Project Report

Machine Translation in China

Qun LIU
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
liuqun@ict.ac.cn

1. Introduction

As in many other countries, the way of machine translation in China is full of twists and turns. In this paper I will introduce the history of machine translation in China in three periods.

2. The first period

The earliest machine translation research in China was carried out from late 1950s to early 1960s. In December 1956, the Chinese government released its "1956-1967 Prospective Plan of Science and Technology Development"[1], in which "automatic translation" was listed as an important task in the item 41. In 1959, researchers in the Institute of Computing Technology (CAS-ICT) and the Institute of Linguistics (CAS-IL) of Chinese Academy of Sciences successfully conducted the first Russian-Chinese machine translation experiment in China. After that, many institutes continued machine translation research for many years.

3. The second period

- Research

The machine translation research was interrupted for about ten years because of the Culture Revolution. In mid-1970s, as part of the activities of a national project - the "748" Project, machine translation research was recovered in many institutes [3].

In early 1980s, personal computers were much power and cheaper and were ready to be used in machine translation research. Many institutes and universities were involved in machine translation researches from this period. Harbin Institute of Technology (HIT) and Northeast University (NEU) began their Chinese-English machine translation researches from mid-1980's. Nanjing University (NJU) started its Japanese-Chinese machine translation research in this period. CAS-ICT recovered its machine translation research from late l980's. The institute of Automata, Chinese Academy of Sciences (CAS-IA) began its research on spoken language translation research and joined the Consortium for Speech Translation Advanced Research (C-STAR) in late 1990's.

From 1987, many Chinese institute and universities were involved in the "Joint study for Multilingual Machine Translation" project, which was supported by Japanese government, and aimed to an ambitious goal of high quality machine translation between five Asian languages, i.e. Japanese, Chinese, Thai, Malay and Indonesia Language [5]. This project lasted about 10 years.

Three machine translations systems obtained the China's National Science and Technology Awards. MT-IR-EC, an English-Chinese title and catalog translation system for journal publishing, which was developed by the Research Institute of Post and Telecommunication Science, won the first National S&T

Award as a machine translation system for the first time. KY-1, an English-Chinese machine translation system, which is developed by the Academy of Military Sciences, won the second class National S&T Awards. The KY-1 system was also the predecessor of the first commercial machine translation system Transtar, which was released by Chinese National Software & Service Co. Ltd. (CS&S) in late 1980s. The Huajian Chinese-English machine translation system, developed by Huajian Co. Ltd, a spin-off from CAS-ICT, won the first class China's National Science and Technology Award in 1996 [2,3].

In this period, almost all the machine translation systems adopt rule-based approach, while some of them took example-based approach as a complement.

- Evaluations & Resources

From 1995 to 2005, the expert group of "863" plan held six open machine translation evaluation in 1994, 1995, 1998, 2003, 2004 and 2005. In the early “863” evaluations, Peking University proposed a check-point based approach for automatic machine translation evaluation [17].

Some public available language resources were also established and released, for example, the Grammatical Knowledge-base for Contemporary Chinese (Peking University, PKU), the People's Daily Corpus for Chinese word segmentation and segmentation (PKU), the Hownet (a lexical semantic knowledge database developed Prof. Zhendong Dong), etc. An open source tool ICTCLAS (CAS-ICT) for Chinese word segmentation and part-of-speech tagging was also released in 2003. These resources and tools greatly promoted the machine translation and other natural language processing researches and applications.

- Applications

CAS-ICT and GSL Co. Ltd. (Hongkong) released the first electronic dictionary which supported machine translation between Chinese and English -- the Instant-Dict EC863 in 1992, which made great commercial success in the market [3]. There were also many machine translation systems appeared in the software market, such as the Huajian System (Huajian Group), the Transtar system (CS&S), the GaoLi system (GaoLi Co. Ltd. & the Institute of Linguistics, Chinese Academy of Social Sciences), the CCID system (CCID Group), the Kingsoft Quick Translation system (Kingsoft Co. Ltd.). The first web-based machine tranlation system - "Read World" (Sunshine Co. Ltd) was released in 1999.

Some research groups, such as HIT group and NEU group, also developed some application systems collaboratively with domestic or foreign companies.

4. The third period

- Research

From 2004, some groups start their statistical machine translation (SMT) researches in China. These groups includes: the CAS-ICT group, the CAS-IA group and the Xiamen University (XMU) group. In 1995, these three groups held the first Symposium of Statistical Machine Translation (SSMT) in Xiamen, which became the precedent of the series of the annual China Workshop of Machine Translation (CWMT). From then on, the SMT researches grew very fast in China.

Many Chinese research groups (CAS-ICT, CAS-IA, HIT, NEU, Microsoft Research Asia, Toshiba China, etc.) published papers in the top conferences (ACL, EMNLP, COLING, etc.) and Journals (Computational Linguistics, Machine Translation) in MT area in recent years [4-16]. CAS-ICT proposed a series of new syntax-based translation model and approach, i.e. maximum-entropy bracketing transduction grammar model [5], tree-to-string model [6], forest-based approach [10,11], joint parsing and translation

approach [14], and dependency-to-string model [16]. These research works are referred or followed by many researchers in the machine translation community.

