A Hybrid Approach to Chinese-English Machine Translation

Liu Ying Liu Qun Zhang Xiang
Institute of Computing Technology
Chinese Academy of Sciences
BeiJing,100080
P. R. China

Chang Baobao
Institute of Computational Linguistics
Peking University
BeiJing,100871
P. R. China

Abstract — A hybrid method to Chinese-English machine translation is presented, a rule-based analysis is combined with statistical data. The rule-based lexical analyzer and syntactic analyzer leave some amount of ambiguity that are resolved using statistical approach. Hidden Markov Model(HMM) is used to return a score for each parts of speech, improved probabilistic context free grammar(PCFG) is used for syntactic scoring, and probabilistic score function is used for semantic aspect. In addition, parameters are usually estimated poorly when the training data is sparse. Smoothing the parameters is thus important in the estimation process, back-off procedure is used to smooth the parameters.

I. INTRODUCTION

Natural language processing, and artificial intelligence(AI) in general, have focused mainly on building rule-based systems with carefully hand-crafted rules and domain knowledge. Our own Chinese-English machine translation system, have used these techniques quite successfully, however, as the application of processing closed text is moved to applications of processing real text, We found challenges that questioned our previous assumptions and suggested combined method(combine rule-based approach with statistics-based approach).

For a rule-based approach, knowledge is induced by linguistic experts and is encoded in terms of rules, and knowledge is represented deeper, and quality and fluency of translation are better. But a huge amount of fine-grained knowledge is usually required to translate well, It is quite difficult for a rule-based approach to acquire such kinds of knowledge. In addition, the maintenance of consistency among the inductive rules is by no means easy, and the coverage is lower. Therefore, a rule-based approach, in general, fails to attain satisfactory performance for large-scale application.

In contrast, a statistics-based approach provides an objective measuring function to evaluate all possible alternative structures in terms of a set of parameters. Generally, parameters are estimated from a training corpus by using well-developed statistical theorems. The linguistic uncertainty problems can thus be resolved by a solid mathematical basis. Moreover, the knowledge acquired by a statistics-based method is always consistent because all the data in the corpus are jointly considered during the acquisition process. But knowledge is represented shallower than rule-based approach, and quality and fluency are worse than rule-based method.

From the above, a rule-based approach and a statistics-based approach all have advantages and disadvantages, so the two approaches are combined.

In a machine translation system, where the underlying grammar is large, there are many sources which may cause the system to become highly ambiguous. The system must choose a better syntax tree among all the possible ones to reduce the load of the subsequent processing. A better solution is to adopt the "Truncation Strategy" for MT system to restrict the number of parsing paths to be tried according to the relative preference of all the possible paths[1]. The truncation strategy is called the "Scoring Function". 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 [3] and a spoken language processing system[4].

The rule-based method uses a great deal of linguistic knowledge about lexical feature, syntax, and semantics.

0-7803-4253-4/97/$10.00 © 1997 IEEE

Linguistic knowledge is used for analysis and transformation processes. For analysis process, there are morphological analysis, syntactic analysis, and semantic analysis. However statistical theory may also be applied at all levels of MT. The hybrid method is used as follows, lexical scoring function based on Hidden Markov Model(HMM) is combined with morphological analysis, syntactic scoring function based on improved probabilistic context free grammar(PCFG) is combined with syntactic analysis[5], and semantic scoring function is combined with semantic analysis.

This paper is based on our Chinese-English MT system, and make use of the correct segmentation and syntactic structure to have supervised training for scoring function. At the same time, scoring functions are used to truncate undesirable analysis and hence improve the analysis speed and accuracy rate.

To avoid the sparse data problem, the parameters are first estimated by various parameter smoothing methods[6] and [7], and a back-off procedure that Katz proposed is used in our system[7].

II. MORPHOLOGICAL ANALYSIS COMBINED WITH LEXICAL SCORING FUNCTION

Non-deterministic algorithm is adopted in our system. Because deterministic algorithm can not provide the ability to backtrack, and the error of any step in translation process will result in failure of translation, but non-deterministic algorithm provide the ability to backtrack and produce all probabilities that include correct result and ambiguous results. The number of possible analyses associated with a given sentence is usually very large due to non-deterministic algorithm and the ambiguous nature of natural languages. But, It is desirable that only the best one be passed to the subsequent processing. In addition, translation time for a sentence is usually limited when a large number of sentences are translated in batch mode. Therefore, it is important to obtain the best syntax tree which has the best annotated semantic interpretation within a reasonably short time. This is only possible with an intelligent parsing algorithm which can truncate undesirable analysis as early as possible and avoid wasting time in parsing those ambiguous constructions that will eventually be discarded.

Because there isn't any word marker in Chinese texts, A function of morphological analysis phase is segmentation. Input strings without word marker are segmented into word strings. Bi-directional maximum matching with backtracking algorithm is used in segmentation. Another function of morphological analysis phase is to assign part of speech to each word. Suffix and overlap of Chinese are processed with rule-based method, and part of speech are assigned to each word by looking up dictionary.

