IEEE Intelligent Systems
www.computer.org/intelligent

Chinese R&D in Natural Language Technology

Chengqing Zong and Qingshi Gao

Vol. 23, No. 6
November/December 2008

Chinese R&D in Natural Language Technology

Chengqing Zong, *Chinese Academy of Sciences*

Qingshi Gao, *Beijing University of Science and Technology*

In the past 50 years, China's R&D in natural language technology has made fruitful advances but still faces many challenges. National programs continue to contribute to these achievements.

Natural language is human society's most basic, direct, and convenient communication tool for exchanging ideas and sentiments. From babies' first cries, people express their intentions using language. With the rapid spread of the Information Era, the languages people use to communicate with each other have become more and more diversified, agile, and ubiquitous.

But how does a human brain carry out the cognitive process of natural language understanding (NLU)? How do we build a computationally logical relationship between language, knowledge, and the impersonal world? How do we implement highly discriminative semantic computing? Why can't people of different races, with the same brain configuration and the same mechanism of vocal organs and ears, naturally understand different languages? So many questions puzzle us. Some experts even say that the language barrier has become one of the most important factors in blocking the globalization of human society in the 21st century. So, for us, the most challenging tasks in computer science and technology involve how to break the barrier of language differences and how to provide human-computer information exchange and interpersonal help and services that are natural, convenient, effective, and personalized.

Human language has two basic attributes: text and speech. So, generally speaking, natural language technology (NLT) includes

- using lexicons, words, sentences, text, and discourse as processing objects to parse, translate, summarize, and classify text; and
- using speech signals as processing objects to recognize and synthesize speech and to identify speakers.

Text and speech technologies are closely related but are relatively independent in both fundamental theory and implementation method. This article deals with parsing, translating, summarizing, and classifying text rather than the details of speech technologies.

In five decades of work on NLT, China has achieved much in the development of language data resources, research on fundamental theory and method, and the application of practical technologies. However, we still face many challenging problems.

Terminology

As we know, NLU has been a research topic ever since AI was proposed in 1956. From the term's connotation, NLU seems to focus more on the cognitive problems that the human brain faces in processing language, but the final goal is to implement practical natural language processing (NLP) systems oriented to performing specific tasks. Therefore, NLU is also called NLP, defined as the discipline that studies the linguistic aspects of human-human and

1541-1672/08/$25.00 © 2008 IEEE
Published by the IEEE Computer Society

human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement processes incorporating such models, identifies methodologies for iteratively refining such processes and models, and investigates techniques for evaluating the resulting systems.[1]

Whether it's called NLU or NLP, the processing object is human language, and the basic task is to solve language problems using computing technologies. So, we also call NLP *computational linguistics* (CL). Thus, the three terms NLU, NLP, and CL have different particular emphases in connotation and extension, but they address overlapping issues. Generally, for simplification, from here on, we call it natural language technology (NLT) if there's no ambiguity. Chinese information processing (CIP) specifically refers to NLT when we consider the Chinese language as the processing object.

China's NLT History

NLT research began with machine translation (MT) in the 1950s and maintained that focus for the next 30 years. In 1956, the Chinese government listed it in its plans for science development as a project named "Machine Translation, Development of Natural Language Translation Rules, and Research on the Mathematic Theory of NLP." In 1957, the Chinese Academy of Sciences' (CAS) Institute of Computing Technology (ICT), in cooperation with the Institute of Linguistics, organized a research team to study Russian-Chinese MT. In 1959, the team performed the first experiment of Russian-Chinese translation (of nine different types of complex sentences) with China's big, type-104 computer system. From the late 1950s to the mid-1960s, the Beijing Language College, Beijing Russian Language College, South China Institute of Technology, and Harbin Institute of Technology developed MT research teams. Since then, various teams throughout China have studied Russian-Chinese and English-Chinese translation.[2]

In the 10 years from 1966 to 1975, MT research almost stopped in China, as in other countries, because of the American Academy of Sciences' Automatic Language Processing Advisory Committee (http://en.wikipedia.org/wiki/ALPAC) report. The US Government established this committee of seven scientists, led by John R. Pierce, in 1964 to evaluate progress in CL, especially MT research. The 1966 report was very skeptical of MT research up to that point and emphasized the need for basic research in CL. This eventually caused the US Government to significantly reduce its funding of the topic.[3] However, the community revived that research in 1975. The Institute of Science and Technology Information of China (ISTIC) established a new MT research team with the Institute of Linguistics and ICT. These experts studied and experimented with a translation system based on the corpora of metallurgy.

