Learning-to-Translate Based on the S-SSTC Annotation Schema

Tang Enya Kong
Enyakong1@gmail.com

Zaharin Yusoff
University Sarawak Malaysia,
Sarawak, Malaysia
zarinby@gmail.com

Christian Boitet
Université Joseph Fourier,
Grenoble, France
Christian.Boitet@imag.fr

Abstract

We present the S-SSTC framework for machine translation (MT), introduced in 2002 and developed since as a set of working MT systems (SiSTeC-ebmt). Our approach is example-based, but differs from other EBMT approaches in that it uses alignments of string-tree alignments, and in that supervised learning is an integral part of the approach. Our model directly deals with three main difficulties in the traditional treatment of MT that stem from its separation from the "translation task" (the 'world'). First, by allowing the system to learn from real translation examples directly, we avoid the need to indefinitely pursue the elusive goal of writing grammars to exactly describe intermediate syntactico-semantic monolingual representations and their correspondences. Second, we make explicit the dependence of the MT system performance on the input from the environment. That is possible only because the learning process uses feedback from the real translation knowledge when constructing its knowledge representation. Third, such MT systems using an inductively learned knowledge base yield a desirable non-regressive behavior by using translation mistakes to improve their knowledge base.

1. Introduction

The S-SSTC-based framework for the construction of MT systems has been introduced in 2002 [Mosleh et. al. 2002] and developed since to an operational state (SiSTeC-ebmt for English-Malay and English-Chinese). In this article, we would like to stress a particular aspect, namely that this approach is better capable of modeling the translation knowledge of human translators than other example-based approaches. Because the translation knowledge is represented as alignments (synchronizations) between string-tree alignments (SSTCs, or structured string-tree correspondences), it is more natural to translators (and post-editors) than direct word-word, string-string or chunk-chunk correspondences used in classical SMT and EBMT models. It is also totally static, hence more understandable than procedural knowledge embedded in almost all RBMT approaches.

The learning process which is an integral part of the development of SiSTeC-ebmt MT systems can in fact be viewed as a special case of the study of reasoning reported in [Khardon&Roth 94], because it combines the interfaces to the 'world' used by known learning models with the reasoning task and a performance criterion suitable for it. In such a framework, the intelligent agent is given access to its learning interface, and is also given a grace period in which it can interact with this interface and construct its representation Knowledge Base (KB) of the 'world'. Its reasoning performance is measured only after this period, when it is presented with 'queries' from some query language, relevant to the 'world', and has to answer whether such 'queries' are implied by the learned 'world' model. In our case, the 'world' is the 'translation task' captured in terms of the parallel texts produced by human translators and enriched by their S-SSTCs, and the 'queries' are simply modeled by a predicate `Translate(ST,TT)` where ST is the source language text and TT is a variable to be instantiated by a target language text if the 'translation' model learned is capable of performing such translation.

Our model directly deals with three main difficulties in the traditional treatment of MT which stem from its separation from the "translation task" (the 'world'). First, by allowing the system to learn from real translation examples directly, we avoid the need to indefinitely pursue the elusive goal of writing grammars to exactly describe intermediate syntactico-semantic monolingual representations and their correspondences. Second, we make explicit the dependence of the MT system performance on the input from the environment. This is possible only because the learning process uses feedback from the real translation knowledge when constructing its knowledge representation. Third, such MT systems using an inductively learned knowledge base yield a desirable non-regressive behavior by using translation mistakes to improve their knowledge base.

Learning to translate is just like any other machine learning task; it is concerned with modeling and understanding learning phenomena with respect to the 'world' — a central aspect of cognition. Traditional theories of Machine Translation systems, however, have assumed that such cognition can be studied separately from learning. It is assumed that the knowledge is given to the system, stored in some representation language with a well-defined meaning, and that there is some mechanism which can be used to determine what source language text can be translated with respect to the given knowledge; the question of how this knowledge might be acquired and whether this should influence how the performance of the machine translation system is measured is not considered. We prove the usefulness of the 'learning-to-translate' approach by showing that through interaction with the world, the agent truly gains additional translating power, over what is possible in more traditional settings.

