many cases already a problem representing just one language (Sahlgren 2005)). Ways of limiting the vocabulary include using stop-word lists to remove "information poor" features, frequency thresholding and conflation into feature classes (for example lemmatization). In word vector spaces the latter is often accomplished by bringing

semantically related words to a common lemma or stem. In the experiments described below conflation was attempted by moving from term frequency classes towards prefix frequency classes, i.e. the leading characters of each token. This way a document's fingerprint effectively is represented by a feature vector containing the frequency of each prefix of a set length *n* occurring in the corpus. This has for example been used in information retrieval for filtering of similar documents written in the same language (Stein 2005). We here attempt to utilize this notion in cross-language text alignment.

4.1 Data sets and experimental setup

In this set of experiments we have used the JRC-Acquis corpus (Steinberger et al. 2006). This corpus consists of European Union law texts, which are domain specific and also very specific in their structure. Many texts are listings of regulations with numerical references to other law texts[4] and named entities (such as countries). The corpus is very large, containing a different amount of texts depending on the language. Here we have investigated the language pair Swedish-English, i.e. we used Swedish as a source language attempting to find the corresponding parallel text in English. We have also used only those documents that have a counterpart in both languages, resulting in a total of 20.145 documents. In Appendix A, a Swedish example file along with its corresponding, parallel, English translation from the JRC-Acquis corpus is given. In order to delimit the search space for the practicality of this experiment we have not compared each Swedish source text with each and every English text. Instead we, in one experiment, compared the similarity between a true positive (the corresponding, parallel, English text) and one true negative (a randomly chosen non-parallel English text), letting the algorithm choose the closest match (as defined by the cosine angle between the feature vectors for each text). In another experiment we repeated the setup, but instead of only using one true negative we used nine. This gave us a random chance of picking the true positive of 50 percent in the case of one true positive and one true negative,

and 10 percent in the case of one true positive and nine true negatives (see Table 2 below). In order to rule out any random fluke in the choice of true negative(s) for each true positive both experiments were carried out 10 times, making new random pairings each time. An average was then taken, calculated over these ten runs.

As in Stein (2005) we have extracted a-priori probabilities of prefix classes from reference corpora. Since we are dealing with the language pair Swedish-English we have used a Swedish reference corpus, the Swedish Parole corpus[5], and an English ditto, the British National Corpus (Aston & Burnard 1998). The Swedish reference corpus is comprised of roughly 20 million words. In order to have a comparable English reference corpus we have only used the first 20 million words of BNC (out of roughly 100 million). These two corpora can be seen as the expected distribution of the prefix classes for each language, while each text's feature vector then is the deviation to the expected distribution. What we thus attempt to model is the hypothesis that a deviation from the expected frequency distribution pattern in one language in the pair could possibly reflect a similar deviation in the other.

In this set of experiments the feature vector for each text was preprocessed in two ways:

1. Using Parole as reference corpus for the Swedish texts and BNC as reference corpus for the English, by calculating the difference in frequency between the occurrences of a prefix in the reference corpus and in each text. The prefixes in these vectors were then sorted by the frequency in each respective reference corpus. The most common feature in the source language corresponds to the most frequent feature in the target language, and so on. The comparison of the text's feature vectors is then based on the deviation from the expected and normalized distribution for each language.
2. No normalization using reference corpora. Instead the raw frequencies are compared directly. However, matching of features is still based on the frequency in each language's respective reference corpus.

As mentioned above, feature vectors were created using the leading *n* characters of each word occurring in each reference corpus, as well as in any of the 20.145 documents used in the tests. A fingerprint was constructed for each reference corpus and each document, in both languages, for *n*=1..3, both using all lower case, (lc), prefixes as well as prefixes maintaining their original capitalization. To be noted here is the fact that the vocabulary size grows at an explosive rate as *n* grows, especially when the original capitalization is preserved.

