Translation Technologies of Portuguese to Chinese Machine Translation System: PCT Assistente

Fai Wong, Mingchui Dong, Chiwai Tang, Francisco Oliveira
Faculty of Science and Technology of University of Macao, INESC Macau
{derek, dmc, kevin, francisco}@inesc-macau.org.mo

Abstract

This paper presents the translation technologies of Portuguese to Chinese machine translation system, PCT Assistente - a hybrid translation model integrated with example-based and rule-based translation paradigms to counterbalance the intrinsic weakness of these approaches by combining the strong features of another. In the example-based model, translation examples are annotated under the schema of Translation Corresponding Tree (TCT) structure that acted as the basic knowledge for performing the translation task. While in the rule-based model, Constraint Synchronous Grammar (CSG) is adopted as the language formalism to analyze the syntactic structure of an input sentence and produce the corresponding translation in target language during the translation process. In this paper, the translation paradigms based on Translation Corresponding Tree and Constraint Synchronous Grammar are discussed. The linkage of two translation modules by combining the strengths of different approaches to improve the quality of the translation system is studied, and finally, the architecture of network based PCT Assistente (machine translation) system is described.

1. Introduction

The objective of any machine translation system is to overcome the language barrier among the people for globalization of information available in different languages. The other important goal is to provide effective tools for the professional translators to improve their daily work in a more efficient way. Many attempts are being made all over the world to develop machine translation systems. However, the main bottlenecks in the development of efficient machine translation systems are inherent ambiguities involved in natural languages which need enormous knowledge for the disambiguation, and difficulty in finding the form for representing the knowledge. Many different approaches to machine translation system design have been proposed in literature including the rule-based MT, example-based MT, and statistic-based MT (Bennett and Slocum, 1985, Sato and Nagao, 1990, Brown et al., 1993, McTait, 2001, Knight and Marcu, 2005, Wong et al., 2006). Each of these approaches has its strength and weakness in application to the development of machine translation alone. The combination of these methods leads to a hybrid system, which seems the way to go, to avoid the intrinsic obstacles of both different translation methods (Carl and Hansen, 1999, Jain et al.,

2001).

In this paper, a hybrid system is designed to integrate the advantages of example-based and rule-based approaches and get rid of their disadvantages. Based on this strategy, a Portuguese to Chinese machine translation system, PCT Assistente, is implemented. The paradigms we consider here are the example-based MT based on Translation Corresponding Tree (TCT) (Wong et al., 2004, Wong et al., 2006a, Tang et al., 2006) and the rule-based MT based on Constraint Synchronous Grammar (CSG) (Wong et al., 2005, Wong et al., 2006b). In example-based system, the representation of translation examples is closely related to algorithms of searching and matching examples (or example fragments) for use to facilitate translation for incoming texts. TCT has been proposed as an alternative representation structure. Furthermore, an intrinsic property of TCT is that it can be used to flexibly annotate translation examples that are not literally translated (Wong et al., 2004). This is especially important to the case of Portuguese and Chinese translation, since quality bilingual corpora of these languages are not available. In the rule-based MT paradigm, the translation accuracy mostly depends on the correct analysis. Therefore the disambiguation of parsing results becomes one of the critical issues in the development. Constraint Synchronous Grammar (CSG) is used as the language formalism in our rule-based MT module for disambiguating structures and for improving the parsing performance, since CSG allows the parser to remove the incorrect (ambiguous) structures as the parsing progresses by making use of various linguistic features.

The remainder of the paper is organized as follows. Section 2 discusses the representation of translation examples in example-based MT system by using the Translation Corresponding Tree structure, and the corresponding translation process based on such representation schema. The description of Constraint Synchronous Grammar and its function in rule-based translation system is presented in section 3. The integration of TCT-MT and CSG-MT modules in the system is detailed in section 4. Finally, the current development of PCT Assistente and its extension to network based translation is reviewed in section 5, and a conclusion is made to end this paper.

2. Translation Based on TCT – The Example-Based MT Approach

In structural example-based MT systems, examples in knowledge base are normally annotated with their dependency structures (Aramaki et al., 2001), and the corresponding relationships between source and target sentences are established at structural level. However, these approaches require bilingual examples that have ‘parallel’ translation or ‘close’ syntactic structures (Grishman, 1994). That is, the source and target sentences have explicit corresponding constituents. Moreover, these methods need two linguistic parsers to analyze the bilingual text, and robust parser is not always available, especially for two different languages. Besides, the management of the ambiguities in two language parsers has to tackle

the combination of overall ambiguities (Watanabe et al., 2000). Thus, we employ the Translation Corresponding Tree (TCT) structure as the annotation schema for describing the translation examples in example-based MT module, since it uses only a single syntactic structure as the representation media, and its description method provides us a large degree of flexibility in describing some linguistic phenomena that are not standard. This, as a result, resolves the problems we concerned.