2. The bilingual knowledge bank base as a set of S-SSTCs

Bilingual parallel texts which encode the correspondences between source and target sentences have been used extensively in implementing the so called example-based machine translation systems [Sato 91, Richardson et. al. 2001, Menezes et. al. 2001, Kawahara & Kurohashi 2010]. In order to enhance the quality of example-based systems, sentences of a parallel corpus are normally annotated with their constituent or dependency structures [Sadler&Vendelmans 90], which in turn allows correspondences between source and target sentences to be established at the structural level. Here, we annotate parallel texts based on the Structured String-Tree Correspondence (SSTC) [Boitet&Zaharin 88]. The SSTC is a general structure that can associate, to strings in a language, arbitrary tree structures as desired by the annotator to be the interpretation structures of the strings, and more importantly is the facility to specify the correspondence between the string and the associated tree which can be interpreted for both analysis and synthesis in the machine translation process. These features are very much desired in the design of an annotation scheme, in particular for the treatment of certain non-standard linguistic phenomena, such as unprojectivity or inversion of dominance [Tang&Zaharin 95].

In this paper, we show how to use the good properties of the SSTC annotation scheme for S-SSTC-based MT, using the example of the SiSTeC-ebmt English-Malay Machine Translation system. We have chosen dependency structures as linguistic representations in the SSTCs, since they provide a natural way of annotating both the tree associated to a string as well as the mapping between the two [Goh 96]. We also give a simple means to denote the translation elements between the corresponding source (English) and target (Malay) SSTCs. Similar arguments also appeared in [Sadler&Vendelmans 90] and [Maxwell&Schubert 89]. The dependency structure used here is in fact quite analogous to the use of abstract syntax tree in most of the compiler implementation. However, we note that the SSTCs can easily be extended to keep multiple levels of linguistic representation (e.g. syntagmatic[1], functional and logical structures) if that is considered important to enhance the results of the machine translation system. Naturally, the more information annotated in an SSTC, the more difficult is the annotation work; that is why one should try to keep only the annotations contributing most to the task at hand.

In the general case, let S be a string (usually a sentence) and T a tree (its linguistic representation). Instead of simply write (S,T), we want to decompose that 'large' correspondence into smaller ones (S1, T1)…(Sn, Tn) in a hierarchical fashion; hence the adjective 'structured' in 'SSTC'. If T is an abstract representation of S, some nodes may represent discontinuous words or constituents (e.g. He gives the money back to her), or some words are not directly represented (e.g. auxiliaries, articles), or some words omitted (elided) in S may have been restored in T. [Boitet&Zaharin 88] have shown how to encode such string-tree correspondences in the tree part (T), through 2 functions, SNODE and STREE, even if the trees are 'abstract', but provided they obey some formal constraints that are in effect verified by all known kinds of linguistic trees. In the SSTC diagrams presented here, any tree node N bears a pair X/Y where X = SNODE(N) and Y = STREE(N). X and Y are generalized (not necessarily connex) substrings of the string S, and are written as minimal[2] left-to-right lists of usual intervals, like 1_3+4_5). SNODE(N) denotes the substring that corresponds to the lexical information contained in node[3], while STREE(N) denotes the (again possibly discontinuous) substring that corresponds to the whole subtree rooted at node N.

[1] by constituents.

[2] That means that any occurrence of $n_1\_n_2+n_2\_n_3$ is replaced by $n_1\_n_3$, n_i being a position between two typographical words, or more generally (to handle writing systems without word delimiters such as Chinese, Japanese, Korean, Vietnamese, Thai, Lao, or Khmer), between two characters.

[3] A lexeme for leaves and nothing for internal nodes in syntagmatic structures, and a lexeme for each node in a dependency structure, where a lexeme might be a compound corresponding to a discontinuous substring, such as *give_back, neither_nor, if_then_else,* etc.

