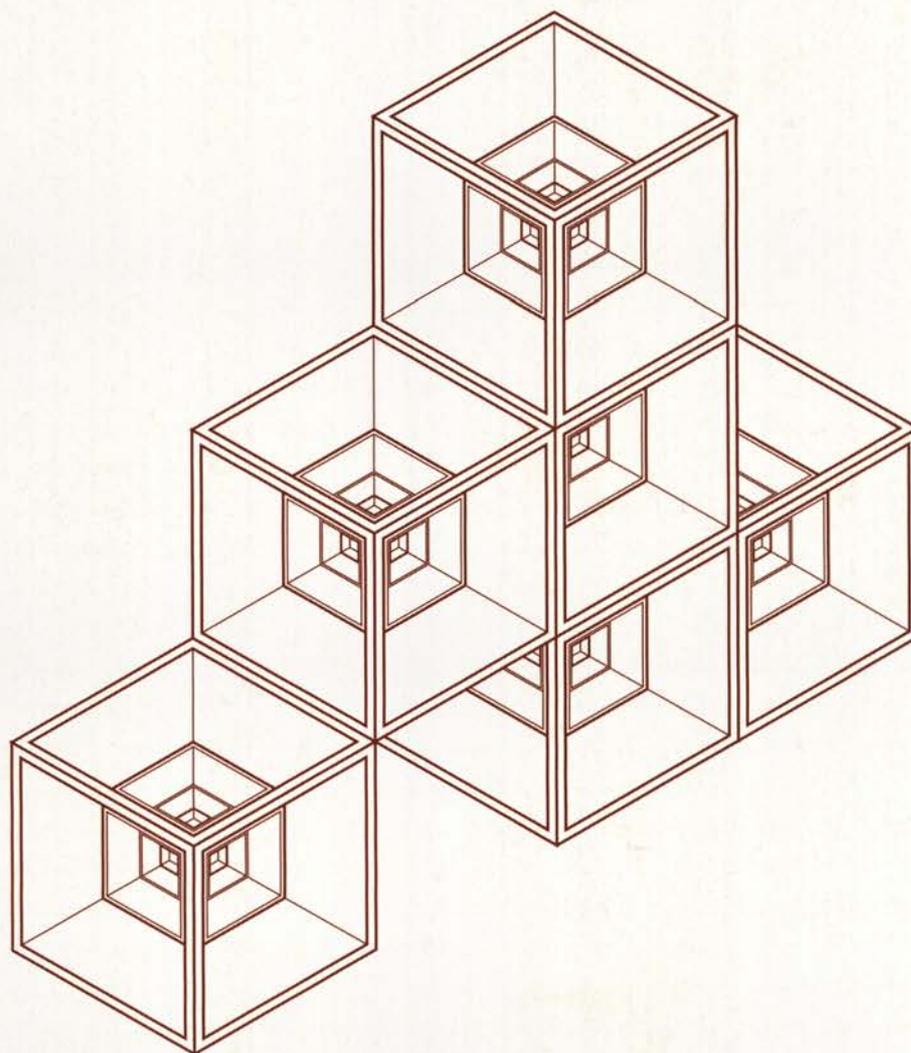


# AAMT

Asia-Pacific Association for Machine Translation

# Journal



October 2008

**No.43**

アジア太平洋機械翻訳協会

# 目 次

巻頭言：	MT 研究の土俵.....飯田 仁..... 1
海外事情：	Progress in Machine Translation in Mainland of China ..... Bo XU, Chengqing ZONG ... 2
	Towards a High-quality Korean-English MT..... .....Mupyo Hong, Chang-Hyun Kim, Young-Gil Kim .... 8
プロジェクト：	MASTER プロジェクト ..... 中村 哲 ..... 13
	北京オリンピック音声翻訳モニター実証実験 ..... 清水 徹、中村 哲 ..... 16
レポート：	COLING2008 における機械翻訳研究の傾向 ..... 清水伸幸、宮尾祐介、河原大輔 ..... 18
	マレーシア会議報告 ..... 村上 嘉陽 ..... 20
	TC シンポジウム報告 ..... 村上 嘉陽 ..... 23
新サービス：	「超多言語」のテキスト翻訳システムを携帯電話で無償サービス開始 ..... ..... 隅田 英一郎 ..... 28
新製品紹介：	The 翻訳® 2008 プレミアム..... 東芝ソリューション(株) .... 31
	ATLAS V14..... 富士通(株)..... 34
委員会報告：	AAMT 機械翻訳システム評価テストセットの概要 ..... AAMT 機械翻訳課題調査委員会 ... 37
会員紹介：	AAMT 会員のひろば (第 3 回) ..... 43
総会報告：	第 18 回通常総会および関連行事の報告 ..... 50
活動報告：	協会活動報告 (2008 年 6 月～9 月) ..... 52
編集後記：	..... 宇津呂 武仁 ..... 54

# CONTENTS

Foreword:	MT Research Sphere of Activity ..... <i>H. Iida</i> ..... 1
Report:	Progress in Machine Translation in Mainland of China ..... <i>Bo XU, Chengqing ZONG</i> ..... 2
	Towards a High-quality Korean-English MT..... ..... <i>Mupyo Hong, Chang-Hyun Kim, Young-Gil Kim</i> ..... 8
Project:	MASTER Project -New Speech and Natural Language R&D Center at NICT ..... ..... <i>S. Nakamura</i> ..... 13
	Speech Translation Field Experiment Systems for Beijing Olympic Game ..... ..... <i>T. Shimizu, S. Nakamura</i> ..... 16
Report:	Research Trend of Machine Translation in COLING2008 .... ..... <i>N. Shimizu, Y. Miyao, D. Kawahara</i> ..... 18
	Malaysia Report ..... <i>Y. Murakami</i> ..... 20
	TC Symposium ..... <i>Y. Murakami</i> ..... 23
New Services:	A service was inaugurated to automatically translate Japanese or English sentences via a mobile phone into 17 languages free of charge..... <i>E. Sumita</i> ..... 28
New Product:	“The Honyaku 2008 Premium” ..... <i>Toshiba Solutions Corp.</i> ..... 31
	“ATLAS V14” ..... <i>Fujitsu Co. Ltd.</i> ..... 34
Committee Report:	AAMT Test-Set: a Method of Quality Evaluation of Machine Translation Sentences ..... ..... <i>MT Committee. for seeking future direction of MT</i> ..... 37
AAMT Members:	AAMT Agora ..... 43
General Meeting:	Report on the 18 <sup>th</sup> AAMT General Meeting and related events ..... 50
AAMT Activities	AAMT Activities (June 2008～September 2008)..... 52
Editor's Note:	Message from the Chair of the Editorial Committee..... <i>T. Utsuro.</i> ..... 54

## MT 研究の土俵

東京工科大学メディア学部  
飯田 仁

2008 年の夏是北京オリンピックで沸いた。スタジアムの外でも善し悪しはともかく賑やかで、話題に事欠かなかった。MT 研究についても、現地になかったものの、話題豊富であったかと思う。というのも、「タクシーに多言語翻訳システムを導入、北京五輪の準備着々」、「携帯型端末による多言語音声翻訳システムの北京五輪での活用を中国に提案」、「世界初 携帯電話を使った動翻訳サービスがスタート」などの報道記事が 07 年の話題を飾った。そして、北京五輪直前の 7 月に「JTB ツアーが携帯翻訳サービス 北京五輪で実証実験」という報道発表があった。思い起こすに、2002 年の日韓共催サッカー・ワールドカップの開催に向けて翻訳サービスの実用試験計画などの提案が求められたと記憶する。そのときは MT の実証実験までには至らなかった。それに対し、近いうちに目にするであろう北京五輪での実験結果に期待する。

MT に関わる活動を探るとき、AAMT のサイトを参照することが、まず第一の行動かと思う。このサイトの日本地域情報からトピック一覧の項を開くと、2002 年から今年 7 月までの約 6 年半の進展を一目で辿ることができる。多くが報道メディアに取り上げられた話題であり、ニュースのヘッドライン程度の記述である（溯って内容を読むには過去の新聞記事などを探る）。とは言え、トピック一覧を 02 年からひとわり眺めてみると、基本的なところでは似た話題の繰り返しが意外と目につく。報道対象が一定の水準までに達していて、翻訳対象の拡張や機能の拡張がトピックになり、また新規計画の採択などが話題になっている。かつて、共立出版社から 1969 年に発刊され、01 年に 499 号で休刊となってしまった雑誌「bit」を 70 年頃から 20 年間表紙にその号のトピックが明示されているから、

12 冊単位で眺めていくとエポックが見えてくる。京都賞を受賞したクヌースの仕事などは顕著に浮かび上がっていた。それに比べ、当の AAMT トピックス一覧から、ダイナミックスさを感じるには無理がある。その違いは、研究目標や実用化目標の違いに起因するのではないだろうか。

MT の研究目標（例えば、言語のバリアフリー実現）を追究することは人間の能力相当（問題解決、思考、状況判断などの能力）をそのままに実現することになり、マイルストーンの設定が難しいであろう。HI や AI においても同様の事情があると考えられる。音声認識研究に置き換えてみると、つぎの変遷を経て、確率的統計手法が中核技術となり、しかしユーザ誰もがつぎに再度使ってみたいと思う技術には至っていない。その間、有限状態ネットワークに依存した認識手法、スペクトログラム・リーディング手法、そして 90 年代に入り、HMM による音素認識と bi-gram による言語モデルとの融合手法。翻って、MT 研究では、現在統計的機械翻訳が主流であり、対訳コーパスがある程度整えば、新たな言語対の MT システムが短期間で作成できる。つまり、この翻訳手法は普遍的な翻訳技術であり、世界共通の土俵で議論が戦わされる。

しかし、広く利用者はローカライゼーションのための作業を要し、TM の充実を図ることになる。このことは、音声認識であれ、MT であれ、確率的統計手法は中核技術であっても、その利用の仕方、あるいは使い勝手が整ってこそ、価値があるといえる。とくに前者については、文脈を含む状況までも掘り上げる統計情報とするための適材サービス対象を明確にする方法論が求められる。「コミュニケーションデザイン」と相俟って MT 次世代が拓かれる夢は悪くない。

## **Progress in Machine Translation in Mainland of China**

Chengqing ZONG

Bo XU

Institute of Automation, Chinese Academy of Sciences (CASIA)

cqong@nlpr.ia.ac.cn

xubo@hitic.ia.ac.cn

---

**Abstract** This paper summarizes the progress in research and development (R&D) of machine translation (MT) in mainland of China over the past decades. The research activities on MT in Institute of Automation of Chinese Academy of Sciences (CASIA) are also introduced. The authors kindly hope to draw more attention and support to R&D of MT technology in China.

**Keywords:** Machine translation, Progress in machine translation, Spoken language translation

### **1. Introduction**

As we know that machine translation (MT) is an earliest and persistent research topic in the area of natural language processing (NLP). NLP research was de facto started from MT. Almost at the same time when MT was studied in the world in 1950s, investigation on MT was did in mainland of China. During the past decades, the situation of MT research has been significantly changed, both in theory and application.

In the recent years, with the fast development of networking and the quick increasing of international intercommunion, the market of MT becomes larger and larger, which attracts more and more enterprisers to dive into the competition of MT technology. Especially, with the rapid development of Chinese economy, the Chinese language draws more and more attention. The language pair of Chinese and one of other languages has become the first choice for machine translation research. In the most important evaluations of MT in the recent years, including NIST Open MT evaluation<sup>1</sup> and IWSLT (International Workshop on Spoken Language Translation), the number of participants to Chinese-to-English MT evaluation is always the biggest. Meanwhile, the quick progress of statistical machine translation strongly promotes the R&D of MT, and the Chinese related MT certainly benefits from the period of quick changing.

After reviewing the history of R&D of MT in China, this paper presents some current work on MT R&D in China. The work on MT in Institute of Automation of Chinese Academy of Sciences (CASIA), is specially introduced. Finally, the concluding remarks are given.

### **2. History of MT in China**

Almost at the same time when Georgetown University studied the first experimental MT system in 1950s, MT investigation was supported by the Chinese government and the project was listed in the planning program of science development in 1956 named as *Machine Translation, Development of Natural Language Translation Rules, and Research on Mathematic Theory of NLP*. Under support of the project, Russian-Chinese MT was studied in 1957 by the Institute of Computing Technology (ICT) in cooperation with the Institute of Linguistics (IL), Chinese Academy of Sciences (CAS). In 1959, the first experiment of Russian-Chinese translation was done with computer system of type 104 made in China. In the experiment, nine different types of complex sentences were translated. In that period, some research teams of MT were built in Beijing Language College, Beijing Russian Language College, South China Institute of Technology, and Harbin Institute of Technology as well,

---

<sup>1</sup> <http://www.nist.gov/speech/tests/mt/>

and Russian-Chinese and English-Chinese translation have been studied and experimented since that time (Feng, 2004).

During the ten years from 1966 to 1975, research on MT experienced a drowsy period in China for some special reasons like in other countries because of the report of ALPAC given by the American Academy of Sciences. However, it has been revived since 1975. A new research team of MT was established in the Institute of Science and Technology Information of China (ISTIC) and a translation system was studied and experimented based on the corpora of metallurgy area.

MT has come into a flourishing period in China since 1980s. The system of *Science Translator NO. 1* became a landmark in the early of 1980s, which was developed by the Military Science Academy of China oriented to the translation of full text and title from English to Chinese. Professor Zhendong DONG proposed the concept of logical semantic structure at that time. Meanwhile, *Gao-Li* English-to-Chinese MT system was developed by Gao-Li Computer Company in cooperation with the Institute of Linguistics, Chinese Academy of Social Sciences.

Professor Qingshi GAO, academician of CAS, and his research team in ICT, CAS, have been engaged with research of MT since 1980 and experienced several stages from the theory research, system experiment, and development of practical system. The intelligent English-to-Chinese MT system (IMT/EC-863) developed by his research team was appraised in 1992 (Chen *et al.*, 1992). Afterward, the first pocket English-to-Chinese translation machine in the world was successfully developed by Dr. Zhaoxiong CHEN and his colleagues. The usufruct of the technology was transferred for US\$740,000, which was the biggest charge in the area of software transferring in China at that time. IMT/EC-863 system gained the first-class National Award of Science and Technology Progress of China in 1995. Based on the MT technology, Dr. CHEN and his colleagues founded the Huajian Group Co., Ltd<sup>2</sup>, and established the Research Center of Computer & Language Information Engineering, CAS. Huajian Group developed quickly and has become a big corporation which not only develops the products of MT, but also successfully takes a big market of multi-lingual translation service. Now Huajian is mainly engaged in technological research, product development, application integration and technical services in the field of computer and language information processing, to provide the government, businesses and individuals with solutions to computer information system applications such as computer information processing, systems integration and information services.

During 1980s, multi-lingual translation had been investigated for eight years by the five Asian countries including Japan, China, Indonesia, Thailand, and Malaysia under ODA program of European Economic Community (EEC), which was presided by the Center of the International Cooperation for Computerization (CICC) of Japan and 6000 millions Japanese yens were supported by Japanese government. Some Chinese universities, companies and institutes, including China National Software & Service Company Limited (CS&S), Information R&D Center of Ministry of Mechanics and Electronics of China, Beijing Language College, ISTIC, Northeast University, Tsinghua University, Nanjing University, and Renmin University of China and so on, joined in this international project (Yao *et al.*, 2002).

After the end of 1980s, especially after coming into 1990s, some new MT companies have been established, such as Kingsoft<sup>3</sup>, and Global View<sup>4</sup> etc. The research and development of MT has come into a flourishing period in China.

The speech-to-speech (S2S) translation, sometime saying spoken language translation (SLT), has been studied in mainland of China since 1990s started. The first experimental speech-to-speech translation system was developed by Sichuan University at the beginning of 1990s (Yang *et al.*, 1992). CASIA has studied the S2S translation since the middle of 1990s and became a partner of C-STAR (Consortium for Speech Translation Advanced Research International<sup>5</sup>) in 2000. Afterward, CASIA has been engaged in SLT research more than decade, joined so many important international activities on SLT, and made much achievement.

So many scientists, linguists and engineers devoted their life into R&D of MT in mainland of China and made much contribution to progress of MT technology, such as Yongquan LIU, Zhuo LIU, Qingshi GAO, Zhendong DONG, Zhiwei

---

<sup>2</sup> <http://www.hjtek.com/en/index.html>

<sup>3</sup> <http://www.kingsoftresearch.com/>

<sup>4</sup> <http://www.gvlinux.com/>

<sup>5</sup> <http://www.c-star.org/>

FENG, Tianshun YAO, Qi YUAN, Huilin WANG etc.

### 3. Current MT in China

At present, there are so many universities, institutes and companies conducting with MT R&D in China, including the following ones:

- Harbin Institute of Technology (HIT)
- Institute of Automation, CAS
- Institute of Computing Technology, CAS
- Institute of Software, CAS
- Inner-Mongolia University
- ISTIC, CAS
- Nanjing University
- Northeast University
- Research Center of Computer & Language Information Engineering, CAS
- Sinkiang University
- Tsinghua University
- Xiamen University
- Huajian Group Co., Ltd.<sup>6</sup>
- China National Software & Service Company Limited

In the recent years, the Chinese government funds more and more for R&D of MT technology. Almost all national programs support this work, including the national key fundamental research program (“973” Program), the national natural science foundation of China (NSFC), the national hi-tech research and development program (“863” Program) of China, and the national key technology R&D program as well.

In 2001, the Chinese government gained the qualification to hold 2008 Olympic Games in Beijing. The Hi-Tech R&D Program of China (“863” Program) funded a big project named as “Research and Development of Multi-lingual Intelligent Information Service System and Application Demonstration System” (Grant No. 2002AA117010 ) in 2002, whose purpose was to study and develop the human language technology oriented to 2008 Olympic Games to provide the technical support to reach the goal of “*Digital Olympic, Personalized Olympic*”. The main tasks included (1) development of information service system on networking and kiosk; (2) development of computer-aided multi-lingual translation system (off-line); (3) development of real-time multi-lingual translation system (on site / on line); and (4) development of multi-lingual service system for mobile terminals. The framework is shown in the following Figure 1.

---

<sup>6</sup> <http://www.hjtek.com/en/index.html>

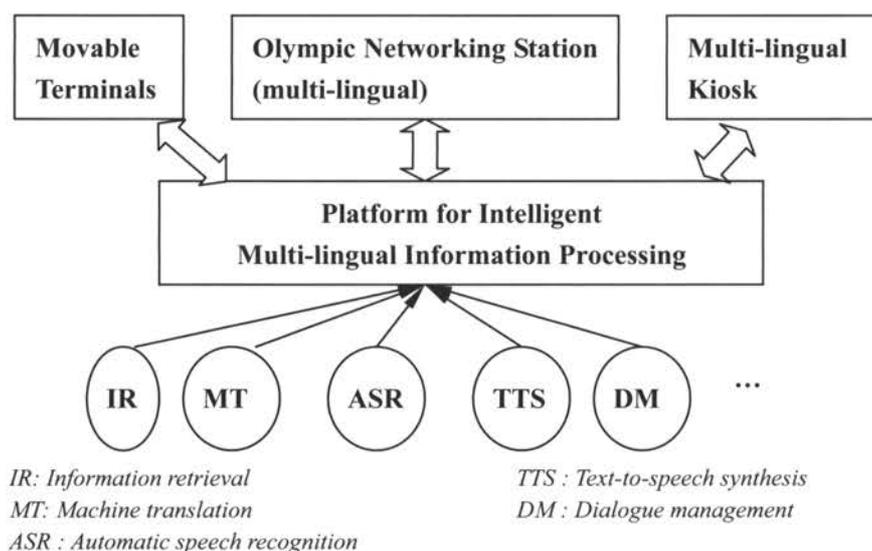


Fig. 1. Frame of Olympic Project

CapInfo Co. Ltd., CASIA, Huajian Group, Beijing University of Post and Telecommunication, and some other universities and companies joined the project. During Beijing Olympic Games, computer-assisted MT technology was successfully applied for multi-lingual information service centre, kiosk service, call-centre and post office to help human interpreters to translate the various types of documents, athletes' resumes, records and all background information, schedules of activities, and the dialogues between clerks and customers in the post offices.

In the recent years, the statistical machine translation (SMT) has been more addressed. Especially, CASIA and ICT of CAS, HIT, and Xiamen University have much achievement on SMT research. The novel translation model based on tree-to-string alignment template (TAT) was proposed by Dr. Yang LIU and his colleague (Liu *et al.*, 2006; Liu, 2007). (Liu *et al.*, 2006) received the Best Asian Language Paper Award conferred by the Asian Federation for Natural Language Processing (AFNLP) in ACL-COLING joint conference, Sydney. The spoken Chinese-to-English MT system developed by CASIA participated in IWSLT evaluations. According to the subjective evaluation results, CASIA system was the best one in IWSLT, 2007 and 2008 (Fordyce, 2007; He *et al.*, 2008).

