Improving Word Alignment in an English – Malay Parallel Corpus for Machine Translation

Suhaimi Ab. Rahman **Normaziah Abdul Aziz, Ph.D**

Language Engineering Research Lab
MIMOS
Malaysia Technology Park
57000 Kuala Lumpur, Malaysia
smie@mimos.my, naa@mimos.my

Abstract

A bilingual parallel corpora is an important resource in constructing an English – Malay Bilingual Knowledge Base that is heavily referred to in our English to Malay machine translation system. We present an approach that we applied at word level alignment from a bilingual parallel corpora to improve the translation quality of our English to Malay Example-based machine translation. Initially, *one-to-one* word alignment was applied against the source and target languages. We revised this method to a *many-to-one* word alignment. The comparison of translation results for both method shows that our *many-to-one* word alignment is capable to improve the translation quality.

1. Introduction

We have compiled an English-Malay bilingual parallel corpus consist of 250,000 words in the domain of agriculture and health. These two domains were put in placed as an initial deployment of the English-Malay machine translation service to the rural community users. This research and development (R&D) project's goal is to address the language barrier issue as part of narrowing the Digital Divide problems in Malaysia[1]. In country such as Malaysia where English is a second language for majority of its people, language is one of the factors that ought to be addressed in the digital divide issues. Hence, the need of tools such as online machine translation systems is important to ensure that the non-English speakers community could broaden their knowledge resources unlimited to their native language as discussed in (Aziz, N., et al, 2002).

We started our applied research work with University of Science Malaysia's (USM) prototype machine translation. USM's works surrounding this research have been described in various technical platforms such as (al Adhaileh, Tang, 2001) and (al Adhaileh, Tang, Zahrin, 2002), among others. We continued the work by upgrading USM's *proof-of-concept* version to a *real usage* version for deployment to the digital divide communities.

Work on alignment of parallel corpus for machine translation, sense disambiguation, information retrieval for multilingual environment and other language related researches have been actively discussed at various perspective and levels of discussions such as in (Chen, 1993), (Dagan, Church and Gale, 1993), (Gaussier, 1998), and (Ahrenberg, et.al., 2000) among others. However, our discussion in this paper is based on an experience that we encountered while developing and testing in upgrading a prototype machine translation and not out of a theoretical research exercise. Referring to our English-Malay parallel corpora, this paper discusses on how we revised the word level alignment from the bilingual parallel corpora to improve the translation quality of the English to Malay Example-based machine translation (EBMT).

2. Parallel Corpora for EBMT

A bilingual of English and Malay parallel corpora is a significant resource in constructing an English – Malay Bilingual Knowledge Base (BKB). This BKB is heavily referred to in our English to Malay example-based machine translation system.

As for the process of alignment in our English-Malay parallel corpora, initially, an auto sentence alignment process together with an English-Malay dictionary mapping are applied to align our English-Malay parallel text. An alignment algorithm that uses English-Malay dictionary mapping offers a potential for higher accuracy of word alignment that leads to better translation quality. After these two processes, we manually review the result of English-Malay bi-texts through post-editing to improve the English – Malay word alignment. The bilingual parallel text will be used in constructing a bilingual knowledge bank (semi-) automatically through

[1] This work is funded by MIMOS' Bridging Digital Divide Programme, under the 8th Malaysia Plan.

available parsers and alignment tools. A representation schema named *Synchronous Structured String-Tree Correspondence* (S-SSTC) is used to annotate the translation example pairs, describing the correspondence relation between the source and target sentences (Al-Adhaileh, 2002).

Referring to the alignment process for content of our parallel corpora, here we are addressing the problem of "many English words to be represented in one Malay word", which then improves the linguistic quality of translation by our English to Malay EBMT. The following are a few cases of English words and its translation of Malay word(s). Note that in these cases, the number of the translated word(s) of the target language is one or lesser than the number of words from its source language. Using such replacement will produce a better translation.

as well as --- juga
as long as --- selagi
in order to --- supaya
such as --- seperti

The following Figure 1 shows some examples with the above words in the bilingual parallel text that is used for our EBMT.