Since the 1980s, MT has flourished in China. A landmark at that time was the Military Science Academy of China's Science Translator No. 1 system, which translated full text and titles from English to Chinese. Also, Zhendong Dong proposed the concept of logical semantics.[4] Meanwhile, the Gaoli Computer Company developed the Gaoli English-to-Chinese MT system in cooperation with the Institute of Linguistics and the Chinese Academy of Social Sciences (CASS).[5]

After 1989, the importance of empirical research began to be recognized around the world, and statistics were introduced into NLP study.

Qingshi Gao (coauthor of this article) and his ICT team have been engaged in MT research since 1980, from theoretical research, to system experiments, to the development of practical systems. In 1982, Gao and his colleagues proposed the preliminary theory of semantic elements.[6] Over the course of more than 27 years, Gao's theory gradually matured,[7] and he and his colleagues built more than 400,000 semantic elements and their bilingual expressions. Based on this theory and these models, Gao's PhD candidate Zhaoxiong Chen developed the intelligent English-to-Chinese MT system—the IMT/EC-863.[8] In that work, the authors proposed the Subcategory (SC) Grammar,[9] which incorporated data and operational processing and included in the rules a test function sensitive to context. Since then, Chen and his colleagues successfully developed the first pocket English-to-Chinese translation machine in the world. (www.hjtek.com/en/About/History.html)

Meanwhile in 1986, under the European Economic Community's Official Development Assistance Program, China, Indonesia, Thailand, Malaysia, and Japan investigated multilanguage translation. The Japanese government funded the project at 6 billion yen, and Japan's Center of International Cooperation for Computerization (CICC) administered it. Many Chinese universities, companies, and institutes—including ISTIC, the China National Software & Service Company, the Information R&D Center of the Ministry of Mechanics and Electronics of China, Beijing Language College, Northeast University, Tsinghua University, Nanjing University, and Renmin University of China—joined this international project.[10]

The flourishing development of MT promoted related NLT research in China. In 1987, according to the requirements of China's Key Technology Research (7·5) Program, 13 organizations (including the Beijing University of Aviation and Spaceflight, Yanshan Corp., Beijing Normal University, and the China Development Corp. of Standard Technology) started to specify the standard for Chinese word segmentation (CWS) under the leadership of Yuan Liu and Nanyuan Liang. Liwei Chen, Yun Wang, and Yongquan Liu took part in the project as advisors. In 1992, China's National Supervisory Bureau of Technology approved the Standard of Contemporary Chinese Word Segmentation for Information Processing (BG13715). The standard was put into effect on 1 May 1993,[11] and the CWS technique and many language databases developed quickly from that point on.

After 1989, the importance of empirical research began to be recognized around the world, and statistics were introduced into NLP study. IBM researchers proposed statistical MT based on the noisy-channel model and developed the Candide experimental system,[12–14] and the hidden Markov model (HMM) was successfully applied in speech recognition. The time when rule-based MT methods overwhelmingly controlled research became history.[15,16]

Table 1. Some Chinese corpora.

Year	Developer	Corpus	Number of characters
1979	Wuhan University	Contemporary Chinese literature	5,270,000
1983	Beijing University of Aviation and Spaceflight	Contemporary Chinese	20,000,000
1983	Beijing Normal University	Textbooks for junior and senior high school students	1,060,000
1983	Beijing Language College	Statistical frequency of contemporary Chinese words	1,820,000

咱们/rr 中国/ns 这么/rz 大{da4}/a 的{de5}/ud 一个/mq 多/a 民族/n 的{de5}/ud 国家/n 如果/c 不/df 团结/a ，/wd 就/d 不/df 可能/vu 发展/v 经济/n ，/wd 人民/n 生活/n 水平/n 也/d 就/d 不/df 可能/vu 得到/v 改善/vn 和{he2}/c 提高/vn 。/wj[1]

Figure 1. An example of contemporary Chinese in the ICL corpus. Each text is segmented into words, and each word is tagged as a part of speech.

