PCT: Portuguese-Chinese Machine Translation Systems

Fai Wong and Sam Chao
Faculty of Science and Technology of University of Macau,
Av. Padre Tomás Pereira S.J., Taipa, Macao
{derekfw, lidiasc}@umac.mo

Abstract: In this paper, the machine translation (MT) tools that have been developed in the University of Macau, with a focus on the languages of Portuguese and Chinese, are presented. These systems, act as a translation tools, can be used to better manage the workflow of professional translators, as well as used as facilities for teaching courses. Machine translation, as the chosen theme in this paper, likes many other fields, it has its theoretical (methodological) and practical parts. For teaching Computer Science students, the MT systems can be used to illustrate problems in language analysis at different levels, especially the different methodologies to the development of a new MT system. While for translation students, it can be used to demonstrate how computer works, what MT can and cannot do, and how to master these tools in their translation workflow. This paper focuses in discussing the underlying architecture, components and methodologies of the developed MT systems, and its use to the teaching purpose.

1. Introduction

The automatic translation of languages from one to another through the use of computers is becoming more and more attractive, not only to the researchers or developers, but also to the language translators and language learners, due to the continue emerging of various machine translation systems in the market being used by many enterprises and educational entities for the purposes of translation and teaching. Especially for a multi-cultural society, like Macao, blending Asian and Western elements, the use of translation tools to overcome the communication barriers tends to be very significant. Macao is indeed a multiple language society, with three writing and four spoken languages. The writing languages include Chinese, Portuguese and English; the spoken languages consist of Mandarin, Cantonese, Portuguese, and English. There are many official documents which are written in either one of the languages, the use of MT tools to translate them into the other language within an acceptable timeframe provides a feasible solution to tackle the ever increasing demands of multilingual documents (Wong *et al.*, 2006*b*). However, the subject of machine translation (or computer aided translation) is not widely taught. The existence of fears, prejudice, and even the over expectation of MT may pose a negative attitude towards MT, due to the lack of knowledge and misinformation. Recently, there is a consensus in teaching machine translation course to different target students (Kenny and Way, 2001), notably students from Computer Science (Somers, 2003; Hearst, 2005), students from translation (Robichaud and L'Homme, 2003), and even students from translator trainee (Gaspari, 2001; Zerfass, 2004). In this paper, the MT tools that have been developed in the University of Macau, with a focus on the languages of Portuguese and Chinese, are presented (Wong *et al.*, 2006*b*). These systems, act as a translation tools, can be used to better manage the workflow of professional translators, as well as used as facilities for teaching courses. Machine translation, as the chosen theme in this paper, likes many other fields, it has its theoretical (methodological) and practical parts. For students from Computer Science, the use of MT can be used to illustrate problems in language analysis at different levels, especially the different methodologies to the development of a new MT system. On the other hand for translation students, it is necessary to understand what MT and related computer aided translation system can and, more important, cannot do. Translators need some

insight into how computer works, why it is difficulty, what kind of translation tasks that computer is appropriate for, what alternative tools are available and how to integrate these tools into their translation workflow.

The paper is organized as follows. The research background and the developed MT systems between Portuguese and Chinese are reviewed in section two and three. Section four discussed the underlying translation architecture and the methodologies. Section five presents a primitive tool that has been developed for the teaching purpose; follows by a conclusion to end this paper.

2. Research Background

The development of Portuguese-Chinese MT was initiated and deployed by University of Macau together with INESC-Macau and Tsinghua University in 1996 under the background Macau is a city that has two official languages, Chinese and Portuguese. Portuguese is considered as an official language in addition to Chinese, even after the handover. Due to the lack of people with bilingual proficiencies, many of the historical documents, official statements, news, technical specifications, operation manuals, proposals, etc. cannot be translated between Portuguese and Chinese without the use of computer translation tools within an acceptable timeframe. The development of practical MT systems focused on Portuguese and Chinese sounds to response to the huge demands of translation from the society of Macau. On the other hand, the solid fundamental knowledge of MT together with the developed translation tools can either be used, as valuable resources, for teaching courses in the university.

3. PCT Systems

In the research, several application systems related to the translation of languages between Portuguese and Chinese have been developed (Wong and Mao, 2003; Wong *et al.*, 2007). This consists of: 1) bidirectional Portuguese and Chinese electronic dictionary (PCTDict) equipped with instant translation and pronunciation functionalities; 2) bidirectional machine aided translation system (PCTAssist) which offers a workbench for translation professional, and 3) network based machine aided translation system (PCTNet) that extended from the desktop version of PCTAssist to the translation based on network infrastructure, which allows users within a working group or department be able to share their translation knowledge through a server.