2.1. TCT Representation

The TCT structure is proposed as an extension of structure string tree correspondence representation from monolingual to bilingual representation (Wong et al., 2004). By using a single general syntactic tree, not only the sentence string of source language can be flexibly associated with, but also the corresponding translation of the sentence can be associated to the same structure, hence to describe the linguistic correspondences between two languages. The TCT structure uses triple sequence intervals [SNODE(n)/STREE(n)/TTREE(n)$\in\sigma$] encoded for each node in the tree to represent the corresponding relationships between the structure of the source sentence and the substrings from both the source and target sentences. Each set of corresponding information is made up of the three interrelated correspondences: 1) to denote the head word of sub-structure; 2) the dominated substrings of the sub-structure; and 3) the corresponding substrings in target language for the same sub-structure. The associated substrings may be discontinuous in all cases. This annotation schema preserves the ability to describe non-standard and non-projective linguistic phenomena for a language (Boitet and Zaharin, 1988). The schema also allows the annotator to flexibly define the corresponding translation from the target sentence to the syntactic structure of the source sentence when necessary, which is the central idea behind the TCT formalism. Figure 1 illustrates an example annotated under the described representation.

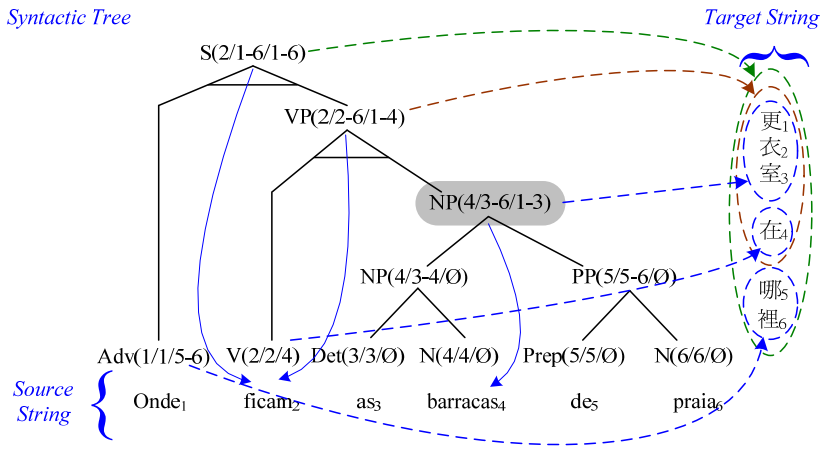

Figure 1 An TCT representation for translation example "Onde ficam as barracas de praia?/更衣室在哪裡?"

The translation equivalents of different constituent units (from lexical level to sentential

level) for an example-pair can be easily retrieved by referencing the related sequence intervals encoded by SNODE(*n*), STREE(*n*) and TTREE(*n*). Another inherited property of TCT structure is that it can be flexibly extended to keep various kinds of linguistic information, if they are useful for specific purpose. For instance, the annotations can capture the crossing dependencies (syntax relationships) for the languages between Portuguese and Chinese (Wong et al., 2001), and is marked with a horizontal line for the nodes to represent the inversion of the translation fragments of its immediate sub-trees, as the VP in Figure 1.

2.2. Translation Based on TCTs

During the translation process, as shown in Figure 2, the input sentence is first analyzed into the syntactic structure with the help of a Portuguese parser. For each possible sub-graph (constituency unit) of the syntactic structure, a list of closely related TCTs (or sub-TCTs) is retrieved from the example base. The searching is based on the criteria: 1) the structural constituents must have similar structures; 2) the grammatical categories of the root nodes and the dominated nodes must be the same as that of the source structure. In addition, the content words of the root node will also be considered to evaluate the degrees of similarities between the examples and the source structure, such that the chosen TCTs (or sub-TCTs) which can best describe the source sentence will be used to reconstruct the target structure. For unmatched terminal nodes, the corresponding translation can be obtained from a bilingual dictionary. If more than one example is found, the system will evaluate the distance between the example candidates and the source sentence based on the edit distance function (Levenshtein, 1966). Finally, the translation for the sentence is generated by traversing through the target tree under the control of syntax constraints.