4.2 Results

model:	1. Parole / BNC		2. no normalization	
Prefix size	mean precision	Lowest - highest	mean precision	lowest – highest
k=2, *n*=1	50 %	0.496 - 0.503	**87 %**	0.865 - 0.872
k=2, *n*=1, lc	50 %	0.497 - 0.502	86 %	0.852 - 0.858
k=2, *n*=2	50 %	0.497 - 0.502	80 %	0.794 - 0.799
k=2, *n*=2, lc	50 %	0.498 - 0.502	76 %	0.756 - 0.762
k=2, *n*=3	50 %	0.496 - 0.502	76 %	0.759 - 0.769
k=2, *n*=3, lc	50 %	0.495 - 0.505	75 %	0.747 - 0.753
k=10, *n=1*	10 %	0.097 - 0.102	**68 %**	0.674 - 0.678
k=10, *n=1, lc*	10 %	0.098 - 0.102	65 %	0.646 - 0.655
k=10, *n=2*	10 %	0.099 - 0.104	54 %	0.534 - 0.543
k=10, *n=2, lc*	10 %	0.098 - 0.103	45 %	0.450 - 0.455
k=10, *n=3*	10 %	0.100 - 0.102	50 %	0.497 - 0.504
k=10, *n=3, lc*	10 %	0.097 - 0.102	44 %	0.438 - 0.442

Table 2: Swedish source, one true positive and one true negative English target (*k*=2); one true positive and nine true negatives (*k*=10). Lower case is abbreviated lc. The precision is calculated over 10 random selections of the non-parallel text(s). Also given is the lowest and the highest result of the ten runs. At *k*=2 baseline-random is 50 percent and our results indicate up to 87 percent precision, at *k*=10 baseline-random is 10 percent and our results indicate up to 68 percent precision.

As can be seen in Table 2 it is far more favorable to compare the raw frequencies of the features in the source and target vectors, rather than comparing the deviation based on the frequency distribution in the reference corpus of the respective languages. This is further supported by the fact that model two stands even stronger, relatively speaking, when pin-pointing the right match out of ten possible target texts.

We can also see that the results are very stable – there is only a slight difference in the

precision between the best and the least good run – even though there is little overlap between the 10 randomly generated lists of pairs. The highest number of pairs that one of the lists has in common with any of the other lists is 12 (out of 20.145). When it comes to the lists containing 10 target words this number is nearly non-existent.

One possible answer for the success of the second model could of course be that the source and target texts always are lexically very alike. This could be the case if they to a high degree share the same vocabulary, for instance named entities. This does not seem to be the case if we take a look at Table 3.

	k=1		*k*=10	
Baseline	mean precision	lowest – highest	Mean precision	lowest - highest
1	50 %	0.496 - 0.503	10 %	0.097 - 0.102
2	50 %	0.497 - 0.503	10 %	0.099 - 0.102
3	50 %	0.497 - 0.504	10 %	0.098 - 0.102

Table 3: Baselines using only basic features, each tracking the number of occurrences of; baseline1={bytes, tokens, dot, comma, percent, digit, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, baseline2={bytes, tokens, dot, comma, percent} and baseline3={tokens, dot, comma}.

The degree of precision and the stability of the results are encouraging. However, for the sake of a fairer comparison one might want to reconsider the baselines used in this experiment as being too naïve. In the next section, a different set of roughly language independent features, as well as some language dependent (relying on the use of a part-of-speech tagger), is presented, tested on some of the Nordic language pairs.

5 Identifying Parallel and Non-parallel Texts in Bilingual Corpora using Simple Frequency Features and Memory-based Learning

In the final experiment on trying to identify whether two texts in different languages in a bilingual corpus are parallel or not, a memory-based machine learning technique was used. The identification problem can be viewed as a classification problem where the possible

classes are *Parallel* and *Non-parallel*. We put forward the hypothesis that simple frequency counts on for instance paragraphs, sentences and words, as well as part-of-speech information, could be valuable features for detecting whether a text pair in two different languages is parallel or not.

The following language pairs were used: Swedish-Danish, Swedish-Finnish and Danish-Finnish (treating the leftmost language in each language pair as the source language, and the rightmost language as the target language). Using language pairs from both related and non-related language families is important in order to investigate if such issues influence the results. Two bilingual corpora for each language pair were created, consisting of an equal amount of *Parallel* and *Non-parallel* instances (only one true positive and negative instance, thus giving a 50 percent random chance of picking the true positive), amounting to in total six corpora. The corpora were extracted from the JRC-Acquis corpus (described in Section 4) and the Hallå Norden corpus (described in Section 3).

