TransEasy: a Chinese-English Machine Translation System Based on Hybrid Approach

Qun Liu[1] and Shiwen Yu[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
P.O.Box 2704, Beijing 100080, China
liuqun@mtgroup.ict.ac.cn
[2] Institute of Computational Linguistics, Peking University,
Beijing 100871, China
yusw@pku.edu.cn

Abstract. This paper describes the progress of a machine translation system from Chinese to English. The system is based on a reusable platform of MT software components. It's a rule-based system, and some statistical algorithms are used as heuristic functions in parsing as well. There are about 50,000 Chinese words and 400 global parsing rules in the system. The system got a good result in a public test of MT system in China in Mar. 1998. It is a research vehicle up to now.

1 Introduction

The current level of Chinese-English machine translation is much lower than that of English-Chinese machine translation in China [1], because the analysis of Chinese is more difficult than that of English. There are several commercial Chinese-English MT systems in the software market, but most of them are of little practical use value.

TransEasy is a new Machine Translation System from Chinese to English we developed. We try to make it a practical Chinese-English MT system.

The basic information of the system is as follows:

- System Name: TransEasy
- Developer: Institute of Computing Technology, Chinese Academy of Sciences
  Institute of Computational Linguistics, Peking University
- System Builders and contacts:
  Qun Liu, Shiwen Yu, etc. (the authors of this paper)
- System Category: research vehicle
- System Characteristics:
  Translation Speed: about 40,000 Chinese Characters per hour
  Domains Covered: no specific domain
  Input Formats: plain text
  Output Quality: about 70% of the translations are understandable

- Resource: Lexicon: 50,000 entries of Chinese Words
  Rulebases: less than 400 rules in the global parsing rulebase
- Hardware and Software: IBM/PC compatible
  Windows 95 or NT with Chinese environment (GB Code)

2 General MT Development Platform

The development of the Chinese-English MT System is based on a General MT Development Platform [7].

The development of a MT system is tedious work. MT systems for different language pairs, or even MT systems of different kinds, have many common data structures and algorithms. Making these data structures and algorithms reusable, will greatly reduce the work of development. The Platform provides this kind of reusability. The Platform contains a lot of software components. These software components implement the most frequently used data structures and algorithms in MT systems. The relations between these software components are clearly defined. These software components can be used independently. The platform supports the development of MT systems of different language pairs, and provides the API functions for specified natural language. The software components are easily extended to fit different types of MT systems.

3 Algorithms

We use the transfer approach in the system. The process of the translation is: morphological analysis, structural analysis, transform, structural generation and morphological generation

The morphological analysis of Chinese is much more difficult than that of English. There are two kinds of ambiguities in this phase: segmentation ambiguity and tagging ambiguity. The morphological analysis of this system includes four steps:

1. overlapped words processing and dictionary consulting
2. rule-based disambiguation of segmentation
3. rule-based disambiguation of tagging
4. statistical disambiguation of segmentation and tagging

In the last step we use a HMM model. A threshold is used to discard the tags with low probability [8]. Because we use a nondeterministic algorithm, we need not to eliminate all the ambiguity in this phase. The ambiguities that cannot be eliminated are kept to the phase of structural analysis.

The structural analysis relies mainly on syntax features and uses the semantic information for reference. Semantic analysis is done simultaneously with the syntactic analysis.

A modified Chart Parsing algorithm is used in the structural analysis phase. We use a statistical algorithm to improve the efficiency of parsing. A probability is given

to each potential node by scoring function, which use PCFG as the model [8]. When the parsing of the source sentence failed, a kind of "soft failure" technique is used to give the most likely result.

In the phases of transform and structural generation we use the algorithm which we call it "local sub-tree transform algorithm".

4 Knowledge Bases

There are 10 knowledge bases of different kinds used in the system.

Language model is very important in machine translation. We defined a special knowledge base called *Language Model* to define all the lexical, syntactic and semantic categories and attributes of source and target languages. All the linguistic symbols used in other knowledge bases must be defined in the Language Model. The language model of this MT system chiefly originates from the "Grammatical Knowledge Base of Contemporary Chinese"[10]. The analysis of this MT system relies mainly on syntax features and uses the semantic information for reference.

