Front. Comput. Sci. China 2007, 1(1): 58–93
DOI 10.1007/s11704-007-0007-2

SURVEY ARTICLE

HTRDP evaluations on Chinese information processing and intelligent human-machine interface

LIU Qun (✉)[1], WANG Xiangdong[1,3], LIU Hong[1], SUN Le[2], TANG Sheng[1], XIONG Deyi[1,3], HOU Hongxu[1,3], LV Yuanhua[2], LI Wenbo[2], LIN Shouxun[1], QIAN Yueliang[1]

1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
2 Institute of Software, Chinese Academy of Sciences, Beijing 100085, China
3 Graduate University of Chinese Academy of Sciences, Beijing 100085, China

© Higher Education Press and Springer-Verlag 2007

Abstract From 1991 to 2005, China's High Technology Research and Development Program (HTRDP) sponsored a series of technology evaluations on Chinese information processing and intelligent human-machine interface, which is called HTRDP evaluations, or "863" evaluations in brief. This paper introduces the HTRDP evaluations in detail. The general information of the HTRDP evaluation is presented first, including the history, the concerned technology categories, the organizer, the participants, and the procedure, etc. Then the evaluations on each technology are described in detail respectively, covering Chinese word segmentation, machine translation, acoustic speech recognition, text to speech, text summarization, text categorization, information retrieval, character recognition, and face detection and recognition. For the evaluations on each technology categories, the history, the evaluation tasks, the data, the evaluation method, etc., are given. The last section concludes the paper and discusses possible future work.

Keywords HTRDP evaluations, Chinese information processing, intelligent human-machine interface

1 Preface

In recent decades, public technology evaluations became effective a boost to the research and development of certain areas of computer science, such as speech recognition, information retrieval and extraction, machine translation, and natural language parsing, etc. There are two types of evaluations: official and unofficial. Official evaluations are projects that are supported and funded by the government, while unofficial evaluations are usually organized by scientific research societies. The most famous official evaluation organizer is the NIST—the National Institute of Standard Technology of the United States. It organizes several evaluations every year, and most of which are supported by the Defense Advanced Research Projects Agency (DARPA). Some of them are very famous and have a long history, e.g., the TREC evaluation conference on information retrieval that was started in 1992 as part of the TIPSTER Text program and had been executed 15 times before the end of 2006. In Europe, the TC-STAR speech translation project organizes evaluations on speech recognition, speech to text and machine translation in recent years.

In China, there is a series of official technology evaluations which are called HTRDP evaluations on Chinese information processing and intelligent human-machine interface. The HTRDP means High Technology Research and Development Program, which is one of the biggest R&D fund supported by the Chinese government. The HTRDP is also called the "863" program, since it is approved by Chinese former leader DENG Xiaoping in March 1986. The Chinese information processing and intelligent human-machine interface is one of the most important areas supported by the HTRDP. In 1991, HTRDP started the evaluations, which is also called the "863" evaluations, to examine the progress of the projects it had funded. However, the HTRDP evaluation is becoming a more and more important occasion for researchers to acquire large scale data and benchmarks, to demonstrate their research progress and to exchange new ideas with each other.

This paper introduces the HTRDP evaluations in detail. The general information of the HTRDP evaluations is described in section 2. Then, the evaluations of Chinese word segmentation, machine translation, speech recognition, information retrieval, and other technologies are described in detail in Section 3 to Section 11. The last section concludes this paper and discusses future work.

Received November 3, 2006; accepted December 25, 2006

E-mail: liuqun@ict.ac.cn

2 General information of HTRDP evaluations

The HTRDP evaluations was started in 1991, while in 1990, there was an informal evaluation in speech recognition. In the first two evaluations in 1991 and 1992, there were only two technology categories evaluated — speech recognition and Chinese character recognition. While in the third evaluation in 1994, two more technology categories had been included: text to speech and machine translation. At present, the HTRDP evaluations cover a wider range of technology categories, which include the following:

- ASR: Automatic Speech Recognition
- TTS: Text to Speech
- MT: Machine Translation
- CWS: Chinese Word Segmentation
- IR: Information Retrieval
- TC: Text categorization
- TS: Text Summarization
- CR: Character Recognition
- FR: Face Detection and Recognition

Table 1 lists the technology categories which was evaluated in the HTRDP evaluations in each year.

Table 1 Technology categories evaluated in each HTRDP evaluation

	1st	2nd	3rd	4th	5th	6th	7th	8th
	1991	1992	1994	1995	1998	2003	2004	2005
ASR	√	√	√	√	√	√	√	√
TTS			√	√	√	√	√	
MT			√	√	√	√	√	√
CWS				√	√	√	√	
IR						√	√	√
TC						√	√	
TS				√	√	√	√	
CR	√	√	√	√	√	√		
FR							√	

For each technology category, several evaluation tasks will be defined. Table 2 gives an example of the evaluation tasks in three technology categories in HTRDP evaluation 2005.

There were five HTRDP evaluations during 1991 to 1998. After a five-year interruption, the HTRDP evaluations resumed in 2003 and were conducted annually from 2003 to 2005.

The general procedure of the HTRDP evaluation is:

1) Discussing technology categories to be evaluated and the tasks in each category;

2) Releasing the evaluation schedule and guidelines. All the official information about the HTRDP evaluation is provided on a website (http://www.863data.org.cn)

3) Registration. After the registration, a mailing list is created for each evaluation category to discuss the problems about the evaluation.

4) Data preparing. In the previous several evaluations, large scale training data are provided to some tasks. All the training data and test data, as well as the evaluation tools, can be licensed via ChineseLDC (http://www.chineseldc.org) after the evaluation.

Table 2 Tasks in 2005 HTRDP Machine Translation Evaluation

Category	Task		Task ID
Machine Translation	Machine Translation	Chinese→English	CEMT
		English→Chinese	ECMT
		Chinese→Japanese	CJMT
		Japanese→Chinese	JCMT
		Japanese→English	JEMT
	Word Alignment	English→Japanese	EJMT
		Chinese↔English	CEWA
Information Retrieval	Relevant web page retrieval	Relevant web page retrieval	WEB
Automatic Speech Recognition	Continuous Desktop Speech Recognition	2X realtime	CSR_PC_2X
		20X realtime	CSR_PC_20X
	Keyword Spotting in Continuous Telephone Speech	2X realtime	KWS_PHONE_2X

5) Testing. The first to the seventh HTRDP evaluations adopt an on-site evaluation method, which means that all the participating sites should take their systems to the specified evaluation spot and run the systems during the specified time period. From the eighth evaluation in 2005, online evaluation has been used. This will encourage more researchers, especially oversea researchers, to participate in the HTRDP evaluations.

6) Workshop. An evaluation workshop has been held after each year's evaluation since 2004. Each participating site is required to report the technical details of its system. This provides a very good chance for the researchers to exchange their ideas and share their experience.

All the HTRDP evaluations were organized by the Institute of Computing Technology, Chinese Academy of Sciences. There were two other co-organizers: The Institute of Software of Chinese Academy of Sciences and the National Institute of Information and Communication Technology of Japan, which provided technical support on the information retrieval and machine translation evaluations.

The HTRDP evaluations are open to any site all over the world, not only those who undertake HTRDP projects. The participants of the HTRDP evaluations include almost all the active research organizations in related areas in China, and some of them come from overseas. Table 3 gives the number of categories and participating systems in each HTRDP evaluations.

Table 3 Number of categories and participating systems in each HTRDP evaluations

	1990	1991	1992	1994	1995	1998	2003	2004	2005
	Pre	1st	2nd	3rd	4th	5th	6th	7th	8th
Number of Categories	1	2	2	4	6	6	8	8	3
Number of Participating Systems	5	16	17	39	65	43	46	113	45

Tables 4–6 listed the participatinggroups in the HTRDP 2005 evaluations on machine translation, automatic speech recognition and information retrieval.

Table 4 Participating groups of HTRDP 2005 evaluation on machine translation

Beijing University of Technology
CCID Cooperation
Futsuji Cooperation (Japan)
Huajian Cooperation
Harbin Institute of Technology
Institute of Automation, Chinese Academy of Sciences
Multran Cooperation
National University of Defense Technology
Nanjing University
Sharp Cooperation (Japan)
Transtar Cooperation
Xiamen University

Table 5 Participating groups of HTRDP 2005 evaluation on automatic speech recognition

School of Information Science and Technology, Beijing Institute of Technology
Electronic Engineering Department of Tsinghua University
PRIS laboratory of Beijing University of Post and Telecommunications
National Laboratory on Machine Perception, Peking University
ThinkIT Speech Laboratory, Institute of Acoustic, Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Initport Incorporation, Shanghai
Speech Processing Laboratory, Harbin Institute of Technology

Table 6 Participating groups of HTRDP 2005 evaluation on information retrieval

Institute of Automation, Chinese Academy of Sciences
Information Retrieval Lab, Harbin Institute of Technology
State Key Laboratory of Intelligent Technology and Systems of Tsinghua University
PRIS laboratory of Beijing University of Post and Telecommunications
Institute of Computer Science and Technology of Peking University

In next sections, we will describe the details of the evaluations for each technology category.

3 Chinese word segmentation

3.1 Introduction

Chinese word segmentation, including named entity recognition and part of speech tagging, is a foundational requirement of almost all application of Chinese language processing, such as machine translation, information retrieval, speech to text, etc. However, it is not a simple work. Lots of researchers have been dedicated to this research area since the 1980s, but there are no high quality and accessible Chinese word segmentation tools for most Chinese NLP researchers, until the statistical approaches are broadly used in this area in recent years.

Besides the HTRDP evaluations, another Chinese word segmentation evaluation series are the SIGHAN Chinese Word Segmentation Bakeoffs. SIGHAN is a Special Interest Group of the Association for Computational Linguistics, providing an umbrella for researchers in the industry and the academia working in various aspects of Chinese Language Processing. Till now, three SIGHAN Backoffs have been executed, attached with the 2nd, 4th, and 5th SIGHAN work- shop in 2003, 2005 and 2006 [1, 2, 3]. In the first and second Backoffs, there were only Chinese word segmentation tracks. However, in the recent 3rd Backoff in 2006, another Named Entity Recognition track was included. POS tagging has not been evaluated in the SIGHAN evaluations. In these Backoffs, no explicit specifications for Chinese word segmen- tation were given. The specifications were implicit in the given training corpus. Usually several different training corpus was given, and the different test corpus was also provided for the different training corpus respectively.

The ACE evaluations [4] which are organized by NIST also have evaluation tracks on named entity recognition on Chinese and other languages. However, the ACE evaluations focus on information extraction problems. It has a more detailed classification of named entities than SIGHAN and HTRDP evaluation do, while it does not concern Chinese word segmentation and POS tagging.

In the HTRDP evaluations, Chinese word segmentation evaluations have been conducted four times, in 1995, 1998, 2003 and 2004 respectively. The first three evaluations were comprehensive evaluations which contain tasks including:

- Chinese word segmentation (CWS)
- Named entity recognition (NER)
- Part of speech tagging (POS)

The 2004 evaluation focused on named entity recognition, with much more detailed specifications. The 2004 evaluation contained data for both simplified Chinese and traditional Chinese, while the previous three evaluations were only for simplified Chinese.

We can see that the early evaluations used very small data, compared with the large data used in recent years.

Figure 1 gives the best results of the word segmentation precision of each evaluation.

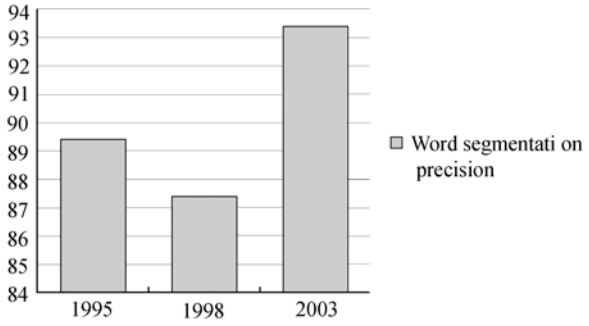

Fig. 1 The best results of word segmentation precision in each year

Figure 2 gives the best results of precision and recall of named entity recognition of each evaluation (the data of 2004 evaluation in this figure is for simplified Chinese).

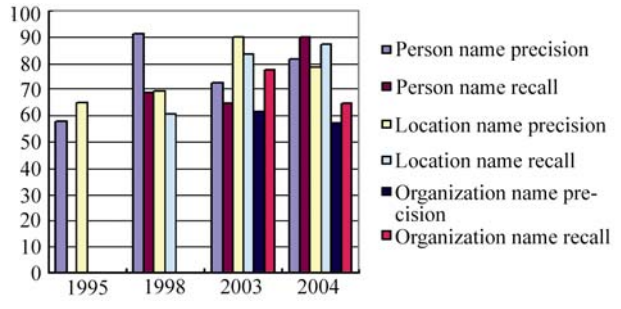

Fig. 2 The best results of name entity recognition in each year

Since the data and the specifications for each evaluation are not the same, the results for these evaluations are not comparable. In the early evaluations, the size of the test data was rather small, and the specifications were very simple. While in recent evaluations, the size of the test data became very large, and the specifications were much more complex and strict. Therefore, although we cannot see significant improvement in the above figures, actually the state of the art of Chinese word segmentation and named entity recognition has advanced greatly in recent years with the broad usage of statistical technology.

3.2 Specifications and metrics

3.2.1 Chinese word segmentation (CWS)

3.2.1.1 Specifications

Since the concept *word* is not a linguistically well defined concept in Chinese, different researchers usually have different definition of the segmentation unit (word) in Chinese word segmentation. There are several famous specifications for Chinese word segmentation, for example, the Chinese standard GB/T 13715-92, the Peking University word segmentation specification for the People's Daily Corpus, the Tsinghua University word segmentation specification, the Chinese Penn Treebank word segmentation specification, and etc. In the HTRDP evaluation, several methods are used to deal with the problem of segmentation unit definition, rather than proposing a new specification:

1. The Chinese standard GB/T 13715-92 and the Tsinghua University specification are recommended as the reference specifications;
2. A small size sample corpus are given to show how to deal with some real ambiguity in word segmentation;
3. An error-tolerance criterion are used in the evaluation, which means that even if a word segmentation piece given by the participating system is not the same with the reference segmentation, it may be also regarded as correct segmentation, if it did not violate the common sense, and the rules used in the sample corpus.

To carry out the error-tolerance criterion, some human judgments should be used in the evaluation process. Firstly, all the word segmentation given by the participating systems are compared with the reference segmentation, by which process a file containing all the different segmentation pieces are generated and be submitted to human judgments. Then, after the human judgments, those word segmentation pieces which are different with the reference segmentations but are deemed acceptable by human experts are added to the reference segmentations. Finally, a software tool will compute the segmentation score for each participating system according to the new reference segmentation, which may contain different acceptable segmentation pieces for the same text. This technology is somewhat like the pooling-technology used in the TREC information retrieval evaluation.

3.2.1.2 Metrics

The metrics for Chinese word segmentation evaluation are precision, recall, and F-score:

$$\text{CWS Precision} = \frac{\text{number of correctly segmented words}}{\text{number of words in output segmentation}}, \tag{1}$$

$$\text{CWS Recall} = \frac{\text{number of correctly segmented words}}{\text{number of words in reference segmentation}}, \tag{2}$$

$$\text{CWS F-Score} = \frac{2\times\text{CWS Precision}\times\text{CWS Recall}}{\text{CWS Precision+CWS Recall}}. \tag{3}$$

3.2.2 Named entity recognition (NER)

3.2.2.1 Specifications

In the first three HTRDP Chinese word segmentation evaluations, the named entity recognition is one of the three tasks in this evaluation category, and there are no specifications for the named entity recognition. The four types of named entity are defined in these evaluations: Person Names, Location Names, Organization Names, and Other Names.

In the HTRDP evaluation of 2004, the named entity recognition evaluation is the only task in the Chinese word segmentation category. A detailed specification is given for the named entity recognition evaluation. There are three types of named entity defined: Proper Names, Temporal Expressions, and Numerical Expressions. For Proper Names, there are three subtypes: Personal Names (PER), Location Names (LOC), and Organizational Names (ORG). For the Temporal Expressions, there are only two subtypes: Dates (DAT) and Times (TIM). For the Numerical Expressions, only one subtype are defined (NUM). For the detailed specifications, please visit the HTRDP evaluation website.

