Application of Translation Corresponding Tree (TCT) Annotation Schema in Example-Based Machine Translation

WONG Fai[†], HU Dong Cheng[†], MAO Yu Hang[†], TANG Chi Wai[‡], DONG Ming Chui[‡]

[†]Speech and Language Processing Research Center, Tsinghua University, 100084 Beijing, China
[‡]Faculty of Science and Technology of University of Macao, PO Box 3001, Macao SAR
derekfw@umac.mo, hudc@mail.tsinghua.edu.cn, myh-dau@mail.tsinghua.edu.cn,
sekevin@inesc-macau.org.mo, dmc@inesc-macau.org.mo

Abstract

In this paper, we present an Example-Based Machine Translation (EBMT) system for Portuguese to Chinese translation. In our approach, the examples used for translation are annotated under the representation schema of Translation Corresponding Tree (TCT). Each Translation Corresponding Tree describes a translation example (a pair of bilingual sentences). It represents the syntactic structure of source language sentence (i.e. Portuguese in our system), as well as denotes the translation correspondences (i.e. Chinese translation) for each node in the representation tree. In addition, syntax transformation rules are also encapsulated at each node in the TCT representation that captures the differentiation of grammatical structure between the source and target languages. With this annotation schema, translation examples are effectively represented and organized in the bilingual knowledge database. In the translation process, the source sentence is parsed. The output, syntactic tree, is then used for finding the similar TCTs or constituency parts of TCTs from the knowledge DB. By referring to the translation information coded in the TCTs, target language translation is synthesized.

Introduction

The construction of bilingual knowledge base, in the development of example-based machine translation systems (Sato and Nagao, 1990), is vitally critical. In the translation process, the application of bilingual examples concerns with how examples are used to facilitate translation, which involves the factorization of an input sentence into the format of stored examples and the conversion of source texts into target texts in terms of the existing translations by referencing to the bilingual knowledge base. Theoretically speaking, examples can be achieved from bilingual corpus where the texts are aligned in sentential level, and technically, we need an example base for convenient storage and retrieval of examples. The way of how the translation examples themselves are actually stored is closely related to the problem of searching for matches. In structural example-based machine translation systems (Grishman, 1994; Meyers et al., 1998; Watanabe et al., 2000), examples in the knowledge base are normally annotated with their constituency (Kaji et al., 1992) or dependency structures (Matsumoto et al., 1993), which allows the corresponding relations between source and target sentences to be established at the structural level. All of these approaches annotate examples by mean of a pair of analyzed structures, one for each language sentence, where the correspondences between inter levels of source and target structures are explicitly linked. However, we found that these approaches require the bilingual examples that have '*parallel*' translations or '*close*' syntactic structures (Grishman, 1994), where the source sentence and target sentences have explicit correspondences in the sentences-pair. For example, in (Wu, 1995), the translation examples used for building the translation alignments are strictly selected based on constraints. As a result, these approaches indirectly limit their application in using the translation examples that are '*free translation*' for the development of example-based machine translation system. In this paper, we overcome the problem by designing a flexible representation schema, called Translation Corresponding Tree (TCT). We use the Translation Corresponding Tree (TCT) as the basic structure to annotate the examples in our bilingual knowledge base for the Portuguese to Chinese example-based machine translation system.

Translation Corresponding Tree Representation

Translation Corresponding Tree structure, as an extension of structure string-tree correspondence representation (Boitet and Zaharin, 1988), is a general structure that can flexibly associate not only the string of a sentence to its syntactic structure in source language, but also allow the language annotator to explicitly associate the string from its translation in target language for the purpose to describe the correspondences between different languages.

