A News-oriented Chinese-English Machine Translation System[1]

Qun Liu[†‡] **Baobao Chang**[†] **Weidong Zhan**[†] **Qiang Zhou**[#]

[†] *Institute of Computational Linguistics, Peking University*
[‡] *Institute of Computing Technology, Chinese Academy of Sciences*
[#] *State Key Laboratory of Intelligent Technology and System, Tsinghua University*
liuqun@ict.ac.cn, {chbb,zhwd}@pku.edu.cn, zhouq@s1000e.cs.tsinghua.edu.cn

Abstract: This paper introduces a News-oriented Chinese-English Machine Translation system. Firstly, general information of the system is given. Secondly, the architecture of the system — the micro-engine pipeline is introduced. Thirdly, the several main micro-engines in the system are discussed, including micro-engines for Chinese lexical analysis, example-based micro-engine, and rule-based micro-engine. Fourthly, linguistics resources used in this system are described. Finally, we will give the conclusion and the future works.

Keywords: **Multi-Engine Machine Translation, Chinese Information Processing**

面向新闻领域的汉英机器翻译系统[1]

刘群[†‡] **常宝宝**[†] **詹卫东**[†] **周强**[#]

[†] *北京大学计算语言学研究所*
[‡] *中国科学院计算技术研究所*
[#] *清华大学智能技术与系统国家重点实验室*
liuqun@ict.ac.cn, {chbb,zhwd}@pku.edu.cn, zhouq@s1000e.cs.tsinghua.edu.cn

摘要: 本文介绍了一个面向新闻领域的汉英机器翻译系统。首先介绍了系统的总体情况。其次，介绍了该系统的总体结构——微引擎流水线。再次，讨论了系统中几个主要的微引擎了，包括词法分析微引擎，基于实例的翻译微引擎和基于规则的翻译微引擎。然后，介绍了系统中所使用的语言资源。最后给出了结论和下一步的工作计划。

关键词：**多引擎机器翻译，中文信息处理**

1 Introduction

The News-oriented Chinese-English Machine Translation System is a project supported by the national key fundamental research schema. Institute of Computational Linguistics of Peking University, Institute of Computing Technology of Chinese Academy of Sciences and State Key Laboratory of

[1] This project is supported by the National Key Fundamental Research Schema.
The project number is G1998030507-4.
Many thanks to Prof. Sun Maosong for his support to this project.

Intelligent Technology and System of Tsinghua University undertake this project.

The leader of the project is Prof. Yu Shiwen. The project started on January 2000. In this system we adopt a multi-strategy approach, which we call it a micro-engine pipeline approach. Now the first version of the system is completed. The system has a dictionary of about 50,000 Chinese words in this version. It can translate 1,000 words per minute on a computer with an Intel Pentium III 500 CPU.

2 Multi-micro-engine Approach

Various approaches are used in different MT systems, such as rule-based approach, statistical-based approach, example-based approach, transfer-based approach, interlingua-based approach, and etc. Each approach has its advantages and disadvantages. A lot of researchers realized that using a hybrid approach would archive better result than using a single approach. Many systems used a multi-engine approach in MT system. However, there are many different methods to integrate different MT engines to a system. [Frederking and Nirenburg 94] proposed a typical diagram of multi-engine MT, where every MT engine translate the input sentence in parallel, and put the translations of the source sentence segments with a score into a Chart-like structure, according to the position of the source constituent. Then the system uses a dynamic program algorithm to find the best combination of the translations. [Zhang and Choi 1999] use different engines in different phases in a transfer-based MT system. In analysis phase, rule- & statistics-based engines are started up. In transfer phase, patter- & statistics-based engines are started up. In synthesis module, rule-based engine is started up.

In our system, we proposed a general-purpose multi-engine MT architecture — Micro-Engine Pipeline, which is defined as below:

- A micro-engine pipeline consist of an array of micro-engines;
- All the micro-engines share a chart data structure;
- There are two kind of micro-engine: recognizer and selector;
- A recognizer should implement two function:
  1) Recognize: To produce a new constituent (edge) according the existing constituents in the chart;
  2) Translate: To translate a constituent produced by itself; the translation function may call the translate function of the recognizer which produce the children constituents of the constituent to be translated.
- A selector should implement one function:
  1) Select: To select best sequences of constituents on the charts;
- The translation algorithm:
  1) Assign the first micro-engine in the micro-engine pipeline to TheEngine;
  2) If TheEngine is NULL, then translation fails, return NULL;
  3) If TheEngine is a recognizer, call its recognize() function repeatedly, until a constituent cover the whole sentence is produced, or no new constituent is produced. In the first case, goto step 6);
  4) If TheEngine is a selector, call its select() function, discards all the constituents which do not appear in the returned sequences.
  5) Assign the next micro-engine in the micro-engine pipeline to TheEngine, goto 2);
  6) Assign the constituent which cover the whole sentence to TheRoot;
  7) Call the translate() function of TheEngine to translate TheRoot;
  8) Translation success, return the target sentence.