In this phase, more than one part of speech for a word and a lot of kinds of segmentation for a sentence may be generated with non-deterministic algorithm, so morphological analysis is combined with lexical scoring function. Many results are deleted using lexical scoring function according to maximum likelihood principle.

Lexical scoring function is based on Hidden Markov Model(HMM). For an input sentence $w=w_1w_2\cdots w_n$, where $w_i(i=1,n)$ stands for the ith word of the input sentence. $c=c_1c_2\cdots c_n$, where c_k (k=1,n) stands for the part of speech assigned to w_k. According to Bayesian theorem, the following equation is right.

$$P(c|w)=\frac{P(c)P(w|c)}{P(w)} \qquad (1)$$

Since P(w) is the same for all possible lexical sequences, it can be ignored without affecting the final results. Like the standard tagging procedures[8], the probability terms P(w|c) and P(c) in (1) can be approximated as follows, respectively:

$$P(w|c)=\prod_{i=1}^{n} P(w_i|c_i)$$

$$P(c)=\prod_{i=1}^{n} P(c_i|c_1^{i-1})$$

$$\approx \prod_{i=1}^{n} P(c_i|c_{i-1}) \qquad \text{bigram model}$$

$$\text{or} \approx \prod_{i=1}^{n} P(c_i|c_{i-1},c_{i-2}) \qquad \text{trigram model}$$

Lexical score function is defined as follows:

$$S_{lex}=\prod_{i=1}^{n} P(c_i|c_{i-1})\times P(w_i|c_i) \qquad \text{bigram model} \quad (2)$$

$$\text{or} = \prod_{i=1}^{n} P(c_i|c_{i-1}, c_{i-2}) \times P(w_i|c_i) \quad \text{trigram model} \quad (3)$$

III. SYNTACTIC ANALYSIS COMBINED WITH SYNTACTIC SCORING FUNCTION

Improved chart parsing algorithm is adopted in syntactic analysis. Word and phrase structure are described with complex feature structure, and lexical information, grammatical information and semantic information may be included in complex feature structure. Unification-based operation is adopted for two different feature structure.

A lot of phrase structures may be generated in analytic phase with non-deterministic algorithm, in order to find the correct structure and improve translation speed, syntactic scoring function is used. Syntactic scoring function is based on improved probabilistic context free grammar(PCFG). For each phase structure reduced in analytic phase, score is given with scoring function. The best phrase structure is retained and the others are canceled according to maximum likelihood principle.

Improved PCFG is derived by improving PCFG. Contexts are not considered for PCFG, but contexts are considered for improved PCFG. Left bottom contexts and right bottom contexts are considered for improved PCFG. The process that phrase structure probabilities are computed is independent of syntactic analysis algorithm for improved PCFG, but depends on syntactic analysis algorithm in [2] and [4].

Syntactic scoring function is defined as follows, Fig.1 represents the given syntactic tree $Tree_x$.

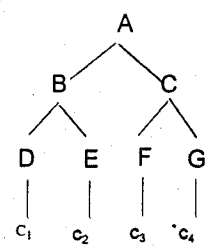

Fig. 1 $Tree_x$

$S_{syn}(Tree_x) = P(A|\{\Phi\}, BC, \{\Phi\}) \times P(B|\{\Phi\}, DE, \{c_3, c_4\}) \times P(C|\{c_1,c_2\}, FG, \{\Phi\}) \times P(D|\{\Phi\}, c_1, \{c_2,c_3,c_4\}) \times P(E|\{c_1\}, c_2, \{c_3,c_4\}) \times P(F|\{c_1,c_2\}, c_3, \{c_4\}) \times P(G|\{c_1,c_2,c_3\}, c_4, \{\Phi\}) = P(A|l_A, BC, r_A) \times P(B|l_B, DE, r_B) \times \cdots \times P(G|l_G, c4, r_G)$ (4)

Where Φ represents the null symbol, l_Z (z =A,B,C,D,E,F,G) and r_Z represent the left and right contexts to be consulted when Z is reduced. For computational feasibility, only a finite number of left and right contextual symbols are considered in (4). One left and right contextual symbols are considered in our system.

IV. SEMANTIC ANALYSIS COMBINED WITH SEMANTIC SCORING FUNCTION

Semantic analysis mainly focuses on case frame analysis in our system. Case frames associated with verbs are used, a set of cases associated with verbs is: nominative, accusative, dative, location, way, instrument. Case frame is a particular organization of knowledge intrinsic to a verb. Semantic analysis using case frame provides a kind of means for various types of disambiguation, for example: multiple noun senses, multiple verb senses, ambiguity of extraction site, ambiguity of prepositional phrase attachment, and so on.