3.1. PCTDict System – Electronic Dictionary

Dictionaries form the fundamental linguistic resources for language learners and language translators. Even in the development of natural language applications, like machine translation, dictionary plays a vital role in the application system. There are large numbers of dictionaries now available in the electronic forms, and some of them are even available on the internet. However, from the human use point of view, the design and creation of PCTDict system differentiates itself from others electronic dictionaries in the following aspects (Wong and Mao, 2003).

First, PCTDict is designed for Portuguese and Chinese languages. By reviewing the bilingual electronic dictionaries that are available in market, most of them are designed for native language to English and English to native language, while multilingual dictionaries are usually developed for languages that are from the same cognates. Secondly, most of the bilingual dictionaries between Chinese and other languages are basically designed for Chinese-speakers. These dictionaries usually do not concern too much the difficulties of the lookup process that may face to non Chinese-speakers. Chinese is non-alphabetic language. It usually relies on some input methods, e.g. Pinyin and the root radical methods (Karel *et al.*, 2002). This depends on user's ability to operate the writing system, and particularly, it can be a problem for non Chinese-speakers if they do not have any knowledge about the Chinese language.

However, these issues are addressed in the design of PCTDict system, which provides a comprehensive dictionary interface (Bilac *et al.*, 2002), that allows both Chinese and non Chinese-speakers easily consult the dictionary through the use of mouse-tracking computer technology (Wong *et al.*, 2003). The system will monitor the movement of the mouse cursor as well as the context underneath of the mouse pointer over text. Corresponding information of the word will be shown on a small popup window once the word is captured and identified. Another and most benefit of this function is that users are not necessary to stop their ongoing tasks and still can consult the meaning of unknown word while reading and editing. Thirdly, a morphological analyzer is integrated to identify an unknown word caused by its inflection and derivation. This arrangement is due to the empirical study of the characteristics of Portuguese, as they have a highly developed morphology. On the other hand, useful linguistic and grammatical information can also be derived from the inflected word and presented to user, in comprehensive way, through a simple interpreter. Moreover, a text-to-speech engine is adopted in PCTDict, which allows user listen to the pronunciation rather than interpret the linguistic representation. Currently, the speaking engine consists of three languages, Portuguese, Mandarin and Cantonese.

In brief, the PCTDict has been designed not only for language learners but also for the language translators. In which users are able to manipulate the dictionary with less interference in their work and the dictionary can be properly integrated in the translation workflow.

3.2. PCTAssist System – Translation System

The use of linguistic knowledge from electronic dictionary cannot fulfill the work of professional translator. The work of a professional translator usually involves the translation of documents that is frequently of a repetitive nature and involves the translation of relatively restricted subject fields. Therefore, how the translation work can be effectively managed and assisted by using the computer technologies to accelerate the translation life cycle is the main intention to the development of PCTAssist translation system, in our case, for Portuguese and Chinese.

The PCTAssist is a hybrid translation model that integrates MT modules with the management of translation knowledge based on translation memory (TM) technology to better manage the translation terminologies. In which, the MT module itself is a hybrid model either, which integrates example-based and rule-based translation paradigms to counterbalance the intrinsic weakness of these approaches by combining the strong features of another (Carl and Hansen, 1999; Jain *et al.*, 2001). In the example-based MT engine, translation examples are annotated under the schema of Translation Corresponding Tree (TCT) structure that acts as the basic knowledge for performing the translation task. While in the rule-based MT module, Constraint Synchronous Grammar (CSG) is adopted as the language formalism to recognize the syntactic structure of an input sentence and produce the corresponding translation in target language during the translation process (Wong *et al.*, 2006*a*; Wong *et al.*, 2006*b*). The structure of the integrated machine translation module is illustrated in Figure 1. The two MT engines are pipelined and ordered according to the confidence of the translation quality produced by different paradigms. In this case, the rule-based MT engine

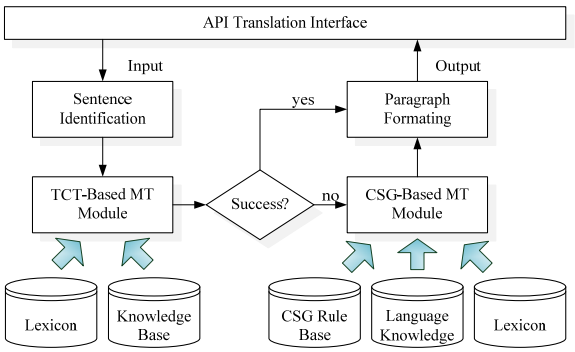

Figure 1. Hybrid approach translation architecture

is used as an universal translation engine and is arranged at the last stage to process the input sentence if it cannot be translated by any preceding MT engine that we believe the translation is more promising.