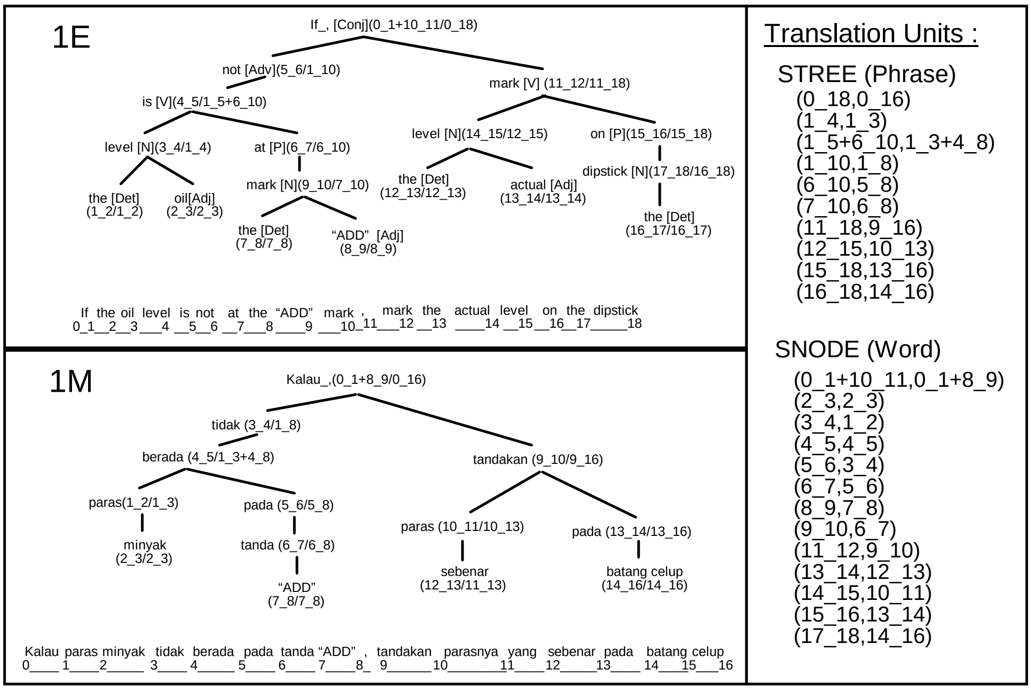

Figure 1: An example pair of English - Malay SSTCs and the corresponding translation elements

As for the correspondences between the source (English) and target (Malay) SSTCs, the translation elements[4] between phrases and words are coded in terms of STREE pairs and SNODE pairs, respectively. To illustrate this, we show in Figure 1 a pair of source (English) and target (Malay) SSTCs and the corresponding translation elements. In the example SSTCs given, an interval is assigned to each word in the sentence, i.e. 0_1 to "if", 1_2 to "the", etc. The node "not" has SNODE = 5_6, meaning that its lexeme corresponds to the word "not" in the sentence. Similarly, the node bearing "is" has STREE = 1_5+6_10, meaning that the subtree it dominates corresponds to the discontinuous substring "the oil level is" + "at the ADD mark".

Figure 2 gives another example of correspondence between source sentence 2E and target sentence 2M. Both translation pairs, (1E, 1M) and (2E, 2M), will serve as running examples of annotated bilingual parallel sentences in the rest of the discussion.

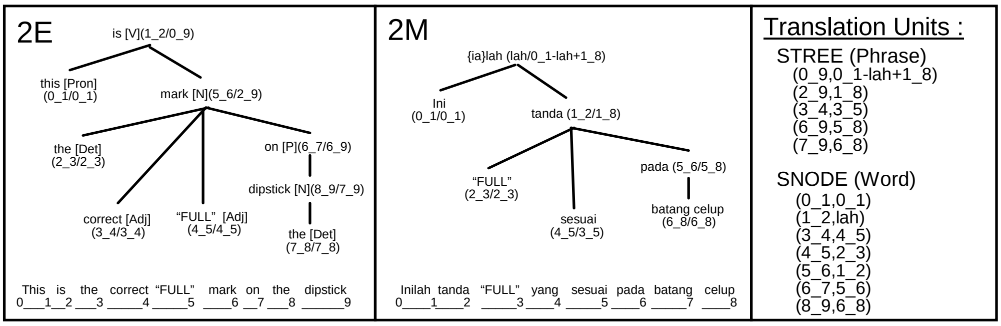

Figure 2: An example annotation between source sentence 2E and target sentence 2M.

3. A learn-to-translate process based on a bilingual knowledge bank (BKB)

The process of *learning-to-translate* begins with the construction of a shared forest structure based on the representation structure of the used source sentences (here 1E and 2E) together with its words index as illustrated in Figure 3 below. The shared forest structure together with its words index is then used to parse a new input source sentence by extracting from the BKB the related substructure of the shared

[4] The term 'translation units' (TUs) has been used in previous publications, but, as the normal sense of TU is 'a minimal unit for human translation', that is, a sentence or a title, an exclamation, etc., we replace it here by 'translation element'.

forest, using the words index as a guide. An example of a shared forest for the new source sentence 3E, constructed based on the shared forest structure of sentences 1E and 2E, is given in Figure 3 below.

