Parsing Extended Constraint Synchronous Grammar in Chinese-Portuguese Machine Translation

Francisco Oliveira[1], Fai Wong[1], Iok-Sai Hong[1], Ming-Chui Dong[1],

[1] Faculty of Science and Technology, University of Macau.
Av. Padre Tomás Pereira, Taipa, Macao
{olifran, derekfw, ma66536, mcdong}@umac.mo

Abstract. This paper presents a design model for parsing an extended version of Constraint Synchronous Grammar (CSG) for Chinese-Portuguese MT, which was initially applied in the reverse way. CSG is a generalization of context free grammars that describes syntactic structures of two languages simultaneously based on the defined constraints. Extensions include the addition of part of speech information in the target syntactic patterns, and the consideration of more information in each constituent's feature structure.

Keywords: Machine Translation, Constraint Synchronous Grammar

1 Introduction

It is always difficult to develop Machine Translation (MT) systems with high quality and efficiency for any domain. In the literature, different methodologies have been proposed. Rule based MT [1] 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. Example based MT [2] analyzes different pieces of bilingual examples stored in parallel corpora for generating the translation. Statistic based MT [3] considers probabilities estimated between the translation of words and the ordering of the sentences extracted from the corpora.

In recent years, researchers proposed solutions based on a stochastic tree-to-tree transduction or a synchronous parsing process in MT. Inversion Transduction Grammar [4] was proposed for defining a single parsing structure based on a set of brackets to account for both languages simultaneously. Multiple Context Free Grammar [5] was used by defining a set of functions for non-terminal symbols in the productions for interpreting the symbols during the generation phase. Deneefe and Knight [6] proposed a practical way in developing a MT system based on Synchronous Tree Adjoining Grammar [7].

In this paper, an extended version of Constraint Synchronous Grammar [8] (CSG) in application to Chinese-Portuguese MT is presented. Originally, CSG is applied for modeling the translation in the reverse way. CSG is used to express syntactic relationships between the source with one or more target sentential patterns. The selection of the most suitable target is based on the defined feature constraints for

each grammar rule. Since CSG uses feature structures for expressing detailed information of each non-terminal constituent (like POS, gender, number agreement, sense, etc), it helps in removing ambiguities during the parsing. When CSG is parsed from Portuguese to Chinese, only one stage of parsing is required for analyzing the syntactical structure of both languages, and the translation can be identified immediately based on feature constraints. However, if the target translation is a language which has a defined morphology, a morphological generation is required for handling the agreements between words in the target output. As a result, in order to extend the capability of CSG, in the target sentential pattern, Part-of-Speech (POS) are added for static variables, and more detailed feature structures related to the target words for each non-terminal constituent are considered. The application of extended CSG in a real Chinese-Portuguese MT System requires a segmentation preprocessing module to insert appropriate spaces between the Chinese words.

This paper is organized as follows. An introduction of CSG is given in section 2. The main extensions to the original version of CSG are introduced in section 3. The design model of Chinese-Portuguese MT based on extended CSG is given in section 4. Evaluation and discussion are presented in section 5, followed by a conclusion.

2 Constraint Synchronous Grammar

Constraint Synchronous Grammar [8] is based on the formalism of Context Free Grammar to the case of synchronous. Each CSG production rule consists of four parts: the reduced symbol, the syntactical pattern of the source, the syntactical pattern of the target, and the feature constraints, as shown in (1).

$$
\begin{aligned}
S \rightarrow\ & NP_1\ PP\ NP_2\ VP^*\ NP_3\ \{ \\
& [NP^1\ VP\ a\ NP^3\ NP^2]\ ; \quad VP_{cat} = vb1\ \&\ PP = \text{“把”}\ \& \\
& \qquad\qquad VP_{s:sem} = NP_{1sem}\ \&\ VP_{o:sem} = NP_{2sem}\ \& \qquad (1) \\
& \qquad\qquad VP_{io:sem} = NP_{3sem} \\
& [NP^1\ VP\ NP^2\ em\ NP^3]\ ; \quad \ldots\}
\end{aligned}
$$

The reduced symbol *S* has two generative rules associated with the source pattern $NP_1\ PP\ NP_2\ VP\ NP_3$. The selection of the most suitable target pattern is based on the defined feature constraints. The one that meets all the conditions defined determines the target pattern which mostly relates to the source. In this case, if the category of *VP* is *vb1*, the preposition is “把”, and the sense of the subject, direct, and indirect objects governed by *VP* corresponds to the sense of NP_1, NP_2, and NP_3, the first target pattern will become associated with the source pattern, and reduced as *S*. The asterisk “*” indicates the head element, and its usage is to propagate all the related linguistic information of the head symbol to the reduced non-terminal symbol.