In practice, PCTAssist has been designed to provide a solution as the translation workbench that allows the translator to carry out his/her translation work in a familiar environment based on a specific workflow in accordance with the nature of translation life cycle. The whole translation process is carried out in the Microsoft Word (MS Word) application. User does not need to learn and manipulate a new translation system. All the related systems are embedded inside the Word application and are hidden from user.

3.3. PCTNet System - Network Based MT

The main disadvantage of standalone MT system is that language resources such as new translation terms, user parameters, and translation preferences cannot be shared among the professional translators within the same organization, department, even the same group, that limits users from cooperation in doing the translation work and exchanging their valuable resources. Which makes machine translation system do not fulfill the need of modern business's trend that characterized by increased competitiveness, reduced cost and increased productivity. According to this concern, the development of PCTNet provides an alternative translation solution based on network infrastructure by combining the networking technologies and MT system to achieve a better translation throughput in terms of translation quality and speed. The basic methodology of PCTNet is the deployment of cooperating and communicating techniques which make use of local and distributed information resources. The language resources of this kind of MT system are intended to be shareable with other translation clients located in different user machines through the central MT server. The most up to date translation data are promised to be delivered to the end users in real time. This, consequently, can keep the documents be consistently translated, in particular when documents (in a series with features and sections in common content) are translated by several people at different time and place.

From the operational point of view, PCTNet has the same translation interface and function as that of PCTAssist, and is embedded inside the MS Word application. While the PCTNet system maintains with an additional connection to the server where centralized the new translation terms and terminologies contributed by other professional translators together with extra parameters that help to identify the source and the status of the translation terms. But all of these configurations are totally transparent to the user.

4. Methodologies & Components

In this section, we focus on discussing the underlying methodologies and components of the rule-based translation engine, as this is main module that we are developing in our recent research. In transfer-based MT system, analysis of the structural deviations of the languages pairs is key to transforming one language into another. This analysis requires a large number of structural transformations, both grammatically and conceptually. The problems of syntactic complexity and word sense ambiguity have been the major obstacles to obtain promising translation results, in particular for the languages between Portuguese and Chinese as they come from different language families and topologies, the problem is more obvious. Therefore, in Portuguese-Chinese MT System, Constraint Synchronous Grammar (CSG) (Wong *et al.*, 2006*b*) is proposed, as language formalism, to model the structural relationships between multiple languages in parallel. During the translation process, the recognition of syntactic structure of source language and the transformation to the target language structure can be accomplished at

the same time using the same set of CSG productions.

4.1. Constraint Synchronous Grammar

Constraint Synchronous Grammar (CSG) is a variation of synchronous grammar to model the syntactic relationship between multiple texts in parallel based on the syntax of Context Free Grammar (CFG). In bilingual case, the CSG formalism consists of a set of production rules that describes the sentential patterns of the source text and target translation patterns in the form of:

$S \rightarrow NP_1\ PP\ NP_2\ VP^*\ NP_3$
{ [NP_1 VP *a* NP_3 NP_2 ; VP_{cat} = *vb1* &
PP = "把" & $VP_{s:sem} = NP_{1sem}$ &
$VP_{o:sem} = NP_{2sem}$ & $VP_{io:sem} = NP_{3sem}$]
[NP_1 VP NP_2 *em* NP_3 ; …]
}

In this production rule, it has two generative rules associated with the sentential pattern of the source *NP_1 PP NP_2 VP 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. 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. Their relationship is established by the given subscripts and the sequence is based on the target sentential pattern. In this model, semantic information is represented by feature descriptors (FD) which give additional flexibility in defining CSG rules for establishing agreements in syntactic and sub-categorization dependencies. Feature unification is performed during the parsing stage. If FDs of each lexical word or lexicon are compatible with each other, i.e. there are no conflicts on the value of all the attributes defined, unification succeeds. Then the production rule is applied and a new FD is constructed for the reduced symbol as a valid syntactic constituent. From the translation point of view, the syntactic structure of target sentence can be determined at the same time once the source sentence is successfully parsed.