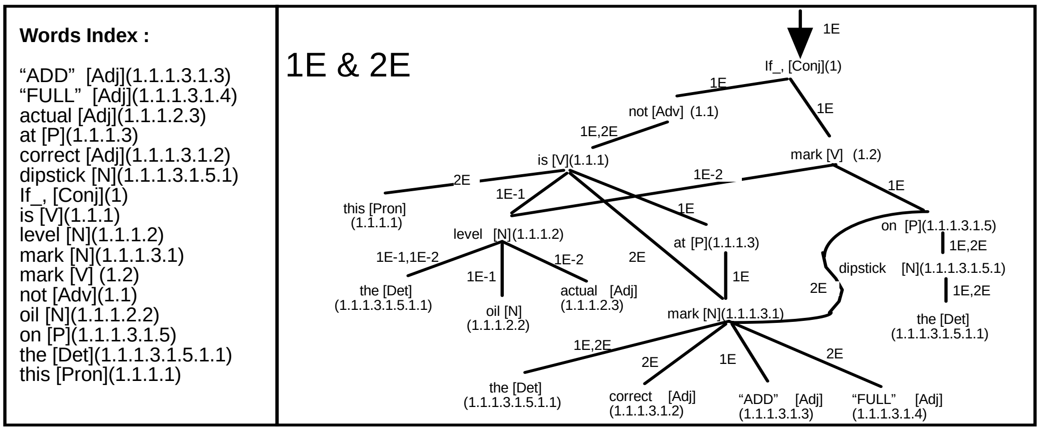

Figure 3: A shared forest structure constructed based on the representation structures of source sentence 1E and 2E with their word index.

The shared forest structure together with its words index is then used to parse a new input source sentence by mean of extracting out the related substructure of the shared forest through the guidance of words index. An example of a shared forest for a new source sentence 3E, constructed based on the shared forest structure of sentences 1E and 2E, is given in Figure 4. The resulting shared forest is then used to construct the corresponding dependency tree of sentence 3E as well as the substring to subtree mappings, as shown on the top part of Figure 5 — the analysis task [Tang 94].

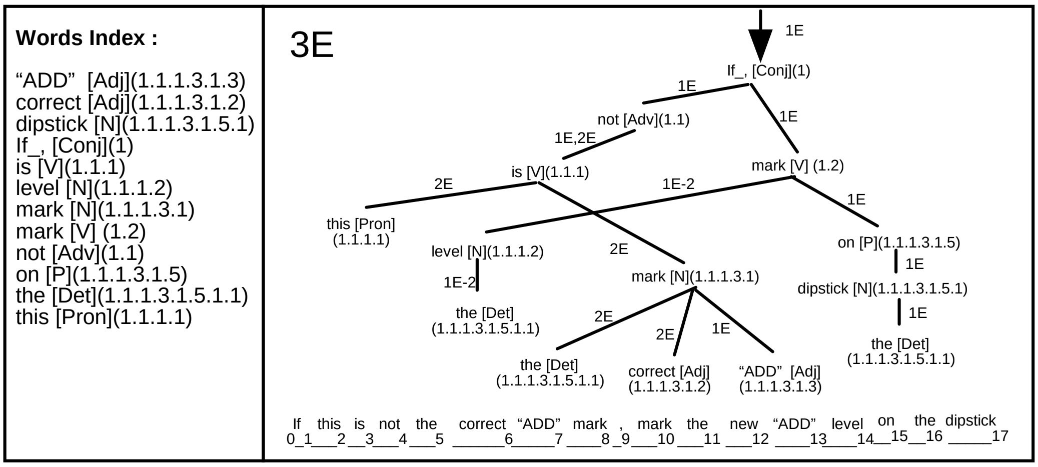

Figure 4: An example shared forest structure for sentence 3E constructed based on structures created in Figure 3.