3.2.2.2 Metrics

For the evaluation of the named entity recognition, the metrics are also precision, recall and F-score. These three metrics are used for each type and subtype, and also for the overall evaluation. Here we just give the definitions of these metrics for person name recognition (PER):

$$\text{PER Precision} = \frac{\text{number of correctly recognized person names}}{\text{number of person names in output data}}, \quad (4)$$

$$\text{PER Recall} = \frac{\text{number of correctly recognized person names}}{\text{number of person names in reference data}}, \quad (5)$$

$$\text{PER F-Score} = \frac{2 \times \text{PER Precision} \times \text{PER Recall}}{\text{PER Precision+PER Recall}}. \quad (6)$$

3.2.3 Part of speech tagging (POS)

3.2.3.1 Specifications

Similar with the uncertainty in Chinese word segmentation, part of the speech tagging is also a problem that is not well defined in linguistics. For some words, different linguists give different part of speech tags. For example, for the word "劳动" in the sentence "劳动创造了人", some linguists regard it as a noun, and some other linguists regard it as a verb. Also, there are different part of speech tagging specifications used by the researchers, such as the PKU part of the speech tagging specifications for the People's Daily corpus, the ICT's part of the speech tagging specification for the ICTCLAS, and the Chinese Penn Treebank's part of the speech tagging specification.

In the part of speech tagging task in the HTRDP evaluations, we did not define specific parts of the speech tagging specifica- tions. To deal with the uncertainty problem of part of speech tagging, three methods are adopted, which is similar with the methods used for word segmentation evaluation:

1. The part of speech tagging specifications released by the Institute of Applied Linguistics of Ministry of Education are used as reference specifications;
2. A small size sample corpus are given according to the reference specifications;
3. An error-tolerance criterion are used in the evaluation, which means that even if a word's part of speech tag given by the participating system is not the same with the reference tag, it may also be regarded as a correct tag, if it can be accepted by some other specifications, and did not violate the rules used in the given sample corpus.

To carry out the error-tolerance criterion, we also have adopted human judgment, which is just like the technology used in word segmentation and is explained in section 3.2.1.1.

3.2.3.2 Metrics

Four metrics are used in part of speech tagging:

$$\text{POS Precision} = \frac{\text{number of correctly tagged words}}{\text{number of words in output data}}, \quad (7)$$

$$\text{POS Recall} = \frac{\text{number of correctly tagged words}}{\text{number of words in reference data}}, \quad (8)$$

$$\text{POS F-Score} = \frac{2 \times \text{POS Precision} \times \text{POS tagging Recall}}{\text{POS Precision+POS Recall}}, \quad (9)$$

$$\text{POS RelativePrecision} = \frac{\text{number of correctly tagged words}}{\text{number of correctly segmented words}} \quad (10)$$

We can find that

$$\text{POS Relative Precision} = \frac{\text{POS Precision}}{\text{CWS Precision}}. \quad (11)$$

3.3 Evaluation Data

3.3.1 Training data and sample data

Up to now, no training data are provided in the HTRDP Chinese word segmentation evaluation. The participants are permitted to use any data available to train their systems.

In the 2003 Chinese word segmentation evaluation, a small sample data were provided, which contained about 1100 words.

In the 2004 named entity recognition evaluation, a sample data which contains 20,000 characters in the traditional Chinese and 50,000 characters in simplified Chinese were provided to all participants.

3.3.2 Test data and reference data

The test data in 1995, 1998 and 2003 contained 10,000 Chinese characters, 100,000 characters and 400,000 characters respectively, which were used for the word segmentation evaluation, and part of which were used for the named entity recognition evaluation and part of speech tagging evaluation.

In 2004, the test data contain 1,000,000 Chinese characters, where one half is for simplified Chinese, and the others for traditional Chinese. All of these data were used for the named entity recognition evaluation.

The test data sizes of each year's evaluation are shown in Fig. 3.

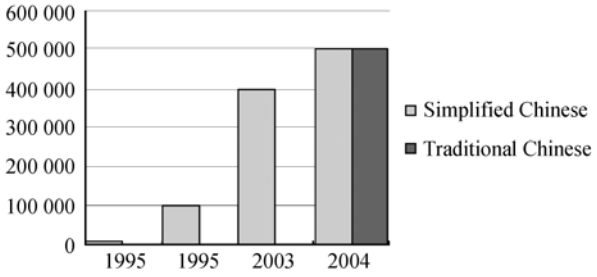

Fig. 3 Test data size for each year's evaluation

All the reference data were made by human experts.

The reference data of 2003 and 2004 evaluation can be licensed via ChineseLDC.

3.4 Results and Analysis

3.4.1 The results of 2003 Chinese word segmentation evaluation

Table 7 gives the precision, recall and F-score for the 2003 Chinese word segmentation evaluation.

We can see that all the systems' F-score are more than 90%; the best system's F-score is 93.46%. It indicates that Chinese word segmentation technology is ready for most of the applications.

Table 7 The results of 2003 Chinese word segmentation evaluation

System ID	Precision	Recall	F-Score
System 1	91.42%	89.27%	90.33%
System 2	93.44%	93.49%	93.46%
System 3	92.04%	93.85%	92.94%
System 4	92.88%	93.40%	93.14%
System 5	93.22%	93.69%	93.45%

3.4.2 The results of 2003 named entity recognition evaluation

The next two tables give the results for the 2003 named entity recognition evaluation. Table 8 gives the overall results, while Table 9 gives the detail results for person names, location names, organization names and other names.

Table 8 The overall result of 2003 named entity recognition evaluation

System ID	Precision	Recall	F-Score
System 2	76.45%	70.15%	73.16%
System 3	34.63%	47.66%	40.11%
System 4	49.42%	63.03%	55.40%
System 5	57.64%	60.74%	59.15%
System 6	69.73%	71.73%	70.72%

From the results we can see that the performances of named entity recognition are still not satisfactory for the current systems. The best F-score for person names, location names, and organization name, and other names are 68.33%, 86.49%, 68.56%, 34.10%.

3.4.3 The results of 2003 part of speech tagging evaluation

Table 10 gives the results of 2003 part of speech tagging evaluation.

We can see that the best performance of part of speech tagging is about 87.5% (absolute precision) and 93.6% (relative precision). The performance is much lower than the performance of English part of speech tagging. We think the main reason is the difference between the languages. As we have indicated, the part of speech in Chinese is not a well defined concept even from a linguistic view. There are a lot of ambiguities in part of speech tagging even when using human judgments, to say nothing of in automatic tagging.

3.4.4 The results of 2004 named entity recognition evaluation

Tables 11−13 give the results of 2004 named entity recognition evaluation for simplified Chinese.

Compared with the results in 2003 evaluation, we can see that the person name recognition had a very large improvement.

Tables 14−16 give the results of 2004 named entity recognition evaluation for traditional Chinese.

We can see that almost all of the results are much lower than that of simplified Chinese. The reason may be that the participants of this task were all coming from mainland China, which only uses simplified Chinese. They did not train their systems using large scale corpus of traditional Chinese.

3.5 Conclusion

The HTRDP conducted Chinese word segmentations in 1995, 1998, 2003 and 2004. In the first three evaluations, the evaluation tracks covered word segmentation, named entity recognition and POS tagging. The last evaluation only focused on named entity recognition. Unlike the SIGHAN Bakeoff on Chinese Word Segmentation, the word segmen-

Table 9 The detailed result of 2003 named entity recognition evaluation

System ID	Person Names			Location Names			Organization Names			Other Names		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
System2	72.35%	64.74%	68.33%	89.72%	83.49%	86.49%	61.54%	77.38%	68.56%	64.74%	23.15%	34.10%
System3	27.27%	43.29%	33.46%	67.72%	78.02%	72.51%	4.65%	10.90%	6.52%	100.00%	0.44%	0.88%
System4	45.36%	61.60%	52.25%	68.64%	87.99%	77.12%	20.36%	31.33%	24.68%	37.67%	20.80%	26.80%
System5	49.79%	68.09%	57.52%	76.72%	84.27%	80.23%	81.51%	10.60%	18.76%	17.99%	24.97%	20.91%
System6	60.59%	78.07%	68.23%	77.95%	86.64%	82.07%	69.31%	60.93%	64.85%	88.89%	10.59%	18.93%

Table 10 The result of 2003 part of speech tagging evaluation

System ID	Precision	Recall	F-Score	Relative Precision
System 2	87.47%	87.52%	87.50%	93.62%
System 3	82.96%	84.59%	83.77%	90.14%
System 4	83.35%	83.81%	83.58%	89.74%
System 5	68.65%	68.99%	68.82%	73.65%

Table 11 The overall results for simplified Chinese

System ID	Recall	Precision	F-Score
System 1	71.34%	63.25%	67.05%
System 2	78.20%	81.93%	80.02%
System 3	77.62%	83.23%	80.33%
System 4	62.54%	58.32%	60.36%
System 5	81.10%	83.69%	82.38%
System 6	71.44%	85.21%	77.72%
System 7	79.56%	83.79%	81.62%
System 8	70.65%	77.42%	73.88%

Table 12 The detailed results for simplified Chinese (part 1)

System ID	Location Names			Person Names			Organization Names		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
System1	73.49%	72.23%	72.85%	74.09%	66.49%	70.08%	47.32%	36.31%	41.09%
System2	76.72%	79.96%	78.31%	87.48%	73.13%	79.66%	33.78%	74.35%	46.45%
System3	72.21%	85.29%	78.21%	81.45%	89.99%	85.51%	40.55%	36.04%	38.16%
System4	61.71%	74.24%	67.40%	67.66%	49.99%	57.50%	31.41%	30.11%	30.74%
System5	79.50%	81.84%	80.65%	84.34%	80.71%	82.48%	39.27%	61%	47.78%
System6	70.22%	86.45%	77.49%	62.19%	84.42%	71.62%	54.67%	56.57%	55.60%
System7	78.43%	87.02%	82.51%	88.47%	81.38%	84.78%	57.41%	64.64%	60.81%
System8	68.78%	84.02%	75.64%	68.83%	75.57%	72.04%	45.10%	43.51%	44.29%

Table 13 The detailed results for simplified Chinese (part 2)

System ID	Dates			Times			Numbers		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
System1	56.76%	74.59%	64.47%	21.48%	4.51%	7.46%	81.27%	67.98%	74.03%
System2	75.39%	86.20%	80.43%	53.70%	73.23%	61.97%	90.15%	90.52%	90.33%
System3	82.12%	88.56%	85.22%	83.70%	85.93%	84.80%	91.71%	92.17%	91.94%
System4	77.50%	66.67%	71.68%	20.37%	7.77%	11.25%	66.86%	59.60%	63.02%
System5	81.73%	86.69%	84.13%	74.07%	81.63%	77.67%	94.12%	91.14%	92.60%
System6	75.78%	88.08%	81.47%	78.52%	82.17%	80.30%	81.81%	93.09%	87.09%
System7	76.71%	81.86%	79.20%	61.48%	63.60%	62.52%	84. 2%	88.57%	86.33%
System8	61.38%	77.40%	68.47%	34.81%	16.32%	22.22%	86.36%	87.93%	87.14%

Table 14 The overall results for traditional Chinese

System ID	Recall	Precision	F-Score
System1	56.04%	54.16%	55.08%
System2	66.63%	74.12%	70.18%
System3	63.10%	73.94%	68.09%
System4	51.51%	51.27%	51.39%
System5	65.84%	73.87%	69.62%
System6	49.80%	79%	61.09%
System7	54.66%	72.96%	62.50%
System8	16.82%	67.41%	26.92%

tation specifications are given explicitly in all the HTRDP Chinese word segmentation evaluation tracks. The metrics of all the tracks are precision, recall and F-score basically, as in the SIGHAN evaluations. In the HTRDP evaluations, some tracks adopt an evaluation method, which partly uses human evaluation for error tolerance, while the SIGHAN Bakeoffs uses a fully automatic evaluation method. The state-of-the-art of Chinese word segmentation has greatly improved in recent years since the broad usage of the statistical method, which reflected the result of the HTRDP evaluation (Fig. 1).

Table 15 The detailed results for traditional Chinese (part 1)

System ID	Location Names			Person Names			Organization Names		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
System 1	61.86%	56.02%	58.80%	57.02%	59.93%	58.44%	25.32%	39.86%	30.97%
System 2	70.04%	68.39%	69.21%	68.06%	61.53%	64.63%	11.22%	78.01%	19.62%
System 3	61.57%	72.11%	66.42%	55.54%	81.85%	66.17%	27.29%	35.99%	31.04%
System 4	52.06%	62.27%	56.71%	51.23%	41.33%	45.75%	14.55%	26.68%	18.83%
System 5	68.49%	62.70%	65.47%	63%	70.47%	66.52%	13.68%	62.93%	22.47%
System 6	51.56%	71.20%	59.81%	37.73%	80.93%	51.46%	14.94%	52.77%	23.29%
System 7	55.10%	71.01%	62.05%	56.92%	66.23%	61.22%	14.36%	57.18%	22.95%
System 8	16.45%	69.52%	26.60%	12.13%	61.88%	20.28%	6.25%	44.35%	10.96%

Table 16 the detailed results for traditional Chinese (part 2)

System ID	Dates			Times			Numbers		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
System 1	22.62%	63.70%	33.39%	27.50%	6.40%	10.38%	79.80%	64.65%	71.43%
System 2	71.40%	82.13%	76.39%	66.37%	77.35%	71.44%	91.78%	86.42%	89.02%
System 3	78.49%	76.50%	77.48%	63.05%	75.63%	68.77%	85.31%	85.60%	85.46%
System 4	64.66%	54.30%	59.03%	34.50%	20.59%	25.79%	68.83%	63.44%	66.02%
System 5	73.75%	80.26%	76.87%	72.50%	79.16%	75.69%	91.79%	83.59%	87.50%
System 6	69.01%	81.09%	74.57%	66.55%	77.87%	71.77%	70.35%	87.78%	78.10%
System 7	61.42%	74.99%	67.53%	27.50%	45.91%	34.39%	74.44%	82.59%	78.31%
System 8	11.78%	63.30%	19.86%	14.36%	20.20%	16.79%	28.36%	81.72%	42.11%

4 Machine translation

4.1 Introduction

Machine translation evaluation is an important method to advance the state of the art of machine translation research. There are several machine translation evaluations in the research community in the world, such as the NIST MT evaluation, the IWSLT evaluation and the TC-STAR MT evaluation.

The most famous machine evaluation series are the NIST MT evaluations [7]. From 2002, the NIST executed a yearly MT evaluation, supported by the DARPA TIDES project. The NIST evaluation uses the automatic evaluation metrics for the first time, such as BLEU and NIST, and provided the large scale training corpus to participants. Now this kind of automatic evaluation metrics are adopted by almost all of the other MT evaluations, such as IWSLT, TC-STAR and HTRDP. The NIST MT evaluations only focus on text translation from Chinese to English and from Arabic to English, while the TC-STAR MT evaluations focus on the spoken language translation between the European languages, and the IWSLT evaluations also focus on spoken language translation between English, Chinese, Japanese and some other languages.

In HTRDP evaluations, the machine translation evaluation is also a traditional category.

Since 1994, six machine translation evaluations have been conduced. In 1994, the first evaluation task of machine translation, there were only two translation directions, which were English to Chinese and Chinese to English. In the evaluation of 2005, we had seven tasks in the HTRDP machine translation evaluation, including a word-alignment subtask.

The machine translation evaluation tasks which have ever been conducted in HTRDP evaluations include:

- Machine Translation from Chinese to English (CEMT)
- Machine Translation from English to Chinese (ECMT)
- Machine Translation from Chinese to Japanese (CJMT)
- Machine Translation from Japanese to Chinese (JCMT)
- Machine Translation from Japanese to English (JEMT)
- Machine Translation from English to Japanese (EJMT)
- Machine Translation from Chinese to French (CFMT)
- Word Alignment between Chinese and English (CEWA)

Table 17 gives the tasks in each machine translation evaluation.

Table 17 The tasks in each machine translation evaluation

	3rd 1994	4th 1995	5th 1998	6th 2003	7th 2004	8th 2005
CEMT	√	√	√	√	√	√
ECMT	√	√	√	√	√	√
CJMT				√	√	√
JCMT		√		√	√	√
EJMT						√
JEMT						√
CFMT					√	
CEWA						√

The number of participators, participating systems and language pairs are given in Fig. 4.

In earlier years, the systems were evaluated in both translation score and system running score. The translation score are subjective score given by several experts. The test sentences were given by translation experts, which covered specific test points. For example, some sentences were given to test the vocabulary of the system, some were used to test the word sense disambiguation, some were for grammatical correctness of target language, etc. The running score are given on the basis of the running procedure. The earlier evaluations were conducted in an on-site way, which means that all the participants should take their system to the same place and run on the test data in the same data. The running score were computed based on the translation speed, the system stability, the friendliness of the interface, and some other aspects. For example, if there were a system corruption in the running, the running score would be reduced.

However, the evaluation method has been greatly changed since 2003. From 2003, the running score are no longer used. The systems are only evaluated on the translation quality. Since IBM researchers proposed the automatic metric BLEU for machine translation, which is successfully used in the NIST evaluation, we also adopt this kind of automatic metrics in HTRDP evaluations, as an addition to subjective evaluation.

