Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema

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

Abstract

In this paper, we describe an Example-Based Machine Translation (EBMT) system for English-Malay translation. Our approach is an example-based approach which relies sorely on example translations kept in a Bilingual Knowledge Bank (BKB). In our approach, a flexible annotation schema called Structured String-Tree Correspondence (SSTC) is used to annotate both the source and target sentences of a translation pair. Each SSTC describes a sentence, a representation tree as well as the correspondences between substrings in the sentence and subtrees in the representation tree. With both the source and target SSTCs established, a translation example in the BKB can then be represented effectively in terms of a pair of synchronous SSTCs. In the process of translation, we first try to build the representation tree for the source sentence (English) based on the example-based parsing algorithm as presented in [1]. By referring to the resultant source parse tree, we then proceed to synthesis the target sentence (Malay) based on the target SSTCs as pointed to by the synchronous SSTCs which encode the relationship between source and target SSTCs.

Keywords: Structured String-Tree Correspondence (SSTC), Example-based Machine Translation (EBMT), Bilingual Knowledge Base (BKB).

1 Introduction

Here, we design an approach for Machine Translation (MT) which depends on related translated examples kept in a Bilingual Knowledge Bank (BKB). This approach is called example-based MT; the linguistic knowledge extracted directly from the example-base will be used to analyze and translate a *source* sentence to the corresponding *target* sentence. Ideally if the sentence is already in the example-base, the translation is found there too, but in most cases, the *source* sentence will not be found in the example-base. In such case, a method is used to retrieve close related examples and use the knowledge from these examples to construct the translation for the *source* sentence. In general, this approach relies on the assumption that if two *source* sentences are “close”, their translations should be “close” too; if the translation of the first one is known, the translation of the other can be obtained by making some modifications in the translation of the first one [4].

The example-based approach has become a common technique for NLP applications, especially in MT as reported in [6], [9] and [7]. However, a main problem normally arises in the current approaches which indirectly limits their applications in the development of a large scale and practical example-based MT system, i.e. the lack of flexibility in representing translation relations between *source* and *target* substrings where the substrings being possibly discontinuous in both cases. In this paper, we propose to overcome the problem by introducing a flexible annotation schema called synchronous Structured String-Tree Correspondence (SSTC) which will be used to annotate translation examples in the Bilingual Knowledge Bank (BKB). We will also present a strategy to translate an English *source* sentence to a Malay *target* sentence based on the synchronous SSTC annotation schema.

2 Structured String-Tree Correspondence (SSTC)

The SSTC is a general structure that can associate, to string in a language, arbitrary tree structure 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 [2]. These features are very much desired in the design of an annotation scheme, in particular for the treatment of linguistic phenomena, which are not-standard, e.g. crossed dependencies [12]. We shall investigate the properties of the SSTC annotation schema here and discuss on its usage toward the construction of a BKB in the next section.

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).

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 [2].

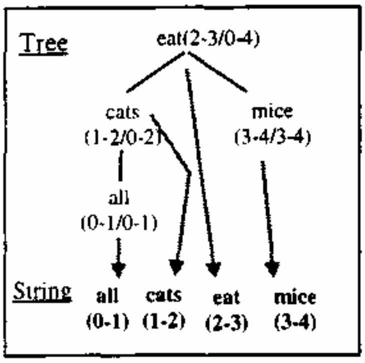


Figure 1: An SSTC recording the sentence **"all cats eat mice"** and its Dependency tree together with the correspondences between substrings of the sentence and subtrees of the tree.

Figure 1 illustrates the sentence **"all cats eat mice"** with its corresponding SSTC. It is a simple projective correspondence. An interval is assigned to each word in the sentence, i.e. (0-1) for **"all"**, (1-2) for **"cats"**, (2-3) for "eat" and (3-4) for **"mice"**. 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 **"cats"** with SNODE interval (1-2) corresponds to the word **"cats"** 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 "eat" with STREE interval (0-4) corresponds to the whole sentence **"all cats eat mice"**.

