

The Application of Machine Translation in China-America Digital Academic Library

Huang Chen, and Chen Haiying

Zhejiang University Libraries, Hangzhou , 310027, PR China

Abstract — This paper briefly introduces the main ideas of machine translation (MT) technique, then discusses the application of MT to the China-America Digital Academic Library (CADAL).

Index Terms — CADAL, machine translation, digital library

I. INTRODUCTION

The China-America Digital Academic Library (CADAL) Project was launched by China-US scientists (<http://www.cadal.cn>) with a goal of digitizing one million books for a digital library. The project is a major element in the China Education Ministry's "Tenth Five-year Plan," and is intended to provide digitized resources for teaching and research as well as to prompt the sharing of those resources.

The aim of the CADAL digital library is to provide an information service. Nowadays users are no longer satisfied with much of the information retrieved through the internet; thus, there is a need for integrated and processed information from various media. The ideal information service should provide the knowledge that the user seeks, as well as a solution to the user's problem. Therefore, CADAL not only provides digitized books, replacing traditional printed ones, but also processes the digitized resources to extract relevant information, thus providing more service to the user. Machine translation (MT) is a service that CADAL intends to adopt to provide bilingual or multi-lingual translations.

II. MACHINE TRANSLATION THEORY AND TECHNOLOGY

As one of the earlier research branches of Natural Language Understanding (NLU), MT is a process used to translate one natural language into another. The software that completes such a task is named Machine Translation System.

Warren Weaver, director of the Natural Sciences Department of the Association of America Rockefeller Fund, published a memorandum in 1949 entitled *Translation*, to raise the issue of machine translation. With the development of both classical linguistic theory and modern computational linguistic theory, some

commercial MT systems appeared in the later 1980s, such as the American SYSTRAN System, the METAL System by Texas University and Siemens Company, ATLAS by HITACHI Company, and the CETA System by Grenoble University.

MT research in China was listed in the governmental "Science Development Compendium" as "MT/Natural Language Mathematics Theory" in 1956. In the later 1980s and early 1990s, two MT systems of practical value appeared: the "KY-1" English-Chinese MT System by the Academy of Sciences of Military Affairs, and the "863-IMT" English-Chinese MT System by the Institute of Computing Technology, a division of the Chinese Academy of Sciences.

In recent years, MT systems are usually installed with professional dictionaries, run on the internet and have a user-friendly interface. MT research for new applications, such as speech translation systems, is also underway.

Traditional MT belongs to Knowledge Based MT (KBMT) [1], also called the Rule-based Method. Linguistic rules, which cover a wider domain than the training corpus, are constructed by specialists. These rules and their resulting systems tend to make more sense for human beings and can be adjusted quickly. However, they suffer from the following drawbacks:

(1) Rule-based methods are too labor intensive, and rule construction requires extensive linguistic training.

(2) Rule consistency is difficult to maintain, even for the same designer. Furthermore, common sense is often difficult to encode.

(3) Rules designed by different experts can sometimes contradict each other and thus affect the overall system performance.

Facing the challenges of KBMT, Professor M. Nagao at Tokyo University proposed an analogy-based MT method in 1984 [2]. Many researchers extended Nagao's method to form a so-called Example-based MT (EBMT). The basic idea of EBMT is simple: given an input passage S in a source language and a bilingual text archive, where text passages S' in the source language are stored, aligned with their translations into a target language, T' , S is compared with the source-language "side" of the archive. The "closest" match for passage S' is selected and the translation of this closest match, the passage T' is accepted as the translation of S .

Statistical MT [30 40 5] can be seen as one variant of EBMT. The basic idea of statistical MT is that the translation is based on the statistical probability of the words of the same text in two languages (parallel corpora). When such texts in two languages exist, the probabilities of the words can be counted, and the translation system can be "taught to translate" by using the probabilities.

III. APPLICATIONS OF MT IN CADAL

A. The Goal

CADAL is making use of MT in a number of ways:

- (1) Important information, such as a book's title or authors, is translated manually, or first translated by MT systems and then verified manually;
- (2) As the cornerstone of CADAL's system, MT provides instant service such as translation of contents indexed by XML;
- (3) Integrating MT with other services, such as multilingual information retrieval and special words retrieval.

In the CADAL server (<http://www.cadal.zju.edu.cn/>), we applied a bilingual service engine to support the metadata retrieval between English and Chinese. This engine provides instant translation of book profiles. On the left of Figure 1 is a book profile in Chinese. When the user clicks the "English Profile" (link words in pink below the image of the book's cover), the system displays its English profile translated by the MT engine. The result is shown to the right of Figure 1.

B. MT evaluation in CADAL

Based on the goal of CADAL, we evaluated a number of existing MT systems. These include systems developed by IBM, Carnegie Mellon University, USC/ISI, RWTH Aachen University, Microsoft (Redmond) and the Institute of Computing Technology, a division of the Chinese Academy of Sciences.

