A Flexible Example Annotation Schema: Translation Corresponding Tree Representation

Fai WONG, Dong Cheng HU, Yu Hang MAO
Speech and Language Processing Research Center,
Tsinghua University, 100084 Beijing
huangh01@mails.tsinghua.edu.cn
{hudc, myh-dau}@mail.tsinghua.edu.cn

Ming Chui DONG
Faculty of Science and Technology
of University of Macao,
PO Box 3001, Macao SAR
dmc@inesc-macau.org.mo

Abstract

This paper presents work on the task of constructing an example base[1] from a given bilingual corpus based on the annotation schema of Translation Corresponding Tree (TCT). Each TCT describes a translation example (a pair of bilingual sentences). It represents the syntactic structure of source language sentence, and more importantly is the facility to specify the correspondences between string (both the source and target sentences) and the representation tree. Furthermore, syntax transformation clues are also encapsulated at each node in the TCT representation to capture 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 that we need for the Portuguese to Chinese machine translation system.

1 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; Aramaki et al., 2001; Al-Adhaileh et al., 2002), 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 selected based on strict constraints. As a result, these approaches indirectly limit their application in using the translation examples that are '*free translation*' to the development of example-based machine translation system. In practice, most of the existing bilingual corpus, the meanings of the source sentences are interpreted in target language in the nature of '*freer*', other than literally translated in a projective manner and stayed as close to the source text as possible, in particular for the languages-pair that are structural divergences, such as Portuguese and Chinese.

[1] Or bilingual knowledge base, we use the two terms interchangeably.

As illustrated in Figure 1, the translation of the Portuguese sentence "*Onde ficam as barracas de praia?*" is interpreted into "*更衣室在哪裡?* (Where are the bathhouses?)" other than straightly translated to "*沙灘帳篷在哪裡* ? (Where are the tents of beach?)". The translations of the words, i.e. "*barracas*" and "*praia*", of the source sentence do not explicitly appear in target sentence. As a result, in the conventional alignment process, to achieve a fully aligned structural representation for such sentences-pair may be problematic. However, we found that such type of examples is very common. We have investigated around 2100 bilingual examples that are extracted from a grammar book "*Gramática da Língua Portuguesa*" (Wang and Lu, 1999), and found that 63.4% of examples belong to the discussed case, where the number of *unmatched* words is more than half the number of words in source sentence. In this paper, we overcome the problem by designing a flexible representation schema, called Translation Corresponding Tree (TCT). We use the TCT as the basic structure to annotate the examples in our example bilingual knowledge base for the Portuguese to Chinese example-based machine translation system.

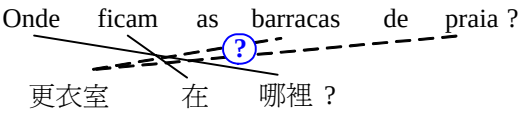

Figure 1. An example of '*free translation*', where the translations of some words in Portuguese sentence do not appear in target Chinese sentence.

2 Translation Corresponding Tree (TCT) Representation

TCT 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.

2.1 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 central idea behind the formalism of TCT.

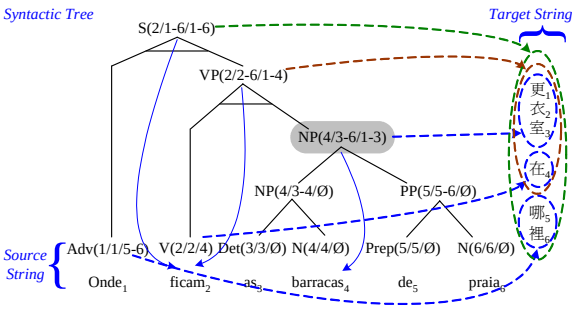

Figure 2. An TCT representation for annotating the translation example "*Onde ficam as barracas de praia? (Where are the bathhouses?) / 更衣室在哪裡?*" and its phrase structure together with the correspondences between the substrings (of both the source and target sentences) and the subtrees of sentence in source language.

As illustrated in Figure 2, 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 (Teahan et al., 2000), 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.

2.2 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; Aramaki et al., 2001), 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 constituent unit from its subtrees when the corresponding translation of the unit is not associated in the TCT representation.