3 Constructing a Bilingual Knowledge Bank based on the synchronous SSTC

In Example-Based Machine Translation system [8], the use of Bilingual Knowledge Bank (BKB) containing bilingual parallel texts which encode the 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 [7], 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. The dependency structure has been chosen as the linguistic representation of the SSTC as it gives a natural way to establish the translation units between the *source* (English) and *target* (Malay) SSTCs; similar arguments also appear in [7], [13] and [5]. However, 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 in the annotated dependency tree is tagged with part of speech (POS) and a sense number as coded in a given dictionary.

In our approach, translation examples are established by mean of a synchronous SSTC editor as illustrated in Figure 2.

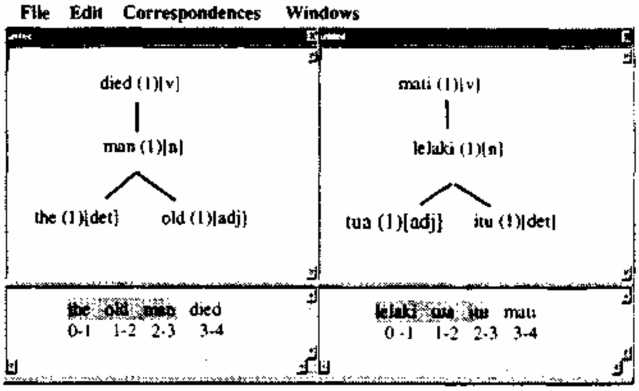


Figure 2: The synchronous SSTC editor.

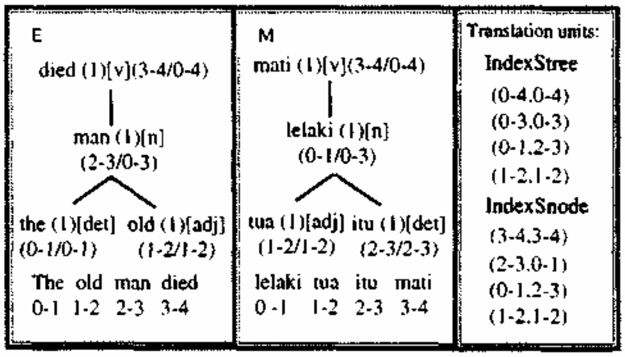


Figure 3: Example SSTCs for English *source* sentence **"the old man died"** and the Malay *target* sentence **" lelaki tua itu mati"** together with their translation units.

