Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System

Budiono, Hammam Riza, Chairil Hakim
Science and Technology Network Information Center (IPTEKnet)
Agency for the Assessment and Application of Technology (BPPT), Jakarta, INDONESIA
budi@iptek.net.id, hammam@iptek.net.id, chairil@iptek.net.id

Abstract

Parallel text is one of the most valuable resources for development of statistical machine translation systems and other NLP applications. However, manual translations are very costly, and the number of known parallel text is limited. Hence, our research started with creating and collecting a large amount of parallel text resources for Indonesian-English. We describe in this paper the creation of parallel corpora: ANTARA News, BPPT-PANL and BTEC-ATR. In order to be useful, these resources must be available in reasonable quantities and qualities to be useful for statistical approaches to language processing. We describe problem and solution as well robust tools and annotation schema to build and process these corpora.

1. Introduction

In recent years, our research focuses in developing Open Source Toolkit for English-Indonesian translation system. We need to build a good quality with reasonable size of parallel corpus in Indonesian-English. We started by collecting Indonesian corpus and perform raw corpus cleaning, translation, alignment and XML tagging. The alignment at sentence levels makes parallel corpora both more interesting and more useful. As long as parallel corpora exist, sentence aligned parallel corpora is an issue which is solved by sentence aligners. In our case, the alignment is performed manually by hand while doing the actual translation.

The task that was carried out by us in gathering corpus was conducted in several stages. Until now, we had several collections from various resources. Among them is the collection of ANTARA News corpus, collection of BPPT-PANL corpus and collection of BTEC-ATR corpus. Respectively this work had various Domain (National News, International News, Business/Economy, Politics, Science, Technology, and Sport) and different sources (News agency, Online Publisher, International institution) leading toward different handling and process.

2. Collection of ANTARA Corpus

ANTARA is the national news agency of Indonesia that has a collection of news articles available in two languages, Bahasa Indonesia and English. ANTARA develop a large news collection for the last 10 years, for various domains, i.e. political news, economics news, international news, national news, sport news, science news and entertainment news. All of these news articles were stored in a database system (Oracle) as comparable corpora and the structure of the database did not have the key pairs between one news article written in Indonesian and the one in English news article.

At the beginning, we had a long tedious process for reaching an agreement between the two sides, ANTARA and BPPT. We asked permission to use these data for our researches to develop automatic translation which in return will help ANTARA's journalist and reporters for translation. In addition, the resulting work will benefit both ANTARA and BPPT in the form of alignment of news articles and key pairs for database improvement.

The main problem is transforming this comparable corpus into parallel corpus. We should distinguish between Parallel Corpora with Comparable Corpora. The latter (comparable corpora) are texts in different

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 92–95,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

languages with the same main topic. A set of news articles, from journals or news broadcast systems, as they refer the same event in different languages can be considered as Comparable Corpora. Consider a news item about the September 11 tragedy. Different newspapers, in different languages will refer the World Trade Center, airplanes, Bin Laden and a lot of other specific words. This is the case of Comparable Corpora which can be used by translators who know day-to-day language but need to learn how to translate a specific term.

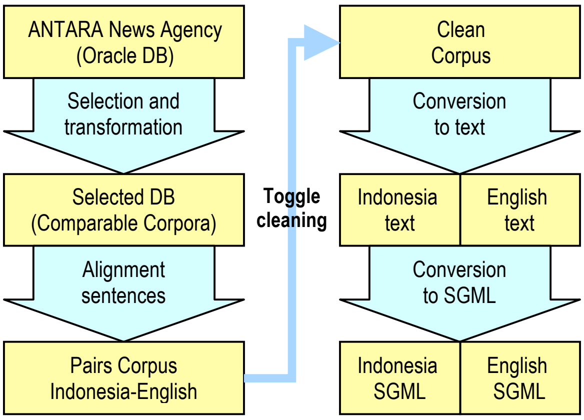

Figure 1. ANTARA Corpus Processing

The number of data's that was used is between the period of year 2000 to 2007 and the articles were taken in stages, in SQL format, amounting to 250.000 sentence pairs (2.5 Million words). These data, afterwards was processed by referring to the news title in respective article to become article pairs. After this article fitting was finished, the next step was to make the pairs of sentence and then the result was store in a new table of database. The work scheme of ANTARA corpus collection is given below in Figure 1.