3 Micro-engines

The actual micro-engine pipeline in our Chinese-English MT system consist of the following micro-engines:

1) Example-based Exact-maching Recognizer: to search the bilingual corpus to find the exact-matched example, and return the translation directly;
2) Dictionary Lookup Recognizer: the recognizer to look up the dictionary and process the overlapping Chinese words. Since no space exists between Chinese words, all the possible words in the dictionary are added to the chart. This process is so-called full segmentation;
3) Named Entity Recognizer: the recognizer to recognize named entity which is not recorded in the dictionary, such as Chinese people names, Chinese place names, foreign names, numbers, and etc.
4) Segmentation and POS Tagging Selector: to do Chinese word segmentation and POS tagging, using a hybrid of rule-based approach and statistical-based approach, which we will introduce in other papers;
5) Example-based Fuzzy-matching Recognizer: to search the bilingual corpus to find the fussy-matched example, and translate the sentence by an example-based approach. The bilingual corpus used here is sentence-aligned. This recognizer is still under construction;
6) Rule-based Recognizer: a traditional rule-based MT engine, using an LFG-like grammar and a chart parser with unification supported. The translation function of this engine also adopts a rule-based approach.

7) Fail-soft Recognizer: if all of the above recognizer do not product a constituent covering the whole sentence, fail-soft recognizer will try to find a best sequence of the exist constituents and produce a new constituent, with all the constituents in the best sequence as the children of the new constituents. The algorithm to find the best sequence is just like the "chart walk" algorithm introduced in [Frederking and Nirenburg, 1994].

4 Resources

In the micro-engine pipeline approach, each micro-engine uses its own resources. For example, the Name Entity Recognizer needs dictionaries of Chinese people names, Chinese place names and foreign names, and the Rule-based Recognizer needs rulebases, and etc. Here we will not example all these resource in detail. We will only introduce the dictionary and the corpus.

A core dictionary and a extend dictionary is used in the system. There are about 50,000 Chinese words in the core dictionary. The information of the core dictionary is very rich. The grammatical information of Chinese words of this dictionary is mainly extracted from the "The Grammatical Knowledge-base of Contemporary Chinese" [Shiwen Yu, et al, 1998], and the semantic information is based on an ontology and semantic frame schema, which was introduced in [Hui Wang, et al, 1998]. The extend dictionary contains about 200,000 Chinese words. Information in the extend dictionary is rather simple, i.e., the Chinese and English word forms and POSs. Now the extend dictionary is under construction and is not integrated to the system yet.

A monolingual corpus and a bilingual corpus are used in the system. The monolingual corpus is the "Peoples Daily Corpus" with Chinese words segmented and POS tagged [Shiwen Yu, 2000]. Now it is used to train a HMM model in Segmentation and POS Tagging Selector. The bilingual corpus consists of two parts. One part contains news, Chinese government white papers and editorials of Xinhua News Agency. There are about 40,000 sentence pairs in this part. The other part contains about 180,000 sentence pairs collected from various sources. The whole bilingual corpus is sentence-aligned and used in the Example-based Exact-matching Recognizer and Example-based Fuzzy-matching Recognizer. Now we are doing further chunk-level alignment on the former part of the bilingual corpus.

5 Conclusion and future work

The News-oriented Chinese-English MT System is developed on the basis of a traditional rule-based MT system [Qun Liu, Shiwen Yu, 199810], which is used as Rule-based Recognizer in the micro-engine pipeline. We have test the two systems in a small test set, and find that the translation of the new system is a little better that of the old system, mainly because the Named Entity Recognizer is added.

Compared with other implementation methods of the multi-engine MT, the micro-engine has many advantages. The system modularity is better. When we add a new micro-engine to the MT system, we need not to change the algorithm of the whole system. The micro-engines could be very specialized; it needs not to be almighty MT engine. That means, it could adopt a very special algorithm to deal with a very special language phenomena. The relation among the micro-engines is more cooperative, rather than competitive. The micro-engines are used in serially, rather than in parallel. So the later micro-engine can make use of the result generated by the former engine.

