PAN Localization Project

RESEARCH REPORT

PHASE 1.2

Research Report on Translation, Alignment, and Other Issues Related to Parallel Corpus Development from Bahasa Indonesia to English

Balai Jaringan Informasi Ilmu Pengetahuan dan Teknologi (IPTEKnet)

BPPT

Research Report on Translation, Alignment, and Other Issues Related to Parallel Corpus Development from Bahasa Indonesia to English

Corpora in general and, particularly, parallel corpora are very important resources for tasks in the translation field like linguistic studies, information retrieval, machine translation systems or natural language processing in general. In order to be useful, these resources must be available in reasonable quantities, because most application methods are based on statistics. The quality of the results depends a lot on the size of the corpora, which means robust tools are needed to build and process them.

In this PANL project, our objective is to build an Open Source system for English-Indonesian translation system. We need to build a good quality with reasonable size of Parallel Corpus Indonesia-English (codename PANL-BPPT). We started by collecting Indonesian corpus and perform raw corpus cleaning, translation, alignment and XML tagging (see Figure 1). 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.

This report is divided into 3 sections. First section reports the translation of Indonesian corpus and describes some definitions for sentence and translation. The second section describes the alignment process and issues related to achieving good quality of translation and alignment. The tagging of corpus using XML schema is given in section 3.

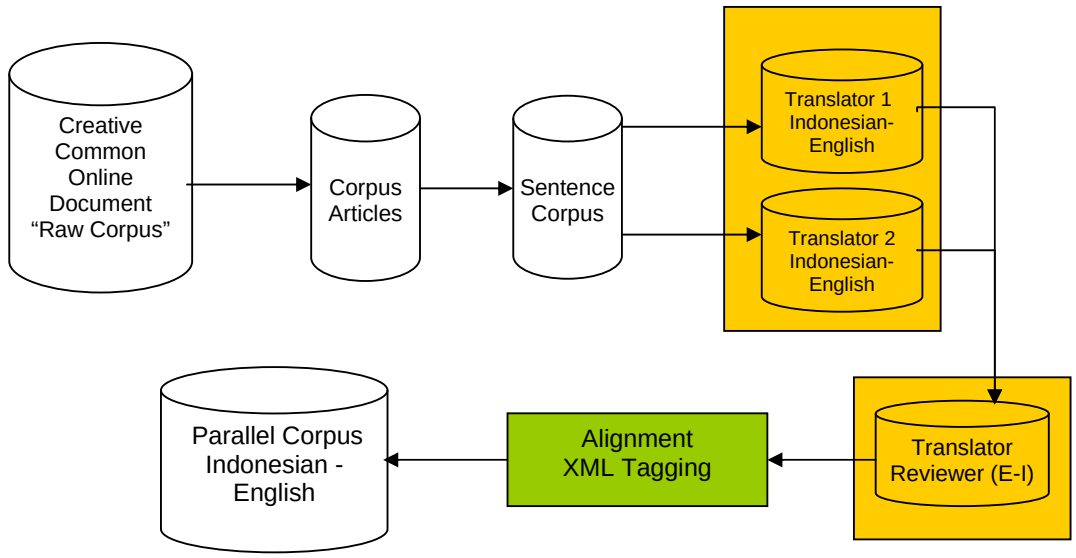

Figure 1. Overview of Translation and Alignment of PANL-BPPT Parallel Corpus

1. Translation of Indonesian Corpus

As we have reported in Phase 1.1, 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 the content.

We intend to build a Parallel Corpora which is formally defined as a collection of texts in different languages, where one of them is the original text (in our case Bahasa Indonesia) and the other are their translations (in our case is English). A similar corpora, for example, the COMPARA (http://www.linguateca.pt/ COMPARA/) project is a parallel corpus of Portuguese/English fiction texts with about 40.000 translation units (at the end of 2003). Another example is the well known Aligned Hansards of the 36th Parliament of Canada (http://www.isi.edu/natural-language/download/hansard/). This is an English/ French corpus with more than one million translation units.

