Hybrid Machine Aided Translation System based on Constraint Synchronous Grammar and Translation Corresponding Tree

Fai Wong, Francisco Oliveira, Yiping Li
Faculty of Science and Technology, University of Macau, Macao
Email: {derekfw, olifran, ypli}@umac.mo

***Abstract*—As the demand of translating large volume of material between Portuguese and Chinese is increasing rapidly nowadays, especially in the city of Macau, the translation work becomes impractical without the support of effective tools. In order to fulfill this gap and build up a translation workbench environment for translators, a Machine Aided Translation System between these languages, PCTAssist, is introduced. It is a hybrid system that applies not only Translation Memory technology but also Machine Translation methodologies, including the annotation schema of Translation Corresponding Tree (TCT) in the representation of bilingual examples, and the language formalism Constraint Synchronous Grammar (CSG) in analyzing the syntactic structure between the languages to accomplish the translation task.**

***Index Terms*—Machine Translation, Constraint Synchronous Grammar, Translation Corresponding Tree**

I. Introduction

The advancement of computer technologies has made many changes in the daily life. As more documents have to be translated daily, human translation becomes impractical without the help of computer tools. These include the use of electronic dictionaries, terminology corpora, translation memory, and automatic translation. They are often combined together as a whole in order to improve translator's daily work, which is classified as Automatic Machine Translation (MT) systems and Machine Aided or Computer Assisted Translation (CAT) systems. Automatic MT systems generate the translation based on the information in the Knowledge Base without human intervention. On the other hand, CAT systems first produce a preliminary translation result, and based on the quality of the translation, translators make necessary changes afterwards.

There are a huge number of systems available in the market nowadays. They differ in the supported file formats, languages, operating systems, functions provided, and price. A list of these systems can be found in [1]. Moreover, different designs to MT have been proposed in the literature. Rule based MT [2] approach is based on a set of linguistic grammar rules for handling the translation, which can be categorized as Direct, Transfer based, or Interlingua based approaches. They differ in the definition of the linguistic context, the knowledge used, and the number of stages needed for translating a sentence. Direct approaches only handle word by word translation, and they ignore all the syntactic and semantic information. There are three modules in Transfer based approach: analyzer module analyzes the source text and converts it into an intermediary representation; transfer module maps the representation into a target language structure based on a set of conversion rules; generation module synthesizes the transferred representation into the corresponding target language. In Interlingua approach, the transfer module is not considered. Example based MT [3][4] analyzes different pieces of bilingual examples stored in parallel corpora for generating the translation. However, it often depends on the quality of the examples and the similarity function applied. As there are more digitized resources available nowadays, Statistic based approaches [5][6] become a new research trend. These approaches take into consideration of probabilities estimated between the translation of words and the ordering of the sentences extracted from the corpora. The accuracy is often highly dependent with the information of the digitized resources.

Each of these approaches has its strength and weakness in application to the development of MT alone. The combination of these methods leads to a hybrid system in order to avoid the intrinsic impediments of different translation methods [7][8].

Although it is so easy to get information and look for these tools, there isn't any practical and commercial Machine Aided Translation System especially developed for Portuguese and Chinese languages. In particular, the use of these two languages plays an important role in the city of Macau, which is considered as official languages.

In this paper, a hybrid Machine Aided Portuguese Chinese system, PCTAssist, is presented. The system is targeted for Portuguese and Chinese, and provides a helpful translation tool for translators who need to work with these languages. Moreover, the system is designed to integrate the advantages of Rule based and Example based approaches and to get rid of their disadvantages.

Since Portuguese and Chinese come from different language families, they are very different in terms of writing and grammar. Table 1 shows some common non-standard linguistic relationships between them.

doi:10.4304/jcp.7.2.309-316

TABLE I.
NON-STANDARD LINGUISTIC PHENOMENA BETWEEN PORTUGUESE AND CHINESE

Linguistic Phenomenon	Bilingual Example	English Translation
Discontinuity Relationships	Ela emprestou ao Pedro uma caneta 她 把 一支鋼筆 借給了 佩德羅	She lent one pen to Peter
Crossing Relationships	Carro bonito 漂亮汽車	Beautiful car
Words that are vanished in the target translation	Universidade de Macau 澳門大學	University of Macau
Words that should appear in the target translation	A escola fundou-se em 1949 學校成立於 1949 年	The school was funded in 1949

All of these examples can be easily found in Portuguese and Chinese translations. As an example, in terms of discontinuity relationships, the verb “emprestar” (to lend) is composed of “把” and “借給了” (to lend) in the Chinese translation. For crossing relationships, many words in Portuguese may have a different sequence order in Chinese. On the other hand, there are many words that should appear or vanish in the target sentence to guarantee the correctness of the translation. For example, the preposition “的” (of) should not appear in the target translation for “Universidade de Macau” (University of Macau), while the word “年” (year) should appear even if the source sentence does not have any word related with the meaning of “ano” (year).