Chinese researchers widely studied and experimented with statistical NLP. As in the rest of the world, this method has two basic characteristics. First, the linguistics corpus grew quickly, and its advancement further accelerated NLP development. So, there are many successful NLP systems based on statistics, such as CWS and the part-of-speech (POS) tagging system, and the Chinese character input system. Second, when speech technology was closely combined with NLP, new research directions appeared—for example, speech-based human-computer dialogue systems and speech-to-speech translation systems. The first experimental system of speech-to-speech translation (SpeechTrans) was developed in 1989.[17] After two years, an experimental system of English-to-Chinese speech-to-speech translation was developed in China.[18] The CAS's Institute of Automation (CASIA), partnering with the Consortium for Speech-to-Speech Translation Advanced Research International, has studied speech-to-speech translation since the mid-1990s. It has organized or joined in some important activities, including the International Workshop on Spoken Language Translation (IWSLT).[19] CASIA, Tsinghua University, Beijing Jiaotong University, and Harbin Institute of Technology have achieved much in human-computer dialogue research.

Meanwhile, with the rapid development of networking, NLT became a hot topic in China as in the rest of the world. New techniques, including information retrieval, information extraction, automatic text summarization, question-answering systems, and speech-document summarization have attracted attention.

After the mid-1990s, China's national power increased quickly, and the number of Chinese-speaking people worldwide increased (www.edu.cn/english_1369/index.shtml). An international array of academicians and entrepreneurs are understandably focusing on CIP. CIP thus has stepped into an unprecedented blossoming stage. Not only do the Chinese government and Chinese enterprises support increased funding, but also other developed countries and large international corporations are investing more and more in the R&D of CIP techniques.

China's NLT Achievements

In the past five decades, Chinese R&D of NLT has resulted in a series of achievements in the following three areas: building large, influential linguistic databases, advancing fundamental research on CIP, and developing practical CIP techniques.

Developing Linguistic Knowledge Bases and Corpus Bases

The corpus base and lexicon knowledge base are the basis for developing NLP systems. Since the end of the 1970s, researchers around the world, including China, have paid a lot of attention to developing corpora and linguistic knowledge bases. Table 1 shows representative Chinese corpora developed between 1979 and 1983.[20]

Peking University's Institute of Computational Linguistics (http://icl.pku.edu.cn), or ICL, has been working on developing a contemporary Chinese corpus since 1992 and has made many advancements. Tsinghua University, Shanxi University, the Harbin Institute of Technology, Beijing Language and Culture University, Northeast University, CASIA, Hong Kong City University, and the Institute of Linguistics at Academia Sinica have also contributed to the R&D of the corpus. R&D in linguistic resources for minorities' languages in China has also made significant progress. Sinkiang University, Sinkiang Normal University, Inner Mongolia University, the Northwest University for Nationalities, and the CASS Institute of Ethnology and Anthropology have made many contributions to Chinese minority language processing.

Following are some representative results.

Peking University's corpus. The ICL developed a corpus on the basis of all the articles in the *People's Daily* published in 1998. The total number of Chinese characters is 26 million. All texts are segmented into words, and each word is tagged as a part of speech according to ICL's basic specification for processing contemporary Chinese.[21] Figure 1 shows an example.

In recent years, this corpus has been the most popular one for R&D of CWS. Also, a Chinese-English bilingual corpus aligned in sentence level has been collected by ICL, which consists of about 20 million Chinese characters and 10 million English words.

HowNet. HowNet (www.keenage.com/html/e_index.html) is a Chinese extralinguistic knowledge system for computing semantic meaning. It unveils concepts' interconceptual and interattribute relationships as they are connoted in Chinese and English bilingual lexicons. Zhendong Dong and Qiang Dong developed HowNet on the basis of the assertions posited by Zhendong Dong.[22]

HowNet is characterized by the following peculiarities:[22,23]

- It's not a lexical database, thesaurus, or semantic dictionary. It focuses on concepts rather than words. HowNet can

deal with words because words are forms of content—that is, concepts.
- HowNet is computer oriented. It wouldn't be necessary to spend as much effort to build an ontology such as WordNet (http://wordnet.princeton.edu).
- Instead of simple classification and natural language definitions, HowNet defines its concepts in a formal language with nonambiguous sememes as basic units and with arguments as the relationship(s) they might represent.

These features guarantee the computing of meanings and the representation of in-depth relationships among concepts. Figure 2 presents an example record in the HowNet dictionary, and Table 2 shows the current volume of HowNet. (The data come from www.keenage.com/html/e_index.html on 31 October 2008.)

In recent years, HowNet has become a popular knowledge resource and is widely used in NLP, such as in MT, word-sense disambiguation, named-entity recognition (NER), and text classification. In October 2007, HowNet released a beta version of word sets for sentiment analysis (www.keenage.com/html/e_index.html).