As stated in Section 4, many texts in the JRC-Acquis corpus contain listings of regulations and numerical references to other law texts, thus containing very short sentences. The Swedish, Danish and Finnish text sets contain around 20.000 texts, where most of the texts also exist in a parallel version in the other two languages.

The Hallå Norden corpus consists of short information texts regarding mobility information in the Nordic region (see Section 3). The corpus is very small (around 200 texts per language pair), but provides a different type of text from a different domain that reflects another type of language use than the texts in the JRC-Acquis corpus. Although the texts are short and may also contain a lot of listed information, they are not as fragmented as the texts in the JRC-Acquis corpus. In Appendix B, a Swedish and a Danish example file from the Hallå Norden corpus are given. These examples illustrate the type of texts this corpus contains, and how they contain sequences that are parallel translations but also sequences that

may be missing. Moreover, they exemplify how differently the texts can be formatted, especially with regards to paragraphs. This text pair was recognized as non-parallel using the simple algorithm for detecting non-parallel files described in Section 3.

5.1 Machine Learning Algorithm

For this experiment the machine learning algorithm used was memory-based learning, using the TiMBL software (see Daelemans et al. (2007) for a reference guide). It was used with the classification algorithm IB1, applying default settings with regards to algorithmical settings. This means that the distance metric used was *Overlap* and the feature weighting used was *Gain Ratio*. A feature selection experiment was performed on these default values, testing different combinations of features. The tests were performed through *10*-fold cross-validation, splitting the entire data sets into 10 parts, equal in size, containing the same amount of *Parallel* and *Non-parallel* classified text pairs, using nine parts for training and one part for testing in turn for each part.

5.1.1 Features

For each text in the bilingual corpora, the following features were extracted:

- Total number of words
- Total number of sentences
- Total number of paragraphs
- Average length of words
- Average (word) length of sentences
- Average (word) length of paragraphs
- The five most frequent part-of-speech bi- and tri-grams

Moreover, the difference (in percent) in the total number of words, sentences and paragraphs between a text pair as well as the difference in the average number of words, sentences and paragraphs between a text pair was calculated and used as features. Here, difference is calculated the following way: $(\max(s\text{-}t))/(s+t)\times 100$, where s is the value of the total number

or average length of words, sentences or paragraphs for the source language text and *t* is the value of the total number or average length of words, sentences or paragraphs in the target language text. In total, each instance in the data set consisted of 39 features (including an instance id, which was never included in the feature selection)

5.1.2 Definitions

A simple approach was used in order to identify words, sentences and paragraphs. Words are defined as a sequence of characters separated by space. No punctuation characters are included as words (a word such as ``EG/EEG'' is replaced with ``EGEEG''), and digits are not counted as words. When calculating the average length of a word the number of characters in each word is used.

Sequences of characters ending with "." and/or newline are defined as sentences. When calculating the average length of a sentence the number of words in each sentence is used. Sequences of characters ending with newline are defined as paragraphs. When calculating the average length of a paragraph the number of words in each paragraph is used. More sophisticated identification of words, sentences and paragraphs could of course be used.

5.1.3 Part-of-speech Tagging

Before extracting words, sentences and paragraphs all texts were part-of-speech tagged. For Swedish Granska[6] was used, for Danish CST's Part-of-Speech Tagger[7], and for Finnish Fintwol[8]. The taggers use different sets of tags, and have, naturally, been evaluated on different corpora. However, they are state-of-the-art tools for the respective languages. Fintwol, for instance, is the only available tool for tagging Finnish and has been used for creating gold data in the Morpho Challenge 2007[9]. For this experiment, the different tag sets were not mapped to a uniform tagset. The idea was that the distribution patterns of part-of-speech bigrams and trigrams for each language would reflect the relationship between the texts.

5.2 Data Set

For each corpus, all features for each text in one language chosen as the source language was paired with the corresponding (true positive) text in the target language, creating an instance with the classification *Parallel*. The source language text was also paired with a randomly picked target text (true negative), creating an instance with the classification *Non-Parallel*.