During the alignment process, the sentences were reviewed manually, by means of election against the quality of the translations. The toggle cleaning stage is used for the process of cleaning of the punctuation mark like [? ! " " ' ' : ; {}]. Afterwards, these sentences pairs was separated into two documents, each for Indonesian and English and put into SGML format. Attention has to be made to keep the consistency of translation from the comparable corpora into parallel corpora.

The ANTARA corpus is used for building machine translation using an open source MOSES SMT. It can be reported here that the BLEU score of 0.76 can be reached by using 1 Million words training set.

3. Collection of BPPT-PANL Corpus

The creation of this corpus is divided into 3 steps [6]. First step is the translation of Indonesian corpus; the second step is the alignment process and resolving issues and followed by tagging of corpus using XML schema in step 3.

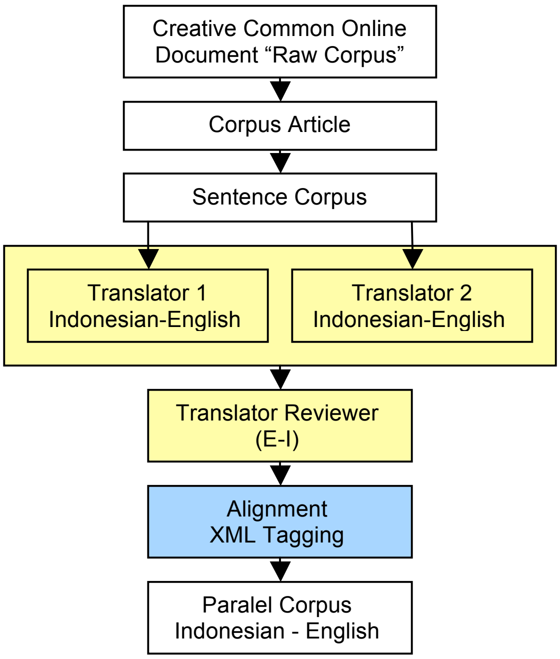
Creative Common Online Document "Raw Corpus"
Corpus Article
Sentence Corpus
Translator 1 Indonesian-English
Translator 2 Indonesian-English
Translator Reviewer (E-I)
Alignment XML Tagging
Paralel Corpus Indonesian - English

Figure 2. BPPT-PANL Corpus Development

3.1 Translation of Indonesian Corpus

We have collected corpora in Bahasa Indonesia covering various domains. This corpus is collected from various online sources which we can apply Creative Commons IPR to its content [2].

Translation, in our project definition, is the semantic and syntactic transfer from a sentence written in Bahasa Indonesia to a sentence in English language. This definition is rigidly constructed, in order to preserve sentence alignment between the original text and the target text in English.

If we are going to translate and align sentences, then obviously we must clarify what we understand by *sentence*. While most people have strong intuitions about what is a sentence and what is not, there is no universal definition of that notion. Before we set out on devising one, however, it should be noted that because PANL-BPPT Corpus is primarily intended to be used as a training text for statistical machine translation systems, both the exact translation and the actual segmentation of the text that results from translation are crucially important.

Our main concern in this regard was to come up with some guidelines for translation that would be both practical for the translators and aligners as well as it is useful for the end-users of the corpus. We started out with something relatively straightforward, which we then expanded as needed.

Of course, given the relative vagueness of the definitions of sentence and translation given above, it was clear that in many situations, arbitrary decisions would have to be made. Our human aligners were instructed to be as consistent as possible. But even then, because of the repetitive nature of the task, errors had to be expected.

3.2 Alignment of Parallel Texts

A parallel text alignment describes the relations that exist between a text and its translation. These relations can be viewed at various levels of granularity: between text divisions, paragraphs, sentences, propositions, words, even characters. While it would certainly have been interesting to produce finer-grain alignments, it was decided that BPPT-PANL Corpus would record correspondences at the level of sentences. This decision was based on a number of factors.

First, sentence-level alignments have so far proved very useful in a number of applications, which could be characterized as *high recall, low precision* applications, i.e. applications where it is more important to have all the answers to a specific question than to have only the good ones.

