QUERY TRANSLATION FOR CROSS-LANGUAGE INFORMATION RETRIEVAL BY PARSING CONSTRAINT SYNCHRONOUS GRAMMAR

FRANCISCO OLIVEIRA[1], FAI WONG[1], KA-SENG LEONG[1], CHIO-KIN TONG[1], MING-CHUI DONG[1]

[1]Faculty of Science and Technology, University of Macau, Macao
E-MAIL: {olifran, derekfw, ma56538, ma66535, mcdong}@umac.mo

Abstract:

With the availability of large amounts of multilingual documents, Cross-Language Information Retrieval (CLIR) has become an active research area in recent years. However, researchers often face with the problem of inherent ambiguities involved in natural languages. Moreover, this task is even more challenging for processing the Chinese language because word boundaries are not defined in the sentence. This paper presents a Chinese-Portuguese query translation for CLIR based on a Machine Translation (MT) system that parses Constraint Synchronous Grammar (CSG). Unlike traditional transfer-based MT architectures, this model only requires a set of CSG rules for modeling syntactic structures of two languages simultaneously to perform the translation. Moreover, CSG can be used to remove different levels of disambiguation as the parsing processes in order to generate a translation with quality.

Keywords:

Cross-Language Information Retrieval, Machine Translation, Constraint Synchronous Grammar

1. Introduction

The main objective of Cross-Language Information Retrieval (CLIR) is to retrieve information written in a language different from the language of the user's input query. This is especially helpful in Macau, where many documents are written in Chinese and Portuguese, since both of them are official languages in the territory. CLIR systems permit users to retrieve documents written not only in their native language but also in the other one. However, it is not easy to obtain query translations with high quality in any domain due to the inherent ambiguities and linguistic phenomena involved in natural languages, and the need of a enormous knowledge for disambiguation.

In the literature, several approaches have been proposed. In bilingual dictionary based approaches [1], query translation is generated by looking at the entries of the lexicon. Although it is efficient in terms of translation, large amount of vocabulary is needed to cover all the words. However, it is not easy to achieve and the problem is even worse with Chinese, which do not have word boundaries. Moreover, words in the dictionary usually have more than one translation, and it is a difficult task for selecting the best translation by just considering the bilingual dictionary.

MT based approaches seems to be the ideal solution for CLIR. It is mainly because MT systems translate the sentence as a whole, and the translation ambiguity problem is solved during the analysis of the source sentence. Rule-based MT [2] uses a method based on a set of linguistic rules, where rules are translated in a linguistic way. Since these rules are universal, they are domain independent. However, this approach often requires a large human cost in formulating rules and it is hard to maintain consistency as the number of rules increases. In statistical based approaches [3, 4], the translation is determined by estimating the probabilities between the translation of words and the ordering of the sentences based on a parallel corpora. However, these approaches suffer from the dependency with the parallel corpora. For Example-based MT [5], it will search for pieces of examples stored in the parallel corpora for generating the translation, but it often depends on the quality of the examples and the similarity function defined.

In recent years, some versions of synchronous grammars [6] are proposed for solving non-isomorphic tree based transduction problem and to provide solutions to Machine Translation. For example, Synchronous Tree Adjoining Grammar [7] was initially applied for semantics but was later considered for translation [8]. Multiple Context-Free Grammar [9] was used by defining a set of functions for non-terminal symbols in the productions in order to interpret the symbols in the target generation. However, it is hard to describe discontinuous constituents in linguistic expression [10]. Melamed [6] modeled the problem of MT as a synchronous parsing based on Generalized Multitext Grammars that maintain two sets of productions as components, one for each language, for modeling parallel texts. Although it can be used to describe

semantic information with details associated with a non-terminal, it is difficult for the development of a practical MT system due to its lack of flexibility.