In addition, *The Chinese Thesaurus*[24] and the Hierarchical Network of Concepts,[25,26] which was introduced in the article "The Outline of HNC,"[25] are also important linguistic knowledge bases and are widely used in NLP R&D in China.

Chinese LDC. The Chinese Language Data Consortium, or Chinese LDC (www.chineseldc.org), was founded in 2003 through China's National Key Fundamental Research Program (the 973 Program) and the National Hi-Tech Research and Development Program (the 863 Program). It's an academic organization within the Chinese Information Processing Society (CIPS). The Chinese LDC aims to create the most systematic, comprehensive Chinese linguistic knowledge base in the world. Now it's constructing and collecting various types of Chinese linguistic and speech resources used by the various CIP fields, including lexicons, dictionaries, thesauri, corpora, and software tools.[27]

The Chinese LDC has more than 80 types of language data, including

- the Chinese-English bilingual corpus for evaluating MTs (used by the 863 Program in 2005 and 2007),
- speech data for automatic speech recognition evaluation (used by the 863 Program in 2005),
- speech data for evaluation of text-to-speech synthesis (used by the 863 Program in 2003 and 2004),
- the general, balanced, contemporary Chinese corpus of the National Language Committee (a syntactic tree bank), and
- the six-regional-accents speech corpus of Chinese.

Figure 3a (see next page) shows the Chinese LDC data distribution, and Figure 3b gives each of five sources' contribution to the corpus.

At present, more than 70 universities, institutes, and companies have joined the Chinese LDC as members. The organization has sold more than 130 data sets and authorized some universities and institutes to use more than 40 of them for evaluation.

Table 2. HowNet's volume.

Feature	Number of instances	Feature	Number of instances
Chinese words	96,744	Chinese entries	111,470
English words	93,467	English entries	117,967
Definition count	28,925	Record count	188,069

```
NO. = 076856
W_C =  买主
G_C = N [mai3 zhu3]
E_C =
W_E = buyer
G_E = N
E_E =
DEF = {human|人: domain={commerce|商业},{buy|买: agent={~}}}
```

Figure 2. An example record in HowNet. The dictionary stores meanings and in-depth relationships among concepts.

Basic Research

In the past five decades, through the effort of scholars in the NLT area worldwide, NLP has reached three milestones:[28]

1. The theory of the complex feature and unification grammars were proposed.
2. Lexicalism (the theory that grammatical information is specified in and projected from the lexicon) was founded in linguistics study.
3. The corpus-based method and the statistical language model became widely used.

In the early years of NLP research, Yongquan Liu, Zhuo Liu, and many other experts contributed a great deal to the study of MT. Some of the methods and approaches they proposed played an important role at that time and exerted a profound influence on MT research in China.

In the 1970s, Zhiwei Feng first studied the information entropy of Chinese characters. He figured out by hand that the entropy of Chinese characters is 9.65 bits, which is close to the result of 9.71 bits that Yuan Liu worked out by computer at the end of the 1980s. Feng's work established solid ground for CIP research.

Around 1983, Feng proposed methods for parsing Chinese sentences using multibranch and multitag trees, and processing the tags using set theory.[29] This was at the same time as and identical to the concept of the complex feature set and the Unification Grammar proposed by Martin Kay. To some extent, the SC Grammar is an extension of the complex feature set and the Unification Grammar.

In the 1960s, Zhuo Liu proposed the idea of building a lexical expert system in a MT system. This was the earliest concept of lexicalism in China. The *Grammar Information Dictionary of Contemporary Chinese* developed by ICL is a typical application of lexicalism. The theory of concept association

