The Construction of Bilingual Knowledge Bank based on the Synchronous SSTC Annotation Schema

Mosleh H. Al-Adhaileh & Tang Enya Kong
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, MALAYSIA
{mosleh, enyakong}@cs.usm.my

Abstract

In this paper, we would like to present an approach to construct a huge Bilingual Knowledge Bank (BKB) from a given bilingual corpus based on the idea of synchronous Structured String-Tree Correspondence (SSTC). The SSTC is a general structure that can associate an arbitrary tree structure to string in a language as desired by the annotator to be the interpretation structure of the string, and more importantly is the facility to specify the correspondence between the string and the associated tree which can be non-projective. With this structure, we are able to match linguistic units at different inter levels of the structure (i.e. define the correspondence between substrings in the sentence, nodes in the tree, subtrees in the tree and sub-correspondences in the SSTC). This flexibility makes synchronous SSTC very well suited for the construction of a Bilingual Knowledge Bank we need for the English-Malay MT application.

Keywords: *Structured String-Tree Correspondence (SSTC), Bilingual Knowledge Bank (BKB), Example-Based Machine Translation (EBMT).*

1 Introduction

Recently, much effort was devoted to the compilation of the bilingual corpora for the purpose of machine translation. There is a strong argument that a bilingual corpus, when appropriately structured, can largely replace conventional dictionaries and grammar rules in machine translation. With this objective in mind, we propose, in this paper, an approach to construct a Bilingual Knowledge Bank (BKB) from a given bilingual corpus. In our approach, we introduce a flexible annotation schema called synchronous Structured String-Tree Correspondence (SSTC), which will be used as the basic structure to annotate translation pairs in the bilingual corpus. The SSTC is a general structure that can associate an arbitrary tree structure to string in a language as desired by the annotator to be the interpretation structure of the string, and more importantly is the facility to specify the correspondence between the string and the associated tree which can be non-projective. The flexibility in the mapping from source to target languages, using synchronous SSTC, makes possible to state direct correspondences without a mediating interlingual representation. By doing this, we are able to match linguistic units at different inter levels of the structure (i.e. define the correspondence between substrings in the sentence, nodes in the tree, subtrees in the tree and sub-correspondences in the SSTC). This flexibility makes synchronous SSTC very well suited for the construction of a Bilingual Knowledge Bank we need for the English-Malay MT application.

In this paper, we will propose an approach to construct a huge BKB by incorporating some of the existing tools in the annotation process. First, bitext alignment tools that have been proven their efficiency on other pairs of languages (i.e. SIMR: a bitext mapping tool and GSA: a segment alignment tool) will be adapted to perform English-Malay bitext alignment. Each English sentence in the aligned bitext will then be annotated with part of speech (POS) and phrase structure tree produced by the Apple Pie Parser (APP) for English. The annotated English sentences will then be compiled into an SSTC structure. Next, the Malay SSTC structure of each Malay sentence will be generated based on the corresponding English SSTC structure and the alignment mapping. Finally, the resultant pair of English and Malay SSTCs will be edited semi-automatically to obtain a

synchronous SSTC, which is the basic element of BKB.

2 Bitext Mapping and Alignment

In our proposed approach, texts that are available in two languages (English-Malay bitexts), are the main source of data. The first step in extracting useful information from bitexts is to find corresponding words and terms in the bitext (i.e. bitext mapping and text alignment). To achive this, bitext alignment tools that have been proven their efficiency on other pairs of languages (i.e. SIMR: a bitext mapping tool and GSA: a segment alignment tool) will be adapted to perform English-Malay bitext alignment. A brief introduction to the SIMR/GSA tools is given in the next subsection.

2.1 SIMR bitext mapping / GSA segment alignments tools

SIMR, the Smooth Injective Map Recognizer, a generic pattern recognition algorithm that is particularly well suited to mapping bitext correspondence. SIMR exploits the correlation between the lengths of mutual translations. Like the *char-align* [4], SIMR infers bitext maps from likely points of corresponde-nce between the two texts, points that are plotted in a two-dimensional space of possibilities. Unlike other methods, SIMR greedily searches for only a small chain of correspondence points at a time.