#### 4. Study on MT in CASIA

The CASIA, as a partner of C-STAR, has studied speech-to-speech translation since the middle of 1990s, and joined in many important activities, including the international workshop on spoken language translation (IWSLT) (Zong *et al.*, 2005). Some landmarks are given as follows:

- 1996, CASIA joined C-STAR as affiliate member
- 2000, CASIA became a C-STAR partner
- 2000, Development of prototype of Chinese-Japanese-English SLT pocket machine in cooperation with Panasonic
- 2002, Development and demonstration of experimental system of Chinese-Korean SLT in common used mobile phone
- 2003, Establishment of Chinese Language Data Consortium (Chinese LDC)<sup>7</sup> in cooperation with other universities and institutes
- 2004, Demonstration of PDA-based Chinese-English bi-directional S2S translation system separately in Beijing International Exposition of Sciences and Technology (May 22-26, 2004) and Spanish International Culture Forum

<sup>7</sup> www.chineseldc.org

(July 16-18, 2004, Barcelona).

- 2007, According to the subjective evaluation results of IWSLT'2007, CASIA Chinese-to-English SLT system was the best one
- 2008, According to the subjective evaluation results of IWSLT'2008, CASIA Chinese-to-English SLT system is the best one (Please refer to Fig. 2 for the system architecture)

Since 2005, CASIA has extended her research interests. Besides SLT, text-to-text MT is also focused. In NIST Open MT Evaluation<sup>8</sup> in 2008, the CASIA Chinese- to-English MT system was the best one among the Chinese participants.

Now CASIA is mainly engaged in SMT methods and has proposed many novel approaches, including the new approached to named entity translation (Chen *et al.*, 2008), extraction of translation template (Hu *et al.*, 2006), word alignment (He *et al.*, 2008c), reordering models (Zhang *et al.*, 2008; He *et al.*, 2008b), and system combination (Li *et al.*, 2008) as well.

## 5. Conclusion

In this paper we review the history of MT R&D and introduce some achievements in the past five decades of China. The following two points may be summarized: (1) Much progress in MT research has been made, but the original models, algorithms, and grammars, especially the effective models and methods regarding Chinese understanding and translation, are wishfully expected; (2) The Chinese government has established so many foundations to support R&D of MT technology, and there is a very large market of MT in China. The high quality MT systems are imminently necessitated.

## References

- Chen, Yufeng, and Chengqing Zong. 2008. A Structure-based Model for Chinese Organization Name Translation. *ACM Transactions on Asian Language Information Processing*, 7(1): 1-30
- Chen, Zhaoxiong, and Qingshi Gao. 1992. A New Context-Sensitive Subcategory (SC) Grammar for Machine Translation (in Chinese). *Journal of Computer*, (11): 801 -808
- Feng, Zhiwei. 2004. *Research on Machine Translation* (in Chinese). China Translation and Publishing Corporation.
- Fordyce, Cameron Shaw. 2007. Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*. Trento, Italy. Oct. 15-16, 2007. Pages 1-12
- He, Yanqing, Jiajun Zhang, Maoxi Li, Licheng Fang, Yufeng Chen, Yu Zhou and Chengqing Zong. 2008a. The CASIA Statistical Machine Translation System for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hawaii, USA. October 20-21, 2008
- He, Yanqing, and Chengqing Zong. 2008b. A Generalized Reordering Model for Phrase-Based Statistical Machine Translation. In *Proceedings of the 8<sup>th</sup> Conference of the Association for Machine Translation in the*

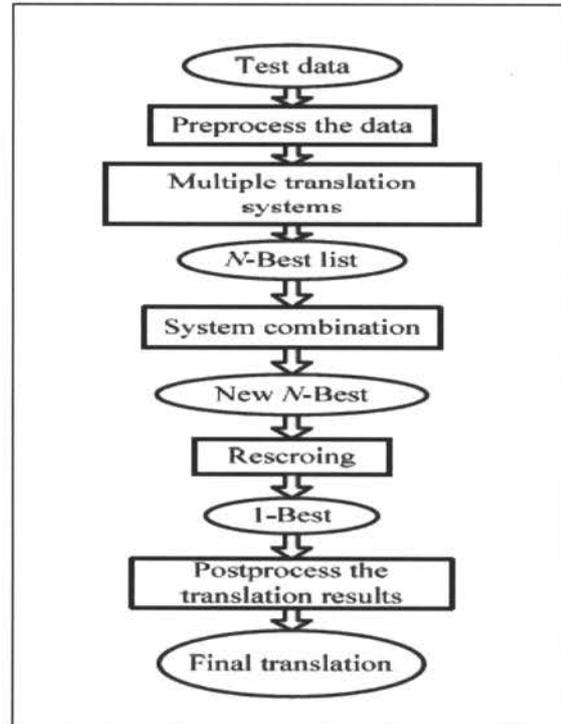


Fig. 2. Architecture of CASIA SLT system

<sup>8</sup> [http://www.nist.gov/speech/tests/mt/2005/doc/mt05eval\\_official\\_results\\_release\\_20050801\\_v3.html](http://www.nist.gov/speech/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html)

- Americas (AMTA)*, October 21-25, 2008. Waikiki, Hawai'i, USA
- He, Yanqing, Yu Zhou, and Chengqing Zong. 2008c. Word Alignment Based on Multi-Grain. To appear in *the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, December 16-19, 2008. Kunming, China
- Hu, Rile, Chengqing Zong, and Bo Xu. 2006. An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment. *IEEE Transactions on Audio, Speech, and Language Processing*. Vol. 14, No.5, September 2006. Pages 1656-1663
- Maoxi Li, and Chengqing Zong. 2008. Word Reordering Alignment for Combination of Statistical Machine Translation Systems. To appear in *the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, December 16-19, 2008. Kunming, China
- Liu, Yang. 2007. *Research on Tree-to-String Statistical Translation Models* (in Chinese). Ph.D. Thesis. Institute of Computing Technology, CAS.
- Liu, Yang, Liu Qun, Lin Shouxun. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proceedings of the 21<sup>st</sup> International Conference on Computational Linguistics and the 44<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, July 2006. Pages 609-616
- Yang, Jiayuan, Daofa Lin, Wanbo Luo *et al.* Design and Implementation of Spontaneous English-to-Chinese Speech-to-Speech Translation System (in Chinese). *Journal of Acoustics*, 1992, 17(5): 327 ~ 333
- Yao, Tianshun, Jingbo Zhu *et al.* 2002. *Natural Language Understanding- A study to make machine understand human language* (in Chinese). Tsinghua University Press.
- Jiajun Zhang, Chengqing Zong, and Shoushan Li. 2008. Sentence Type Based Reordering Model for Statistical Machine Translation. In *Proceedings of Conference on Computational Linguistics (COLING)*, August 18-22, 2008. Manchester, UK. Pages 1089-1096
- Zong, Chengqing, Mark Seligman. 2005. Toward Practical Spoken Language Translation. *Machine Translation*, 19(2): 113-137

## Towards a High-quality Korean-English MT

Munpyo Hong  
Sungkyunkwan University, Korea  
[skkhmp@skku.edu](mailto:skkhmp@skku.edu)

Chang-Hyun Kim & Young-Gil Kim  
ETRI, Korea  
{[chkim](mailto:chkim@etri.re.kr), [kimyk](mailto:kimykim@etri.re.kr)}@etri.re.kr

### 1. Introduction

Since early 2000, MT has been paid much attention in Korea due to the fast growth of the Internet. However, most of the efforts to bring out the off-the-shelf products into the market have failed mainly because of the relatively poor quality of the MT. One of the few MT systems actively used not only in Korea but also in foreign countries is the Korean-English Patent MT system.<sup>1</sup> The performance of the patent MT system is reported to be good enough for foreign patent examiners to retrieve and read the Korean patent documents in which they are interested in English. (cf. Hong et al. (2005))

The quality of the Korean-English patent MT might be just good enough for the cross-lingual information retrieval, however not good enough, if it is to be used for other purposes like English scientific paper authoring. If an MT system is to be employed for academic paper writing in English, the quality of the translation must be far superior to that of the patent MT system. However, we are well aware of the difficulties and obstacles in improving the quality of Korean-English MT in a short period of time.

This paper addresses two issues in developing the Korean-English scientific paper MT system based on the Korean-English patent MT system. Given the assumption that the authors will inevitably have to post-edit the MT results in order to submit an English paper to a conference or an academic journal, the reduction of the time and the efforts for the post-editing is very important for the success of the system.

One way to reduce the cost for the post-editing is to employ a controlled language for MT. One might argue that using a controlled language for MT is just to shift the load from the post-editing to pre-editing. However, empirical studies on controlled language show that the load for the pre-editing based on a controlled language is generally less than that for the post-editing where a controlled language is not employed. (cf. Lehrndorfer & Mangold(1997), Lehrndorfer & Schachtl(1998))

The other way is to allow the users to interact with the MT system to get the better translation result. By allowing the users to have an access to each of the intermediate results of the translation, i.e. morphological-, syntactic analysis, and target word selection, the maximum use of the MT system can be made. However, a question arises in this respect whether a user might want to intervene in the system at all.

In section 2, we will introduce the concept of the Controlled Korean. We will share our experience in designing the Controlled Korean for Korean-English MT. In Section 3, the interactive MT system and the relevant issues will be discussed. In section 4, we will report about the beta service of the scientific paper MT system. In section 5 we will discuss about the future works.

### 2. Controlled Korean

Controlled Language is a sublanguage of a natural language designed to improve the readability and the translatability of a text. Originally, the concept of a controlled language was introduced in the field of technical documentations to prevent the misunderstanding of texts. In the last couple of years the necessity of a controlled language has increased significantly, especially in the technical documentation domain, as the number of pages to be translated has increased enormously.

To design a controlled language for MT, the purpose of the application must be properly understood. Most of the controlled languages introduced so far have without exceptions limited lexicons and writing rules. (cf. AECMA(1995)) This is possible and useful, when the controlled language and the MT system are used by a homogeneous group of users like employees in a company. However, in the academic paper authoring setting, it is almost impossible to enforce the distributed users to use only the allowed lexicon and not to use any words they like. We therefore gave up introducing the restricted Korean lexicon and rather focused on the writing rules which impact on the translation quality most.

To do this, we manually checked 40,000 Korean-English machine translation pairs. We scored all the machine translated results from 0(poor) to 4(very good). The translations above 3 points(good) were excluded. We assumed that the syntactic analysis and the transfer of those sentences succeeded without fatal errors. Then we focused on the translations that were scored below 3. Many of them failed in the morphological or syntactic

<sup>1</sup> Korean-Japanese MT system is the most widely used MT system in Korea

analysis for some reasons. We were only interested in those cases where, though the morphological and the syntactic analysis succeeded, the translation was awkward. In such cases generally the so-called 'Konglish (English in Korean style)' was produced. We assumed that the reason why such 'Konglish' was generated was that the source Korean sentence was stylistically poor.

We found out 8 most frequent writing errors that Korean authors commit and that have the most significant effects on the quality of MT:

- **Subject-Predicate Mismatch:** Semantic mismatches between the subject and the predicates in a sentence
- **Topic Markers in the Sentence Initial Position:** Sentence initial NPs with topic markers are underspecified w.r.t. their case
- **Ambiguous Functional expressions:** Use of specific ambiguous function words
- **Spoken language expressions:** Use of spoken language type expressions
- **Double subject/object construction:** Though these constructions are legitimate in Korean grammar, they are difficult to analyze correctly
- **Punctuation:** If a sentence is long, the correct punctuation helps the analysis
- **Minor grammatical errors:** Though the grammatical errors are not supposed to be treated by a controlled language checker, some grammatical errors are so frequent and not recognized by the authors that they have to be checked before the syntactic analysis
- **Light-verb expressions:** Frequent use of light-verb constructions that can be substituted by a single verb

The philosophy in designing the Controlled Korean was to give the authors as much freedom as possible. The authors write sentences that will be automatically checked against Controlled Korean writing rules by a Controlled Korean Checker: cf. Fig.1



Fig. 1: Controlled Korean Checker

The learnability of the controlled language is a very important issue in designing a controlled language. Lehrndorfer(1996), for example, criticized the AECMA Simplified English for its poor learnability. Especially, in the academic paper authoring setting, it will be very difficult to teach the users the Controlled Korean. Therefore, the application of the above rules must be automated with a Controlled Korean Checker. However, the formalization of above rules is not always simple. For example, the semantic mismatch of the subject and the predicates in a sentence is very difficult to detect automatically. It would be very difficult without deep semantic processing. However, the deep semantic processing technique is currently not available. Therefore we collected lexical clues with which we can detect the semantic mismatches on the surface level. Currently about 6,000 manually constructed lexical rules and metarules are stored in the Controlled Korean Checker.

Our experiment with 200 test sentences revealed that by employing the Controlled Korean, the translation accuracy could be improved for 4.25%. Considering that only the style errors that were automatically detected by the Controlled Korean Checker were corrected, the translation quality is expected to improve significantly in the future, as we build more lexical rules for the checker. Among the 8 writing error types, "the ambiguous functional expression", "subject-predicate mismatch" and "double subject/object" errors were 3 most significant error types that effect on the translation quality most.

### 3. Interactive MT

The current Korean-English academic paper MT system is customized from the Korean-English Patent MT system. The initial academic paper MT system was equipped only with a Controlled Korean Checker. The MT system has been open to beta testers for six months and the feedbacks from the testers have been collected. Overall the beta testers were satisfied with the MT output, however, many of them wanted to know what caused the mistranslations and how could they fix them. Besides, the MT outputs still contained many errors, even though the source Korean sentences were rewritten guided by the Controlled Korean rules.

The idea of interactive MT is not new in the MT research history. For example, in Foster et al (1997) and Langlais et al (2000), the target sentence under construction serves as the medium of communication between an MT system and its user. In such an environment, human translators interact with a translation system that functions as an assistance tool and dynamically provides a list of translation candidates. To extend a type of translation models, a hybrid approach was suggested. (cf. Yamabana(1997))

We revised the initial paper MT system reflecting the most frequent user feedbacks: the maximization of user's engine control, the control level management by the users, the offering of sufficient information for error corrections, and the user-friendly interface. For the advanced paper MT users, we allowed them to control the intermediate steps of the translation. They can intervene in the morphological- as well as syntactic analysis. If they find something is wrong in the analysis, they can fix it by themselves in a user-friendly interface. Fig.2, for example, shows the window displaying the morphological analysis result. If a user doesn't like the translation, he/she can backtrack and revise the analysis error, if there is one. If the error lies in the POS tagging, the user can fix it and retry the translation as shown in Fig.2.

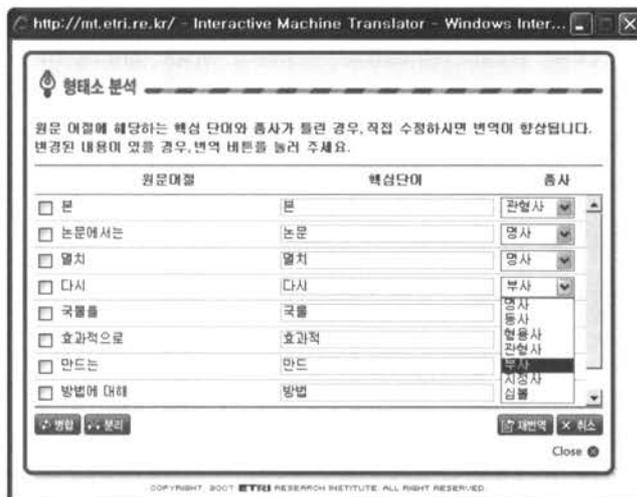


Fig. 2: Interactive Morphological Analysis Window

They can even access to the target word selection step. The correct target word selection is based on the correct word sense disambiguation (WSD). However, the correct WSD is not always possible. The user can select the appropriate target word by himself, if he doesn't like the one suggested by the system as in Fig.3.



Fig. 3: Interactive Target Word Selection Window

By allowing the advanced users to have the access to the intermediate steps, they can make full use of the MT system and thus have the best possible translation result that the MT system can offer.

While some users might want to have the full control of the translation steps, some others might not need it. They might just want to have their source text get translated simply by clicking on the button or after just examining the POS tagging result. Thus the users are given the right to choose the control level they want. If a user is relatively poor in English, he/she may depend more on rewriting of the Korean sentence or the correction of Korean analysis error. If the user is relatively good at English, he/she can intervene in the target word selection step.

#### 4. The Analysis of the Beta Service

We have run the paper MT server for the employees of ETRI for 6 months. The total number of the users who have visited the site at least once was 694. Considering that the total number of the researchers at ETRI is around 1,000, the needs for the paper MT system seem to be high. 694 users have translated 39,506 sentences with the MT system. The average number of the translated sentences per user is thus about 56. Among those 694 users there must be the users who just tried the system for fun or just for trial. We were interested in finding out which function of the system do the users find most useful. We thus narrowed our attention only to those users who have at least once tried the additional functions of the MT system such as the Controlled Korean Checker. We will call such users as "active users".

The number of the active users was 344, comprising around 50% of the whole users. The number of the sentences that the active users translated with the system was 38,096. Among the 38,096 sentences, 30,954 sentences (81.25%) were translated without using any additional functions. The active users simply accepted the translation results generated by the MT system. Only when translating 7,142 sentences (18.75%), the additional functions were put into use.

Among the additional functions of the MT system, the users accessed to the Controlled Korean Checker most frequently. In section 2 we showed that by employing the Controlled Korean Checker the accuracy can be improved for 4.5%. The function that the users used second most frequently was the case structure revision function. In the Korean syntactic analysis, the normalization of the topic markers, or in other words, the case disambiguation of the topic markers is very critical to the correct analysis of the case structure. In 236 sentences, the users corrected the case structure analysis error in the intermediate step. Although the users were provided with a user-friendly interface for the interaction with the system, the other functions were not so frequently used by the users contrary to our expectation. cf. Table 1

Total number of re-translation	Controlled Korean Checking	Morphological Analysis Revision	Dependency Structure Revision	Case Structure Revision	Target Word Generation Revision
7,142	6,725	37	47	236	97

Table 1: The number of the re-translated sentences

## 5. Conclusion

In this paper we presented our approach to develop a high-quality Korean-English MT system for scientific paper translation. In order to improve the translation quality enough for the scientific paper authoring in English, we proposed the concept of the 'Controlled Korean' and the 'Interactive MT'. In designing the Controlled Korean we didn't introduce the controlled lexicon as done in other related researches, but rather focused on the writing rules. As the learnability and the usability are very important factors for the acceptance by the users, we tried to formalize the Controlled Korean rules as much as possible for the automatization. We have built around 6,000 rules for the Controlled Korean Checker. As a result, we could improve the translation accuracy of the Korean-English MT system for 4.25% by employing the Controlled Korean Checker.

At the same time, by allowing the users to interact with the system in the analysis as well as in the target word selection step, we gave the users the chance to reflect his/her linguistic intuition on the translation result. However, on the contrary to our expectation, most of the active users didn't use the interactive functions quite often except the Controlled Korean Checker. It could imply that the users were satisfied with the raw translation result on one hand. On the other hand, it could mean that the user interface for the interactive MT is still not user-friendly enough for the users to take advantage of the interactive system. Among the interactive functions, the active users found the case structure revision function most useful after the Controlled Korean Checking. The improvement of the coverage of the Controlled Korean Checker seems to have the biggest impact on the overall translation quality at the current stage of the development. Also the user-interface of the interactive functions must be reconsidered as the users don't actively make use of them. In order to find out if the interactive functions are really useful as the researchers may expect or hope, a more thorough user study must be performed.

## References

- AECMA (1995): AECMA Simplified English, A Guideline for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language, Issue I
- Foster, G, Isabelle, P., & P. Plamondon (1997): Target-Text Mediated Interactive Machine Translation, *Machine Translation*, Vol12
- Hong, M., Kim, Y., Kim, C., Yang, S., Seo, Y., Ryu, C. & S. Park (2005): Customizing a Korean-English MT System for Patent Translation, Proceedings of MT-Summit X
- Langlais, P., Foster, G., & G. Lapalme (2000): TransType : a computer-aided translation typing system. In Workshop on Embedded Machine Translation Systems.
- Lehrndorfer, A. (1996): Kontrolliertes Deutsch: linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation. Tübingen: Narr.
- Lehrndorfer, A. & R. Mangold (1997): How to Save Money in Translation Cost, TC-Forum 97-2, URL: <http://www.techwriter.de/tc-forum/pdf/editions/tcf972s.pdf>
- Lehrndorfer, A. & S. Schachtl (1998): Controlled Siemens Documentary German and TopTrans, TC-Forum 98-3, URL: <http://www.tc-forum.org/topictr/tr9contr.htm>
- Mitamura, T. & Nyberg, E. H. (1995): Controlled English for Knowledge-Based MT: Experience with the KANT System, Proceedings of TMI-95.
- Mitamura, T. (1999): Controlled Language for Multilingual Machine Translation, Proceedings of MT-Summit 1999
- Möller, M. (2003): Grammatical Metaphor, Controlled Language and Machine Translation, Proceedings of EAMT/CLAW 2003
- Yamabana, K., Kamei, S., Muraki, K., Doi, S., Tamura, S. & K. Satoh (1997): A Hybrid Approach to Interactive Machine Translation – Integrating Rule-based, Corpus-based, and Example-based Method, Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.

