Rule base combined Linguistics knowledge with Corpus*

Ying Liu
Lab of Computational Linguistics,
Department of Chinese Language and
Literature,Tsinghua University,
BeiJing, China
yingliu@mail.tsinghua.edu.cn

Chengqing ZONG
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
cqzong@nlpr.ia.ac.cn

Abstract - *This paper proposes a new approach to construction of rule bases for the transferred- based machine translation. In our approach, the rule bases are constructed in combination of the linguistics knowledge and large scale of corpora. On the one hand the lexical knowledge, the syntactic knowledge and the semantic knowledge are all used in the rules. on the other hand the knowledge is used for the statistics and self-learning of rules. In each rule base, all rules are scored and ranked. Thus an impersonal choice for the sentence can be made. The preliminary experimental results show that the approach may increase the speed to build the rule base and improve the quality of rules.*

Keywords: Chinese-English machine translation, rule base, linguistics knowledge, corpus

1 Introduction

The rule bases are very important for the transfer-based machine translation (MT), in which the analyzer, transfer and generator are all working based on the rules. There are many factors determining the quality of an MT system, including whether the design of the rule bases is appropriate, whether the rules may cover a large scope of linguistics phenomena, and whether the different rules keep consistent and so on. If the rules are summarized from a small size of language material, or the rules are constructed with the introspective method, the rules must be incomprehensive. Therefore, the reasonable rules should be constructed by investigating large numbers of language materials, and the large-scale corpora are necessary. We

believe that it is a right way to combine the linguistics knowledge and the large-scale corpora.

Recently, there has been a rebirth of empiricism in the field of natural language processing. The corpus-based approach has been successfully used in many different areas of natural language processing, such as in part-of-speech tagging and speech recognition and so on. The corpus-based approach is also widely used in rule bases construction and parser development[3][5]. The scoring function is provided in [2]. To resolve ambiguity problems the scoring function has been successfully applied to an English-Chinese machine translation system, Chinese-Chinese machine translation system and a spoken language processing system [2] [9] [10][11][12] .

Because the linguistics-based method and the corpus-based method each has advantages and disadvantages, Combination of the two methods is widely used for segmentation, tagging, analysis, and machine translation[1][4][6][7][14][15]. Constructing and maintenance of rule bases should make use of the combination of the two methods

There are the rule bases for the word segmentation, part-of-speech(POS) tagging, analysis , transfer, and synthesis in a transfer-based Chinese-English machine translation system. In our approach, the rule bases are constructed in combination of the language knowledge

* 0-7803-7952-7/03/$17.00 © 2003 IEEE.

and large-scale corpora. We make use of the language knowledge such as the lexical knowledge, morphological knowledge, syntactic knowledge, semantic knowledge and the commonsense knowledge. In another words, all kinds of rules may use the lexical knowledge, POS knowledge, morphological knowledge, syntactic knowledge, and the semantic knowledge and so on. On the one hand, the corpora are used to help to construct rules by capturing knowledge correlative with rules. However, the simple rules, including the word segmentation rules and the POS tagging rules etc., can be captured by using the rule templates, and the corpora are used to fill the rule templates. On the other hand, the corpora are used to judge whether the rules are correct. When the rules are judged, the probability is evaluated to the rules. The probability scores of these rules are the correct rate of rules. However, when the tree can not be judged correct, the probability scores of both analysis rules and transfer rules are related to the scores of trees. The scores of trees takes into account the lexical information, syntactic information, and also the semantic information.

2 Rule bases

As we know, there is not any word boundary marker in the Chinese sentence, so the word segmentation usually causes the intersection ambiguities and combination ambiguities. In our Chinese-English MT system, the word segmentation rules are mainly utilized to deal with the two kinds of ambiguities. The tagging rules are utilized to disambiguate for the word with more than one POS. There may be more than one syntactic tree in the analysis phase. So the analysis rules are utilized to disambiguate and parse the sentences. A Chinese syntactic tree may be transferred into more than one English syntactic tree. So, the transfer rules are designed for disambiguating and transferring[8].

The synthesis rules are utilized to disambiguate and synthesize the English word string. The morphology rules are used for morphology synthesis and disambiguating. Figure 1 gives the transfer-based Chinese-English machine translation process.

The basic policy for constructing the rules are summarized as follows:

(1) *Integrality*: The typical language phenomena should be covered by the rules, such as the typical intersection ambiguities, combination ambiguities, ambiguities of POS tagging, typical grammar and semantic phenomenon, word and structural transfer ambiguities, and synthesis ambiguities and so on.

