Using Parallel Corpora to Create a Greek-English Dictionary with Uplug

Konstantinos Charitakis
Department of Computer and Systems Sciences (DSV)
KTH-Stockholm University
164 40 Kista
Stockholm, Sweden
kcha@kth.se

Abstract

This paper presents the construction of a Greek-English bilingual dictionary from parallel corpora that were created manually by collecting documents retrieved from the Internet. The parallel corpora processing was performed by the Uplug word alignment system without the use of language specific information. A sample was extracted from the population of suggested translations and was included in questionnaires that were sent out to Greek-English speakers who evaluated the sample based on the quality of the translation pairs. For the suggested translation pairs of the sample belonging to the stratum with the higher frequency of occurrence, 67.11% correct translations were achieved. With an overall 50.63% of correct translations of the sample, the results were promising considering the minimal optimisation of the corpus and the differences between the two languages.

1 Introduction

Due to the diversity of the known languages and the vast amount of resources required to produce a bilingual dictionary, people have turned their efforts towards the automation of the task. The emergence of statistical methods has shown promising results and they have given results accurate enough, with less effort and resources required that could be used for the task of automated dictionary extraction (Brown et al., 1990). Parallel corpora, which are texts aligned together with their translation in one or more languages, are extensively used in statistical translation methods as they contain a vast amount of bilingual lexical information (Veronis, 2000). After the emergence of statistical translation methods many corpora processing systems and tools have been implemented and have been applied to parallel corpora of most of the popular natural languages. However there are not many projects on automated creation of a dictionary between the Greek and English language pair.

Similar work of extraction of Greek-English dictionary was performed by Piperidis et al. (1997; 2005), although in both cases the approach was different as it employed statistical techniques coupled with linguistic information for better results and it was applied on a corpus in software domain and on a corpus consisting of official EU documents respectively.

Related work with the use of the same system is the work described by Dalianis et al. (2007) where they used Uplug on Scandinavian and English parallel corpora and obtained 71% and 93% for precision and recall respectively, for Swedish-English dictionaries.

The primary focus of this paper is on the extraction and evaluation of a Greek-English dictionary created from parallel corpora using the Uplug system. The task was performed without the use of linguistic information and without the use of optimised sentence aligned corpora for the Greek-English language pair.

2 Dictionary Extraction and Evaluation

2.1 The Uplug System

For the processing of the corpora, the Uplug word alignment system was used. Uplug origins from a project in Uppsala University and provides a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora (Tiedemann, 1999).

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 212–215

Uplug uses language-specific pre-processing modules if available. In other case Uplug uses the basic pre-processing modules that run the general tokenizer, the sentence splitter and add simple XML markup.

The word aligner implemented in the Uplug system is the *Clue Aligner* which is based on the combination of word alignment clues. The idea is that features like frequency, part-of-speech, parsing and word form, together with similarity and frequency measures are taken into account and are considered as association clues between words. All these association clues are then combined together in order to find links between words in the source and target languages (Tiedemann, 2003). Uplug uses the word Clue Aligner to iterative size reduction and alignment of the corpora.

2.2 Collection of Parallel Corpora

There are many available public corpora over the web. One of the most interesting attempts is the OPUS corpus (Tiedemann and Nygaard, 2004). However the corpora provided in most cases are already aligned, most often at sentence level and tagged using XML format. There were concerns about the optimised corpora available, in the way that optimised corpora would give optimised results while our intention was to work with as realistic input elements as possible. In order to test the full potential of the Uplug system including its sentence alignment process, it was thought necessary the use of raw text parallel corpora. Therefore a manually created corpus was used.

The English and Greek translated documents included in the corpus were mainly collected from the web site of the European Union.

All web documents were stripped from their HTML format and were included in plain unformatted text in one single text source file. Documents available in PDF format had to be included also in unformatted text in the text source file, after they have been manually aligned at document and paragraph level to their original state.

Corpus	Size	Words	Unique words
English	1.23 MB	196,048	10,450
Greek	2.46 MB	204,043	18,117

Table 1. Characteristics of parallel corpora.

The final bilingual corpus created constituted the Greek text, which contained 204,043 words, and the English text which contained 196,048 words. The Greek text contained 18,117 different unique words while the English text 10,450 unique words (see Table 1 above).