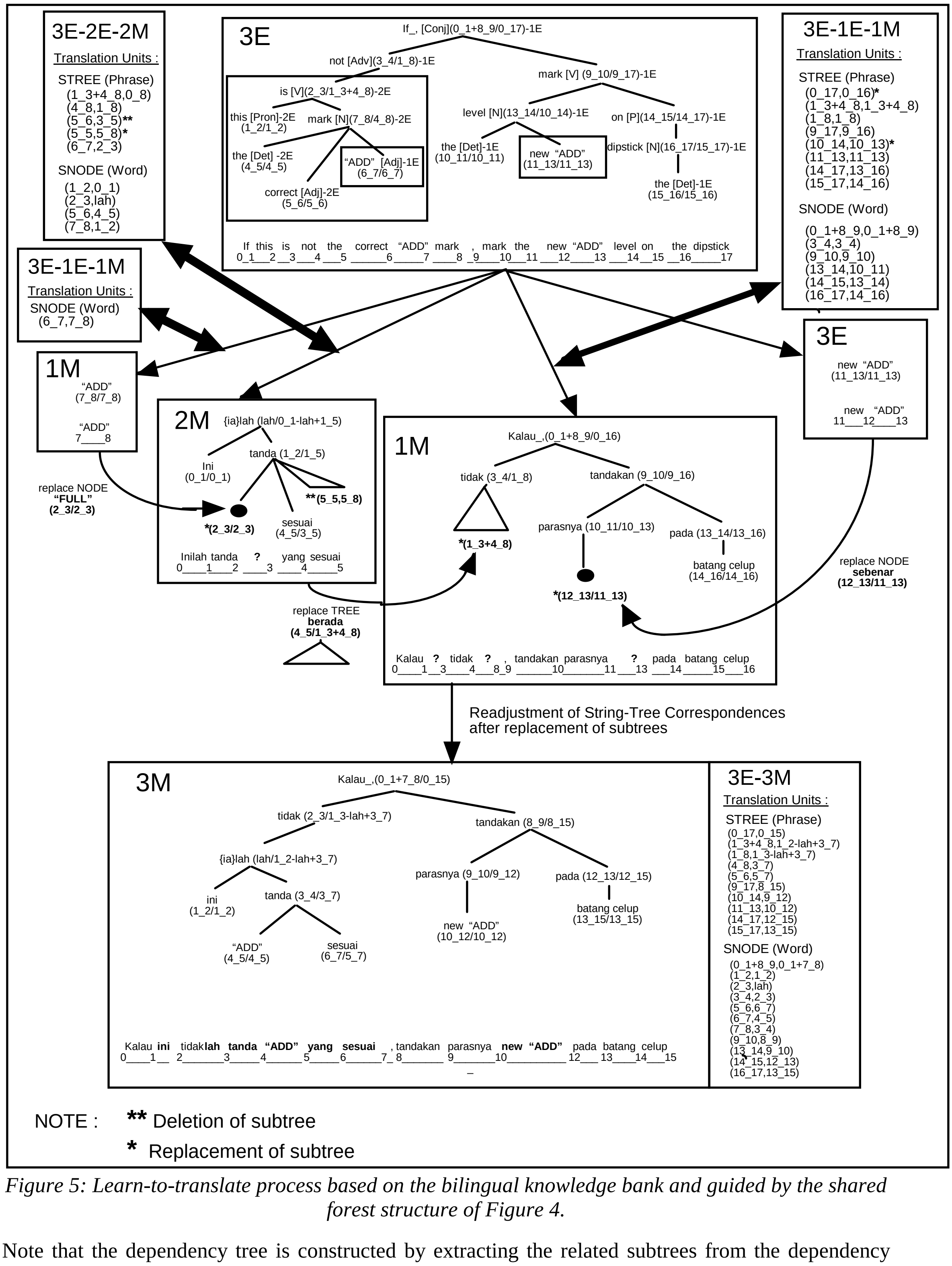

Figure 5: Learn-to-translate process based on the bilingual knowledge bank and guided by the shared forest structure of Figure 4.

Note that the dependency tree is constructed by extracting the related subtrees from the dependency trees of both sentences 1E and 2E. Note also that a substring which has not been treated before, e.g. "new ADD", will be set to correspond to a node in the dependency tree; the location of this node in the dependency tree is decided based on its context (i.e. the surrounding words). The resulting sub-SSTC of sentence 3E is then used to retrieve the related target language sub-SSTC based on the translation elements stored in the bilingual knowledge bank — the transfer task. The target sub-SSTCs are then merged to form a complete SSTC for the translated sentence, as shown at the bottom of Figure 5. Such a merging process can be considered as a kind of synthesis process in order to construct the target sentence [Heng 95].

4. Learning from corrected translation mistakes

In order to improve the performance of an MT system, not only do we need to fix detected errors, we also need to increase at the same time the translation knowledge (encoded in the bilingual knowledge bank). To do that, we feed back to the system the information as to whether the previous translation has been done correctly or not. In the case of error translation, it is also necessary to correct the BKB by making the necessary adjustments to reflect the error correction process (hence the need for an integrated post-editing environment), so that a similar error will not occur again. In the case of perfect translation, we may reinforce scores attached to the used translation elements, or do nothing, treating the result as a simple confirmation. In the given example, the resulting target SSTC appears to have some errors and it is corrected as highlighted in Figure 6. The corrected SSTC is then added to the system in order to enrich the bilingual knowledge bank. Here, we add a shared forest structure constructed from the representation structures of source sentences 1E, 2E and 3E, as shown in Figure 7.