As a result, in order to better handle the translation and describe the relationships between Portuguese and Chinese language, Translation Corresponding Tree (TCT) [9][10] and Constraint Synchronous Grammar (CSG) [11][12] are proposed as the underlying technologies of PCTAssist. In Example based paradigm, TCT serves as an alternative representation structure for facilitating the searching and matching of the fragments between bilingual texts. Furthermore, TCT has an intrinsic property to deal with bilingual examples that are not literally translated [9]. In Rule based paradigm, CSG is used as the language formalism for modeling the relationship between the source and target languages simultaneously based on a set of constraints defined.

This paper is organized as follows. Sections 2 and 3 discuss about the underlying technologies of TCT and CSG. Section 4 presents the overall design, architecture, and the main functions of PCTAssist. Evaluations are detailed in Section 5, and a conclusion is followed in section 6.

II. TRANSLATION CORRESPONDING TREE TRANSLATION PARADIGM

In structural Example-based MT systems, relationships between bilingual sentences stored in the Knowledge Base are established at structural level, and examples are normally annotated with dependency structures [13]. Moreover, these systems need two linguistic parsers, and they require all the bilingual examples to be parallel translations with each other or to have close syntactic structures [14]. However, for languages like Portuguese and Chinese, from different language families and topologies, they do have very different syntactic structures, and do not have explicit corresponding constituents. Moreover, robust parsers for both languages are not always available, and ambiguities management in two parsers has to consider the combination of overall ambiguities [15].

As a result, PCTAssist considered the use of Translation Corresponding Tree (TCT) structure as the annotation schema for describing bilingual examples in the Knowledge Base. In TCT schema, only one parser is required, and a single syntactic structure is used to model the relationship between two languages.

A. Representation of Translation Corresponding Tree

TCT structure is an extension of the structural string tree correspondence representation to the bilingual case. Only one syntactic tree is required for modeling the source and target language string. TCT structure uses triple sequence intervals SNODE(n), STREE(n), and TTREE(n) which are encoded for each node in the tree to represent the relationships between the structure of the source sentence and the substrings from both the source and target sentences. Each set is made up of three interrelated correspondences to denote: the head word of sub-structure; the dominated substrings of the sub-structure; and the substrings in target language with the same sub-structure. The associated substrings may be discontinuous in all cases. Fig. 1 illustrates a bilingual example annotated by TCT schema. The equivalent translation of different constituent units (from lexical to sentential level) are retrieved by referencing sequence intervals.

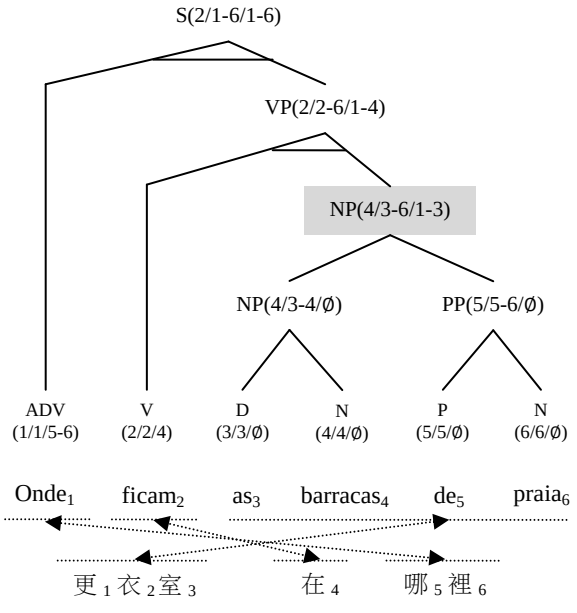


Figure 1. TCT representation example for “Onde ficam as barracas de praia/更衣室在哪裡”

As an example, *NP(4/3-6/1-3)* in grey color has the following meanings: numberings denote the index of the word in the sentences; index 4 denotes the head and source word of the sub-structure; indices 3-6 denote the sequence of words in the source sentence, from the third to the sixth word; and indices 1-3 denote the sequence of words in the target sentence, from the first to the third word.

Horizontal bars in the nodes represent the inversion of the translation fragments in its immediate sub-trees. The horizontal bar in *VP* indicates that the source sentence nodes "ficam" (are) and "as barracas de praia" (locker rooms) should have an inverted sequence order in the corresponding translation, in this case "更衣室" (locker rooms) comes before "在" (are) in the target sentence.