But good results can not be achieved only using rule-based method, semantic scoring function is combined with semantic analysis. The semantic information for each phase structure comes from the most important N children of the structure via some mechanism, say semantic unification. For simplicity, the most important child of the structure is considered in our system, which is called the head phrase. The head phrase is marked using symbol" !" in syntactic analysis phase. For a terminal word, the semantic information is assigned a semantic tag from the lexicon. So it is easy to explore the feature co-occurrence frequencies among the phrase structure. After a syntax tree is properly annotated, we could then evaluate the semantic score of $Tree_X$ in Fig.1 approximately as:

$S_{sem}(Tree_x) = P(A'|\{\Phi\}, B'C', \{\Phi\}) \times P(B'|\{\Phi\}, D'E', \{c_3, c_4\}) \times P(C'|\{c_1,c_2\}, F'G', \{\Phi\}) \times P(D'|\{\Phi\}, c_1, \{c_2,c_3,c_4\}) \times P(E'|\{c_1\}, c_2, \{c_3,c_4\}) \times P(F'|\{c_1,c_2\}, c_3,$

$\{c_4\})\times P(G'|\{c_1,c_2,c_3\}, c_4, \{\Phi\})=P(A'|\ l_A, B'\ C', r_A)\times P(B'|\ l_B, D'\ E', r_B)\times\cdots\times P(G'|l_G,c4,r_G)$ (5)

where Z' (Z=A,B,C,D,E,F,G) represents the phrase Z annotated with semantic information.

During the course of estimating parameters with statistical approach, A probability of zero is assigned to any structure or sequence of tags that did not occur in the training data. But such sequence or structure may occur if other texts are considered. A probability of zero for a sequence creates problems because the model becomes useless for such sentences. To avoid the sparse data problem, a back-off procedure that Katz proposed is used in our system[7].

V. EXPERIMENT RESULTS

Grammatical Information Dictionary of Contemporary Chinese is used in our system[9], and there are twenty-six grammatical classification in the dictionary. Semantic classification is mainly aimed at verb, noun, and adjective.

Three corpora are used to train with the different approach, which are syntactic analysis of rule-based method, syntactic analysis of hybrid method, and semantic analysis of hybrid method. Corpus t_1,corpus t_2, and corpus t_3 are selected from the sentences that are translated in our system.

We will evaluate the three corpora in three measures: accuracy rate, ambiguity rate, and selection power. The measure of accuracy rate and ambiguity rate of parse tree has been widely used in the literature. Selection power(SP) is proposed in [10], which is defined as the average selection factor of the disambiguation mechanism on the task of interest.

From the following three table, we can see that accuracy rate is the lowest when sentences are analyzed only using rule-based method(see Table I), and the accuracy rate is the highest when sentences are analyzed using syntactic analysis, semantic analysis and their corresponding scoring functions. Moreover, The selection power of Table III becomes smaller than Table II when semantic analysis and semantic scoring function are used, which imply better disambiguation power .

TABLE I TEST CORPORA AFTER SYNTACTIC ANALYSIS OF RULE-BASED APPROACH

Text	Words	Ambiguity rate	Accuracy rate
t1	1601	17.2%	75.8%
t2	1433	15.7%	74.9%
t3	2240	16.9%	75.1%

TABLE II TEST CORPORA AFTER SYNTACTIC ANALYSIS OF HYBRID APPROACH

Text	Words	Accuracy rate	Selection power
t1	1601	81.1%	0.21
t2	1433	82.3%	0.20
t3	2240	83.5%	0.22

TABLE III TEST CORPORA AFTER SEMANTIC ANALYSIS OF HYBRID APPROACH

Text	Words	Accuracy rate	Selection power
t1	1601	87.7%	0.20
t2	1433	86.9%	0.19
t3	2240	88.6%	0.18

VI. CONCLUSIONS AND FUTURE WORK

This paper provides a hybrid method to a Chinese-English machine translation system. The hybrid method combines rule-based method and statistics-based method. When the sentences are analyzed with rule-based method, score acquired with statistics-based method is made use of truncate undesirable analysis, and hence improve translation speed and accuracy rate.

VII. REFERENCES

[1] 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., 1987,pp.580-586.

[2] Su Keh-Yih and Chang Jing-Shin, " Semantic and syntactic aspects of

score function" , In Proceedings, 12th International Conference on Computational Linguistics, Budapest, 1988, pp. 22-27.

[3] 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, 1991, pp.33-40.

[4] Su Keh-Yih, Chang hung-Hui and Lin,Yi-Chung(1991)," 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.

[5] P.Laface and R.De Mori, *Speech Recognition and Understanding*, NATO ASI Series, Vol. F 75, pp. 345.

[6] I.J. Good, " The population frequencies of species and the estimation of population parameters" , Biometrika,40, 1953,pp. 237-264.

[7] Katz Slava M., " Estimation of probabilities from sparse data for the language model component of a speech recognition" , IEEE Transactions on Acoustics, Speech and Signal processing, ASSP-35,1987, pp. 400-401.

[8] Church Kenneth, " A stochastic parts program and noun phrase for unrestricted text" , In Proceedings, IEEE 1989 International Conference on Acoustic, Speech, and Signal Processing. Glasgow, pp. 695-698.

[9] Yu Shiwen et al., " The Specification of the Grammatical Information Dictionary of Contemporary Chinese" , journal of Chinese information processing Vol.10 No.2, pp.1-22(in Chinese).

[10]Tung-Hui Chiang, Yi-Chung Lin and Keh-Yih Su, "Robust learning, Smoothing, and Parameter Tying on Syntactic Ambiguity Resolution", Computational Linguistics,Volume 21, Number 3, pp.321-349