For the evaluation of the Japanese related machine translation systems, we begin the cooperation with the NICT (National Institute of Information and Communication Technology, Japan) in 2005. The NICT helped us on the

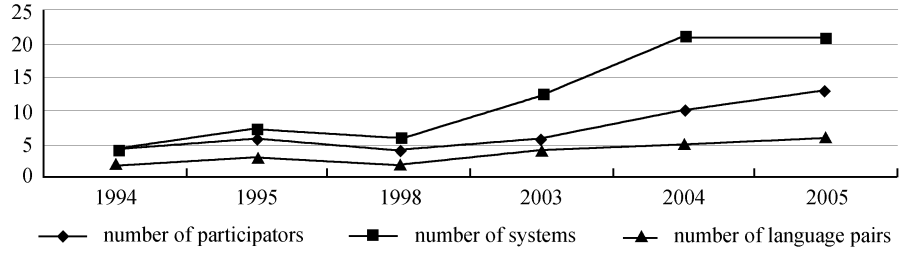

Fig. 4 Participators, systems and language pairs

construction of the test corpus and the reference data, and the human judgments of the translation results of some Japanese related evaluation tasks.

4.2 Evaluation method

In early HTRDP machine translation evaluations, the systems were evaluated in both the translation score and system running score (such as speed and robustness of the systems). However, in the evaluations after 2003, only the translation quality is evaluated. In this paper, we will only focus on the translation quality evaluation.

Machine translation quality was evaluated in human assessment and automatic metrics [12, 13, 16].

4.2.1 Human assessments of machine translation task

In the evaluations before 2004, single metric was used in the human assessments, which was called translation score or intelligibility. In the evaluation of 2005, both ***Adequacy*** and ***Fluency*** were used. The scoring criterion for adequacy and fluency are shown in Table 18 and Table 19 [7].

The results were presented to human judges, for each sentence, in a random order of participants' translations. They read both the source text and the translations given by the participating systems and then gave an adequacy score and a fluency score for each translation according to the scoring criterion. The score is scaled from 0 to 5 with no more than one decimal. The average adequacy and fluency scores over all the sentences by human judges were calculated.

Table 18 Adequacy: scoring criterion for human assessment

Score	Description
0	None
1	Almost no information
2	Little information
3	More information
4	Most information
5	All information

Table 19 Fluency: scoring criterion for human assessment

Score	Description
0	Incomprehensible
1	Difficult to comprehend
2	Disfluent
3	Non-native
4	Good
5	Flawless

4.2.2 Automatic metrics of machine translation task

In the middle 1990s, the organizer of the HTRDP machine translation evaluation proposed a kind of automatic evaluation method based on a test point [5]. A test set which contained thousands of test points is proposed and an automatic evaluation system MTE based on these test points is developed. However, MTE had not been used in the HTRDP evaluation officially.

In 2000s, *n*-gram based automatic machine translation evaluation metrics were proposed, such as BLEU [7] and NIST. HTRDP evaluations adopt these automatic metrics from 2003. In 2003, only the NIST metric was used. In 2004, many automatic metrics, such as BLEU, NIST, GTM [9], mPER, mWER [12], were used.

In 2005, an additional automatic metric called the ICT [18], which was developed by the organizer, was also used. The MT systems were only scored using case sensitive reference translations. For evaluation of the Chinese and Japanese translation, the metrics were applied at a character level, instead of a word level.

BLEU is a metric based on *n*-gram [15, 16], higher is better. The range is from 0 to 1.

$$\begin{aligned} \text{score} &= BP * \exp\left(\sum_{n-1}^{N} w_n \log p_n\right), \\ BP &= \min\left\{1, \exp\left(1 - \frac{L_{\text{ref}}}{L_{\text{sys}}}\right)\right\}, \end{aligned} \tag{12}$$

P_n is count of *n*-gram matched references divided the count of all *n*-gram in translation. *BP* is the length penalty, L_{ref} is length of the most similar reference, L_{sys} is the length of translation. *N* is the maximize *n*-gram length and W_n is the weight of *n*-gram.

NIST is also a metric based on n-gram, higher is better. Its value is always greater than 0, and the upper limit is about 12 to 14.

$$\begin{aligned} \text{score} = \sum_{n=1}^{N} &\left\{ \sum_{\substack{\text{all } w_1 \cdots w_n \\ \text{that co-occur}}} Info(w_1 \cdots w_n) / \sum_{\substack{\text{all } w_1 \cdots w_n \\ \text{in sys output}}} (1) \right\} \\ &* \exp\left\{\beta \log^2\left[in\left(\frac{L_{\text{sys}}}{\overline{L}_{\text{ref}}}, 1\right)\right]\right\}. \end{aligned} \tag{13}$$

β is a constant as a experiential threshold. $\overline{L}_{\text{ref}}$ is the average length of references, and the other parameters are the same as BLEU.

As BLEU and NIST are the main metrics of evaluation, we give out GTM, mWER, mPER and ICT for reference.

The GTM is a metric based on text similarity, higher is better. Its range is from 0 to 1.

$$\text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{14}$$

$$\text{precision} = \text{MMS}/L_{\text{sys}}, \quad \text{Recall} = \text{MMS}/\overline{L}_{\text{ref}}, \tag{15}$$

MMS is the max match length.

mWER is a metric based on edit distance, lower is better. Its values are greater that 0, and mostly less than 1.

$$\text{score} = \min_{\text{all refs}} \{\text{edit distance} / L_{\text{ref}}\} \tag{16}$$

mPER is similar to mWER, but it discards word ordering. The lower the better. Its values are greater than 0, and mostly less than 1.

ICT, is developed by the Institute of Computing Technology [18]. Here we will give some comparison results between the ICT measurement and others in this paper.

In word alignment evaluation, aligning quality will be measured automatically by precision, recall, F1 measure and AER [10, 11].

The gold-standard annotated alignments has two kinds of alignment links, a G_S(sure) alignment links set, which contains alignment links that are unambiguous, and a G_P(possible) alignment links set, which contains ambiguous and unambiguous alignment links. G_S is a subset of G_P. If the result of word alignment system is A, then metrics are defined as follows.

a) Precision

$$P = \frac{|A \cap GP|}{|A|}. \tag{17}$$

b) Recall

$$R = \frac{|A \cap Gs|}{Gs}. \tag{18}$$

c) F1

$$F1 = \frac{2 * P * R}{P + R}. \tag{19}$$

d) AER (Alignment Error Rate)

$$AER = 1 - \frac{|A \cap GP| + |A \cap Gs|}{|A| + |Gs|}. \tag{20}$$

4.2.3 Evaluation toolkits

For human assessment automatic evaluation, we developed a toolkit.

We used the automatic evaluation toolkit of NIST evaluation mteval-v11a.pl [7], and made some modification:

a) Add support of Unicode;

b) Add support of Chinese and Japanese;

c) Add support of GTM, mWER, mPER, etc.

4.3 Evaluation data

4.3.1 Training data

In the HTRDP machine translation evaluations before 2004, no training data were provided to the participants.

In 2005, a training corpus was provided for the CEMT, ECMT and CEWA tracks. The corpus was collected by the Institute of Computing Technology, CAS and some participants, which contains about 830,000 pairs of sentences.

4.3.2 Test data

In the evaluations before 1998, the test data are man-made sentences which cover certain sets of test points. For example, tens of sentences are used to test the ability of word sentence disambiguation of the system, in each of which there is at least one typical multi-sense word. Other tens of sentences are used to test the ability of parsing of the system, so each of the sentences would contain at least one typical syntax structure ambiguity. Since these sentences are man-made sentences, they are usually quite simple, compared to the real text sentences. The average length of these sentences is 10 to 15 words.

From 2003, we used real text sentences in the machine translation evaluations.

In 2005, the machine evaluation task included six language pairs, as listed in Table 17. For each language pair, the test data contains two types: dialogues and essays. The domain of the essay data is news, and the dialogue data is Olympic related (includes sports, transport, travel, weather, etc.).

Considering that one system will process Chinese, English and Japanese, we adopted Unicode (UTF16, little endian) as the corpus encoding.

The amount of test data is shown in Table 20.

Table 20 Size of test corpus

Language pair		Sum	
		Sentence number	Character/Word number
Chinese	Dialogue	About 460	About 9,400
	Essay	About 490	About 21,000
English	Dialogue	About 450	About 4,700
	Essay	About 490	About 12,000
Japanese	Dialogue	About 460	About 11,000
	Essays	About 490	About 21,000

The test data was collected from the Internet, books and teaching material. Chinese and English data are made by ICT, CAS and the Japanese data are made by the NICT (National institute of Information and Communications Technology, Japan).

4.3.3 Reference data

In 2005's machine translation evaluation, four references are given for each test sentence. We selected 4 experts whose mother tongue is the target language to make the reference translations. They translated the source data individually.

Chinese-Japanese, English-Japanese and Japanese- English reference data are made by the NICT, and the other data are all made by the ICT.

4.4 Results and analysis

In 2005's machines translation evaluation, the human assessment metrics were primary metrics. Human judges accessed machine translation with emphasis on the Adequacy and Fluency. The automatic metrics included NIST and BLEU, but we gave out some other metrics, like GTM, mWER, mPER and ICT.

4.4.1 Results

4.4.1.1 Human assessment evaluation

In 2005's machine translation evaluation, a subset of sentences were selected for human assessment evaluation per track; it's about 20 to 40% of all the test data. As shown in Table 21, the scale of human assessment evaluation can satisfy the request.

Table 21 Work of human assessment evaluation

Language pair	System number	Sentence number	Work time (hours)
Chinese-English	8	200	12
English-Chinese	6	300	11
Japanese-Chinese	2	400	7

The selected sentences are given to 4 experts, the sequence of answers of test systems are shuffled. Each expert gives out adequacy score and fluency score for each sentence and system.

4.4.1.2 Results

(1) Chinese-English
  (a) Dialogue (Table 22)
  (b) Essay (Table 23)
(2) English-Chinese
  (a) Dialogue (Table 24)
  (b) Essay (Table 25)
(3) Chinese-Japanese
  (a) Dialogue (Table 26)
  (b) Essay (Table 27)
(4) Japanese-Chinese
  (a) Dialogue (Table 28)
  (b) Essay (Table 29)
(5) Japanese-English
  (a) Dialogue (Table 30)
  (b) Essay (Table 31)
(6) English-Japanese
  (a) Dialogue (Table 32)
  (b) Essay (Table 33)

Table 22 Results of Chinese-English dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.1392	0.2506	0.7158	0.6192	0.4843	0.4091	65.38	64.25
System 2	6.2097	0.1747	0.6677	0.6717	0.5351	0.3357	57.42	52.49
System 3	5.7794	0.1524	0.6277	0.6942	0.5602	0.3197	51.56	47.06
System 4	5.8981	0.1544	0.6472	0.6881	0.5485	0.3155	56.96	53.72
System 5	5.5226	0.1454	0.5795	0.7357	0.6078	0.3509	53.41	51.59
System 6	5.9216	0.1814	0.6478	0.7134	0.5514	0.3518	50.42	57.16
System 7	6.0509	0.1714	0.6161	0.7175	0.5813	0.3589	55.58	55.02
System 8	4.2273	0.0710	0.5179	0.7683	0.6437	0.224	39.74	33.02

Table 23 Results of Chinese-English essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	6.9015	0.1843	0.7053	0.7228	0.5337	0.2343	61.72	55.90
System 2	6.2120	0.1361	0.6452	0.7560	0.5727	0.2090	53.97	47.28
System 3	5.3211	0.1073	0.5946	0.7860	0.6121	0.1743	43.90	38.72
System 4	5.9200	0.1287	0.6645	0.7612	0.5702	0.1851	52.81	46.97
System 5	4.9876	0.0718	0.5268	0.8412	0.6729	0.1863	41.23	32.30
System 6	5.7906	0.1188	0.6463	0.8307	0.5936	0.2087	37.33	39.33
System 7	5.5237	0.1056	0.5745	0.8077	0.6297	0.1926	40.65	36.08
System 8	4.1341	0.0550	0.4944	0.8385	0.6946	0.1292	36.52	30.31

Table 24 Results of English-Chinese dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.6444	0.3506	0.7302	0.5631	0.4261	0.4581	78.01	71.61
System 2	6.6385	0.2657	0.6917	0.6129	0.4644	0.3690	70.41	64.47
System 3	7.0142	0.2958	0.7096	0.5914	0.4535	0.4123	73.55	67.36
System 4	7.8703	0.3776	0.7470	0.5321	0.4156	0.4677	82.59	78.24
System 9	5.6119	0.2063	0.5972	0.6795	0.5651	0.2563	74.64	69.17
System 10	6.8419	0.2913	0.7135	0.5853	0.4529	0.3912	73.62	68.16

Table 25 Results of English-Chinese essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	8.4334	0.3447	0.7537	0.6544	0.4170	0.3051	51.24	43.57
System 2	8.2600	0.3246	0.7629	0.6519	0.4191	0.2834	51.22	42.47
System 3	7.7755	0.2876	0.7333	0.6840	0.4435	0.2632	47.05	37.95
System 4	8.7453	0.3709	0.7930	0.6162	0.3934	0.3137	55.78	47.85
System 9	5.8304	0.1804	0.6205	0.7523	0.5581	0.1267	43.16	33.90
System 10	6.6745	0.2281	0.6998	0.7236	0.4946	0.1959	41.16	31.45

Table 26 Results of Chinese-Japanese dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.1158	0.3512	0.7792	0.6483	0.4421	0.3197	53.44	44.87
System 2	6.9879	0.3069	0.7637	0.7071	0.4771	0.2782	47.56	37.28
System 3	Absent							

Table 27 Results of Chinese-Japanese essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.6785	0.3036	0.7851	0.6904	0.4363	0.2321	42.37	33.02
System 2	8.5858	0.3750	0.8265	0.6450	0.3886	0.2788	44.74	35.29
System 3	Absent							

Table 28 Results of Japanese-Chinese dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.7098	0.3302	0.7302	0.6030	0.4430	0.4767	67.94	67.03
System 3	Absent							
System 11	6.3052	0.2292	0.6656	0.6626	0.5019	0.3781	58.44	56.88

Table 29 Results of Japanese-Chinese essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 1	7.9797	0.3007	0.7170	0.6748	0.4636	0.3249	50.41	44.58
System 3	Absent							
System 11	6.7836	0.2277	0.6862	0.7066	0.4969	0.2591	43.84	37.00

Table 30 Results of Japanese-English dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 3	Absent							
System 12	5.3656	0.1529	0.5878	0.7392	0.5983	0.3495	65.81	52.77

Table 31 Results of Japanese-English essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 3	Absent							
System 12	5.5193	0.1309	0.6139	0.8213	0.5984	0.2295	55.58	38.09

Table 32 Results of English-Japanese dialogue

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 3	Absent							
System 12	8.0045	0.4875	0.7934	0.5320	0.3818	0.4314	63.31	46.69
System 13	7.1239	0.3915	0.7663	0.5995	0.4377	0.3562	63.80	46.81

Table 33 Results of English-Japanese essay

ID	NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
System 3	Absent							
System 12	9.1112	0.4581	0.8167	0.6406	0.3766	0.3071	45.70	29.25
System 13	8.6910	0.4464	0.8223	0.6463	0.3938	0.2831	44.93	27.66

(7) Word Alignment (Chinese-English)

Table 34 Results of word alignment

ID	Precision	Recall	F measure	AER
System 3	0.4993	0.5186	0.5088	0.4918
System 5	0.8087	0.7220	0.7629	0.2348

4.4.2 Analyses

4.4.2.1 System type and performance

In 2005, there were 21 participating systems, in which the 19 systems were rule-based machine translation (RBMT) systems, one system was an example-based machine translation (EBMT) system, and the other system was a statistical machine translation (SMT) system. The maximum and the average BLEU scores of each type are given in Fig. 5.

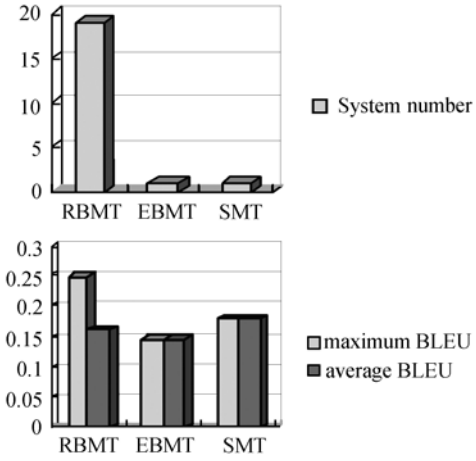

Fig. 5 System type and score

As shown in Fig. 5, we can find that, the RBMT is the main technology used in China currently. The best system is a RBMT system. However, as we can see, the SMT gains a good score above average. Considering recent years' trend, we think more SMT systems will appear and will get better results in the near future.