4.2. Language Analysis Components

Reviewing the translation components of the MT system as depicted in Figure 2, besides the modules of Sentences Segmentation & Tagging and Document Binding, to identifying the boundaries of sentences for feeding into the (sentence-based) translation engine and reassembling the target document in the correct format after translation, the translation kernel consists of six major components: Word Segmenter, Morphological Analyzer, Part-Of-Speech (POS) Tagger, Syntax Parser (and Transformer), Morphological and Language Generator.

4.2.1. Word Segmentation

Unlike Western languages, Chinese is written in a continuous way without any delimiter to explicitly specify the word boundaries. Therefore, for the translation from Chinese to Portuguese, the first and most essential step is to find out the words boundaries of sentences. In our case, the strategy is to pick the *N*-best segmentation candidates based on probabilistic *N*-shortest-path model (Leong *et al*., 2006). The objective is to have a segmentation module with high recall rate that includes the correct segmentation in the candidate set. Then let the upcoming analysis processes determine and choose the correct sentence based on subsequent analytical information.

4.2.2. Part-of-Speech Disambiguation

In MT, knowing the *part-of-speech* (POS) of words is an important step in discovering the linguistic structure of sentences. This information facilitates the higher-level analysis, such as recognition of noun phrases and other syntactic patterns in text. This seems to be a simple problem, but it is actually a very difficult problem. It suffers

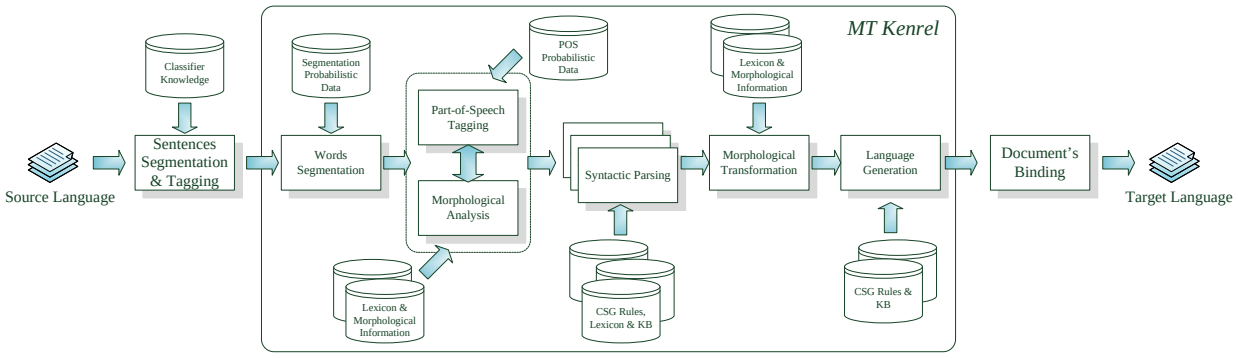

Figure 2 Different analytical processes in MT

from serious problems of part-of-speech ambiguities and unknown words due to variations of morphology. In our system, the tagger is constructed based on probabilistic model with an extension to interpolate the orthographic features of words for predicting the correct POS. Details of the work can be found in Wong *et al.*, 2004.

4.2.3. Morphological Identification

The function of morphological analysis in MT is to recover the canonical form of lexical item for Portuguese, as well as to analyze the associated linguistic information based on word morphemes. Actually, the analysis of compositional word is considered as syntactic analysis, since the inflectional morphology defining the possible variations on the root of word reflects the grammatical meaning and semantic information, such as gender, person, tense, mood, number, and grammatical category, etc. Besides the objective to conclude the conventions of morpheme variations hence to reduce the size of lexical dictionary from keeping the inflectional paradigms in full, another important intention is to resolve the unknown word problem by using the software analysis approach to dynamically recover the canonical form of lexicon and extract the embedded grammatical information. In PCT, different sets of knowledge formulated as rules are used to identify the morph-syntactic information of lexicon.