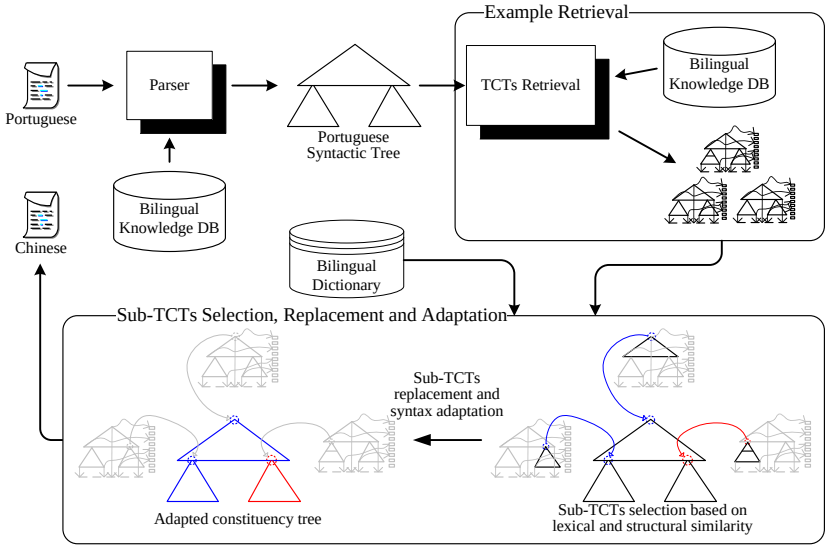

Figure 2 Translation process based on TCT examples as the knowledge base

3. Translation Based on CSG – The Rule-Based MT Approach

In rule-based machine translation, to analyze the structure deviations of languages pair hence to carry out the transformation from one language into another as the target translation is the kernel part in a translation system, and this requires a large amount of structural transformations in both grammatical and concept level. The problems of syntactic complexity and word sense ambiguity have been the major obstacles to produce promising quality of translation. Different from transfer-based and unification-based MT systems, we employ an alternative formalism, Constraint-Based Synchronous Grammar (CSG), as the grammar formalism in our rule-based MT system.

3.1 Definition of Constraint-Based Synchronous Grammar

Constraint-Based Synchronous Grammar (CSG) (Wong et al., 2005) is a variation of synchronous grammars (Lewis and Stearns, 1968) that is based on the formalism of Context Free Grammar (CFG). In CSG formalism, it consists of a set of production rules that describes the sentential patterns of the source text and target translation patterns. Every production rule of CSG is in the form of:

S → *source sentential pattern* { [*target sentential pattern*; control conditions] ,
[*target sentential pattern*; control conditions] ,
...
}

In the left hand side, *S* is the reduced syntactic symbol. In the right hand side of the production, it is divided into two components: the sentential pattern of the source language, and the translation pattern of the target language. Furthermore, in each sentential pattern of the source language, it may consist of one or more translation patterns associated with control conditions based on the features of non-terminal symbols of the source rule for describing the possible generation correspondences in target translation. These conditions are not only used for inferring the structure of source input in the parsing module, but also for structuring the target output pattern in the generation module.

$S \rightarrow NP_1\ VP^*\ NP_2\ PP\ NP_3$ {[$NP_1\ VP^1\ NP_3\ VP^2\ NP_2$; $VP_{category}$ = vb1,
$VP_{sense\ of\ subject} = NP_{1\ sense}$,
$P_{sense\ of\ indirect\ object} = NP_{2\ sense}$, $VP_{sense\ of\ object} = NP_{3\ sense}$],
[$NP_1\ VP\ NP_3\ NP_2$; VP =vb0,
$VP_{sense\ of\ subject} = NP_{1\ sense}$,
$VP_{sense\ of\ indirect\ object} = NP_{2\ sense}$]

}

Figure 3 An example of CSG production

Figure 3 shows an example of CSG production. In this production rule, it has two generative rules associated with the sentential pattern of the source NP_1 VP* NP_2 PP NP_3. The determination of the suitable generative rule is based on the control conditions defined by rule. The one satisfying all the conditions determines the relationship between the source and target sentential pattern. For example, if the category of the verb is vb1, and the sense of the subject, indirect, and direct objects governed by the verb, VP, corresponds to the first, second, and the third nouns (NP), then the source pattern NP_1 VP* NP_2 PP NP_3 is associated with the target pattern NP_1 VP_1 NP_3 VP_2 NP_2. Their relationship is established by their given subscripts and the sequence is based on the target sentential pattern. In other words, in the production S→ NP_1 VP* NP_2 PP NP_3 [NP_1 VP NP_3 NP_2], although the first NP and the verb corresponds to each other in the same sequence, the sequence for the second and third NP in the source are changed in the target sentential pattern. The asterisk "*" indicates the head element, and its usage is to propagate all the related features/linguistic information of the head symbol to the reduced non-terminal symbol in the left hand side. The use of the "*" is to achieve the property of features inheritance in CSG formalism.