Different languages often have structural differences between each other, including the syntactic order between the languages, discontinued constituents, and constituents that may vanish or appear in the target language translation. All these issues can be handled by CSG. 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, consider the following bilingual sentence: "她把兩支鋼筆借給了佩德羅" / "Ela emprestou ao Pedro duas canetas" (She lent two pens to Peter). Moreover, suppose that this sentence is going to reduce to the symbol *S* in (1). In this case, there is a discontinuity of the words in Chinese, where "把" and "借給了" should be associated with the verb "emprestou" (to lend). Since the sentence matches the source pattern of (1), it will be associated with the first target pattern. The consideration of constituents that are disappeared or shown in the target is handled in a similar way.

$$\text{NP} \rightarrow \text{NP}_1\ \text{NP}_2\ \{\ [\text{NP}^2\ \textit{de}/\text{Prep}\ \text{NP}^1]\ ;\ \text{NP}_{1\text{sem}} = \textit{place}\ \&\ \text{NP}_{2\text{sem}} = \textit{institution}\} \qquad (\mathbf{2})$$

For instance, in production (2), it often happens that the preposition "的" (of) in Chinese is vanished. However, it is necessary to add the word "de" (of) in the target pattern to have a correct translation. As an example, in the bilingual sentence: "澳門大學" / "Universidade de Macau" (University of Macau), although "的" (of) in Chinese is vanished, it still requires the preposition "de" (of) in the target sentence.

3 Extensions of Constraint Synchronous Grammar

The design of CSG is originally targeted for the translation from Portuguese to Chinese. However, it can be further extended in the adaption to other language pairs. The main objective of this paper is to highlight the extensions required for the application of CSG in modeling the translation from Chinese to Portuguese.

$$\text{S} \rightarrow \text{NP}_1\ \text{VP}\ \text{NP}_2\ \text{NP}_3 \quad \{\ [\text{NP}^1\ \text{VP}\ \text{NP}^3\ \textit{a}/\text{Prep}\ \text{NP}^2]\ ;\ \ldots\ \} \qquad (\mathbf{3})$$

For each target sentential pattern, POS associated with each static word is defined. In the production rule (3), a POS tag is assigned to the static word "a" (to) as a preposition for morphological synthesis, which is later used in the generation of the correct word form in the target pattern associated. One might wonder the necessity of the POS associated with the static word. If the static word "a" (to) is not going to be changed in the target pattern, then it doesn't need to explicitly specify that "a" (to) is a preposition. However, this is not true, because the word is ambiguous. For instance, suppose that the following sentence matches the production rule (3): "他 給了 安娜 三 張 票" (He gave three tickets to Anna). If POS is not attached with the static word, the system could misinterpret it as an article instead of preposition, which would make a big difference in the translation result. In this case, if "a" (to) is a preposition, the system will generate correctly as "Ele deu três bilhetes à Ana" (He gave three tickets to Anna), where "à" is the contracted result of "a" (to) as a preposition and "a" as the article associated with "Ana" (Anna).

More linguistic information are considered for giving extra flexibility in establishing agreements for syntactical and sub-categorization dependencies in CSG rules. Each linguistic information is represented by Feature Descriptor (FD), in the type of “attribute=value”, and the value can be either an atomic symbol or recursively another FD. In general, each lexical constituent has several FDs and all of them are stored in Attribute Value Matrices for disambiguation and morphological synthesis.

Different POS have different FDs in general, and they are used for two purposes. Firstly, FDs help in selecting the most appropriate translation based on the feature constraints defined in CSG rules. Only FDs compatible with each other are considered as succeeded, and a new FD is then constructed based on the unification. In CSG formalism, a preference measurement mechanism is considered. Weight values are assigned to features structures, and during the unification process, they can either be rewarded or punished depending on different factors, and the one that has the highest value is considered as the most preferable content. Secondly, FDs provide necessary information for morphological conversions in generating a correct translation.

4 Design Model of Chinese-Portuguese MT based on CSG

An overview of the proposed Chinese-Portuguese MT design model based on extended CSG rules is given in Figure 1. The core part of the parsing module can be used in both translation directions, but different modules before and after processing, and extended CSG rules are required in the translation from Chinese to Portuguese.

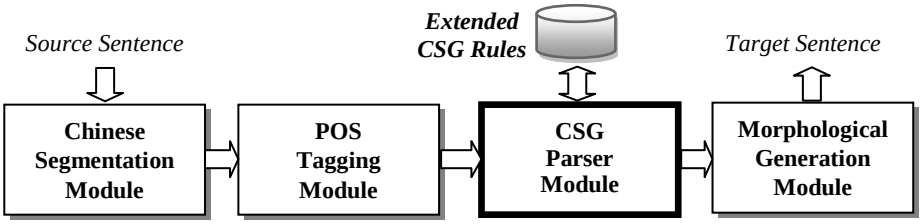

Fig. 1. Design Model of Chinese-Portuguese MT based on Extended CSG rules

The Chinese source sentence is first segmented, and each word is assigned with a POS by CSAT [9], an integrated application with Chinese segmentation and tagging ability. Extended CSG rules are then parsed by a modified version of generalized LR algorithm [10], 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. In the generated parse tree, each node has an associated target sentential pattern which is used to generate the final translation accordingly. Morphological synthesis is then applied to render the

translation result, which is based on the feature unification. If there is a failure in the unification between two different FDs, the module performs necessary changes in terms of gender, number, tense, categories of person to make it successful. Finally, articles restoration and contraction between surrounded words are performed.