Secondly is the automatic acquisition of information about translation, as was proposed in [1] as part of a project to build a machine translation system entirely based on statistical knowledge. Such statistical models need to be *trained* with large quantities of parallel text. Intuitively, the ideal training material for this task would be parallel text aligned at the level of words. Yet, because these models picture the translation process in an extremely simplified manner, reliable statistical estimates can nevertheless be obtained from much less precise data, such as pairs of sentences.

For all these reasons, we decided that it would be more appropriate initially to concentrate on sentence-level alignments. Furthermore, we decided to restrict ourselves to "non-crossing" alignments, which is a parallel segmentation of the two texts, into an equal number of segments, such that the *nth* segment in one text and the *nth* segment in the other text are translations of one another [4].

It was suggested that all the texts would be aligned twice, each time by a different aligner. The resulting alignments would then be compared, so as to detect any discrepancies between the two. The aligners were then asked to conciliate these differences together. Because the entire BPPT-PANL corpus was aligned by the same two aligners, this way of proceeding not only minimized the number of errors; it also ensured that both aligners had the same understanding of the guidelines.

3.3 Corpus Tagging

SGML and XML played a major part in the BNC project [3] which serve as an interchange medium between the various data-providers, as a target application-independent format; and as the vehicle for expression of metadata and linguistic interpretations encoded within the corpus.

From the start of the project, it was recognized that we have to choose a standard format such as TEI P4 or XML in order to maintain the corpus for long term storage and also enable distribution of the data. The importance of XML as an application independent encoding format is also becoming apparent, as a wide range of applications for it begin to be realized.

The basic structural mark up of texts may be summarized as follows. Each of the documents or text source articles making up the corpus is represented by a single <corpus> element, containing a header <domain> and <language>, and followed by sentence ID <number>.

The header element contains detailed and richly structured metadata supplying a variety of contextual information about the document (its domain, source, encoding, etc., as defined by the Text Encoding Initiative).

Sample tagging for English as follows:

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<corpus>
  <national>
    <language>english</language>
    <id>1</id>
    <sentence>The Indian government is
    providing scholarships to 20 Indonesian
    students annually including for university
    graduate and post-graduate
    studies.</sentence>
  </national>
</corpus>
```

4. Collection of BTEC-ATR Corpus

BTEC was the abbreviation from Basic Travel Expression Corpus. This corpus this was the everyday normal conversational speech mostly use in traveling and tourism. The source corpus was in monolingual English belonging to NICT-ATR Japan.

As part of A-STAR project cooperation [5], our task was to do the manual translation from English to Indonesian. Similar to the method that was used in developing BPPT-PANL corpus collection; we developed 153.000 utterances into parallel corpora. Additionally, we developed POS Tagging, syllabification and word-stress into this corpus. The main difference was BPPT-PANL originated in monolingual Indonesian that was taken from the domain international, national, economics, sport and science whereas BTEC-ATR originated in speech of English in the travel domain.

5. Summary

After many attempts for having a reasonable size of parallel text for statistical machine translation experiments, we are now having a good quality of parallel corpus collection in Bahasa Indonesia and English as follows:

Name	*Size*	*Domain*	*Origin*	*Annotation Scheme*
ANTARA	250K sentences	News (National Economy)	ANTARA News Agency	TEI P4
BPPT-PANL	500K words	News (Busines, Science)	Online Publisher	TEI, XML, TMX
BTEC-ATR	153K sentences	Travel	NICT-ATR	XML
INC-IX	100K sentences	Parliament Report	BPPT	GDA

References

[1] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The Mathematics of Machine Translation: Parameter Estimation. Computational Linguistics, 19(2).

[2] Wikipedia Creative Commons Website, http://en.wikipedia.org/wiki/, retrieved August 08

[3] Aston, G. and Burnard, L. The BNC Handbook Edinburgh: Edinburgh University Press., 1998

[4] Simard, M. and Plamondon, P. (1996). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. In Proceedings of AMTA-96, Montréal, Canada.

[5] Sakriani Sakti, Eka Kelana, Hammam Riza (BPPT), Satoshi Nakamura, Large Vocabulary ASR for Indonesian Language in the A-STAR Project, 2007.

[6] Riza, Hammam, et.al, PAN Localization Project Report. BPPT, 2008-2009.