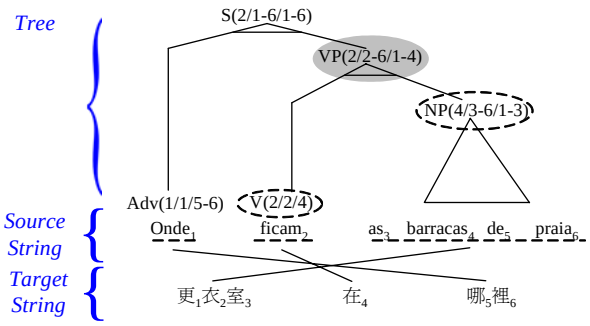

Figure 3. The transfer relationships between the sentence-constituents of source language and its translation in target language are recorded in TCT structure.

Figure 3 demonstrates the crossing relations between the source and target constituents in an TCT representation structure. In graphical struc-

ture annotation, a horizontal line is used to represent the inversion of translation fragments of its immediate subtrees. For example, the translation substring "*更衣室在*" of the shaded node, *VP*, can be obtained by inverting the order of the corresponding target translations "*在*" and "*更衣室*" from the dominated nodes *V* and *NP*. Therefore, such schema can serve as a mean to represent translation examples, and find structural correspondences for the purpose of transfer grammar learning (Watanabe et al., 2000; Matsumoto et al., 1993; Meyers et al., 1998).

3 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. The main different and advantage of our approach is that it uses a single language parser to process other than two different parsers, one for each language (Tang and Al-Adhaileh, 2001).

In our example base, each translation pairs is stored in terms of an TCT structure. The construction 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.

3.1 The TCT Generation Algorithm

In the overall construction processes, the task to compile the syntactic structure of source sentence into the TCT representation by linking the translation fragments from the target sentence is the vital part. The following steps present the complete process to generate an TCT structure for a translation example *"Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行為*".

Parsing Portuguese Sentence

The process begins by parsing the Portuguese sentences with a Portuguese parser. The parsing result is a phrase structure in terms of bracketed annotation. Each bracketed constituent of the structure tree is attached with a grammatical category. Figure 4 shows the resultant parsed structure of the Portuguese sentence.

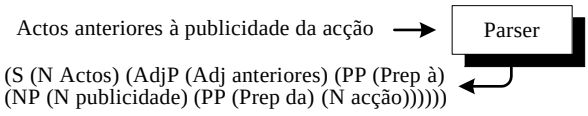

Figure 4. Portuguese sentence is analyzed by a linguistic parser, and its output is the phrase structure expressed in bracket notation.

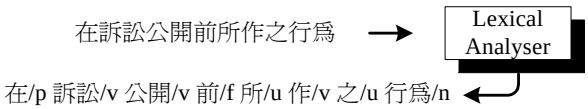

Figure 5. The analyzed lexical items for Chinese sentence.

Analyzing Chinese Sentence

The construction of TCT structure is fundamentally based on the syntactic structure of Portuguese sentence. The finding of translation units between the sentences pair is relying on structure tree of Portuguese sentence and the sequences of lexical words from Chinese sentence. Thus, instead of analyzing the Chinese sentence in deep, we analyze the Chinese sentence in the lexical

level by using the Chinese Lexical Analysis System (ICTCLAS) (Zhang, 2002). Each Chinese word is delimited with spaces and assigned with a part of speech as illustrated in Figure 5.

Constructing Correspondence Structure for Portuguese Sentence

After parsing and obtaining the syntactic structure of Portuguese sentence, next step is to compute the correspondences for the structure against the surface strings of the source sentence, which includes the corresponding phrase for a constituent unit in the tree and the corresponding content word that headed the constituent unit, both of these correspondences are denoted by the sequence intervals of the substrings spanning across the sentence fragments. In finding the corresponding phrasal substrings for subtrees, we start associating the lexical words to the corresponding terminal nodes of the structure tree by assigning the related offsets to SNODE(n) and STREE(n) of the nodes. Then we proceed to next upper level constituent units in the tree where the corresponding substrings are derived by connecting the lexical words from the nodes in the lower level it dominated. Theoretically, if node, N, has m daughters, $N_1 \ldots N_m$, then the sequence interval for N will be STREE(N) = STREE(N_1) $\cup$ STREE(N_2) $\cup \ldots \cup$ STREE(N_m), the interval is bounded by spanning nodes of its immediate subtrees. To identify the lexical head for a constituent unit, we use simple rule to determine it by considering the grammatical category of the phrasal unit, and choose the word that owns the same category from the daughter nodes, then assign the interval of chosen to SNODE(N). Figure 6 shows the structure produced in this stage.