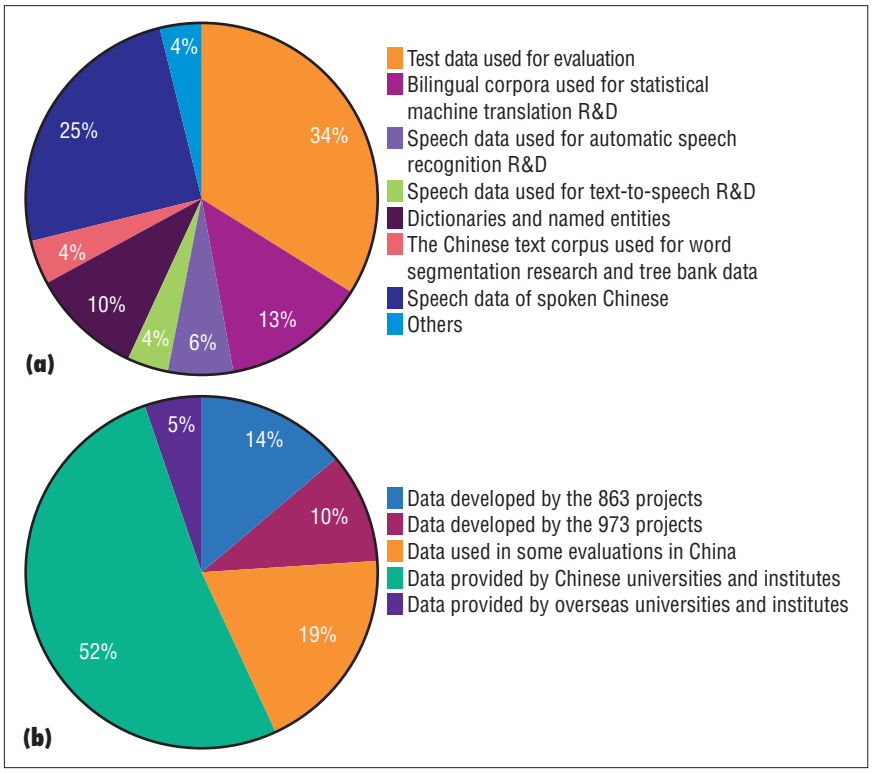

Figure 3. The Chinese Language Data Consortium's (a) data distribution and (b) data sources. The organization aims to create the world's most comprehensive Chinese linguistic knowledge base.

and concept-attribute association proposed in HowNet is really the further development and promotion of lexicalism.

In addition, Xingguang Lin proposed many new ideas regarding dependency grammars, case grammars, and the valence of Chinese verbs, in which he carefully considered the characteristics of the Chinese language. He compiled the *Dictionary of Contemporary Chinese Verbs* based on his ideas of valence.[30] Chinese-language researchers have widely applied Lin's work.

Ever since corpus-based methods became the main NLP technique in the 1990s, Chinese scholars have attempted to study them on the basis of statistics and apply them to the R&D of CIP. The most remarkable results are reflected in two ways. First, researchers proposed and developed many new approaches to CWS, including the language-model-based approach, the HMM-based approach, and the CRF-based (condition-random-field-based) approach. The evaluation results announced by SIGHAN (the Association of Computational Linguistics' special interest group on Chinese language processing) show that CWS system performance has significantly improved in recent years.[31]

The second major result was the extensive study of statistical MT. For instance, Yang Liu and his colleagues proposed a novel translation model based on the tree-to-string alignment template.[32,33] For this work, they received the Best Asian Language Paper Award from the Asian Federation for NLP in 2006.[32] In recent years, some MT systems performed well during international evaluations of spoken language translation (IWSLT).[34,35]

Unfortunately, few widely recognized theories or models have been proposed in China. However, our NLP work has come a long way in the past 20 years. Within the last decade, some universities and institutes have set up an NLP (or CL) specialty. We estimate that more than 3,000 students, including undergraduates, graduate students, and PhD candidates, are studying or working on NLP in China.

Applied R&D

The development of practical NLT has had fruitful results. Three significant projects were the Chinese Laser Typesetting System developed by Xuan Wang, the Chinese character card developed by Guangnan Ni, and the optical-character-recognition system developed by CASIA. Some might not consider them NLT, but the following results were definitely generated from applied R&D in NLP and CIP.

Chinese Steno Machine. Yawei Tang, a stenography expert and professor, originated and developed the Yawei Chinese stenograph, the first electronic steno machine in China. In 1934, when he was just 19, he proposed his first ideas about the steno machine. He started to develop the machine on a computer in the early 1980s, proposed the first complete coding scheme and keyboard design for the machine in 1993, and exhibited the first machine on 19 May 1994. According to the design, about 600 Chinese characters can be input per minute. The Yawei steno machine is widely used in court, captioning, meetings, real-time translation, and other applications. In 2005, Tang was given the National Qian Weichang Award of Science and Technology, the highest award in the area of CIP.

Chinese Input System. Xiaolong Wang was the first to propose a method for inputting Chinese pinyin sentences into a computer. (Pinyin, a system for transliterating Chinese characters into the Roman alphabet, was introduced in 1959 and adopted by the People's Republic of China in 1979.) At the beginning of the 1990s, he and his colleagues developed the first input system, called InSun, based on minimum word segmentation.[36] InSun changed the traditional Chinese input method, which used either characters or fixed words as input units. Since then, many new Chinese input methods have been proposed and developed, such as the Ziguang and the Sougou methods. So, it's getting easier and easier to input Chinese characters.