4.4.2.2 Relation of data type and score

All the participators' systems used the same translation engine for dialogue and essay. From Fig. 6, we can see that all of them perform better in dialog data than in essay data. The reason may be that dialogue sentences are shorter than the essay sentences. So we conduct another experiment to show the relevance between the length of sentence and different metrics.

We computed the relation of sentence length and the scores as

$$\text{correl}(x, y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2 \sum_{i=1}^{n}(y-\bar{y})^2}}. \quad (21)$$

A positive correlation of x and y means that x is large and y is large too, and a negative correlation means that x is large and y is small. A larger correlation (positive or negative) shows stronger relationship between the two variables, and a value near 0 means they have little relation.

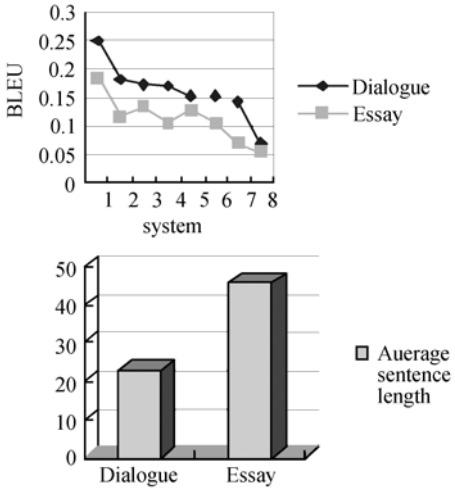

Fig. 6 Comparison of dialogue and essay

Table 35 Relation of metrics and sentence length

NIST	BLEU	GTM	mWER	mPER	ICT	Adequacy	Fluency
−0.06	−0.148	0.031	0.368	0.213	−0.4463	−0.316	−0.451

As shown in Table 35, NIST, BLEU and GTM are almost length irrelative, but the ICT, mWER, mPER and the human assessment metrics are length relative. The ICT and human assessment are negatively correlated with the sentence length, which means the longer sentence will receive a lower score, but the mWER and mPER is positively correlated with the sentence length. This is because the mWER and mPER are error rate metrics, which means that the higher the score is, the poorer the translation quality is.

What should be pointed out is that, in the above experiment, the NIST, BLEU and GTM are computed on single sentence, to see the correlation between the metrics and the sentence length. However, normally they are computed on the whole test data. This makes some difference, but we think it does not have a significant effect on the conclusion.

4.4.2.3 Relation of automatic metrics

In recent evaluation, we used several evaluation metrics. Here we computed the correlation between these metrics.

Table 36 lists 6 metrics' correlation by final score, i.e. the score on whole test set. In this table, we can find that all these metrics are strongly positively correlated (mWER, mPER are negative with others since they are error rate metrics), it means a system with a higher BLEU score should or will also have a higher NIST score, a higher GTM score, etc.

Table 36 Correlation of automatic metrics by final score

	NIST	BLEU	GTM	mWER	mPER
BLEU	**0.9818**				
GTM	0.9576	0.9348			
mWER	−0.9203	−0.8898	−0.9517		
mPER	−0.9355	−0.9201	**−0.9913**	**0.9734**	
ICT	0.9275	0.9356	0.8123	−0.7288	−0.7633

4.4.2.4 Relation of human assessment metrics and automatic metrics

As shown in Table 37, BLEU has both highest correlation with fluency and adequacy, and mWER also a highest with adequacy in average.

4.4.2.5 Relation of two human assessment metrics

In Table 38, the very high correlation shows that in this evaluation, the two human assessment metrics, adequacy and fluency, have a very high correlation, especially in the English-Chinese tasks. Due to the very high correlation between fluency and adequacy in the English-Chinese tasks, it seems that for evaluation of those machine translation systems whose target language is Chinese, it is unnecessary to use the two human assessment metrics. Maybe one of them would be enough.

4.4.2.6 Evaluations based on words and based on characters

In automatic metrics, the computing of score is based on a unit. The more units match with the reference, the higher the score will be. The unit used in English is a word. But, in Chinese and Japanese, there are different choices. The unit may be a word or a character. Because of the ambiguity of the word segmentation in Chinese and Japanese, we use the automatic metrics based on the characters in the HTRDP evaluation, rather than words. However, we would like to know if there is any difference between these two methods. Here we give the result of an experiment to show the difference between the scores based on characters and those based on words. In this experiment, we segmented the Chinese sentence using ICTCAS, a Chinese word segmentation and tagging tool developed by the Institute of Computing Technology, Chinese Academy of Sciences [7], and then computed the automatic metrics based on the segmented sentences. Finally we got the correlation between these two scores of all the automatic metrics, as shown in Table 39.

We can see that the correlation of the two methods is very high. The score by character could represent the performance of the translation system well. Due to the character based method avoid the effect of uncertain of word segmentation and has very low performance loss, the character based method is reasonable.

Table 40 gives the BLEU scores of the Chinese-English machine translation (based on words) and the English-Chinese machine translation (based on words and characters).

As results shown in the table, the BLEU score of English-Chinese based on characters is much greater than that based on word. This is because the character is a smaller unit than word. For the same cooccurrence string between the output data and the reference data, there are longer n-grams based on chararcters than those based on word.

There is another interesting issue. Some people have suggested that Chinese-English machine translation is much more difficult than English-Chinese machine translation, because there are no strict grammatical restrictions in the Chinese language. Is this true? This question is quite difficult to answer, since we cannot compare these two kinds of system in the same test set. However, in the HTRDP evaluation, since the test set for these two translation directions are selected in the same conditions and have the same data size, we think that the results are somewhat comparable. As we know, it is unfair to compare the word-based BLEU score of the Chinese-English machine translation and the character-based BLEU score of the English-Chinese machine translation. In this experiment, we also gave the word-based BLEU scores for English-Chinese machine translation systems. From the above table, we can see that the average BLEU score of the English-Chinese machine translations is higher than that of the Chinese-English machine translations by 3 percent, when both the BLEU scores are computed based on words. It seems to give support to the suggestion.

4.5 Conclusion

From 1994 to 2005, the HTRDP had conducted 6 machine translation evaluations. Seven evaluation tracks concerning translations between the Chinese and English, Japanese and etc., are included in the evaluations. There is also a Chinese-English word alignment track in 2005. Both human assessment and automatic evaluation metrics are used in HTRDP machine translation evaluations. For automatic evaluation metrics, the popular metrics such as BLEU, NIST, mWER, mPER, GTM are used. There is also a new metric

Table 37 Relation of assessment metrics and automatic metrics

		NIST	BLEU	GTM	mWER	mPER	ICT
CE Dialogue	Adequacy	**0.9556**	0.9092	0.8885	–0.8982	–0.8663	0.8814
	Fluency	0.9420	0.9525	0.8787	–0.7725	–0.8346	**0.9609**
CE Essay	Adequacy	0.8280	0.8548	0.7753	**–0.9589**	–0.8230	0.6450
	Fluency	0.9369	**0.9688**	0.9380	–0.9458	–0.9593	0.7658
EC Dialogue	Adequacy	0.6477	**0.7370**	0.4397	–0.6240	–0.4283	0.5501
	Fluency	0.6073	**0.7021**	0.4113	–0.6013	–0.3953	0.5057
EC Essay	Adequacy	0.9072	0.9327	0.8416	**–0.9531**	–0.8715	0.8665
	Fluency	0.9048	0.9330	0.8309	**–0.9468**	–0.8655	0.8659
Average	Adequacy	0.8346	**0.8585**	0.7362	**–0.8585**	–0.7472	0.7357
	Fluency	0.8477	**0.8891**	0.7647	–0.8166	–0.7636	0.7745

Table 38 Relation of two human assessment metrics

	Chinese-English	English-Chinese
Dialogue	0.8956	0.9922
Essay	0.9374	0.9982

Table 39 Correlation of character-based scores and word-based scores

	NIST	BLEU	GTM	mWER	mPER	ICT
Correlations	0.9981	0.9969	0.9952	0.9876	0.9953	0.9964

Table 40 BLEU score of Chinese-English and English-Chinese

	Max.	Min.	Ave.
Chinese-English (word-based)	0.2506	0.071	0.1627
English-Chinese (character-based)	0.3776	0.2063	0.2979
English-Chinese (word-based)	0.2614	0.1079	0.1903

ICT used in 2005, which is proposed by the organizer.

The RBMT systems perform very well in the HTRDP machine translations evaluation. For the SMT, the recent hot research area in the world, although there is only one SMT system that participated in the 2005 evaluation, it got a fairly good result. We can see that there are more SMT researches and some progress appearing in recent years in China. They are expected to get better results in future evaluations.

5 Automatic speech recognition

5.1 Introduction

In the field of automatic speech recognition, it is well accepted that evaluations play an important role in improving the techniques. From the middle of the 1980s, the NIST started a series of evaluations covering a wide range of speech recognition tasks, such as recognition of reading speech, spontaneous speech, dialogue, broadcasting news, etc. [19] Through more than 20 years, the NIST has proposed various evaluation methods and established a standard procedure of evaluation, and has boosted the development of speech techniques considerably. In Europe, the evaluation of speech recognition systems is a part of the TC-STAR project [20], which aims to assess the speech recognition techniques used in speech-to-speech translation.

The HTRDP automatic speech recognition (ASR) evaluation is more like the NIST evaluations, aiming to evaluate the performances of the ASR systems in a wide range of tasks. The first ASR evaluation was launched in 1991, when the first HTRDP evaluation campaign began. Ever since then, it remained a very important part of the HTRDP evaluations and was included in all the following evaluations from 1992 to 2005.

In earlier years, there was usually only one task in each evaluation, such as isolated word recognition or continuous speech recognition. In recent years, many subtasks were defined in one evaluation to test speech recognition systems for different tasks with speech in different languages and collected through the different channels. The subtasks in the ASR evaluations from 2003 to 2005 are shown in Table. There are four major tasks involved: syllable recognition, large vocabulary continuous speech recognition (LVCSR), key word spotting (KWS) and isolated word (command) recognition. The collection of the test data and running of the systems are both done on the PC for the former two tasks, while the KWS systems are also run on the PC but with the test speech collected through the telephone channel, and the command recognition subtask is required to run on a PDA with the test data recorded using the same device. In 2003 and 2005, all subtasks are defined as recognition of the Chinese Mandarin speech, while the evaluation in 2004 also included LVCSR subtasks for English. To encourage new techniques which are time-consuming, the LVCSR task are further divided into 2-times-realtime and 20-times-realtime subtasks, and other subtasks are all required to be fulfilled under 2-times-realtime restriction.

Table 41 The different subtasks carried out in evaluations from 2003 to 2005

Subtasks			2003	2004	2005
Syllable Recognition			√		
LVCSR on PC	Chinese	2X real-time		√	√
		20X real-time	√	√	√
	English	2X real-time		√	
		20X real-time		√	
KWS for telephone speech				√	√
Command recognition on PDA				√	

All subtasks in each evaluation are determined according to the state and requirement of the applications after discussion of the researchers in the related fields. For example, the KWS and the command recognition tasks are introduced to the evaluations because of its wide application in the real world.

Not only the subtasks, but also the evaluation procedures and methods have been changing during the years. The ASR evaluations before 2005 adopted an on-site evaluation method, which means that all the participating sites should take their systems to the specified evaluation spot and run the systems in the specified time period. From the eighth evaluation in 2005, the online evaluation method was used and all test data and recognition results were transferred through the Internet. This lowered the cost, avoided the problem of malfunction of systems, and encouraged more researchers, especially oversea researchers, to take part in the evaluation. Since training and adaptive data play an important role in the ASR systems, the development data have been provided from 2004 and training data were given in the evaluation of 2005 to help the researchers adapt their systems. From 2004, in the LVCSR subtask, word error rate (WER) has been adopted as major evaluation metric instead of word accuracy used in former evaluations, because WER is more suitable and popular among researchers. In 2005, the DET curve [21] was used in the KWS subtask instead of the recognition precision used in 2004, because it is a better assessment of the overall performance.

Though the HTRDP ASR evaluations are similar to those of the NIST, there are still differences. The NIST evaluations emphasize algorithms and techniques, while the HTRDP evaluations focus on both techniques and applications. For example, in the HTRDP evaluations, test data are collected in various noisy environments instead of adding noise to speech, and tasks on the PDA and other application are included.

It is well accepted among participants that by providing comparison and discussion of different systems for the same subtask, the HTRDP ASR evaluations have been very important in facilitating communication of researchers and accelerating progress. Fig. 7 gives the character accuracies of three sites for the Chinese LVCSR 20X subtask in 2003 and 2004. From the figure, we can see that great improvement had been made within one year and the recognition accuracies of the three systems have increased considerably.

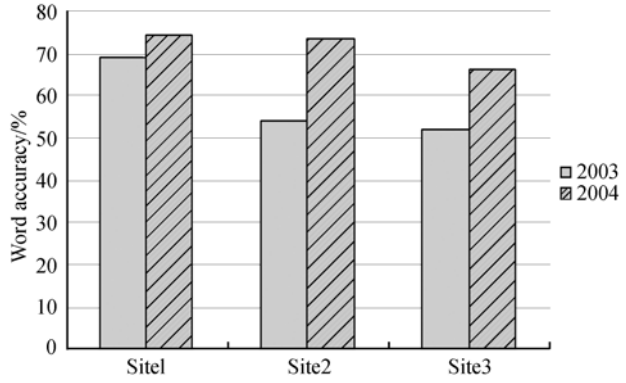

Fig. 7 Character accuracy of systems from three sites for the Chinese LVCSR 20X subtask in 2003 and 2004

In the next chapters, details of the ASR evaluations are presented, with the 2004 evaluation detailed as an example.

5.2 Evaluation methods

As mentioned in 5.1, the ASR evaluation of 2004 consisted of the three major tasks: continuous speech recognition on PC desktop (PC engine), keyword recognition for continuous telephone speech (phone engine), and command word recognition on embedded devices (embedded-device engine). The PC engine task aims at speaker-independent continuous speech recognition on the PC desktop platform with unlimited words. The phone engine task is defined as speaker-independent keyword recognition for continuous telephone speech. The sentences are generated using given word list and grammar. While the embedded-device engine task aims at command word recognition on embedded devices, such as a PDA, with limited command words selected from an unlimited vocabulary. The evaluation metrics for the three tasks are defined as follows, which are popular metrics in similar evaluations.

5.2.1 Evaluation metrics for the PC engine task

The evaluation program uses dynamic programming (DP) to search for a global minimization of the Levenshtein distance (also known as edit distance) between recognition results and the reference to determine the numbers of correct words, substitute words and deleted words. Notice that for Chinese speech, "words" actually means Chinese characters.

We define

$$\#\text{ Reference words} = \#\text{ Correct words} + \#\text{Substituted words} + \#\text{ Deleted words}, \quad (22)$$

$$\#\text{ Output words} = \#\text{ Correct words} + \#\text{Substituted words} + \#\text{ Inserted words}, \quad (23)$$

$$\#\text{ Error words} = \#\text{ Substituted words} + \#\text{Deleted words} + \#\text{ Inserted words}. \quad (24)$$

Based on the above parameters, the five main metrics to evaluate the performance of a speech recognition systems are:

$$\text{Word Error rate} = \frac{\#\text{ Error words}}{\#\text{ Reference words}} \times 100\% \quad (25)$$

$$\text{Error rate of substituted words} = \frac{\#\text{ Substituted words}}{\#\text{ Reference words}} \times 100\%, \quad (26)$$

$$\text{Error rate of inserted words} = \frac{\#\text{ Inserted words}}{\#\text{ Reference words}} \times 100\%, \quad (27)$$

$$\text{Error rate of deleted words} = \frac{\#\text{ Deleted words}}{\#\text{ Reference words}} \times 100\%, \quad (28)$$

$$\text{Correct sentence rate} = \frac{\#\text{ Correct sentences (all words are correct)}}{\#\text{ all sentences}} \times 100\%. \quad (29)$$

5.2.2 Evaluation metrics for the telephone speech task

According to the grammar used in this task, each sentence recorded can be described with a frame, which contains several slots, and each slot can be resolved to a sequence of keywords. The aim of this task is to extract slots from the speech. A slot is recognized correctly if and only if all keywords are recognized correctly. If any insertion errors, substitution errors, or deletion errors occur, the slot will be treated as not recognized correctly. In other words, the result of the recognition of a slot is a string of words, and only if this string is exactly the same as the answer, the slot is labeled as "right". A sentence is labeled "right" only if all slots are recognized correctly. The metrics used for this task are:

$$\text{Correct slots rate} = \frac{\#\text{ Correct slots}}{\#\text{ All slots}} \times 100\%, \quad (30)$$

$$\text{Correct sentences rate} = \frac{\#\text{ Correct sentences}}{\#\text{ All sentences}} \times 100\%. \quad (31)$$

5.2.3 Evaluation metrics for embedded engine

The major metric used is correct command rate, which is defined as follows:

Correct commands rate

$$= \frac{\#\text{ Correct commands}}{\#\text{ Reference commands}} \times 100\%. \quad (32)$$

5.3 Evaluation data

5.3.1 Evaluation data for the PC engine task

Test corpora of the Chinese speech data for the PC engine task are as shown in Table 42. Sixty utterances were provided as a development set one month before the evaluation. The data of the English subtasks are similar. The size of the test set in English is the same with the Chinese's.