English (E)	**Malay (M)**
E1 : Wild flowers ***such as*** orchids and primroses are becoming rare.	M1 :Bunga-bunga hutan ***seperti*** orkid dan primros semakin jarang ditemui.
E2 : ***As long as*** you maintain your diet, you don't need to worry about your health.	M2 ***: Selagi*** anda menjaga pemakanan anda, anda tidak perlu risau mengenai kesihatan anda.

Figure 1: Example of the English-Malay Bilingual Corpus.

3. Word Level Alignment

It is necessary to align the two texts of the target and source language to extract information from the parallel corpora. The alignment process is meant to associate chunks of text in the source language document with the ones of the translated version in the target language as discussed in (Somers, H.) In our work, the alignment is done at sentence and word level.

The initial auto alignment algorithm at word level splits each word in a sentence, one by one. We refer to this approach as *one-to-one* word alignment method. Due to the nature of Malay and English at linguistic level, there are instances where several words in English are best represented or translated to one Malay word. This is also true at instances where one English word needs to be represented in a few Malay words, when translated. However, here we are addressing "English phrases that is to be represented in one Malay word" alignment.

3.1 Many-to-One Approach

The processes that are involved in our many-to-one word alignment are as follows:

1. Get the aligned source sentence.
2. Generate the list of word-form from the source sentences from lexicon parser process.
3. In order to get many-to-one word, the process will refer to the lexicon parser, which contains phrases based on English grammar.
4. The logical dictionary mapping is used to retrieve the meaning for each word-form.
5. Improve the word-level alignment output, when necessary by manual post-editing.

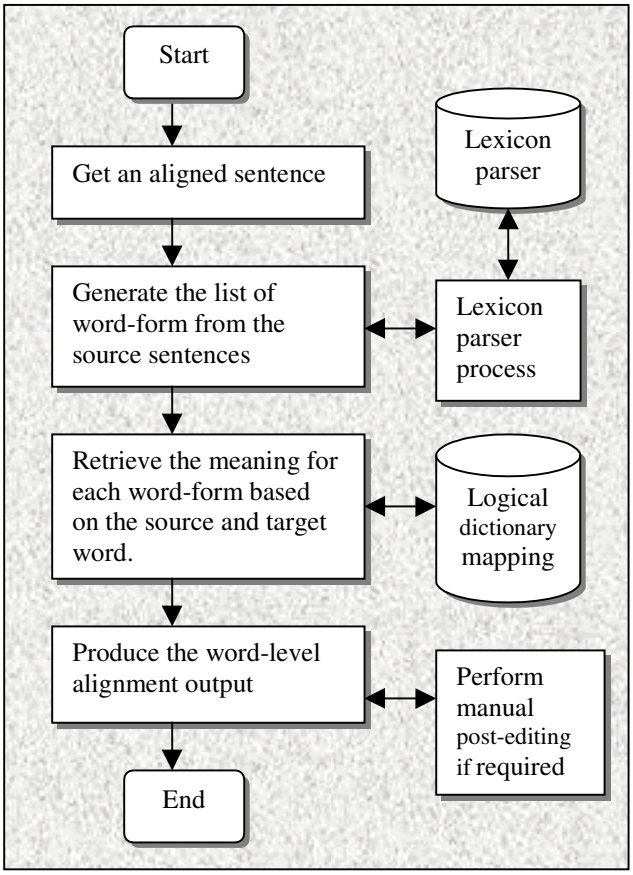

Figure 2: The process involved in the many-to-one approach word alignment

In this process, the lexicon parser is of great use where we can add the phrases related to this *many-to-one* word alignment issue. It contains the identified phrase together with its lexicon tag.

Besides that, we further improve the algorithm which then eliminates redundancy data and thus, reduce its run-time. In other words, i) after compiling the *many-to-one* word-level alignment, we manage to reduce the number of alignment between source word and target word; ii) reduce the occurrence number of *null words* returned after the *many-to-one* alignment being made; where, null word is the word-level alignment that carries no meaning. With these steps taken, we have a better version of parallel corpus to work on.