TCT annotation schema captures complex structural relationships between Portuguese and Chinese easily by defining a simple constraint for each constituent in the tree to restrict the sequence order of the translation. Immediate sub-trees are only allowed to cross in the inverted order.

B. Translation based on Translation Corresponding Tree

The translation process based on TCT is shown in Fig. 2. The input sentence is first preprocessed by the morphological analyzer in restoring the original format of the words, and then by the Part-of-Speech tagger in assigning syntactic tags. Based on the analyzed output, a Portuguese parser is used in generating the corresponding syntactic tree. This structure is then decomposed into sub-graphs and the system searches for related TCT structures stored in the Knowledge Base by considering the following criteria: 1) structural constituents must have similar structures; 2) grammatical categories of the root nodes and the dominated nodes must be the same as those in the source sentence's nodes. Furthermore, content words of the root node are considered to estimate the similarity between the examples and the source structure such that the reconstruction of the target structure can be accomplished based on the best TCT structures (or sub-structures) selected. For unmatched terminal nodes, the translation is obtained from the bilingual dictionary. On the other hand, if there are more than one TCT structure or sub-structure found, the system evaluates the distance between the example candidates and the source sentence based on the edit distance function [16]. Finally, the translation of the source sentence is generated by traversing through the target tree under the control of syntax constraints.

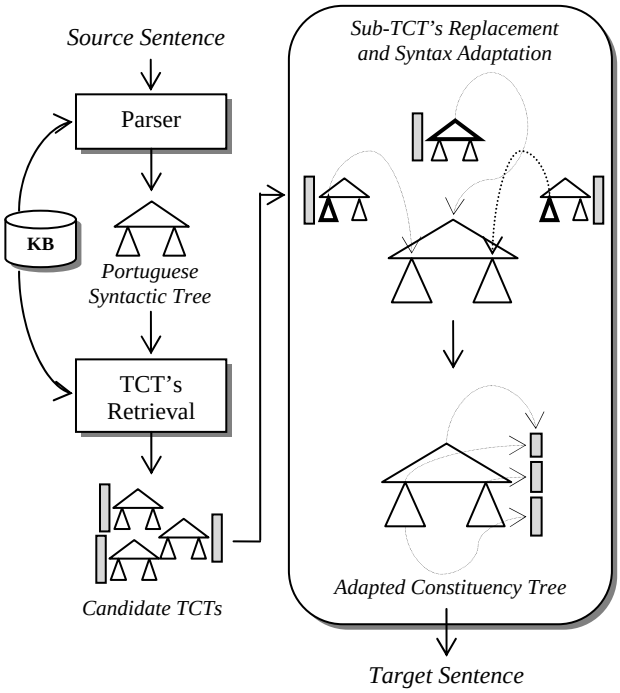


Figure 2. Translation process based on TCT examples stored in the Knowledge Base

III. CONSTRAINT SYNCHRONOUS GRAMMAR TRANSLATION PARADIGM

In Rule-based MT systems, the analysis, transformation, and generation from one language into another require a deep understanding of both languages and enough knowledge to ensure the quality of the translation for any domain. However, it is not easy to achieve the goal due to the syntactic and word sense ambiguity of the languages. PCTAssist applied Constraint Synchronous Grammar (CSG) [11] in expressing syntactic relationships between the source and one or more target sentential patterns, and the selection of the most suitable target is based on the feature constraints defined for each grammar rule.

A. Definition of Constraint Synchronous Grammar

Constraint Synchronous Grammar is based on the formalism of Context Free Grammar to the case of synchronous. In CSG formalism, it consists of a set of production rules that describes the sentential patterns of the source text and target translation patterns. In CSG, every production rule is in the form of the example below.

$$
\begin{aligned}
&S = NP_1\ VP^*\ NP_2\ PP\ NP_3\ \{ \\
&\quad [NP_1\ VP^1\ NP_3\ VP^2\ NP_2;\ C_1] \\
&\quad [NP_1\ VP\ NP_3\ NP_2;\ C_2] \\
&\qquad \} \\
&C_1 = \{\ VP_{\text{ category}} = vb1, \\
&\qquad\quad VP_{\text{ sense subject}} = NP_{1\text{ sense}}, \\
&\qquad\quad VP_{\text{ sense indirect object}} = NP_{2\text{ sense}}, \\
&\qquad\quad VP_{\text{ sense object}} = NP_{3\text{ sense}}\ \} \\
&C_2 = \{\ VP_{\text{ sense subject}} = NP_{1\text{ sense}}, \\
&\qquad\quad VP_{\text{ sense indirect object}} = NP_{2\text{ sense}}\ \}
\end{aligned}
\qquad (1)
$$

In production rule (1), 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 NP₂ PP NP₃* is associated with the target pattern *NP₁ VP¹ NP₃ VP² NP₂*.

