Creating Sentence-Aligned Parallel Text Corpora from a Large Archive of Potential Parallel Text using BITS and Champollion

Kazuaki Maeda, Xiaoyi Ma, Stephanie Strassel

Linguistic Data Consortium
University of Pennsylvania
3600 Market St., Suite 810
Philadelphia PA, 19104 USA
{maeda, xma, strassel}@ldc.upenn.edu

Abstract

Parallel text is one of the most valuable resources for development of statistical machine translation systems and other NLP applications. The Linguistic Data Consortium (LDC) has supported research on statistical machine translations and other NLP applications by creating and distributing a large amount of parallel text resources for the research communities. However, manual translations are very costly, and the number of known providers that offer complete parallel text is limited. This paper presents a cost effective approach to identify parallel document pairs from sources that provide potential parallel text – namely, sources that may contain whole or partial translations of documents in the source language – using the BITS and Champollion parallel text alignment systems developed by LDC.

1. Introduction

Parallel text is one of the most valuable resources for development of statistical machine translation systems and other NLP applications (Brown et al., 1993). The Linguistic Data Consortium (LDC) has supported research on statistical machine translations by creating and distributing a large amount of parallel text resources for the research communities (Ma and Cieri, 2006). However, manual translations are very costly, and the number of known providers that offer complete parallel text is limited. It is a cost effective approach to identify parallel document pairs from sources that provide potential parallel text – namely, sources that may contain whole or partial translations of documents in the source language.

LDC has recently created a large corpus of sentence-aligned parallel text data for the DARPA GALE [1] Phase 3 MT participating research sites from a large archive of potential parallel text. This corpus contains more than 1 million segment pairs in Arabic-English parallel text and Chinese-English parallel text. In creating this corpus, we utilized the document alignment module of the BITS (Bilingual Internet text Search) system and the Champollion sentence alignment tool, both developed by LDC.

This paper describes the approach LDC took in creating this corpus as well as the contents of this corpus. The source materials and methodologies described in this paper are related to that of Munteanu and Marcu (2005), which describes a method for finding parallel sentences in non-parallel corpora. This paper describes a similar attempt to identify and extract parallel text from non-parallel corpora using existing software written by LDC.

2. BITS and Champollion

The BITS system was developed to find and collect parallel text without human intervention (Ma and Liberman, 1999; Ma and Cieri, 2006). It consists of three components: the language identifier, the webpage retriever and the document aligner. While BITS was designed as a complete system, these components may be used individually, and the document aligner module is, in fact, still extremely useful in identifying parallel documents in an existing bilingual text archive.

A related tool also developed by LDC is Champollion, a robust sentence aligner for parallel text. BITS system does not include a sentence alignment module, so the Champollion tool is a complement to the BITS system.

Parallel text taken from sources that were not designed as complete translations of the source documents is inevitably noisy. The Champollion tool is capable of aligning sentences in noisy data, making it an ideal sentence alignment tool for such data. The Champollion toolkit (CTK) is available for download as opensource from `http://champollion.sourceforge.net/`.

3. Parallel Text Found in LDC's Newswire Collection

LDC regularly collects multilingual news articles from newswire agencies, such as AFP (Agence France Presse) and Xinhua News Agency via direct feeds. News articles in multiple languages, such as Arabic, Chinese and English, are available from these news agencies, and some articles are translations from another language.

LDC collects newswire articles from these sources, applies appropriate processes and formats the articles into a predefined format with SGML markups. A large segment of this collection of newswire articles, the Gigaword corpora, have been published as LDC general publications (Graff et al., 2007; Graff, 2007a; Graff, 2007b).

4. Harvesting Potential Parallel Text from Web Sites

During the past decade, improvement in search engine technologies, high-speed Internet connections, and increase of newspaper web sites in many languages have allowed us to explore potential parallel text in various efficient ways.

[1] Global Autonomous Language Exploitation (http://projects.ldc.upenn.edu/gale/)

News articles from bilingual or multilingual web sites provide additional sources of potential parallel text.

Identifying potential sources of parallel text may be done using search engines, wikipedia and other means. Downloading documents can be done using available crawling tools and harvesting tools. Many newspaper web sites keep old articles in their archives, and in some cases, the URL locations of the archived articles are predictable from the URL path names. In other cases, archived article documents are encoded in such a way that they cannot be easily predicted. In such cases, the Internet Archive (http://www.archive.org) is useful in finding the URLs of archived articles. For example, if one types in "http://www.bbc.co.uk" in the Internet Archive Wayback Machine, the results from more than 2000 copies of the web site between 1996 and 2007 will be displayed. Not all articles are saved for each date, but the top page provides links to the articles from that date. This provides one way of harvesting old documents that may not be linked from the current pages.

It should be noted, however, the use and redistribution of harvested data is subject to intellectual property rights. The distribution rights of the sources included in the GALE found parallel text corpus was negotiated between LDC and the providers, and appropriate agreements were reached in order for us to distribute the data.