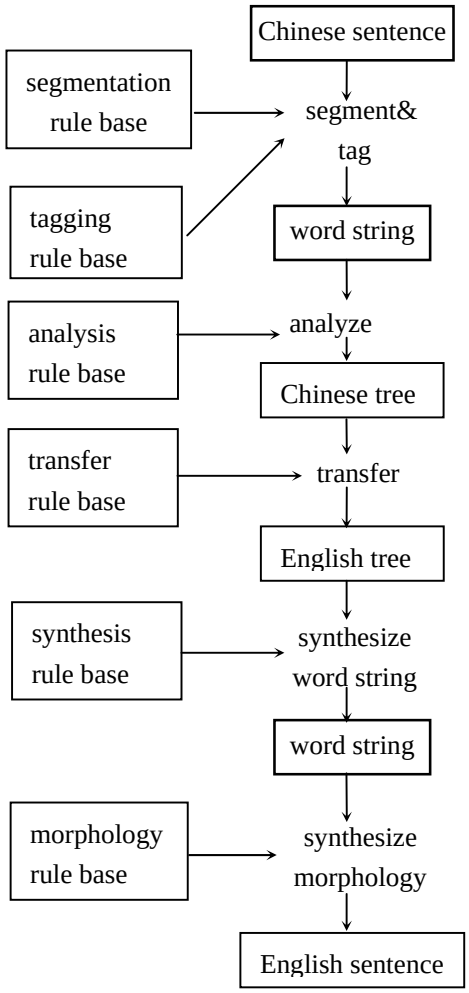

Figure 1 Transfer-based Chinese English Machine Translation

(2) *Consistency*: The different rules in the same rule base should keep the consistency, such as the symbol and the expression of rule.

(3) *The relation between general rule and exceptional rule*: The general rules should express the general language regulations, but the exceptional rules express the special use of some words.

(4) *Difference between two kinds of language*:

There are a lot of differences between Chinese and English not only in the word but also in the sentence structure. So, the difference must be considered enough for both the transfer rule and synthesis rule.

(5) *Rule based on large scale of corpora*. The rules should be changed with the language's change.

(6) *Lexical information, syntactic information and semantic information*. In all kinds of rules, the lexical information, syntactic information, and the semantic information should be made enough use of, thus the ambiguities are probably disambiguated.

3 Linguistics knowledge

In our experimental system, the rules are expressed with complex feature set, and the unification operation is adapted for the complex features. Some examples are given as follows.

Rule-1: ABC(A+BC, AB+C)
IF %A.POS=r, %BC.POS=a|f|d
THEN A+BC

This rule expresses: The string ABC may be segmented into A+BC or AB+C(A, BC, AB and C are all word).

IF the part of speech of A is a pronoun, the part of speech of BC is an adjective or orientation word or adverb, the string ABC is segmented as A+BC.

For example: 这里面有许多好看的衣服。(There are a lot of beautiful cloth inside.) According to the rule, the string "这里面" is segmented into "这 里面", but not "这里 面".

The rule includes parts of speech knowledge such as r(pronoun), a(adjective), f (orientation word) and d (adverb).

Rule-2: 比(n-v-p)

IF 比.LeftNeighbor.yx=的|之 THEN n ELSE v|p

This rule expresses: The part of speech of "比" may be a noun, a verb or a preposition. If its left adjoining word is "的" or "之" , the word is tagged as noun. Otherwise, the word is tagged as verb or preposition. In the rule, the actual words "的" and "之" are included.

Rule-3: np→ap !np
%ap.dingyu=yes, %np.zhxyu1=yes, %ap.zhuti=%np
IF ap.neibujiegou=word, %np.neibujiegou=word,
%np.semantickind = goujian | chouxiangshiwu,
THEN $.dingyu=yes.
ELSE $.dingyu=no ENDIF

This rule includes the syntactic knowledge and semantic knowledge. Here, the syntactic knowledge means "dingyu", "zhxyu1", "zhuti" and "neibujiegou". The semantic knowledge refers to "goujian" and "chouxiangshiwu". Please refer to [13] for details on the symbol "dingyu", "zhxyu1", "zhuti", "neibujiegou", "goujian" and "chouxiangshiwu".

4 Statistical knowledge

In the transfer-based Chinese-English machine translation system, one rule can be used for any phase, such as the segmentation, tagging, analysis, transfer and synthesis and so on. And there may be many analytical results in any phase. So, we score the rules and make them ranked. The scored rules are used to truncate unnecessary results, and thus the speed of machine translation is increased. The rule score makes enough use of the large scale of corpora and the lexical knowledge, syntactic knowledge and semantic knowledge. Thus the accuracy of the MT system is improved.