The Hallå Norden-corpus consisted of the following corpora:

Swedish-Danish, 191 text pairs

Swedish-Finnish, 196 text pairs

Danish-Finnish, 239 text pairs

The JRC-Acquis corpus consisted of the following corpora:

Swedish-Danish, 14 231 text pairs

Swedish-Finnish, 14 226 text pairs

Finnish-Danish, 23 238 text pairs

The Swedish-Danish and Swedish-Finnish data sets from the JRC-Acquis corpus were smaller than the Finnish-Danish due to part-of-speech tagging problems on the Swedish texts. Each data set was divided into *10* subsets for the *10*-fold cross-validation process, containing an equal amount of *Parallel* and *Non-parallel* instances.

5.3 Results

Test	Description
1	Default, all features except first feature (instance id), used as baseline
2	Total number and average length of words, sentences and paragraphs
3	All part-of-speech features
4	Part-of-speech bigrams
5	Part-of-speech trigrams
6	Difference in total number and average length of words, sentences and paragraphs
7	Difference in total number of words, sentences and paragraphs
8	Difference in average length of words, sentences and paragraphs
9	Difference in total and average number of words
10	Difference in total number and average length of sentences
11	Difference in total number and average length of paragraphs

Table 4: Feature test descriptions. The extracted features were grouped in different sub-groups.

In Table 4 the performed feature tests are described. In total, eleven feature tests were performed on each data set. The extracted features were divided into the following sub-groups: total numbers and average lengths of words, sentences and paragraphs, part-of-speech tag information, and differences between each text with respect to total numbers and average lengths of words, sentences and paragraphs. These groups of features were tested independently. Also, the sub-groups were further divided into smaller subsets of features, in order to test which feature(s) produced the best results. Test 1, which includes all features except the instance id, was used as the baseline. The groups of features and the baseline was chosen based on intuition, and should of course be scrutinized and tested further in future developments.

The results for the Hallå Norden data sets were surprisingly good (see Table 5)**,** despite the small size of the corpora. It is interesting to note that the part-of-speech information yielded very poor results. Perhaps this could be improved by mapping the different tag sets into a uniform tag set. Moreover, choosing the five most frequent part-of-speech bi- and trigrams may not distinguish parallel and non-parallel text pairs very well, as they may be common in all texts. Extracting discriminative part-of-speech patterns would be desirable. However, the features containing information about the differences between the number of, or average length of, words, sentences and paragraphs in the text pairs yielded

Test	Swedish-Danish	Swedish-Finnish	Danish-Finnish
1	74.7	52.0	69.8
2	7.9	9.6	13.8
3	9.5	14.5	16.9
4	20.1	33.4	30.1
5	8.7	13.2	16.7
6	79.9	65.9	73.7
7	82.4	68.1	73.7
8	76.9	60.1	67.8
9	85.3	63.0	68.5
10	72.3	68.3	77.7
11	59.0	55.2	76.3

Table 5: Results, Hallå Norden, average accuracy (in percent) of the *10*-fold cross-validation tests, one can see that all tests from 6 to 11 yield good results.

promising results. In particular, the feature test where all information about differences between the texts (Test 7) produced good results for all language pairs.

Test	Swedish-Danish	Swedish-Finnish	Finnish-Danish
1	92.2	90.1	88.1
2	25.0	24.9	22.7
3	37.0	46.8	50.5
4	59.4	65.1	66.5
5	52.6	54.7	54.2
6	92.7	90.3	88.6
7	93.2	90.7	89.2
8	93.1	90.5	88.5
9	93.3	89.7	88.5
10	89.3	89.7	85.9
11	93.1	89.2	89.0

Table 6: Results, JRC-Acquis, average accuracy (in percent) of the *10*-fold cross-validation tests, one can see that all tests from 6 to 11 yield good results.

The results for the JRC-Acquis data sets are given in Table 6. The results are very encouraging. As in the tests on the Hallå Norden corpora, using the features that reflect the differences in the total number and average length of words, sentences and paragraphs produced good results for all language pairs. Using the information about the total number and average length of words for each text separately did not yield good results for any data set. Perhaps normalizing them in some way would be advantageous.

Overall the result patterns are similar for the two different corpora, even though the results for the JRC-Acquis corpora are better than the results for the Hallå Norden corpora. It is interesting to note that the patterns are so similar despite the different characteristics of the text sets (in size, domain type and text type for instance).