The TCT Structure

The TCT representation uses a triple sequence intervals [SNODE(*n*)/STREE(*n*)/STC(*n*)] encoded for each node in the tree to represent the corresponding relations between the structure of source sentence and the substrings from both the source and target sentences. In TCT structure, the correspondence is made up of three interrelated correspondences: 1) one between the node and the substring of source sentence encoded by the interval SNODE(*n*), which denotes the interval containing the substring corresponding to the node, 2) one between the subtree and the substring of source sentence represented by the interval STREE(*n*), which indicates the interval of substring that is dominated by the subtree with the node as root, and 3) the other between the subtree of source sentence and the substring of target sentence expressed by the interval STC(*n*), which indicates the interval containing the substring in target sentence corresponding to the subtree of source sentence. The associated substrings may be discontinuous in all cases. This annotation schema is quite suitable for representing translation example, where it preserves the strength in describing non-standard and non-projective linguistic phenomena for a language (Boitet and Zaharin, 1988; Al-Adhaileh et al., 2002), on the other hand, it allows the

annotator to flexibly define the corresponding translation substring from the target sentence to the representation tree of source sentence when it is necessary. This is actually the idea behind the formalism of Translation Corresponding Tree.

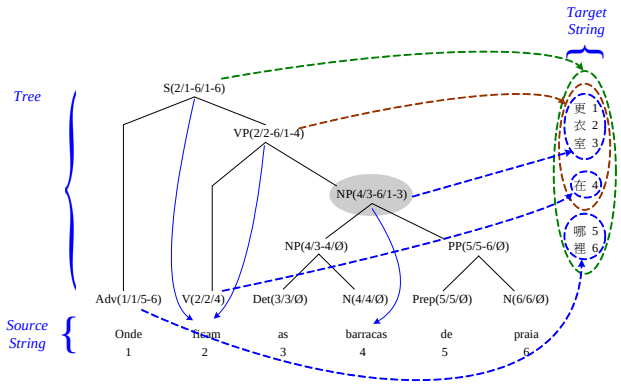

Figure 1: An TCT representation for annotating the translation example "*Onde ficam as barracas de praia? (Where are the bathhouses?)*/*更衣室在哪裡*?" and its phrase structure together with the correspondences between the substrings (of both the source and target sentences) and the subtrees of sentence in source language.

As illustrated in Figure 1, the translation example "*Onde ficam as barracas de praia?*/ *更 衣 室 在 哪 裡 ?*" is annotated in a TCT structure. Based on the interpretation structure of the source sentence "*Onde ficam as barracas de praia?*", the correspondences between the substrings (of source and target sentences) and the grammatical units at different inter levels of the syntactic tree of the source sentence are expressed in terms of sequence intervals. The words of the sentences pair are assigned with their positions respectively, i.e. "*Onde* (1)", "*ficam* (2)", "*as* (3)", "*barracas* (4)", "*de* (5)" and "*praia* (6)" for the source sentence, as well as for the target sentence. But considering that Chinese uses ideograms in writing without any explicit word delimiters, the process to identify the boundaries of words is considered to be the task of word segmentation, instead of assigning indices in word level with the help of word segmentation utility, a position interval is assigned to each character for the target (Chinese) sentence, i.e. "*更* (1)", "*衣* (2)", "*室* (3)", "*在* (4)", "*哪* (5)" and "*裡* (6)". Hence, a substring in source sentence that corresponds to the node of its representation is denoted by the intervals encoded in SNODE(*n*) for the node, e.g. the shaded node, *NP*, with interval, SNODE(*NP*)=4, corresponds to the substring "*barracas*" in source sentence that has the same interval. A substring of source sentence that corresponds to a subtree of its syntactic tree is denoted by the interval recorded in STREE(*n*) attached to the root of the subtree, e.g. the subtree of the shaded node, *NP*, encoded with the interval, STREE(*NP*)=3-6, corresponds to the substring "*as barracas de praia*" in source sentence. While the translation correspondence between the subtree of source sentence and substring in the target sentence is denoted by the interval assigned to the STC(*n*) of each node, e.g. the subtree rooted at shaded node, *NP*, with interval, STC(*NP*)=1-3, corresponds to the translation fragment (substring) "*更衣室*" in target sentence.