## MASTAR プロジェクト

独立行政法人 情報通信研究機構 中村 哲

### 1. はじめに

独立行政法人情報通信研究機構（以下、NICTと記述）では、ユビキタスネット社会の実現に向けた総務省の「u・Japan 政策」に基づき、人と人との豊かなコミュニケーションを実現するためのユニバーサルコミュニケーションの研究を進めています。ユニバーサルコミュニケーションでは、言語・文化・能力などの壁を越えて自由にコミュニケーションが行える環境を実現するための技術開発を行います。この壁の中で、人と人との言葉の壁、人とコンピュータの言葉の壁を越える技術を研究開発すべく、平成 20 年 4 月から開始されたプロジェクトが MASTAR (Multi-lingual Advanced Speech and Text Research) プロジェクトです。

### 2. MASTAR プロジェクトの狙い

近年、音声・言語処理は飛躍的な進歩を遂げています。この理由の一つとして挙げられるのが、大量のデータ（コーパス）を収集し、用例や統計モデルと機械学習により、自動的に処理系を構築する方法論（コーパスベース技術）が確立されたことです。このことは、これまで、研究開発フェ

ーズでデータを収集し、システムを構築、その後、実用化に結びつけるといった従来の開発プロセスでなく、実際に使用される場面でのデータを直接収集し、それを機械学習に用いることで、研究開発フェーズから実際の場面での性能向上を直接行える、新しい研究開発プロセスの到来と考えることが出来ます。さらに、WEB の普及、発達は、さらなる進歩を生むと考えられています。Web の仕組み、Web 上の情報を利用することで、世の中にある固有名詞の取り込み、多言語辞書の構築やコーパス収集、単語の関係抽出、信頼性などの解析を行うことも可能になります。

### 3. MASTAR プロジェクトの構成

本プロジェクトの目標は、Web、ネットワークを利用し、産業界、社会とリンクした形で持続的に、音声・言語資源を蓄積、成長していく研究開発の仕組みを作ることです。具体的には、以下の 4 つの研究開発を行います。

- ① 総合科学技術会議の社会還元加速プロジェクトの一つに選定されたネットワーク音声翻訳に関する技術開発、社会実験、社会還元を行う。

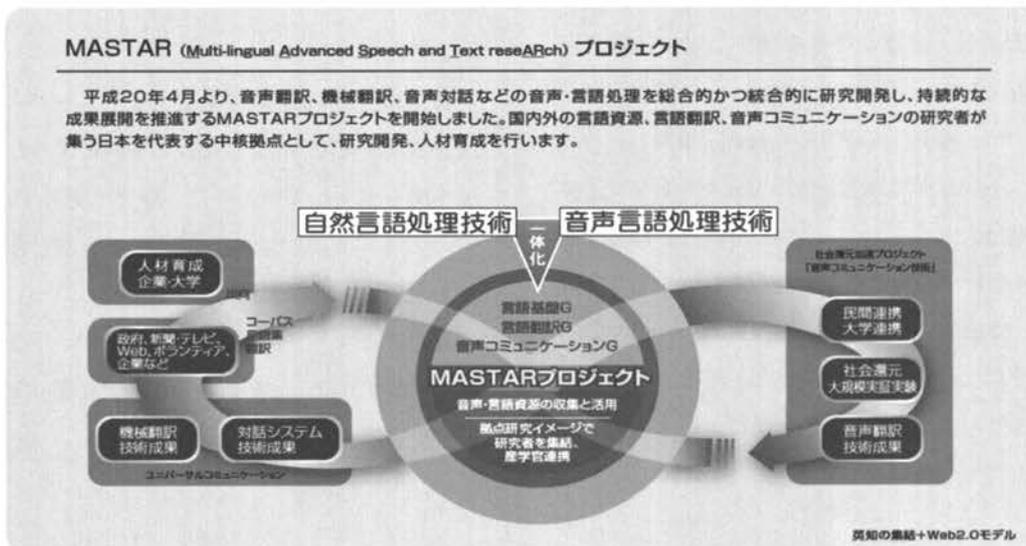


図1 MASTAR プロジェクトの概要

- ② 産業界とマニュアルなどを対象とした WEB2.0 型の機械翻訳サービスを行い、共通辞書、共通コーパスの蓄積、翻訳技術研究のポジティブ成長サイクルを確立する。
- ③ ユニバーサルコミュニケーションの一環として、自治体などを通じあらゆる利用者へ情報を届けるための音声対話インタフェース技術の技術開発、社会実験、社会還元を行う。
- ④ 世界的言語資源を構築し、配信、還元する。

上記の研究開発を進めるため、NICT では、NICT 起源の自然言語処理の研究グループ、ATR 起源の音声・言語処理グループを融合し、アカデミア、産業界からも新しくメンバーを迎えて、新しく 5 年間のプロジェクトを構築しました。プロジェクトは、言語基盤グループ、言語翻訳グループ、音声コミュニケーショングループから構成され、有機的に協力し 4 つの研究開発を進める形となっています。

#### 4. MASTAR プロジェクトのテーマ

##### ① 音声翻訳研究

音声翻訳は、音声認識、機械翻訳、音声合成の 3 つのモジュールから構成されます。不特定の話者の自由発話をいかに認識し、翻訳し、自然な音声を作成するか、多くの音声・言語処理の課題を含むプロジェクトでしたが、現在、比較的短い文章からなる典型的な旅行会話に対して TOEIC600 点相当の人間の翻訳品質と等価な性能を達するまでに研究開発が進みました。この技術を核に、5 年間のプロジェクトでは、種々の実証実験を計画、検証し、バックキャスト的な研究開発、社会還元を行います。技術的課題は大きく 2 つで、一つめは固有名詞の問題で、膨大な物や場所の名前の登録をいかにを行い、音声認識、機械翻訳をいかに高精度に行うかです。二つめの課題は対象言語の増加です。現在の日英中の 3 言語に加えて、より多くの言語に対応する方法論の確立を行う必要があります。

##### ② テキスト翻訳

用例ベースや、統計ベースのコーパスベース技術では、対訳データを収集することで翻訳エンジンを学習することが出来ます。本プロジェクトでは、汎用の対訳データを Web から対訳を自動的に抽出する技術、対訳データを多言語化することで翻訳エンジンの多言語化を行います。一方、翻訳性能は学習コーパスに大きく依存するため、実際に利用される場面でのデータを収集することが不可欠です。MASTAR プロジェクトでは、実際のユーザと共同で、既存の対訳データによる翻訳エンジンの学習、翻訳、人手ポストエディット、翻訳エンジンの学習というループを構成し、翻訳性能の改善、ポストエディットの削減、実用化を一体的に進めます。

##### ③ 音声対話システム

音声による多言語対話システムの研究開発を行います。音声に加えて、非言語的な情報を考慮した話し手と息の合った対話を行い、あらかじめ設定されている情報のスロットを埋める対話だけでなく、内容を理解して、Web、知識ベースから情報を取り出し、適切な形で情報を提示するプロアクティブな対話システム技術を研究開発します。音声対話システムの構築でも、コーパスベース技術を取り入れ、Web 上の知識を適切に提示する音声対話インタフェース技術の確立を目指します。

##### ④ 言語資源

音声翻訳、機械翻訳で必要となるコーパスや辞書などの言語資源、およびそれらを利用するためのツール群を開発します。特に、辞書に関しては Web 上のテキストから世界最大規模の辞書を自動構築し、日々更新することを目指します。さらに、大量の汎用対訳コーパス、対訳辞書も自動、手動等の手段を問わず作成し、多数の団体、研究者に配信することを目指します。

### ⑤言語グリッド

MASTAR プロジェクトで作成するものも含め、世の中には多種多様な言語資源やツールが存在しています。言語グリッドとはそれらの言語資源やツールをインターネット上で共有し組み合わせることで、現場に応じた精度の高い翻訳サービスを実現する基盤技術です。

### 5. 成果例

図2に音声翻訳システムの図を示します。現在、小型のPCに日英中の音声認識、音声合成、機械翻訳モジュールがインストールされた小型スタンドアロンシステムと、携帯電話をフロントエンドとしネットワークで接続されたサーバで音声認識、機械翻訳、音声合成をするネットワーク接続携帯電話システムを開発しています。このシステムを用いて、H20年8月の北京オリンピック期間に日本からの旅行者に対する実証実験も行い、有効性を確認しています。

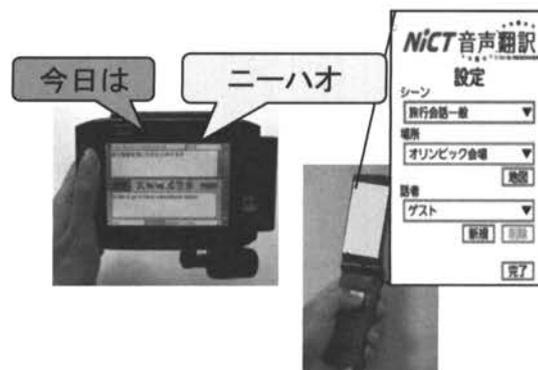


図2 日中音声翻訳システム  
(スタンドアロン版、携帯電話版)



図3 多言語テキスト翻訳システム

図3に統計翻訳による多言語テキスト翻訳システムを示します。現在、17言語間の旅行会話テキストの翻訳が可能になっています。

図4には、音声対話システムを示します。小型PCに音声認識、理解、対話管理、音声合成、表示を行うシステムです。

図5にはWebからの概念辞書獲得システムの様子を示します。昨年度のWeb情報から得られた情報の中にすでに残留農薬への関連が見いだせることを示しています。



図3 ネットワーク音声対話システム

### 5. 終わりに

MASTAR プロジェクトでは、産学官の開かれた共同研究体制を構築し、音声・言語に関する拠点を構築すべく研究開発を行っていきます。この中で、実世界データとネットワークを活用した成長的研究開発を進め、技術の社会還元を進めます。また、いろいろな企業からの出向、共同研究を行い、研究開発のみならず、人材育成を行っていきます。

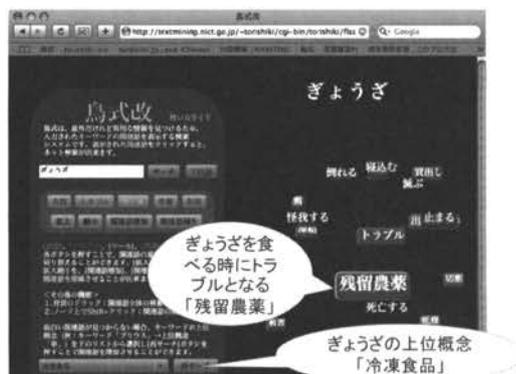


図4 Webからの概念辞書自動獲得

## 北京オリンピック音声翻訳モニター実証実験

独立行政法人 情報通信研究機構 清水 徹、中村 哲

独立行政法人情報通信研究機構（以下、NICTと記述）MASTARプロジェクトでは、ユビキタスネット社会の実現に向けた総務省の「u-Japan 政策」に基づき、人と人との豊かなコミュニケーションを実現するためのユニバーサルコミュニケーションの研究を進めています。

言語の違いはコミュニケーションを自由に行うことの障害となるため、経済活動等のグローバル化に伴い、音声・言語の研究は、ますます重要になってきています。言葉の壁を乗り越え、さまざまな話題や環境で話された日常の話し言葉を、正しく認識したり、翻訳・合成したりする技術や、インターネット上に流通する様々な情報を翻訳する技術は、言語の壁を乗り越えた新しいコミュニケーションを可能にする技術として大きな期待が寄せられています。

本プロジェクトの活動の一環として、音声翻訳の研究開発ならびに成果展開の推進を目的とした北京市内における音声翻訳モニター実証実験を8月7日から8月12日まで行いました。

この実験のために、携帯電話に向かって海外旅行会話を中心とした日常会話を話しかけると、中国語への翻訳結果が音声で再生される世界で初めての「携帯電話音声翻訳サービス」を新たに開発し、北京オリンピック観戦ツアーなどに参加する日本人旅行者や北京在住の日本人のモニターを募集し、実際に北京市内で利用してもらいました。モニターには、観光、ショッピング、道案内等で、このサービスを使って中国の人とコミュニケーションしていただくとともに、その体験についてアンケートにも答えてもらい、音声認識性能、翻訳性能、使用性の両面から、音声翻訳の実用化に向けた課題を明らかにしていく予定です。図1は、携帯電話音声翻訳サービスのイメージを示したものです。



図1 携帯電話音声翻訳サービス

このサービスでは、次のような新しい機能が導入されています。

- (1) 予めいくつかの文を発声することでユーザの声の特徴を登録することができます。
- (2) 場所や場面に応じて、それぞれの条件に合った言葉の辞書に切り替えることができます。
- (3) 翻訳結果の音声を合成することができます。
- (4) 翻訳結果を再度もとの言語に翻訳して、翻訳結果が正しいかどうかを確認することができます。

また、今回の実験のために、中国の清華大学と協力して北京市内の観光、ショッピング、移動、食事に用いられる数千語の固有名詞辞書を新たに整備しました。図2に音声翻訳システムの構成を示します。

本実験では、「携帯電話音声翻訳サービス」以外に、ビジネス手帳大の小型のパソコンに音声翻訳の全ての機能を内蔵した「音声翻訳専用機」での実験も行いました。これは、従来は大型サーバを用いていた音声翻訳機能を、手に載るサイズに小型化したものです（図3）。

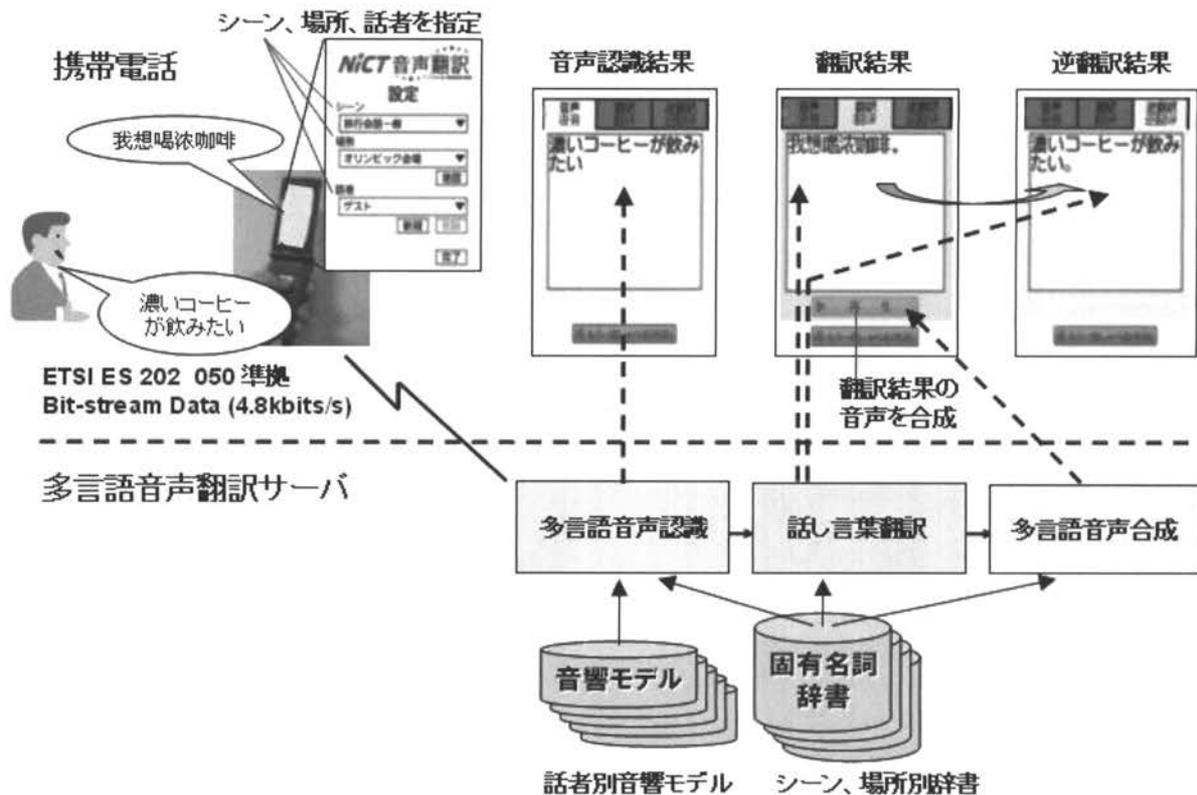


図2 携帯電話音声翻訳サービスの構成

モニターからのアンケート結果によれば、音声を音声に翻訳する技術への期待は大きく、自動音声翻訳機を早く実用化して欲しいという声が多かったほか、携帯電話で音声からの音声への翻訳が実現されていることについて、技術がここまで進展していることについての驚きの声が開かれました。その反面、様々な言い回しを正しく音声認識・翻訳して欲しい、レストランや品物の名前など単語をもっと入れて欲しい、スムーズな会話が可能なように処理速度を速くして欲しい、初めての人でも使い易い画面インターフェースにして欲しい、相手の声も精度良く認識して欲しいなどの様々な意見が寄せられました。

今後、ソフトウェアの改良を進めながら、日本を訪れる中国人観光客や中国を訪れる日本人観光客向けの実験により、性能検証を継続していく予定です。このような実証実験と開発の両輪を回していくことにより、より使い易いシステムを実現して行きます。



図3 音声翻訳機能をビジネス手帳大のPCに全て内蔵した音声翻訳専用機（日本語、中国語、英語間の相互の音声翻訳が可能）

参考文献：

清水、他、“携帯電話を用いた日中音声翻訳実証実験システムの試作”、日本音響学会平成20年度秋季研究発表会 2-1-6、2008年9月

## COLING2008 における機械翻訳研究の傾向

東京大学 清水 伸幸、宮尾 祐介

独立行政法人 情報通信研究機構 河原 大輔

## はじめに

COLING2008 は、8 月 18 日から 22 日にかけて UK の Manchester で開催されました (図 1)。連日 35 度に達していた日本と比べると、Manchester の 18 度程度という気温は肌寒く、夏であるにもかかわらず長袖の衣服が必要でした。日本から来た研究者の幾人かは風邪の兆候が見受けられ、健康管理の大切さを感じる会議となりました。

機械翻訳の観点からみると、本会議では 7 つものセッションが機械翻訳で占められており、いずれも大きなホールが割り当てられていました。機械翻訳に対する注目度は大変高かったように思います。

本稿では、まず全体の機械翻訳の傾向を述べてから、筆者らが特に興味を持った個別の発表についてお伝えしたいと思います。

## 全体の傾向

BLEU という機械翻訳の評価手法が一般に使われるようになってから、Giza++, Pharaoh など基礎的なベースラインの開発を容易にするツールの公開もあり一気に広まったように感じられる Phrase-based SMT ですが、今回の会議でも、大半の論文がこの手法に関係しており、このモデルの一部分、特に弱点である翻訳後の語順の改善的を絞った論文が多く見受けられました。これには幾つかの手法があり、あらかじめソース言語のテキストをターゲット言語の語順に近づけてから SMT を走らせる、SMT の一部分に機械学習の分類器を組み込んで、語順に関する特別なケースを見分ける、などの手法もありますが、特に、Phrase-based SMT の延長線上で、並行コーパスからフレーズのアライメントと共に、ソースとターゲット言語間の語順変化も統計的に学習する Hierarchical SMT が、SMT の大きな流れになりつつあるようです。また、幾つ



図 1 会場の Manchester 大学

かの機械翻訳の出力を何らかの形で組み合わせる手法も徐々に増えてきたと感じました。また、機械翻訳の評価手法について近年活発な研究が行われていますが、COLING においても評価に関する研究が複数ありました。特に、性質の異なる機械翻訳システムを横断的かつ公平に評価する研究や、機械翻訳システムの特徴を詳細に分析できる評価手法など、機械翻訳の細かい性質に深く踏み込んだ研究が多いのが特徴的でした。

## 個別の発表

他分野との関連があり、機械翻訳を含む発表で筆者らが特に興味を持ったのは以下の通りでした。

まず、Shawe-Taylor による Keynote lecture では、Kernel regression を機械翻訳に使った自身の研究についての言及がありました。非常に機械学習らしい手法を応用したもので、機械学習の立場から機械翻訳にアプローチしたらこうなるだろうな、という雰囲気に興味深かったです。ただ、まだ機械翻訳のような大量に並行コーパスを処理しなければならない分野では実用性に疑問符がつく印象があり、データマイニングや、そこに含まれる教師なし学習のようなパターン発見型の分野の応用を視野に入れた方が有望なのではないかと感じました。

この点で、[Lopez]は、データマイニング的な Fast Lookup の手法を用いて、大規模なコーパスからテストデータの翻訳に必要な Translation Model をオンデマンドで計算し高速化する手法でした。また、それに合わせて可能になる Hierarchical SMT の語順やフレーズ組み合わせのパターンの拡張について実験しています。

