Creating a reusable English – Afrikaans parallel corpora for bilingual dictionary construction

Aldin Draghoender, Mattias Kanhov
Department of Computer and Systems Science, (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
aldr-dra@dsv.su.se, kanhov@dsv.su.se

Abstract

This paper investigates the possibilities in creating a bilingual English – Afrikaans dictionary by building a parallel corpus and using the Uplug tool to process it. The resulting parallel corpus with approximately 400,000 words per language was created partly from texts collected from the South African government and partly from the OPUS corpus. The recall and accuracy of the bilingual dictionary was evaluated based on the statistical data collected. Samples of translations were generated, compiled as questionnaires and then assessed by English – Afrikaans speaking respondents. The results yielded an accuracy of 87.2 percent and a recall of 67.3 percent for the processed dictionary. Our English – Afrikaans parallel corpora can be found at the following address: http://www.let.rug.nl/tiedeman/OPUS/

1. Introduction

Whether it is for business intelligence, shopping or for communicating in social websites such as Facebook, the Internet has become the largest source of information thus creating a platform for multilingual information retrieval. South Africa is a country with eleven official languages where most of the population only speaks a small percentage of all the languages and could therefore benefit from multilingual information retrieval. For this reason the need of a multilingual dictionary is of great importance.

In this paper we present our work where we created a parallel corpus, ran it through the Uplug tool, generated a dictionary and then finally processed and evaluated it.
Previous research using Uplug for word alignment of parallel corpora was performed by for example Dalianis et al (2009) with 71 percent average frequency and an average recall of 93 percent on Swedish - English. There was also no confirmation that POS-tags improve word alignment. Charitakis (2007) had a Greek-English parallel corpus which comprised of about 200 000 words per language. The conclusion based on their quality was that 51 percent (f>3) of the translations were correct while with higher frequency (f>11) 67 percent was achieved.

2. Creating a reusable corpus

Because of the lack of parallel corpora, we decided to create our own corpus by mining multiple English – Afrikaans bilingual texts from the Internet. However, during the corpus creation process we received a portion of the OPUS corpus by Tiedemann and Nygaard (2004).

This meant that our final corpus would be partly from the OPUS corpus and partly from a parallel corpus that we created by sourcing publications from the South African government website (South African Government Information, 2010). These publications were converted from PDF format to plain text and then manually aligned at paragraph level. Only small modifications were needed after that as the texts already were aligned at sentence level for the most part. The final corpus contained 421,587 Afrikaans words and 397,757 English words respectively and covering three domains: Law, public speeches and technical documentation. Around 200,000 words (roughly 50%) per language originated from the OPUS corpus.

3. Uplug and word alignment

The Uplug system is an application with the purpose of providing a modular platform for the integration of text processing tools (Uplug, 2010). The reason why Uplug was the system of choice is because it has been used in many similar projects and it is fairly easy to get acquainted with. The resulting dictionary contained a total of 87,388 lines of word pairs (translations) with one pair per line after a total runtime of 9 hours 22 minutes and 54 seconds. The dictionary however contained many duplicate words and some wrong character encoding, so it needs to be cleaned. The cleaning was done manually because the errors in the dictionary were often unique, making automated cleaning difficult to configure. The translations with frequency of 2 or less were seen as unreliable and therefore removed from the dictionary.
After removing these duplicates and words with a frequency of 2 or less, we finally got a “cleaned” dictionary with 6,450 word pairs which was a 91 percent decrease from the original size.

4. Evaluation

Finally to evaluate the original- and cleaned dictionary, three different sample texts in English were used along with three different types of measuring techniques. The sample texts were chosen as to cover several domains in order to get reliable results. The following measuring techniques were used:

English words found – to measure the amount of words from the sample texts which were present in the dictionary.

Accuracy – the amount of words found in the sample texts that were present in the dictionary and were correctly translated. The words not found in the dictionary would be ignored.

Recall – the amount of correctly translated words that were found in the sample texts. The words not found would be considered as incorrect translations.

Dictionary	English words found in dictionary	Accuracy	Recall
Original	85.48%	79.11%	71.71%
Cleaned	75.27%	87.16%	67.31%

Table 1. The summarized results.