Application of MT. As we mentioned earlier, in the 1990s Chen and his colleagues developed the IMT/EC-863, for which they received China's National Award of Science and Technology Progress. On the basis of the MT technology, Chen and his colleagues founded the Huajian Group Company (www.hjtek.com/en/index.html) and the CAS Research Center of Computer and Language Information Engineering. Huajian Group has developed some MT products and a lot of other application software; it not only sells MT products but also

has a big share of the multilingual translation service market.

Information Retrieval. Information retrieval has become an attractive research topic in recent years, especially since Google became popular. Successful Chinese companies in this area include Baidu (www.baidu.com.cn), TRS (www.trs.com.cn/en), and Zhongsou (www.zhongsou.com), which all make use of CIP techniques. The search engines have also employed new CIP approaches, such as CWS, NER, and out-of-vocabulary processing.

In addition, many companies are applying information retrieval methods in combination with NLP and other technologies—for example, text-to-speech systems, computer-assisted teaching, digital libraries, and so on.

Problems Yet to Be Solved

Of course, some problems continue to challenge NLT development in China. The following problems are especially difficult obstacles.

No Authoritative National Standard for Developing Data Resources

As we know, standards and criteria are important in the development of online language data resources. The Chinese government has organized experts to work out these standards and criteria for CIP, including ones for CWS and POS tagging, the Chinese API specification for hand-held devices for personal information processing, and so on. However, because the standards aren't authoritative or well implemented, many corpora use different tags. This means that Chinese corpora are limited in size and quality and they can't be aggregated easily or widely shared. In summary, there's no authoritative national corpus.

Weak Basic Research

In the NLP research area, many grammars, models, and methods are widely used. Unfortunately, few of them are especially proposed for CIP. As we know, there are many differences between Chinese and English, in both syntax and semantics. However, almost all grammars, models, and algorithms used in CIP come from studies of English processing. Some of these are useless; we look forward to having effective theories, models, and algorithms for CIP.

Ignoring Linguistic Study

In recent years, especially since corpus-based approaches became the mainstream in NLP, researchers have been employing more and more models and algorithms of machine learning. Studying specific language phenomena is ignored. In our opinion, NLT is not only mathematics but also linguistic engineering. So, we strongly suggest paying more attention to linguistic study and carefully combining mathematical methods with specific language-phenomena processing.

National Programs and Projects

In the past two decades, the Chinese government has paid a great deal of attention to NLT R&D. Many projects have been funded.

The National 973 Program

The National Key Fundamental Research Program (the 973 Program), proposed in March 1997, aims to solve the key fundamental problems of science and technology that are needed for national economic development. In the past decade, the government has funded 384 projects, several of which deal with NLT. For example, the Image, Speech, Natural Language Processing, and Knowledge Mining project, which received support from 1998 to 2002, financed the Chinese LDC. Songde Ma was the principal investigator, and many universities and institutes, including CASIA, Tsinghua University, Northeast University, the Institute of Acoustics, CAS, and the CASS Institute of Linguistics, participated. Another NLT project, called the Theory and Method of Digital Content Understanding, was funded in 2004.

The Natural Science Foundation

The National Natural Science Foundation of China (NSFC) supports basic scientific research. In the NFSC, three departments support projects regarding NLT. In recent years, almost all NLT problems, including lexical analysis, syntactic parsing, and semantic computing, as well as problems in minority-language information processing, have been involved in these projects.[37]

The National 863 Program

China's National Hi-Tech R&D Program (the 863 Program), proposed in March 1986, focuses on the R&D of high technology oriented to practical applications. In recent years, the 863 Program has funded many NLT projects, especially in developing language data resources and Internet techniques. Five years ago, the 863 Program funded important R&D projects for technologies to support the 2008 Olympic Games, including multilingual MT, personalized information services based on the Internet, mobile communications, and kiosks. Some of the final products were used in the Olympics.

National Key Technology R&D Program

This program supports projects to turn current techniques and resources into practical systems and to build a high-quality, large-scale corpus. Projects include multilingual information services, development of a multilingual corpus, and Internet content management.

Other Support

The Chinese government also has funds and programs to support NLT R&D, such as the Innovation Fund for Technology-Based Firms and international cooperation funds. In addition, all the provinces and municipalities have their own natural science foundations or similar programs that support NLT R&D in various ways and levels.