We should distinguish between Parallel Corpora with Comparable Corpora. The latter (comparable corpora) are texts in different 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 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. This kind of corpora can be used by translators who know day-to-day language but need to learn how to translate a specific term. In this case, the translator can find specific comparable corpora where the term is used, and read the translation to search for the translation of that specific term. They can also be used for terminology studies, comparing words used in different languages for similar concepts.

What is a translation?

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. In the followings, we describe the translation and alignment procedure that we adopt in this project.

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 the 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. Essentially, these were the guiding principles:

- *A Sentence is a syntactically autonomous sequence of words, terminated by a full-stop punctuation.*

  The term *full-stop punctuation* naturally includes periods (`.'), exclamation marks (`!') and question marks (`?'), but we also admitted the possibility of a sentence ending with a colon (`:') or semicolon (`;'), as long as the sentence could stand on its own syntactically (this is what we mean by *syntactically autonomous*). In general, we consider that the symbol that explicitly marks the end of a sentence (if such a symbol exists) belongs to that sentence.

- *Titles are sentences.*

  This applies to chapter titles, section titles, table titles, figure titles, etc. even though these generally do not end with full-stop punctuation.

- *Enumerators are sentences.*

  Any number, Roman or Arabic, or letter that appears in front of a title (chapter title, section title, etc.) or paragraph, is a sentence. This is also true of the ``N.B.'' or ``Note:'' that precedes notes.

- *Items of an enumeration are sentences*

  As is, of course, the ``header'' of the enumeration. This rule only applies when the items in the enumeration are separated by semicolons, or when the presentation clearly suggests that this is an enumeration, such as, for example, when all items appear on separate lines.

- *Each cell in a table is a sentence*

  Some documents contained tables. In most cases, however, the formatting of the table was lost, and all that remained was the content of the cells, separated by arbitrary markers (for example, pairs of commas or vertical bars).

The definition of sentence given above suggested a very straightforward way of producing alignments by hand: read both texts in parallel, and segment them as you go along, in such a way that:

1. Segment boundaries always coincide with sentence boundaries;
2. The n^{th} segment in one text and the n^{th} segment in the other are translations of one another;
3. Segments are always as small as possible.

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.

We needed to provide the translators and aligners, with some criteria for determining what constitutes a translation. In general, we found it satisfactory to say that segments of text *A* and *B* were translations of one another if they conveyed the same ``ideas" or ``concepts", at least to an acceptable point. The main practical problems we had to solve revolved around situations where the translation deviated from its usual ``linear" progression.

Parallel text translation problems are categorized into the following classifications:

- Type A: translations cannot be viewed as parallel corpora. The translator often changes the order of sentences and some content as soon as they maintain the basic idea behind the text;
- Type B: translations give reasonable results on word alignment, as most specific terms from the corpora will be coherently translated between sentences;
- Type C: translations are the best type of parallel corpora for alignment.

As this type of parallel corpora is normally composed of institutional documents with laws and other important information, translation is done accurately, so that no ambiguities are inserted in the text, and they maintain symmetrical coherence;

Considering the automatic translation objective, stylistic and semantic translation types can have problems. Stylistic approach makes the translator look for some similar sound, sentence construction, rhythm, or rhyme. This means that the translator will change some of the text semantic in favor of the text style. The semantic approach has the advantage that the text message and semantic is maintained, but the type of language can change (as the translation will be addressed to an audience that differs significantly from the one of the original text).

2. Alignment of Indonesian-English Parallel Texts

The first step in the process of building a corpus of hand-aligned parallel text is to clarify what we understand by the term *alignment*. Essentially, this entails describing the objects that the alignment connects, and defining how the alignment connects them. Based on the answers to these questions, a set of *guidelines* can then be devised, which the human aligners will be instructed to follow when producing the alignments.

What is an Alignment?

Although we can add morphological analysis, word lemmas, syntactic analysis and so on to parallel corpora, these properties are not specific to parallel corpora. The first step to enrich parallel corpora is to enhance the parallelism between units on both texts. This process is called “alignment”. Alignment can be done at different levels, from paragraphs,

sentences, segments, words and characters. Usually, alignment tools perform the alignment at sentence and word levels.