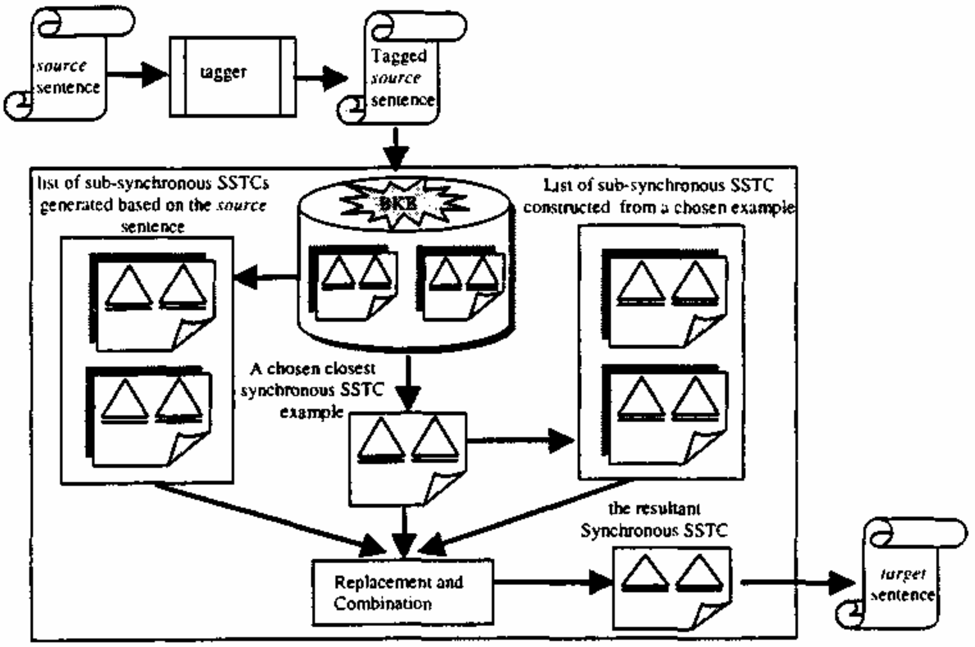

source sentence
tagger
Tagged source sentence
list of sub-synchronous SSTCs generated based on the source sentence
BKB
List of sub-synchronous SSTC constructed from a chosen example
A chosen closest synchronous SSTC example
the resultant Synchronous SSTC
Replacement and Combination
target sentence

Figure 4: Example-Based Machine Translation Based on the synchronous SSTC.

Based on the notations used in the SSTC, we denote the translation units between the *source* (English) and the *target* (Malay) SSTCs in terms of STREE pairs (for phrases) and SNODE pairs (for words) [11]. For instance, as illustrated by the synchronous SSTC given in Figure 3, the fact that **"died"** is translated to **"mati"** is expressed by (3-4, 3-4) under the index SNODE of the translation units. Whereas, the fact that **"the old man"** is translated to **"lelaki tua itu"** is expressed by (0-3.0-3) under the index STREE of the translation units. Note that this approach is quite similar to the synchronous Tree-Adjoining Grammar presented in [10]. 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 not-standard [12]. This flexibility is very much desired in establishing translation units between source and target substrings which being possibly discontinuous in both cases.

4 Example-Based Machine Translation Based On the SSTC

In this section, we shall informally present the general schema of the EBMT, highlighting the various components involved, and give an example to illustrate the process of translating a *source* (English) sentence to the corresponding *target* (Malay) sentence.

As illustrated in Figure 4, the process of translation begins with the use of a sense tagger [3] to tag each word in the *source* sentence with a sense number together with its POS. The tagged *source* sentence will then be parsed to establish a single rooted representation tree based on the example-based parser presented in [1]. The parser first constructs the sub-SSTCs for all phrases IT the *source* sentence by referring to some close related examples in the BKB (i.e. examples that contain some words tagged with similar sense number as in the tagged *source* sentence).

To each sub-SSTC constructed for the *source* sentence, the corresponding *target* sentence sub-SSTC can be determined based on the translation units as established by the synchronous SSTCs in the BKB. The *source* and *target* sub-SSTCs generated together with the corresponding translation units identified will form a list of sub-synchronous SSTCs. Next, we will proceed to determine an example synchronous SSTC in the BKB which can be used as a reference to combine the generated list of sub-synchronous SSTCs in order to form a complete synchronous SSTC containing both SSTCs for the *source* and *target* sentences. A list of sub-synchronous SSTCs is constructed from the chosen example, which will be replaced by their corresponding sub-synchronous SSTCs generated from the *source* sentence to form the final synchronous SSTC through a combination process. Finally, the *target* sentence as appeared in the *target* SSTC will be retrieved and outputs as the translation of the *source* sentence.

In the following, we shall present an example to illustrate the process of translation as described above Suppose the system intends to find the translation for the English *source* sentence ***"the old man picks the green lamp up"***, based on the set of examples representing the example-base (BKB) in Figure 5.