Table 42 Corpora of Chinese speech for the PC engine task

Corpora Size	Speakers	Record condition and data storage
200 sentences	10 males and 10 females. Each speaker reads 10 sentences in Chinese mandarin (maybe with slight accent). Each sentence is read only once	Real environment with noise Speech data are sampled at the rate of 16 kHz with 16 bit quantization Each sentence is stored as one wav file

5.3.2 Evaluation data for the telephone speech task

The data of telephone speech task are related to the Olympic Games, which are divided into 5 settings, as shown in Table 43.

Fifty utterances (10 for each setting) will be provided as a development set one month before the evaluation. All sentences are generated according to the given word table and grammar. The word table and grammar are released one month after the release of this plan.

5.3.3 Evaluation data for the embedded task

A total of 600 command words will be used, mostly person names, place names, and operating commands of PDA. Words beyond the range of these 600 words are not included in the test corpora. The test corpora include 600 command words, with each word including 2 to 5 Chinese characters. All the Chinese characters used in the words are within the range of the GB-2312 secondary code table. There are no different words with the same pronunciation. The vocabulary of the test corpora is provided at evaluation spot. All the data are sampled at the rate of 8 kHz with 16 bit Quantization. The utterance of each command word is stored as one .wav file. Ten command words are provided as samples one month before the evaluation.

5.4 Results and analysis

Tables 44–47 gives the results of all the Chinese speech subtasks in evaluation of 2004.

Table 43 Test corpora for the telephone speech task

Domain	Settings	Corpora Size	Speakers	Record condition and data storage
Olympic-oriented Domain	Public transportation information query	40 sentences	10 males and 10 females. 2 males and 2 females for each setting Each speaker reads 20 sentences with Chinese mandarin (maybe with slight accent). Each sentence is read only once	Real environmental noise Speech data are sampled at the rate of 8 kHz with 16bit quantization Each sentence speech is stored as one wav file
	Weather forecast query	40 sentences		
	Travel information query	40 sentences		
	Catering information query	40 sentences		
	Sports game information query	40 sentences		

Table 44 Results of the PC engine Chinese speech 2X task in 2004

	System1	System 3	System 5	System 6	System 7
Error rate of substituted words	56.4%	27.7%	36.3%	24.5%	33.5%
Error rate of inserted words	4.3%	0.7%	2.5%	0.2%	3.0%
Error rate of deleted words	3.2%	2.4%	2.7%	3.9%	1.6%
Word error rate	63.9%	30.8%	41.5%	28.6%	38.1%
Correct sentence rate	0.0%	5.5%	3.0%	7.5%	4.0%

Table 45 Results of the PC engine Chinese speech 2X task in evaluation of 2004

	System 1	System 3	System 4	System 5	System 6	System 7
Error rate of substituted words	55.5%	23.5%	70.5%	33.3%	22.7%	29.3%
Error rate of inserted words	4.8%	0.8%	2.9%	1.3%	0.3%	2.8%
Error rate of deleted words	2.5%	1.5%	6.3%	2.9%	3.4%	1.8%
Word error rate	62.9%	25.8%	79.7%	37.5%	26.4%	33.9%
Correct sentence rate	0.0%	7.0%	0.0%	3.5%	10.0%	6.0%

Table 46 Results for the telephone speech task in evaluation of 2004

		System1	System 3	System 4	System 5	System 6
Correct slots rate	Transportation	36.2%	61.7%	53.9%	54.6%	47.5%
	Catering	25.8%	80.9%	36.0%	59.6%	69.7%
	Sports	21.3%	30.3%	17.2%	27.9%	14.8%
	Travel	33.3%	73.8%	53.6%	71.4%	70.2%
	Weather	15.0%	77.2%	9.2%	62.1%	46.6%
	Total	26.7%	62.2%	34.0%	53.4%	47.1%
Correct sentence rate	Transportation	15.0%	45.0%	32.5%	40.0%	27.5%
	Catering	15.0%	72.5%	17.5%	42.5%	57.5%
	Sports	2.5%	25.0%	2.5%	15.0%	5.0%
	Travel	25.0%	57.5%	30.0%	57.5%	55.0%
	Weather	2.5%	65.0%	2.5%	50.0%	37.5%
	Total	12.0%	53.0%	17.0%	41.0%	36.5%

Table 47 Results for the embedded-device task in evaluation of 2004

	System 1	System 2	System 3	System 4	System 5	System 6
Correct commands rate	1.3%	54.2%	73.2%	9.0%	67.5%	72.2%

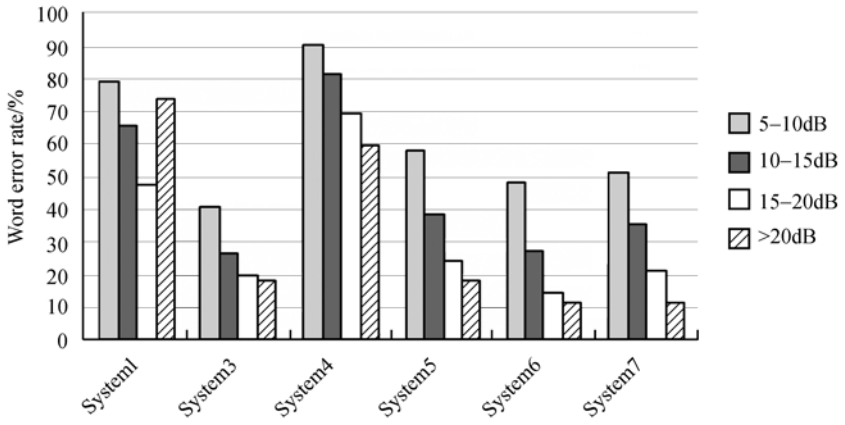

Fig. 8 Word error rates for data of different SNRs of the 20X systems

The tables show that for the PC engine task for Chinese speech, the best word (character) error rate is 28.6% and 26.4% for 2-times-realtime and 20-times-realtime systems respectively. The main reason of the relatively high word error rate is that the test data are collected in a noisy environment and the signal noise ratios (SNRs) are relatively low (less than 15dB) for most data. Fig. 8 shows the word error rates for data of different SNRs of the 20X systems, indicating that error rate falls considerably when SNR increases.

For the telephone speech task, the best system achieved a correct slots rate of 62.2% for the total test set. The challenge lies in the difficulty of exploring the grammar in addition to acoustic recognition. It can be seen from Table 46 that the correct slot rates vary among the different settings. That is due to the different degree of complexity of the grammar defined for the different settings.

For the command recognition on the embedded devices task, the highest correct rate is 73.2%, with the major challenge of a low SNRs.

5.5 Conclusion

The HTRDP ASR evaluation is one of the first conducted categories of the HTRDR evaluations. It covers a wide range of tasks, focusing on both technique and application. Through the years, the performances of the systems increased considerably due to the comparison and communication brought by the evaluations.

Since tasks or test data are different from that of other international evaluations such as the NIST evaluation, the results are not comparable. But from the results of 2004, we can see that performances of participant systems are quite good considering the challenges and difficulties. But they are obviously not satisfying for real application, with the

biggest challenge of the variety and diversity of the speech and recording conditions, such as kind of environmental noise, SNR, and the style of speaking.

6 Text to speech

To our best knowledge, the public evaluation campaigns for text to speech (TTS) organized by a third party are not easily available. The main reason for this is due to the subjectivity of the TTS evaluation. This subjectivity will cause many other problems. For example, the cost to train listeners, to design and prepare test materials is both quite high. The second problem is fairness. Since the outputs of the different systems will be put together to listeners. It is possible that the listeners will learn from better outputs and give unreasonably high scores for worse outputs. We call it the learning effect, which will be discussed further in the following sections.

There are two famous TTS evaluation campaigns that were organized internationally. The first one was organized by the third ESCA TTS Workshop [22]. However, this evaluation was not formal because all the listeners were the participants themselves and the experimental procedure was short. In this evaluation, three types of different texts were designed: newspaper text, semantically unpredictable sentences and telephone directory listings. To avoid the learning effects and therefore bias, the listening experiments were carefully designed. There were about 70 systems for the total 19 different languages which participated in this campaign. But the final results were not public.

The second one was organized in the past several years by the TC-STAR Workshop on Speech-to-Speech Translation [23]. The evaluation included several different subtasks, such as the prosody evaluation and full system evaluation. Since the evaluation was organized with the other two evaluations—spoken language translation (SLT) and automatic speech recognition (ASR)—texts from the outputs of ASR and SLT modules were also used as inputs for the TTS systems other than clean text inputs.

Compared with these two evaluations, the TTS evaluations organized by HTRDP have two major differences and specialties:

1) Pair comparison, instead of the traditional 5-scale Mean Opinion Score, is used. We believe that pair comparison is more objective and therefore can reduce learning effects greatly.

2) Test materials from special domains are used. The special domains contain very different vocabularies. This will force the TTS systems on the applications in special domains, for instance, domains related to the Olympic Games.

6.1 Introduction

The first evaluation campaign of the text to speech category under the HTRDP was launched in 1994, although other category evaluations began in early 1991. At that time, there were 9 systems from 4 sites participating in the evaluation campaign. Since then, the text to speech (TTS) evaluation category was included as a tradition in HTRDP evaluations for five times: year 1994, year 1995, year 1998, as well as the two recent years 2003 and 2005.

There are several different subtasks in each TTS evaluation. Table 48 shows the different subtasks carried out in each evaluation. In the first TTS evaluation, a word-level and a sentence-level naturalness were evaluated using the MOS measure on words and sentences. However, the experiments show that paragraph texts that contain diverse types of sentences are better as a test corpus for naturalness rather than individual and independent sentences. Therefore only paragraph level naturalness test were carried out in later evaluations.

Table 48 The different subtasks carried out in each evaluation.

Subtasks	1994	1995	1998	2003	2004
Syllable articulation	√	√	√		
Word level intelligibility	√	√	√	√	
Sentence level intelligibility	√	√	√	√	√
Naturalness (word/sentence/paragraph-level)	√	√	√	√	√
Linguistic test		√	√	√	
Systematic evaluation	√	√	√		

The method for naturalness assessment was also changed from MOS to pair comparison which will be described in section 6.2. Generally speaking, the MOS (Mean Opinion Score) standard is difficult to be operated by listeners during the listening test. Additionally, the MOS is also subject to learning effects. Therefore, in the recent two years 2003 and 2004, the pair comparison which is designed to avoid learning effects was used as the major measure for naturalness evaluation instead of the MOS.

For linguistic test, we focus on the evaluation of word segmentation, recognition of polyphones, numbers, special symbols, measurement units and so on. This subtask was designed to test the TTS systems and its capability of processing complicated linguistic problems before converting text into speech. Table 49 shows that the average accuracy of the linguistic test improved from 74.3 to 79.2. That is why the linguistic test was not implemented once again in 2004.

Table 49 The average accuracy of linguistic test from 1995 to 2003

	1995	1998	2003
Average accuracy of Linguistic test	74.3	76.8	79.2

In the first several years, the systematic performance besides the quality of synthesized speech was also assessed. This kind of systematic evaluation included system installation, required resource, and system robustness and so on. However, this evaluation was optional.

For the intelligibility test, great improvement was gained in each evaluation. Systems have improved in intelligibility by using advanced technologies and there were no significant differences between systems in this test. Therefore the

syllable articulation and system word level intelligibility test were cancelled gradually.

Table 50 shows the average sentence level intelligibility accuracy for all evaluations. From the table, we can see that the performance was greatly improved all through except for the year 2004. This indicates that in the past several years, the technologies used indeed improve the intelligibility on one hand and that the current method for sentence level intelligibility test was not suitable on the other hand. Before 2004, meaningful and syntactic sentences were used to test the intelligibility. However, in 2004, semantically unpredictable sentences were used and that is why the average score for this year has decreased compared with that of the previous year.

Table 50 The average accuracy of sentence level intelligibility

	1994	1995	1998	2003	2004
Average accuracy of sentence level intelligibility	77.4	83.7	84	96.4	94.6

The major differences between the evaluations in 2003 and 2004 and the evaluations before 2003 are:

1) both the desktop and embedded with the TTS systems were evaluated while only desktop systems were evaluated before 2003;

2) The test corpora used for evaluation covered both the general domain and Olympic-oriented domain (including sports reports, weather forecast, urban transportation, travel and food).

All these evaluation expansions were made because there appeared more and more embedded TTS systems for research and the industrial applications and also because there is an urgent need to make high-quality TTS systems in the special domain—Olympic-oriented domain—available due to the 2008 Olympic Games in Beijing.

From all the TTS evaluations, the quality of the synthesized speech and the performance of the TTS systems are greatly improved. This is due to the advanced technologies used. However, the evaluation technologies are also being developed to satisfy the development of the TTS, to really reflect the progress and to lead the research in this field.

In the next sections, the evaluation of year 2004 will be described in detail. The evaluation methods, the evaluation data, as well as results and analysis are all presented.

6.2 Evaluation methods

There were two subtasks carried out in 2004 TTS evaluation: sentence level intelligibility test and paragraph level naturalness test.

6.2.1 Sentence level intelligibility test

In the sentence level Intelligibility test, ***Semantically Unpredictable Sentences (SUS)*** [22] are used. Each sentence contains one or two blanks which may be content words, function words or phrases. Listeners are asked to fill the blanks according to what they've heard. The intelligibility is measured with the percentage of correct response.

6.2.2 Paragraph level naturalness test

The ***Pair Comparison*** method is used for the naturalness evaluation. All of the results from the participants will be compared in pairs. In each pair, each side will be marked as A or B. The listeners are asked to make the judgment of which they prefer between A and B according to the rhythm and prosodic fluency, and give the score in five levels, -2, -1, 0, +1, +2 (see Table 51). The results in each comparison pair will be played only once in random order, AB or BA. The sentences played for testing are also ordered in random list. To reduce learning effects and get a fairer evaluation result, the speech outputs from each participant will have equal opportunities for playing, listening and testing. For example, if participant P1 is compared with participant P2, there will be half amount of the sentences played in order P1 and P2, and the rests are played in inverse order. The final statistical analysis based on the recording results will offer rating, confidence interval of 95%, and standard deviation.

Table 51 The scale of score for pair comparison

Comparison pair	Score
A is better than B	+2
A is slightly better than B	+1
A is the same as B	0
A is slightly worse than B	−1
A is worse than B	−2

6.3 Evaluation data

According to the assessment methods listed above, there are two types of corpora: sentence corpus (for sentence level intelligibility test) and paragraph corpus (for paragraph level naturalness test). The corpora are for both the general domain test and the Olympic-oriented test.

6.3.1 Sentence corpus for general domain and Olympic-oriented test

The corpus involves semantically unpredictable sentences of length at most 15 characters. These sentences are automatically generated using different, common syntactic structures with words randomly selected from a lexicon of frequency tables. The syntactic structures are shown as follows:

Subject - Verb - Direct object
Subject -Intransitive Verb
Subject -Transitive verb -Direct object – Indirect object
Sentences with Q-words
Sentences with 把(BA)
Sentences with 被(BEI)

Words for the general domain are selected from the word list with different part of speech (POS) tags. Table shows the general domain word list by POS tags of words. Words for the Olympic-oriented domain are randomly selected from the word list which is related to the Olympic Games and contains a high frequency of words from sports reports,

weather broadcast, urban transportation, travel and food. Table 53 shows the Olympic-oriented domain word list by their different sub-domains.

Table 52 The word list of general domain

	Noun	Verb	Adj.	Conj.	Num.	Pron.	Prep.	Quan.
Word Num.	981	991	860	7	19	16	8	86

Table 53 The word list of Olympic-oriented domain

	weather broadcast	food	sports reports	urban transportation	Travel
Word Num.	60	103	155	485	102

6.3.2 Paragraph corpus for general test

The paragraphs used are news reports, stories, essays, comments and so on. Some English words used frequently in Chinese might be embedded in the texts for assessment. The lengths of the paragraphs are no more than 200 characters/words.

6.3.3 Paragraph corpus for Olympic-oriented test

The paragraph will come from sports report, transportation, travel, weather broadcast, food and so on. It might be simple dialogues including three to five questions and answers, greetings, fragments of information service and so on. The length of the paragraph will be no more than 200 characters. The sizes of each corpus will be announced during the test.