In this paper, we applied Constraint Synchronous Grammar (CSG) [10], a variation of synchronous grammar, as the kernel of the MT system in performing query translation for Chinese-Portuguese CLIR. CSG can be used to express detailed feature structures like gender, number, agreement, etc in each non-terminal constituent for performing necessary disambiguation in each level, and CSG can express non-standard linguistic phenomena, including crossing dependencies, and discontinuous constituents in the inference rules. Moreover, CSG allows the parser to remove the ambiguous parse trees as the parsing progresses by making use of various linguistic features defined.

This paper is organized as follows: the design model of a Chinese-Portuguese query translation for CLIR by parsing CSG is given in section 2. Chinese word segmentation is presented in section 3. The parsing of Constraint-based Synchronous Grammar and the generation of the translation are detailed in sections 4 and 5. Finally, the conclusion is given in section 6.

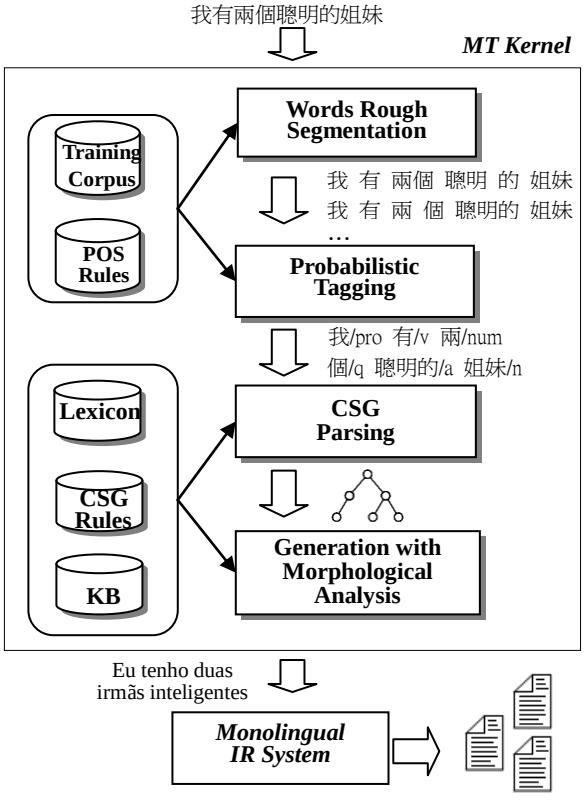

Figure 1. Kernel design of the proposed MT system

2. Design Model of Chinese-Portuguese CLIR System

The kernel of the MT system proposed in application to CLIR is shown in Figure 1. The translation process begins with word segmentation of the given Chinese query, and it is then tagged by a probabilistic tagger. Based on the segmented and tagged result, the source sentence is further analyzed by using a modified generalized LR parser [11] for inferring the syntactic structure of the input, guided by CSG rules, in order to determine the target language sentential pattern. This pattern is then morphologically analyzed in order to generate the translation of the source language. The sentence translated is then passed to a monolingual IR system for retrieving documents written in target language.

3. Chinese Segmentation and Tagging Module

Whether Chinese words are segmented effectively and correctly is vital in obtaining a good translation result in MT systems involving Chinese translation. This is mainly because Chinese sentences, unlike other Western languages such as Portuguese, there are no delimiters between words in the sentence. Moreover, there are many ambiguity problems in correctly segmenting a Chinese sentence.

In our design model, we applied N-Shortest-Paths method [12] for generating a set of rough segmented results of Chinese sentences.

3.1. Words Rough Segmentation Model

For a given Chinese sentence, a directed graph is constructed with each of its atomic characters as the vertices (V_1, V_2, …, V_n). Edges between the vertices are determined by probabilities of the atomic characters or the combinations of the words obtained in the Chinese corpus. Let W be one of the possible results of the segmentation for the Chinese sentence C, then the probability of W, given C is defined as:

$$P(W \mid C) = \frac{P(W)P(C \mid W)}{P(C)} \quad (1)$$

Since the probability of the Chinese sentence $P(C)$ is a constant, and the probability of C, given W must be 1, the objective is to determine the N different segmentations which have the N largest probabilities of $P(W)$. Suppose that a possible segmentation sequence W consists of w_1, w_2, …, w_m words, then the probability $P(w_i)$ can be approximated as:

$$P(w_i) \approx \frac{(k_i + 1)}{(\sum_{j=0}^{m} k_j + V)} \quad (2)$$

k_i is the number of occurrences of w_i and V is the number of word types in the training corpus. Smoothing is applied by adding a constant in the numerator by taking into consideration that w_i may not appear in the training corpus. By assuming that the context within the sentence is not considered for simplicity, the best word sequence W can be computed as

$$\begin{aligned} \arg\max_{W} P(W) &= \arg\max \prod_{i=1}^{m} P(w_i) \\ &\approx \arg\max(\prod_{i=1}^{m} \frac{(k_i + 1)}{\sum_{j=0}^{m} k_j + V}) \end{aligned} \quad (3)$$

Based on the segmented candidate Chinese sentences, these are going to be tagged by a probabilistic tagger [13] based on Hidden Markov Model [14] to determine the final segmented sentence and the best POS tag for each word.

4. Parsing Constraint Synchronous Grammar

Constraint Synchronous Grammar [10] is based on the formalism of Context Free Grammar (CFG) 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.

4.1. Definition of CSG

In CSG, every production rule is in the form of

S → 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. For example, if the category of *VP* is *vb1*, the preposition given is "*把*", and the sense of the subject, direct, and indirect objects governed by the verb *VP* corresponds to the first, second, and the third nouns (*NP*), then the source pattern NP_1 *PP* NP_2 *VP* NP_3 is associated with the target pattern NP_1 *VP a* NP_3 NP_2.

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

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, NP_1 *VP a* NP_3 NP_2, although the first *NP* in the source pattern corresponds to the first *NP* in the target one, the verb, the second and third noun phrases in the source are changed in the target sentential pattern.

Understanding the ordering of constituents in the target sentential pattern is very important because it affects not only in the correctness of the sentence in terms of grammar but also in terms of meaning. For example, suppose that the sentence "*貧窮的人*" (a poor man) is going to be translated. If word by word translation is applied, the sentence will be translated as "*pobre homem*". Although the sentence is translated correctly in terms of grammar, it is not correct in terms of the meaning. This happens because the positioning between adjectives and nouns in Portuguese language may produce different meanings. In this case, "*pobre homem*" means a pitiful man and not a poor man. This problem can be easily solved by defining a CSG production rule that has different generative rules associated with the same source sentential, where each of these rules are controlled by different conditions. As a result, the source sentence "*貧窮的人*" will be translated as "*homem pobre*" (a poor man) instead of "*pobre homem*".

4.2. Feature Descriptors in Attribute Value Matrix

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 descriptors related to a single lexical word or a constituent are encoded in Attribute value matrices (AVM). Each FD is a set of pairs in the type of "*a* = *v*", where "*a*" is an attribute and "*v*" is a value, either an atomic symbol or recursively a FD. Moreover, feature unification is performed during the

parsing stage. If FDs of each lexicon word or lexical are compatible with each other, i.e. there are no conflicts on the value of all the attributes defined, unification succeeds and a new FD is constructed.

As an example, consider that a new noun phrase is going to be reduced based on the words "*探測*" (probe) and "*石油*" (petroleum) and below, it shows their AVMs.

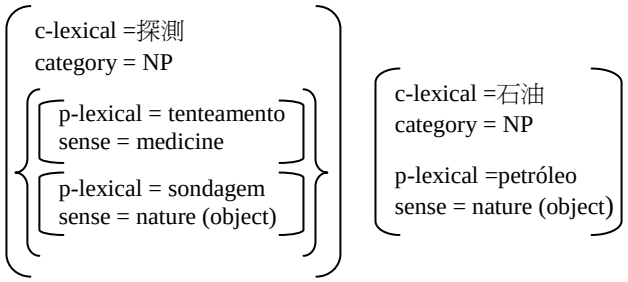

Figure 2. AVMs of the words "*探測*" and "*石油*"

If the control condition defined by the rule requires that the senses of the noun phrases must be equal to each other, then the unification will select the meaning of "*sondagem*" (probe) since this sense can be unified with the one of "*petróleo*" (petroleum).