もう一つ、機械学習的なアプローチでは、[Xu et al.]らが、機械翻訳に必要な中国語の単語境界線を半教師ありのペイズの手法で単一コーパスと並行コーパスの両方から学習し、機械翻訳の結果が向上することを実験で示しました。このような手法の応用はまだ多くはないですが、計算能力の向上に伴い、これから徐々に増えていくのではないかと思います。

次に、機械翻訳の評価についてですが、[Zhou et al.] は、言語現象（言語的チェックポイント）ごとに機械翻訳の精度を測定するという手法です。例えば、訳語の選択、イディオムの訳出、動詞と目的語の共起、前置詞句の訳出などに着目し、それぞれのチェックポイントごとに精度を測定します。BLEU や NIST などの単一のスコアを使った場合には隠れてしまう個々のシステムの特徴が、この手法により明らかになると期待できます。

この研究の面白い点は、テストデータを自動生成するところです。具体的には、辞書マッチング、単語アライメント、両言語のパーザ（句構造と依存構造）などを用いて、各チェックポイントのインスタンスを自動生成しています。これらの自動処理に依存することで当然エラーによる影響も考えられますが、各処理のエラーの影響を小さくする工夫や、エラーがどの程度最終的な評価に影響するかの分析も行っています。実験では、統計的機械翻訳システム間の比較、および統計ベースシステムとルールベースシステムとの比較を行っています。例えば、前処理でフレーズの並べ替えを行うシステムと行わないシステムの比較では、BLEU スコアではほとんど差がありませんが、提案手法による評価では動詞と主語や目的語の関係の精度が向上していることが示されています。また、統計ベースシステム

とルールベースシステムとの比較では、後者の方がイディオムや代名詞の訳出で精度が高いことが示されています。このような詳細な分析は、異なるシステムの比較だけでなく、システム開発においてどの部分を改良すべきかの指針を与えてくれるものとしても期待されます。

[Zollmann et al.] は、異なるアーキテクチャの統計的機械翻訳手法を系統的に比較した研究です。具体的には、phrase-based, hierarchical, syntax-augmented モデルを複数の言語対や異なるデータサイズで比較し、それぞれの特徴を観察しています。実験では、中国語-英語翻訳やウルドゥ語-英語翻訳では hierarchical と syntax-augmented による精度向上が見られるがアラビア語-英語翻訳ではほとんど効果がないこと、大きな言語モデルを利用することでどのモデルでも同程度の精度向上が見られること、などが示されました。近年では単純な統計的機械翻訳ではなく句構造や依存構造などより複雑な構造を用いる手法が活発に研究されていますが、この研究はそのような構造による効果がどのように現れるのかを詳細に検証している点で興味深い研究です。

## おわりに

今回の COLING は、2010年8月23日から27日に北京で行われる予定です。2008年の北京においてはオリンピックが行われ、機械翻訳技術がさまざまところで活躍したようですが、2年後にはさらなる実用化・発展が期待でき、非常に楽しみです。

## 参考文献

- Lopez, A. Tera-scale Translation Models via Pattern Matching. COLING2008.
- Xu, J. et al. Bayesian Semi-supervised Chinese Word Segmentation for Statistical Machine Translation. COLING2008.
- Zhou, M. et al. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. COLING 2008.
- Zollmann, A. et al. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. COLING 2008.

マレーシア会議報告 (11<sup>th</sup> International Conference on Translation)

株式会社ナビックス 村上 嘉陽

約1年前になるが、平成19年11月20日(火)～23日(金)の日程で、11<sup>th</sup> International Conference on Translation への参加のため、マレーシア・クアラルンプールを訪問した。

20日には夜間に宿舎に到着し、23日の帰国の際には昼頃のフライトとなったため、実際には2日半程度の滞在となった。

都合により会議の最終日である23日を移動日としたため、3日間の会議(最終日は半日)の初日と2日目のみに参加し、2日目の午前のセッションで20分程度のプレゼンテーションを行った。

会議はITNMB(The Malaysian National Institute of Translation)が主催し、メインホールは300人程度を収容できる規模であった。

テーマは翻訳であり、その内容は文芸翻訳から機械翻訳に至るまで、実に様々であった。

初日は文芸翻訳が中心となったが、2日目の午前部はCAT(Computer Aided Translation)がテーマであり、私も含めて4名が登壇した。因みに私以外の登壇者は工科大の教授クラスが2人と、あとは工科大のマスターコースの学生が1人で、順番は私がトリだった。

初日よりも2日目の参加者が多く、主催者側の発表によると、2日目の午前中は200名以上の参加者がいたとのことである。(2日目の午後は、さらに参加者が増えた。)

今回のプレゼンテーションの主旨は、日本の翻訳事情の紹介だったが、とりわけ最近の新たな動きの一つとして、制限言語(Controlled Language)を使うことで、機械翻訳や翻訳メモリの活用をさらに効果的なものにするるとともに、英語から更に多言語へ翻訳展開する際のミスやロスをなくすというお話をご紹介させて頂き、概ね好意的に受け取られた。

また京都大学の石田教授にご了承を頂き、マレーシアのマラヤ大学も参加した日中韓馬の4カ国による国際コラボレーション実験(ICE2002)のお話も併せてご紹介させて頂いた。

同じセッションで最初の講演者となったテナガ国立大のZaharin Yusoff教授は、事務局の方のお話によると、マレーシアではMT関係で有名な先生とのことだったが、講演終了後に「お互い同じようなことを考えていますね。」と言われ、相互にパワーポイントのプレゼンテーション資料を交換した。

またプレゼンテーションの終了直後には、アイルランドから参加されたスティーブン・ドッド博士(英語学や英語教育学の観点から語彙の制限を研究)から声がかかり、大変興味深い講演だったとお褒めの言葉を頂いた。博士のお話では、ナビックスはマニュアル制作等に代表されるテクニカルコミュニケーション分野での語彙限定を考えているため、より現実的なアプローチとして成功できるのではないかと思うとのことだった。因みに博士は分野限定をせずに、日常生活全般を対象にした語彙の限定を考えておられるため、(語彙を限定して言葉を)簡単にするための取り組みなのに、大変困難なテーマだと苦笑されていた。

今回のホストであったITNMBの所長のご夫人も私のプレゼンテーションを聴講されていたらしく、「あなたの考えは銀行送金でバッチリだと思うわ」との一風変わったご指摘を頂いた。これはICE2002に対するコメントだったわけだが、ご夫人が仰りたかったことは、銀行送金とは言うまでもなく一定の金額を送金する大変重要な行為であり、間違いが許されない。したがって間違いや誤解を招くリスクを回避するためにも、コミュニケーションをする両者の言語が、例えそれが何語であってもシ

ンプルであるべきであり、そのためには双方が Simple Language を強く意識すべきだと仰りたかったようである。セッション終了後の休憩時間にこうしたお話を頂いたわけだが、「新しい技術に取り組む人は常に間違い扱いされるなど、大変な思いをすることが多いけれど、必要な技術であれば必ず達成できるから、今にあなたはヒーローになれると思うの。だからどうか頑張って！」とも激励して頂いた。ITNMB から招聘された理由を確認してみたところ、2つあった。1つは日本の通訳や翻訳事情をマレーシアで紹介することだったが、もう1つは、会議の一年半前に ITNMB の職員が日本を訪問して翻訳関係の企業や団体をインタビューして回った際に、ナビックスだけが MT に関して否定的なことを言わなかったからとのことだった。因みに日本でのインタビューは、翻訳会社の関係者など 20 人以上に及んだそうだが、どこも「MT なんて使い物にならない」といった雰囲気だったのだそうで、それに加えて翻訳関係の仕事に携わっているが、英語でのインタビューに対応できない人が多く、ITNMB 側は大変不思議且つ不便に感じたのだそうである。

インドネシアやマレー半島の文化研究を専門とする研究者で、偶然フェアウェルパーティーで隣席に座ったフランス人の女性からは、「MT はまだ文芸翻訳では使えないのかしら？」と真顔で質問された。彼女の質問からは MT そのものへの技術的な関心というよりも、むしろ文芸を通じて世界中の人々と感情を分かち合いたいとの思いが感じられた。日頃「MT は使えない」といった批評を耳にすることが多いが、このように世界には MT への期待感も存在していることを肌で感じた次第である。またマレーシアでは（お国柄もあってか）MT に関しては比較的寛容な態度で、使えるのなら新技術として積極的に採り入れようとした機運も感じられ、少なくとも日本の翻訳業界ほど毛嫌いしていないように思われた。

2 日目の午後の分科会では、会議の前年に ITNMB の職員が来日した際の報告を、マレー語で行っていた。会場では 30 人程度が耳を傾けており、ITNMB からも是非会場に顔を出して欲しいと要請されたため、私も分科会の会場に足を向けたが、ITNMB の職員がナビックスを訪問した際の写真がスクリーンに映し出された瞬間に（マレー語の分からない私が会場にいたため）、聴衆から拍手が沸き起こり、思わぬ所で心温まる歓迎を受けることができた。マレーシア国民学校で日本語を教えているマレーシア人の教員とそのお弟子さんからも、プレゼンテーション終了後に挨拶をされた。その際に名刺をお渡ししたところ、自分たちには名刺が無いのにもかかわらず、日本人から名刺を頂くことができたといふに喜ばれたが、その光景は、今でも大変新鮮に感じたことを思い出す。

初日の基調講演によると、マレーシアでは著名な文学作品などのマレー語への翻訳や出版のためのプロジェクトが 2006 年から始まっており、これが 2010 年まで続くとのことである。ITNMB では 2006 年に教育省から約 4,000 万円をプロジェクト推進のための費用として既に受け取っており、2007 年中にはこれを使い切る予定とのことだったが、アラビア語や欧州言語などからのマレー語への翻訳を充実させたいと考えている様子であった。

（今回の講演の謝礼代わりに、ITNMB で編集・出版した F・フクヤマの *The End of History and the Last Man* のマレー語訳の書籍を頂いた。）

教育省の肝入りもあり、また国会議員が ITNMB の Chairman をしていることもあって、今後は翻訳の社会的な認知度を上げていきたいとの機運が感じられた。

そのマレーシアの国会議員で、ITNMB の Chairman をされている Dr. Wan Hashim Wan Teh からは「あなたが日本からお越しになった方ですね？」と声をかけられ、ご挨拶をすることができた。博士曰く日本のことにも触れた本を書いたとの

ことで、是非とも東京に郵送したいとお申し出を頂き、快諾した。(後日著書は郵送されてきた。)

マレーシアはマレー系、インド系、中国系の人々が混在した国であり、訪問した 2007 年は建国 50 周年に当たる記念の年でもあった。

KLIA (クアラルンプール国際空港) は、マレー語、英語、中国語とともに日本語にも対応しており、エアポートライナーの乗り場では日本語のアナウンスも流れるほどで、実に快適な旅であった。しかし帰国直後に、会場に程近いクアラルンプールの中心地付近でインド人による暴動が発生したことを新

聞報道で知った。実はその地域が、ついその数日前に私が足を向けた界限だったため、にわかには信じがたく、最初は他の国の間違いではないのかと報道内容を疑ったほどであったが、その後の報道を追いかけながら、まだまだ訪問した国への基本的な理解が欠如していたことを痛感した次第である。

また訪れる機会もあることと思われるが、その際にはマレー語を話すことはできないまでも、せめて今回お世話になったことに感謝しつつ、マレーシアについての理解を深めてから改めてお邪魔したいものである。

## TC シンポジウム報告

株式会社ナビックス 村上 嘉陽

去る8月26、27の2日間にわたり、テクニカルコミュニケーター協会（以後 TC 協会）主催によるTC シンポジウムが都内で開催された。

今回のテーマは「意匠×情報」であり、その意図するところは、世の中にはデザインの数だけ情報があり、また情報の数だけデザインがあるとの認識から、新たなマニュアルのデザインを追求するということではないかと思われる。

今年は約50セッションが会期中に開催されたが、第20回となることを記念して、ドイツと韓国からテクニカルコミュニケーションの団体を招聘したことや、マニュアルコンテストでは今年から新設された組込マニュアル部門の日本ビクター（液晶テレビの組み込みマニュアル）がマニュアル・オブ・ザ・イヤーを受賞したこともあり、話題の多いシンポジウムとなった。

初日の午前には Web コンテンツクリエイターの三井英樹氏による基調講演で始まったが、この基調講演に加えて、ドイツの団体（tekomp）によるプレゼンテーションやパネルディスカッションが両日開催されたため、今回は tekomp のセッションを中心に聴講した。

## ＜基調講演＞

内容としては、製品取説情報をよりよいユーザー体験とするために、テクニカルコミュニケーターは Web の RIA（Rich Internet Application）技術から何を学び取ることができるのかといった視点での話であった。

印象的だったのは、Web の場合は情報を提供して理解してもらうだけでは最早ユーザーに満足してもらうことはできず、提供した情報をユーザーが活用して共感するところまで至らなければならなく

なっており、その段階まで達しなければ、情報として見なされなくなっているとの報告であった。さらに「情報」とは「知っていて楽しくなる情報」と「知っていると助かる情報」、そして「知っておくべき情報」などに分類されるため、Web の世界では最適な技術や表現、センス、テキスト量などのいくつかの検討すべきポイントがあるとの指摘も、併せて印象的であった。

聞けば Web の世界も、日々作り上げるものの表現力や機能性、操作性などに腐心しながらも、制作段階では徹夜仕事が珍しくないらしく、実は他ならぬ自分たちこそが果たして sustainable（持続可能）なのか、即ち仕事を続けて行くことができるのか、そしてそのような3Kまがいの業界で、若い人材が果たして育つのかといった課題も横たわっているとの話もあり、労働集約的な作業に携わるローカライズベンダーとしては、それこそ深く共感するような内容であった。

とにかくユーザーに対するホスピタリティを忘れず、「見て分かる情報」から「活用して知る情報」を提供できるように日々考え続けることは必須であるがゆえに、結論としてはマニュアルを制作するテクニカルコミュニケーターにとっての答えは「頑張ることしかない」となったが、ただ関心を持ったのは、Web の世界では（不仲で有名な）デザイナーとエンジニアの双方の役回りができる人材が増えてきているという話である。つまりデータベースが分かるデザイナー、デザインの分かるソフトエンジニアといった具合に、従来いずれかであった職種が融合したかのような存在の若者が、確実に増えてきているのだそうで、ご本人は喜ばしいとのことであったが、当方からすれば、うらやましい限りの話であった。

講演の最後には、真珠は貝の体内に異物が混入することで天然真珠を生成する話を持ち出し、新しい人材は古い組織には馴染まず、また新しい技術も古い習慣には馴染まないことに触れながら、何事も変えることを恐れてはならないとのコメントを残しての降壇となった。

#### <VDE>

午後は、スイスのチューリッヒに拠点を持つ VDE という会社の Dr. Konstantin による使いやすいマニュアルの提案と題したプレゼンテーションと、tekompによる自らの組織や活動に関するセッションにそれぞれ参加した。

VDE のプレゼンテーションの内容は、Dr. Konstantin 自身がプロジェクトマネージャーを務める SMART Manual の紹介であった。端的に言えば、製品のテストを主業務としてきた VDE が、人間工学的に好ましいマニュアルのデザインを企画して実現したマニュアルの制作事例ということらしく、サンプルを拝見する限りでは、マニュアルの文書構造やレイアウトをシンプルにして見やすくすることで、説明の読みにくさや分かりにくさの解消、翻訳コスト・制作コスト・印刷コストの圧縮といった諸課題の克服を目指す取り組みであるように思う。

因みに VDE は、日本では大阪と東京にオフィスを構えているが、日本では製品のテスト業務しか行っていないのだそうで、ゆえに SMART Manual は全て欧州で取り仕切っているとのことであった。このため、もし SMART Manual に興味がある場合には後日連絡させてもらうこともあるかもしれないと話したところ、Dr. Konstantin には歓迎された。

#### <tekomp>

tekomp は、2 日目には全日にわたって Dr. Konstantin らの VDE のメンバーを交えるなどしてパネルディスカッションを開催したが、とりわけ

午前のパネルディスカッションで、欧州のローカライズ事情について行われたいくつかの報告は、大変参考になった。

上述の Dr. Konstantin からは言語の使用状況に関する報告があったが、それによると欧州では現在 31 言語が使われているそうである。そしてそのうち、広く使われている言語を調査した結果、上位は以下の通りとなった。

	A	B	C
1. 英語	13%	38%	51%
2. ドイツ語	18%	14%	32%
3. フランス語	12%	14%	26%
4. イタリア語	13%	3%	16%
5. スペイン語	9%	6%	15%
6. ポーランド語	9%	1%	10%
7. オランダ語	5%	1%	6%
8. ロシア語	1%	6%	7%
9. スウェーデン語	2%	1%	3%

A: 母国語として使用

B: 外国語として理解できる

C: 外国語として話せる

上記の通り、EU 人口の 51% が英語を使うことができると回答していることからして、最も広く使われている言語は英語であることが分かるが、より詳細に国別で見ると、スウェーデンでは 89%、マルタでは 88%、オランダでは 87%、デンマークでは 86% の人々が英語を母国語以外の言語として使っているとの報告もあった。

また EU 拡大に伴って、フランス語とドイツ語のバランスに変化が出てきたことも指摘され、要するに新加盟国は旧東欧圏でドイツ語に習熟している国民が比較的多いため、国民の 24% がフランス語を使用しているルーマニアを除き、欧州ではドイツ語を使用する人々が増えたとの内容であった。

ただしテクニカルコミュニケーションの観点から重要なのは、フランスやドイツのような大国よりも、むしろ小国のクロスコミュニケーションの状

況、換言すれば、ひとりで複数言語を話す状況こそが現地の事情を端的に現すがゆえに、そうした点に着目することが重要であるとの指摘は印象に残った。例えばルクセンブルグでは 92%の国民が複数言語を話すことができるそうだが、バルト三国においては、使うことができるはずのロシア語を、国のアイデンティティー保持のために敢えて禁止した国もあるそうで、いずれにしても個々のマニュアルを欧州に合った文書にするための努力は、常に各国からのフィードバックを反映しなければならないとの指摘には、頷くところが多かった。

tekom の Executive Director である Dr. Fritz からは、こうした言語的な視点とともに、法的な視点や文化的な視点も重要であるため、tekom としては翻訳ではなく、ローカライゼーションのサポートをテーマとして、会員サービスを行っているとのことがあった。

因みに Dr. Fritz からは、文化的な視点の事例として以下のような指摘があった。

#### 日欧のマニュアルを見て感じた相違点

(博士が瞬時に思い出せたことのみ：順不同)

1. ビジュアルレイアウト
2. グラフの使い方
3. 写真(被写体が人の場合、欧州人は欧州のマニュアルには欧州人が写っているものと思っ  
ているため、黒人が写っているとアメリカの会社  
の、またアジア人が写っているとアジアの会社  
のマニュアルかなと思ってしまう)
4. 数の数え方(ドイツ人は1を人差し指ではなく  
親指から数える→イラストで人差し指だけが  
立っていると意味が分からない)
5. 色(東アジアでは赤はお祝いのイメージがある  
が、欧州では警告を意味する)
6. 技術の擬人化(日本人は技術を人や動物などに  
例えて表現することがあるが、欧州ではそうし  
た感覚はなく、違和感を感じる)
7. マニュアルの頁数(マニュアルは数百ページで  
はなく、せめて数十ページ)

計算言語学を専門とする同じく tekom の Dr. Siegel からは、ローカライゼーションのための Simplified English の有効性に関する言及があった。司会者からの質問は Simplified English は、ローカライゼーションにとってベストソリューションなのか否かとの問いかけだったが、Dr. Siegel はベターではあってもベストではないと言い切り、それよりも(とりわけ日本語は主語や目的語が省略されても文として成立してしまうため)、主語や目的語を必ず明記することや、文の長さを短めにするなどして、ソース言語を分かりやすく誤解のないように仕上げることの方が、単語が制限されて表現が不自由となる Simplified English よりもよほど効果的とのアドバイスがあった。(tekom では Dr. Siegel がワーキンググループを主宰しており、20 語以内で一文とするための取り組みをしているとの報告もあった。)

このパネルディスカッションの最後には、欧州の 10 年前と現在の比較に関する言及が Dr. Fritz からあった。それによると、10 年前にはローカライゼーションは欧州では議論もされなかったが、現在は数多くの国や言語を抱える集団に EU になったことにより、重要な議題のひとつとなり、TC (テクニカルコミュニケーション) 技術も大きく注目されるようになったのだそうである。

このように欧州人でさえも多様化する欧州の実情を掴みきれなくなってきた、これまでとは違った対応を心がけようと模索し続けていることを考えると、日本で行う欧州向けのローカライゼーションが、単に日本語や英語からの言語展開だけで終わるのではなく、常に現地からのフィードバックを反映させる過程であるべきとのアドバイスは正論と思われた。

欧州、とりわけドイツでは新製品を評価する専門誌が以前からあり、マニュアルも含めてそこでの評価が低ければ、その製品が売れることは絶対はないという。これに加えて、消費者保護などの観点から厳しい規制のある欧州の実情を考えると、今後メーカ