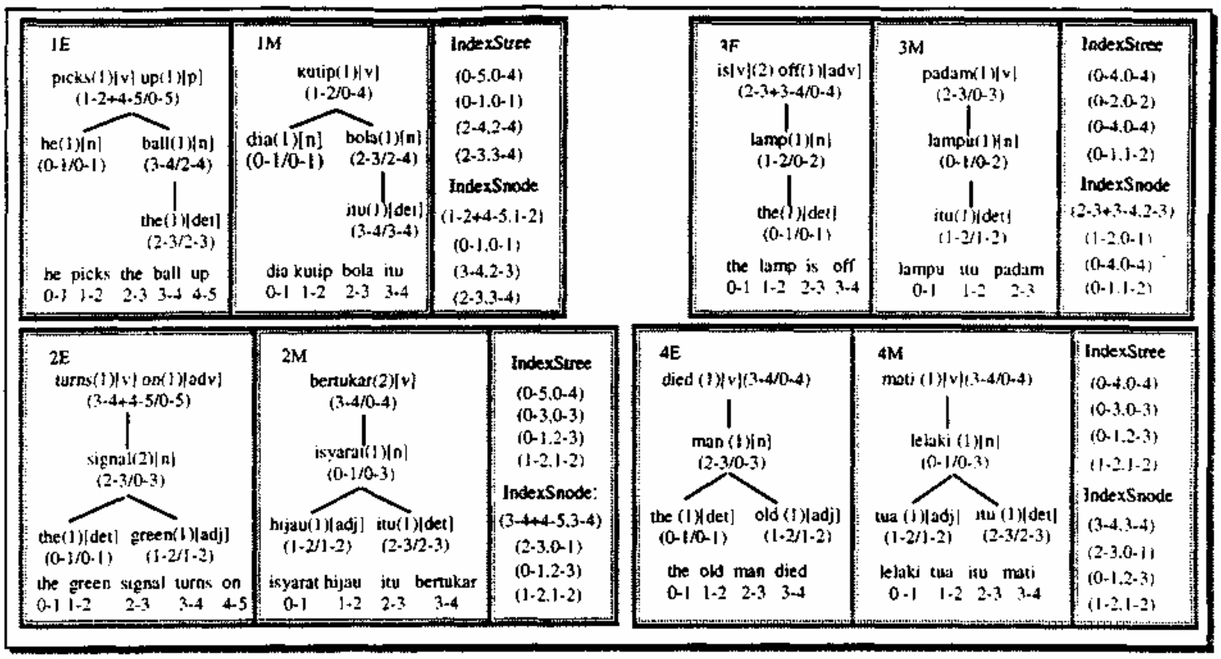
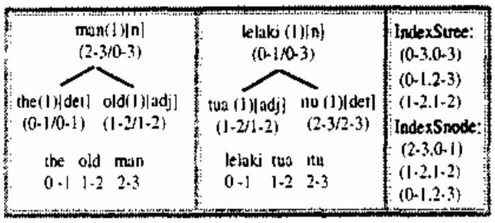


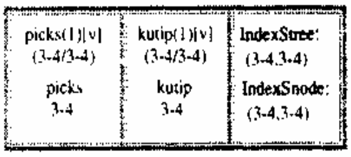
Figure 5: A set of synchronous SSTCs contains the English sentences, the Malay translation sentences and their translation units, representing the BKB.

From the *source* sentence, the following list of sub-synchronous SSTCs are generated based on the set of related examples found in the BKB:

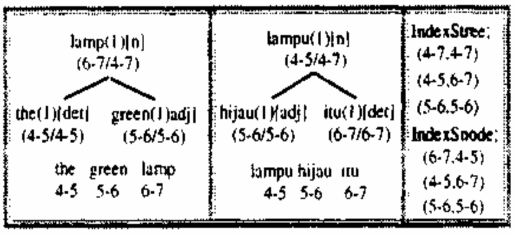
(1)



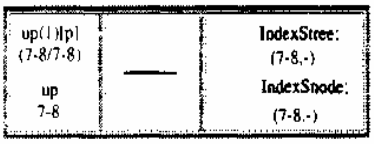
(2)



(3)



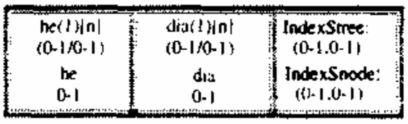
(4)



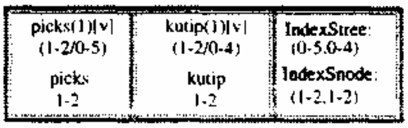
In order to combine the sub-synchronous SSTCs to form a complete synchronous SSTC, the system first finds examples in the BKB which contain a word in the *source* sentence appearing as the root word of the dependency tree in its *source* SSTC. If more than one example are found (in most cases), the system will calculate the distance between the *source* sentence and these examples [1], and the closest example (namely the one with minimum distance) will be chosen as a reference to combine the list of generated sub-synchronous SSTCs to form a complete synchronous SSTC.