In our approach, the segmentation rule score, tagging rule score, and the morphology rule score are all computed according to the same principle. The segmentation rule score is the correct rate of the rule when a large scale of Chinese texts are segmented with the rule, which is counted automatically based on a given correctly segmented corpora. The tagging rule score is the correct rate of the rule when a large scale of corpora are tagged with the rule, which is also counted automatically based on a given correctly tagged corpora. The morphology rule score is the correct rate of the rule when a large scale of English texts are synthesized with the morphology rule, which is also counted automatically.

The analysis rule score, transfer rule score and synthesis rule score are also computed according to the same principle:

When there is a correct tree base, the score of the analysis rule $H \rightarrow G_1\ G_2$ is the probability of the rule, signed as $P(H \rightarrow G_1\ G_2)$.

$$P(H \rightarrow G_1\ G_2)=f(H \rightarrow G_1\ G_2)/f(H) \quad (1)$$

Where $f(H \rightarrow G_1\ G_2)$ is the frequency of the rule $H \rightarrow G_1\ G_2$ in the correct tree base, and f(H) is the sum of frequency of all rules rewriting H.

When there is not a correct tree base, if we can evaluate the score of each tree, we can think of the score of the tree as degree of confidence, then for each use of a rule, we could augment the corresponding counter by the confidence in the tree. This would result in the following algorithm:

(1) Find all parses $P_1, P_2, \ldots, P_k$ of each sentence, and compute their scores $S(P_i)$.

(2) For each use of the rule $H \rightarrow G_1\ G_2$ in parse P_i, add $S(P_i)$ to the corresponding counter:

$$C(H \rightarrow G_1\ G_2)=C(H \rightarrow G_1\ G_2)+S(P_i) \quad (2)$$

(3) Estimate the rule score:

$$P(H \rightarrow G_1\ G_2)= C(H \rightarrow G_1\ G_2)/C(H) \quad (3)$$

Where C(H) is the sum of the counters associated with all rules rewriting H. $S(P_i)$ is computed using the following formula:

$$S(P_i) = P(A \mid \Phi, BC, \Phi) \times P(B \mid \Phi, DE, \{F, G\}) \times P(C \mid \{D,E\}, FG, \Phi) \quad (4)$$

Referring to the Figure 2, A, B, C, D, E, F, G are respectively tagged syntactic and semantic information. Letter Φ represents the null symbol. $P(X_1|X_2, X_3X_4, X_5)$ is the score of X_1 rewritten with X_3X_4 when the left context is X_2 and the right context is X_5. For the computational feasibility, only a finite number of the left

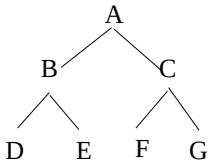

Figure 2. A parsing tree

and the right contextual symbols are considered. The detailed process may be seen in [9].

The algorithm presents the main problem, in which $S(P_i)$ depends on $P(X_1|X_2, X_3X_4, X_5)$. It is reasonable to apply this procedure starting from the rule probabilities of the correct tree base and improving the estimation in a loop.

The process to proofread errors is also proposed: First a threshold value for each kind of rule is set, all rules whose scores are under the threshold value are treated as wrong. Second, all sentences that related to the wrong rule are selected automatically from the corpora. We can find the errors according to the selected sentences and rectify the errors.

5 Self-learning of rules

IF rules are constructed only with introspective method, rules may be restricted. So large scale of corpora are used for learning new unrestricted rules. For segmentation rules, tagging rules, and morphology rules, rule template[16] is designed for learning. The rule templates are related to the 5 words of preceding and subsequent to the ambiguity string. And also, the word, part of speech ,the syntactic knowledge and the semantic knowledge are used in the template. The word, part of speech ,the syntactic knowledge and the semantic knowledge can be looked up from the lexicon automatically. All preceding and subsequent 5 words of the ambiguous strings are found with the large scale of corpora.

Because the analytic rules, transfer rules and synthesis rules are all complicated, the rules are acquired under the computer-aided process. An analysis rule is made up of context free grammar and condition. The actual example is written as Rule 3 of in Section 3. The condition is complex, and can not be learned automatically. The computer-aided process is performed as the following: e.g. For VP→V+N, or NP→V+N, both of them may be correct in actual text. We may select all sentences related to V+N from corpora automatically, then we also can formalize the rule deeper, such as condition or restriction for the rule according to the selected sentences.

