Creating a Reusable English-Chinese Parallel Corpus for Bilingual Dictionary Construction

Hercules Dalianis, Hao-chun Xing, Xin Zhang

Department of Computer and Systems Sciences (DSV)
KTH-Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: {hercules,haoc-xin,xin-zhan}@dsv.su.se

Abstract

This paper first describes an experiment to construct an English-Chinese parallel corpus, then applying the Uplug word alignment tool on the corpus and finally produce and evaluate an English-Chinese word list. The Stockholm English-Chinese Parallel Corpus (SEC) was created by downloading English-Chinese parallel corpora from a Chinese web site containing law texts that have been manually translated from Chinese to English. The parallel corpus contains 104 563 Chinese characters equivalent to 59 918 Chinese words, and the corresponding English corpus contains 75 766 English words. However Chinese writing does not utilize any delimiters to mark word boundaries so we had to carry out word segmentation as a preprocessing step on the Chinese corpus. Moreover since the parallel corpus is downloaded from Internet the corpus is noisy regarding to alignment between corresponding translated sentences. Therefore we used 60 hours of manually work to align the sentences in the English and Chinese parallel corpus before performing automatic word alignment using Uplug. The word alignment with Uplug was carried out from English to Chinese. Nine respondents evaluated the resulting English-Chinese word list with frequency equal to or above three and we obtained an accuracy of 73.1 percent.

1. Introduction

Today, with the advent of the Internet and the publishing of information digitally in several languages it has become more difficult to access information unless one has the right search word. This is especially true if one looks for information in a language that is not mastered so well. Many people living and working in an international environment have a passive understanding of one or several languages and an active understanding of just one or two. Passive understanding means that one can read a language but might not be so good at writing it. Active understanding means that one can both read and write a language well. To carry out multilingual information retrieval one has to be able to translate the query in a search engine from one language to another language of which the user has passive understanding in order to retrieve information. The term Multilingual Information Retrieval (MLIR) refers to the ability to process a query for information retrieval in any language. (Hull & Grefenstette 1996). There have been two main approaches to multilingual information retrieval. One is document translation and the other is query translation using bilingual dictionaries. We focus on the latter, specifically on the Uplug word alignment system (Tiedemann 2002, Uplug 2010) for creating bilingual dictionaries. The Uplug word alignment tool has never to our knowledge been used to align Chinese with any other language, and therefore we decided to test Uplug on Chinese and English and compare our results with other approaches.

2. Background

We carried out a survey to find out which language pairs the Uplug word alignment tool has been used on, which other systems exist and what research has been carried out on aligning Chinese words with other languages,

Table 1, above, provides an overview of previous research on the use of parallel corpora and word alignment to create bilingual dictionaries and their accuracy.

3. Aligning Chinese

The word is the smallest independent element of many languages. Chinese words, however, use a single character as the basic written unit. Unlike Western languages such as English, Spanish or Swedish, Chinese language does not have any delimiters to mark word boundaries, (Zhang et al., 2003). To treat Chinese with natural language processing tools (NLP tools) one needs to perform word segmentation (tokenization) before processing it with NLP tools. NLP tools do contain word segmentation, but it is a trivial task in most European languages.

3.1 Word segmentation

Since there are no delimiters to mark word boundaries, and no explicit definitions of words in Chinese, we need to separate words from the continuous character string. There are, however, some common ambiguities, for example, '马上', that can signify one word meaning 'quickly' or 'immediately', and may also signify two words, '马' and '上' that together mean 'horse' and 'up'. This type of ambiguity is very typical of Chinese (Gao et

al., 2003) and will of course influence the result of Chinese word segmentation.

Author	Tools	Corpus 1	Corpus 2	Accuracy
Charitakis (2007)	Uplug	200 000 Greek words	200 000 English words	67%
Megyesi & Dahlqvist (2007)	Uplug	150 000 Swedish words	126 000 Turkish words	69%
Velupillai & Dalianis (2008)	Uplug	80 000 Swedish words 80 000 Swedish words	69 000 Finnish words 80 000 Norwegian, Danish words	65.4% 93.1%
Piao (2002)	POS	30 000 English words	87 000 Chinese characters	89%
Nyström et al. (2006)	ITools	174 000 Swedish words	153 000 English words	76%
Martin et al. (2005)	UMIACS Limited JHU.AER Emphasis. II	2 000 000 Inuktitut words	4 000 000 English words	50% definite 90% probable
	USheffield	70 000 Hindi words	60 000 English words	77%

Table 1. Comparison of related work in parallel corpora and word alignment

3.2 Undefined words

During word segmentation undefined words can be problematic. A lexicon cannot contain all the place names, institution names and personal names that can occur, such as *Kista, Adecco, Jason, Peter*, etc., but word segmentation for Chinese needs automatically to identify all of those words. For language processing of Chinese, lexical analysis is therefore of vital importance.

