APPLICATION OF MACHINE TRANSLATION IN CHINA-AMERICA DIGITAL ACADEMIC LIBRARY

Huang Chen Chen Haiying
Zhejiang University Libraries，Hangzhou，310027, PR.China

Abstract: This paper briefly introduces the main ideas of Machine Translation (MT) techniques，then discusses the application of MT in the China-America Digital Academic Library (CADAL)

Key words: CADAL, machine translation, digital library

1. INTRODUCTION

The China-America Digital Academic Library (CADAL) Project was launched by China-US scientists (http://www.cadal.cn), aiming at digitizing one million books for the digital library. The project is also one of the key projects of the China Education Ministry for the “Tenth Five-year Plan”, intended to provide digitized resources for teaching and research, and to prompt the sharing of those resources.

The aim of the digital library is to provide information service. Nowadays users are no longer satisfied with the information retrieved though the internet; what they need is the integrated and processed information in different media. The service may even contain the knowledge and solution to the problem users have. CADAL therefore not only provides

digitized books, replacing traditional printed ones, but also process the digitized resources to extract relevant information, and provides more services to the users. Machine Translation (MT) is a service that CADAL intends to adopt to provide bilingual or multi-lingual translations.

2. MACHINE TRANSLATION THEORY AND TECHNOLOGY

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

Warren Weaver, director of Natural Sciences Department of Association of America Rockefeller Fund, published a memorandum entitled "Translation" to raise the issue of Machine Translation in 1949. With the development of both classical linguistic theory and modern computational linguistic theory, some commercial MT systems appeared later 80's, 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 80's and early 90's, 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 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 [3，4，5] can be seen as one variant of EBMT. The basic idea in statistical MT is that the translation is based on the statistical probabilities 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.

3. APPLICATIONS OF MT IN CADAL

3.1 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 on the rightof Figure 1.

Figure 1. A book profile in both Chinese and English

3.2 MT evaluation in CADAL

Based on the goal of CADAL, we evaluated a number of existing MT systems. These included 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 though 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.

3.3 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 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 standard to provide translation services at different levels. CADAL will use different engine for different tasks. For instance, the translations of the author, book title, sentences, paragraphs, abstract, full text are carried out through different translation strategies, CADAL allows 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.

4. CONCLUSIONS

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 Publishers, 1992.

2. Nagao, M. (1984), **A** framework of a mechanical translation between Japanese and English by analogy principle, *in* `Artificial and Human Intelligence: edited review papers at the International NATO Symposium on Artificial and Human Intelligence sponsored by the Special Programme Panel held in Lyon, France, October, 1981', Elsevier Science Publishers, Amsterdam, chapter 11, 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. In Proc. of the Joint SIGDAT Conf. 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)