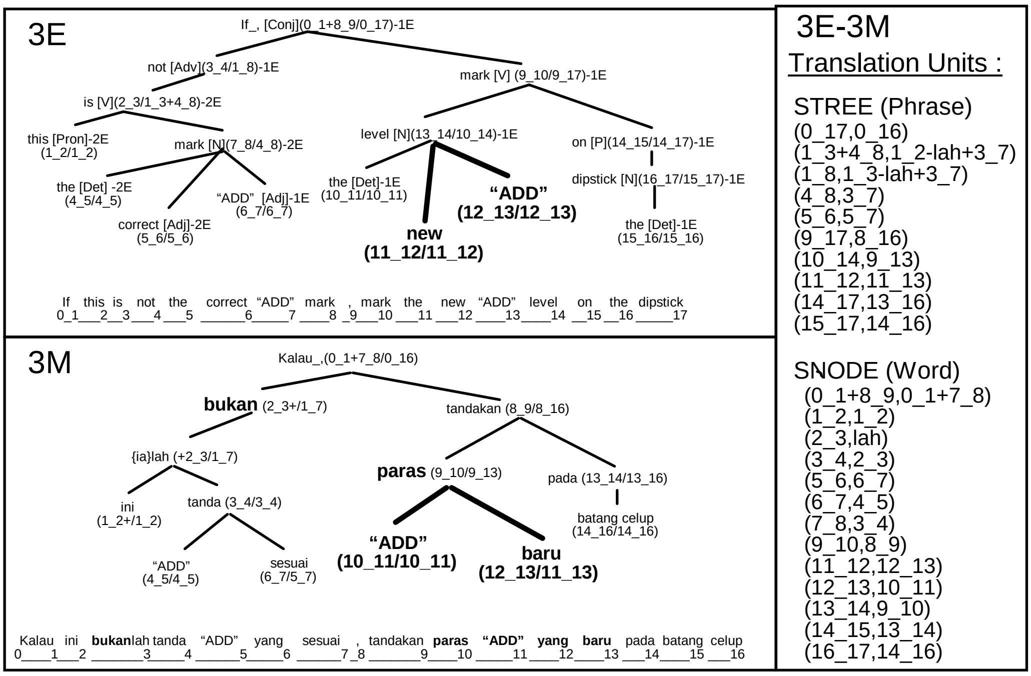

Figure 6: An example annotation between source sentence 3E and target sentence 3M after going through the correction done by the linguist on the improper annotations produced by the MT system.

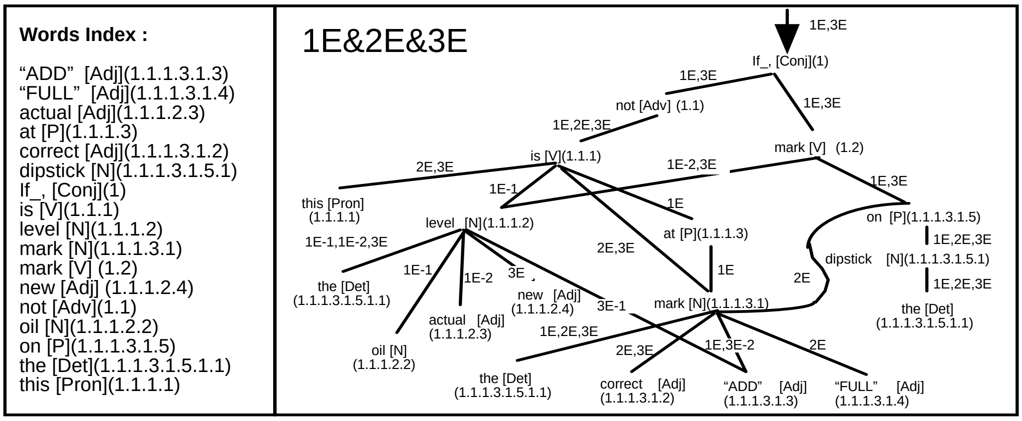

Figure 7: A shared forest structure constructed based on the representation structures of source sentence 1E, 2E and 3E with its words Index

5. Implementation notes

The main purpose of the project described in this paper is to build a general software package that provides an integrated environment for the construction of S-SSTC-based EBMT systems. In this project, we put emphasis on the development of an English->Malay MT system in the domain of computer science texts. However, the same methodology can be adapted to develop MT systems for any other typology of texts, and naturally also for any other language pairs. The current SiSTeC-ebmt platform consists of four major subcomponents (as shown by the diagram given in Figure 8), namely (1) the preparation of an annotated bilingual parallel texts to be used for the initial learning process, (2) a set of acquisition tools used to construct the initial bilingual knowledge bank, (3) a general MT system to translate new input sentences (using the bilingual knowledge bank) into the target language, together with all the related annotation, (4) the post-editing process to make corrections (if any) on the translation as well as on the annotations, which in turn will be used by the learning tools to confirm the well translated parts and adjust the translation elements of the BKB corresponding to the corrected parts.