4.2.4. Sentence Generation

The generation process takes the responsibility to render the translation of the input sentence by referencing the set of generative target sentential patterns that were determined by the parser. Moreover, in order to ensure that the system generates perfectly the translation in target language grammatically, the unification of Functional Descriptors (FD) is employed as a validation operation for each node, which was constructed for each constituent node in the parsing stage. This includes the change of word morphology based on the set of grammatical agreements such as number, gender, tense, and categories of person, and the render of target sentence based on syntactic and semantic constraints, as examples shown in Figure 3.

$$FD_1 = \begin{bmatrix} \text{c-lexical} = \text{兩} \\ \text{category} = \text{q} \\ \text{p-lexical} = \text{dois} \\ \begin{bmatrix} \text{gender} = \text{male} \\ \text{number} = \text{plural} \end{bmatrix} \end{bmatrix} \qquad FD_2 = \begin{bmatrix} \text{c-lexical} = \text{支 鋼筆} \\ \text{category} = \text{NP4} \\ \text{p-lexical} = \text{caneta} \\ \begin{bmatrix} \text{gender} = \text{female} \\ \text{number} = \text{singular} \end{bmatrix} \end{bmatrix}$$

Figure 3. AVMs of words "*兩*" and "*支 鋼筆*"

Unification of FD_1 and FD_2 will fail as the gender and the number are different. In such a case, necessary conversions are performed so that FD_1 and FD_2 will be compatible with each other. Therefore, the generated result for "*兩支鋼筆*" is "*duas canetas*" (two pens).

4.3. Modulization of Translation Tasks

MT is a big engineer work, consisting of many analytical tasks and it is quite difficult to predefine (hardcode) the flow of analysis sequence of language at different levels during the system design time. Take the syntactic analysis as an

example, the recognition of a specific type of sentence constituents may be triggered multiple times and interlaced with the recognition of another type of constituent, due to the fact that constituents are often nested with each other, and constituent extracted can be arbitrarily deep in nested embedded clauses. In PCT, the flow of analytical tasks is driven by an external *action* file for guiding the translation process. This allows designer to flexibly experience and configure the processing behavior of the translation system. On the other hand, the system can also be used as a teaching platform or development base for students to experience the problems in language analysis at different level and further enhance a specific part of the translation system.

5. An CSG Based Translation Tool

The most critical section in a transfer-based MT system is the parsing and transformation of syntax for one language to another, and this is the philosophy behind the formalism of CSG. Thus, the simplest way to teach MT for students is to start from this kernel part which can simply present the main idea to students how translation is done through the use of computer based on synchronous grammar. In order to achieve this, we have developed a *simple translator* model based on CSG formalism, as shown in Figure 4. In this tool, we do not include any other analytical tasks for lexicons, e.g. morphological analysis and POS assignment. All that the students have to do is to construct CSG productions for bilingual text according to the formalism features of CSG to establish the translation relationship between source and target sentences, if necessary students are also allowed to add extra information in the target patterns such as modifiers and quantifiers for rendering the translation in target language. In execution, the syntax of the productions are checked and validated, and a parse table is generated. After that, the CSG translator is ready for parsing an input sentence and producing the corresponding translation in target language once the sentence is successfully recognized by the grammar, together with a parsed syntactic tree is displayed (Figure 4). As described in section 4.1, CSG is ready for use to parse and translate sentence based on the same set of productions. The main different between this tool and a real PCT translation system students experiencing is that this primitive tool does not provide any preprocessing modules for lexicons and is language independent. Students are free to design the translation for any pair of languages according to their preferences. Based to our teaching experience, we found that students can easily capture the central idea and technique of transfer-based MT system, and are stimulated to widen their concerns from the construction of syntactical relationship of bilingual text to the analysis of lexical morphology, POS assignment, and the problems of language ambiguous. This helps instructors to easily guide students to further explore the related analytical tasks or components in the development of a new MT system.

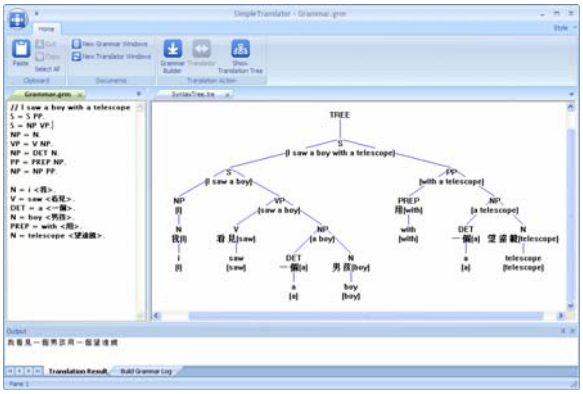

Figure 4 Main interface of CSG-based translator

6. Conclusion

This paper presents various machine translation systems and tool that have been developed in University of Macau with a focus on the languages of Portuguese and Chinese. These tools can be either used for the purpose of teaching machine translation course for students from different

majors, or used to better manage the workflow of professional translators.