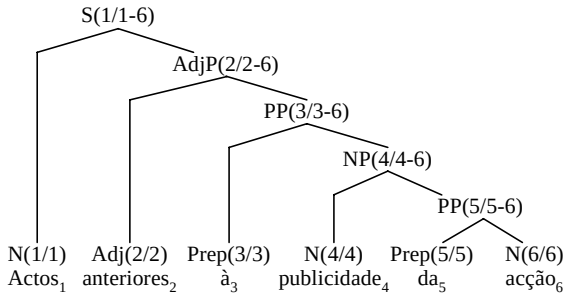

Figure 6. The Portuguese correspondence structure.

Associating Translation Correspondences

In this process, we adopt a search for alignments between constituent units of Portuguese sentence and the corresponding translation fragments from Chinese sentence, proceeding bottom-up through the tree. It makes use of the information about possible lexical correspondences from a bilingual dictionary and the grammatical categories of the lexical words, tagged in previous stage, to generate initial candidate alignments. Figure 7 presents the initial lexical alignments.

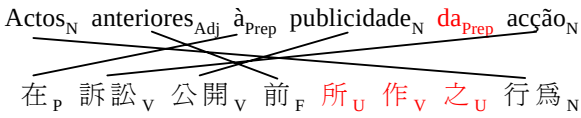

Figure 7. Initial candidate alignments of corresponding words.

Based on the possible word correspondences, the associated structure of the Portuguese sentence, together with the grammatical categories information, the search proceeds to align phrases by gradually increasing length (phrasal correspondences in different levels of constituent tree) based on the following criterions.

First, for any un-aligned words sequence "w_{ua}" being bounded by aligned words of daughter nodes "$w_{a\text{-}left}$" and "$w_{a\text{-}right}$", we take the whole fragment "$w_{a\text{-}left} w_{ua} w_{a\text{-}right}$" (including the bounding words or phrases) as the corresponding substring for the parent node that immediately dominates the daughter nodes, such that STC(N) = STC(N_{left}) $\cup$ STC(N_{right}).

Second, for the case that the un-aligned fragment is not bounded by any aligned units, our approach relies on the assumption that if two set of sentence constituents (source and target sentences) are corresponding, their grammatical categories as well as the number of constituents should be consistent. The essential idea of the search is to look for inter levels where the constituent units of the structure of Portuguese sentence and the lexical words in Chinese sentence can be projected in one-to-one manner. We use the previous example "*Onde ficam as barracas de praia? (Where are the bathhouses?)*/ *更衣室在哪裡?*" to illustrate the searching strategy. Beside the corresponding lexical items, e.g. "*Onde* / *哪裡*" and "*Ficam* / *在*", that can be de-

termined with the aid of a given dictionary, the process proceeds bottom-up and searches through the tree by considering only the unmatched items that if the assumption hold or not. For example, at the leaf level, the different numbers of the lexical items ("as_{Det}, $barracas_{N}$, de_{Prep}, $praia_{N}$" and "$更衣室_{N}$") violates the assumption. The process repeats the investigation in next upper level in the representation structure of Portuguese sentence. As illustrated in Figure 8, the alignment can be identified only at the level where the number and the part of speech of constituent unit of Portuguese ("$[as\ barracas\ de\ praia]_{NP}$") are consistent to that of the lexical item in Chinese sentence ("$[更衣室]_{N}$"). Consequently, the correspondences between the associated structure of Portuguese sentence and the translation fragments of Chinese sentence can be determined and established. For any node in the structure which has no translation equivalent is assigned with "*empty (Ø)*" interval to STC(*N*).

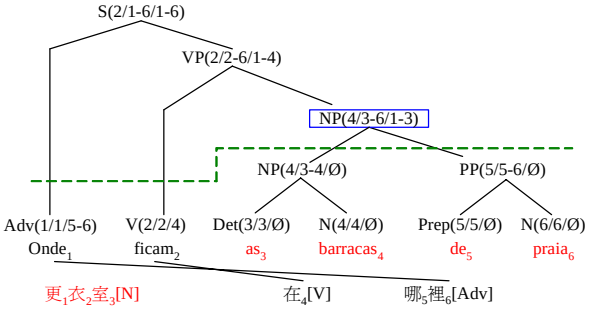

Figure 8. Finding the alignment for unbounded words.

Third, for acquiring the crossing constraint for a constituent node in the representation tree, which is determined by examining the order of the translation correspondences of the spanning nodes against the sequence of those appeared in Chinese sentence. For any node that representing Portuguese phrase whose corresponding translation is derived from its daughters by inverting the corresponding translations is denoted by assigning a Boolean value to INVERT(*N*) attached to the node. In graphical annotation, a horizontal line is used as a sign for indicating the inversion. As demonstrated in Figure 9, the corresponding translations of the daughters of node *S* are crossed between the sentences of Portuguese and its translation in Chinese. The corresponding translation "*在訴訟公開前*" of its second daughter appears prior to that "*行為*" of the first daughter node in the target translation of Portuguese sentence. Hence the inversion property for the constituent node in the syntactic structure of source sentence is consequently determined.