5. Formatting Potential Parallel Text Downloaded from Web Sites

LDC's newswire collection processes from web sources use Python scripts written for formatting the documents. The Python scripts uses the BeatifulSoap html parsing library, and are updated whenever the format of the original web source (html, etc) files change.

A common problem with formatting web-harvested documents is that the formats change over time, and it takes a large amount of programming effort to create many versions of the formatting scripts.

In creating this corpus, we took the following approach.

1. Convert html files to "markdown" text format[2] using an existing utility. The "markdown" format is a light weight markup format, which is highly human readable.

2. Write simple Perl or Python scripts to extract the essential components of the news articles from the markdown text format.

This approach turned out to be an efficient approach for our purpose because Step 1 generated fairly human-readable text files. Step 2, the cleaning effort required relatively small amount of programmers' time for each source, and we were able to convert the harvested files into our regular SGML-based newswire format with relatively little effort.

6. Identifying Translation Document Pairs Using BITS

As previously mentioned, the BITS system consists of the following three modular components:

- Language Identifier
- Webpage Retriever
- Document Aligner

For the creation of this corpus, only the document aligner module of BITS was used. For the documents that were already in LDC's news archive, there was no need to identify the language or retrieve documents from websites.

The document aligner module of BITS returns a document similarity score for the document in Language A and the document in Language B. It tokenizes both documents and uses a translation lexicon to return a document similarity score based on the ratio of identified translation token pairs. This process is applied for all pairs within a given set of documents, document pairs that had higher scores than the predetermined threshold were judged as parallel text.

This document alignment could be very time-consuming if the given pool of document are very large. The comparison windows - how many days of documents should be compared - were determined on based on a tradeoff between the speed and the likelihood of finding parallel text. Newswire articles are normally time sensitive and translated articles are normally distributed without much delay. On the other hand, translations of news columns from newspaper web sites may be published days or even weeks later, and the number of articles found on web sites may not be as large as the major newswire agencies, such as AFP and Xinhua. In such cases, a larger comparison window was used.

This document matching process was by far the most time-consuming process in creating this corpus. The document matching processes were distributed to 20 workstations and run during the non-business hours of LDC.

7. Sentence Alignment Using Champollion

All of the documents identified as parallel text were also processed with the Champollion sentence aligner. Champollion is a lexicon-based sentence aligner developed by LDC for robust alignment of noisy data (Ma, 2006). Champollion is ideal for the alignment tasks as the documents here were not initially designed as parallel text, and are inevitably *noisy* – sentences may not have one-to-one mapping, and there may be deletions or additions.

Champollion tokenizes input files, and applies a light stemmer for Arabic and English. It then uses a dynamic programming method to find the optimal alignment which maximizes the similarity of the source text and the translation. The similarity scores are computed in terms of *stf*, the segment-wide term frequency, and *idtf*, the inverse document-wide term frequency. The combined *stf-idtf* measure evaluates the importance of a translated word pair to the alignment of two segments, but is offset by how common the word is in the entire documents. The similarity score is further adjusted with the *alignment penalty*, which is assigned according to the alignment type, such as 1-1 and 1-2, and the *length penalty*, which is a function of the length of the source segment the length of the target segment.

All paired documents in the corpus were segmented into sentences using punctuation-based automatic sentence segmenters, and each segment was marked with the SGML tag *seg*.

[2] http://daringfireball.net/projects/markdown/

```
<seg id="1">First sentence.</seg>
<seg id="2">Second sentence.</seg>
...
```

The Champollion tool was then applied to each document pair. The segment mapping was stored in the alignment files (.align) using the following format. The numbers on the left-hand side indicate the segment IDs in Language A, and the numbers on the right-hand side indicate the segment IDs in Language B. The token *omit* indicates that there was no corresponding segments.

```
1 <=> 1,2
2 <=> 3
3 <=> omit
4 <=> 4,5
```

8. GALE Phase 3 Found Parallel Text Corpus

This corpus was created as part of the training data for the DARPA GALE Phase 3 MT Evaluation program, as a supplement to manual translations. The newswire sources included in this corpus include documents taken from LDC's regular newswire collection as well as documents harvested from web sites. The corpus contains over 57,000 document pairs, 540,000 sentence pairs, of Arabic-English parallel text and over 44,000 document pairs, 690,000 sentence pairs, of Chinese-English parallel text. Part of the source data overlaps with the newswire materials covered by the ISI Parallel Text corpora (Munteanu and Marcu, 2007a; Munteanu and Marcu, 2007b); however, this corpus also covers new material, namely, 1) AFP and Xinhua documents newer than January 2005, and 2) sources other than AFP and Xinhua.

The following bilingual sources were considered as potential sources of Arabic-English parallel text for this corpus.