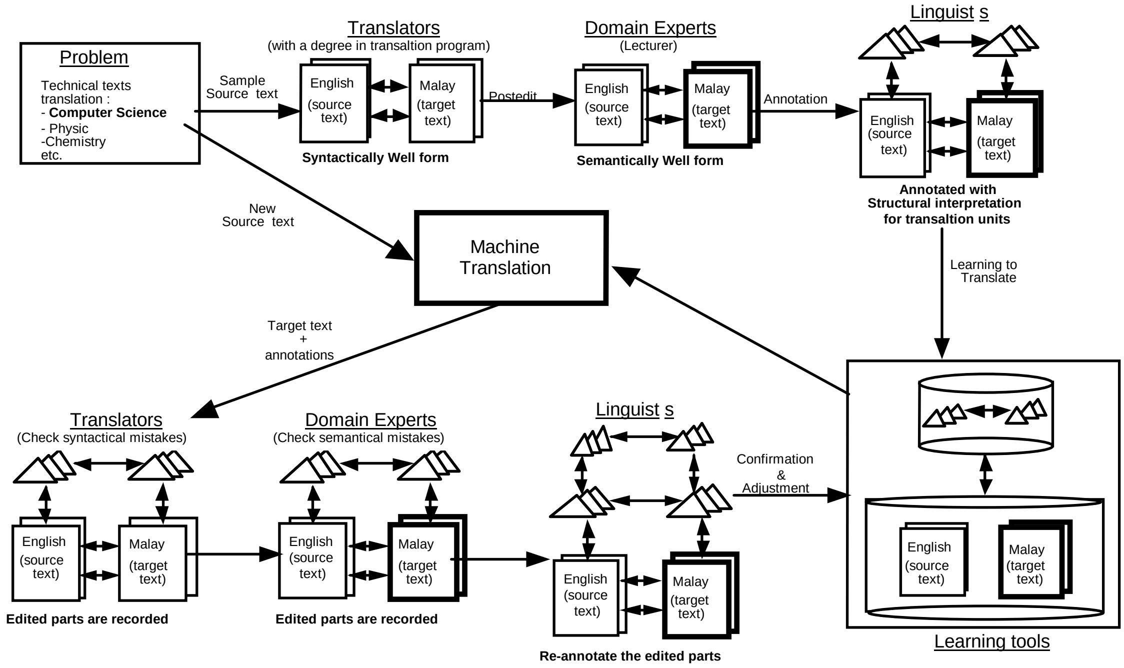

Figure 8: An overview of the implementation for the "learn-to-translate" S-SSTC-based model

An English->Malay MT system with 100,000 translation examples annotated in the S-SSTC has been constructed based on the implementation frame as described above. To provide an overview of the performance of this system, a quick comparison of the MT results produced by Google Translate and our SiSTeC-ebmt is given in the table below.

Sample English Text	**Translation to Malay by Google Translate**	**Translation to Malay by SiSTeC-ebmt (100,000 S-SSTCs)**
The main purpose of the project described in this paper is to build a general software package that provides an integrated environment for the construction of S-SSTC based EBMT systems. In this project, we put emphasis on the development of an English->Malay MT system in the domain of computer science texts. However, the same methodology can be adapted to develop MT systems for any other typology of texts, and naturally also for any other language pairs.	Tujuan utama projek yang dihuraikan dalam kertas kerja ini adalah untuk membina satu pakej perisian umum yang menyediakan persekitaran bersepadu bagi pembinaan sistem S-SSTC EBMT berasaskan. Dalam projek ini, kami meletakkan penekanan kepada pembangunan bahasa Inggeris> MT sistem bahasa Melayu dalam domain teks sains komputer. Walau bagaimanapun, kaedah yang sama boleh disesuaikan untuk membangunkan sistem MT bagi mana-mana tipologi teks lain, dan secara semulajadi juga untuk mana-mana pasangan bahasa lain.	Tujuan utama daripada projek itu digambarkan di dalam kertas ini untuk membina perisian umumnya pakej yang menyediakan mengintegrasikan S-SSTC persekitaran bagi pembinaan berdasarkan EBMT sistem. Dalam projek ini, kami meletakkan teks sains sistem menekankan pembangunan English->Malay MT di domain komputer. Walau bagaimanapun, metodologi sama boleh disesuaikan mengikut merangka sistem Tm untuk tipologi lain teks, dan secara semula jadi juga untuk sebarang pasang bahasa-bahasa lain.

We provide also in the following table a comparison of the results produced by our SiSTeC-ebmt system with different size of its bilingual knowledge bank.