Their relationship is established by the given subscripts and the sequence is based on the target sentential pattern. As an example, in the first generative rule, although the first *NP* in the source pattern corresponds to the first *NP* in the target one, 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.

Non-standard linguistic phenomenon can be handled by defining appropriated CSG rules. The ordering of the constituents is modeled easily by using the subscripts and the sequence defined in CSG production rule. The discontinuity between words in different languages is solved by defining non-terminal symbols that appear in the source but not the target pattern or vice-versa. As an example, in the first target sentential pattern in production (1), two verbs (VP^1 and VP^2) are associated with the source one. Similarly, the consideration of constituents that are disappeared or shown in the target syntactical pattern is handled in a similar way.

Every lexical word or constituent has its semantic information represented in feature descriptors (FD) for the control of constraints checking at the parsing stage. The syntax of FD is in the type of "attribute = value", and the value can be either an atomic symbol or recursively a FD. A set of FDs related to a single entry is encoded in Attribute Value Matrices (AVM). Fig. 3 shows a simplified AVM of the adjective "rica" (rich) and verb "dar" (to give).

The AVM of "rica" (rich) contains its POS information and two different meanings ("豐富的" and "有錢的"), each with a different sense. On the other hand, in the AVM of "dar" (to give), besides its POS information and meaning, it also contains the verb type, sense of the subject and object, verb category, tense, person, and number. These values may change during the unification process.

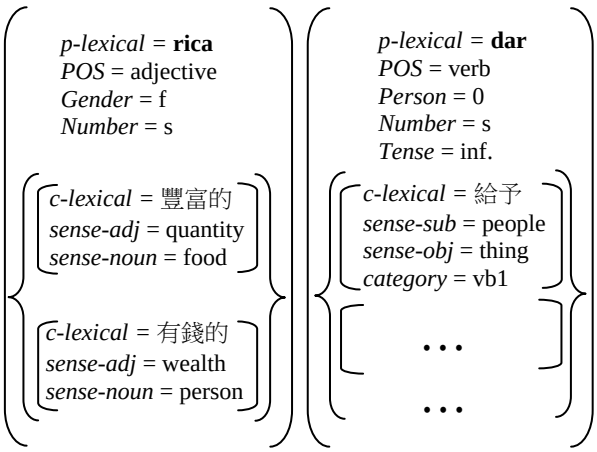


Figure 3. Attribute Value Matrices of "rica" (rich) and "dar" (give)

B. Parsing and Translation based on Constraint Synchronous Grammar

Extended CSG rules are parsed by a modified version of generalized LR algorithm [17], a shift-reduce approach based on an extended LR parsing table. Besides having the actions to be accomplished (shift, reduce, accept), and the state of the parser at different stages of parsing, the table is extended by taking feature's constraints and target rules into consideration. In other words, as the parser identifies CSG productions through the normal shift actions, it checks the associated constraints to determine if the current reduce action is valid or not.

During the construction of a valid parse tree, since the proposed model heavily relies on the synchronous relationship between the source and target languages, different types of ambiguity may occur in the parsing process. First, the system needs to disambiguate different possible target translations with the same source tree structure. Suppose that the sentence to be analyzed is "comida rica" (rich food). Based on (2) and the AVM of "rica", since the constraint requires that the sense of the being modified noun of the adjective to be the same as the sense of noun, the meaning of "豐富的" will be selected instead of "有錢的".

$$\begin{aligned} &NP = NP_1\ AP_1\ \{\ [AP_1\ NP_1;\ C_1] \\ &\qquad\qquad \cdots\ \} \\ &C_1 = \{\ AP_{1\ \text{sense noun}} = NP_{1\ \text{sense}}\ \} \end{aligned} \tag{2}$$

Second, depending on the sense of the current word or phrase, the target sequential order may vary. If the sentence to be analyzed is "comida rica em proteínas" (protein rich food), ambiguous target sentential patterns can be easily removed based on (3). In this case, since the sense of "comida rica" matches the sense of "proteínas", the target pattern *NP₂ NP₁* is associated with the tree structure of "comida rica em proteínas" after reduce action is performed.