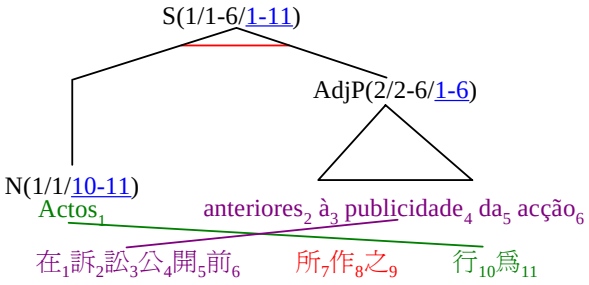

Figure 9. Determination of crossing dependency between the translation correspondences

Finally, in case the representation of TCT generated in previous process needs further editing, an TCT editor can be used to perform the necessary amendment. Figure 10 presents the final TCT structure describing a translation example.

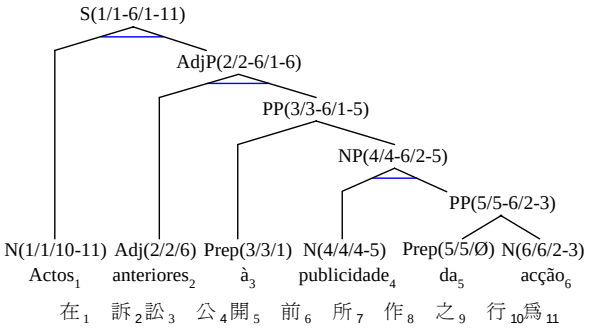

Figure 10. An TCT structure constructed for the translation example "*Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行為*".

3.2 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 10, 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. Each TCT structure is being indexed by its nodes in the bilingual knowledge base, in order that the representation examples can be effectively consulted.

4 Conclusion

In this paper, a novel annotation schema for translation examples, called Translation Corresponding Tree (TCT) structure, is proposed and has been applied to the construction of bilingual knowledge base (example base) to be used for the Portuguese to Chinese machine translation system. The TCT representation provides a flexible nature to describe the corresponding relations between the inter levels of the structure against its substrings in a sentence, in particular the corresponding translation fragments (substrings) from the target translation sentence are explicitly expressed in the structure. We have proposed a strategy to semi-automate the example base construction process. A preliminary TCT structure for a translation example is first produced by the system, then the representation structure can be further modified manually through an TCT editor to get the final structure.

Acknowledgement

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

References

Mosleh Hmoud Al-Adhaileh, Enya Kong Tang, and Yusoff Zaharin. 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.

Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe. 2001. *Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation*. In Proceedings of MT Summit VIII, pp.27-32.

Christian Boitet, and Yusoff Zaharin. 1988. *Representation trees and string-tree correspondences*. In Proceeding of COLING-88, Budapest, pp.59-64.

Ralph Grishman. 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.

Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. 1992. *Learning Translation Templates from Bilingual Text.* In Proceeding of COLING-92, Nantes, pp.672-678.

Yuji Matsumoto, Hiroyuki Isimoto, and Takehito Utsuro. 1993. *Structural Matching of Parallel Texts*. 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp.23-30.

Adam Meyers, Roman Yangarber, and Brown Ralf. 1998. *Deriving Transfer Rules from Dominance-Preserving Alignments*. In Proceedings of Coling-ACL (1998), pp.843-847.

Satoshi Sato, and Magnus Nagao. 1990. *Toward Memory-Based Translation*. In Proceeding of Coling (1990): pp.247-252.

Enya Kong Tang, and Mosleh Hmoud Al-Adhaileh. 2001. *Converting a Bilingual Dictionary into a Bilingual Knowledge Bank based on the Synchronous SSTC*. In Proceedings of Machine Translation Summit VIII, Spain, pp.351-356.

Suo Ying Wang, and Yan Bin Lu. 1999. *Gramática da Língua Portuguesa*. Shanghai Foreign Language Education Press.

Fai Wong, Yu Hang Mao, Qing Fu Dong, and Yi Hong Qi. 2001. *Automatic Translation: Overcome the Barriers between European and Chinese Languages*. In Proceedings (CD Version) of First International UNL Open Conference 2001, SuZhou China.

Dekai Wu. 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.

Hua Ping Zhang. 2002. ICTCLAS. Institute of Computing Technology,Chinese Academy of Sciences: http://www.ict.ac.cn/freeware/003_ictclas.asp.