3.2 Translation by Parsing CSG

CSG formalism can be parsed by any known CFG parsing algorithm including the Earley (Earley, 1968) and GLR algorithms (Tomita, 1991). In our case, an extended GLR algorithm is adapted for parsing by handling the features constraints and the inference of target sentential patterns. In the view point of formal description for parsing synchronous grammar, to parse a source sentence based on CSG productions can be viewed as to translate the sentence (Wong et al., 2006b). A set *P* of productions is said to *accept* an input string *s* iff there is a derivation sequence *Q* for *s* using source rules of *P*, and any of the constraint associated with every *target component* in *Q* is satisfied. Similarly, *P* is said to *translate s* iff there is a synchronized derivation sequence *Q* for *s* such that *P* accepts *s*, and the link constraints of associated *target rules* in *Q* is satisfied. The derivation *Q* then produces a translation *t* as the resulting sequence of terminal symbols included in the determined target rules in *Q*. The translation of an input string *s* essentially consists of three steps. First, the input string is parsed by using the source rules of productions. Secondly, the link constraints are propagated from source rule to target component to determine and build a target derivation sequence. Finally, translation of input string is generated from the target derivation sequence.

4. Linkage of TCT and CSG Translation Modules

In the developed Portuguese to Chinese machine translation system, PCT Assistente, an

integrated translation approach is adapted by combining the translation modules of TCT-based and CSG-based translation paradigms, hence to counterbalance the weakness of each module as discussed previously. Another consideration about the development of hybrid system involves analyzing the degree of how the two methods are being coupled, weakly or strongly. In a weak linkage manner, it is simpler to design the structure and to maintain the independence of involved systems. The advantage of this arrangement is that it provides an easier control and maintenance. Any improvement to individual component can prevent from distorting the normal operation of another. But the same structures and resources among the component may not be fully shared and utilized is the shortage of this approach. On the other hand, whatever which translation method applied in the development for an individual translation system or engine, there have been more or less involved with other methods to implement part of the process components, such as morphological analysis based on rules, POS tagging based on statistic model, phrasal recognition based on examples, etc. Therefore in this study, we carry out the integration from a whole point of view, that is, to integrate two individual translation systems with different translation mechanisms to improve the overall translation quality.

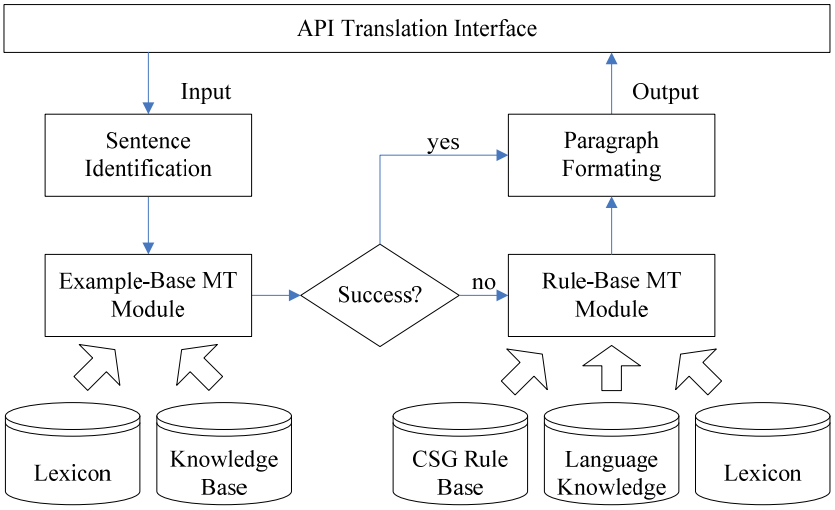

Figure 4 Hybrid approach translation system

Figure 4 demonstrates the architecture of the hybrid approach translation engine of the PCT Assistente system, where the example-based MT module is first called for translation. If translation can be successfully done, the result will be promptly returned to the user; otherwise, the rule-based MT module is used to do the translation and outputs the result as the target translation. From the empirical translation experiment for the administrative documents and legal statements, the overall translation accuracy improves around 8% with the new hybrid translation paradigm.