$$\begin{aligned} &NP = NP_1\ P_1\ NP_2\ \ \{\ [NP_2\ NP_1;\ C_1] \\ &\qquad\qquad\qquad\qquad [NP_1\ P_1\ NP_2;\ C_2] \\ &\qquad\qquad\qquad\qquad \cdots\ \} \\ &C_1 = \{\ P_1 = \text{“em”}, \\ &\qquad NP_{1\ \text{sense}} = NP_{2\ \text{sense}}\ \} \\ &C_2 = \{\ P_1 = \text{“em”}, \\ &\qquad NP_{1\ \text{sense}} = \text{object}, \\ &\qquad NP_{2\ \text{sense}} = \text{location}\ \ \} \end{aligned} \tag{3}$$

Finally, the system may often generate more than one valid tree structure of the same source sentence. As an example, the sentence "ricos pontos turísticos" (rich tourism spots) has two different tree structures, [[ricos/Adj pontos/N]/NP turísticos/Adj]/NP, and [ricos/Adj [pontos/N turísticos/Adj]/NP]/NP respectively. This process effectively removes ambiguities often generated during the traditional parsing process.

Unlike transfer-based MT architectures, instead of carrying out the translation process in sequence by the

analyzer, transfer, and the generation module, in the description of parsing Constraint Synchronous Grammar, the parsing process can be viewed as to translate the sentence in one stage [11]. Similar to the translation process of TCT, the input sentence is first preprocessed, and then, it is parsed by referencing source sentential patterns in the CSG productions, and feature constraints are used to remove ambiguities, determine and build the target derivation sequence. As a result, translation is being generated during the parsing process of the source sentence.

IV. Design of PCTAssist

The proposed technologies especially targeted for modeling Portuguese and Chinese language have been applied in PCTAssist. The design, architecture, and main functions of the system are detailed in the following sections.

A. Translation Life Cycle of PCTAssist

The workflow of PCTAssist is guided by a translation life cycle, as shown in Fig. 4. Furthermore, the design is based on the idea of Case-based Reasoning [18], which is adapted to the nature of the translation work.

When the document is given to the system for translation, it is first analyzed, and the sentences are compared with the past TCT examples stored in the Knowledge Base. All the matched examples are retrieved, scores are assigned, and a list is generated according to the degree of similarity. The recalled list of examples is used to facilitate the translation by referencing the translations of the analog examples. In order to adapt to the new translation, suitable TCT fragments are extracted and recombined in generating the target language. If none of the generated results are higher than the predefined threshold, PCTAssist automatically generates the translation by parsing CSG rules in the Knowledge Base. The proposed translation is then presented to the user for further revision and correction if necessary. This allows the user to interfere the translation process. After revision,