SIMR can be used with the Geometric Segment Alignment (GSA) algorithm. Given a sequence of segments boundaries for each half of a bitext, the GSA algorithm reduces sets of correspondence points to segment alignments. A set of correspondence points, supplemented with segment boundary information, which are produced by the SIMR, expresses segment correspondence, which is a richer representation than segment alignment. For more details on SIMR/GSA algorithms, see [7], [8]. Figure 1 gives an example to illustrate the output from the processes bitext mapping and alignment.

3 The Construction of BKB based on Synchronous SSTC

In Example-Based Machine Translation system [10], the use of Bilingual Knowledge Bank (BKB) containing the bilingual parallel texts encoded with correspondences between the *source* and the *target* sentences is quite popular in implementing such EBMT systems. Sentences in the BKB are normally annotated with their constituency or dependency structures [9]; which in turn allow the correspondences to be established at the structural level. Here, to facilitate such structural annotation, we use the Structured String-Tree Correspondence (SSTC) to annotate the examples in our BKB.

Furthermore, the SSTC structure can easily be extended to keep multiple levels of linguistic information, if they are considered important to enhance the performance of the machine translation system. For instance, in our case here, each node representing a word in the annotated tree structure is tagged with part of speech (POS).

In this section, we shall first introduce the concept of SSTC. It followed by the description of a bitext synchronous parsing technique used to generate both the English and Malay SSTCs for a given aligned translation pair. Finally, we show how the resultant pair of English and Malay SSTCs can be edited semi-automatically to obtain a synchronous SSTC which is the basic element of BKB.

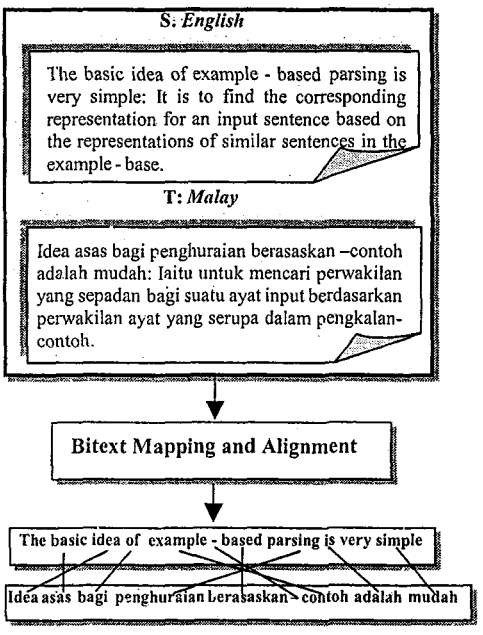


Figure 1: Example output of the processes of bitext mapping and alignment

3.1 Structured String-Tree Correspondence (SSTC)

The SSTC is a general structure that can associate an arbitrary tree structure to string in a language as desired by the annotator to be the interpretation structure of the string, and more importantly is the facility to specify the correspondence between the string and the associated tree which can be non-projective [3]. These features are very much desired in the design of an annotation scheme, in particular for the treatment of linguistic phenomena, which are non-standard, e.g. crossed dependencies [15].

In the SSTC, the correspondence between the sentence on one hand, and its representation tree on the other hand, is defined in terms of finer sub-correspondences between substrings of the sentence and subtrees of the tree. Such correspondence is made of two interrelated correspondences, one between nodes and substrings, and the other between subtrees and substrings, (the substrings being possibly discontinuous in both cases). It can be treated as an extended chart structure [5], which is capable of handling non-projective correspondences between the string and its representation tree.

The notation used in SSTC to denote a correspondence consists of a pair of intervals X/Y attached to each node in the tree, where X(SNODE) denotes the interval containing the substring that corresponds to the node, and Y(STREE) denotes the interval containing the substring that corresponds to the subtree having the node as root [3].

Figure 2 illustrates the sentence **"John picks the ball up"** with its corresponding SSTC. It contains a non-projective correspondence. An interval is assigned to each word in the sentence, i.e. (0-1) for **"John"**, (1-2) for **"picks"**, (2-3) for **"the"**, (3-4) for **"ball"** and (4-5) for **"up"**. A substring in the sentence that corresponds to a node in the representation tree is denoted by assigning the interval of the substring to SNODE of the node, e.g. the node **"picks up"** with SNODE intervals (1-2+4-5) corresponds to the words **"picks"** and **"up"** in the string with the similar intervals, the node **"ball"** with SNODE interval (3-4) corresponds to the word **"ball"** in the string with the similar interval. The correspondence between subtrees and substrings are denoted by the interval assigned to the STREE of each node, e.g. the subtree rooted at node **"picks up"** with STREE interval (0-5) corresponds to the whole sentence **"John picks the ball up"**, the subtree rooted at node **"ball"** with STREE interval (2-4) corresponds to the phrase **"the ball"** in the string.