5. Network Based PCT Assistente

The previous version of PCT Assistente system is a PC-based (standalone) machine translation system, and its main disadvantage is that language resources such as new translation pairs, user parameters, and translation preferences cannot be shared among the translators within the same organization, department, even the same group, which limits users from cooperation in doing the translation work and exchanging their experiences. Thus, an extension of PCT Assistente is developed to support the network-based translation mechanism by combining different networked computing technologies and translation engines to achieve a better translation throughput in terms of translation quality and processing speed. The basic methodology of network-based PCT system is the deployment of cooperating and communicating techniques which make use of local and distributed information resources. The language resources of this new kind of MT system are intended to be shareable with other translation clients located in different user machines. Currently, the network-based MT system is deployed in Intranet applications, which in future, can be promoted to Internet applications that provides different translation services for the public as well as the organizations and enterprises.

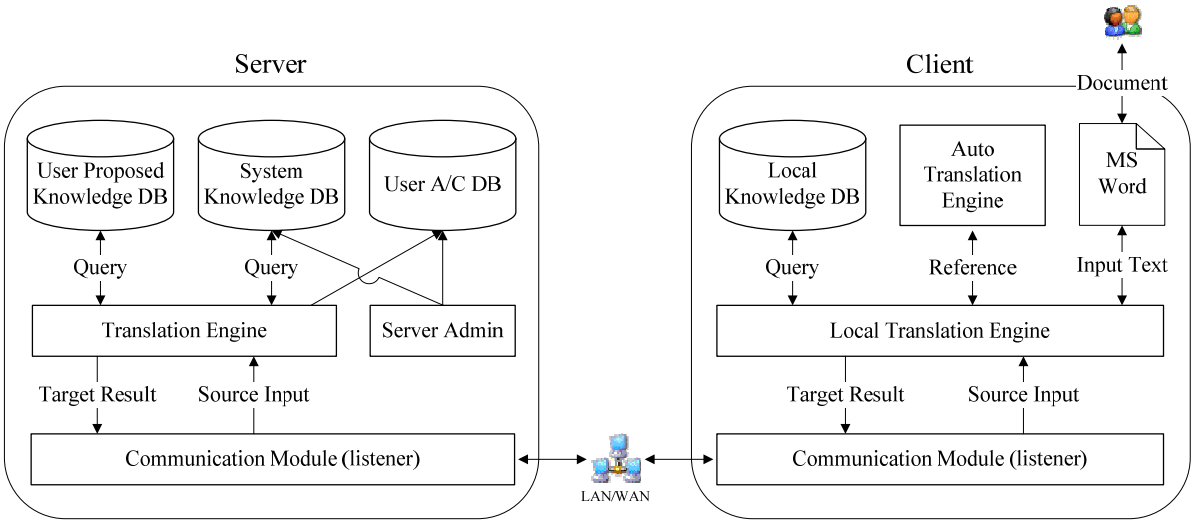

Figure 5 Translation flow in network-based PCT system

In view point of network computing, the best way to maximize the translation throughput is to distribute the translation process in different machines to share the load burden. As a consequence, in our design, both the client and server sites are assigned with a translation engines with different configurations. In the client machine, a light translation engine based on example-based MT approach is set, so that each client can contribute by making use of his/her translation resource to process a trail translation. The translation request is then triggered and sent to the server only if the translation fails locally. In the server site, the system is configured with full translation feature by using all available language resources to carry out the translation task for the input text. The flow of translation process and data is

illustrated in Figure 5, and the content of the system components for different sites are depicted by the left and right diagrams, with a communication channel of the Local Area Network (LAN) or Wide Area Network (WAN).