3.3 Word segmentation software

There is a large amount of word segmentation software and many ways of carrying out word segmentation for Chinese. We compared different word segmentation software and methods (see Xing & Zhang 2008, Gao et al., 2003, Goh et al., 2005). After thorough study we decided to use ICTCLAS (from the Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS 2010) since it was fast, easy to use and gave high accuracy of 98.45 percent.

```
<p id="1">
<s id="s1.1">
<w id="w1.1.1"></w>
<w id="w1.1.2">Order</w>
<w id="w1.1.3">of</w>
<w id="w1.1.4">China</w>
<w id="w1.1.5">Securities</w>
<w id="w1.1.6">Regulatory</w>
<w id="w1.1.7">Commission</w>
<w id="w1.1.8">No</w>
<w id="w1.1.9">.</w>
</s>
<s id="s1.2">
<w id="w1.2.1">53</w>
</s>
</p>
```

```
<p id="1">
<s id="s1.1">
<w id="w1.1.1">中国</w>
<w id="w1.1.2">证券</w>
<w id="w1.1.3">监督</w>
<w id="w1.1.4">管理</w>
<w id="w1.1.5">委员会</w>
<w id="w1.1.6">令</w>
<w id="w1.1.7">第53</w>
<w id="w1.1.8">号</w>
</s>
</p>
```

Figure 1. After the preprocessing step in Uplug we can se an example on an English and a Chinese sentence respectively, that are aligned to each other and where the words are segmented. All in XML-format

3.4 Processing characters and sentences

We also searched for an appropriate English-Chinese parallel corpus to work with, but we did not find any. Instead we found the multilingual Chinese Government law website (BLR 2008), containing law texts in Chinese and their translation in English. We downloaded the multilingual web pages from the web site. We processed the HTML pages to remove HTLM tags. We processed the Chinese text with the ICTCLAS word segmentation tool (ICTCLAS 2010) for Chinese. After we had carried out word segmentation on the Chinese corpus we discovered that our two parallel corpora; the English and the Chinese were very noisy. This means that one Chinese sentence could correspond to two sentences or to a half sentence in the target language. This noise presents a considerable challenge to automatic sentence alignment processing.

In fact, there is professional sentence alignment software, such as Multiconcord (Multiconcord 2010), which is well known for its high precision in dealing with Western languages, but for Chinese, the best accuracy rate is no more than 60 percent (Xie 2004). We also tried Uplug for the sentence alignment, but the result was still not acceptable. Therefore, we decided to carry out the sentence alignment manually to avoid bias from the wrong sentence match. Furthermore, we added one more carriage return after each sentence as a sentence boundary.

Following the steps given above, we took almost 60 hours of manual work to preprocess 100 000 Chinese characters.

The parallel corpus contains now 104 563 Chinese characters equivalent to 59 918 Chinese words, whereas the corresponding English corpus contains 75 766 English words.

```
<cesAlign version="1.0">
<linkGrp targType="s"
 toDoc="Chinese_Text1.xml" fromDoc="English_Text1.xml">
<link certainty="-1466" xtargets="s1.1 s1.2;s1.1" id="SL0.1">
 <wordLink certainty="0.0698212002188447"
lexPair="Order;中国 令" xtargets="w1.1.2;w1.1.1+w1.1.6" />
 <wordLink certainty="0.0487610552735559"
lexPair="of China Commission;委员会" xtargets="w1.1.3+w1.1.4+w1.1.7;w1.1.5" />
 <wordLink certainty="0.142609161125"
lexPair=";号" xtargets="w1.1.1;w1.1.8" />
 <wordLink certainty="0.0630355014339485"
lexPair="Securities .;证券" xtargets="w1.1.5+w1.1.9;w1.1.2" />
 <wordLink certainty="0.05"
 lexPair="No 53;第53" xtargets="w1.1.8+w1.2.1;w1.1.7" />
 <wordLink certainty="0.104928209116383"
 lexPair="Regulatory;监督 管理" xtargets="w1.1.6;w1.1.3+w1.1.4" />
</link>
```

Figure 2. The XML-file in Figure 1, after word alignment showing link certainties between words.