6.4 Results and analysis

Table 54 and Table 55 show the accuracy of the sentence level intelligibility test using semantically unpredictable sentences for the desktop systems and the embedded systems, respectively. Since the embedded TTS systems were restricted by the limited resources, RAM and CPU, they obtain lower performance compared with desktop systems. However, these results were still promising compared with the desktop results of sentence level intelligibility (SLI) test in 1998. Compared with the scores of the SLI test in 2003, the scores this year were slightly lower. This is because SUS sentences instead of meaningful sentences were used for this test. The SUS sentences make listeners harder to subconsciously predict the next words according to the previous contexts.

Table 54 The accuracy of sentence level intelligibility test for desktop systems

	System 1	System2	System 3	System 4
SUS accuracy (%)	96.7	94.5	93.9	93.4

Table 55 The accuracy of sentence level intelligibility test for embedded systems

	System 1	System 5	System 6
SUS accuracy (%)	88.7	86.5	87.9

Table 56 and Table 57 show the results of naturalness test by pair comparison for desktop and embedded systems, respectively. In these two tables, all the possible comparison pairs (A, B) are listed in the first column (A) and second column (B). The average ratings recording how the listeners prefer A to B are shown in the third column. The remaining two columns show the standard deviation and confidence interval of 95%, respectively.

Table 56 The results of naturalness test for desktop systems

systems	Systems to be compared with	Average score	Standard deviation	95% Confidence interval
System 1	System 3	1.61328	0.25538	[1.4772, 1.74936]
	System 2	1.19922	0.53775	[0.91267, 1.48577]
	System 4	1.24609	0.39822	[1.0339, 1.45829]
System 3	System 1	−1.61328	0.25538	[−1.74936, −1.4772]
	System 2	−0.49219	0.28401	[−0.64353, −0.34085]
	System 4	−0.42578	0.26143	[−0.56508, −0.28648]
System 2	System 1	−1.19922	0.53775	[−1.48577, −0.91267]
	System 3	0.49219	0.28401	[0.34085, 0.64353]
	System 4	−0.04297	0.30424	[−0.20509, 0.11915]
System 4	System 1	−1.24609	0.39822	[−1.45829, −1.0339]
	System 3	0.42578	0.26143	[0.28648, 0.56508]
	System 2	0.04297	0.30424	[−0.11915, 0.20509]

Table 57 The results of naturalness test for embedded systems

Systems	Systems to be compared with	Average score	Standard deviation	95% Confidence interval
System 1	System 5	0.96484	0.30188	[0.80399, 1.1257]
	System 6	1.07813	0.32636	[0.90422, 1.25203]
System 5	System 1	−0.96484	0.30188	[−1.1257, −0.80399]
	System 6	0.18359	0.25563	[0.04738, 0.31981]
System 6	System 1	−1.07813	0.32636	[−1.25203, −0.90422]
	System 5	−0.18359	0.25563	[−0.31981, −0.04738]

Compared with MOS, pair comparison cannot provide absolute scores reflecting the listeners' judgment on single systems. However, it seems to be more "objective" than MOS especially with respect to reducing learning effects. Additionally, if TTS buyers want to choose among the different TTS systems, pair comparison can offer direct information for this.

7 Text summarization

The Document Understanding Conference (DUC) [23] is the most important international evaluation in multi-document summarizations. Automatic evaluation metrics, such as ROUGE, are used in DUC evaluations.

The HTRDP text summarization evaluations have been conducted four times in 1995, 1998, 2003 and 2004. Up to now, we only conducted single document summarization evaluation. The subjective assessment method was used in each evaluation. Compared with other evaluation categories, the text summarization evaluation is rather simple. Here we give an introduction to the 2004 text summarization evaluation.

There were two tasks in 2004 evaluation: text summarization and key word extraction.

Twenty documents were given to each participant, and the limits of the length of the summarization were given in the same time. Each participant system was required to generate a summarization and several keywords (no more than 5) for each document.

All the outputs were submitted to 4 human experts. They will give scores to the summarizations and key words generated for all documents by each participating system. The score for summarization is an integer from 1 to 5, and the score for key words is an integer from 1 to 3. The criterion for the scoring on summarization and key words extraction are given in Table 58 and Table 59.

Table 58 Criterion for text summarization evaluation

Score	Description
0	The summarization is not fluent and totally has not reflected the topic of the original article. It is hard to understand what the summarization wants to express
1	The summarization is not fluent and partly reflects the topic of original article. Many key-points of the articles are missed
2	The meaning between sentences are not coherent. The logic is not clear. But the summarization is helpful to understand something about the original article
3	The sentences are fairly fluent. The meaning between the sentences is not consecutive. The summarization partly reflects the topic of the original articles. Most key-points have been summarized
4	The sentences are fluent and coherent. The topic of the original articles is reflected. Only a few key-points are omitted. The logic is reasonable
5	The sentences are fluent and coherence. The topic of the original articles is reflected. No key-points of the articles are omitted. The logic is reasonable. The logic is clear. We can grasp the main idea of original article from the summarization

The final metrics for each participating systems are the sum of its scores on the all 20 documents, as shown in Table 60.

Table 59 Criterion for key word extraction evaluation

Score	Description
0	None of the keywords reflects the topic of the original article
1	Some of the keywords reflect the topic of the original article, but the others do not
2	All the keywords reflect the topic of the original article, but not all important points are covered
3	All the keywords reflect the topic of the original article, and all important points are covered

Table 60 Results of text summarization and key word extraction

System ID	Text summarization	Key word extraction
System1	294.4	190.2
System2	344.4	219.5
System3	322.9	218.4
System4	314.1	148.1
System5	296.3	175.2

8 Text categorization

8.1 Introduction

The purpose of text categorization evaluation is to accelerate research within the text categorization community by providing the infrastructure necessary for a large-scale evaluation of text categorization methodologies and help advance the state of text categorization technology [24, 25]. There are 4 and 9 organizations respectively participated in 2003 and 2004. The evaluations were both on the spot, which means that all participators should gather together with their systems assessed by the operators in identical evaluation environment.

The 1st text categorization evaluation was held in Oct. 28, 2003. The adopted evaluation metrics included the classifier time, micro average of precision, micro average of recall, macro average value of F1, and classifier general score.

The 2nd text categorization evaluation was held in Oct. 19, 2004. The 9 systems were also evaluated on the micro average value, besides the metrics used in the last evaluation. The overall evaluation results were much better than the year before. Specifically, the maximal macro average value of F1 and the maximal classifier general score in the 1st evaluation both are much higher than that of the 2003 evaluation. Moreover, there are 8 of the 9 participants in the 2nd year whose macro average value of F1 and classifier general score both exceeded the best corresponding values of the last evaluation.

The rest of this section will take the 2nd Text Categorization Evaluation as an example to describe the details of the evaluation. Furthermore we will discuss the strength and weakness of this evaluation according to the evaluation results.

8.2 Evaluation Metrics

Performance was measured by MacroF1 (macro average value of F1) [28−30] and classifier general score, which are defined as follows:

(1) P_j—the precision of the jth category

$$P_j = \frac{l_j}{m_j}, \tag{33}$$

where l_j is the number of texts precisely categorized to the jth category, m_j is the number of texts categorized into the jth category by the system.

(2) R_j—the recall of the *jth* category

$$R_j = \frac{l_j}{n_j}, \tag{34}$$

where l_j is the number of texts precisely categorized to the jth category, n_j is the number of texts categorized into the jth category by the expert.

(3) $F1_j$—the F1 of the jth category

$$F1_j = \frac{P_j \times R_j}{P_j + R_j}. \tag{35}$$

(4) Macro P—the macro average of precision

$$\text{Macro}P = \frac{1}{n}\sum_{j=1}^{n} P_j, \tag{36}$$

where n is the number of all categories defined by the expert.

(5) MacroR—the macro average of recall

$$\text{Macro}R = \frac{1}{n}\sum_{j=1}^{n} R_j. \tag{37}$$

(6) MacroF1—the macro average value of F1

$$\text{Macro}F1 = \frac{\text{Macro}P \times \text{Macro}R}{\text{Macro}P + \text{Macro}R}. \tag{38}$$

(7) MicroP—the micro average of precision

$$\text{Micro}P = \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} m_j}. \tag{39}$$

(8) MicroR—the micro average of recall

$$\text{Micro}P = \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} n_j}. \tag{40}$$

(9) MicroF1—the micro average value of F1

$$\text{Micro}F1 = \frac{\text{Micro}P \times \text{Micro}R \times 2}{\text{Micro}P + \text{Micro}R}. \tag{41}$$

(10) Classifier general score

The classifier general score is computed as shown in Table 61. For one document, the first column is the result given by the expert, the second column is an output given by the classifier, and the third column is the corresponding score. Character *A* indicates the first category that the expert gives, and Character *B* is the second category that the expert gives. Character *X* and *Y* indicate the categories generated by the text categorization system, and they are neither identical with *A* nor *B*.

$$\text{Classifier General Score} = \sum_{j=1}^{n} \text{Score}_j, \tag{42}$$

where n is number of files.

Table 61 Score distribution scheme on one document

Expert	Classifier	Score_j
〈A〉	〈A〉	1
	〈A,X〉	0.75
	〈X,A〉	0.5
	〈X〉	0
	〈X,Y〉	0
〈A,B〉	〈A,B〉	1
	〈A,X〉	0.75
	〈A〉	0.75
	〈B, A〉	0.75
	〈B, X〉	0.5
	〈B〉	0.5
	〈X,A〉	0.5
	〈X,B〉	0.3
	〈X,Y〉	0
	〈X〉	0

It should be pointed out that precision, recall and F1 will be computed only for the first result, while classifier general score was computed for both the two results. The time that TC systems take to finish the categorize process is also used to measure the performance.

8.3 Evaluation Data

The Chinese Library Classification (version 4) was adopted to define the predefined categories, which is shown in Table 62. As it is difficult to determine whether a file belongs to Category T or Category Z, both the categories are excluded from the predefined categories. That is to say that there are just 36 predefined categories. Another point that should be pointed out is that the evaluation allows multiple categories, while restricting every system from generating more than two results on one document. It means that one document at most belongs to two categories. The results should be ranked from high to low.

Table 62 Chinese Library Classification (version 4)

A 马列主义、毛泽东思想	B 哲学	C 社会科学总论
D 政治、法律	E 军事	F 经济
G 文化、科学、教育、体育	H 语言、文字	I 文学
J 艺术	K 历史、地理	N 自然科学总论
O 数理科学和化学	P 天文学、地球科学	Q 生物科学
R 医药、卫生	S 农业科学	TB 一般工业技术
TD 矿业工程	TE 石油、天然气工业	TF 冶金工业
TG 金属学、金属工艺	TH 机械、仪表工艺	TJ 武器工业
TK 动力工业	TL 原子能技术	TM 电工技术
TN 无线电电子学、电信技术	TP 自动化技术、计算技术	TQ 化学工业
TS 轻工业、手工业	TU 建筑科学	TV 水利工程
U 交通运输	V 航空、航天	X 环境科学、劳动保护

The evaluation corpus was collected mainly from different Internet resources, including digital libraries, news, journals, magazines, and so on. Finally, 3600 documents were selected to form the evaluation corpus, at an average of 100 documents for each category. (Actually, because of the difference in amount among various categories, it is not exactly 100 documents for each category.) The evaluation documents were chosen according to the following rules:

(1) The data should not depend much on only one data source.

(2) The content of one document should be neither too rich nor too poor, and should be restricted in a reasonable scope.

(3) Overlapping with the evaluation corpus in the 2003 Text Categorization Evaluation should be avoided.

After the evaluation documents were chosen, they were classified into at most two categories manually. At last the questions and standard answers were worked out based on the 2004 Text Categorization Evaluation Plan.

8.4 Evaluation results and analysis

The classifying results were evaluated using an automatic evaluation software, and the detailed evaluation results are shown in Table 63.

Table 63 The evaluation results of 2004 Text Categorization Evaluation

System ID	Classifying Time	the macro average value of F1	the micro average value of F1	Classifier General Score
System1	235s	73.96%	74.44%	2654.75
System2	182s	73.76%	73.72%	2654.75
System3	133s	72.44%	73.28%	2641
System4	3600s	71.04%	71.22%	2561
System5	240s	67.47%	67.58%	2474.05
System6	1925s	68.07%	68.33%	2458.5
System7	402s	66.90%	67.31%	2426.5
System8	900s	66.53%	65.86%	2372
System9	120s	49.69%	46.19%	1671.25

The overall evaluation results were much better than last year. Specifically, the maximal macro average value of F1 and the maximal classifier general score in the 1st evaluation respectively were 61.06% and 2208.75, and the corresponding values in this evaluation reached 73.96% and 2654.75. Moreover, there are 8 of the 9 participants in the 2nd year whose macro average value of F1 and the classifier general score both exceeded the best corresponding values of the last evaluation. These significant improvements were mainly attributed to that we opened the test corpus used in the 1st evaluation as training corpus for the classifier systems in the 2nd evaluation. It is indicated that training corpora play an important in the evaluation through the comparison of the twice evaluations. So we plan to continue increasing the volume of the training corpora, and it is expected that the systems would have an even better performance.

All categorization systems considered, we can see that the evaluation results are related good in some categories, such as A、H、S、TD、TV、U; Whereas the results are also a little weak in some other categories, such as K、N、TB、TH、TQ and TS. This phenomenon is a chief result from of the use of the Chinese Library Classification. A broad scope is concerned in the Chinese Library Classification; however, there are documents of some subcategories that do not agree with the Text Categorization Tasks, for example, the subcategory N2 of the category N. Hence, when collecting the test corpus, the first thing that should be taken into account is to determine the subcategory of the documents in the corresponding category. This would make the test corpus more reasonable.

8.5 Conclusion

It is pointed out that problems such as nonlinearity, skewed data distribution, labeling bottleneck, hierarchical categorization, scalability of algorithms and the categorization of web pages are the key problems to the study of text categorization.

The first and foremost challenge is delivering high accuracy in all applicative contexts. While highly effective classifiers have been produced for applicative domains such as the thematic classification of professionally authored texts (such as newswires), in other domains reported accuracies are far from satisfying. Such applicative contexts include the classification of web pages, spam filtering, and authorship attribution, etc.

A direction is investigating the scalability properties of text classification systems, i.e., whether the systems stand up to the challenge of dealing with a very large number of categories (e.g., in the tens of thousands). The labeling bottleneck, i.e., labeling examples for training, is expensive. There should be an increasing attention in text categorization by semi-supervised machine learning methods, i.e., by methods that can bootstrap off a small set of labeled examples and leverage on unlabeled examples too. However, the problem of learning text classifiers mainly from unlabelled data is still, unfortunately, open.

The current evaluations are supported by the national high technology research and the development program of China (863 program) and fulfill the requirements of the national long-term strategies in Information science. Many key technologies and breakthroughs have been successfully developed during the program. In the current evaluations, four evaluation indexes are used which are the macro average value of F1, the micro average value of F1, classifier general score and classifying time. The first three evaluation indexes are effectiveness indexes while the last one is the efficiency index. Among these four evaluation indexes, the macro average value of F1 and the micro average value of F1 are the general international evaluation indexes. The introduction of the classifier general score can achieve the evaluation objectives with more flexiblity and more accurately. The HTRDP text categorization evaluations promote the national self-determination creative activities in text information Processing.

9 Information retrieval

Information overload on the World-Wide Web (WWW) is a well recognized problem. While existing search engines and information retrieval techniques do a good job of retrieving results in assisting users for the information they are seeking, they often fail to satisfy them. That is to say that there are still necessity and room to improve the performance of the current information retrieval technology. Concentrated information retrieval evaluations are conducted and have been proven to be an efficient means to advance the development of information retrieval technology.

The TREC (Text Retrieval Conference) [31], co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense Advanced Research and Development Activity (ARDA), In the past 14 years, has been the most influential information retrieval evaluation. TREC began with a focus on the evaluation of western languages retrieval, but later, Chinese, Arabic, etc. were also involved. The TREC provides a very large test collection and encourages communications among various research groups in a friendly evaluation forum [32]. In 2005, up to 117 participating groups from 23 different countries participated in the TREC [33]. As a matter of fact, actually Retrieval system effectiveness approximately doubled in the first six years of TREC, and most of today's techniques applied to commercial search engines are first proposed in TREC.

The NTCIR and CLEF are another two important information retrieval evaluations. The NTCIR (NII Test Collection for information retrieval Systems) [34] was started in 1999, co-sponsored by the NACSIS (National Center for Science Information Systems) and the JSPS (Japan Society for the Promotion of Science). The NTCIR workshop is a series of evaluation workshops designed to enhance research in Information Access technologies including Information Retrieval, Question Answering, Text Summarization, Information Extraction, etc. The NTCIR focused especially on the processing of several Asian languages, such as Japanese, Chinese and Korean. The CLEF (The Cross-Language Evaluation Forum) [35] was started in 2000, which has provided evaluation tasks on European languages in both monolingual and cross-language contexts, including monolingual information retrieval, cross-language information retrieval, multilingual information access, domain-specific retrieval, and interactive retrieval.