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 the PANL-BPPT 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. One example of such an application is bilingual lexicography. When a lexicographer is examining a bilingual concordance, with a view to mapping out the various meanings or contexts of use of a particular term or expression, she seeks exhaustively. In other words, she is willing to tolerate a relatively high number of irrelevant or redundant examples (noise), in order to make sure that she doesn't overlook anything.

Automatic or machine-assisted translation verification is another such application. A system that does translation verification will look for specific translation errors, such as omissions on the part of the translator, the use of false cognates, inconsistent use of terminology, etc. If translation verification is anything like spelling or grammar checking, we can expect users to be ready to tolerate a fair amount of noise, just to make sure they don't miss out on glaring errors.

A final example is the automatic acquisition of information about translation, as was proposed in (Brown et al., 1993) as part of a project to build a machine translation system entirely based on statistical knowledge. The statistical models at the heart of these projects are now used for our PANL project, Indonesia country component, Other example are interactive MT projects (Foster et al., 1997) and text alignment systems (Simard and Plamondon, 1996). 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.

This explains why a lot of the research effort in this domain has so far focused on sentence-level alignments. Of course, this is not to say that reference alignments at a finer level would not be a useful thing, in the contrary. Besides, a word-level alignment could be made to incorporate the sentence-level alignment as a by-product.

Unfortunately, producing such a thing as a word-level alignment turns out to be a much more difficult problem: while there is often a one-to-one correspondence between the sentences of a text and its translation, matters get a lot more complicated when we get down to the level of ``words''. The main reason is that, at this level, syntactic and stylistic constraints in the target language affect the content and structure of the translated text at least as much as does the source text. As a result, in order to accurately describe the

complex relations that exist between the words of a text and its translation, we will likely need a fairly elaborate alignment scheme. Finally, it is clear that producing hand-made word-alignments for more than a few sentences is going to be a very costly proposition.

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:

- An *alignment* is a parallel segmentation of the two texts, into an equal number of segments, such that the n^{th} segment in one text and the n^{th} segment in the other text are translations of one another.

We refer to such alignments as ``non-crossing" because of the impossibility to explicitly account for *inversions*, i.e. situations where the order of sentences is not the same in the two texts. This type of alignment nonetheless covers the vast majority of situations encountered in real-life texts. Furthermore, this is the type of output that is actually produced by most existing sentence alignment programs.

Issues Related to Alignment

There are several issues with the alignment. First, there were the cases of *omissions* and *insertions*, i.e. situations where some segment in one text does not appear to have a corresponding counterpart in the other text. In these cases, we allowed for the existence of ``empty" segments in the alignment. This way, a sentence that does not have an equivalent in the other text can be aligned with an empty segment.

There were some situations where we chose to ignore an omission (or insertion), for the benefit of recording a larger correspondence. This would happen, for example, if a single sentence *A* in one language was translated as two sentences A_1' and A_2', between which a third, untranslated sentence *B*' was interpolated. In this case, we would simply align *A* with the sequence A_1'*B*'A_2', regardless of the fact that *B*' has no equivalent in *A*.

Then, there was the case of *inversions*. This happens when the order of the sentences is not the same in the source and translated texts. As mentioned earlier, our definition of alignment makes it impossible to explicitly account for inversions. Two different strategies were adopted, depending on the nature of the inversion.

For simple inversions, we opted for a strategy of ``under-segmentation": when a pair of contiguous sentences *AB* appeared as *B*'*A*' in the other text, we chose not to segment the texts after sentences *A* and *B*', but rather to keep *A* and *B* together within the same segment, and then do the same for *B*' and *A*'.

For more complex inversions, we usually chose to treat the inverted segments as omissions. For example, given some sequence of sentences A_1 A_2 A_3 ... A_n translated as A_2' A_3' ... A_n' A_1', we would consider A_1 and A_1' to be ``omitted" segments (align them with empty segments), and then align $A_2$ with $A_2$', $A_3$ with $A_3$', etc. Although this was clearly not the correct way of aligning the texts, it was felt that in the end, such an alignment would be more ``useful".