3.5 Word Alignment

The Uplug word alignment system performs sentence alignment in the first step and produces an XML formatted file (see Figure 1), and then it performs word alignment with certainties (see Figure 2).

4. Evaluation

We obtained 18 999 entries with frequencies ranging from 1 to 322 and we selected the ones with frequency equal to or to above three for the evaluation. That gave us 2 118 entries to evaluate. The evaluation of the results from the word segmentation and the word alignment were evaluated manually.

Charitakis (2007) used five categories for his evaluation scheme: *Sure, Unsure, Somewhat Correct, Undecided* and *Somewhat Incorrect.* We decided, however, to use a simpler evaluation scheme with only three categories, *Accurate, Unsure and Wrong*, so our respondents would have an easier task.

We preferred to make our standard simpler and easier for respondents to handle. We therefore constructed a questionnaire to be used by our nine respondents (see Table 2).

We also extracted 800 random word pairs from the original 2 118 pairs, and evaluated both lists ourselves using the same evaluation scheme as in Table 2.

To make a numeric calculation of the answers we used the following computational formula:

$$\text{Accuracy} = \frac{1 * \text{No of Accurate} + 0.5 * \text{No of Unsure}}{\text{No of total evaluated word pairs}}$$

For example, if there are three word pairs, and one is accurate, one is unsure and the last one is wrong, then the accuracy will be (1*1+0.5*1)/3=50%. The evaluation result is presented in Table 3.

There is some variation in the different accuracies of the respondents, probably because we did not train the respondents enough but also because we should have shown the top contexts of the words, the so-called concordances, to make it much easier to decide if the translation was correct.

English	Chinese	Accurate	Unsure	Wrong	Comment
Article	条				
securities	证券				
assets	资产				
and	和				
compliance	合规				
may	可以				
insurance	保险				
following	下列				
firm	公司				
management	管理				
enterprise	企业				
investment	投资				
contract	合同				
asset	资产				
in	中				
related	有关				
goods	货物				
food	食品				
enterprises	企业				

Table 2. A sample of the questionnaire

No.	Accurate	Unsure	Wrong	Evaluated pairs	Accuracy
0	581	171	48	800	83.3%
1	484	280	36	800	78.0%
2	395	171	232	798	60.2%
3	241	317	236	794	50.3%
4	247	326	221	794	51.6%
5	643	72	84	799	85,0%
6	598	196	7	801	86.9%
7	530	250	20	800	81.9%
8	432	336	32	800	75,0%
9	635	148	16	799	88.7%
Avg.	**478.6**	**226.7**	**93.2**	**798.5**	**74.1%**

Table 3. Our evaluation (No. 0) and our nine respondents' feedbacks

5. Conclusions and Future work

5.1. Conclusions

The main goal of our work was to use a parallel corpus and Uplug to create an English-Chinese dictionary. To do this, we needed to perform the following steps:

- Step 1. Gather bilingual texts to create the parallel corpus.
- Step 2. Prepare the corpus well, including word segmentation and sentence alignment.
- Step 3. Use Uplug to do the word alignment and receive the word pair list.
- Step 4. Evaluate and optimise the word pair list to constitute the final bilingual dictionary.

The average accuracy of the English-Chinese bilingual dictionary we created was 74.1 percent. This result compared with the results of languages that are more closely related, for example Greek-English, is good, but not quite as good as the word alignment results for English-Chinese achieved by Piao (2002).

A sentence aligned word segmented parallel English-Chinese corpus is now available for future research, the Stockholm English-Chinese Parallel Corpus (SEC) or in the OPUS [1] parallel corpora collection (Tiedemann 2008).

5.2 Future work

In our resulting dictionary, we found several English words appearing in several inflected forms. Singular forms and plural forms of nouns in English, such as 'service' and 'services' respectively, both translate as the Chinese word '服务'. Therefore, lemmatising English words before using Uplug may increase accuracy (see Piao 2002). We also noticed that there are several English synonyms translated as the same Chinese word, but there are no Chinese synonyms in the result list. We do not know the reason for this yet. One way to discover it would be to switch alignment order and let Chinese be the source language and English the target language.

We also believe that there may be other factors that affect accuracy. What is the effect on accuracy of increasing the corpus size? Martin et al. (2005) observed that languages with scarce resources, for example, small corpora, obtain better results, adding extra resources as external dictionaries, though for languages with large corpora these additional resources do not make a difference. How does sentence length in the parallel corpus affect the results? Do corpora with short sentences give better word alignment results? Using non-processed comparable corpora from the Internet often causes problems since the corpora are not completely parallel. We need a system that can detect these non-parallel elements and remove them (see Velupillai et al., 2008).