What has occurred in NLT work in the last 50 years? Although some corpora and linguistic knowledge bases have been developed and are widely used, we don't yet have authoritative standards. We've also made much progress in basic research, but we still need original models, algorithms, and grammars for CIP. Because the Chinese market is so large, we will be able to use additional practical NLT technologies as soon as they're developed. ■

Acknowledgments

We thank Zhiwei Feng for information on the history of NLP research in China. This article was partially funded by the National Natural Science Foundation of China under grants 60575043 and 60736014, the National Key Technology R&D Program under grant 2006BAH03B02, and the Hi-Tech Research and Development Program of China under grant 2006AA01Z194 and 2006AA010108-4.

The Authors

Chengqing Zong is a professor in natural-language technology at the National Laboratory of Pattern Recognition and the deputy director of the National Laboratory of Pattern Recognition, which is part of the Chinese Academy of Sciences' Institute of Automation. He's also a guest professor at Tsinghua University and the Graduate University of the Chinese Academy of Sciences. His research interests include machine translation, information extraction, and human-computer dialogue systems. Zong received his PhD from the Chinese Academy of Sciences' Institute of Computing Technology. He's a director of the Chinese Association of Artificial Intelligence and the Society of Chinese Information Processing, and is an executive member of the Asian Federation of Natural Language Processing. Contact him at cqzong@nlpr.ia.ac.cn.

Qingshi Gao is a professor in the Computer Science Department at the Beijing University of Science and Technology and an academician in the Chinese Academy of Sciences. His research interests include the architecture of large-scale computer systems, parallel algorithms, human-intelligence simulation and application, and natural language processing. Gao received his bachelor's degree in mathematics from Peking University. Contact him at qsgao@public.bta.net.cn.