Expressiveness of Linguistic Information

Another inherited characteristic of TCT structure is that it can be flexibly extended to keep various kinds of linguistic information, if they are considered useful for specific purpose, in particularly the linguistic information that differentiating the characteristics of two languages which are structural divergences (Wong et al., 2001). Basically, each node representing a grammatical constituent in the TCT annotation is tagged with grammatical category (part of speech). Such feature is quite suitable for the describing specific linguistic phenomena due to the characteristic of a language. For instance, in our case, the crossing dependencies (syntax transformation rules) for the sentence constituents between Portuguese and Chinese are captured and attached to each node in the TCT structure for a constituent that indicates the order in forming the corresponding translation for the node from the subtrees it dominated. In many phrasal matching approaches, such as constituency-oriented (Kaji et al., 1992; Grishman, 1994) and dependency-oriented (Matsumoto et al., 1993; Watanabe et al., 2000), crossing constraints are deployed implicitly in finding the structural correspondences between pair of representation trees of a source sentence and its translation in target. Here, in our TCT representation, we adopted the use of constraint (Wu, 1995) for a constituent unit, where the immediate subtrees are only allowed to cross in the inverted order. Such constraints, during the phase of target language generation, can help in determining the order in producing the translation for an intermediate constituency unit from its subtrees when the corresponding translation of the unit is not associated in the TCT representation.

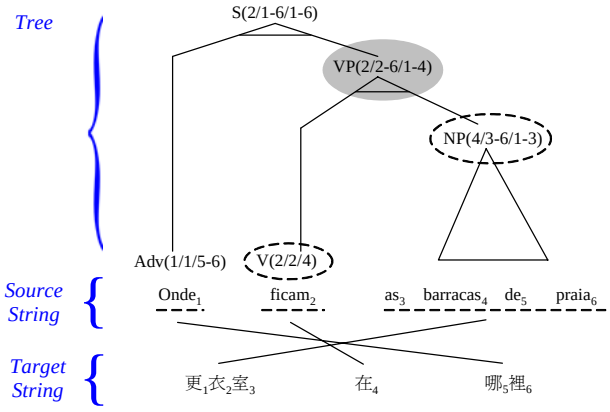

Figure 2: The transfer relationships between the sentence-constituents of source language and its translation in target language are recorded in TCT structure.

Figure 2 demonstrates the crossing relations between the source and target constituents in an TCT representation structure. In graphical structure annotation, a horizontal line is used to represent the inversion of translation fragments of its immediate subtrees.

Construction of Example Base

In the construction of bilingual knowledge base (example base) in example-based machine translation system (Sato and Nagao, 1990; Watanabe et al., 2000), translation examples are usually annotated by mean of a pair

analyzed structures, where the corresponding relations between the source and target sentences are established at the structural level through the explicit links. Here, to facilitate such examples representation, we use the Translation Corresponding Tree as the basic annotation structure.

TCT Generation

In our example base, each translation pairs is stored in terms of an TCT structure. Conceptually speaking, the construction of the example base can be viewed as the process in building the TCT structures for the example cases. To a translation example, the system will automatically process and generate a preliminary TCT representation structure for it. The resultant annotation tree is then further edited by human through the use of an TCT editing program if any amendment to the representation structure is necessary.

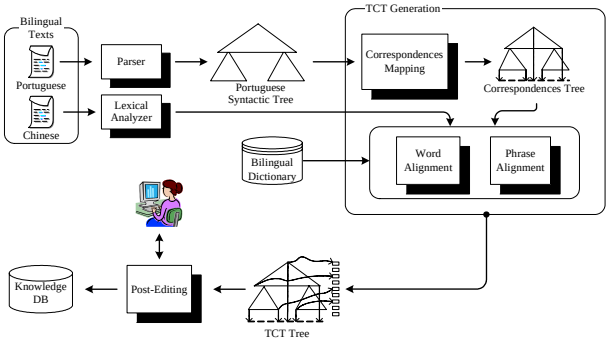

Figure 3: The construction of bilingual knowledge base based on the representation structure of TCT.

In the generation process, it starts by analyzing the grammatical structure of Portuguese sentence with the aid of a Portuguese parser, and a shallow analysis to the Chinese sentence is carried out by using the Chinese Lexical Analysis System (ICTCLAS) (Zhang, 2002) to segment and tag the words with a part of speech. The grammatical structure produced by the parser for Portuguese sentence is then used for establishing the correspondences between the surface substrings and the inter levels of its structure, which includes the correspondences between nodes and its substrings, as well as the correspondences between subtrees and substrings in the sentence. Next, in order to identify and establish the translation correspondences for structural constituents of Portuguese sentence, it relies on the grammatical information of the analyzed structure of Portuguese and a given bilingual dictionary to search the corresponding translation substrings from the Chinese sentence. Finally, the consequent TCT structure will be verified and edited manually to obtain the final representation, which is the basic element of the knowledge base. The overall process in constructing the bilingual knowledge base is depicted in Figure 3, and Figure 4 illustrates the example "*Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行爲*" with its corresponding TCT structure.