In traditional unification based approaches [15], if FDs of each lexicon word or a constituent are not compatible with each other during the unification process, nothing is returned. However, if only one of the FDs' unification fails, then all the related candidate words will be rejected without any flexibility in choosing the next preferable or probable candidate. Thus, in our design, each feature is associated with an initial weight and ranking is performed during the parsing stage for choosing the most suitable candidate word. Suppose that the AVMs of the words "*死屍*" (corpse) and "*漂浮*" (to fluctuate) are shown below:

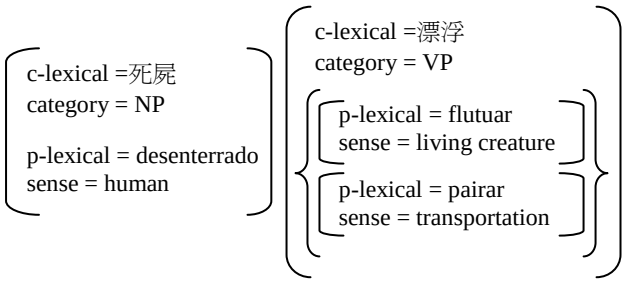

Figure 3. AVMs of the words "*死屍*" and "*漂浮*"

During the parsing stage, if the control condition requires that the sense of *NP* must be equal to the sense of the subject governed by *VP*, weights are assigned during the validation process and the one that has the highest weight will be selected for unification. The assignment of weights is based on the following policies: if unification can be performed between the senses of the lexical words or constituents, then the weight is increased by 1; if unification fails, but if the sense of a word is an inherited hypernym of the other or vice-versa, the weight is increased by 0.5. FDs with the highest weight are chosen as the most preferable content. In this example, traditional unification approach will just return failure. Although there are no exact matches between the senses of "*死屍*" (corpse) and "*漂浮*" (to fluctuate), since the sense human is hyponymic to the sense of living creature, FD of the Portuguese word "*flutuar*" (to fluctuate) will still be unified with FD of "*desenterrado*" (corpse) and selected as the most suitable candidate.

4.3. Expressiveness of CSG

As mentioned previously, CSG can be used to describe non-standard linguistic phenomena. For example, consider the bilingual sentence:

她/NP1 把/PP 兩支鋼筆/NP 借給了/VP 佩德羅/NP
Ela emprestou ao Pedro duas canetas
(She lent two pens to Peter)

It is often that many linguistic expressions will not appear in the translation of the other language. For instance, the preposition *PP* does not appear in any of the target rules. Moreover, the Chinese preposition "*把*" and the verb "*借給了*" should be paired with the Portuguese verb "*emprestar*" (to lend).

These observations show that CSG can be used to express not only structural deviations between two different languages, but also discontinuous constituents' relationships in the Chinese component.

4.4. CSG Parser

CSG formalism is parsed by a modified version of generalized LR algorithm [11] that takes the features constraints and the inference of the target structure into consideration. The main reason for choosing this algorithm is due to the considerable efficiency over the Earley's parsing algorithm [16] which requires a set of computations of LR items at each stage of parsing [11]. Furthermore, the parsing table used is extended by adding features constraints and the target rules into the actions table.

5. Generation of the translation

Once the parse tree is constructed, the translation of the input sentence is generated by referencing the set of generative target sentential patterns that were selected previously.

In each node of the parse tree, there is an associated target sentential pattern, which is used to generate the corresponding translation. Moreover, in order to ensure that the system generates perfectly the translation in Portuguese grammatically, we employ unification of Functional descriptors (FD) as a validation operation for each node.

Since AVMs for each node was constructed for each constituent node in the parsing stage, these will be reused during the generation phase. Since most of the Portuguese words defined in FDs are in their original word-form, they need to be changed based on a set of grammatical agreement rules. Thus, extra FDs will be added accordingly to the AVM, depending on its part-of-speech, for checking the dependency between Portuguese words in order to generate the target translation correctly. These extra attributes include number, gender, tense, and categories of person.