The results are very promising. Even for a small data set such as the Hallå Norden corpora, it is possible to detect parallel and non-parallel text pairs on simple frequency features. However, more tests would need to be performed in order to verify the results properly. In particular, both text sets are very homogeneous, which might affect the results. The texts are similar in both their content and structure. The method should also be evaluated

on more diversified text sets.

Even though the Swedish-Danish and Swedish-Finnish JRC-Acquis corpora were smaller than the Finnish-Danish, the results were similar. It would be interesting to investigate at which point in the size of the data set results seem to decrease. Perhaps fairly small corpora are sufficient in order to obtain good results.

Experiments with other language pairs should also be performed. For instance, part-of-speech information might prove more valuable to other language pairs. Moreover, as stated above, other approaches to using the part-of-speech information should be investigated. Also, the length measures for paragraphs and sentences used here are not normalized in any way. An interesting experiment would be to use language normalized number of characters instead of measuring the raw word lengths. Furthermore, other settings in the chosen machine learning algorithm should be tested. Parameter optimization tests using other distance metrics or weighting schemes might yield improved results. Given the features used, perhaps a different machine learning algorithm such as SVM (support vector machine), might produce better results.

6 Conclusions and Future Work

In the experiments described above we have shown that our methods for identifying and deleting non-parallel texts from different corpora covering different language pairs show great potential. However, the results are, unfortunately, currently not comparable. In future experiments, we will apply the methods on the same corpora and language pairs, and evaluate the results in a comparable manner.

Methods for identifying parallel texts or sequences in texts can be used for many natural language processing tasks, including machine translation systems and dictionary construction. Evaluating and comparing such methods is difficult, as they are developed on different types of corpora and languages. Moreover, there are many evaluation metrics that

can be used, depending on both the availability of gold standard corpora and the purpose of the studies.

We have developed methods with the intention of keeping them as language-independent as possible. For the fingerprint method (described in Section 4), the only language-dependent feature is the use of a reference corpus for each language. Such corpora may, unfortunately, still be difficult to obtain for very small languages with scarce resources. The use of language-dependent part-of-speech information for the simple frequency method (described in Section 5) did not improve results. However, this information should probably be used differently. It is interesting to note that the best results in this experiment were obtained through the purely language-independent frequency features.

Moreover, in further work all our experiments on the identification of parallel text pairs should be run on more language pairs, preferably such that contain languages belonging to different language groups (as has, for instance, been carried out with the combinations with Finnish in the memory based learning experiments). An obvious observation here is that the language pairs should also be tested reversely; that is, if one is to investigate the performance on for instance the language pair Swedish-English, it should also be evaluated on the corresponding pair English-Swedish. Also, the experiments should be re-run on other corpora than the JRC-Acquis and Hallå Norden in order to discern that we are not just investigating peculiarities of these specific corpora.

In a real-world setting, attempting to identify whether a text in one language is parallel with a text in another means that it needs to be compared with many texts in the target language. For instance, the method described in Section 5 should be tested against several true negatives, as the fingerprint-method, described in Section 4, was. We also intend to investigate and develop methods for reducing the search space for candidate translations.

An important aspect of developing methods for cross-language tools or resources is

the possible need for preprocessing tools, such as part-of-speech taggers, covering all languages. This may be difficult to obtain, and different tools use different formatting and tagging schemes. Moreover, they might differ in robustness, which also affects the end results. Evaluating the performance of such preprocessing steps might be desirable.

Creating parallel corpora from Internet resources is both practical and convenient, as many texts are freely available. It is, however, not always trivial to extract the necessary sequences of web texts. Methods for utilizing the structure(s) of different site maps and removing tags and other web-specific formatting details are needed in order to minimize manual work. Moreover, many alternative sources for finding parallel corpora exist, such as digital libraries.

Parallel corpora covering different language pairs and text types are still very scarce, especially for small languages. Such corpora are important for many aspects of translation studies and need to be compiled. Moreover, the access of freely available parallel corpora provides the possibility of creating gold standard corpora that could be used for evaluating and comparing different methods. However, the difficulty of evaluating methods that are needed and used for different purposes still remains.

---

[1] See: http://www.hallonorden.org

[2] See: http://wt.jrc.it/lt/Acquis/

[3] http://www.nist.gov/speech/tests/mt

[4] Referencing systems do however differ between languages. For example, while some use Hindu-Arabic numerals others use Roman.

