Identification of Parallel Text Pairs Using Fingerprints

Martin Hassel †
† DSV, KTH - Stockholm University
Forum 100
164 40 Kista, Sweden
xmartin@dsv.su.se

Hercules Dalianis †‡
‡ Euroling AB, SiteSeeker
Igeldammsgatan 22c
112 49 Stockholm, Sweden
hercules@dsv.su.se

Abstract

When creating dictionaries for use in for example cross-language search engines, one often uses a word alignment system that takes parallel or comparable text pairs as input and produces a word list.

Multilingual web sites may contain parallel texts but these can be difficult to detect. In this article we describe an experiment on automatic identification of parallel text pairs.

We utilize 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).

Keywords

Cross Language Information Retrieval, Identification of Parallel Text, Prefix Frequency Distribution, A-priori Probability.

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.

Word alignment tools are often used for the creation of bilingual word lists. These tools need parallel corpora to work properly. One source is Internet and the multilingual web sites there. Unfortunately these web sites are often only parallel with regard to web pages.

In [6] and in [2] are described different heuristics to download and identify parallel text. However, these methods are not enough since the downloaded parallel text still can be very noisy.

For example [13] found only 45 percent parallel text pairs on the multilingual parallel web site Hallå Norden (Hello Scandinavia) that was intended to be completely parallel and the parallel pages contained 5 percent non-parallel elements.

Therefore, we found a need to develop and evaluate a new method for identifying parallel and non-parallel texts in corpora covering different language pairs.

2. Related Work

The distinction between a parallel and a comparable corpus is very important and has been discussed in for example [10] and also in [3].

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 [7] 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 authors reported significant improvements over the baseline for Arabic-English and for Chinese-English

In [3] 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 [5] 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. Identifying 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 [9]. 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 [8]). 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.

Fingerprinting using prefix frequencies has for example been used in information retrieval for filtering of similar documents written in the same language [11]. We here attempt to utilize this notion in cross-language text alignment.

4. Data sets and experimental setup

In this set of experiments we have used the JRC-Acquis corpus [12]. 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[1] and named entities (such as countries). 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 document pairs.

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, compare 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 setup 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. 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 [11] 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 [4], and an English ditto, the British National Corpus [1]. 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.

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. We would like to find if 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] Referencing systems do however differ between languages. For example, while some use Hindu-Arabic numerals others use Roman.

Table 1: 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.

model:	1. Parole / BNC normalization using reference corpora		2. no normalization using reference corpora	
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

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 feature with the highest frequency in the source language thus 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, i.e. we still sort the features based on respective feature's frequency in the reference corpus. As stated 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.

5. Results

As can be seen in Table 1 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, however, not seem to be the case if we take a look at Table 2.

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.

Table 2: 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}

	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

6. Conclusions and Future Work

In the experiments described above we have shown that our method for identifying and deleting non-parallel texts from corpora covering different language pairs show great potential. In future experiments we plan to use other language pairs from languages that are not closely related as for examples Swedish and Finnish.

Moreover, further experiments on the identification of parallel text pairs should be carried out on more language pairs, preferably such that contain languages belonging to different language groups. 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 corpus in order to discern that we are not just investigating peculiarities of this specific corpus. Yet another point to be taken is that when taking care so that reference corpora are of equal size one should perhaps not simply use the first n words in the larger corpus, but instead do a random sampling of the desired amount of words.

We believe methods such as ours will improve, for instance, automatic bilingual dictionary construction from unstructured corpora and our experiments will be further developed and evaluated along these lines.

7. References

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

[2] Chen, J and J-Y Nie. 2000. Parallel Web Text Mining for Cross-Language IR. Proceedings of RIAO-2000: Content-Based Multimedia Information Access, pp 62-77.

[3] Fung, P. and B. Cheung (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.

[4] Gellerstam, M., Y. Cederholm and T. Rasmark. (2000) The bank of Swedish. In Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000, pp. 329–333, Athens, Greece, 2000.

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

[6] Ma, Xiaoyi and M. Y. Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web. In Proceedings of MT Summit VII, September, pp. 538-542.

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

[8] 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.

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

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

[11] 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.

[12] Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga, (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.

[13] Velupillai, S. and H. Dalianis (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.