Our future work include:
1) Complete the Example-based Fuzzy-matching Recognizer.
2) Improve the Rule-based Recognizer by add statistical information to the LFG-like rulebases and research on the WSD technology to select the best translations of the words;
3) Add a Template-based Recognizer to system. The translation templates will be acquired from the bilingual corpus automatically and be corrected manually [Jiang Zhang, 2000];
4) Add a Chunk Recognizer to the system;
5) Pay more attention to the resource construction, especially on the bilingual corpus.

References

[1] Robert Frederking and Sergei Nirenburg. Three Heads are Better than One. In Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, Germany.

[2] Robert Frederking et al. Integrating Translations from Multiple Sources with the Pangloss Mark III Machine Translation System. In Proceedings of the First Conference for Machine Translation in the Americas (AMTA), Columbia, Maryland, October, 1994.

[3] Vasileios Hatzivassiloglou, and Kevin Knight, Unification-Based Glossing, In: Proc. 14th Int. Joint Conf. Artificial Intelligence, 1995

[4] Manny Rayner and David Carter. Hybrid Processing in the Spoken Language Translator. In Proceedings of ICASSP-97, pages 107-110, Munich, Germany. 1997.

[5] Christopher Hogan and Robert E. Frederking. An Evaluation of Multi-engine MT Architecture. In David Farwell et al., editors, Machine Translation and the Information Soup, pages 113-123, Third Conference of the Association for Machine Translation in Americas (AMTA), Langhorne, PA. USA, October 1998.

[6] 周明，中－日机器翻译系统 J－北京，收录于：黄昌宁，董振东主编，计算语言学文集，第 312～319 页，清华大学出版社，1999.10
Ming Zhou, J-Beijing Chinese-Japanese Machine Translation System, In: Changning Huang, Zhengdong Dong, eds, A collection on Computational Linguistics, pages 312-319, Tsinghua University Press, Oct. 1999

[7] Qun Liu and Shiwen Yu. TransEasy: A Chinese-English Machine Translation System Based on Hybrid Approach. In David Farwell et al., editors, Machine Translation and the Information Soup, pages 514-517, Third Conference of the Association for Machine Translation in Americas (AMTA), Langhorne, PA. USA, October 1998.

[8] 刘群，俞士汶，汉英机器翻译的难点分析，International Conference on Chinese Information Processing，黄昌宁主编，1998 中文信息处理国际会议论文集，第 507-514 页，清华大学出版社，1998.11
Qun Liu, Shiwen Yu, Difficulties in Chinese-English Translation, In: Proceedings of Conference on Chinese Information Processing 1998, pages 507-514, Tsinghua University Press, Nov. 1998

[9] Qun Liu, A Chinese-English Machine Translation System Based on Micro-Engine Architecture, An International Conference on Translation and Information Technology, Hong Kong , Dec. 2000

[10] Min Zhang and Key-Sun Choi, Pipelined multi-engine Machine Translation: accomplishment of MATES/CK system, In: Proceedings of TMI'99, page 228, 1999

[11] 俞士汶，朱学锋，王慧，张芸芸，现代汉语语法信息词典详解，清华大学出版社，1998
Shiwen Yu, Xuefeng Zhu, Hui Wang, Yunyun Zhang, the Grammatical Knowledge-base of Contemporary Chinese – A Complete Specification, Tsinghua University Press, 1998

[12] 王慧，詹卫东，刘群，《现代汉语语义词典》的概要及设计，International Conference on Chinese Information Processing，黄昌宁主编，1998 中文信息处理国际会议论文集，第 507-514 页，清华大学出版社，1998.11
Hui Wang, Weidong Zhan, Qun Liu, Design and Essentials of the Semantic Knowledge-base of Contemporary Chinese, In: Changning Huang, eds, Proceedings of International Conference on Chinese Information Processing, pages: 507-514, Tsinghua University Press, Nov.1998

[13] 俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范，多语言信息处理国际会议，2000'ICMIP 论文集，18-24, 2000 年 8 月，新疆乌鲁木齐
Shiwen Yu, Xuefeng Zhu, Huiming Duan, Specification of large-scale modern Chinese corpus, In: Proceedings of ICMIP'2001, pp 18-24, Aug. 2000, Urumqi

[14] 张健，基于实例的机器翻译的泛化方法研究，中国科学院计算技术研究所硕士论文，2001
Jian Zhang, Research on the Generalization Method of the Example-based Machine Translation, Master's dissertation of Institute of Computing Technology, Chinese Academy of Sciences, 2001