2.3 Processing of Parallel Corpora

All processes after the input of the parallel corpora from pre-processing to dictionary extraction were performed automatically by Uplug. The result of the sentence alignment was used as is, without any corrections. The result was neither filtered nor altered. Pre-processing of the parallel corpora was performed using the basic preprocessing modules due to lack of language specific tools for the Greek language. Word alignment was performed by the Clue Aligner.

2.4 Extraction of Sample Data

The system extracted many translation pairs with frequency of occurrence (f) less than three (f=2 and f=1). These translations were not considered worth evaluating as they were not containing any sign of consistency. The majority of these translations were incorrect although there are exceptions of a few correct ones.

For the evaluation of the results a sample of the output data was used. The sample was extracted using the stratified sampling method. In this method the population is divided into non overlapping categories (stratums). Then random sampling is used to select a sufficient number of elements from each stratum.

The results had to be filtered and only the translation pairs with frequency of occurrence (f) above a threshold included in the evaluation, in order to avoid evaluation of pairs with occurrence that might be based on chances. That threshold was decided to be a frequency of occurrence above or equal to three ($f \geq 3$). Therefore translation pairs with frequency of occurrence less than three were excluded from the process of extraction of the sample and evaluation.

Following this method the population of translation pairs with frequency of occurrence above three was divided in five stratums. The five stratums consisted of pairs with frequency of occurrence:

- f equal to 3 (f=3)
- f equal to 4 (f=4)
- f equal to 5 (f=5)
- f equal to 6 and up to 10 ($6 \leq f < 11$)
- f equal to 11 and up to maximum ($11 \leq f$)

Then a random sample of approximately 100 suggested translation pairs from each stratum was

drawn and five different tables were created. Each table contained the translation pairs that were collected randomly from one of the five categories mentioned earlier.

The total number of pairs with frequency of occurrence above or equal to three ($f \geq 3$) was 1,276 pairs and 498 of them comprised the sample included in the questionnaires.

2.5 Evaluation of Results

There are different ways to evaluate extracted dictionaries. Some of the most common metrics used are precision and recall calculations. However, the use of the above metrics is difficult when the alignments are not just one-to-one (Merkel and Ahrenberg, 1998) like it happens in the extracted dictionary as a result of this project. Therefore the evaluation method used was based on the judgment of fluent Greek-English speakers on the quality of extracted translation pairs. This is a quite common way to evaluate automatically created bilingual dictionaries (Sjöbergh, 2005).

The results of the dictionary were evaluated by classifying suggested translation pairs of the sample into categories depending on their translation quality. It was performed by 12 fluent Greek-English speakers who classified each translation pair of the sample according to one of the five following choices.

1. Accurate – suggested translation is an accurate translation of the source word.
2. Somewhat correct – correct but not accurate translation where someone will understand the meaning of the original word.
3. Undecided – person evaluating is undecided or not familiar with a term.
4. Somewhat incorrect – suggested translation is not correct but can still be useful for a reader to understand the general meaning of a word in a text.
5. Wrong – suggested translation is just plain wrong.

The evaluation rules were left open and the tables containing the randomly collected pairs from the respective stratums were included in the questionnaire in random order regarding the frequency of occurrence, so the evaluation would be as unbiased as possible and reviewers would not realise a pattern in the quality of the translation pairs.

2.6 Results

The results of the analysis of the questionnaires are given in the Tables 2 and 3 below.

	$11 \leq f$ (%)	**$6 \leq f < 11$** (%)	**f=5** (%)	**f=4** (%)	**f=3** (%)
Accurate	42.98	43.27	30.51	23.29	20.06
Somewhat Correct	24.12	19.69	18.72	16.21	14.29
Undecided	2.08	2.28	2.25	1.70	1.58
Somewhat Incorrect	7.84	8.79	10.70	10.92	13.04
Wrong	22.95	25.95	37.79	47.86	51.00

Table 2. Analytical distribution of the evaluation results for each stratum of the sample.

The sum of the percentages of the categories "Accurate" (42.98%) and "Somewhat correct" (24.12%) from the stratum of the sample with the higher frequency ($11 \leq f$) of occurrence is a total of 67.10% of correct translations. Based on the results presented in table 2 above, the overall distribution of the suggested translations based on their quality is given in Table 3 bellow.