As an example, consider the parse tree of the sentence "*她把兩支鋼筆借給了佩德羅*" shown in Figure 4.

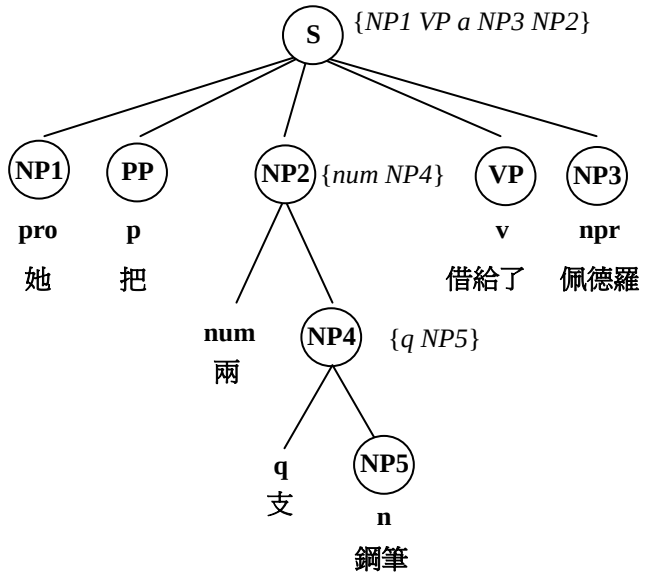

Figure 4. Example of a parse tree

Suppose that the translation of the noun phrase "*兩支鋼筆/NP2*" (two pens), with the target pattern *num NP4*, is going to be generated. The meanings obtained from the bilingual dictionary of the words "*兩*" (two) and "*鋼筆*" (pen) are "*dois*" and "*caneta*" respectively. Moreover, FDs of "*兩*" and "*支 鋼筆*", and their related information are shown below.

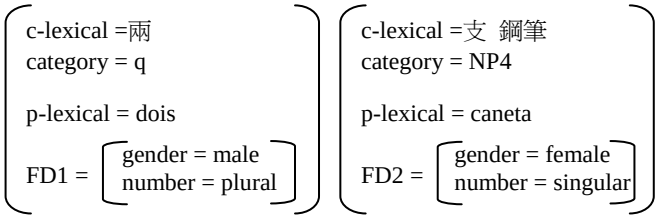

Figure 5. AVMs of the words "*兩*" and "*支 鋼筆*"

Unification of FD1 and FD2 will fail because the gender and the number are different. In such a case, necessary conversions are performed so that FD1 and FD2 will be compatible with each other. Therefore, the generated result for "*兩支鋼筆*" is "*duas canetas*" (two pens). Similarly, since the verb phrase "*借給了/VP*" (lent) must be in agreement with *NP1* and it must have a correct tense, the Portuguese word "*emprestar*" (to lend) should be converted to "*emprestou*" (third person in past tense). Besides unification, articles may need to be restored for each noun phrase is necessary. For example, the noun phrase "*佩德羅/NP3*" (Peter) needs to add an article "*o*" before the Portuguese word "*Pedro*".

After all the unifications and article restorations, the sentence becomes "*Ela emprestou a o Pedro duas canetas*". However, the generated sentence is still not totally correct. It is mainly because some words can be contracted in the Portuguese grammar. In this case, the preposition "*a*" and the article "*o*" should be contracted as one word "*ao*". Thus, an extra module that checks if there is a need for contractions is called at last, and the output of the generation module is "*Ela emprestou ao Pedro duas canetas*" (She lent two pens to Peter).