Here, the word "pick" is the only word in the *source* sentence, which appears as the root word in the given example BKB, namely in the example (El. Ml). The system will first construct the sub-synchronous SSTCs derived from example (El, Ml):

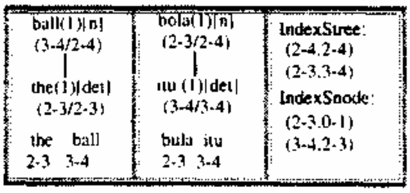
(1)



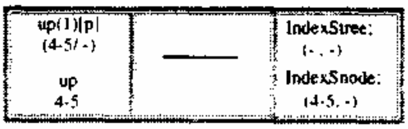
(2)



(3)



(4)



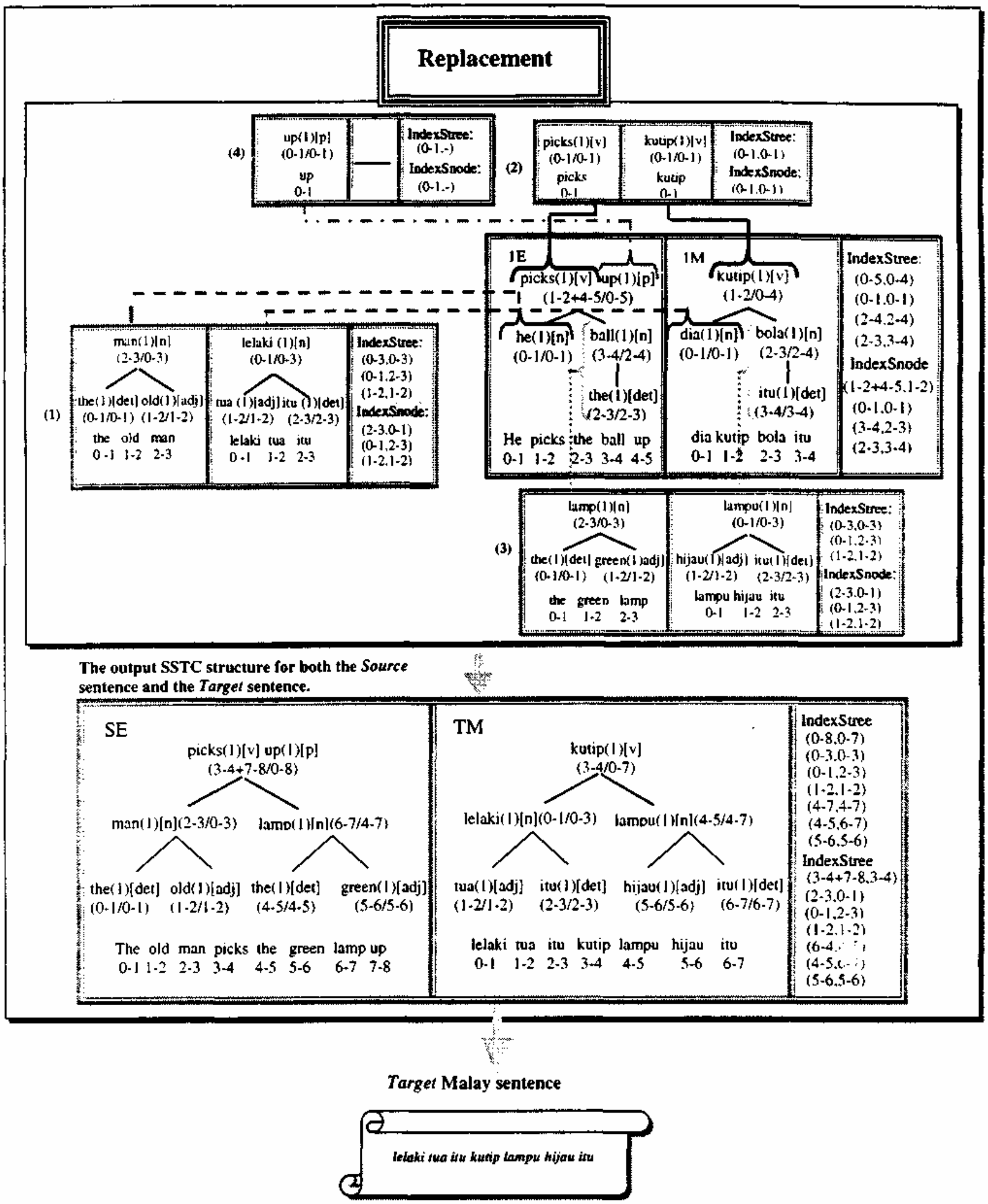


Figure 6: The replacement on the synchronous SSTC example and the translation for the *source* sentence " ***the old man picks the green lamp up*** ".

Next, the sub-synchronous SSTCs derived from example (El, Ml) will be replaced by the corresponding sub-synchronous SSTCs generated from the *source* sentence to form a complete synchronous SSTC. The replacement process is done by traversing in parallel both the *source* and *target* SSTC trees from example (E1, M1) in the manner of preorder traversal, and replace each sub-synchronous SSTC found during the traversal with the corresponding sub-synchronous SSTC generated from the *source* sentence, as illustrated in Figure 6. This approach is analogous to top down parsing technique. Note that before making the replacement, the system must first check that the root word, of all for sub-synchronous SSTCs in both the example and the *source* sentence have the same POS, and that these sub-synchronous SSTCs are occurred in the same order which can be determined easily from their STREE intervals.

Finally, the *target* sentence as appeared in the *target* SSTC of the resultant synchronous SSTC will be output as the translation of the *source* sentence, i.e. " *Lelaki tua itu kutip lampu hijau itu"* as illustrated in Figure 6.

5 A Note on Implementation

In this paper, we have proposed an Example-Based Machine Translation (EBMT) system for English-Malay translation based on the synchronous SSTC annotation schema. A graphic editor for the synchronous SSTC (complete with syntax verification) has been implemented as well as various indexing routines to generate indexes on the data captured in the BKB. A SSTC parser has also been implemented and tested on data covering a wide range of linguistic phenomena [1]. This parser will, in future, be extended further to implement the synchronous SSTC interpreter which being the core of our EBMT proposed here.