For these reasons, 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 PANL-BPPT 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.

In this regard, it is interesting to note that the vast majority of disagreements between the two aligners revolved around questions of sentence segmentation rather than questions of translational equivalence. Considering that, as discussed earlier, the actual segmentation on which the alignment is based is not crucially important, this suggests that it would probably have been a good idea to first have the texts segmented by a single person, and then have the aligners produce the alignments based on that segmentation.

The following shows the alignment tools that is used to help the two aligners to perform the sentence alignment.

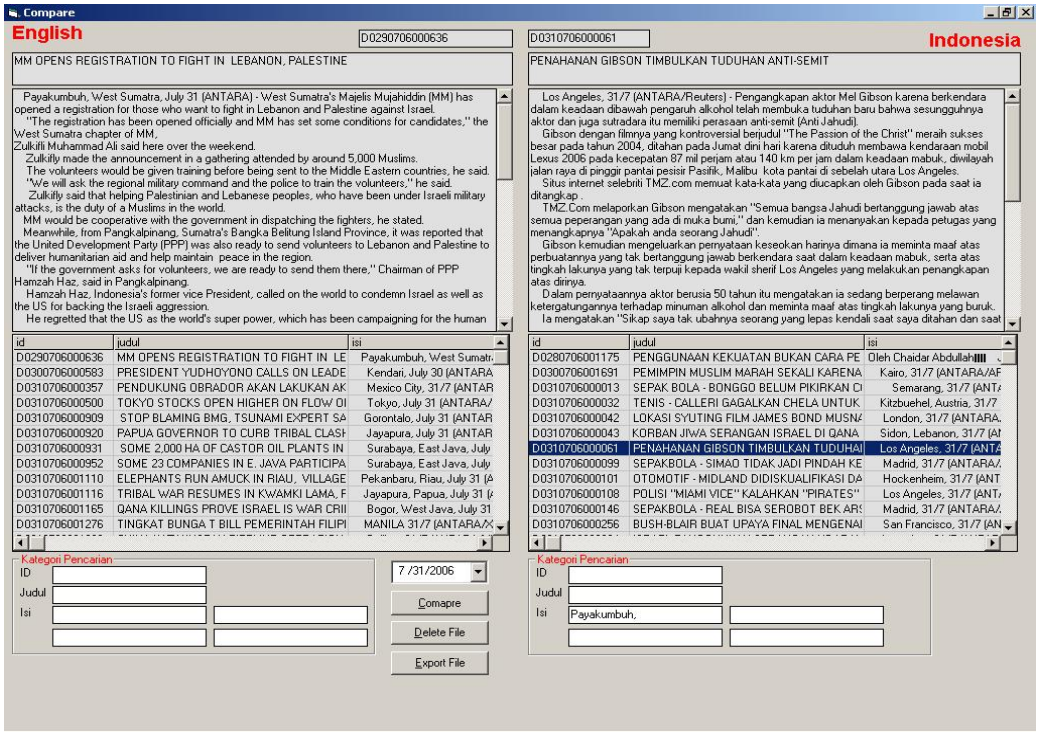

Figure 2. Screen shot of Corpus Alignment Tools

3. Corpus Tagging

In consideration with a standard Indonesian corpus format, we have look into SGML as the mark up language. SGML and XML played a major part in the BNC project 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 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).

The headers were automatically generated from information managed within a relational database constructed from a raw corpus.

At this stage, we did not adopt the idea for written text divided into paragraphs, possibly also grouped into hierarchically numbered divisions. Below the level of the paragraph, all texts are composed of <s> elements, marking the automatic linguistic segmentation, and each of these is divided into <w> (word) or <c> (punctuation) elements. At the subsequent processing of the corpus, we expect that POS Tagging performed at University of Indonesia (UI) will insert a POS (part of speech) annotation attribute to each <w> and <c> elements.