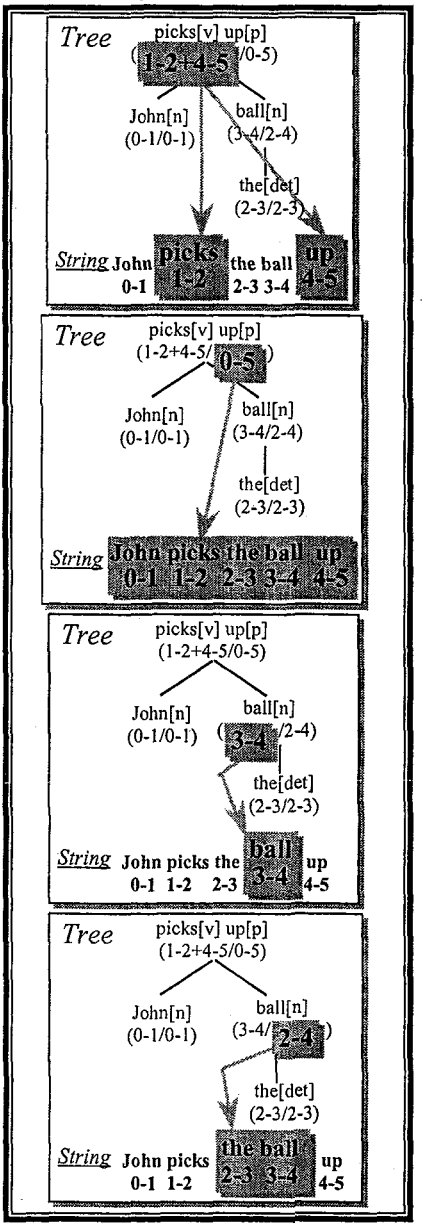
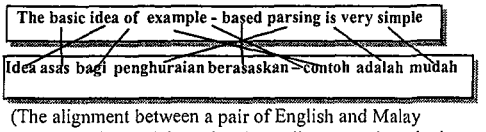


Figure 2: An SSTC recording the sentence **"John picks the ball up"** and its dependency tree together with the correspondences between substrings of the sentence and subtrees of the tree.

3.2 Bitext Synchronous Parsing Technique

Here we describe how to construct the SSTC for the Malay sentence by mean of a bitext synchronous parsing technique. The basic idea is to automatically generate the SSTC for the English sentence through the use of existing English parser. As no parser is currently available for Malay, we propose a synchronous parsing technique to parse the Malay sentence based on the English sentence parse tree together with the alignment result obtained from the bitext alignment algorithm as described earlier. The merit of this proposed technique is to use the output of the parser in one language (e.g. English) which can achieve a good result to parse another language (e.g. Malay).

The following steps describe the bitext synchronous parsing process:



(The alignment between a pair of English and Malay sentences obtained from the bitext alignment algorithm)

- ***English sentence parsing***: After the text is being aligned at different levels (i.e. sentence, phrase, word), each English sentence is passed to a parser. Any available English parser may be used to parse the English sentence. In our case, we choose the Apple Pie Parser (APP) [11] according to the availability. The parsing result of APP is a partial phrase structure tree with simple noun phrases being treated as a single node in the parse tree. The parse tree of the example English sentence is as given below.

```
(S (NP (NPL The basic idea) (PP of (NPL example - based parsing))) (VP is (ADJP very simple)))
```

- ***English sentence SSTC construction:*** In order to obtain the English sentence SSTC structure, we need to compute the string-tree correspondences [13] between the sentence and the parse tree as represented by the SSTC structure illustrated in Figure 3 below.

- ***Lexical transfer:*** In this process, a duplicate copy of the English SSTC created above is generated to be the basic structure for Malay SSTC. First, the English sentence is replaced by the Malay sentence. It followed by the replacement of all English word in the SSTC structure by its corresponding Malay word obtained from the alignment step. In the case of a node containing more than one word, the words will be rearranged according to their order in the Malay sentence. Note that the node represented by an English word which has no Malay equivalent will be deleted. Similarly, English word in the node representing a phrase which has no Malay equivalent will also be deleted. Figure 4 illustrates the SSTC structure for the *Malay* sentence after lexical transfer.