However, analysis suggests that there are some particular factors that may weaken the performance of Chinese information retrieval, e.g., Chinese Word Segmentation, Chinese new word recognition, Chinese abbreviations, Multi-language query processing such as the mixed query of English and Chinese, etc. These problems have not been well addressed. Considering both the difficulties the information retrieval technology is facing and the characteristic of Chinese information retrieval when designing the evaluation, the HTRDP Chinese information retrieval evaluations were conducted to investigate the current research status and the system validity of the Chinese information retrieval system under the circumstance of mass data of the Web.

The HTRDP evaluation of Chinese Information Retrieval is a series of evaluations designed to enhance research in Chinese Information Retrieval technologies including Web search, passage retrieval, etc. The aim is to encourage research in Chinese information retrieval technologies by providing large-scale test corpora reusable for experiments and a common evaluation infrastructure allowing cross-system comparisons [36].

9.1 Introduction

The objective of HTRDP Chinese information retrieval evaluations is to investigate the current research status and the system validity of the Chinese information retrieval system under the circumstance of mass data of the WEB. The organizers consider both the difficulties the information retrieval technology is facing and the characteristic of Chinese information retrieval when designing the evaluation [37].

The HTRDP evaluation of Chinese information retrieval usually provides test corpora (data sets usable for experiments) and unified evaluation procedures for experiment results. Each participating group conducts research and experiments using the common data provided by the information retrieval evaluation organizer with various approaches. The importance of reusable large-scale standard test corpora in Chinese information retrieval research has been widely recognized and an evaluation workshop is now recognized as a new style of active research project that facilitates research by providing the data and a forum for research idea exchange and technology transfer.

For the first HTRDP information retrieval evaluation, the process was started from October, 2003. Three groups from different universities conducted the tasks and submitted the results. For the second evaluation, the process was started from October, 2004. Four groups have registered for the tasks and submitted the results for one or more tasks. The process of the third HTRDP information retrieval evaluation was started from October 2005 and the workshop was held on November, 2005 in Beijing and five groups submitted the results. Table lists the basic information of the each evaluation.

Table 64 Basic information of each 863 IR evaluation

	Time	Num of Groups	Tasks	Test Collection
First Workshop	October, 2003	3	Web Retrieval;	2G (Chinese Web page)
Second Workshop	October, 2004	4	Web retrieval; Passage Retrieval	15G (Chinese Web page)
Third Workshop	October, 2005	5	Web Retrieval	90G (Chinese Web page)

The HTRDP evaluations of Chinese information retrieval show the following development trends:

1. The evaluation procedure is more and more like the famous international information retrieval evaluation while giving prominence to characteristics of Chinese Information Retrieval;

2. Test corpus has been gradually expanded from 2G to

90G which is near the circumstance of mass data of the WEB;

3. Both the content and format of the test topics tend to describe the information need of real life users;

4. Evaluation metrics and relevance judgments are much more similar with the famous international information retrieval evaluation (TREC) [38];

In the rest of this section, we use the evaluation of 2005 as an example to describe the HTRDP evaluation of Chinese Information Retrieval.

9.2 Evaluation Metrics

- **MAP (Mean Average Precision)**

Average precision for a single topic is the mean of the precision after each relevant document is retrieved. *Mean Average Precision* (MAP) for a set of topics is the mean of the average precision scores for each topic.

This is a single-valued measure that reflects the performance over all relevant documents. It favors systems that retrieve documents quickly (highly ranked). When a relevant document is not retrieved at all, its precision is assumed to be 0.

As an example, consider there are two topics. One topic has four relevant documents, which are retrieved at ranks 1, 2, 4, and 7. Another has five relevant documents and three of them are retrieved at ranks 1, 3, and 5. For the first topic, the average precision is (1/1+2/2+3/4+4/7)/4=0.83. For the second topic, the average precision is (1/1+2/3+3/5+0+0)/5 = 0.45. As a result, the MAP of the two topics would be (0.83+0.45)/2=0.64.

- **R-Precision**

The R-Precision for a single topic is the precision after R documents have been retrieved, where R is the number of relevant documents for the topic. The average R-Precision for a set of topics is computed by taking the means of R-Precisions of the individual topics.

For example, assume a run consists of two topics, one with 50 relevant documents and another with 10 relevant documents. If the retrieval system returns 17 relevant documents in the top 50 documents for the first topic, and 7 relevant documents in the top 10 for the second topic, then the R-Precision for the two topics would be (17/50+7/20)/2 = 0.52

- **P@10**

The P@10 for a single topic is the precision after ten documents have been returned. The P@10 for a set of topics is computed by taking the means of P@10's of the individual topics.

9.3 Evaluation data

- **Corpus**

The evaluation data of 2005 contains only test corpus, namely **CWT100g** (The Chinese Web Test collection with 100 GB web pages), which is provided by the Computer Network and Distributed Systems Laboratory of Peking University [39]. CWT100g consists of 5,712,710 Web pages (about 90GB in size) crawled from 17,683 websites in China in June, 2004. Every page in the collection has a "text/html" or "text/plain" MIME type returned from the corresponding HTTP server.

- **Topics**

A topic is a statement of information need designed to mimic a real user's need. Each topic is formatted by a standard method to allow easier construction of queries. The 2005 information retrieval evaluation distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). A topic generally consists of four sections: an identifier, a title, a description, and a narrative. In the 2005 information retrieval Evaluation, the topics are presented in a TXT file. The encoding for Chinese character is GB2312. An example topic is shown as follows:

<top>
<num> 编号：001
<title> 自然语言处理
<desc> 描述：
文档应当涉及在中国得到研究和开发的自然语言处理技术.
<narr> 叙述：
一篇相关的文档应当涉及以下内容：自然语言处理技术；研究自然语言处理技术的公司或者研究机构；利用自然语言技术开发的产品.
</top>

The 2005 information retrieval evaluation distinguishes between two major categories of query construction techniques: automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention and what else are manual methods. In the 2005 IR Evaluation, participants are free to use any indexing and query construction techniques. All information in the topic is free to be exploited.

9.4 Results and Analysis

There are 2 different query construction methods in the 2005 evaluation: automatic method and manual method. The evaluation results are given in Table 65 and Table 66.

Table 65 Chinese IR Evaluation Result-automatic

Metrics	System 1	System 2	System 3	System 4	System 5
MAP	0.2727	0.1862	0.3107	0.3175	0.2858
R-PRECISION	0.3320	0.2554	0.3672	0.3605	0.3293
P@10	0.5300	0.5180	0.6240	0.5540	0.6280

Table 66 Chinese IR Evaluation Result-manual

Metrics	System 1	System 2	System 3	System 4	System 5
MAP	0.3257	0.1705	0.3538	0.2673	0.3671
R-PRECISION	0.3826	0.2327	0.4078	0.3185	0.4140
P@10	0.5580	0.4640	0.6840	0.4800	0.7040

Compared with the previous evaluation result (Fig. 9), the performance of the 2005 information retrieval system has increased a lot. We thought the following factors contributed to the better performance:

1. Since the corpus has been expanded to 90G and much more information such as link information are provided, the participated groups could use advanced relevant evaluation technologies such as the link analysis, anchor text analysis, etc., which leads to more accurate search results [40, 41];

2. As with last year's evaluation data as a training corpus, the participated groups could make use of these training sets to effectively overcome some difficult points of the Chinese IR system such as NER. They can also obtain a more stable and effective retrieval model by adjusting the system parameters [42, 43];

3. Effective use of advanced IR model or technology such as relevant feedback and re-ranking method also help to improve the search results [44];

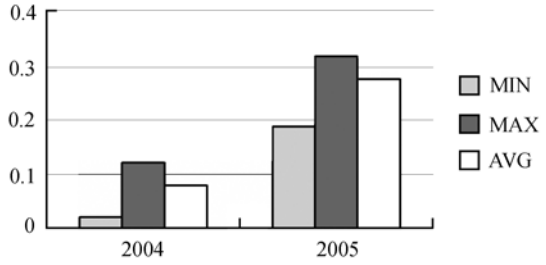

Fig. 9 Comparison of MAP (2004, 2005)

After the concrete analysis of the evaluation result, we can draw the conclusion that the following problems which once had effect on the performance of Chinese IR system have nearly been overcome:

1. Chinese segmentation error, especially for simple NER problems;

2. Chinese new word recognition;

3. Chinese abbreviations;

4. Multi-language query processing such as the mixed query of English and Chinese.

However, there still exist some problems:

1. Mismatch of query words and document words;

2. Some complex NER problems still need further research;

3. The appropriate weight schema of query words.

9.5 Conclusion

The HTRDP evaluation of Chinese information retrieval, sponsored by National High Technology Research and Development Program, is the most influential information retrieval evaluation of China. It is composed of a series of evaluations designed to enhance research in Chinese information retrieval technologies including Web search, passage retrieval, etc. These evaluations have enhanced the communication among industry, academia and government, and advanced the transfer of technologies from promising ideas to commercial products. Moreover, through the HTRDP evaluation of the Chinese Information Retrieval, several related techniques have come to the forefront in the world.

10 Character recognition

10.1 Introduction

With the ever-increasing application of character recognition, the evaluation of character recognition systems is becoming more important as it is of great necessity to develop and compare various character recognition technologies. It can predict performance, monitor progress, provide scientific explanations and identify open problems about such systems [43].

However, it is difficult to compare the performance of different character recognition systems because the systems are usually optimized for different applications and are thus adapted to different character sets and writing styles [56]. Therefore, it is necessary to setup a common sample database on a large scale and to evaluate and compare the performances on the database.

Several recent research works promote the sharing of samples [56]. Examples of widely used databases in the field of handwriting recognition include offline sample databases such as the CEDAR, NIST, CENPARMI, ETL9 (Japan), and PE92 (Korea), and online sample databases such as the UNIPEN and TUAT databases [44] [45]. The online sample databases of TUAT are becoming more and more popular in Japan for system design and evaluation [44]. Two TUAT databases are available [44]: The Nakayosi database is usually used for training, while the Kuchibue database is usually used for evaluation. Many researchers have reported results on the Kuchibue database [44]. But all the above databases are based on English characters, Japanese JIS and Korean KSC Chinese characters, instead of simplified and traditional Chinese characters, which are being used by over one billion Chinese people.

The Chinese government has been evaluating Chinese character recognition systems developed under the National High Technology Research and Development Program of China (863 Program) for seven times since 1991. Before 1997, the previous four evaluations mainly focused on printed character recognition and offline handwritten character recognition as shown in Table 67. However, due to the urgent need from the market of the Online (Handwritten) Chinese Character Recognition (OLCCR) systems, OLCCR evaluation was added in 1998 [46]. With the support of the 863 Program, the offline sample database HCL2000 [47] was established in 1998 and widely used in China.

Table 67 Chinese character recognition evaluations supported by Chinese 863 Program

	1991	1992	1994	1995	1997	1998	2003
Printed character recognition	√	√	√	√	√	√	
Off-line handwritten characters recognition	√	√	√	√	√	√	
Online handwritten character recognition					√	√	√

In July 2002, in order to support and orientate the research and development of the OLCCR systems, a national standard "Requirements and test procedure of on-line handwriting Chinese ideogram recognition" (GB/T18790-2002) [48] was issued in 2002. After the issues of the Chinese Internal Code Specification GB2312-1980 [49] and GB13000-1993, a national standard GB18030-2000 [50], i.e., "Information technology-Chinese ideograms coded character set for information interchange-Extension for the basic set" was issued in 2000. It includes 27533 Chinese characters and is a fundamental standard that should be followed by computer systems in China.

Consequently, it was necessary to setup a test sample database based on 27533 Chinese characters of the GB18030-2000 [50] and OLCCR standard GB/T18790-2002 [48] in order to provide an objective and fair evaluation for OLCCR systems. In 2003, we set up the online OLCCR-2003 sample database based on GB18030-2000 and organized OLCCR-2003 evaluation for the purpose of promoting the communication and collaboration among developers and speeding up the process of application of relative achievements. The rest of this section details the latest OLCCR-2003 evaluation.

10.2 Evaluation metrics

As shown in Fig. 10, the main procedure of the OLCCR-2003 evaluation consists of the four steps: collection of standard sample database, establishment of test sample database, test and evaluation.

The evaluation of an OLCCR system should be in accordance with the national standard "Requirements and test procedure of on-line handwritten Chinese ideogram recognition" (GB/T18790-2002)[48]. The performance of an OLCCR system is usually measured in terms of recognition accuracy, and recognition speed, etc., so the evaluation metrics are listed as follows.

(1) Character Recognition Rate (RS_i)

Character Recognition Rate RS_i denotes the rate of the number of test samples correctly recognized by the OLCCR system to be tested to the total number of test samples for the i^{th} character, which is defined by:

$$RS_i = \frac{CS_i}{NS_i}$$

where NS_i is the total number of test samples of the i^{th} character in the test database, and CS_i is the number of test samples of the i^{th} character, which are correctly recognized by the OLCCR system to be tested.

(2) Character Recognition Rate of Top Ten Match (RS_i^{10})

Character Recognition Rate of Top Ten Match RS_i^{10} denotes the rate of the number of test samples of top ten match to the total number of test samples of the i^{th} character, which is defined by:

$$RS_i^{10} = \frac{CS_i^{10}}{NS_i}$$

where NS_i is the total number of test samples of the i^{th} character in the test database, and CS_i^{10} is the number of test

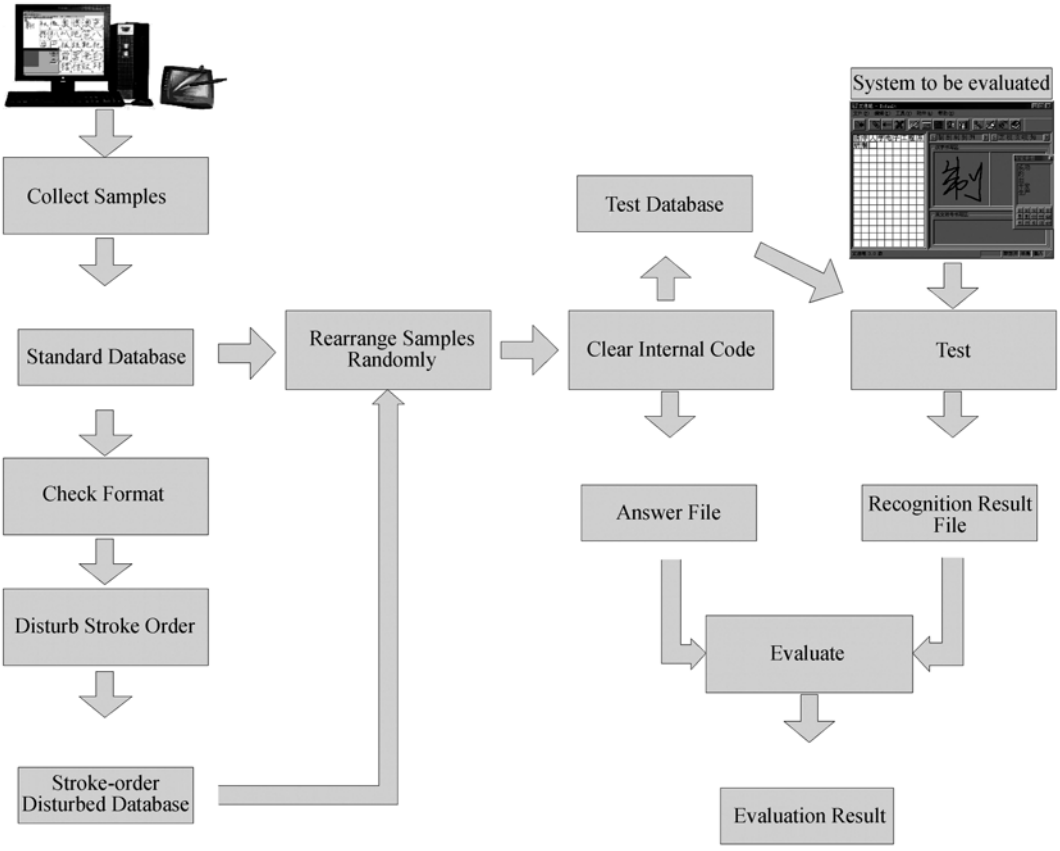

Fig. 10 Main procedure of the OLCCR-2003 evaluation

samples of the i[th] character of the top ten match .