6 Experiment Results

We have summarized 43 segmentation rules and 30

tagging rules with introspective method. The segmented corpora are made up of 3201 Chinese sentences. There are errors in 13 segmentation rules according to the corpora, and 40 segmentation rules are automatically learned from the corpora. The rate of correct of segmentation rules which are captured with introspective method is 69.77%, but can be improved with corpora. The 3201 Chinese sentences are also tagged, There are errors in 15 tagging rules according to the corpora, and 18 tagging rules are automatically learned from the corpora. The rate of correct of tagging rules which are captured with introspective method is 50%, but can be improved with corpora. We believe that the quality of rules can be improved when the scale of corpora becomes larger.

7 Conclusions

In this paper we propose a new approach to construction of rule base in combination of the linguistics knowledge and large-scale corpora. In our approach, the segmentation rule, tagging rule, analysis rule, transfer rule, synthesis rule and also the morphologic rule are made use of the lexical knowledge, syntactic knowledge and the semantic knowledge. The corpora are utilized to improve the quality of the rules and learn the new rules. The preliminary experimental results show that our approach may improve the speed to build the rule base and the quality of rules.

Acknowledgement

The paper is supported by the open task of National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences.

References

[1] Brill, Eric, “Some advances in transformation-based part of speech tagging”,In Proceedings of the Twelfth National Conference on Artificial Intelligence, pp. 722-727,1994

[2] Chen Shu-Chuan, Chang Jing-Shin, Wang Jong-Nae and Su Keh-yih, “ArchTran: A corpus-based statistics-oriented English-Chinese machine translation system”, in Proceedings of Machine Translation Summit III, Washington, D.C, pp.33-40, 1991.

[3] G. DeRose, “Grammatical Category Disambiguation by Statistical Optimization”, Computational Linguistics, Vol 14, No. 1, 1998.

[4] Gao Sheng, Jia Wen-ju et al, “A Mutual-information- based Approach to Rule Quantification”, Journal of Computer Research & Development, Vol.37, No.8, pp984-989, Aug, 2000.

[5] Harman, Robert, Jelinek, Fred, and Mercer, Robert, “Generating a grammar for statistical training.” In Proceedings,1990 Darpa Speech and Natural Language Workshop,1990.

[6] John Hutchins, “machine translation: past, present, future”, Ellis horwood limited, England，1986.

[7] Li Su-Jian, Liu Qun and Bai Shuo, “Chinese Chunking Parsing Using Rule-based and Statistics-based Methods”, Journal of Coputer Research and Development. Vol 39,No.4, Apr.2000.

[8] Liu Qun, Zhan Wei Dong and Chang Bao Bao, “computational model and language model of a Chinese-English machine translation”, Intelligent computer interference and application advancement, the publishing company of electronic industry, pp. 253-258,1998.

[9] Liu Ying, “Syntactic score and semantic score”, Journal of Chinese Information Processing, vo1 14,No. 4, pp.17-24. Apr. 2000.

[10] Su Keh-Yih, Chang hung-Hui and Lin,Yi-Chung,“A robustness and discrimination oriented score function for integrating speech and language processing”, In Proceedings, 2nd European Conference on Speech

Communication and Technology, Geneva, pp.207-210, 1991.

[11] Su Keh-Yih, J.N. Wang, W.H. Li and J.S. Chang,“ A New Parsing Strategy in Natural Language Processing Based on the Truncation Algorithm”, Proc. of Natl. Computer Symposium(NCS), Taipei, R.O.C., pp.580-586, 1987.

[12] Su Keh-Yih,Chang Jing-Shin, “Semantic and syntactic aspects of score function”, In proceedings 12th International Conference on Computational Linguistics, Budapest,pp. 22-27, 1988.

[13] Zhan WeiDong, A study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing, Tsinghua University Press ,GuangXi Technology Press, BeiJing, 2000.

[14] Zhou Ming, Hunag Changning et al, “A Chinese Parsing Model Based on Corpus Rules and Statistics”, Coputer Research and Development. Vol. 31,No.2, Feb. 1994.

[15] Zhou Qiang, Yu Shi wen, “A Kind of Multilevel Processing Method of Segmentation ang POS Tagging for Chinese Corpus”, Research and Application on Computational Linguistics,BeiJing Language University Press, BeiJing , pp. 126-131,1993.

[16] Zong Chengqing, Taiyi Huang and Bo Xu, “An Improved Template-Based Approach to Spoken Language Translation”, In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, pp. 440-443, October 2000.