To sum up, our five research directions are:

(1) Analyse the optimum size of the parallel corpus to obtain a dictionary with high accuracy.
(2) Lemmatise the non-Chinese part of the parallel corpus before performing the word alignment.
(3) Identify parallel text pairs to detect parallel and comparable sentences in corpus.
(4) Experiment with the direction, changing the source and target language of the word alignment.
(5) Experiment with the sentence length in the parallel corpus.

We have some further ideas we would like to test, and therefore more in-depth experiments are needed to support or refute our conclusions.

[1] http://www.let.rug.nl/tiedeman/OPUS/

6. Acknowledgements

We would like to offer our thanks to all anonymous reviewers who gave us advice on how to improve the early version of our manuscript. We would also like to thank our respondents, who gave us enthusiastic help in evaluating our results and finally Martin Hassel that did a final check on the structure of our manuscript.

7. References

BLR 2008. Bilingual legal resources: Laws and regulations from Shanxi Province, http://cq.netsh.com/eden/bbs/751605/

Charitakis, K. 2007. Using parallel corpora to create a Greek-English dictionary with Uplug. In the *Proceedings of Nodalida 2007, The 16th Nordic Conference of Computational Linguistics*, 25-26 May 2007 in Tartu, Estonia.

Gao, J., M. Li and C-N. Huang 2003. Improved Source-Channel Models for Chinese Word Segmentation. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL-2003*. Sapporo, Japan, pp. 7-12.

Goh, C-L., M. Asahara and Y. Matsumoto 2005. Chinese Word Segmentation by Classification of Characters. *Computational Linguistics and Chinese Language Processing* Vol.10, No.3, September 2005, pp. 381-396.

Hull, D. A. and G. Grefenstette 1996. Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp 49-57. Association for Computing Machinery, 1996.

ICTLAS 2010. http://ictclas.org/

Martin, J., R. Mihalcea and T. Pedersen. 2005. Word Alignment for Languages with Scarce Resources. *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, MI, June 2005.

McEnery, T. and A. Wilson 2001. *Corpus Linguistics. An Introduction*. Second Edition. Edinburgh. Edinburgh University Press.

Megyesi, B. and B. Dahlqvist 2007. The

Swedish-Turkish Parallel Corpus and Tools for its Creation, in *Proceedings of Nodalida 2007- The 16th Nordic Conference of Computational Linguistics*, 25-26 May 2007 in Tartu, Estonia.

Multiconcord 2010. Multiconcord: the Lingua Multilingual Parallel Concordancer for Windows http://artsweb.bham.ac.uk/pKing/multiconc/l_text.htm

Nyström, M., M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson and H. Åhlfeldt. 2006. Creating a Medical English-Swedish Dictionary using Interactive Word Alignment. *BMC medical informatics and decision making*, 6:35.

Piao, S. 2002. Word Alignment in English–Chinese Parallel Corpora. *Literary and Linguistic Computing,* Vol. 17, No. 2, 2002.

SEC. Stockholm English-Chinese Parallel Corpus. http://people.dsv.su.se/~hercules/SEC

Tiedemann, J. 2002. Uplug A Modular Corpus Tool for Parallel Corpora. Language and Computers, Parallel corpora, parallel worlds. *Selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April, 1999. Edited by Lars Borin., pp.181-197.

Tiedemann, J. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia

Uplug 2010, http://uplug.sourceforge.net/

Velupillai, S. and H. Dalianis 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. Coling 2008: *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, pp. 10–16, Manchester, August 2008.

Velupillai, S., M. Hassel and H. Dalianis 2008. Automatic Dictionary Construction and Identification of Parallel Text Pairs. In: *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS),* September 25-27, Hangzhou, China.

Xie, J-C. Construction and application of a small English and Chinese parallel corpus. *Journal of PLA University of Foreign Languages*, Vol. 27, No. 3, May 2004.

Xing, H. and X. Zhang. 2008. Using parallel corpora and Uplug to create a Chinese-English dictionary, Master thesis, DSV/KTH-Stockholm University.

Zhang, H-P., Q. Liu, H-K. Yu, X-Q. Cheng and S. Bai 2003. Chinese Named Entity Recognition Using Role Model. *The Association for Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 2, August 2003, pp. 29-60.