	Average (%)
Accurate	32.02 %
Somewhat Correct	18.61 %
Undecided	1.98 %
Somewhat Incorrect	10.26 %
Wrong	37.11 %

Table 3. Overall distribution of translations of the extracted sample based on their quality.

Table 3 contains the calculation of the average of each row from table 2. It shows the average for each category (accurate, somewhat correct, etc.) for all stratums ($11 \leq f$, $6 \leq f < 11$ etc.).

Therefore the correct translations could be summed up to 50.63% of the extracted sample of suggested translations, calculated by adding the percentages of categories "Accurate" (32.02%) and "Somewhat correct" (18.61%).

Table 4 below contains some random examples of suggested translations as appeared in the final output.

96	Commission	Επιτροπή
58	may	μπορεί
4	ensure	εξασφαλίζει

Table 4. Random examples of the suggested translations with their frequency of occurrence.

3 Conclusion

The objective of the project was to use parallel corpora for automated extraction of a bilingual Greek-English dictionary using the Uplug system without the use of linguistic information. The corpora used contained documents in English and Greek retrieved from the Web. The resulted translations of the dictionary were evaluated by Greek-English speakers in order to assess the quality of the suggested translations.

For the suggested translation pairs of the sample belonging to the stratum with the higher frequency (f >11) of occurrence, 67.10% correct translations were achieved.

It was interesting to notice that characteristics such as the quality and the frequency of occurrence of translation pairs are directly proportional (see table 2). In other words one can notice a decrease of the percentage of correct translations as the frequency of occurrence of translation pairs decreases and on the other hand one can notice an increase of the percentage of wrong translations as the frequency of occurrence decreases.

This implies that larger corpora with a bigger collection of documents in the same domain that use the same vocabulary and have a high frequency of usage of the same words, are more appropriate in order to achieve better word alignment quality.

From the analysis of the evaluation of the extracted dictionary sample, it can be concluded that 50.63% of accurate and correct translations has been achieved. This is a respectful percentage of correct translations if someone considers the minimal optimisation of the corpora used, the relatively small size of corpora (400,091 words) and the difference in morphology between the language pair. Of course the different alphabet used by the two languages is also an issue, having in mind that String Similarity measures are used to identify translation equivalents.

Acknowledgments

Many thanks to associate professor Hercules Dalianis for his guidance and enthusiastic supervision and also special thanks to Martin Rimka for introducing me to Uplug.

References

F. Peter Brown, John Cocke, A. Stephen Della-Pietra, J. Vincent Della-Pietra, Frederick Jelinek, D. John Lafferty, L. Robert Mercer and S. Paul Roossin. 1990. *A Statistical Approach to Machine Translation*. Computational Linguistics, 16(2):79-85.

Hercules Dalianis, Martin Rimka and Viggo Kann. 2007. TvärSök - Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian. Presented at the Workshop of the Nordisk Netordbog, Copenhagen, Denmark, April 26, 2007

Magnus Merkel and Lars Ahrenberg. 1998. *Evaluating Word Alignment Systems. PLUG report*. Department of Computer and Information Science, Linköping university, Sweden.

Stelios Piperidis, Sotiris Boutsis and Iason Demiros. 1997. *Automatic translation lexicon generation from multilingual texts*. In Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'97), In 15th International Joint Conference on Artificial Intelligence (IJCAI'97), (pp 57-62). Nagoya, Japan.

Stelios Piperidis, Panagiotis Dimitrakis and Irene Balta. 2005. *Lexical Transfer Selection Using Annotated Parallel Corpora*. In Fifth International Conference on Recent Advances in Natural Language Processing – RANLP 2005, 25 September 2005, Borovets, Bulgaria.

Jonas Sjöbergh. 2005. *Creating a free digital Japanese-Swedish lexicon*. In proceedings of PACLING '05, (pp 296-300). Meisei University, Japan.

Jörg Tiedemann. 1999. *Uplug - a modular corpus tool for parallel corpora*. In the Parallel Corpus Symposium (PKS99). Uppsala University, Sweden.

Jörg Tiedemann. 2003. *Combining clues for word alignment*. In Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03) Budapest, Hungary.

Jörg Tiedemann and Lars Nygaard. 2004. *The OPUS corpus - parallel & free*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

Jean Veronis. 2000. *From the Rosetta stone to the information society: A survey of parallel text processing*. Parallel Text Processing, Jean Veronis (editor), Kluwer Academic Publishers, pp. 1-25.