一各社からは、より現地の状況を踏まえた上でのきめ細やかな対応を望まれるようになる気がした。

#### <リコーの多言語翻訳支援システム>

2日目の午後は、リコーのセッションに参加した。タイトルが「多言語翻訳における効率的業務プロセス構築および支援ツール開発」であったため、内容はまさしく多言語翻訳支援ツールの開発報告なのかと考えたが、実際にはさらに幅が広く、製品改革からマニュアル改革、そして組織改革へと話が及ぶ、過去10年間のかなり大きなプロジェクトについての報告であった。

100名近く収容できる会場はメーカー各社から参加者が詰めかけるなどして、概ね満員の盛況ぶりだった。リコーの場合は99年当時、製品のデジタル化やカラー機の投入、プリンタ事業の立ち上げなどで大忙しだったようだが、これだけ新しい動きが出ると輸出のための翻訳作業も激増して、ただでさえ処理し切れていなかった日本での翻訳対応が、完全に破綻してしまうとの危機感が相当高まったとのことである。このため会社としては欧州言語への対応工数を半減させ、そのローカライズ費用を30%ダウンさせようと、多言語展開業務改革プロジェクトが発足したそうである（下記参照）。

2011年 プロジェクト発足

2012年 JLC（Japanese Language Center：日本の多言語翻訳管理・発注機関）設置

ELC（European Language Center：欧州の多言語翻訳受注・統括機関）設置

2002年 専任翻訳校正（ネイティブによる全言語の翻訳チェック）開始

※ 海外販社各社でネイティブチェッカーを採用し、人件費は日本側が負担。

2003年 用語翻訳スタート（翻訳者を1名から2名体制に拡充）

2004年 CLC（Chinese Language Center：中国）設置

2005年 TLC（Taiwanese Language Center：台湾）設置

多言語翻訳システム開発開始

2006年 トルコ語対応開始

2007年 多言語翻訳システム運用開始

この10年間で数億円単位の投資をした結果、リコーが手にした成果は以下の通りである。

1. カスタマイズしたMS-Wordでの英文執筆
  - ※ 英文は日本人ライターが直接起稿
  - ※ 日本人ライターがWord2003をカスタマイズしたオリジナルツールを使用
2. Web配信、紙マニュアル、MS-Word、PDFなどへのマルチメディア対応
3. 翻訳資産の共有による翻訳品質の向上
4. 翻訳資産の有効活用による短納期化とコスト低減の実現
5. ページレイアウト処理の完全自動化
  - ※ XSL Formatter+Oracle
  - ※ サーバ上で自動組版、事後の微調整なし
6. 修正及び最終確認作業の大幅削減
  - ※ 英文版のスタートを意図的に遅らせ、特に仕様変更などに伴う修正を極力減らす
  - ※ 手作業を極力減らすことで、最終確認作業を極限まで減らす

リコーの認識では、これらの成果は「組織、言語、距離の壁を破り、情報パスを簡素化した」証ということのようだが、究極的なことを言えば、結論は以下の二点に集約できるように思われる。

1. 多言語に翻訳する際の核となる英文の品質向上、翻訳校正業務の平準化といった具合に克服すべき課題は存在し、改革は未だ道半ばである
2. 飽く迄目的は業務プロセスの改革にあり、ツールはその実現手段に過ぎない

10年前のリコーはかなり混沌とした状況であったことからすると、参加者の関心事は当然のように社

内ルールの構築やその運用などに移ったが、管理運用はかなり厳格になされているようで、(ここまでは自社では到達できないという意味からか)一部の参加者からはため息が漏れた。

こうした革新的な取り組みの報告は、常に社外秘情報と隣り合わせとなって肝心なところがブラックボックス化するため、必ずしも参考になるとは限らない。その意味では今回のリコーの事例も例外ではないが、単なるマニュアル改善に止まらないスケールの大きさを考慮すると、参加者によってはかなり刺激的な情報となった可能性もあるように思う。

#### <TC シンポジウムを終えての感想>

会場では毎年必ずメーカーを中心としたクライアント企業の方々とお会いするが、今回も両日にわたって何人もの方々とお話をする事ができた。

その際に、とりわけここ数年の TC 協会の活動はパワーアップして内容的にも充実してきているとの感想が数多く聞かれたが、個人的には全く同感である。

しかしながら平日の開催ということもあって、業務の都合で時間が割けないケースも多々あり、実務担当者からすると内容が年々充実する一方で、足が遠のく現状が悩ましいことと思われる。

今回は欧州のローカライゼーションに関する情報に触れることができたが、MT に関連して言えば、tekom の Executive Director である Dr. Fritz がパネリストとして何気なく語られたことが、今でも気になっている。それは、10 年前に tekom にやって来た頃には MT を使うなどということは全く考えていなかったが、今では使う必要があると考えているといった内容のお話であった。

博士とは幸いにしてご挨拶ができたものの、残念ながらこのことをお聞きするまでには至らなかった。しかし博士から直々に tekom への入会を勧められ、自社が法人会員となることができたため、再会できることを祈念しつつ、欧州のテクニカルコミュニケーターたちが MT をどう使おうと考えているのかについて、今後も注視して行きたいと考えている。

# 「超多言語」のテキスト翻訳システムを携帯電話で無償サービス開始

ATR 音声言語コミュニケーション研究所

隅田 英一郎

## 1 概要と開発の背景

(株)国際電気通信基礎技術研究所(ATR)は、2008年3月31日より、携帯電話で利用可能な日常旅行会話を対象とした「超多言語」のテキスト翻訳のモニターサービスを無償で開始いたしました。ATRでは、「コーパスベース翻訳技術」と「世界最大の100万文からなる対訳コーパス」を開発し、従来の「ルールベース翻訳技術」の抱える開発コストや多言語展開への困難さといった問題点を克服いたしました。このコーパスベース翻訳技術を用い、上記対訳コーパスに新たな言語の翻訳文を追加し、ヨーロッパ系の9言語(英、独、デンマーク、オランダ、仏、イタリア、

スペイン、ポルトガル、ブラジルポルトガル語)、アジア系の9言語(日本、中国、韓国、ロシア、アラビア、インドネシア、マレー、タイ、ベトナム語)、併せて、18言語の間のすべての組合せ306通りのテキスト翻訳を実現しました。これは他に類例のない超多言語の翻訳システムと言えます。

本モニターサービスでは、日本語(または英語)から17の異なる言語へのテキスト翻訳とその逆方向のテキスト翻訳、合計66通りを公開しています。本サービスがカバーする国や地域(対象18言語が主要語である国と地域)は158ヶ国で世界の194ヶ国のほぼ80%に相当します。

携帯電話向け多言語翻訳サービス  
『超多言語翻訳』機能説明

本サイトで出来ること

【合計66通りの翻訳機能】

アラビア語  
ドイツ語  
\*英語  
スペイン語  
フランス語  
インドネシア語  
イタリア語  
\*日本語  
韓国語  
ポルトガル語  
タイ語  
中国語  
デンマーク語  
オランダ語  
ブラジルポルトガル語  
マレー語  
ロシア語  
ベトナム語

日本語 or ⇒ 英語

17言語同時翻訳の例

“こんにちは”

- アラビア語 ⇨ مرحبا
- ドイツ語 ⇨ hallo
- 英語 ⇨ hello
- フランス語 ⇨ hola
- ポルトガル語 ⇨ bonjour
- 中国語 ⇨ halo
- インドネシア語 ⇨ halo
- イタリア語 ⇨ buon pomeriggio
- 韓国語 ⇨ 안녕하세요
- ポルトガル語 ⇨ olá
- 中国語 ⇨ 您好
- 中国語 ⇨ 您好
- デンマーク語 ⇨ goddag
- オランダ語 ⇨ hallo
- ポルトガル語 ⇨ halo
- アラビア語 ⇨ صباح الخير
- ロシア語 ⇨ добрый день
- ベトナム語 ⇨ xin chào

同時逆翻訳の例

“はじめまして”を  
ブラジルポルトガル語へ

入力文 [ ]

はじめまして

翻訳結果 [ ]

come está

翻訳結果の日本語訳 [ ]

初めまして

“Nice to meet you.”を  
フランス語へ

Input sentences [ ]

Nice to meet you.

to [ French ] [ Translating ]

Translation result [ ]

enchanté

Re-translation result [ ]

nice to meet you

対応機種

NTTドコモ(FOMA)、KDDI au(WIN)、SoftBank(3G) それぞれの端末

※英語・日本語に関しては、入力モードによって切り替わる

図 1 機能概要

ATRは20年を超える音声翻訳の研究活動をベースに、2006年より所有する技術の事業化を進め、2007年11月26日に、世界で初めて携帯電話による日英間の旅行会話音声翻訳の商用サービスを、関連会社ATR-Trekを通じて始めました(本誌42号P2-4)。

今回、更なる成果展開を図るべく、超多言語テキスト翻訳のモニターサービスを開始しました。本公開モニターサービスを広く使っていただき利用者の御意見を取り入れ一段の翻訳性能向上とインターフェースの改善をはかり、一年後を目途に本格的なサービス(有償)を開始する予定です。

## 2 基本機能の説明

本サービスのあらましを図1に示しました。

- 対象分野：
  - ・ 旅行会話を中心とした日常会話
- 機能概要：
  - ・ ①並行して多言語に翻訳する機能
  - ・ ②翻訳結果を原文方向に翻訳し翻訳精度をチェックする機能(逆翻訳)

## ● 利用方法：

右のQRコードを利用するか、直接<http://atr-langue.jp/smlt/>を入力してサイトにアクセスするだけで利用できます。



## ● 対応端末

インターネット利用が可能な(インターネット接続機能とブラウザ機能を持つ)携帯電話であれば日本のどの携帯電話会社でも利用可能です。

次に、図2の実際の画面で、利用の流れをご紹介します。

- ① サイトのTOP画面から、「日本語から17言語に翻訳！」ボタンを押すと上段に示した入力画面になります。ここで、旅行場面の文を利用者携帯に備わった入力手段(キー入力や手書き文字認識など)で入力して、翻訳ボタンを押すと数秒で全言語の翻訳結果が同時に表示されます。言語数が多いので適宜スクロールして見るようになります。

### 携帯電話向け多言語翻訳サービス 『超多言語翻訳』機能説明

**日本語から17言語に翻訳!**

1. 翻訳したい文を入力      2. 翻訳ボタンを押下      3. 翻訳結果表示!! (実際は一画面に表示されます)

**訳文を逆翻訳でCHECK!**

1. 翻訳したい文を入力      2. 翻訳方向言語を選び、翻訳ボタンを押下      3. 翻訳結果表示!!      4. 続けて他の言語にも翻訳可能!!

図2 利用の流れ

- ② TOP画面から、「訳文を逆翻訳でCHECK！」ボタンを押すと下段に示した入力画面になる。ここで、旅行場面の文を入力して、翻訳先の言語を選択して翻訳ボタンを押すと1秒程度で、訳文と逆翻訳の結果が表示されます。原文と逆翻訳の結果が（必ずしも一致しなくてもよい）ほぼ同義な場合に、訳文はほぼ正しいことがわかっていますので、安心して相手に訳文を提示できます。

### 3 コーパスベース翻訳技術の仕組みと利点

コーパスベース翻訳技術とは対訳コーパス（原文と訳文の対を大量に集めたもの）から、翻訳システムの知識（確率付き対訳辞書）を自動的に構築する技術です（図3）。

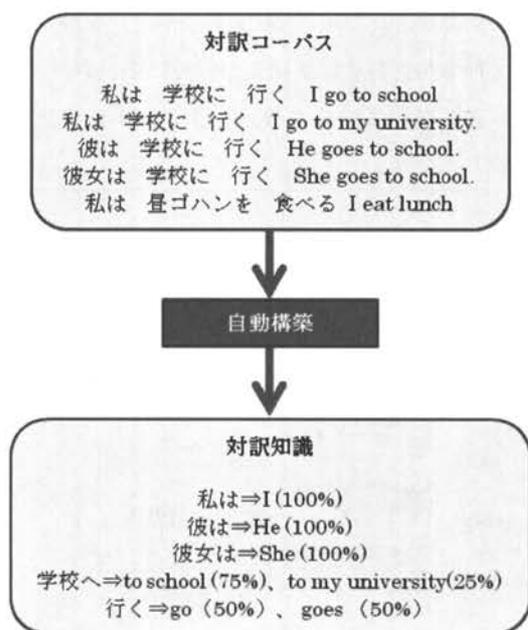


図3 コーパスベース翻訳の仕組み

この手法には2つの利点があります。

- ① N個の言語の対訳コーパスを用意すれば、全ての組合せである  $N * (N-1)$  個の翻訳システムが自動的に構築できること。上述の18言語の場合、 $18 * 17$  で306通りの翻訳システムが出来たわけです。
- ② 新しい分野の翻訳システムを作るには、その分

野のコーパスを集めれば済むこと。例えば、旅行会話の対訳コーパスを集めれば、旅行会話用の翻訳システムが、新聞の対訳データを集めれば、新聞用の翻訳システムが構築できます。今後、様々な分野の高精度の翻訳システムをご提供できると考えています。

### 4 携帯電話での多言語表示

携帯電話は、通常メモリーが小さく、多言語のフォントを内蔵していません。せっかく翻訳できても、表示できなくては意味がありません。この対策として、PCのサーバ側で、画像化して、これを携帯電話で取得して表示しています（図4）。

### 5 今後の展望

A TRでは音声翻訳技術の世界展開すべく、まずはアジアでA TRを中心にした A-STAR<sup>1</sup>コンソーシアム（メンバーは中国、台湾、韓国、タイ、インドネシア、インド、ベトナム、シンガポールの代表的研究機関）において、アジア言語間の音声翻訳実験を実施する予定としており、アジア言語における標準化活動も精力的に進めています。

これらの成果も順次事業化していく予定です。

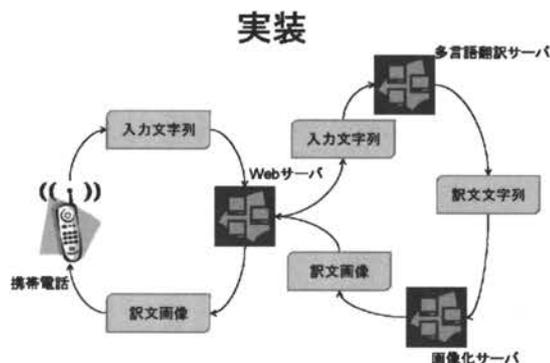


図4 多言語表示の仕組み

<sup>1</sup> Asian Speech Translation Advanced Research の略で、2006年の設立されたアジアにおける音声翻訳の共同研究のためのコンソーシアム (<http://www.slcr.atr.jp/ASStar/>)。

# 「The 翻訳® 2008 プレミアム」

東芝ソリューション株式会社

## 1. はじめに

英日/日英翻訳ソフト「The 翻訳 2008 シリーズ」の、プロフェッショナル向けの上位版「The 翻訳 2008 プレミアム」では、新開発の「ダブル知識翻訳」を翻訳エンジンに搭載し、より自然で読みやすい訳文を実現しました。さらに、翻訳メモリ機能を使用して翻訳作業効率を高める「用例ベース翻訳」、文書の文脈情報 (Context) や書式情報 (Format) を利用する CF エンジンをメール翻訳に適用して精度を向上させる「メール CF 翻訳」を搭載しています。

## 2. The 翻訳 2008 プレミアムの主な特長

### (1) ダブル知識翻訳

翻訳ソフトは原文を解析し訳文を生成する翻訳エンジンが翻訳精度の高さを決めます。従来の The 翻訳シリーズに採用していた意味トランスファ方式では、入力した原文を解析するために知識を活用することで高い翻訳精度を実現してきました。

The 翻訳 2008 プレミアムでは、「原文を解析するための知識」と「訳文を生成するための知識」の2つの知識を使って翻訳精度をさらに高める「ダブル知識翻訳」技術を開発しました。これによって、より自然で読みやすい訳文を実現しています。例えば、「Clicking on one of the areas will take you

to a detailed description of this function.」という原文でも、英語の主語の動名詞句を「～すれば」と訳し、他動詞表現「take you～」を日本語では自動詞表現で訳出することによって、「そのエリアのうちの1つをクリックすれば、この機能の詳細な記述に移動するでしょう。」と自然な表現で訳すことができます。

### (2) 用例ベース翻訳

従来の翻訳メモリ機能を用いた翻訳では、翻訳メモリ辞書に、翻訳する文と全く同一の用例が登録されていなくても、一部が異なる用例 (類似用例) が登録されていれば、その類似用例の訳文を参照し、異なる部分を編集して完成文を作成することができます。ただし、翻訳者は類似用例の訳文中の編集が必要な語句を自分で見つける必要がありました。

The 翻訳 2008 プレミアムでは、こうした編集作業をより効率化するために、類似用例の訳文中の編集が必要な語句とその置き換え候補 (原文の異なる部分の訳語候補) をわかりやすく提示し、訳文の完成を支援する「用例ベース翻訳」を搭載しました。

The 翻訳では、従来より、翻訳後に翻訳する原文中の、類似用例の原文と異なる語句が色づけ表示されます。用例ベース翻訳では、翻訳後に色づけされた原文中の語句をダブルクリックすると、訳文中の

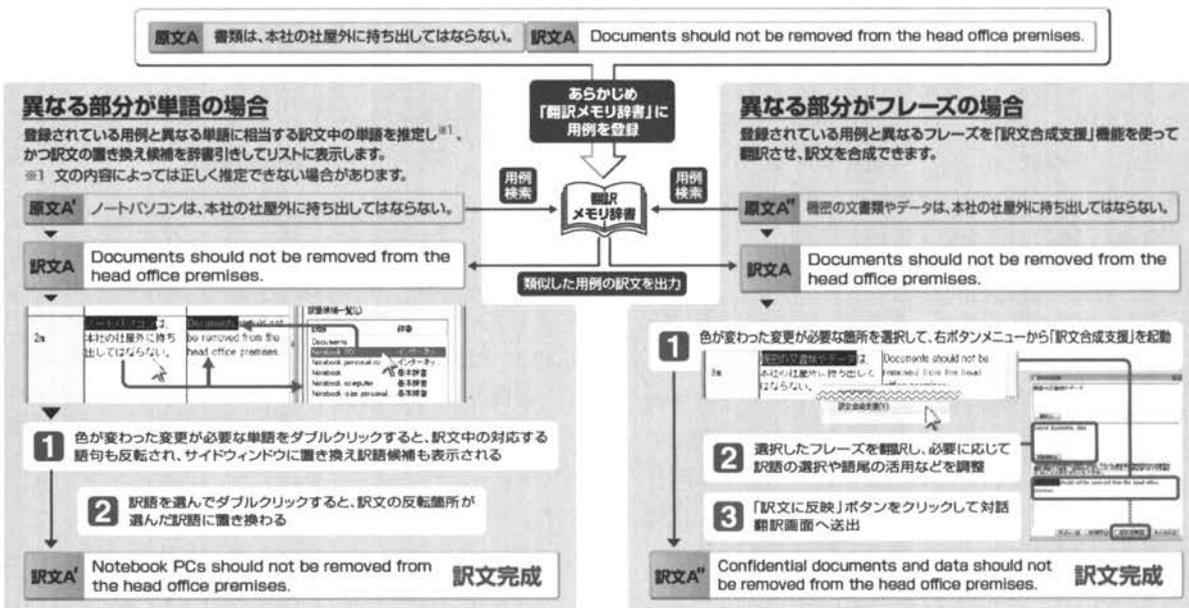


図 1 用例ベース翻訳の例

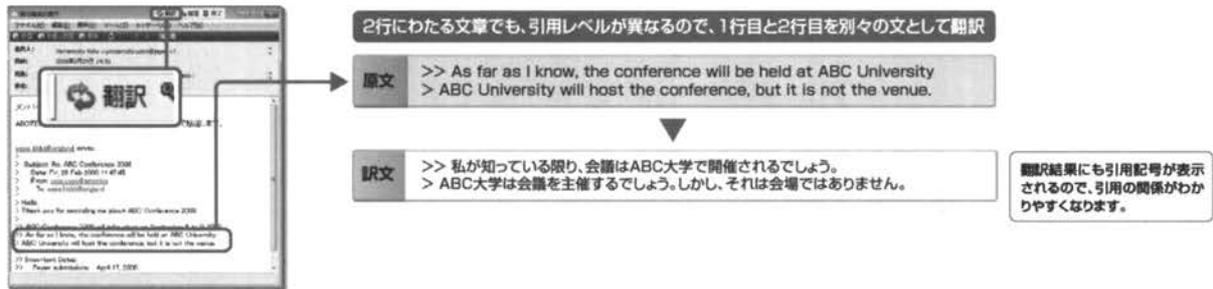


図 2 メールCF 翻訳の例

対応箇所が推定され反転表示されます<sup>1</sup>。同時に、その反転表示された語句に置き換えるべき訳語の候補が表示されるので、候補から希望の訳語を選択して置き換えることができます(図1の「異なる部分が単語の場合」)。従来の翻訳メモリ機能では必要であった、変更が必要な箇所を翻訳者が自分で見つける作業を不要とするため、作業効率を大幅に向上できます。