3.2 The Parallel Corpus Alignment

Below is an example of the different word-level alignment using *one-to-one* and *many-to-one* for the following English-Malay corpus.

Source (English) :
$_0$The $_1$doctor $_2$advises $_3$her $_4$to $_5$rest $_6$***as*** $_7$***long*** $_8$***as*** $_9$she $_{10}$needs$_{11\text{-}12}$

Target (Malay) :
$_0$Doktor $_1$menasihati $_2$perempuan $_3$itu $_4$untuk $_5$berehat $_6$selagi $_7$dia $_8$perlu$_{9\text{-}10}$

The alignment of both source and target sentences are described in Figure 4 a) in a one-to-one alignment approach and Figure 4 b) in a many-to-one alignment method. The dependency tree for each approaches are also shown respectively.

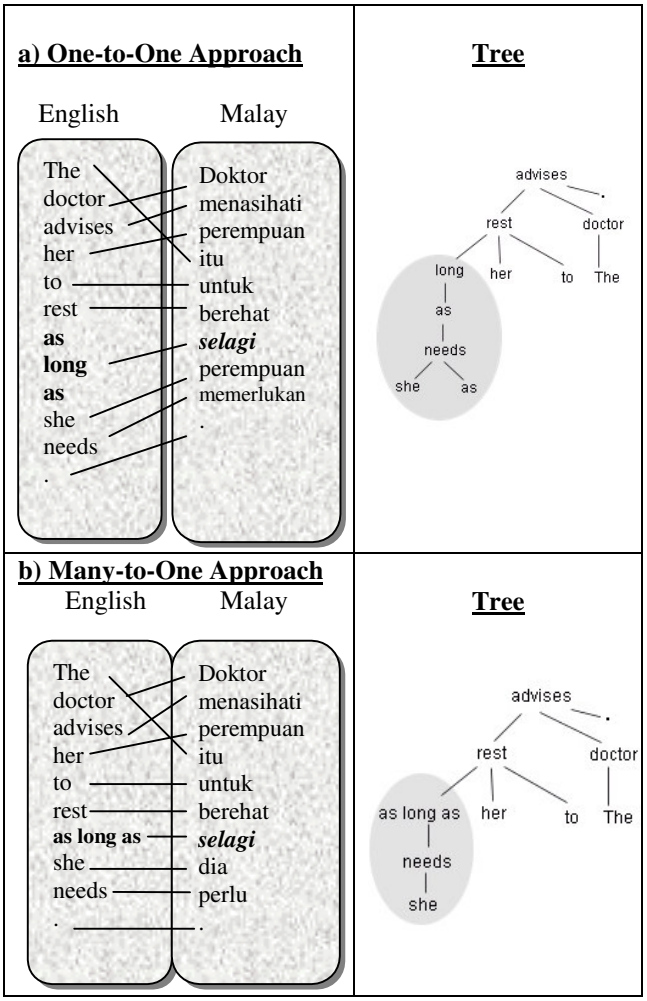

Figure 4: Word-level alignment for the translation pair (a) one-to-one approach. (b) many-to-one method.

4. Test Results

We assign to the English corpus *E* translating to the Malay corpus *M* with a particular alignment. For example, sentence E_1 corresponds to the target sentence M_1. From the parallel corpus (E_1,M_1) in Figure 4, it shows the difference alignment output generated by using *one-to-one* and *many-to-one* methods. The phrase word $_0$***such as***$_1$ from one-to-one method is separated into two words: $_0$***such***$_1$***as***$_2$. Meanwhile, many-to-one method combined the phrase word $_0$***such***$_1$***as***$_2$ into one word $_0$***such as***$_1$. The combination of this phrase word $_0$***such as***$_1$ is produced in the lexicon parser process. Other sentences which contain identified phrase that are in the lexicon parser will go through the same process as described.

The following Figure 5 shows the different results of word level alignment using one-to-one and many-to-one word alignment method.