Translation to Malay by SiSTeC-ebmt (1,500 S-SSTCs)	**Translation to Malay by SiSTeC-ebmt (25,000 S-SSTCs)**	**Translation to Malay by SiSTeC-ebmt (100,000 S-SSTCs)**
Tujuan utama projek itu memerikan dengan kertas ini untuk membina bungkusan perisian jeneral yang memberikan mengintegrasikan persekitaran untuk pembinaan S-SSTC menempatkan sistem EBMT. Dalam projek ini, kami menyimpan penekanan terhadap perkembangan-perkembangan English->Malay MT sistem dalam kawasan kekuasaan komputer teks sains. Walau bagaimanapun, metodologi sama boleh menjadi disadur untuk berkembang MT sistem untuk sebarang typology yang lain (-lain) teks, dan semula jadinya juga untuk sebarang pasangan bahasa yang lain (-lain).	Tujuan sesalur projek itu dikatakan dengan kertas ini untuk membina perisian jeneral bungkusan memberikan yang mengintegrasikan persekitaran untuk senibina S-SSTC berasaskan sistem EBMT. Dalam projek ini, kami meletakkan penekanan terhadap perkembangan English->Malay MT sistem dalam domain komputer teks sains. Walau bagaimanapun, perkaedahan yang sama boleh menjadi disesuaikan memajukan sistem MT untuk typology yang lain (-lain) teks, dan semula jadinya juga untuk bahasa yang lain (-lain) pasang.	Tujuan utama daripada projek itu digambarkan di dalam kertas ini untuk membina perisian umumnya pakej yang menyediakan mengintegrasikan S-SSTC persekitaran bagi pembinaan berdasarkan EBMT sistem. Dalam projek ini, kami meletakkan teks sains sistem menekankan pembangunan English->Malay MT di domain komputer. Walau bagaimanapun, metodologi sama boleh disesuaikan mengikut merangka sistem Tm untuk tipologi lain teks, dan secara semula jadi juga untuk sebarang pasang bahasa-bahasa lain.

Acknowledgments

The work reported in this paper is supported by a Brain Gain Research Grant from the Malaysian Ministry of Science, Technology & Innovation. We thank also to the French Embassy at Malaysia and the ANR Traouiero project for the supports provided to the first author to have one month stay at Université Joseph Fourier (UJF) as a visiting researcher.

References

[Boitet&Zaharin 88] Zaharin bin Yusoff, Christian Boitet, Representation trees and string-tree correspondences, COLING-88, Budapest, August 1988, pp. 59-64.

[Goh 96] Goh Chooi Ling, Pengunting Papan, Laporan Projek Tahun Akhir Sains Komputer, USM, 1996.

[Heng 95] Heng Ai Looi, Natural Language generation in Machine Translation (MT) Based on the String-Tree Correspondence Grammar (STCG), Dissertation submitted in fulfillment of the M.Sc., 1995.

[Kawahara & Kurohashi 2010] Daisuke Kawahara and Sadao Kurohashi, Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation, In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), 2010.

[Khardon&Roth 94] Roni Khardon, Dan Roth, Learning to Reason, Proceedings of AAAI-94.

[Maxwell&Schubert 89] Dan Maxwell, Klaus Schubert (eds), Metataxis in practice: Dependency syntax for multilingual machine translation. Dordrecht/Providence : Foris. DLT 6, 1989.

[Menezes et. al. 2001] Menezes, A., Richardson, S.: A Best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001, Toulouse, France, 2001.

[Mosleh et. al. 2002] Mosleh H. Al-Adhaileh & Tang Enya Kong & Zaharin Yusoff. “A Synchronization Structure of SSTC and ITS Applications in MachineTranslation”. COLING 2002 Post-conference Workshop "Workshop on Machine Translation in Asia", Taipei, Taiwan, September, 2002

[Richardson et. al. 2001] Richardson, S., Dolan, W., Menezes, A., Pinkham, J.: Achieving commercial-quality translation with example-based methods. In: Proceedings of MT Summit VIII, Santiago de Compostela, Spain, 2001

[Sadler&Vendelmans 90] Victor Sadler, Ronald Vendelmans, Pilot Implementation of a bilingual Knowledge Bank, COLING 90, Helsinki, 1990.

[Sato 91] Sato, S., Example-Based Translation Approach to Machine Translation, Proceedings of the International Workshop on Fundamental Research for Future Generation of Natural Language Processing, ATR interpreting Telephony Research Laboratories, 1991.

[Tang 94] Tang Enya Kong, Natural Language Analysis In Machine Translation (MT) Based On The String-Tree Correspondence Grammar (STCG), Dissertation submitted in fulfillment of the Ph.D., 1994.

[Tang&Zaharin 95] Tang Enya Kong, Zaharin Y., Handling Crossed Dependecies with the STCG, proceedings of NLPRS'95, Sofitel Ambassador Hotel, Seoul, Korea, Dec. 4-6, 1995.