[5] http://sprakbanken.gu.se/parole

[6] www.nada.kth.se/theory/projects/granska/

[7] http://www.cst.dk/online/pos_tagger/uk/index.html

[8] http://www2.lingsoft.fi/doc/fintwol/

[9] http://www.cis.hut.fi/morphochallenge2007/

Acknowledgements

We would like to thank Pernilla Näsfors (for her work with producing the Hallå Norden corpora) and Björn Andrist (for his inspiration regarding using the fingerprint algoritm), both at Euroling AB and SiteSeeker. The experiments described in Section 5 were carried out within the course Machine Learning, organized by GSLT (Swedish National Graduate School of Language Technology). We thank our supervisor Joakim Nivre for advice and support.

References

Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A. (2007). *TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide*. Technical Report, ILK Research Group Technical Report Series no. 07-07.

Fung, P. and Cheung, B. (2004) Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. Barcelona, Spain 25 – 26 July 2004.

Katsnelson, Y. and Nicholas, C. (2001). Identifying Parallel Corpora Using Latent Semantic Indexing. In *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK 30 March – 2 April 2001.

Munteanu, D.S. and Marcu, D. (2006). Extracting Parallel Sub-sentential Fragments from Non-parallel Corpora. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics*. Sydney, Australia 17 – 21 July 2006, pp. 81-88.

Munteanu, D. S and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4), pp. 477-504.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19-51.

Sahlgren, M. (2005). An Introduction to Random Indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Copenhagen, Denmark August 16 2005.

Salton, G. ed. (1971). *The Smart Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Somers, H. (2001). Bilingual Parallel Corpora and Language Engineering. In *Anglo-Indian Workshop "Language Engineering for South-Asian Languages" (LESAL)*. Mumbai, India April 2001.

Stein, B. (2005). Fuzzy-Fingerprints for Text-Based Information Retrieval. In Tochtermann, K and Maurer, H., eds. *Proceedings of the I-KNOW '05, Graz 5th International Conference on Knowledge Management Journal of Universal Computer Science*. Graz, Austria: Know-Center, pp. 572-579.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. Genoa, Italy 24 – 26 May 2006.

Tiedemann, J. (2003). *Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD Thesis, Acta Universitatis Upsaliensis: Studia linguistica upsaliensia.

Velupillai, S. and Dalianis, H. (2008). Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. In *The proceedings of the 2nd MMIES Workshop: Multi-source, Multilingual Information Extraction and Summarization*. Manchester, UK, 23 August 2008.

Appendix A Example Files from the JRC-Acquis Corpus (Swedish and English)

(Apart from some minor differences we see that the files are very parallel translations. Also, we see the specificity of the text type: short sentences, named entities and many listings.)

Swedish:

```
2006/796/EG: Rådets beslut av den 13 november 2006 om evenemanget Europeisk
kulturhuvudstad år 2010

Rådets beslut
av den 13 november 2006
om evenemanget Europeisk kulturhuvudstad år 2010
(2006/796/EG)
EUROPEISKA UNIONENS RÅD HAR BESLUTAT FÖLJANDE
med beaktande av fördraget om upprättandet av Europeiska gemenskapen,
med beaktande av Europarlamentets och rådets beslut nr 1419/1999/EG av den
25 maj 1999 om att inrätta en gemenskapsåtgärd för evenemanget Europeisk
kulturhuvudstad för åren 2005 till 2019 [1], särskilt artikel 2.3 och 2.4,
med beaktande av den rapport från juryn från april 2006 som lagts fram för
kommissionen, Europaparlamentet och rådet i enlighet med artikel 2.2 i
beslut nr 1419/1999/EG,
med beaktande av att kriterierna i artikel 3 och bilaga II i beslut nr
1419/1999/EG,
med beaktande av kommissionens rekommendation av den 23 oktober 2006.
HÄRIGENOM FÖRESKRIVS FÖLJANDE.
Artikel 1
Essen och Pécs skall utses till europeiska kulturhuvudstäder 2010 i
enlighet med artikel 2.1 i beslut nr 1419/1999/EG.
Artikel 2
Istanbul skall utses till europeisk kulturhuvudstad 2010 i enlighet med
artikel 4 i beslut nr 1419/1999/EG.
Artikel 3
De tre städerna skall vidta alla åtgärder som krävs för att säkerställa att
artiklarna 1 och 5 i beslut nr 1419/1999/EG genomförs på ett effektivt
sätt.



Utfärdat i Bryssel den 13 november 2006.
På rådets vägnar
S. Huovinen
Ordförande
[1] EGT L 166, 1.7.1999, s. 1. Beslutet ändrat genom beslut nr 649/2005/EG
(EUT L 117, 4.5.2005, s. 20).
-------------------------------------------------
```