文中の他の語句との関係によっては、翻訳する原文と類似用例の原文の異なる部分のみでなく、その前後を含んだ部分に対応する訳文の編集が必要な場合があります。この場合、上述の方法では訳文が完成されません。これは、翻訳する原文と類似用例の原文の異なる部分が複数の語句からなるフレーズの場合によく見られます。こうした状況に対応するため、原文中の編集したい部分を指定して部分翻訳させ、用例の訳文と簡単に合成できるインターフェースも用意しています(図1の「異なる部分がフレーズの場合」)。

### (3) メールCF 翻訳

「メールCF 翻訳」は、電子メールの文脈情報(Context)と書式情報(Format)から、メールの行頭に多く用いられる引用記号(「>」や「|」など)を正しく解釈し、引用記号付きで訳出する機能です。従来商品では難しかった引用部分の前後の文のつながりを判断できるため、元のメールの内容を引用して返信したメールを翻訳する場合に有効です。メールCF 翻訳の例を図2に示します。

### 3. その他の機能

The 翻訳 2008 プレミアムは上記に加え、以下のような多彩な機能を揃えています。

①対話翻訳(高機能翻訳エディタ): 様々な形式の文書を読み込み、原文と訳文を比べながら翻訳できます。メモリ辞書の活用を強力に支援する絞り込み検索・編集機能、任意の文字列や数字を含む表現を\*や#を使って検索できる「ワイルドカード検索」機能、一文ごとにメモや付箋が付けられ

るメモ機能など様々な補助機能があります。また、補助ウィンドウであるサイドウィンドウ、情報ウィンドウ、ボトムウィンドウは、表示/非表示、対話翻訳画面からの切り離し、配置変更が可能です(図3)。



図 3 対話翻訳の画面

②アプリケーション連携翻訳: Microsoft® Word、Microsoft® Excel、PowerPoint®、一太郎、秀丸エディタなどのアプリケーションに翻訳機能をアドインして翻訳作業を行えます。

③PDF 翻訳: PDF 文書を Adobe® Reader®でテキストを取り出して翻訳できます<sup>2</sup>。さらに、Adobe® Acrobat®があればPDF 文書のレイアウトを保ったままで翻訳でき<sup>2,3</sup>、イメージでできている文書

2 ファイルのセキュリティの設定によっては翻訳できない場合があります。  
3 Adobe® Acrobat® 9.0/8.1/8.0/7.0/6.0 が必要です。無償配布の Adobe® Reader® 9.0/8.1/8.0/7.0/6.0 では、連携テキスト翻訳のみとなります。レイアウトは保たれません。Adobe® Acrobat® 9.0/Adobe® Reader® 9.0を使用する場合は、The 翻訳 2008 プレミアム レビジョン2サービスパッチを適用する必要があります。

1 文の内容によっては正しく推定できない場合があります。

も文字認識して翻訳できます<sup>4</sup>。

- ④ ホームページ翻訳：Microsoft® Internet Explorer®、Firefox®などのアプリケーションでホームページを表示しながら翻訳できます。
- ⑤ セレクトコーパス翻訳：翻訳したい分野の文書を登録することで、その分野に適した訳語を自動学習することができます。
- ⑥ 豊富な辞書構築・編集機能：辞書登録候補の抽出、登録用属性の自動付与、辞書編集エディタなど、ユーザ用辞書を構築・編集するためのツールが整備されています。
- ⑦ 高機能 OCR ソフト「Express Reader Pro LE」を搭載：スキャナで読み取った紙文書のテキスト認識結果はテキスト、HTML、RTF の各形式で保存できます。また、直接対話翻訳に取り込み、すぐに翻訳できます。

#### 4. 商品ラインアップ

The 翻訳 2008 プレミアムシリーズは、以下のラインアップとなっています。

- The 翻訳 2008 プレミアム
  - ・搭載辞書語数：477 万語（専門用語辞書 11 種類を含む）
  - ・搭載翻訳メモリ用例数：15 万例
- The 翻訳 2008 プレミアム 専門辞書パック
  - ・搭載辞書語数：736 万語（専門用語辞書 28 種類を含む）
  - ・搭載翻訳メモリ用例数：55 万例
- The 翻訳 2008 プレミアム 特許エディション
  - ・特許文書向け翻訳機能を搭載した最上位商品
  - ・搭載辞書語数：757 万語（専門用語辞書 28 種類および特許辞書を含む）
  - ・搭載翻訳メモリ用例数：57 万例

また、ビジネスユース向けに、ビジネスで利用度の高い機能を選びすぐって構成した商品を用意しています。

- The 翻訳 2008 ビジネス
  - ・搭載辞書語数：477 万語（専門用語辞書 11 種類を含む）
  - ・搭載翻訳メモリ用例数：15 万例

その他、以下のオプション商品があります。

- The 翻訳 2008 英日専門用語辞書
- The 翻訳 2008 日英専門用語辞書

<sup>4</sup> 特殊文字、特殊フォントが使用されている場合や、イメージ部分の元原稿の状態によっては、正しく認識されない場合があります。

#### 5. 動作環境

対応 OS (すべて日本語版)
Windows Vista® Home Basic(32 ビット版) / SP1 Windows Vista® Home Premium(32 ビット版) / SP1 Windows Vista® Business(32 ビット版) / SP1 Windows Vista® Ultimate(32 ビット版) / SP1 Windows XP® Home Edition/Professional SP2 / SP3 Windows® 2000 Professional SP4

対応機種	PC/AT互換機
CPU	Pentium®以上/Celeron®以上推奨
メモリ	<ul style="list-style-type: none"> <li>■ The 翻訳 必要分：128MB以上 (256MB以上を推奨)</li> <li>■ OS使用分を含む推奨値 Windows Vista®：768MB以上 Windows XP®：384MB以上 Windows® 2000：320MB以上</li> </ul>
ディスク	<ul style="list-style-type: none"> <li>■ The 翻訳 2008 プレミアム：940MB</li> <li>■ The 翻訳 2008 プレミアム 専門辞書パック：1.8GB</li> <li>■ The 翻訳 2008 プレミアム 特許エディション：1.9GB</li> </ul> <p>各商品では、下記を必要に応じてインストールできます。</p> <ul style="list-style-type: none"> <li>◎ OCRソフト (Express Reader PRO LE)：60MB</li> <li>◎ 研究社 リーダーズ英和辞典 (第2版)：110MB</li> <li>◎ 研究社 新英和中辞典 (第7版) / 新和英中辞典 (第5版)：550MB</li> </ul> <ul style="list-style-type: none"> <li>■ The 翻訳 2008 ビジネス：940MB</li> </ul> <p>下記を必要に応じてインストールできます。</p> <ul style="list-style-type: none"> <li>◎ 研究社 新英和中辞典 (第7版) / 新和英中辞典 (第5版)：550MB</li> </ul>

#### ■ 製品情報 URL

<http://hon-yaku.toshiba-sol.co.jp/>

- Microsoft、Windows、Windows XP、Windows Vista、PowerPoint、Internet Explorerは、米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。
- Adobe、Acrobat、Adobe Reader、Adobe Acrobatは、Adobe Systems Incorporatedの登録商標または商標です。
- The 翻訳、CFエンジン、セレクトコーパス翻訳は東芝ソリューション株式会社の登録商標です。
- 記載されている商品の名称はそれぞれ各社が商標として使用している場合があります。
- 本製品の仕様は予告なく変更される場合があります。

# 英日・日英翻訳ソフト ATLAS V14

富士通株式会社

## 1. はじめに



ATLAS V14

「ATLAS V14」は、業界最大級の286万語の辞書と長年の技術で培った優れた翻訳エンジンで、プロ仕様の高精度な翻訳結果を導き出す英日・日英翻訳ソフトウェアです。

Microsoft Office、Acrobat 連携などの業務利用からメールやホームページでの翻訳まで幅広い翻訳シーンで活用できます。

## 2. 機能

### 対訳エディタ

対訳エディタとは、原文と訳文を左右に表示し、編集作業を行いながら翻訳を可能とする機能です。編集と翻訳を同一画面で行うことができ、効率の良い作業を実現できます。

番号	原文	訳文
1	次世代コンピューティングとは	What is the next generation computing?
2	気象予報やDNA解析といった膨大な計算を必要とする分野ではスーパーコンピュータの利用が欠かせません。	The use of the super computer doesn't lack it in the field where a huge calculation like the meteorological prediction and DNA analysis, etc. is needed.
3	私達は、1976年から次世代スーパーコンピュータの開発に取り組んでいます。	We have been working on the development of the next-generation supercomputer since 1976.
4	「ペタ」は単位の接頭語で、10の15乗を表します。	"Peta" shows the 15th power of ten by the prefix of the unit.
5	1秒間に計算できる浮動小数点演算の速さを「フロップス」と呼びます。	The speed of a computable floating point arithmetic will be called "FLOPS" in one second.
6	ペタスケールはこのフロップスが10の15乗を超えるペタフロップスの性能であることを意味しています。	It means [petascale] is a performance of [petaflops] for this FLOPS to exceed the 15th power of ten.
7	3ペタフロップスというのは、クロック周波数が3GHzのデスクトップコンピュータ50万台分の演算性能です。	In three-peta FLOPS, the clock frequency is 500,000 desktop computer 3GHz's worth of an operation performance.
8	2000年頃、私達はベクトル型からスカラ型のスーパーコンピュータへと、アーキテクチャーを変更しました。	We changed scalar type's super computer and architecture from the vector type in around 2000.
9	スカラ型は電力や設置面積で有利です。	The scalar type is advantageous in the electric power and space required for installation.

### 連携翻訳

Microsoft Office (Word、Excel、PowerPoint) Acrobat (PDF 文書) に ATLAS の翻訳機能をアドイン。それぞれのアプリケーション上で原文ファイルのレイアウトそのままに翻訳が行なえます。Windows Vista、2007 Microsoft Office system の新しいドキュメント形式 (docx、xlsx、pptx など) やリボンインターフェースにも対応しています。

## ステップ翻訳

Word 文書の原文を先頭から翻訳。原文のレイアウトを確認しながら1文単位で訳文に置き換えることができます。



Word 文書の原文画面



ステップ翻訳画面



ステップ翻訳反映画面

## 翻訳メモリ

原文と訳文を対にした翻訳例を翻訳メモリとして登録できます。翻訳メモリに蓄えられた翻訳例文は、原文のマッチ率に応じて、青、赤、黒の色で表示され作業効率が向上します。

### 対訳エディタからの翻訳メモリ検索例

番号	原文	訳文
1	次世代コンピューティングとは	What is the next generation computing?
2	気象予報やDNA解析といった膨大な計算を必要とする分野ではスーパーコンピュータの利用が欠かせません。	[Supercomputer is indispensable for the areas in which highly calculation-intensive works are required. 50%]
3	私達は、1976年から次世代スーパーコンピュータの開発に取り組んでいます。	We have been working on the development of the next-generation supercomputer since 1976.
4	「ペタ」は単位の接頭語で、10の15乗を表します。	[Peta denotes 1 quadrillion. 73%]
5	1秒間に計算できる浮動小数点演算の速さを「フロップス」と呼びます。	Floating point operations per second indicating the speed of computer is called "FLOPS".
6	ペタスケールはこのフロップスが100の15乗を超えるペタフロップスの性能であることを意味しています。	It means [petascale] is a performance of [petaflops] the performance to exceed the 15th power of ten.
7	ペタフロップスというのは、クロック周波数が3GHzのデスクトップコンピュータ50万台分の演算性能です。	[The computer with a 3-FetaFLOPS speed is almost equivalent to 500,000 desktop computers with 3GHz-clock frequency. 92%]
8	2000年頃、私達はベクトル型からスカラ型のスーパーコンピュータへと、アーキテクチャーを変更しました。	We changed scalar type's super-computer and architecture from the vector type in around 2000.
9	スカラ型は電力や設置面積で有利です。	[The scalar computer has the advantages of lower power consumption. 96%]

対訳エディタ画面

辞書名	利用者辞書0001 / スコア	一致率
辞書名: 利用者辞書0001 / スコア: 70 / 一致率: 73%	「ペタ」は単位の接頭語で、10の15乗を表します。 Peta denotes 1 quadrillion.	73%
辞書名: 利用者辞書0001 / スコア: 70 / 一致率: 73%	「ペタ」は単位の接頭語で、10の15乗を表します。 Peta is 10 of the 15 power of ten.	73%
辞書名: 利用者辞書0001 / スコア: 41 / 一致率: 58%	「ペタ」は単位の接頭語で、10の15乗を表します。 Peta is a prefix that denotes one quadrillion.	58%
辞書名: 利用者辞書0001 / スコア: 8 / 一致率: 31%	「ペタ」は単位の接頭語で、10の15乗を表します。 新製品は、2008年提供を目指しています。 We are aiming to bring our new product in 2008.	31%
辞書名: 利用者辞書0001 / スコア: 8 / 一致率: 19%	「ペタ」は単位の接頭語で、10の15乗を表します。 ペタフロップスは、クロック周波数が3GHzのデスクトップコンピュータ50万台分の演算性能です。	19%

翻訳メモリ検索画面

## 文アライメント支援ツール

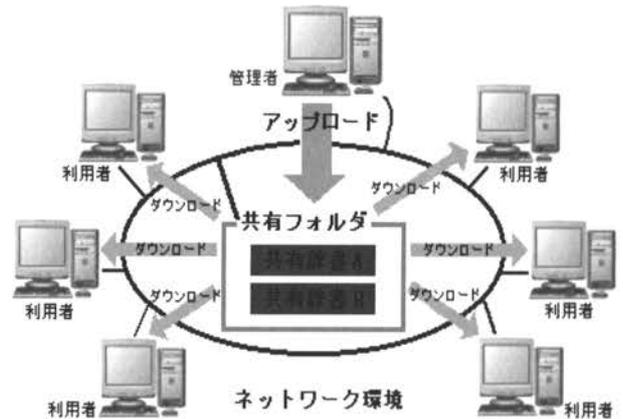
過去に翻訳した原文と訳文の対応を取る「文アライメント支援ツール」を提供しています。

原文	訳文
このたびは、Math M1 Suggestion V1.0 をお買い上げいただきまして、誠にありがとうございます。	Thank you for purchasing "Math M1 Suggestion V1.0".
本製品は、基本分析、行動分析、キーワード分析など、さまざまな機能を提供しています。これからの機能を最大限に活用していただくことにより、効果的な分析が可能です。	This product is equipped with various functions such as Basic Analysis, Activity Analysis and Keyword Analysis. By using these functions according to your needs, efficient analysis is possible.
本書がみなさまのお役に立つことを願っております。	We hope that this manual is of help to all users.
2007年9月	Sep-2007
本書の内容と使いかた	Contents and Use of This Manual
本書の構成	Organization of The Manual
本書は、本製品をこれからインストールする方が初めてのお使いになる方を対象に記述しています。本書は「初級」「中級」「上級」「応用編」「付録」という構成になっています。	This manual is intended for users installing the product, and using the product for the first time. The manual is composed of "Getting Started", "Basic", "Advanced Technique" and "Appendix".
お読みになる前	Getting Started
本書では本製品の概要とインストール方法について説明しています。	This chapter explains the overview of and method for installing this product.
基本編	Basic
本書では機能の概要と操作方法について説明しています。	This chapter explains preliminary operations of each function.
応用編	Advanced Technique
本書では本製品をより効果的に使うためのテクニックについて説明しています。	This chapter explains operators for more effective use of this product.

文アライメント支援ツール画面

## 共有辞書ツール

ネットワーク上に構築された共有フォルダ内に共有辞書をアップロードすることにより、グループ内の管理者や利用者が同一の辞書を利用できます。この機能を利用することで、翻訳品質の均質化を実現するだけでなく、知的財産を利用者全員で共有することが可能です。



辞書共有の概念図

## ファイル翻訳

ファイル名を指定して、複数のファイルをまとめて翻訳できます。比較的量が多い文書をまとめて翻訳する場合に有効です。レイアウトも保持されます。

原文ファイル	ファイルの種類	方向	出力形式	環境
G:\Users\Admin\Documents\ビジネス検討会議\ビジネス計画_2007下期.doc	Microsoft Off.	自動	訳文のみ	一般
G:\Users\Admin\Documents\ビジネス検討会議\次世代スーパーコンピュータ.doc	Microsoft Off.	自動	訳文のみ	一般
G:\Users\Admin\Documents\ビジネス検討会議\中期ビジネス計画.xlsx	Microsoft Off.	自動	訳文のみ	一般
G:\Users\Admin\Documents\ビジネス検討会議\売上分析_北米.xlsx	Microsoft Off.	自動	訳文のみ	一般
G:\Users\Admin\Documents\ビジネス検討会議\2007 Annual Report.pptx	Microsoft Off.	自動	訳文のみ	一般

ファイル翻訳画面

## インターネットアップデート

旅行、スポーツ、音楽、時事用語など様々な分野の辞書（エンターテインメント辞書）を、無料でダウンロードできます。

## その他のツール

その他、メール・ホームページ・ワープロなどあらゆるアプリケーション上でワンクリックで翻訳するマウス翻訳、入力と同時に翻訳するキータイプ翻訳、レター作成支援ツールの「おまかせ記文」などの文書作成に便利なツールを取り揃えています。

## 専門用語辞書

専門分野向けに 28 分野、557 万語（英日 281 万語・日英 276 万語）の専門用語辞書をオプションで用意しています。4 分野 73 万 5 千対の翻訳メモリ用例文も収録しています。

### 専門用語辞書一覧

分野	英日	日英
1. 情報処理	174,000 語	175,000 語
2. 電気・電子	125,000 語	127,000 語
3. 物理・原子力	176,000 語	178,000 語
4. 機械	115,000 語	117,000 語
5. 工業化学	226,000 語	220,000 語
6. プラント	63,000 語	64,000 語
7. 土木・建築	90,000 語	91,000 語
8. 金属	46,000 語	47,000 語
9. 地学・天文	81,000 語	83,000 語
10. 輸送	63,000 語	64,000 語
11. 自動車	43,000 語	45,000 語
12. 軍事	26,000 語	28,000 語
13. 農林水産	70,000 語	71,000 語
14. 生物	162,000 語	164,000 語
15. 医学（生化学）	205,000 語	207,000 語
16. 医学（薬学）	83,000 語	84,000 語
17. 医学（解剖学）	57,000 語	59,000 語
18. 医学（疾患・症状）	216,000 語	218,000 語
19. 医学（精神医学）	32,000 語	33,000 語
20. 医学（医療機器）	52,000 語	54,000 語
21. 金融・経済	65,000 語	66,000 語
22. 法律	18,000 語	19,000 語
23. ビジネス	149,000 語	151,000 語
24. 人名・地名	209,000 語	211,000 語
25. 環境	35,000 語	36,000 語
26. ダイアログメッセージ集	76,000 語	76,000 語
27. バイオ・生化学	31,000 語	32,000 語
28. ビジネス技術実用大辞典	120,000 語	39,000 語

## 医学・薬学用語を収録した辞書パック



ATLAS 医学翻訳ステッドマン  
＋  
南山堂パック V14

「ステッドマン医学大辞典改訂第 6 版」と南山堂「医学英和辞典 第 12 版」の 2 大辞書を収録。医学・薬学分野の翻訳を強力にサポート。辞書

916 万語、翻訳例文（翻訳メモリ辞書）84 万文を搭載しています。さらに、2 大辞書を電子辞書としても収録しているので、翻訳だけでなく医学辞書としても活用できます。

## 3. 動作環境

動作 OS :

日本語 Microsoft Windows Vista/Microsoft Windows XP SP2 以降/Microsoft Windows 2000 SP4 以降

必要メモリ :

256MB 以上

（翻訳メモリ使用時は、512MB 以上推奨）

Windows Vista の場合 512MB 以上

（翻訳メモリ使用時は、1GB 以上推奨）

必要ディスク容量 :

ATLAS 本体（翻訳スタンダード相当）600MB

## 4. 価格（税込）

ATLAS 翻訳スーパーパック V14 134,400 円

ATLAS 翻訳スタンダード V14 92,400 円

ATLAS 専門辞書（技術・ビジネス・医学）50,400 円

ATLAS 医学翻訳ステッドマン+南山堂パック V14  
186,900 円

「ATLAS」の旧版ユーザー様向けには、バージョンアップ製品を販売しています。詳しくは下記の URL を参照ください。

<http://software.fujitsu.com/jp/atlas/versionup.html>

### ■お問い合わせ先

富士通株式会社

プロダクトマーケティング本部

プロダクトマーケティング統括部 ミドルウェア部

〒105-7123 東京都港区東新橋 1-5-2

（汐留シティセンター）

TEL : 03-6252-2661

受付時間： 9 : 00～12 : 00/13:00～17:00

（土日祝日を除く）

<http://software.fujitsu.com/jp/atlas>

技術的なお問い合わせ

ATLAS サポートセンター

TEL : 03-5462-1934 FAX : 03-5462-2344

受付時間： 9:00～12:00/13:00～17:00

（土日祝日除く）

E-mail : [contact-atlas@cs.jp.fujitsu.com](mailto:contact-atlas@cs.jp.fujitsu.com)