From teaching point of view, using the CSG based translation tool as the primitive platform of the course appears to be a good solution to help students acquire the basic concepts of machine translation, as well as knowledge of different language analytical processes required to the development of a new MT.

Acknowledgements

The research work reported in this paper was partially supported by Research Committee of University of Macau under grant CATIVO: 5868 and it was also supported by Science and Technology Development Fund of Macau SAR under grant 041/2005/A.

Reference

[1] Bilac S., Baldwin T. and Tanaka H. (2002). “Bringing the Dictionary to the User: the FOKS system.” *The 19th International Conference on Computational Linguistics (COLING2002)*, pages 89-95.

[2] Carl, M., and S. Hansen. (1999). "Linking Translation Memories with Example-Based Machine Translation." *Machine Translation Summit VII*, Singapore, 617-624.

[3] Gaspari, F. (2001). "Teaching Machine Translation to Trainee Translators: a Survey of Their Knowledge and Opinions." *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela, Spain.

[4] Hearst, M. (2005). "Teaching Applied Natural Language Processing: Triumphs and Tribulations." *the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*, Michigan, USA, 1–8.

[5] Jain, R., and A. J. R.M.K.Sinha. (2001). "ANUBHARTI - Using Hybrid Example-Based Approach for Machine Translation." *Symposium on Translation Support Systems (STRANS-2001)*, 86-102.

[6] Kenny, D., and A. Way. (2001). "Teaching Machine Translation & Translation Technology: A Contrastive Study." *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela, Spain, 13-17.

[7] Karel S. and Robert W. (2002). “Using Affordances in Electronic Chinese/English Dictionaries for Non Chinese-Speakers”, *Technical Reports, Cognitive Science*, Carleton University.

[8] Leong, K. S., F. Wong, C. W. Tang, and M. C. Dong. (2006). "CSAT: A Chinese Segmentation and Tagging Module Based on the Interpolated Probabilistic Model." *The Tenth International Conference on Enhancement and Promotion of Computational Methods in Engineering and Science (EPMESC-X)* Sanya Hainan, China, 1092-1098.

[9] Robichaud, B., and M.-C. L’Homme. (2003). "Teaching the automation of the translation process to future translators." *MT Summit IX Workshop on Teaching Translation Technologies and Tools*, New Orleans, USA, 27-34.

[10] Somers, H. (2003). "Prolog models of classical approaches to MT." *MT Summit IX Workshop on Teaching Translation Technologies and Tools*, New Orleans, USA, 35-43.

[11] Wong, F., and Y. H. Mao. (2003). "Framework of Electronic Dictionary System for Chinese and Romance Languages." *Automatique des Langues (TAL)*, 44(2), 225-245.

[12] Wong, F., M. C. Dong, Y. P. Li, and Y. H. Mao. (2003). "Semantic & Morphological Analysis and Mouse Tracking Technology in Translation Tool." *Computational Methods in Engineering and Science (EPMESC IX)*, Macao, 525-532.

[13] Wong, F., S. Chao, M. C. Dong, and Y. H. Mao. (2004). "Interpolated Probabilistic Tagging Model Optimized with Genetic Algorithm." *Third International Conference on Machine Learning and Cybernetics*, Shanghai, China, 2569-2574.

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

[15] Wong, F., M. C. Dong, and D. C. Hu. (2006). "Machine Translation by Parsing Constraint-Based Synchronous Grammar." *Tsinghua Science and Technology*, 11(3), 295-306.

[16] Wong, F., M. C. Dong, C. W. Tang, and F. Oliveira. (2007). "Translation Technologies of Portuguese to Chinese Machine Translation System: PCT Assistente." *The 4th Chinese Digitization Forum (CDF)*, Macau SAR, China.

[17] Zerfass, A. (2004). "Teaching Translation Tools over the Web." *The Twentieth International Conference On Computational Linguistics (COLING-2004): Second International Workshop on Language Resources for Translation Work, Research & Training*, Geneva, Switzerland, 61-70.