Translation Equivalents

Through the notation of translation corresponding structure for representing translation examples in the bilingual knowledge base, the translation units between the Portuguese sentence and its target translation in Chinese are explicitly expressed by the sequence intervals STREE(*n*) and STC(*n*) encoded in the intermediate nodes of an TCT structure, that may represent the phrasal and lexical correspondences. For instance, from the translation example being annotated under the TCT representation schema as shown in Figure 4, the Chinese translation "*訴訟* " of Portuguese word "*acção*" is denoted by [STREE(*n*)=6/STC(n)=2-3] in the terminal node. For phrasal translation, we may visit the higher level constituents in the representing structure of TCT and apply the similar coding information to retrieve the corresponding translation for the unit that representing a phrasal constituent in a sentence. In order that the representation examples can be effectively consulted, each TCT structure is being indexed by its nodes in the bilingual knowledge base. Thus, all the possible sub-TCTs (translation units) or the constituency structures of an TCT can be easily retrieved for reference.

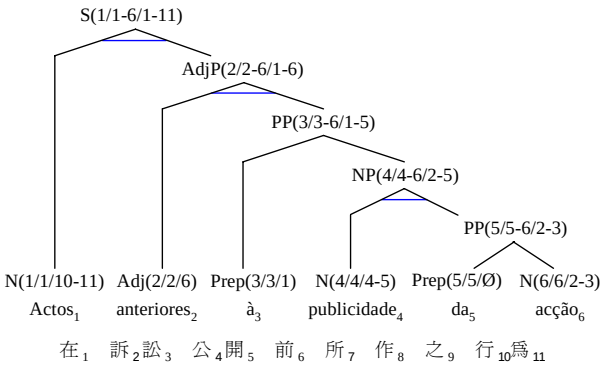

Figure 4: A TCT structure constructed for the translation example "*Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行爲*".

Example-Based Translation Based on TCT

In example-based machine translation systems, a corpus of translation examples used to facilitate the translation rather than linguistic rules is the significant component (Sato and Nagao, 1990). In our approach, translation examples are annotated under the representation structure of TCT. Each TCT structure consists of a sentence in source language, e.g. Portuguese in our case, an associated constituency structure that describing the source sentence, the mapping between the inter levels of abstracted structure and its surface string of the sentence, as well as the corresponding relations against its translation in target language, e.g. Chinese, including the translation fragments and the constraints of crossing dependencies between the source and target phrasal units. During the translation process, a new input sentence is first analyzed into the form of representation structure, followed by retrieving the related examples that contain the same words or comprise the same constituency structures as the input sentence from the example base, and use them to synthesize the final translation for the input sentence

guided by the syntactic information of sentential constituents and the translation correspondences of the referenced examples. The overall picture of the translation processes is depicted in Figure 5.

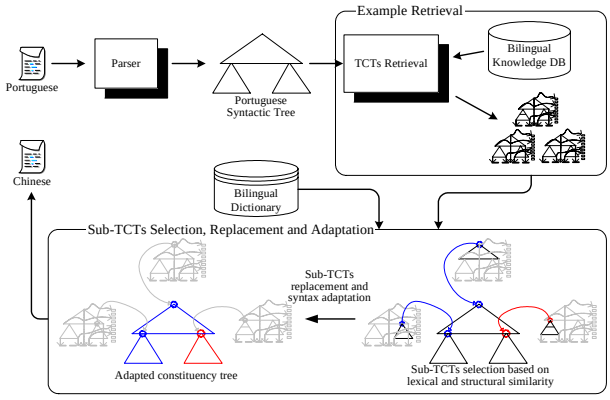

Figure 5: The overall translation processes by using the TCT representation examples as the bilingual knowledge base (example base).