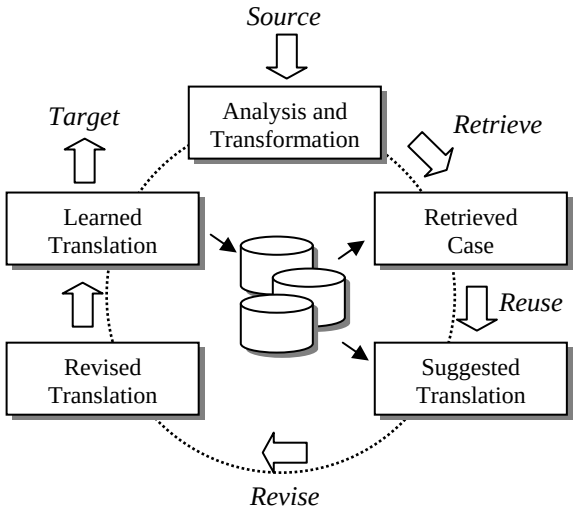


Figure 4. Translation Life Cycle

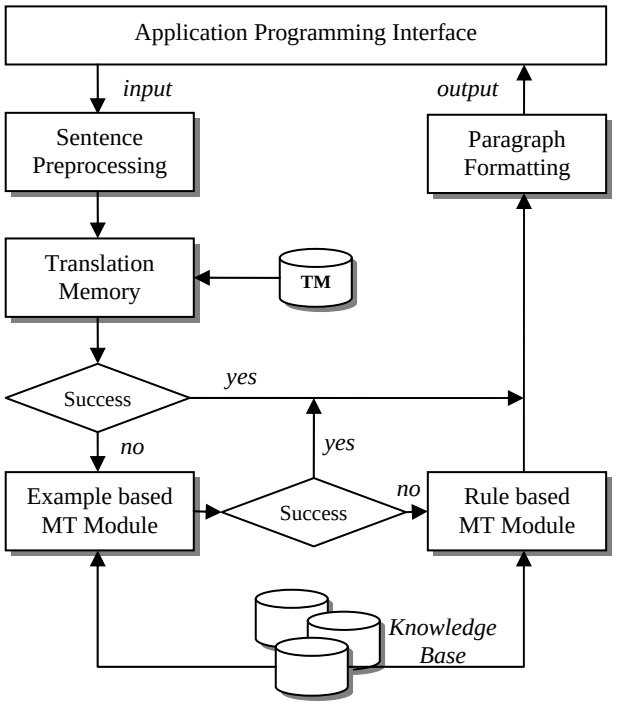


Figure 5. High Level Architecture of PCTAssist

valuable new translations can be retained and stored in the Knowledge Base for future use. This functional behavior allows PCTAssist to accumulate more translation knowledge during the translation work. Gradually, as the system gains more knowledge, it can achieve more promising translation results.

B. Architecture of PCTAssist

The overall architecture of PCTAssist is shown in Fig. 5. Through the application programming interface, sentences are first preprocessed and the system searches for any candidate words or phrases in the translation memory as long as they are higher than the predefined threshold. These are returned to the users for final selection. If there is a failure, Example based MT module based on TCT representation schema is called for the construction of a possible TCT tree based on sub-tree's replacement and syntax adaptation. If this process fails, then Rule based MT module based on CSG is used to parse the source sentence and generate the target translation.

In the proposed design, translation engines are kept independent of each other for easier management. Any changes in one component will not produce side effects in the normal operation of the other. Although the resources cannot be fully shared, many of the translation methodologies are applied in the same way on different engines, such as Portuguese morphological analysis, Part-of-Speech tagging, word sense disambiguation, phrasal recognition, Chinese word segmentation, etc.

C. Main Functions Provided in PCTAssist

PCTAssist provides a simple interface to do the translation, which is embedded in Microsoft Word environment.

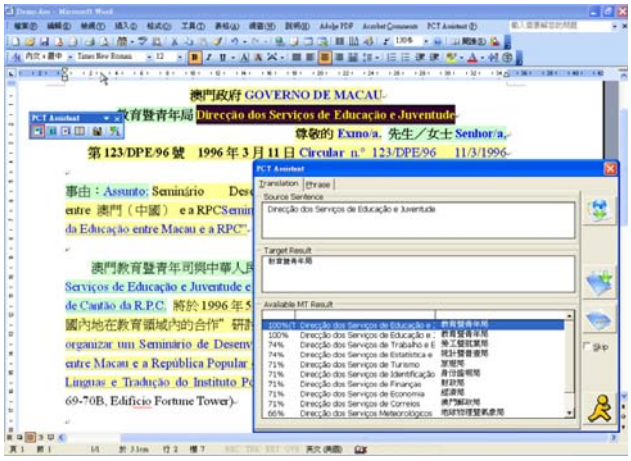


Figure 6. Screenshot of PCTAssist

The communication between PCTAssist and the document processing system is through the API, and all the logic is hidden from the users. The linguistic knowledge of PCTAssist contains more than 100,000 entries, and a screenshot of the system is shown in Fig. 6.

Users are allowed to create and select suitable databases for handling the translation task according to different domains. Depending on the type of document to be translated, it is always useful to load technical terms and specific knowledge in order to get better translated results. As a result, a simple interface is provided in PCTAssist to create, manage, and switch between different domain specific databases.

The system provides two types of translation. In interactive translation, based on the selected sentences, the system first analyzes the sentences boundaries, removes all the formatting information, and for each identified sentence, they are translated one by one through the system by an interface provided. On the other hand, in automatic translation, PCTAssist performs a preliminary translation of the whole document or selected sentences without any interaction between the system and the user. A progress bar appears and the translation that has the highest score calculated by the System for each source sentence is replaced in the document.

Once the translation task is finished, source sentences are replaced by the translation. In order to compare the translated results with the source ones, PCTAssist provides a function to highlight each bilingual pair with different colors to distinguish the boundaries. This allows users to post-edit the translations and do not need to switch between the translated and the original document.

After the revision, users may retain all the verified bilingual pairs into the Knowledge Base for future use. This functional behavior increases the data in the Knowledge Base, and gradually, the system can create user's own personal translation style, better keep the consistency in the translation, and achieve more promising translation results.

V. Evaluation and Discussion

In order to evaluate the applicability of PCTAssist, different experiments in terms of translation efficiency and quality are carried out. The translation quality is measured by human assessments, and three automatic evaluation metrics, including NIST [19], BLEU [20], and METEOR [21]. The average value of three human evaluated results is considered in human assessments, and the quality of the translated sentences is classified as Good, Acceptable, and Bad. The size of the MT's knowledge applied for the evaluation is shown in Table 2.