## AAMT 機械翻訳システム評価テストセットの概要

機械翻訳課題調査委員会ワーキンググループ 1

### 0. はじめに

現在、日本語・中国語を対象とした機械翻訳システムがインターネット上を始め、様々な形式で利用されるようになった。今後も日本語・中国語翻訳システムの利用者が増加の傾向にあるため、翻訳システムの一層の改良が必要と思われる。このような機械翻訳システムの中国語への関心を背景として、アジア太平洋機械翻訳協会の機械翻訳課題調査委員会（以下、AAMT 調査委員会）は、現状の日本語・中国語の翻訳システムによる翻訳文の達成度を評価し、また、現状の翻訳システムにおける問題点を調査することを目的として活動を推進してきた。本稿は、AAMT 調査委員会の機械翻訳評価活動の基盤である AAMT 機械翻訳システム評価テストセット（以下、AAMT テストセット）を紹介する。

### 1. AAMT 機械翻訳システム評価テストセットのねらい

AAMT 機械翻訳システム評価テストセットは、機械翻訳システムの技術的な問題点を検出することを目的とする。

これまでの機械翻訳システム研究では、機械翻訳システムが達成すべきと考えられる様々な技術的課題が指摘されてきた。AAMT テストセットを用いた評価法では、機械翻訳システムをこれらの技術的課題の達成度に基づいて評価を行う。そのため各テスト文には翻訳システムが技術的に解決しておくべきと想定される言語形式と、評価者向けの設問が含まれている。評価者は、あるテスト文に対する翻訳結果に基づいて設問に回答（Yes/No）するだけでよい。この回答結果により、機械翻訳システムの評価、あるいは技術的問題点の検出が可能となる。

一般的な機械翻訳システムの評価法として、翻訳の流暢さおよび適切さという観点からの主観評価が挙げられる。主観評価では評価者が主観的に翻訳結果に任意の評点（五段階）を与える。一方、AAMT テストセットの場合、評価者は設問に対して回答するだけでよい。そのため、主観評価で問題とされる評価の揺れを抑えることができると考えられる。

また、AAMT テストセットを用いると、テスト文に付与された設問に Yes/No で回答するだけで、翻訳が困難と考えられる言語現象が適切に翻訳されているかどうか判定できる。そのため、評価者が翻訳における言語的問題点や機械翻訳システムの技術的問題を把握している必要はない。そのため、複数の評価者間における言語現象に対する知識や気づきの差といった問題が回避できる。

### 2. AAMT 機械翻訳システム評価テストセットの構成

本稿で紹介する AAMT テストセットは、日中機械翻訳を評価対象とするものである。この AAMT テストセットは、日中機械翻訳システムが処理すべき言語現象を含む日本語テスト文約 370 文とその中国語模範訳例、および評価観点となる設問からなる。評価項目は表 1 の通りである。言語現象として、(1) 品詞、(2) 文の部分構造、(3) 文構造、(4) 構文、(5) コロケーションを扱い、さらに各現象の低位項目として、数種類の項目が設定されている。

そして、その各項目に対して、テスト文や設問が付与されている。実際の項目の書式は、表 2 が示すように、【番号】、【例文】、【訳文】、【質問】、【訳出例】、【関連文】、【参照項目】、【解説】から構成される。

部分名称	項目名称
(1) 述部	(1-1) 述部の訳し分け
	(1-2) 断定文
	(1-3) 体言述語
	(1-4) 複合述部
	(1-5) 訳が一用言となる並列用言
	(1-6) 用言の副詞（句）化
	(1-7) 補助動詞
	(1-8) 基本動詞の訳し分け
(2) 名詞	(2-1) 名詞の訳し分け
	(2-2) 複合名詞
	(2-3) 「名詞1の名詞2」という構造を持つ名詞句の処理
	(2-4) 「名詞1の名詞2の名詞3」という構造を持つ名詞句の処理
	(2-5) 並列構造を持つ名詞句の処理
	(2-6) 疑問表現の名詞節の処理
	(2-7) 用言性名詞（サ変名詞）
	(2-8) 英語における数の扱い
	(2-9) 固有名詞表現
	(2-10) 形式名詞
	(2-11) 関係を示す名詞
(3) 副詞	(3-1) 副詞のタイプ
	(3-2) 副詞句
	(3-3) 擬音語・擬態語
(4) 連体修飾語句	(4-1) 非活用連体修飾
	(4-2) 用言性連体詞
	(4-3) 格助詞相当句
	(4-4) 埋め込み文修飾
(5) 助詞	(5-1) 助詞の訳し分け
	(5-2) 深層格の認定
(6) 接辞	
(7) テンス、アスペクト、モーダル	(7-1) テンスの処理
	(7-2) アスペクトの処理
	(7-3) モーダルの処理
	(7-4) ボイスの処理

(8) 特殊構造表現	(8-1) 慣用表現の処理
	(8-2) 四字熟語
	(8-3) 呼応表現
	(8-4) 天候・気象表現
	(8-5) 無生物主語構文
	(8-6) 「はが」構文
	(8-7) 比較表現
	(8-8) 比喩表現
	(8-9) 部分否定、二重否定、倒置文
	(8-10) 敬語
	(8-11) 引用・伝聞表現
	(8-12) 例示・列挙表現

表 1：設問項目

【カテゴリー】	項目名称（述部など）
【文番号】	1～377
【日本語】	彼は、誤りを犯したことがわかった。
【中文】	他明白自己犯了个错误。 他，明白了违犯了错误的事。
【設問（日本語）】	既に起きた動作を表す「了」が使われていますか？
【設問（中国語）】	表示已经完成的动作的“了”是否使用？

表 2：項目例

### 3. AAMT 機械翻訳システム評価テストセットを用いた評価結果

実際に AAMT テストセットを用いて日中機械翻訳システムの評価実験を行った。その結果、AAMT テストセットの適正さを確認することができた。本実験では AAMT テストセットを利用した評価結果の信頼性を検証するために、同一システム、同一文に対して複数の評価者が評価を行い、その結果を比較した。AAMT テストセットの評価結果が複数の評価者間において高い相関関係を示せば、AAMT テストセットの信頼性が高いと考えられる。

次に、AAMT テストセットの評価結果の妥当性を検証するために、日中機械翻訳システムの翻訳結果と日英機械翻訳システムの翻訳結果に対する評価結果を比較した。一般的には、日英機械翻訳システムの方が日中機械翻訳システムより翻訳精度が高いと考えられる。日英機械翻訳システムの評価結果が日中機械翻訳システムの評価を上回れば、AAMT テストセットによる評価法の妥当性が確認できたと考えられる。

本実験で用いた機械翻訳システムは、ウェブ上で無償利用可能な日中機械翻訳システムと日英機械翻訳システム、それぞれ6システムである。翻訳対象となる文は AAMT テストセットにおける全テスト文 (337 文) である。評価者は日本語を第二言語とする中国語母語話者である。評価者は評価結果に対して、AAMT テストセットによる評価法と流暢さ・適切さによる主観評価 (五段階) の評価を行った。

(なお、公開用 AAMT テストセットには各テスト文の主観評価の平均点が付与されている。)

日中機械翻訳システムと日英機械翻訳システムにおける各システムの評価結果は表 3 のとおりである。表 3 において、AAMT テストセットの回答結果は Yes を 1, No を 0 に変換されている。変換した結果が「設問」として示されている。また、主観評価の結果は「流暢さ」と「適切さ」がそれぞれ 1 から 5 の評点により示されている。

システム	設問	流暢さ	適切さ
日中機械翻訳システム 1	0.43	2.71	3.02
日中機械翻訳システム 2	0.42	2.73	2.99
日中機械翻訳システム 3	0.25	2.17	2.44
日中機械翻訳システム 4	0.39	2.55	2.79
日中機械翻訳システム 5	0.48	2.87	3.06
日中機械翻訳システム 6	0.46	2.83	3.11
日英機械翻訳システム 1	0.69	3.40	3.71
日英機械翻訳システム 2	0.80	3.60	4.01
日英機械翻訳システム 3	0.70	3.20	3.65
日英機械翻訳システム 4	0.85	3.67	4.1
日英機械翻訳システム 5	0.68	3.14	3.63
日英機械翻訳システム 6	0.56	2.70	3.3

表 3 : 機械翻訳システムの評価結果

まず、AAMT テストセットの評価結果の信頼性を検証するために、各翻訳システムにおける複数の評価結果について相関分析を行った。AAMT テストセットの設問による評価結果は表 4 のとおりである。各システムの評価結果 1 と 2 における結果の分布は図 1 である。この評価結果に対して相関分析を行った結果、統計的に有意な相関係数が得られた ( $r=0.98$ ,  $p<.01$ )。

この結果から、機械翻訳システム評価において AAMT テストセットによる評価法は信頼性があると考えられる。評価結果に強い相関関係が認められたが、評価 1 と評価 2 において評点に差がみられた。評価 1 と評価 2 の評価結果に対して t 検定を行った結果、統計的に有意な差が認められた ( $t(5)=14.70$ ,  $p<.01$ )。今後、評価結果を文レベルで詳細に分析し、評価結果の違いについて調査を行う必要があると考えられる。

システム	評価 1	評価 2
日中機械翻訳システム 1	0.33	0.43
日中機械翻訳システム 2	0.35	0.42
日中機械翻訳システム 3	0.17	0.25
日中機械翻訳システム 4	0.28	0.39
日中機械翻訳システム 5	0.39	0.48
日中機械翻訳システム 6	0.38	0.46

表 4：評価 1 と評価 2 における評価結果

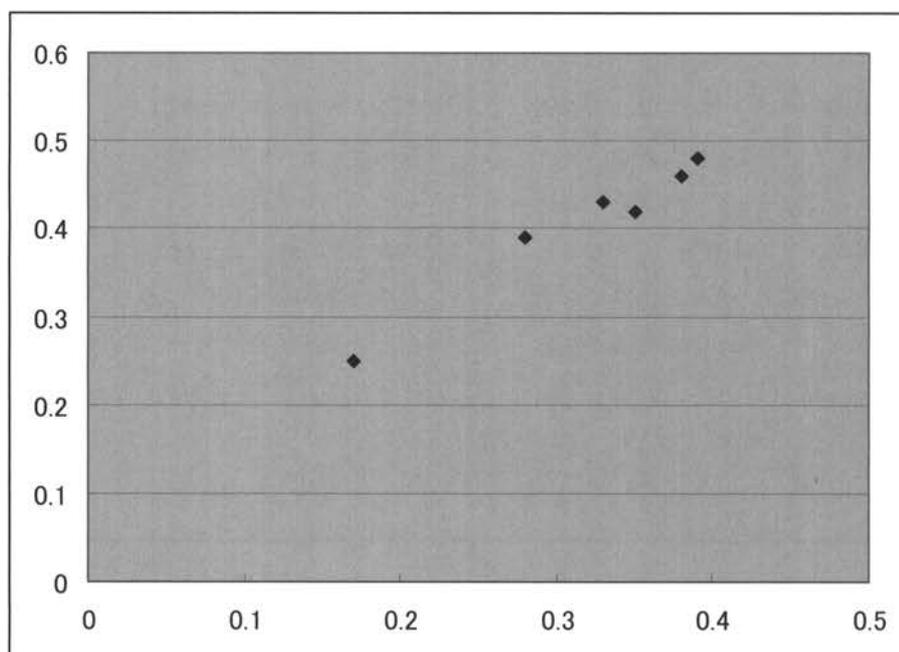


図 1：評価 1 と評価 2 における評価結果の散布図

信頼性に関して、主観評価結果に対しても相関分析を行った。その結果、「流暢さ」による評価結果 ( $r=0.96$ ,  $p<0.01$ ), 「適切さ」による評価結果 ( $r=0.95$ ,  $p<0.01$ ) のどちらの場合も統計的に有意な高い相関係数が得られた。今後、主観評価との比較を詳細に行うことで、さらに AAMT テストセットによる評価法の特徴が明らかになると思われる。

次に、AAMT テストセットの評価結果の妥当性を検証するために、日中機械翻訳システムと日英機械翻訳システムに対する評価結果を比較した。AAMT テストセットの設問による評価結果は表 5 のとおりである。表 5 において、「設問」は AAMT テストセットによる評価結果を示し、「流暢さ」と「適切さ」はそれぞれの主観評価結果を示す。

主観評価「流暢さ」と「適切さ」のいずれにおいても日英機械翻訳システムによる翻訳結果が日中機械翻訳システムによる翻訳を上回っていることが明らかになった。そして、AAMT テストセットによる評価結果においても、同様に日英機械翻訳システムの翻訳結果が日中機械翻訳システムの翻訳結果を上回っていることがわかった。

この結果から、機械翻訳システム評価において AAMT テストセットによる評価法は主観評価の場合と同程度に妥当であると考えられる。今後、評価結果から適切に技術的問題点を検出できているかを確認するために、評価結果を詳細に調査する必要があると考えられる。

システム	設問	流暢さ	適切さ
日中機械翻訳システム	0.40	2.62	2.86
日英機械翻訳システム	0.71	3.30	3.77

表5：日中機械翻訳システムと日英機械翻訳システムの評価結果

表5において、AAMT テストセットによる評価が主観評価と類似していることがわかる。そこで、AAMT テストセットによる評価結果が主観評価において「流暢さ」と「適切さ」のどちらとの類似性が高いかを判断するために、各システムにおいて AAMT テストセットの評価結果と主観評価結果（「流暢さ」と「適切さ」）に対して相関分析を行った。

その結果（表6）、いずれのシステムにおいても「適切さ」との相関係数が高いことがわかった。さらに、この「流暢さ」と「適切さ」に対する相関係数の差に対して t 検定を行った結果、統計的に有意な差が認められた ( $t(5)=3.60, p<.05$ )。したがって、AAMT テストセットの評価は主観評価における「適切さ」との類似性が高いといえる。

システム	設問対「流暢さ」	設問対「適切さ」
日中機械翻訳システム1	0.72	0.78
日中機械翻訳システム2	0.75	0.79
日中機械翻訳システム3	0.66	0.67
日中機械翻訳システム4	0.74	0.77
日中機械翻訳システム5	0.71	0.82
日中機械翻訳システム6	0.74	0.79

表6：AAMT テストセットの主観評価結果の相関係数

#### 4. おわりに

AAMT 調査委員会は、日本語・英語が対象であった JEIDA テストセットを日本語・中国語対象の評価セットへと改良を終え、AAMT テストセットを作成した。この AAMT テストセットを利用して、インターネット上で無償利用できる

日中機械翻訳システムを対象とした評価実験を行った。その結果、AAMT テストセットの信頼性と妥当性が確認された。

今後、この AAMT テストセットの改良を始め、多言語への展開など様々な取り組みを行う。

# AAMT会員のひろば

AAMT 会員の皆さまのいっそうの親密な交流の場を、AAMT Journal 誌面上で提供するべくスタートいたしました「AAMT 会員のひろば」、第三弾の今号では、学識経験者および翻訳家の 4 名の個人会員の皆さまからのご寄稿をご紹介します。

学識経験者として、研究者として、翻訳家として、それぞれの分野において第一線でご活躍の皆さまより、機械翻訳との関わり、機械翻訳および翻訳業界への熱い思いなど、数々の貴重なご意見、ご要望を寄せていただきました。

AAMT Journal では、会員の皆さまからのご寄稿を心よりお待ちしております。

ご寄稿・お問い合わせは AAMT 事務局(E-mail: aamt-info@aaamt.info)まで宜しく願いいたします。

---

## 個人会員（敬称略・50 音順）

---

### 会員名・所属

安藤 進（翻訳家）／ANDO Susumu

---

### 自己紹介

#### 【機械翻訳とのかかわり】

1982 年、富士通研究所で機械翻訳システム ATLAS の研究開発に参加したのがこの世界に入るきっかけでした。日本語を解析した結果として渡される意味構造から英語を生成するのが私に与えられた仕事でした。

具体的には、ノードからノードへとアークをたどりながら単語を出力していく生成文法のルールを作成する仕事でした。いま、この当時の資料を調べてみると、ATLAS の生成文法は 1985 年の時点で約 4,000 行でした。

ATLAS が市場化された後、富士通の関連企業である FIE で、特許庁からの委託研究として、特許抄録文の翻訳にかかわる仕事をしました。わたしの担当は、ATLAS の翻訳結果に対する評価をまとめることでした。この報告書では、特許抄録 500 件、約 2,000 文について、システムの課題と辞書の課題に分けて詳細に説明しています。

#### 【翻訳とのかかわり】

FIE では、英文マニュアルの作成に取り組みました。英文テクニカルライティングとの出会いは非常に新鮮でした。10 人ほどの英語のネイティブスピーカーのグループに入り、日々、日英間の表現の相違、文化の相違に苦闘しました。「意味を伝える」が当時の共通認識でした。

その後、翻訳会社十印では、マニュアルのほかに、特許やビジネス関連の資料などさまざまな翻訳にかかわりました。自分で翻訳するだけでなく、ほかの人が翻訳したものをチェックしたり評価したりする仕事も増えました。

その後、フリーの翻訳者として独立し、Java や JavaScript などの技術解説書の出版翻訳を 10 数冊手がけました。

### 【AAMT とのかかわり】

十印の翻訳部長として産業界のさまざまな方々と知り合うこともできました。1991 年ごろ、現在の AAMT の前身である機械翻訳協会の設立準備委員会が何回か開催されました。

その折、サイマルの翻訳部長（当時）の早良哲夫氏が「日本機械翻訳協会」より「アジア太平洋機械翻訳協会」のほうがよいのではないかと提案されました。長尾眞先生が「なるほど」とおっしゃり「Asia-Pacific」が英語名称に採用されたことを思い出しました。

こうして、AAMT とは発足当初からの長いかかわりになります。現在も、研究会に時折参加しています。

### 【人手翻訳は正解なのか？】

わたしは最近、産業界における翻訳の正解は、多数の利用者が違和感なく受け入れてくれる翻訳ではないかと考えるようになりました。

特許など、産業翻訳の世界では、翻訳の良し悪しの判断は、審査を担当する人の数だけあるからです。判断の基準も、専門用語の訳語選択から権利関係にかかわる微妙な表現の適否にいたるまで多様です。

このような現状を踏まえると、機械翻訳の精度は、模範とされる正解との類似度だけでよいのでしょうか。

### 【「米在日大使」vs「米駐日大使」】

たとえば、先日、ある研究会で、英日翻訳の評価に使われたニュース記事の模範訳にある「米在日大使」が気になりました。これは人手翻訳によるとのことですが、機械翻訳の「米駐日大使」のほうがよいのではないかと思いました。自宅に帰り、Google でフレーズ指定にして検索すると、次のようになりました（検索日は 2009 年 9 月 9 日）。

"米駐日大使" 11,600 件

"米在日大使" 8 件

上記のヒット件数が 2 つの用語の使用頻度をある程度反映しているとする、正解として使われた人手翻訳は誤りであることとなります。だとすると、その結果得られた数字の信憑性にも疑問が残ります。

これまで、機械翻訳の評価では人間による翻訳が正解として使用され、この正解と比較して翻訳の精度が語られるのが通例でした。このやり方でいいのでしょうか。

たとえば、機械翻訳の出力結果について、インターネット上に公開されている膨大な文例を参照して、文レベルや語句レベルなどでその妥当性を自動的に検証してスコアを出すといった方法はないものかなどと考えています。

### 【ちょっとご紹介】

機械翻訳と人間翻訳とのかかわりについては、国際交流基金が発行している『をちこち（遠近）』（2008 年 6 月号）「翻訳のアナログ思考をデジタル技術が支える」（安藤進、同書 p.40）で詳しく紹介しています。

インターネットそのものを表現辞典として利用する具体的な手法については「ちょっと検索！翻訳に役立つ Google 表現検索テクニック」（丸善、2007 年 8 月発行）で詳しく紹介しています。参考まで。

---

## 会員名

石崎 俊（慶應義塾大学環境情報学部兼大学院政策・メディア研究科）

／ISHIZAKI SHUN

---

## 自己紹介

### 【自己紹介とMTとの関わり】

かなり前のことですが、大学を卒業してからは、電総研(現在の産総研)で音声情報処理、自然言語処理、その応用として機械翻訳の研究を進めていました。慶応大学の SFC というキャンパスに 1992 年に移ってからは、自然言語処理とパターン情報処理などの工学的なアプローチばかりでなく認知科学的な側面に力を入れ、2000 年代になってからは脳科学にも興味を持っています。学部の研究会(ゼミ)や修士課程、博士課程の学生は、他の研究室に比べて人数も多く、多彩な研究テーマで賑やかに研究を進めています。