Table 1. Sample Indonesia and translated English corpus

Indonesia	**English**
Irak akan tegas hadapi militan	Iraq is to be tough on militants
Perdana menteri Irak Nouri al-Maliki mengatakan dia bertekat untuk bertindak melawan milisi-milisi ilegal yang dituduh mendorong pertumpahan darah sektarian di Irak	The Iraqi prime minister Nouri al-Maliki says he's determined to act against illegal militias which have been blamed for growing sectarian bloodshed in Iraq
Maliki mengatakan hanya pasukan pemerintah yang berhak membawa senjata dan dia akan menghantam kelompok mana pun yang menantang otorita negara atau bertindak melanggar hukum	Mr Maliki said only government forces had the right to carry weapons and he would strike hard against any group that challenged state authority or acted outside the law
Dutabesar Amerika di Irak Zalmay Khalilzad pada hari Selasa mengatakan bahwa para pemimpin Irak telah menyepakati jadwal bagi langkah-langkah politik dan keamanan	The American ambassador in Iraq Zalmay Khalilzad said on Tuesday that Iraqi leaders had agreed to a timetable of political and security measures
Ini termasuk tindakan melawan milisi Dan dia memperkirakan kemajuan besar akan terjadi dalam 12 bulan ke depan	These include action against the militias And he expected significant progress within the next 12 months

Sample Tagging for monolingual English and Bahasa Indonesia 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>
- <national>
  <language>english</language>
  <id>2</id>
  <sentence>By establishing good cooperation with Bali, hopefully that more Indians
     could spend holidays in Bali, she said.</sentence>
    </national>
    </corpus>

<?xml version="1.0" encoding="iso-8859-1" ?>
- <corpus>
- <national>
  <language>indonesia</language>
  <id>1</id>
  <sentence>Pemerintah India memberikan beasiswa kepada 20 mahasiswa
     Indonesia setiap tahunnya untuk melanjutkan ke jenjang
     pendidikan.</sentence>
    </national>
- <national>
  <language>indonesia</language>
  <id>nas2</id>
  <sentence>Adanya kerjasama yang baik dengan Bali, diharapkan masyarakat India
     dapat menikmati liburan ke Bali, kata dia.</sentence>
    </national>
    </corpus>
```

Sample Tagging for English and Bahasa Indonesia using the Omega-T as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
  <!DOCTYPE tmx SYSTEM "tmx11.dtd">
  <tmx version="1.1">
    <header
     creationtool="OmegaT"
     creationtoolversion="1.8.1"
     segtype="sentence"
     o-tmf="OmegaT TMX"
     adminlang="EN-US"
     srclang="ID"
     datatype="plaintext"
    >
    </header>
     <body>
       <tu>
         <tuv lang="ID">
```

```
        <seg>Inflasi 1,04 persen pada Desember 2006.</seg>
      </tuv>
      <tuv lang="EN-US">
        <seg>Inflation clocked in at 1.04 percent in December 2006.</seg>
      </tuv>
    </tu>
    <tu>
      <tuv lang="ID">
        <seg>Semoga masalah ini hanya sementara.</seg>
      </tuv>
      <tuv lang="EN-US">
        <seg>Hopefully, these downturns will be mere temporary.</seg>
      </tuv>
    </tu>
  </body>
</tmx>
```

4. Reference

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. Foster, G., Isabelle, P., and Plamondon, P. (1997). Target-Text Mediated Interactive Machine Translation. *Machine Translation*, 21(1-2).
3. Wikipedia Creative Commons, http://en.wikipedia.org/wiki/, retrieved August 08
4. Commons Wikipedia, http://commons.wikimedia.org/wiki/Main_Page, Sept 2008
5. Aston, G. and Burnard, L. *The BNC Handbook* Edinburgh: Edinburgh University Press., 1998
6. Simard, M. and Plamondon, P. (1996). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. In *Proceedings of AMTA-96*, Montréal, Canada.
7. Garside, R., Leech, G., and McEnery, T. *Corpus annotation: linguistic information from computer text corpora*, Harlow: Addison-Wesley Longman, 1997.