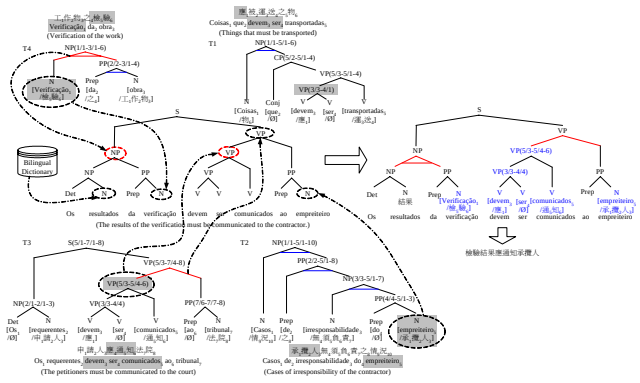

Figure 6: Translation by matching and replacing.

To translate a Portuguese sentence, in our system, can be viewed as the process to construct an TCT structure for describing the input sentence guided by the collection of annotated TCT representations of examples from the example base, follow by traversing the resultant representation structure according to the order being controlled by the crossing constraints encoded in each node (grammatical unit) to produce the target translation for the source sentence in Chinese. During the process, the internal structure of the source sentence is first analyzed with the help of a parser and a syntactic representation tree of the sentence is produced as the parsing result. Then for each subgraph (constituency unit) of the constructed tree, the system retrieves a list of close related TCTs or sub-TCTs from the example base based on the constraint that the constituency units (TCTs or sub-TCTs) that have similar grammatical structure (as well as the grammatical categories labeled for the root nodes and the dominated nodes) as that of the source sentence are recalled. In addition, the content words of the root node of the constituency unit will also be considered for determining the examples that are completely matched to the source sentence. After the related examples are identified and obtained from the example base, the next step is to select the set of TCTs or sub-TCTs to form a complete TCT structure that can best describe the source sentence by replacing the subtrees of source sentence with the chosen sub-TCTs. For those of unmatched terminal nodes, the corresponding Chinese translation can be consulted from a given bilingual dictionary and filled to complete the construction of TCT structure for the sentence. In the case if more than one example is found, the system will evaluate the distance between the chosen examples and the source sentence based on the edit distance function. The replacement process to construct the target TCT for the source sentence is demonstrated in Figure 6. Finally, the corresponding translations appeared in the resultant TCT structure are combined to form the target translation in Chinese.

Acknowledgement

The research work reported in this paper was supported by the Research Committee of University of Macao under grant CATIVO:3678.

References

Al-Adhaileh, M.H., Tang, E.K. & Zaharin, Y. (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.

Boitet, C. & Zaharin, Y. (1988). *Representation trees and string-tree correspondences*. In Proceedings of COLING-88, Budapest, pp.59-64.

Grishman, R. (1994). *Iterative Alignment of Syntactic Structures for a Bilingual Corpus*. In Proceedings of Second Annual Workshop on Very Large Corpora (WVLC2), Kyoto, Japan, pp.57-68.

Kaji, H., Kida, Y. & Morimoto, Y. (1992). *Learning Translation Templates from Bilingual Text*. In Proceedings of CoLING-92, Nantes, pp.672-678.

Matsumoto, Y., Isimoto, H. & Utsuro, T. (1993). *Structural Matching of Parallel Texts*. 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp.23-30.

Meyers, A., Yangarber, R. & Ralf, B. (1998). *Deriving Transfer Rules from Dominance-Preserving Alignments*. In Proceedings of Coling-ACL (1998), pp.843-847.

Sato, S. & Nagao, M. (1990). *Toward Memory-Based Translation*. In Proceeding of Coling (1990), Vol.3, pp.247-252.

Watanabe, H., Kurohashi, S. & Aramaki, E. (2000). *Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation*. In Proceedings of COLING-2000.

Wong, F., Mao, Y.H., Dong, Q.F. & Qi, Y.H. (2001). *Automatic Translation: Overcome the Barriers between European and Chinese Languages*. In Proceedings (CD Version) of First International UNL Open Conference 2001, SuZhou China.

Wu, D. (1995). *Grammarless extraction of phrasal translation examples from parallel texts*. In Proceedings of TMI-95, Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, v2, Leuven Belgium, pp.354-372.

Zhang, H.P. (2002). *ICTCLAS*. Institute of Computing Technology, Chinese Academy of Sciences: http://www.ict.ac.cn/freeware/003_ictclas.asp.