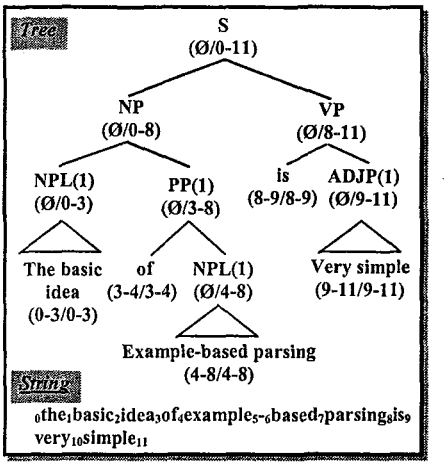


Figure 3: An SSTC for the *English* sentence **"the basic idea of example-based parsing is very simple"**.

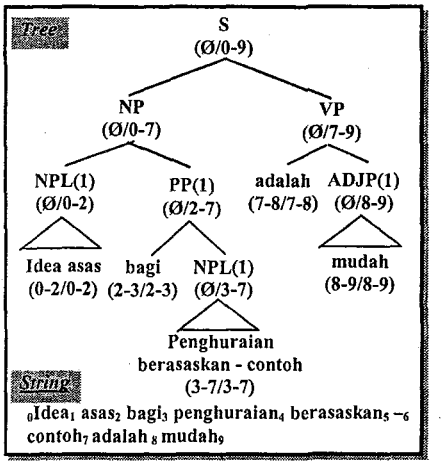


Figure 4: An SSTC construction for the Malay sentence **"idea asas bagi penghuraian berasaskan-contoh adalah mudah"** after lexical transfer.

3.3 Synchronization of SSTC

In this process, the resultant pair of English and Malay SSTCs will be edited semi-automatically to obtain a synchronous SSTC which is the basic element of BKB. Based on the notations used in the SSTC, the translation units between the English and the Malay SSTCs can be constructed in terms of STREE pairs (for phrases) and SNODE pairs (for words) [14]. For instance, as illustrated by the synchronous SSTC given in Figure 5, the fact that **"very simple"** is translated to **"mudah"** is expressed by (9-11,8-9) under the index SNODE of the translation units. Whereas, the fact that **"is very simple"** is translated to **"adalah mudah"** is expressed by (8-11,7-9) under the index STREE of the translation units. Note that this approach is quite similar to the synchronous Tree-Adjoining Grammar presented in [12]. The main difference between our approach and the synchronous TAG is the flexibility provided by the SSTC in the treatment of some linguistic phenomena, which are non-standard [15]. This flexibility provided by the SSTC is very much desired in establishing translation units between source and target substrings, which is possibly discontinuous in both cases. In case the representation of synchronous SSTCs generated need further editing, a synchronous SSTC editor as illustrated in Figure 6 can be used to perform the necessary amendment. Figure 7 gives an overall picture of the processes involved in the construction of a BKB from a given bitext.

4. Conclusion

In this paper, we described an approach to construct a Bilingual Knowledge bank (BKB) from a given bilingual corpora. We introduced a flexible annotation schema called synchronous Structured String-Tree Correspondence (SSTC), which has been used to annotate translation examples in the BKB. The flexibility in the mapping from English to Malay sentences, using synchronous SSTC, makes possible to state direct correspondences without a mediating interlingual representation. By doing this, we are able to match linguistic units at different inter levels of the structure (i.e. define the correspondence between substrings in the sentence, nodes in the tree, subtrees in the tree and sub-correspondences in the SSTC). We also have proposed a synchronous parsing technique to parse the Malay sentence based on the English sentence parse tree together with the alignment result obtained from the bitext alignment algorithm. A graphic editor for the synchronous SSTC (complete with syntax verification) has been implemented. Finally the constructed BKB (see Figure 7) can be used as an example-base for the EBMT [2], besides we can also derive an example-base parsing for Malay which is very much needed for Malay language processing [1].