- Evaluations & Resources

From 2007, an open machine translation evaluation was held accompanying with almost every CWMT[1], where all the training data and test data were shared on ChineseLDC[2]. In recent years, many new language pairs, include some minority languages in China, were include in the CWMT MT evaluation, such Mongolian, Tibetan, Uygur, Kazakh, Kirgiz, etc. More than 20 research groups have ever participated in the CWMT evaluations, including some groups from foreign countries or foreign invested enterprises in China, such as Systran (USA), Microsoft Research Asia (US-China), I2R (Singapore), Sharp Co. Ltd. (Japan), NICT-ATR (Japan), Dublin City University (Ireland), Fujitsu R&D Center (Japan-China), NTT (Japan), etc.

In 2006, five research groups, including CAS-ICT, CAS-IA, XMU, Institute of Software of Chinese Academy of Sciences (CAS-IS), HIT, jointly released an open source phrase-based SMT system: Silk Road version 1.0 in the second SSMT. In 2011, NEU released a new open source machine translation system named NiuTrans.

Some Chinese MT groups gained good results in the international machine translation evaluations. In the NIST open MT Evaluations, CAS-ICT ranked No.5 in 2006 in Chinese-English track, and No.3 in 2009 in the progress test of system combination Chinese-English track. CAS-IA and CAS-ICT also ranked No.1 in the IWSLT spoken language translation evaluation for many times.

- Applications

Besides the traditional machine translation companies which adopt rule-based approach in their products, such as Huajian Group, CCID and CS&S, some companies or institutes began to adopt statistical machine translation products in real applications. Netease Co. Ltd. released its web-based Youdao Translation products in August 2008. Baidu Co. Ltd., the most big search engine company in China, also released its web-based machine translation system in 2011. CAS-ICT and the Eastlinden Co. Ltd. jointly developed a patent translation system and utilize it for large scale patent translation. In Beijing 2008 Olympic Games, machine translation technologies were also used to provide automatic or computer-aided translation services.

5. Conclusion

Machine translation has a history of more than 50 years in China. A number of institutes, universities and companies have been involved in machine translation researches and applications. In recent years, Chinese researchers achieved great progress in statistical machine translation research. We believe the machine translation researches and applications will have a better prospect in the future in China, along with the growing of the Chinese economy and the exchanges between China and the world.

Acknowledge:

Thanks to Dr. Yajuan Lü for her valuable suggestions and information.

[1] http://nlp.ict.ac.cn/new/CWMT/index.php
[2] http://www.chinese.org

References:

[1] 科学规划委员会，1956-1967 年科学技术发展远景规划纲要，1956 年 12 月

[2] 董振东，中国机器翻译的世纪回顾，中国计算机世界，2000 年第 1 期，2000 年 01 月 03 日

[3] 黄果，再建巴比塔——冯志伟研究员谈我国机器翻译发展历程，计算机世界报，1999 年 9 月 27 日

[4] Yajuan Lü, Ming Zhou, Sheng Li, Changning Huang, Tiejun Zhao, Automatic Translation Template Acquisition Based on Bilingual Structure Alignment, Computational Linguistics and Chinese Language Processing, Vol. 6, No. 1, February 2001, pp. 83-108

[5] Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In Proceedings of COLING/ACL 2006, pages 521-528, Sydney, Australia, July.

[6] Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In Proceedings of COLING/ACL 2006, pages 609-616, Sydney, Australia, July.

[7] Hua Wu and Haifeng Wang, Pivot language approach for phrase-based statistical machine translation, Machine Translation, Volume 21, Number 3, 165-181, 2007

[8] Dongdong Zhang, Mu Li, Chi-Ho Li, Ming Zhou, Phrase Reordering Model Integrating Syntactic Knowledge for SMT, in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning, pp. 533–540, Prague, June 2007.

[9] Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou, Minghui Li, Yi Guan, A Probabilistic Approach to Syntax-based Reorderingfor Statistical Machine Translation, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 720–727,Prague, Czech Republic, June 2007

[10] Haitao Mi and Liang Huang. 2008. Forest-based Translation Rule Extraction. In Proceedings of EMNLP 2008, pages 206-214, Honolulu, Hawaii, October.

[11] Haitao Mi, Liang Huang and Qun Liu. 2008. Forest-Based Translation. In Proceedings of ACL 2008, pages 192-199, Columbus, Ohio, USA, June.

[12] Jiajun Zhang, Chengqing Zong, Shoushan Li, in: Proceedings Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), Manchester, UK, August 2008.

[13] Hua Wu, Haifeng Wang and Chengqing Zong, Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora, in Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, August 2008.

[14] Yang Liu, and Qun Liu. 2010. Joint Parsing and Translation. In Proceedings of COLING 2010, Beijing, China, August.

[15] Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative Word Alignment by Linear Modeling. Computational Linguistics, 36(3).

[16] Jun Xie, Haitao Mi, Qun Liu, A novel dependency-to-string model for statistical machine translation, in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 216--226, Edinburgh, Scotland, UK. July 2011

[17] Yu Shiwen, Automatic evaluation of output quality for Machine Translation systems, Machine Translation, Volume 8, Numbers 1-2, pages 117-126, 1993, DOI: 10.1007/BF00981248