We compiled a questionnaire from the English words found and their translations that were evaluated by English/Afrikaans speaking respondents as well as Google Translate. The respondents evaluated the word pairs by deeming them either Correct, Partly correct or Wrong.

These results were then used to calculate *accuracy* and *recall*. Google Translate was used because of the small number of evaluating people. The English translation of the word pairs was entered into the translator, if the translation corresponded to the Afrikaans word in the word pair they were considered correct. If the translator produced a different word, that word was then entered into Google Translate. If the English word produced corresponded to the English word in the word pair, it was considered correct or partly correct depending on the accuracy.

5. Results

The average values for the evaluations done of the original and cleaned dictionary are seen in Table 1.

Evaluator	Correct	Partly correct	Wrong
Google translate	85.26%	6.17%	8.57%
Person A	87.35%	8.04%	4.61%
Person B	91.04%	5.91%	3.06%
Person C	91.37%	4.86%	3.77%
Person D	80.77%	5.32%	13.91%
Average	**87.15%**	**6.06%**	**6.78%**

Table 2. Accuracy evaluations for the cleaned dictionary.

The decrease of English words found is understandable as the majority of the translations in the dictionary are low frequency and therefore removed during the cleaning process.

The accuracy for the cleaned dictionary had an average improvement of around 8 percentage points compared to the original dictionary, showing the importance of manual dictionary cleaning.

6. Conclusions and future work

When creating a parallel corpus, we found that many errors can occur when PDF documents are converted to plain text, therefore it is important that the whole text is thoroughly reviewed to identify errors. The texts must also manually be paragraph aligned (and preferably also sentence aligned) to get a good result but it demands a lot of time as most corpora are composed of several thousand sentences or more.

Uplug was a very effective tool when processing the corpus. Except for some duplicate- and double translations as well as an error with wrong character encoding, the whole process worked very well.

The results showed a clear connection between how many English words found from the sample texts, recall and accuracy when comparing the original dictionary with the cleaned one. The size of the dictionary was reduced to 9 percent of its original size after cleaning it, the amount of English words found was reduced to 75.5 percent from the original 85.5 percent while the accuracy increased from 79.1 percent to 87.2 percent, showing that a huge number of the translations with frequency of 2 or less were faulty and unnecessary.

The fact that Afrikaans is closely related to English and in addition to a large corpus, we got a relatively high overall accuracy compared to similar research. We also found that manually processing and cleaning the dictionary is an important step to ensure high accuracy.

For future work, a good idea may be to use a lemmatizer to get the base form of the word which could lead to better results. As we did not find an Afrikaans lemmatizer, one idea could be to use a Dutch lemmatizer since the languages share the same language structure.

For further reading see Draghoender & Kanhov (2010).

Acknowledgement

We would like to thank our supervisor Hercules Dalianis and our respondents who took their time to fill out the translation questionnaires.

References

Charitakis, K. (2007). Using parallel corpora to create Greek -English dictionary with Uplug. In Proc. of Nodalida, 2007, 16th Nordic Conference of Comp. Ling., 25-26 May 2007. Tartu, Estonia.

Dalianis, H., Rimka, M. and Kann, V. (2009). Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian languages. In Proc. of Nodalida, 17th Nordic Conference on Comp. Ling, May 15-16 2009. Odense, Denmark.

Draghoender, A. and Kanhov, M. 2010. Creating a reusable English - Afrikaans parallel corpora for bilingual dictionary construction, B.Sc thesis. Department of Computer and Systems Sciences, (DSV), Stockholm University.

South African Government Information. (2010). [Online] Available at http://www.info.gov.za/ [Accessed 17 March 2010].

Tiedemann, J. and Nygaard, L. (2004).The OPUS corpus - parallel and free. In Proc. of the Fourth International Conference on Language Resources and Evaluation, (LREC), May 26-28, 2004. Lisbon, Portugal.

Uplug, (2010). The Uplug homepage. [Online] Available at: http://www.let.rug.nl/~tiedeman/Uplug/ [Accessed 20 January 2010].