- AFP (Agence France Presse) (source ID: AFP)
- Xinhua News Agency (source ID: XIN)
- Al Ummah Newspaper (source ID: UMH)
- Al Hayat Newspaper (source ID: HYT)
- Al-Asharq Al-Awsat (source ID: AAW)

Among these sources, AFP, Xinhua and Al Ummah are part of LDC's regular newswire collection. Al Ummah news articles are regularly delivered to LDC in a bitext format – i.e., the original articles in Arabic and their English translations were distributed together; therefore, there was no need for document alignment.

The Al Hayat and Al-Asharq Al-Awsat are collected from their respective web sites. LDC has implemented harvesting processes for these sites and added them to its regular newswire collection in November 2006. All data prior to November 2006 was harvested separately for the purpose of creating this corpus.

The following bilingual sources were considered as potential sources of Chinese-English parallel text for this corpus.

- AFP (Agence France Presse) (source ID: AFP)
- Xinhua News Agency (source ID: XIN)
- China News Service (source ID: CNS)
- People's Liberation Army Daily (source ID: PLA)
- HKSAR - Press Releases from the Government of the Hong Kong Special Administrative Region of the People's Republic of China (source ID: HKS)

For the Chinese-English parallel text, AFP and Xinhua were in LDC's regular data collection. The CNS, PLA and HKSAR data were collected from web sites.

Table 1 summarizes the data volumes in the corpus.

9. Discussion

The methods presented in this paper proved to be effective in reducing the cost of creating a parallel text corpus. A large portion of the programming effort was spent on writing a basic toolkit to harvest, format, and run the document and sentence alignment systems on distributed machines. Then, a smaller amount of programmers' time was spent on modifying the basic toolkit for individual sources. In terms of time, not effort, the most time-consuming part of this project was to run the document alignment system. The larger the pool of the documents to align, the longer it took. We partially solved this problem by reducing the size of pools to compare based on the individual properties of the sources and by distributing the processes to multiple workstations, running the process during non-work hours at LDC.

The amount of found parallel documents varied greatly by source. There were sources for which a large amount of effort was made to harvest the data, and little parallel documents were found. There were also sources which contained a large proportion of parallel text.

The accuracy of document-level alignment was controlled using the following method: 1) a small subset of the data was aligned at the document level using the BITS system; 2) bilingual speakers of both source and target languages judged whether document alignments are correct, 3) the similarity scores generated by BITS were compared against the human judgments, and a threshold for the similarity scores was determined so that the document pairs that scored higher than the threshold were judged as parallel documents, and 4) the threshold was used for the document alignment process.

10. Summary

We described the process for identifying parallel text and creating a sentence-aligned parallel text corpus from an archive of potential parallel text using BITS and Champollion, both developed by LDC. As an alternative to creating parallel text resources by manually translating text, our approach proved to be cost effective in creating a parallel text corpus. The GALE Parallel Text corpus mentioned in this paper is scheduled to be released as an LDC general publication in 2008. The Champollion toolkit (CTK) is available for download as opensource from `http://champollion.sourceforge.net/`. The BITS system is currently not available as opensource software, but

lang-pair	source name	source ID	#docs	#segments	#tokens
arabic-english	Al-Asharq Al-Awsat	AAW	26	1005	19095
arabic-english	Agence France Presse	AFP	6374	68222	987983
arabic-english	Al Hayat Newspaper	HYT	363	13771	261649
arabic-english	Al Ummah Newspaper	UMH	1695	20392	387448
arabic-english	Xinhua News Agency	XIN	48846	437746	8560775
chinese-english	Agence France Presse	AFP	1821	21842	412180
chinese-english	China News Service	CNS	226	2409	36218
chinese-english	HKSAR Press Releases	HKS	24454	485998	9236166
chinese-english	People's Liberation Army Daily	PLA	606	5392	102448
chinese-english	Xinhua News Agency	XIN	17012	177756	3377364

Table 1: Data Volumes in GALE Phase 3 Found Parallel Text Corpus

we plan to make parts of the technologies used in this system available to the community. Both the data and the tools described in this paper are representative of LDC's effort in making more parallel text resources available to the community.

11. Acknowledgment

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

12. References

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword third edition. Published as LDC publication: LDC2007T07.

David Graff. 2007a. Arabic Gigaword third edition. Published as LDC publication: LDC2007T40.

David Graff. 2007b. Chinese Gigaword third edition. Published as LDC publication: LDC2007T38.

Xiaoyi Ma and Christopher Cieri. 2006. Corpus support for machine translation at LDC. In *Proceedings of LREC-2006*.

Xiaoyi Ma and Mark Liberman. 1999. BITS: A method for bilingual text search over the web. In *Proceedings of the Machine Translation Summit VII*.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC-2006*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2007a. The ISI Arabic-English automatically extracted parallel text. Published as LDC publication: LDC2007T08.

Dragos Stefan Munteanu and Daniel Marcu. 2007b. The ISI Chinese-English automatically extracted parallel text. Published as LDC publication: LDC2007T09.