6. Conclusion

In this paper, we proposed Chinese-Portuguese query translation for CLIR based on a MT system that parses Constraint Synchronous Grammar. In this architecture, based on the given Chinese sentence, a set of rough segmented results is generated and after tagging all of these candidate sentences, the one with the highest score will be selected. The sentence is then parsed for inferring the syntactic structure based on Constraint-based Synchronous Grammars. Unlike transfer-based MT architectures where the translation process is carried out in sequence by different analytical phases, by parsing CSG rules, the corresponding target sentential pattern can be inferred

immediately, so that our approach can reduce information loss during the transfer process. After constructing the parse tree, it is used for generating the translation with the assistance of the unification between functional descriptors defined to guarantee the correctness of the grammar and the quality of the translation.

The proposed MT model can remove different types of ambiguity at different stages for enhancing the quality of the translation: the creation of word boundaries in the segmentation module removes ambiguity between Chinese words; Part-of-speech ambiguity is removed by probabilistic tagger; structural ambiguities in parse trees can be removed by parsing CSG; and lexical ambiguities, where words may have more than one meaning, usually referred as the problem of word sense disambiguation, can be solved through CSG parsing through the analysis of surrounded neighbors of the ambiguous word in question.

Acknowledgements

The research work reported in this paper was supported by "*Fundo para o Desenvolvimento das Ciências e da Tecnologia*" (Science and Technology Development Fund) under grant 041/2005/A and it was also supported by Research Committee of University of Macau under grant CATIVO:2372.

References

[1] Ballesteros, L., Croft, W. B., "Dictionary-based Methods for Cross-Lingual Information Retrieval", Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pp. 781-801.

[2] Bennett, W. and Slocum, J., "The LRC Machine Translation System", Computational Linguistics, Vol. 11, No. 2-3, pp. 111-121, 1985.

[3] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A Statistical Approach to Machine Translation", Computational Linguistics, Vol. 16, No. 2, pp. 79-85, 1990.

[4] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Mercer, "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics, Vol. 19, No. 2, pp. 263-311, 1993.

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

[6] Melamed I. D., "Multitext Grammars and Synchronous Parsers", Proceedings of NAACL/HLT 2003, Edmonton, pp. 79-86, 2003.

[7] S. M. Shieber and Y. Schabes. "Synchronous Tree-Adjoining Grammars", Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), Helsinki, Finland, pp. 253-258, 1990.

[8] A. Abeillé, Y. Schabes, A. Joshi, "Using lexicalized TAGs for machine translation", Proceedings 13rd International Conference on Computational Linguistics (COLING-90), Helsinki, Finland, Vol. 3, pp. 1-7, 1990.

[9] Seki, H., Matsumura, T., Fujii, M., Kasami, T., "On multiple context-free grammars", Theoretical Computer Science, Vol. 88, No. 2, pg. 191-229, 1991.

[10] Wong F., Hu D. C., Mao Y. H., Dong M. C., and Li Y. P., "Machine Translation Based on Constraint-Based Synchronous Grammar", Proceedings of the Second International Joint Conference on Natural Language (IJCNLP-05), Vol. 3651, Jeju Island, Republic of Korea, pp. 612-623, 2005.

[11] Tomita, M., "An efficient augmented-context-free parsing algorithm", Computational Linguistics, Vol. 13, No. 1-2, pp. 31-46, 1987.

[12] Zhang HP, Liu Q., "Model of Chinese words rough segmentation based on N-shortest-paths method", Journal of Chinese Information Processing, Vol. 16, No. 5, pp. 1-7, 2002.

[13] Leong K. S., Wong F., Tang C. W., and Dong M. C., "CSAT: A Chinese Segmentation and Tagging Module Based on the Interpolated Probabilistic Model", Proceedings in Computational Methods in Engineering and Science (EPMESC-X), Sanya, Hainan, China, pp. 1092-1098, 2006.

[14] Rabiner L., "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol. 77, No. 2, pp. 257–286, 1989.

[15] K. Ronald, "The Formal Architecture of Lexical-Functional Grammar", Journal of Information Science and Engineering, Vol. 5, pp. 305-322, 1989.

[16] Early J., "An Efficient Context-Free Parsing Algorithm", Communications of the Association for Computing Machinery, Vol. 13, No. 2, pp. 94-102, 1970.