5 Evaluation and Discussion

In order to evaluate the applicability of the extended CSG rules in Chinese-Portuguese MT, two experiments are carried out. The first one is based on BLEU [11] and NIST [12] metrics. The second one is based on the average value of human assessments evaluated by three linguistics. The test suite includes 100 sentences selected from a grammar book [13], with an average of 10.16 word, in a close domain.

Table 1. BLEU, NIST, and Human Assessment Experiment Results

Metrics	BLEU	NIST	Human Assessment		
			Good	Acceptable	Unacceptable
Score	0.6630	6.0110	65%	20%	15%

Table 1 shows the evaluation results. The scores are directly affected by several factors. Since MT involves a chain of processes (segmentation, POS tagging, CSG parsing, morphological generation) to be accomplished, if one of them gives an incorrect result, no doubt, an incorrect translation is generated. Moreover, the ellipsis and the omission of words in Chinese are so common as to produce many difficulties for MT systems in guaranteeing the quality of the translation. BLEU and NIST metrics don't have the same effect as human assessment. For example, consider the following correct bilingual pair: "學生們遲到了" / "Os alunos chegaram atrasados" (Students arrived late), and the translation generated by the System is "Os estudantes atrasaram" (Students were late). In terms of human assessment, this sentence is ranked as Good, because it is fluent, adequate, and correct in terms of grammar and meaning but it has different words compared with the reference translation. It is very often to have these cases in reference translations that directly affect the scores of BLEU and NIST metrics since they are based on n-gram co-occurrence precision.

Although it is always hard to have enough extended CSG rules in covering all the cases for any domain, more can be defined either manually or through machine learning approaches in the future to compensate this issue. This evaluation shows that new considerations in CSG mentioned can be fully applied in Chinese Portuguese MT.

6 Conclusion

This paper presents necessary extensions of Constraint Synchronous Grammar in application to Chinese-Portuguese MT, including: the assignment of POS tags in static words of the target patterns for each CSG production rule; the consideration of

more feature descriptors in attribute value matrix for each lexical word or constituent. On the other hand, in a practical Chinese-Portuguese MT System based on extended CSG rules, a segmentation module is considered for assigning delimiters between Chinese characters, and a morphological module is used for rendering the final translation result based on the agreement rules defined.

Acknowledgments. This research work was supported by the Research Committee of University of Macau under Ref. UL019/09-Y1/EEE/DMC01/FST Cativo: 5868.

References

1. Bennett, W.S., Slocum, J.: The LRC Machine Translation System. Computational Linguistics 11(2-3), 111--121 (1985)
2. Brown, R.D.: Example-Based machine translation in the Pangloss system. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 169--174. Copenhagen, Denmark (1996)
3. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A Statistical Approach to Machine Translation. Computational Linguistics 16(2), 79--85 (1990)
4. Wu D.: Grammarless extraction of phrasal translation examples from parallel texts. In: Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 354--372. Leuven, Belgium (1995)
5. Seki, H., Matsumura, T., Fujii, M., Kasami, T.: On multiple context-free grammars. Theoretical Computer Science 88(2), 191--229 (1991)
6. Deneefe, S., Knight, K.: Synchronous Tree Adjoining Machine Translation. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP). Singapore (2009)
7. Shieber, S.M., Schabes, Y.: Synchronous Tree-Adjoining Grammars. In: Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), pp. 253--258. Helsinki, Finland (1990)
8. Wong F., Hu D. C., Mao Y. H., Dong M. C., Li Y. P.: Machine Translation Based on Constraint-Based Synchronous Grammar. In: Proceedings of the 2nd International Joint Conference on Natural Language (IJCNLP-05), pp. 612--623. Jeju Island, Republic of Korea (2005)
9. Leong K. S., Wong F., Tang C. W., Dong M. C.: CSAT: A Chinese Segmentation and Tagging Module Based on the Interpolated Probabilistic Model. In: Proceedings in Computational Methods in Engineering and Science (EPMESC-X), pp. 1092--1098. Sanya, Hainan, China (2006)
10. Tomita, M.: An efficient augmented-context-free parsing algorithm. Computational Linguistics 13(1-2), 31--46 (1987)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311--318. Philadelphia (2002)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138--145. San Diego, California (2002)
13. Wang, S., Lu, Y.: Gramática da Língua Portuguesa. Shanghai Foreign Language Education Press (1999)