English:

2006/796/EC: Council Decision of 13 November 2006 on the European Capital of Culture event for the year 2010

Council Decision
of 13 November 2006
on the European Capital of Culture event for the year 2010
(2006/796/EC)
THE COUNCIL OF THE EUROPEAN UNION,
Having regard to the Treaty establishing the European Community,
Having regard to Decision No 1419/1999/EC of 25 May 1999 of the European Parliament and the Council establishing a Community action for the European Capital of Culture event for the years 2005 to 2019 [1], and in particular Articles 2 paragraph 3 and 4, thereof,
Having regard to the Selection Panel report of April 2006 submitted to the Commission, the European Parliament and the Council in accordance with Article 2 paragraph 2 of Decision 1419/1999/EC,
Considering that the criteria laid down in Article 3 and Annex II of Decision No 1419/1999/EC are entirely fulfilled,
Having regard to the recommendation from the Commission of 23 October 2006,
HAS DECIDED AS FOLLOWS:
Article 1
Essen and Pécs are designated as %quot%European Capital of Culture 2010%quot% in accordance with Article 2 paragraph 1 of Decision No 1419/1999/EC as amended by Decision No 649/2005/EC.
Article 2
Istanbul is designated as a %quot%European Capital of Culture 2010%quot% in accordance with Article 4 of Decision No 1419/1999/EC as amended by Decision No 649/2005/EC.
Article 3
All cities designated shall take the necessary measures in order to ensure the effective implementation of Articles 1 and 5 of Decision 1419/1999/EC as amended by Decision No 649/2005/EC.

Done at Brussels, 13 November 2006.
For the Council
The President
S. Huovinen
[1] OJ L 166, 1.7.1999, p. 1. As amended by Decision No 649/2005/EC (OJ L 117, 4.5.2005, p. 20).

Appendix B Non-Parallel Example Files from the Hallå Norden Corpus (Danish and Swedish)

(The underlined parts of Danish text are missing in the Swedish translation, and the two first sentences are juxtaposed. Also, the second last sentence in the Swedish file is missing in the Danish translation)

Danish:

Stemmeret i Danmark
Kun danske statsborgere med fast bopæl i Danmark som er myndige og fyldt 18 år har stemmeret til folketingsvalg.
Du har stemmeret til kommunalvalg, hvis du er over 18 år, har fast bopæl, er dansk statsborger eller har boet i landet uafbrudt de seneste tre år. Det betyder, at indvandrere og flygtninge kan stemme ved kommunal- og amtsrådsvalg, selv om de ikke har dansk statsborgerskab. Ophold regnes fra den dag man registreres i folkeregistret.
Statsborgere fra EU-lande, Island og Norge kan stemme ved kommunal- og amtsrådsvalg, hvis de har fast bopæl i Danmark. Det samme gælder personer, der arbejder for staten i udlandet eksempelvis diplomater og soldater, samt i enkelte tilfælde deres ægtefælle eller samlever.
Borgere fra andre EU-lande har stemmeret til EU-parlamentet, hvis de har fast bopæl i Danmark og er fyldt 18 år.

Senest opdateret: 16-11-2006

Swedish:

Rösträtt i Danmark

Alla myndiga personer över 18 år som är fast bosatta i Danmark har rösträtt i kommunala val.

Endast danska medborgare har rösträtt i valet till det danska folketinget.

Medborgare i andra EU-länder har rösträtt i EU-parlamentsvalet om de är fast bosatta i Danmark och har fyllt 18 år.

För mer information, se lag 730 av den 9 oktober 1998 på www.retsinfo.dk.

Senast uppdaterad: 24-11-2006