The first experiment evaluates the efficiency of the system. A test suite (TS) of 2070 sentences is extracted from government web pages [22], with an average of 19 words. The total translation time is 484.26 seconds, and the average translation time per sentence is 0.24 seconds.

Different sets of data and domain are considered in the evaluation of the translation quality. Test suites 1 and 2 contain sentences randomly extracted from administrative government documents [23], and test suite 3 includes sentences randomly downloaded online from government web pages [22]. The first test suite contains 200 sentences which have been tuned for the domain specified before the translation task is accomplished. Test suite 2 also contains 200 sentences but some of them are closed while others are unseen from the same domain. Since MT systems are impossible to handle the translation well in every domain, the purpose of the last test suite is to evaluate the quality of the system when the knowledge is out of domain. Automatic and Human evaluation results are shown in Tables 3 and 4.

TABLE II.
Resources Used For Evaluation Purpose

Resources	Size
Bilingual Dictionary	110,000 entries
Morphological and POS correction Rules	330 entries
Specific patterns	76 rules
TCT Structures	500 structures
CSG rules for Full parsing	785 rules

TABLE III.
Automatic Evaluation Results

TS	Domain	Sen-tence	Avg. Word Length	NIST	BLEU	METE OR
1	Close	200	13.89	4.9669	0.3439	0.5897
2	Mix	200	14.85	4.9598	0.2921	0.5365
3	Open	500	16.09	5.4956	0.2486	0.5487

TABLE IV.
Human Assessment Results

TS	Domain	Sen-tence	Avg. Word Length	Good	Accept able	Bad
1	Close	200	13.89	47%	40%	13%
2	Mix	200	14.85	37%	35.5%	27.5%
3	Open	500	16.09	14.8%	28.3%	56.9%

The scores are directly affected by several factors. Since MT involves a chain of processes in the translation process (segmentation, morphological analysis, POS tagging, TCT searching, CSG parsing, etc), if one of them gives an incorrect result, no doubt, an incorrect translation is generated. On the other hand, BLEU and NIST metrics do not have the same effect as human assessment. Since they are based on n-gram co-occurrence precision, it may generate low scores even if the translated sentence is correct when compared with the reference translations that use different synonyms. In order to overcome this issue, METEOR was proposed by introducing more factors like synonyms and stemming.

In terms of human assessment, when PCTAssist is tested in a close domain, it gets an accuracy of 87% for Good and Acceptable cases. By adding some sentences of open domain to the close one actually decreases at a rate of 14.5%. When most of the sentences in the test suite 3 are out of domain, the accuracy is 43.1% for Good and Acceptable cases. These results show that the translation quality of PCTAssist decreases drastically when more out-of-domain documents are used for translation, which is a common issue in all the translation systems. This can be compensated by retaining new bilingual pairs to the system, adding new TCT structures and CSG rules, applying machine learning approaches in getting more linguistic information.

Besides the entries stored in the Translation Memory, the system is highly dependent on TCT structures and CSG rules stored in the Knowledge Base. Since it is always difficult to have bilingual data aligned to each other between structurally different languages, in this case, Portuguese and Chinese, automatic acquisition and construction of these structures and rules are always difficult. As a result, our construction process is done semi-automatically. First, we make use of available skeletal bracketing structures and off-the-shelf machine learning tools in the extraction and conversion of Context Free Grammar rules, and then these are later extended manually by linguistics. Since our target is to provide a translation environment to the users in the daily work, the help and verification of linguistics can effectively ensure the quality of the entries in the Knowledge Base as well as the translation quality.

VI. Conclusion

In order to fulfill the high demand of translation work in Macau, a practical and hybrid Machine Aided Translation System, PCTAssist is developed. In this paper, the overall design, architecture, functions, and translation methodologies are presented and reviewed in details. The system applies Translation Corresponding Tree structure for annotating bilingual examples, and Constraint Synchronous Grammar as the formalism for analyzing the syntax of bilingual texts. Both paradigms are specially designed to handle languages which are structurally different, in our case, Portuguese and Chinese. Moreover, they can easily express non-standard linguistic phenomena, including the syntactic sequence order, crossing relationships, discontinued constituents, and words that may vanish or appear in the other language.

There are still rooms for improvement in the current system. Since the system is installed in one computer, if the translation task is done by a group of translators, it is always useful to share the translation knowledge and vocabulary between them in order to accelerate the translation task and make the translation results more consistent. The architecture of the system can be extended to a Client-Server model in which a centralized Knowledge Base is resided in the server side and clients can access it through the network.