Example : Bilingual Corpus (E,M): E_1 : Wild flowers ***such as*** orchids and primroses are becoming rare. M_1 : Bunga-bunga hutan seperti orkid dan primros semakin jarang ditemui.	
One-to-one word alignment method	**Many-to-one word alignment method**
Wild -- > hutan	Wild -- > hutan
flowers -- > Bunga-bunga	flowers -- > Bunga-bunga
such -- > seperti	***such as -- > seperti***
as -- > null	orchids -- > orkid

orchids -- > orkid	and -- > dan
and -- > dan	priroses -- > primros
primroses -- > primros	are -- > null
are -- > null	becoming -- > semakin
becoming -- > semakin	rare -- > jarang ditemui
rare -- > jarang ditemui	. -- > .
Test:	
New input sentence E1: You have to eat more vegetables ***such as*** salad, spinach and mustard.	
Results:	
Translation results using one-to-one word alignment M1a: Anda ada untuk makan lebihan banyak sayuran sebagai seperti salad, bayam dan sawi.	
Translation results using many-to-one word alignment M1b: Anda perlu makan banyak sayuran seperti salad,bayam dan sawi.	

Figure 5: Word level alignment, testing and result using *one-to-one* and *many-to-one* approach.

Referring to the test above, there are two different results generated form the EBMT system. The example of the input sentence focusing to the phrase word $_{0}$***such as***$_{1}$. By referring *one-to-one* method, it shows that the translation is more on word-to-word translation. This is because i) the dependency tree or sub-tree for the phrase word $_{0}$***such as***$_{1}$ is not found in our Bilingual Knowledge Base (BKB), in the context of the input sentence; and ii) the phrase word $_{0}$***such as***$_{1}$ is not in the lexicon parser. Meanwhile, for *many-to-one* method, the translation is more accurate because i) the dependency tree or sub-tree of the phrase word $_{0}$***such as***$_{1}$ found in the BKB; and ii) the lexicon parser process found the phrase word $_{0}$***such as***$_{1}$ in the lexicon parser.

We revised the *one-to-one* auto alignment to a *many-to-one* word alignment for relevant cases. After running several test data of 100 English sentences with such words (e.g. as long as, such as, years old, in order to), we discovered that this *many-to-one* word alignment manage to ensure the construction of a more accurate of our Bilingual Knowledge Base, thus better quality of translation result i.e. from the Malay linguistic perspective.

5. Conclusion

The discussion above shows that the translation improvement could be made via a *many-to-one* word alignment of a bilingual parallel corpus, in the context of an English and Malay parallel text. The improvement is significant to us when we are refining the translation quality (from the perspective of Malay language). At the same token, we managed to reduce the processing time at a factor of 4 for searching the proper word alignment between the source and target word in the bilingual parallel corpora.

References

Al-Adhaileh, M.H. (2002). Synchronous Structured String Tree Correspondence (S-SSTC) and its Application for Machine Translation, PhD Thesis, University of Science Malaysia.

Al-Adhaileh Mosleh H. & Tang Enya Kong. (2001). Converting a Bilingual Dictionary into a Bilingual Knowledge Bank based on the Synchronous SSTC. Proceedings of MT Summit VIII, Santiago de Compostela, Spain, 18 Sept 2001.

Al-Adhaileh, Mosleh H., Tang Enya Kong and Zaharin Yusoff. (2002). A Synchronization Structure of SSTC and its Applications in Machine Translation. The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan.

Aziz, N., et al. (2002). Is Machine Translation Still Relevant?, in MIMOS 2002 Tech-Symposium Proceedings.

Chen, S. F. (1993) . Aligning sentences in bilingual corpora using lexical information. In Proceedings of ACL-93, Columbus OH.

Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust Bilingual Word Alignment for Machine Aided Translation. In Proceedings of the Workshop on Very Large Corpora: Acad. & Industrial Perspectives, Columbus OH.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In Proceedings of COLING-ACL-98, Montreal, (pp. 444-450).

Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein & Jörg Tiedemann (2000). Evaluating Word Alignment Systems. Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000), Athens, Greece, 31 May - 2 June, 2000, Volume III: 1255-1261.

Somers, H., Bilingual Parallel Corpora and Language Engineering, available at http://www.emille.lancs.ac.uk/lesal/somers.pdf