Results show that the performance of MT Systems created by RWTH Aachen University, CMU and ISI is superior to even that by SYSTRAN, but strategies of MT vary from system to system. RWTH Aachen University adopted the SBMT model, and improved the traditional noise channel based paradigm into the maximum entropy model [6], Their MT System also further enhanced the words- based alignment model to a phrase-based alignment model. Mega2RADD by CMU integrates SBMT with EBMT through a translation engine, and provides the optimized translation result. Re2Write by ISI takes IBM-4 statistical model as the prototype; the translation quality is improved by adding grammar analysis and KBMT. The models used and the improvement of quality in those systems show that a single translation strategy, whether rule-based or based on statistical data, is only a partial solution, and integration of multiple translation strategies is the common feature of those systems.

C. MT strategy in CADAL

In light of the foregoing evaluation and current research in MT, we believe that the hybrid translation strategy is the most appropriate for MT in CADAL.

Firstly, we intend to collaborate with CMU by using their Mega2RADD system as the basic framework, and adopting the idea of RWTH Aachen University, which is to improve the source-channel based paradigm into the maximum entropy model. In this model, the parameters are estimated by large-scale samples. Among them, $P(E)$, the priori probability that E happens, can be estimated by constructing appropriate English linguistic model, while $P(F/E)$, the conditional probability of F given E can be estimated by the text-allied source text and the corresponding target text. To this end, in addition to the hardware and software (natural language parsing and synthesis), MT in CADAL involves a dictionary, grammar rules, and dynamic correlation between texts.

Secondly, under the framework of multiple engines, CADAL will take mtSDK (<http://lan.cpip.net.cn/>) as the

Field	Chinese Profile	English Profile
Title	孟德斯鸠法意·(一)	• of meaning of law of Meng Desi turtledove - one -
Author	著: 孟德斯鸠	Author: Meng Desi turtledove
Contributor	贡献者: 严复	Contributor: Yan Fu
Subject	主题: 法学 法的理论	Subject: Law The theory of the law
Description	描述: 孟德斯鸠法意[6册]	Description: Meaning of law of Meng Desi turtledove [6 books]
Publisher	出版社: 商务印书馆·上海	Publisher: Business affairs conforms book house · Shanghai
Create Date	创建日期: 1758-01-01 00:00:00	Create Date: 1758-01-01 00:00:00
Type	书籍类型: 民国图书	Type: Books of the Republic of China
Format	格式: Image/Djvu(.djvu)	Format: Image/Djvu(.djvu)
Source	来源:	Source:
Language	语言: Chinese	

Fig. 1. A book profile in both Chinese and English

standard to provide translation services at different levels. CADAL will use different engine for different tasks. For instance, translations of author, book title, sentences, paragraphs, abstract, and full text are carried out through different translation strategies, and CADAL will allow users to ask for the translation service and highlight the text that they would like translated. The translation engine of course must “understand” the individual words to translate in this way.

Thirdly, from automatic machine translation to human translation, there are human-assisted machine translations and machine-assisted human translations, to which CADAL will pay more attention. Under the framework of level translation, human intervention is allowed to improve the translation quality in CADAL.

IV. CONCLUSION

CADAL will make full use of the advantages of different MT systems to provide MT services to its users. The system will adopt multiple translation strategies, including rule-based, example-based and statistics-based strategies; manage various information used during the translation by employment of an object-oriented multiple type database; and provide a user interface which allows manual intervention to the resultant translation of MT. In order to obtain the linguistic resources required by KBMT, CADAL will also pay attention to the construction of its word library based on ontology, drawing on the research of Semantic Web.

As a digital library shared globally, CADAL will make use of state-of-the-art information technology to provide users of different levels with appropriate services, and let users study and work with digitized resources effectively. With the development of computer technology, we cherish the hope that MT can not only translate the textual information into the language with which the user is most familiar, but also achieve semantic information retrieval between different languages.

REFERENCES

- [1] Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman, *Machine Translation: A Knowledge-Based Approach*, San Mateo, CA: Morgan Kaufmann, 1992.
- [2] Nagao, M. (1984), “A framework of a mechanical translation between Japanese and English by analogy principle,” *Artificial and Human Intelligence*: edited review papers at the International NATO Symposium on Artificial and Human Intelligence sponsored by the Special Programme Panel. Lyon, France, October, 1981, Elsevier Science Publishers, Amsterdam, pp. 173-180.
- [3] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, *A Statistical Approach to Machine Translation*, Computational Linguistics, 1990.
- [4] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical

Machine Translation: Parameter Estimation,” *Computational Linguistics*, Vol. 19, No. 2, 1993.

[5] F. J. Och, C. Tillmann, and H. Ney. “Improved alignment models for statistical machine translation,” Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20-28, University of Maryland, College Park, MD, June 1999.

[6] Franz Josef Och, Hermann Ney, *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation*, ACL2002.

[7] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu, “A Method for Automatic Evaluation of Machine Translation,” IBM Research, RC22176 (W0109-022) September 17, 2001.

[8] LIU Qun, “Survey on Statistical Machine Translation,” *Journal of Chinese language information*, No. 4, 2003 (in Chinese)