機械翻訳との具体的な関わりは、1980 年から 1981 年まで米国イエール大学に国から派遣されて滞在したときに、それまでの音声情報処理から自然言語処理に研究テーマを変えて、イエール大学で盛んに研究されていたテキストの意味表現法や意味解析システムに基づいて、日本語生成システムを作ったことに始まります。帰国後には、自分で英文や和文の意味解析システムを作って機械翻訳システムとして統合し、それからしばらくして当時の国の予算で行われたアジア 5 カ国による機械翻訳プロジェクトに参加しました。それが大きなきっかけになって、AAMT の設立時から関係して、現在も理事を務めています。

---

## 機械翻訳および翻訳業界に期待すること

### 【今後の機械翻訳技術に期待すること】

私が滞在したイエール大学では、当時の世界的な潮流であった文法重視の研究方針には懐疑的で、人間が行っているやり方を参考にすべきだと主張していました。意味表現法や意味解析を重視し、また、事例推論(CBR:Case-Based Reasoning)という考え方を当時から主張していました。人間の推論法をモデル化して応用しようという考え方です。ただし、このような意味を重視する方法は、当時の工学的レベルではかなり高度であって、実用性のある応用システムの構築は難しい状況でした。

しかし、20 年以上たった現在では、形態素解析や統語解析技術が進み、実用性のある意味解析システムの構築も夢ではなくなっていると思います。さらに、意味解析技術が段階的に進歩するたびに、様々な応用システムが開発されて幅広く使われていくと思います。Web の普及に伴って、自然言語処理の適用範囲は機械翻訳ばかりでなく、情報検索や情報抽出などを中心に格段と広がっているからです。

意味を扱う場合は、従来の工学的な手法の延長線上で高度な意味解析が実現できるかどうかは不明のように思います。よく知られているように、格関係には表層格と深層格があり、表層格を扱う場合は形態素解析や統語解析と同様に効率的な処理が実現すると思いますが、深層格の場合は、どのような深層格を使用するかは別にして固有の難しさが内在しています。もちろん、人間並みの意味理解機能の実現は未来の課題として良いと思いますが、意味の曖昧性などを扱う場合は、必然的に文脈が関係し背景的な知識やいわゆる常識も関係してきます。従って、知識工学やオントロジーなども研究のスコープに入るかもしれません。それらを使用する実用的な機械翻訳システムの実現に期待したいと思っています。

このときに過去の研究の経緯を考えると、基礎研究と応用研究の違いはあるものの、長期的な研究と短期的な研究を使い分けて考え、どちらが重要かというよりは、両方の研究をしっかりと進めるべきという考え方に行き着くと思います。

私自身が期待しているのは、1990年代後半から急速に発展してきた脳科学です。さまざまな脳機能計測装置が開発・実用化され、言語に関する脳科学も発展してきています。私の研究室でも数年前から光を使ったNIRSという脳機能計測装置を使用して研究を進めています。まだ難しい問題が多くありますが、人間の脳における言語情報処理プロセスに関するヒントがリアルタイム計測で得られないかと博士課程や修士課程の学生を中心に研究を進めています。

今後のMT研究の発展を若い研究者に期待しています。さまざまなアプローチから大きな成果が得られると思います。

---

## 会員名

江原暉将（諏訪東京理科大学 システム工学部 電子システム工学科）／EHARA Terumasa

---

## 自己紹介

私がなぜ機械翻訳の研究をしているのかをつきつめて考えると、自分の誕生に遡る。私は1944年(昭和19年)11月に現在の台湾・台北市で生まれた。私の生まれた日の新聞には「陸海の神驚 相次ぐ必殺の體當り 一相呼應してレイテ灣の敵艦船を猛攻」とある。私は戦争の恐怖におびえながら、最幼年期を過ごしたことになる。以来、いかにして争いのない世界を作り出していくかを求めて、今に至っている。機械翻訳の研究もその一環である。

言語の多様性は文化の多様性を生み[1][2]、一点集中を避けて安定を保つ一方で、人々の間での意思疎通を阻害し[3]、さまざまな軋轢を生ずる。言語の多様性を保ちつつ、かつ円滑な意思疎通を可能にする技術手段として、機械翻訳を研究開発してきた。世界の現状を見ると、人口問題・環境問題など大きな問題が山積している。これらの問題を戦争によらず解決するには、話し合い(コミュニケーション)によるほかない。より良いコミュニケーション環境を実現するために、今後とも機械翻訳をはじめとする自然言語処理の研究開発に微力を尽くしたい。

[1] 金谷武洋：日本語文法の謎を解く、ちくま新書、2003。

[2] 芳賀綏：日本人らしさの構造一言語文化論講義、大修館書店、2004。

[3] ウンベルト・エーコ(上村忠男、廣石正和 訳)：完全言語の探求、平凡社、1995。

---

## AAMT へのご要望

英語や中国語、日本語といった大言語のみならず、小言語を対象にした機械翻訳システムの研究開発を支援してほしい。たとえば、以下のような活動が考えられる。(1) 留学生や日本人学生などが研究を手軽に始められるように、機械翻訳のオープンプラットフォームを提供する。(2) フェアユースの観点から、辞書やコーパスなど既存の著作物を容易に研究開発に使えるようにする。(3) 小言語の処理に関する国際シンポジウムを開催する。(4) 小言語間のフリーな機械翻訳サービスを提供する。

---

---

## 会員名

吉川潔(翻訳家)/KIKKAWA Kiyoshi

---

## 自己紹介

25年前に家庭用のパソコンが発売されました。ワープロ機能と通信機能を用いれば、地方でも東京の翻訳会社の間で原稿が送受信可能、在宅で生活費を稼ごうと横着なことを始めました。当時、翻訳会社に10人前後の翻訳者が常駐していました。そこに、安値の翻訳単価を提示して外注化を要請し、翻訳原稿を横取りしてきました。翻訳会社も翻訳者用の部屋が不要になるなど、メリットが多く、外注化が急速に広まりました。SOHO、テレワーク、在宅勤務のはしりです。今は、特許事務所から米国出願用の原稿を電子メールで受信し、英訳後に返信しています。

### 【私と機械翻訳】

昭和50年代後半、京都大学のグループが機械翻訳を開発というニュースはショックでした。SOHO翻訳が軌道にのる頃に機械翻訳も普及し、我々は追放かなという不安な気持ちでした。最初に御世話になったIBS社の桜井社長は、電子メールを試用する翻訳ネットワークを全国的な規模で構築しただけでなく機械翻訳の実用化にも熱心でした。

そこで、H社ソフトのモルモット役を命じられ実態を体験できました。一つの英文に対して約10の和訳文が出力し、ユーザが選択するという一匹狼のソフトハウスにも出会いました。同社には他の大会社も試訳依頼し、他の翻訳会社が様子見に來訪していました。多くの人が八王子市の同社を訪問しているでしょう。

---

## 機械翻訳および翻訳業界に期待すること

### 【機械翻訳の問題点】

このような経過から機械翻訳に関心を持っていました。そこで、25年の体験から体系的に選んだ860の英文を、3社の翻訳ソフトを用いて和訳し結果(137頁)を3年間にわたり発表してきました。本年は数量表現-1(40頁)に重点をおいた短文の英訳を試み、その結果を関係者に配布しました。

その試訳から、日常的な行動様式を短く表す文章はほぼ正確に訳すが、少し紛らわしい数量表現に出会うと誤訳することに気付きました。その欠陥を修正していただければ、妥協したレベルで実用化が可能と「アラ探しをしてケチを付けている」間に気付きました。

例えば、「私は数分後に病院へ行く」 --- これは誰でも英訳できます。

あるいは、「薬の量を1/2に減らせ」 --- 翻訳ソフトは正確に訳します。

ところが、「薬の量を数分の1に減らせ」- (数分の1)を正訳するソフトは皆無です。

n a t i v eによると「a fraction of ~」が(~の数分の1)に該当するそうです。

1/2, 1/3 -- のように数学的に推論しても fraction は思い浮かびません。

次に、「これは、一つ前のパックである」 ←→ This is a pack one pack earlier.

この例文を正確に英訳または和訳できる市販のソフトはありません。この2例は、重箱の隅から探した例外的な文でなく、普通に使われている表現です。

そういう文例を上記のように調査し発表しました。他にも辞書や参考書に未記載だが、時々用いられる数量表現に実務翻訳者が気付いたときに、ソフト屋さんに提示し対策していただく、すなわち、双方向で情報を交換することがベストな解決方法でしょう。

機械翻訳の研究者や技術者が、どういう考えで翻訳ソフトを開発しているかを知るために、昨年3月に催された「crest」、4月に催された「edr」に出席しました。その時に用いられていた例文の単語数は、僅か10語程度でした。

年に一度、日本翻訳連盟の有志と共にソフトメーカを訪問し意見を交換しています。ワープロソフトの文法ミス指摘機能が長文になると機能不全になると同様に、翻訳ソフトにも限度があるそうです。ところが、翻訳業界に依頼がくる原稿は100語、特許や貿易の契約書では300語以上になります。これでは、誤訳の多い結果になるわけです。

我々が指摘したミスを直すと、別のところで異なるミスが発生し、彼方立てれば此方が立たずの状況のようです。しかし、一般大衆が指摘した基本的なミスを放置し、理論だけ突っ走っても自動的に解決すると思えません。こちらの対策を優先してほしいです。

---

## AAMT へのご要望

(会員特典に関するご希望はじめ、忌憚のないご意見をご記入ください。)

### 【機械翻訳に期待すること】

私は4年前の新潟県の水害と地震の被災者です。その際に在留外国人(例えば、新潟県には出稼ぎ者を含めて1万人)の対応に支援者が困りました。被災者である翻訳者として、言語の違いによるトラブルを傍観できません。地方にも英語や韓国語や中国語を話せる人が散在し、未熟であっても在留外国人を支援しています。その支援活動に役立つツールとしての機械翻訳ソフトの登場を要請し手伝いたいです。

災害や医療や裁判や海外旅行の現場で、数量の誤訳はトラブルを生じます。その現場で用いられる会話や文章は、ほとんどが短文です。そこで、

① 実態に即した、短文(単語は10字程度、「when」や「if」や関係代名詞が含まれる複文は対象外)で正確な数量表現に特化した「翻訳ソフト」の開発を要請します。

具体的には「短文で正確な数量表現に特化した翻訳ソフト」を、NICTの言語グリッド

又は市販の電子辞書やケータイにインストールして、キー操作可能とします。

通訳マシンを購入しましたが、音声入力が騒然とした現場では十分に機能しない不安があります。音声認識度も改善しているようですが、我々は市販品でしか判断できません。

② 海外資料の検索に応用

6月の総会の前に、国際出願で高名な弁理士を訪問しました。午後に機械翻訳の総会に出席するが、要望はと尋ねたら「海外資料の検索に使えたらねー」と語っていました。

出願予定の特許に抵触する、海外の技術資料の調査に使えたらという真意でしょう。幾つかの日本語のキーワード入力を機械翻訳で英訳し、関係する海外の発明を抽出し、正確な翻訳を翻訳業界に依頼するように進んで欲しいというのは、虫の良い願望でしょうか。

③ 高齢者が活用できるように

団塊の世代が第一線を退く時代になりました。一部の人はパソコンを操作できます。海外資料（例えば、政治経済情報）をインターネットから入手して和訳し、社会的な活動に活用したいという要望があります。現状の市販ソフトは一括翻訳しても、再び微細にわたって確認する必要があります。それならば、上記のように、ディスプレイ上で短文（文節ごとに）を黒塗り（範囲指定）して和訳したほうが得策です。長文や複文への対応は、受験英語の体験があれば可能です。脳力ゲームよりボケ防止に効果的でしょう。

機械翻訳の開発に、多くの人が大変な御苦勞をかけていることは分かっています。上記のメーカ訪問の印象から、技術者が過勞死寸前の状態で頑張っています。あらためて産学協同でという噂も聞きます。関係する皆様が力を合わせて、別の角度から「短文で正確な数量表現に特化した翻訳ソフト」を開発されるように要請します。

AAMTのホームページに、会長が「開発された機械翻訳システムも使われなくては意味がありません。システムを開発する人とシステムを使う人との両輪で翻訳という知的作業が実現」と論じておられます。

真の実用化とは、ユーザが役立つと感ずることです。田舎の一介のSOHO翻訳者の出過ぎた発言かもしれません。しかし、25年前の「機械翻訳の開発」のショックなニュースとモルモット役を担ったときを思い浮かべて、私なりに手伝いたいと願っています。

---

## 第 18 回通常総会および関連行事の報告

AAMT 事務局

当協会の第 18 回通常総会が 2008 年 6 月 26 日（木）13 時 30 分より東京・御茶ノ水の総評会館で開催されました。総会后、各委員会からの報告会、講演会、そして第 3 回 AAMT 長尾賞授与式と受賞者による記念講演会が盛況のうちに行われました。

### 第 18 回通常総会

1. 開会の辞
2. 会長挨拶（独）情報通信研究機構 言語翻訳グループ 上席研究員 井佐原 均
3. 来賓挨拶
4. 出席会員の確認
5. 議事録署名人の選出
6. 議案  
第 1 号議案 2007 年度事業報告（案） 第 2 号議案 2007 年度決算報告（案）  
第 3 号議案 2008 年度事業計画（案） 第 4 号議案 2008 年度収支予算（案）  
第 5 号議案 役員改選について（案） その他
7. 閉会の辞

### 報告会

- |                       |                        |
|-----------------------|------------------------|
| 1. 機械翻訳課題調査委員会        | 委員長 内元 清貴（（独）情報通信研究機構） |
| 2. AAMT/Japio 特許翻訳研究会 | 委員 江原 暉将（諏訪東京理科大学）     |
| 3. インターネット WG         | リーダー 富士 秀（（株）富士通研究所）   |
| 4. 編集委員会              | 委員長 宇津呂 武仁（筑波大学）       |

### 講演会

講演：「言語グリッド：インターネット上の多言語サービス基盤」

村上陽平 氏（（独）情報通信研究機構 言語グリッドプロジェクト）

### 第3回 AAMT 長尾賞授与式・記念講演会

受賞者：(1) シャープ(株)情報通信事業本部

要素技術開発センター 機械翻訳チーム

受賞理由：翻訳メモリとハイブリッド翻訳方式を中核としたテキスト翻訳ならびに

音声翻訳システムなどの各種形態の機械翻訳技術の実用化

受賞者：(2) 日本電気および株式会社高電社

受賞理由：携帯電話向け多言語自動翻訳サービスの事業化

選考委員長：飯田仁（東京工科大学）

選考委員：石崎 俊（慶應義塾大学）

中川 裕志（東京大学）



▲ スピーチする長尾真国立国会図書館館長



▲ 受賞記念講演を行うシャープ佐田いち子氏



▲ 受賞記念講演を行う NEC 三浦貢氏

### 懇親会

本会後の懇親会は、多数の参加者にお集りいただき、学識経験者、研究者、翻訳家等、幅広い分野の方々が活発な意見交換を行い有意義な交流の場となりました。

# 協会活動報告

(2008年6月～2008年9月)

## 第18回通常総会

2008年6月26日

- ①第1号議案 2007年度事業報告(案)
- ②第2号議案 2007年度決算報告(案)
- ③第3号議案 2008年度事業計画(案)
- ④第4号議案 2008年度収支予算(案)
- ⑤第5号議案 役員改選について(案)
- その他・会員提案事項

## 報告会

2008年6月26日

- ①機械翻訳課題調査委員会
- ②AAMT・Japio 特許翻訳研究会
- ③編集委員会
- ④インターネットWG

## 講演会

2008年6月26日

講演：「言語グリッド：インターネット上の多言語サービス基盤」  
村上陽平 氏 ((独)情報通信研究機構 言語グリッドプロジェクト)

## 第3回 AAMT 長尾賞授与式・記念講演会

受賞者：(1) シャープ(株)情報通信事業本部 要素技術開発センター 機械翻訳チーム  
受賞理由：翻訳メモリとハイブリッド翻訳方式を中核としたテキスト翻訳ならびに  
音声翻訳システムなどの各種形態の機械翻訳技術の実用化  
受賞者：(2) 日本電気および株式会社高電社  
受賞理由：携帯電話向け多言語自動翻訳サービスの事業化

## 機械翻訳課題調査委員会

2008年6月16日(2008年度 第3回)

- ①前回委員会の議事録の確認
- ②各WGの活動について(各WGに分かれて議論)
- ③活動内容の報告(各WGから)
- ④活動内容についての議論
- ⑤まとめと次回委員会について

2008年7月14日(2008年度 第4回)

- ①前回委員会の議事録の確認
- ②各WGの活動について(各WGに分かれて議論)
- ③活動内容の報告(各WGから)
- ④活動内容についての議論
- ⑤まとめと次回委員会について

2008年9月8日(2008年度 第5回)

- ①影浦峽先生のご講演
- ②前回委員会の議事録の確認

- ③各 WG の活動について（各 WG に分かれて議論）
- ④活動内容の報告（各 WG から）
- ⑤活動内容についての議論
- ⑥まとめと次回委員会について

#### AAMT/Japio 特許翻訳研究会

2008 年 5 月 16 日

- ①前回記事録の確認
- ②今年度の研究内容について
- ③次回の開催について

2008 年 6 月 27 日

- ①前回記事録の確認
- ②今年度の研究内容について
- ③次回の開催について

2008 年 9 月 5 日

- ①前回記事録の確認
- ②今年度の研究内容について
- ③次回の開催について

#### 編集委員会

ML にて最新号の進捗打合せ及び次回委員会の開催日調整。

#### インターネット WG

- ①AAMT ホームページの更新（毎月）
- ②AAMT Forum メイリングリストの管理
- ③AAMT Forum への情報発信（月数回）
- ④委員会活動におけるメイリングリストの管理
- ⑤会員専用ホームページ開設にむけた準備
- ⑥その他事務局ネットワークのインフラ管理全般

AAMT ジャーナル編集委員会委員長  
筑波大学大学院システム情報工学研究科  
知能機能システム専攻  
宇津呂 武仁

AAMTジャーナル43号をお送りします。

8月に北京で開催されましたオリンピック、先日の日本人研究者ノーベル賞受賞と、この数ヶ月、国際的なイベントがあったことは、まだ記憶に新しいところです。特に北京オリンピックのような、多数のスポーツ関係者や報道関係者が世界中の国々から集う祭典こそ、MT技術が本領を発揮すべき機会であることは言うまでもありません。

(独)情報通信研究機構では、そのような未来の姿を現実のものとするべく、北京オリンピックの期間中に、北京市内にて携帯電話音声翻訳のモニター実証実験を行いました。今号では、そのモニター実験の様子をいち早く会員の皆様にご紹介することを狙って、紹介記事を企画しております。

また今号の巻頭言は、AAMT副会長であります、東京工科大学の飯田仁先生にご執筆頂きました。飯田先生は、まさに今回、音声翻訳モニター実証実験を担当された研究所にて、永きに渡って、企画・指導者的なお仕事をしておられたご経歴があり、ご自身の巻頭言の中でもそのことに触れておられます。

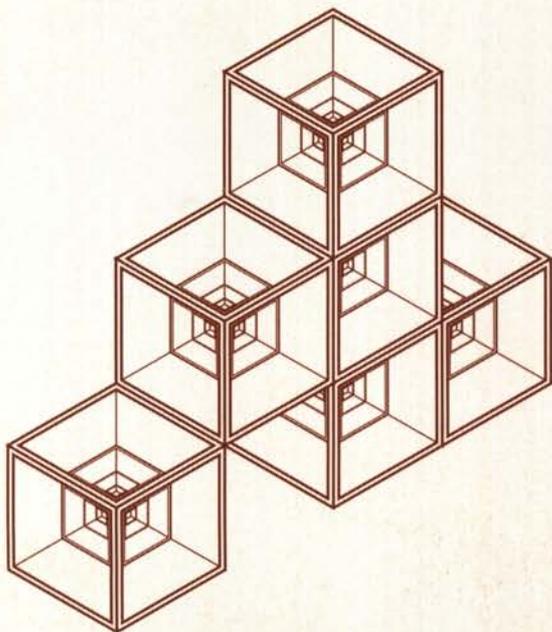
その他、今号におきましては、通常の号と同様の企画を掲載しております。機械翻訳課題調査委員会の活動成果としましては、「AAMT機械翻訳システム評価テストセットの概要」の速報をご執筆いただきました。「AAMT会員のひろば」の企画におきましては、個人会員の紹介文4件を掲載しました。

*MEMO*

*MEMO*



# AAMT



## AAMTジャーナル No.43

---

発行：アジア太平洋機械翻訳協会（AAMT）

ホームページ：<http://www.aamt.info>

住所：〒619-0289 京都府相楽郡精華町光台3丁目5番地

独立行政法人 情報通信研究機構 言語翻訳グループ内

phone：0774-93-4625 fax：0774-93-4627

編集委員会：宇津呂 武仁 後藤功雄 大倉 清司

熊野 明 三浦 貢 村上 嘉陽

事務局：神崎 享子 喜田 あゆみ

印刷所：株式会社ナビックス

---

Asia-Pacific Association for Machine Translation

c/o Computational Linguistics Group, NICT

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0289

Phone:+81-774-93-4625 FAX:+81-774-93-4627

URL:<http://www.aamt.info>