6. Conclusion

The translation technologies of Portuguese to Chinese machine translation system, PCT Assistente, are explored and reviewed in this paper. Different from other example-based and rule-based translation methods, our example-based translation paradigm is based on Translation Corresponding Tree (TCT) annotated structure, where the translation examples are being annotated with these structures and stored as the example base for the translation system. Moreover, the corresponding translation process based on this format of knowledge examples is also designed. While in the rule-based translation module, Constraint-based Synchronous Grammar (CSG) is adapted as the language grammar for analyzing the syntax of translation text, which is different from other language formalisms. CSG models languages in a synchronous approach, where both the source and target languages' relationships are being modeled and described by such synchronous grammar. The advantage of CSG is that once the source text is being analyzed, the corresponding target translation text can be inferred immediately, and this can reduce the information loss caused by pipelining different analytical phases in normal transfer-based MT systems. The translation architecture of PCT Assistente system based on the hybrid approach is also described in this paper through the discussion of the interrelatedness and interdependencies between approaches in different coupling levels. Finally, an extension of PCT Assistente system to the network-based translation by combining the network computing and translation technologies to maximize translation throughput in terms of cooperation and speed is discussed and presented.

References

[1] Aramaki, E., Kurohashi, S., Sato, S. et al. Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation. In Proceedings of MT Summit VIII pp.27-32. 2001.

[2] Bennett, W., Slocum, J. The LRC Machine Translation System. Computational Linguistics, Vol.11 (2-3), pp.111-121. 1985.

[3] Boitet, C., Zaharin, Y. Representation trees and string-tree correspondences. In Proceeding of COLING-88, Budapest pp.59-64. 1988.

[4] Brown, P., Pietra, S. A., Pietra, V. J. et al. The Mathematics of Machine Translation: Parameter Estimation. Computational Linguistics, Vol.19 (2), pp.263-311. 1993.

[5] Carl, M., Hansen, S. Linking Translation Memories with Example-Based Machine Translation. Proceedings of Machine Translation Summit VII pp.617-624. Singapore:

1999.

[6] Earley, J. An Efficient Context-Free Parsing Algorithm. PhD, Computer Science Department, Carnegie-Mellon University, 1968.

[7] Grishman, R. 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. 1994.

[8] Jain, R., Sinha, R. M., Jain, A. ANUBHARTI - Using Hybrid Example-Based Approach for Machine Translation. Proceedings of Symposium on Translation Support Systems (STRANS-2001) pp.86-102. Feb. 15-17, 2001.

[9] Knight, K., Marcu, D. Machine Translation in the Year 2004. Proceedings of the IEEE International Conference on Acourstics, Speech, and Signal Processing pp.45-50. March 18-23, 2005.

[10] Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. Cybernetics and Control Theory, Vol.10 (8), pp.707-710. 1966.

[11] Lewis, P., Stearns, R. Syntax-directed transduction. Journal of the Association for Computing Machinery, Vol.15 (3), pp.465-488. 1968.

[12] McTait, K. Translation Pattern Extraction and Recombination for Example-Based Machine Translation. PhD, Centre for Computational Linguistics, Department of Language Engineering, UMIST, 2001.

[13] Sato, S., Nagao, M: Toward Memory-Based Translation. In Proceeding of Coling-90, Vol.3 (1990) 247-252

[14] Tang, C. W., Wong, F., Leong, K. S. et al. Application of Translation Corresponding Tree (TCT) Annotation Schema for Chinese to Portuguese Machine Translation. Proceedings of the Tenth International Conference on Enhancement and Promotion of Computational Methods in Engineering and Science (EPMESC-X) pp.1105-1109. Sanya: Aug. 21-23, 2006.

[15] Tomita, M. Generalized LR Parsing. Carnegie Mellon University: Kluwer Academic Publishers, 1991.

[16] Watanabe, H., Kurohashi, S., Aramaki, E. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In Proceedings of The 18th International Conference on Computational Linguistics pp.906-912. 2000.

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

[18] Wong, F., Hu, D. C., Mao, Y. H. et al. A Flexible Example Annotation Schema: Translation Corresponding Tree Representation. In Proceedings of the 20th International Conference on Computational Linguistics pp.1079-1085. Switzerland, Geneva: 2004.

[19] Wong, F., Hu, D. C., Mao, Y. H. et al. Machine Translation Based on Constraint-Based Synchronous Grammar. In Proceedings of The Second International Joint Conference on Natural Language (IJCNLP-05): Vol.3651: LNAI Springer,(612-623). Jeju Island,

Republic of Korea: 2005.

[20] Wong, F., Dong, M. C., Hu, D. C. Machine Translation Based on Translation Corresponding Tree Structure. Tsinghua Science and Technology, Vol.11 (1), pp.25-31. February, 2006.

[21] Wong, F., Dong, M. C., Hu, D. C. Machine Translation by Parsing Constraint-Based Synchronous Grammar. Tsinghua Science and Technology, Vol.11 (2), pp.295-306. 2006.