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,* the 36th Annual Meeting of the Association for Computational Linguistic and the 17th International Conference on Computational Linguistic, Vol. I, Montreal.

[2] Boitet, C. and Zaharin, Y. (1988). "Representation trees and string-tree correspondences", In Proceedings of *COLING-88,* Budapest.

[3] Guo, C. M. (1995)."Machine Tractable Dictionaries: Design and Construction". Ablex: Norwood, NJ.

[4] Lepage, Y. and Shin-ichi, A. (1996). "Saussurian analogy: a theoretical account and its application". In Proceeding of *COL1NG-96,* 2, Copenhagen.

[5] Maxwell, D. and Schubert K. eds. (1989). "Metataxis in practice: Dependency Syntax for Multilingual machine Translation". Dordrecht/Providence: Foris. DLT 6.

[6] Nagao, M. (1984)."A Framework of a mechanical translation between Japanese and English by analogy principle". In A. Elithorn, R. Banerji, (Eds.), Artificial and Human Intelligence, Elsevier: Amsterdam.

[7] Sadler, V. and Vendelmans, R. (1990). "Pilot implementation of a bilingual knowledge bank". In Proceedings of *Coling-90,* 3, Helsinki.

[8] Sato, S. (1991). "Example-Based Translation Approach to Machine Translation". In Proceedings of International Workshop on Fundamental Research for Future Generation Natural Language Processing, ATR interpreting Telephony Research Laboratories.

[9] Sato, S. and Nagao, M. (1993). "Example-based Translation of technical Terms". In Proceedings of *TMI-93,* Kyoto.

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

[11] 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.

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

[13]. Tang, E. K. and Zaharin, Y. (1996). "Learning to Translate Based on the SSTC Annotation Schema". UTMK Document, School of Computer Sciences, USM, Malaysia, 1996.