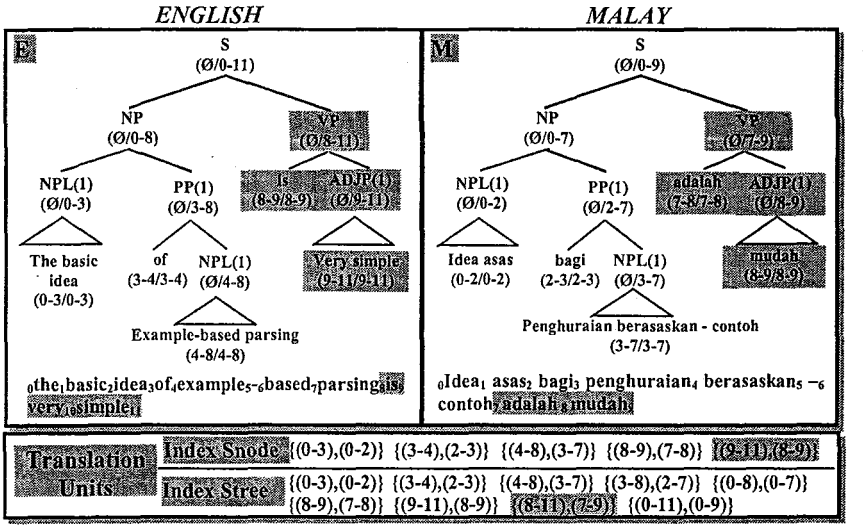


Figure 5: Example synchronous SSTC for the *English source* sentence **"the basic idea of example-based parsing is very simple"** and the Malay *target* sentence **"idea asas bagi penghuraian berasaskan-contoh adalah mudah"** together with their translation units.

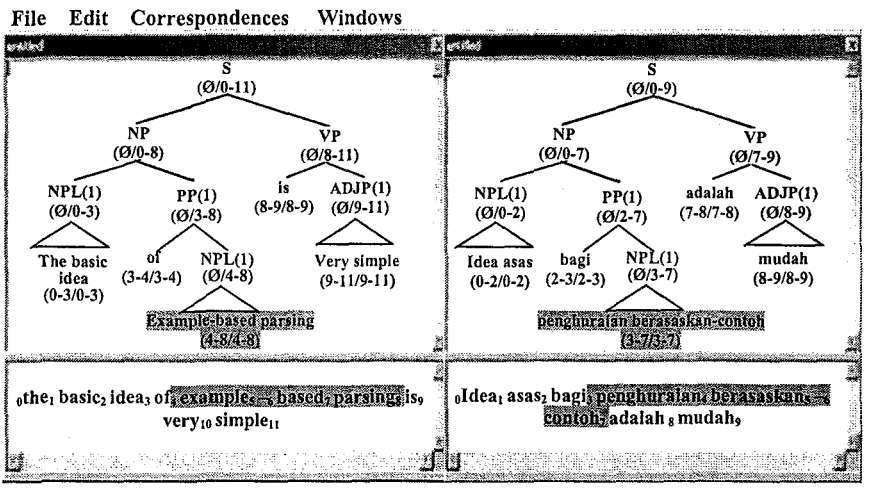


Figure 6: The synchronous SSTC editor.