References

1. B. Manaris, "Natural Language Processing: A Human-Computer Interaction Perspective," *Advances in Computers*, vol. 47, 1998, pp. 2–68.
2. Z. Feng, *Research on Machine Translation*, China Translation and Publishing Corp., 2004 (in Chinese).
3. J.R. Pierce et al., *Language and Machines—Computers in Translation and Linguistics*, Automatic Language Processing Advocacy Committee (ALPAC) report, Nat'l Academy of Sciences, 1966; http://en.wikipedia.org/wiki/ALPAC.
4. Z. Dong, "Logical Semantics and Its Application in Machine Translation," *China's Machine Translation*, Y. Liu, ed., Knowledge Press, 1984, pp. 25–45 (in Chinese).
5. Z. Liu, "Overview of the English-Chinese Machine Translation System JFY-II," *The Chinese Language*, no. 3, 1981, pp. 216–220, and no. 4, 1981, pp. 279–285 (in Chinese).
6. Q. Gao and Z. Chen, "The Principle of Human-Like Machine Translation," *J. Computer Research & Development*, vol. 26, no. 2, 1989, pp. 1–8.
7. Q. Gao, X. Gao, and Y. Hu, "Semantic Language and Multi-Language MT Approach Based on SL," *J. Computer Science & Technology*, vol. 18, no. 6, 2003, pp. 848–852.
8. Z. Chen and Q. Gao, "Intelligent English-Chinese Machine Translation System IMT/EC," *Sciences in China*, vol. A, no. 2, 1989, pp. 186–194 (in Chinese).
9. Z. Chen and Q. Gao, "A New Context-Sensitive Subcategory (SC) Grammar for Machine Translation," *Chinese J. Computers*, no. 11, 1992, pp. 801–808 (in Chinese).
10. T. Yao et al., *Natural Language Understanding: A Study to Make Machines Understand Human Language*, Tsinghua Univ. Press, 2002 (in Chinese).
11. Y. Liu, Q. Tan, and X. Shen, *The Standard of Contemporary Chinese Word Segmentation for Information Processing and Word Segmentation Methods*, Tsinghua Univ. Press and Guangxi Press of Science and Technology, 1994 (in Chinese).
12. P.F. Brown et al., "A Statistical Approach to Machine Translation," *Computational Linguistics*, vol. 16, no. 2, 1990, pp. 79–85.
13. P.F. Brown et al., "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 263–309.
14. A.L. Berger et al., "The Candide System for Machine Translation," *Proc. ARPA Conf. Human Language Technology* (HLT 94), Assoc. for Computational Linguistics, 1994, pp. 157–162.
15. J. Hutchins, "Machine Translation over Fifty Years," *Histoire, Epistémologie, Langage*, vol. 23, no. 1, 2001, pp. 7–31.
16. "Machine Translation," *NSF Report on Multilingual Information Management: Current Levels and Future Abilities*, ch. 4, 1999; www.cs.cmu.edu/~ref/mlim/index.html.
17. H. Kitano, *Speech-to-Speech Translation: A Massively Parallel Memory-Based Approach*, Kluwer Academic Publishers, 1994.
18. J. Yang et al., "Design and Implementation of Spontaneous English-to-Chinese Speech-to-Speech Translation System," *J. Acoustics*, vol. 17, no. 5, 1992, pp. 327–333.
19. C. Zong and M. Seligman, "Toward Practical Spoken Language Translation," *Machine Translation*, vol. 19, no. 2, 2005, pp. 113–137.
20. Z. Feng, "The History and Current Status of Corpus Base Development in China—Review and Problem Analysis of Corpus Based Research," *Proc. Int'l Conf. Chinese Computing* (ICCC 01), 2001, pp. 29–46 (in Chinese).
21. S. Yu et al., "The Specification for Basic Processing of the Contemporary Chinese Corpus at Peking University," *J. Chinese Information Processing*, vol. 16, nos. 5–6, 2002, pp. 49–64, 58–65 (in Chinese).
22. Z. Dong and D. Qiang, *HowNet and the Computation of Meaning*, World Scientific Publishing Company, 2006.
23. Z. Dong and Q. Dong, "HowNet—a Hybrid Language and Knowledge Resource," *Proc. Int'l Conf. Natural Language Processing and Knowledge Eng.*, IEEE Press, 2003, pp. 820–824.
24. J. Mei et al., *The Chinese Thesaurus*, Shanghai Dictionary Press, 1996 (in Chinese).
25. C. Huang, "The Outline of HNC," *J. Chinese Information Processing*, vol. 11, no. 4, 1997, pp. 11–20 (in Chinese).
26. Y. Jin, *Language Technology of HNC and Its Application*, Science Press, 2006 (in Chinese).
27. J. Zhao et al., "Construction and Development of Chinese LDC," *Some Important Issues in Chinese Information Processing*, B. Xu et al., eds., Science Press, 2003, pp. 218–225 (in Chinese).
28. C. Huang and X. Zhang, "Milestones of Natural Language Processing Technology," *Foreign Language Teaching and Research*, no. 3, 2002, pp. 180–187 (in Chinese).
29. Z. Feng, "Multi-Label and Multi-Branch Tree for Automatic Analysis of Chinese Sentences," *J. Artificial Intelligence*, no. 2, 1983 (in Chinese).
30. X. Lin et al., *Dictionary of Contemporary Chinese Verbs*, Beijing Language and Culture Univ. Press, 1994 (in Chinese).
31. G.-A. Levow, "The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition," *Proc. 5th SIGHAN Workshop Chinese Language Processing*, Assoc. for Computational Linguistics, 2006, pp. 108–117.
32. Y. Liu, Q. Liu, and L. Shouxun, "Tree-to-String Alignment Template for Statistical Machine Translation," *Proc. 21st Int'l Conf. Computational Linguistics and 44th Ann. Meeting Assoc. Computational Linguistics*, Assoc. for Computational Linguistics, 2006, pp. 609–616.
33. Y. Liu, *Research on Tree-to-String Statistical Translation Models*, doctoral dissertation, Inst. Computing Technology, Chinese Academy of Science, 2007 (in Chinese).
34. C.S. Fordyce, "Overview of the IWSLT 2007 Evaluation Campaign," *Proc. Int'l Workshop Spoken Language Translation* (IWSLT 07), 2007, pp. 1–12; www.mt-archive.info/IWSLT-2007-TOC.htm.
35. Y. He et al., "The CASIA Statistical Machine Translation System for IWSLT 2008," *Proc. Int'l Workshop Spoken Language Translation* (IWSLT 08), 2008, pp. 85–91; www.slc.atr.jp/IWSLT2008.
36. X. Wang, "Chinese Pinyin Sentence Input System InSun," *J. Chinese Information Processing*, vol. 7, no. 2, 1993, pp. 45–54 (in Chinese).
37. L. Xu and T. Zhao, "Summarization of Results of Program Funded by NSFC in the Field of Natural Language Processing in Recent Years," *J. Software*, vol. 16, no. 10, 2005, pp. 1853–1858 (in Chinese).