There are six *rulebases* used in different phases of the translation. The rules in rulebases are global rules, which are available in all situations. Local rules, which are restricted with certain word, are stored in the dictionary. The format of global rules and local rules are the same. The basic format of all the rules is: *Pattern + Unification Equations,* somewhat like the format of rules in LFG grammar.

A bilingual *dictionary* is used in the MT system. It contains not only the meaning items, but also the local rules restricted with the words.

The *example base* is actually a corpus for the MT system. On one hand, it stores collected bilingual text. On the other hand, it stores the entire sentences that have been translated by the MT system. The source text, target text, source trees and target trees are stored. Users can modify the translations and trees stored in it. The example base is used to evaluate the translation of the system and to generate the statistical data that is used in the scoring functions in lexical analysis and parsing. We are developing another example-based translation engine for this system at the same time.

The statistical database stores the statistical data used in the scoring functions in lexical analysis and parsing. The data is generated from the example base.

5 Current State and Future Works

A training set of 4,000 Chinese sentences has been used to train the system. These sentences are selected on purpose, which cover the most frequently used patterns of sentences in Chinese. About 90% of the translations of the sentence in the training set are understandable, which means people may know the meaning of the source text from the target text. In the open test of Chinese-English Machine Translation Systems held by the Steering Committee of Chinese Hi-Tech R&D Plan in March 1998, this system got a good result, about 71% of the translation is understandable. Usually

the result English sentences are not very proper in grammar. The most common mistakes are on articles, number of nouns and tense of verbs, because there are no corresponding linguistic features in Chinese and it is very difficult to generate these features in the translation properly in high correctness.

The future work we plan to do includes: use more Chinese sentences (about 10,000) to train the system, expand the corpus, improve the statistical algorithm, and add an example-based translation engine.

6 Acknowledgements

This project is supported by the Chinese National High-Tech R&D Plan. The development of this system is under the leadership of Prof. Xiang Zhang. We would like to thank all the members of our research group, especially Weidong Zhan, Ying Liu, Baobao Chang, Bin Wang, Hui Wang, Qiang Zhou and Yu Ye for their hard work on the project.

References

1. Duan Huimin and Yu Shiwen, Test Report on MT Systems, on (Chinses) Computer World Newspaper, 1996.3.25, p183. (In Chinese)
2. Feng Zhiwei, New Discussion on Machine Translation of Natural Language, Press of Language and Character, 1995. (In Chinese)
3. Feng Zhiwei, The Computer Processing of Natural Language, Shanghai Press of Foreign Language Education, 1996. (In Chinese)
4. Gazdar G., Mellish C., Natural Language Processing in Lisp, Addison-Welsley Publishing Company, 1989
5. Kay, Martin, Unification in Grammar, In V.Dahi & P.Saint-Dizier(Ed.) Natural Language Understanding and Logic Programming, Elseview Science Pub, 1985.
6. Liu Qun, Zhan Weidong, Chang Baobao, Liu Ying, The Computational Model and Linguage Model of a Chinese-English Machine Translation System, Advance on Intelligent Computer Interface and Application, Press of Electronic Industry, 1997. (In Chinese)
7. Liu Qun, Zhang Xiang, Research Method Based on Software Component for Machine Translation, Language Engineer, Press of Tsinghua University, 1997. (In Chinese)
8. Liu Ying, Liu Qun, Zhang Xiang, Chang Baobao, A hybrid approach to Chinese-English machine translation, IEEE ICIPS'97 Conference, 1997.
9. Shen Yang, Zhen Ding'o (Ed.), Research on Valence Grammar on Modern Chinese, Press of Peking University, 1995. (In Chinese)
10. Yu Shiwen, Specification on "Syntax Information Dictionary on Modern Chinese", Journal of Chinese Information Processing, Vol 10, No 2, pp1-22, 1996. (In Chinese)
11. Yu Shiwen, Some Aspect on Computational Linguistics, Applied Linguistics, No. 3, 1993. (In Chinese)