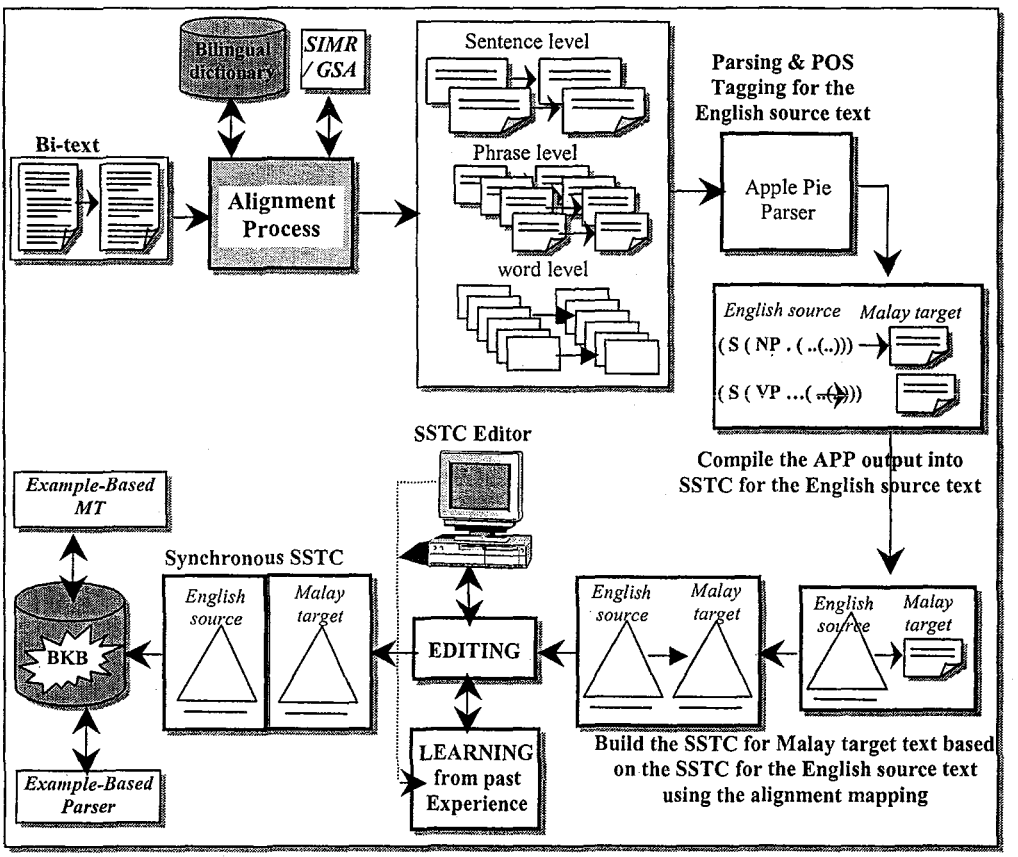


Figure 7: The construction of the BKB based on the synchronous SSTC.

Acknowledgment
We would like to thank I. Dan Melamed for providing us with the SIMR/GSA tools and for his help in porting them to English-Malay language.

References

[1] Al-Adhaileh, M. H. and Tang, E. K. 1998. *A Flexible Example-Based Parser Based on the SSTC.* In Proceedings. of COLING-ACL'98, Vol. I, Montreal, Canada.

[2] Al-Adhaileh, M.H. and Tang, E.K. 1999. *Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema.* In Proceedings of MTS-VII (Machine Translation SUMMIT VII). Singapore.

[3] Boitet, C. and Zaharin, Y. 1988. *Representation trees and string-tree correspondences.* In Proceedings of COLING-88, Budapest. Hungary.

[4] Church, K. 1993. *Char_align: a program for aligning parallel texts at the character level.* In Proceedings of ACL93, Ohio.

[5] Kay, M. 1973. *The MIND system.* In R. Rustin (ed), Natural Language Processing. New York: Algorithmics Press.

[6] Kay, M. 1980. *Algorithm schemata and data structures in syntactic processing.* CSL-80-12, Xerox Corporation. Reprinted in RNLP.

[7] Melamed, I. D. 1997. *A portable algorithm for mapping bitext correspondence.* In Proceedings of ACL35/EACL8.

[8] Melamed, I. D. 1999. *Bitext Maps and Alignment via Pattern Recognition.* Computational Linguistic, 25(1).

[9] Sadler, V. and Vendelmans, R. 1990. *Pilot implementation of a bilingual knowledge bank.* In Proceedings of COLING-90, 3, Helsinki, Fenland.

[10] Sato, S. 1991. *Example-Based Machine Translation.* Ph.D. thesis, Kyoto University, Japan.

[11] Sekine, S. 1996. *Apple Pie Parser.* http:// cs.nyu.edu/ cs/ projects/ proteus/ app/.

[12] Shieber, S.M. and Schabes, Y. 1990. *Synchronous Tree-Adjoining Grammars.* In Proceedings of COLING-90, 3, Helsinki, Fenland.

[13] Tang E. K. 1994, *Natural Language Analysis In Machine Translation (MT) Based On The String-Tree Correspondence Grammar (STCG),* Dissertation submitted in fulfillment of the Ph.D., Universiti Sains Malaysia, Penang, Malaysia.

[14] Tang, E. K.1996. *Interactive Disambiguation in Multilevel Parallel Texts Alignment towards the construction of a Bilingual Knowledge Bank.* In Proceedings of MIDDIM-96, Post-COLING seminar on Interactive Disambiguation, Ch. Boitet (ed), pp. 101-106.

[15] Tang, E. K. and Zaharin, Y. 1995. *Handling Crossed Dependencies with the STCG.* In Proceedings of NLPRS'95, Seoul, Korea.