Acknowledgment

The authors are grateful to Prof. Rui P. Martins for his help and support. This work was partially supported by the Research Committee of University of Macau under grant UL019/09-Y2/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

[1] Computer-assisted translation, http://en.wikipedia.org/wiki/Computer-assisted_translation

[2] W.S. Bennett, J. Slocum, "The LRC Machine Translation System," *Computational Linguistics*, vol. 11, nos. 2-3, 1985, pp. 111-121.

[3] R.D. Brown, "Example-Based machine translation in the Pangloss system," *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, 1996, pp. 169-174.

[4] K. McTait, "Translation Pattern Extraction and Recombination for Example-Based Machine Translation," *PhD Thesis*, Centre for Computational Linguistics, Department of Language Engineering, UMIST, 1996.

[5] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin, "A Statistical Approach to Machine Translation," *Computational Linguistics*, vol. 16, no. 2, 1990, pp. 79-85.

[6] A. Lopez, "Statistical Machine Translation," *ACM Computing Surveys*, vol. 40, no. 3, article 8, 2008.

[7] M. Carl, S. Hansen, "Linking Translation Memories with Example-Based Machine Translation," *Proceedings of Machine Translation Summit VII*, Singapore, 1999, pp. 617-624.

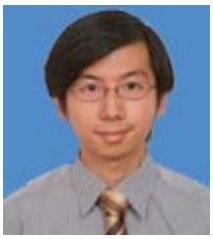
[8] R. Jain, R.M. Sinha, A. Jain, "ANUBHARTI - Using Hybrid Example-Based Approach for Machine Translation," *Proceedings of Symposium on Translation Support Systems (STRANS-2001)*, Kanpur, India, 2001, pp. 86-102.

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

[10] C.W. Tang, F. Wong, K.S. Leong, "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)*, Sanya, 2006, pp.1105-1109.

[11] F. Wong, D.C. Hu, Y.H. Mao, M.C. Dong, Y.P. Li, "Machine Translation Based on Constraint-Based Synchronous Grammar," *Proceedings of the 2nd International Joint Conference on Natural Language (IJCNLP-05)*, Jeju Island, Republic of Korea, 2005, pp. 612-623.
[12] F. Oliveira, F. Wong, K.S. Leong, C.K. Tong, M.C. Dong, "Query Translation for Cross-Language Information Retrieval by Parsing Constraint Synchronous Grammar," *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 2007, pp. 4003-4008.
[13] E. Aramaki, S. Kurohashi, S. Sato, H. Watanabe, "Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation," *Proceedings of MT Summit VIII*, 2001, pp. 27-32.
[14] R. Grishman, "Iterative Alignment of Syntactic Structures for a Bilingual Corpus," *Proceedings of Second Annual Workshop on Very Large Corpora (WVLC2)*, Kyoto, Japan, 1994, pp. 57-68.
[15] H. Watanabe, S. Kurohashi, E. Aramaki, "Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation," *Proceedings of the 18th International Conference on Computational Linguistics*, Germany, 2000, pp. 906-912.
[16] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Cybernetics and Control Theory*, vol. 10, no. 8, 1966, pp. 707-710.
[17] M. Tomita, "An efficient augmented-context-free parsing algorithm," *Computational Linguistics*, vol. 13, no.s 1-2, 1987, pp. 31-46.
[18] A. Aamodt, E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications*, vol. 7, no. 1, 1994, pp. 39-52.
[19] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California, 2002, pp. 138-145.
[20] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002, pp. 311-318.
[21] S. Banerjee, A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan, 2005, pp. 65-72.
[22] Macao Special Administrative Region Government Portal, http://www.gov.mo
[23] Government Printing Bureau, http://pt.io.gov.mo

Fai Wong received the B.Sc. and M.Sc. degrees in Software Engineering from the University of Macau in 1995 and 1999, and Ph.D. degree in Automation from Tsinghua University in 2005. He is currently an Assistant Professor in the Department of Computer and Information Science at University of Macau, with a secondary appointment in the Instituto de Engenharia de Sistemas e Computadores de Macau. His active and diverse research interests span areas of natural language processing and computational linguistic. His current research interests include machine translation, snapshot translation and machine learning to natural language processing. He is a member of IEEE and ACM.



Francisco Oliveira received the B.Sc. and M.Sc. degrees in Software Engineering from the University of Macau, in 2003 and 2006 respectively. He is currently a Research Administrative Officer and Ph.D. student in the University of Macau. His research areas include natural language processing, machine translation, and mobile computing technologies.

Yiping Li received the Ph.D. degree in applied mathematics from the University of Washington in 1987. Currently he is a professor of Computer and Information Science in the University of Macau. His research area includes machine translation and perturbation methods.