(3) Overall Recognition Rate (R)

Overall Recognition Rate R denotes the average recognition rate of all the characters in the test database, which is defined by:

$$\mathrm{R} = \frac{\sum_{i=1}^{n} RS_i}{n}$$

where $\mathrm{RS_i}$ is the aforementioned Character Recognition Rate for the i[th] character, and n is the total number of all the characters in the test database.

(4) Overall Recognition Rate of Top Ten Match ($\mathrm{R^{10}}$)

Overall Recognition Rate of Top Ten Match ($\mathrm{R^{10}}$) denotes the average recognition rate of top ten match for all the characters in the test database, which is defined by:

$$\mathrm{R}^{10} = \frac{\sum_{i=1}^{n} RS_i^{10}}{n}$$

where $\mathrm{RS_i^{10}}$ is the aforementioned Character Recognition Rate of Top Ten Match for the i[th] character, and n is the total number of all the characters in the test database.

(5) Recognition Speed (V)

Recognition speed V should be calculated according to the equation:

$$V = \frac{N}{T}$$

where N is the total number of test samples of all the characters in the test database, and T is the recognition time of all the test samples cost by the OLCCR system to be tested.

10.3 Evaluation data

According to the national standard GB18030-2000[50], the characters to be tested and the number of sample sets are shown in Table 68. The evaluation group used devices such as compression tablet or electromagnetic tablet to collect samples to set up the OLCCR-2003 standard database according to Table 68.

10.4 Evaluation results

The OLCCR-2003 evaluation is designed to measure algorithm performance for OLCCR systems based on Chinese characters of GB18030-2000 and OLCCR standard GB/T18790-2002. There were two systems from China that participated in the evaluation. The evaluation results of the two systems are listed in Table 69 and Table 70.

Table 68 Characters to be tested and number of sample sets

Character set database		Digits & letters (62 Chars)	B1 (GB18030 2-byte code, 2nd partitions, 6763 Chars)	B2 (GB18030 2-byte code, 3-4th partitions, 14240 Chars)	B3 (GB18030 4-byte code, 6530 Chars)
Standard	Number of sets	60	60	30	30
	Number of samples	3,720	405,780	427,200	195,900
Stroke-order-disturbed	Number of sets	None	10	10	10
	Number of samples		67,630	142,400	65,300
Total	Number of sets	60	70	70	70
	Number of samples	1, 307, 930			
Remarks		1. The character set B3 is optional while the others are mandatory 2. The ratio of regular samples to fluent samples should be about 2:1 3. Disturb the stroke order of samples randomly 4. The format of the test sample file should be consistent with the Appendix B of the OLCCR standard GB18790-2002 [48]			

Table 69 Evaluation Results of System I

Database		Character set			
		Digits & letters (62 Chars)	B1 (GB18030 2-byte code, 2nd partitions, 6763 Chars)	B2 (GB18030 2-byte code, 3-4th partitions, 14240 Chars)	B3 (GB18030 4-byte code, 6530 Chars)
Standard	Number of sets	60	60	30	30
	Number of samples	3,720	405,780	427,200	195,900
	Recognition rate	81.45%	98.55%	98.00%	96.69%
	Recognition rate of top ten match	99.52%	99.94%	99.96%	99.90%
Stroke-order-disturbed	Number of sets	None	10	10	10
	Number of samples		67,630	142,400	65,300
	Recognition rate		98.59%	98.36%	96.88%
	Recognition rate of top ten match		99.93%	99.96%	99.91%
Total	Number of sets	60	70	70	70
	Number of samples	1,307,930			
	Recognition rate	97.85%			
	Recognition rate of top ten match	99.94%			
	Execution time	9,549 Second			
	Recognition speed	136.97 Chars/ Second			

Table 70 Evaluation results of System II

Database		Character set			
		Digits & letters (62 Chars)	B1 (GB18030 2-byte code, 2nd partitions, 6763 Chars)	B2 (GB18030 2-byte code, 3-4th partitions, 14240 Chars)	B3 (GB18030 4-byte code, 6530 Chars)
Standard	Number of Sets	60	60	30	None
	Number of Samples	3,720	405,780	427,200	
	Recognition Rate	70.16%	99.30%	98.17%	
	Recognition Rate of Top Ten Match	98.01%	99.97%	99.97%	
Stroke-order-disturbed	Number of Sets	None	10	10	
	Number of Samples		67,630	142,400	
	Recognition Rate		99.40%	98.00%	
	Recognition Rate of Top Ten Match		99.96%	99.98%	
Total	Number of Sets	60	70	70	0
	Number of Samples	1,046,730			
	Recognition Rate	98.43%			
	Recognition Rate of Top Ten Match	99.97%			
	Execution Time	29,674 Second			
	Recognition Speed	35.27 Chars/ Second			

Character recognition rate of each character for each system is omitted due to the limitation of page length.

10.5 Conclusion

During the OLCCR-2003 evaluation, we set up a large-scale test sample database based on the 27533 Chinese characters of the GB18030-2000 and OLCCR standard GB/T18790-2002 in order to provide an objective and fair evaluation for OLCCR systems. The sample database can be used both for training and testing for OLCCR research works. Based on the database, we tested two OLCCR systems that achieved good performances in terms of recognition accuracy, and recognition speed. To increase the difficulty of testing in the future, we should not constrain the writing style of samples so that most of the characters can be written fluently. Additionally, since the establishment of the large-scale Chinese character test sample database is very laborious, we may consider automatic generation of samples by computer in the future.

11 Face detection and recognition

11.1 Introduction

In recent years, face recognition has become an active area of research in pattern recognition, computer vision and psychology [51, 52]. The rapid development is due to a combination of factors: active development of algorithms, the availability of a large database of facial images, and methods for evaluating the performance of face-recognition algorithms [53, 54].

So far, there has been no official performance evaluation of the face detection technology. Based on important application of face recognition technology in the military, security and law, the United States funded the evaluation project named FERET (FacE REcognition Technology Test) by DARPA from 1993 to 1997, and organized three performance evaluations of face recognition technology. Test results show that the best performance of the first recognition rate is 95 percent on 1196 individuals set with similar conditions in training set and test set; however, to different cameras and different illumination conditions, the highest recognition rate drops to 82 percent, to face images acquired with one year's interval, the highest recognition rate is only around 51%. This shows that the face recognition algorithm is not very well for different light conditions, different pose, different cameras, and the ability to adapt to aging.

After the FERET project, a number of commercial face recognition systems have been developed. The U.S. Defense department organized further commercial face recognition systems evaluation named Face Recognition Vendor Test (FRVT). It has been held twice: FRVT2000 and FRVT2002. Some well-known face recognition systems participated in the FRVT 2002 test. FRVT 2002 included three evaluation tasks such as face identification, verification, and watch list task. For face verification task, there are two metrics: the false accept rate and the verification rate. Face identification ion is a closed universe evaluation, only gives recognition results in the gallery. The identification rate at rank k is the fraction of probes that have rank k or higher. Identification performance is plotted on a cumulative match characteristic (CMC). Watch list task is open universe identification, for this task, the algorithm should first judge whether the face image is in watch list pools; if it is then given recognition results. For watch list task, FRVT2002 give a false alarm rate to compare the identification rate among different algorithms. Under ideal conditions as frontal visa images, for a

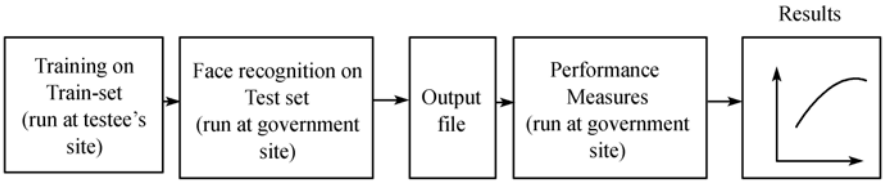

Fig. 11 Schematic of Face recognition testing procedure

total of 121,589 face images from 37,437 individuals for Face Identification, the highest recognition rate for the first is 73 percent, while for Face Verification, the error rate is about 6 percent.

The HTRDP 2004 Face Detection and Face Recognition Evaluation (FD&FR-04) has been established for the first time in China. The primary objectives are to assess the state-of-the-art face recognition technologies especially for Asians, identify future research areas and measure algorithm performance.

The FD&FR-04 supports five subtasks including face detection (FD), automatic face identification (AFI), partially automatic face identification (PAFI), automatic face verification (AFV) and partially automatic face verification (PAFV).

Fig. 11 presents a schematic of the testing procedure for face recognition task. Face detection test procedure is similar to this without training part. In the rest of this section, we will give more details such as test corpora, performance measures and FD&FR-04 test results.

11.2 Evaluation methods

As mentioned above, FD&FR-04 includes tasks of face detection, face identification and face verification. Face detection is to find all faces in an image where there can be multiple or no faces in an image. For the identification task, a system is presented with an unknown face that is to be identified, whereas, for verification task, a system is presented with a face and a claimed identity, and the algorithm either accepts or rejects the claim. Partially automatic face recognition tasks are given a facial image and the coordinate of the centers of eyes. Fully automatic tasks are only given facial images.

11.2.1 Evaluation metrics for the face detection task

To measure the result of face detection, we should give the reference location and size of faces in test images. Some definitions are needed here.

Definition 1 is Region of Interest (ROI). Here, ROI of a face image is the minimum rectangle including the centers of two eyes and the center of the mouth. The region in the black frame is the ROI as shown in Fig. 12. ROI, as the core region of the face, is used to evaluate the accuracy of face detection systems performance. The centers of eyes and mouth are labeled by advanced manual work.

Definition 2 is Region of Face (ROF). Here, ROF is the output rectangle of the detected face, which includes the location of top and left point, and width and height of the face.

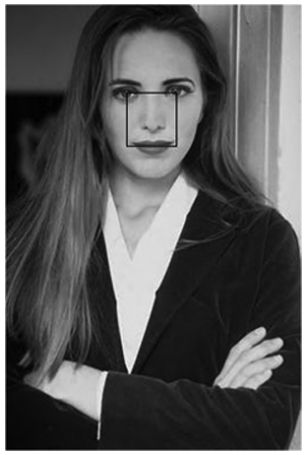

Fig. 12 ROI samples

Definition 3 is Correct Detect Face. ROF, the output of detected face, should include mouth center and at least one eye center. ROF should overlay at less 80% of the area of ROI, and area of ROF should be less than 6 times of that of ROI.

Based on the above three definitions, evaluation metrics for face detection are defined as follows:

- **Correct Detection Rate**

$$\mathrm{CD} = \frac{\text{Correct face number}}{\text{Total face number}} \times 100\%. \tag{43}$$

- **False Detection Index (FD): false detect number**
- **Average Process Time (Unit: ms/frame, a referenced measure)**

11.2.2 Evaluation metrics for the face identification task

For face identification and face verification tasks, gallery set and probe set are generally needed. A gallery set is a collection of images of known individuals against which testing images are matched. A probe set is a collection of probe images of unknown individuals that need to be recognized. For face recognition, the basic models for evaluating the performance of an algorithm are the closed and open universes. In the closed universe, every probe is in the gallery. In an open universe, some probes are not in the gallery. The FD&FR-04 is a closed universe model, which allows one to ask how good an algorithm is at identifying a probe image.

In face identification task, formally for each probe p from probe set we sort the similarity scores against gallery G, and

obtain the rank of the match. First Selection Result is the one whose gallery match is at rank one. Nth Selection Result is the one whose gallery match is at rank n.

Evaluation metrics for face Identification are as follows.

- **Correct Recognition Rate on First Selection**

$$\mathrm{FR} = \frac{\text{number of first selection is correct result}}{\text{Total number of samples in Probe Set}} \times 100\%. \quad (44)$$

- **Correct Recognition Rate on Top Ten Selection**

$$\mathrm{FR10} = \frac{\text{number of top ten selection has correct result}}{\text{Total number of samples in Probe Set}} \times 100\%. \quad (45)$$

- **Cumulative Match Characteristic Curve from top one to top fifty**
- **Average Process Time (Unit: ms/frame, a referenced measure)**

11.2.3 Evaluation metrics for the face verification task

In face verification task, for each p from Probe set we calculate the single similarity score against each g in the gallery. If p and g belong to the same subject and their similarity score is greater than threshold t, we call p as correctly verified. If p and g belong to different subjects and their similarity score is greater than threshold t, we call p as false accepted.

Evaluation metrics for face Verification are as follows.

- **Correct Verification Rate**

$$\mathrm{CVR} = \frac{\text{Number of correctly verified samples}}{\text{Total number that p and q come from same subject}} \times 100\%. \quad (46)$$

- **False Alarm Rate**

$$\mathrm{FAR} = \frac{\text{Number of false accepted samples}}{\text{Total number of samples}} \times 100\%. \quad (47)$$

- **Receiver Operator Characteristic curve:** The ROC curve with X-axis is FAR and Y-axis is CVR.
- **Equal Error Rate**
- **FAR100:** CVR when FAR=1%
- **FAR1000:** CVR when FAR=0.1%
- **Average Process Time (Unit: ms/frame, a referenced measure)**

11.3 Evaluation data

11.3.1 Evaluation data for the face detection task

A total of 2000 images of RGB color file were provided for FD tasks. These images came from the Internet, TV or movies. The scale of image size was different from 100*100 pixels to 1000*1000 pixels. Each image may include single face, multiple faces or no face. The size of the face is from 20*20 pixels to 300*300 pixels. To test the detailed performance of face detection, some test images used were with complex backgrounds, and some had varying accessories, lighting, pose and expression.

11.3.2 Evaluation data for the face recognition and identification tasks

Face images including human head and shoulder are collected for face recognition evaluation. Each image includes a single face. All the individuals are from China. The test data of face verification is similar to that of face identification besides the different sorted format. Test corpora for face identification are as follows.

1) Training Sample Corpora

The training set is a collection of images that is used to generate a generic representation of faces and/or to tune parameters for an algorithm. In the FD&FR-04 evaluation, the training set contains face images from 20 individuals. For each subject, one camera is used to capture the image in front of him/her. Besides one image with an office background, other images are captured under a white background. Each subject is also asked to look up, down and to the side to capture 3 face images. We also considered 3 kinds of expressions, 3 kinds of accessories or distance, and 3 kinds of lighting directions. Each individual has 15 face images including variations of background, accessories, lighting, pose and expression as Table 11.1 shows. This gives a total of 3000 face images in the training set. For PAFI and PAFV tasks, a reference text file with exact locations of two eyes of each face image is given.

2) Test Sample Corpora

Test corpora for face identification include Gallery corpora and Probe corpora as mentioned in part 11.2. Each has face images from 500 individuals. Some face images between Gallery set and Probe set are the same.

For face identification task, Gallery set has 500 face images, single frontal face image under white background for each individual. Probe set has a total of 3400 face images. One part of the probe set has 100 subjects, 14 face images including 14 kinds as Table 71 shows for each individual. Another part of probe corpora has 400 subjects, 5 face images for each individual include face images of background subset and one out of other four subsets including accessories, lighting, pose and expression subset. To ensure that matching is not done by file name, we give the images random names.

11.4 Results and analysis

The FD&FR-04 evaluation is designed to measure algorithm performance for face detection, face identification and face verification tasks. There were five sites from China that participated in the FD&FR-04 evaluation. Four sites selected the whole five sub tasks, another one selected the FD and PAFI sub task. Thus, a total of 22 test systems will be reported in the following. Table 72 to Table 76 and to Fig. 16 give the results for the FD&FR-04.

From the above results, in the face detection task, the best CD rate is 93.582% with 689 FD index by system2 and the second CD rate is 91.154% with 304 FD index by system4. As we know, the performance is better if the result has high CD rate and low FD index. But these two measures are correlated. Increasing CD rate will also increase FD index. Therefore, it's difficult to say which system is better by only one measure.

Table 71 Introduction for 14 kinds of face image for face recognition task

Subset	Details	Characters	Explain
Background	White	Frontal face with white Background (capture two images)	Frontal face with different background
	Office	Frontal face with office Background	
Accessories and distance	Cap	Select one cap from several caps	Frontal face with single color background
	Glasses	Select one glasses from several glasses	
	Distance	Different distances from the camera, such as go back 20 to 40cm	
Expression	Close Eye	Close Eye	Frontal face with single color background
	Laugh	Laugh	
	Surprise	Surprise	
Illumination	M_0	Middle camera's axis, 0° azimuths, 0° elevations	Frontal face with single color background
	M_45	Middle camera's axis, 45° azimuths, 0° elevations	
	M_45	Middle camera's axis, 45° azimuths, 45° elevations	
Pose	Up	Up face	A little pose various with single color background
	Down	Down face	
	Side	Side face	

Table 72 Face detection results in FD&FR-04

Id	System1	System2	System3	System4	System5
CD	86.687%	93.582%	71.639%	91.154%	45.967%
FD	81	689	1834	304	1452
AT(ms/f)	1197.5	2167.5	477.5	1145	1491.5

Table 73 Automatic face identification results in FD&FR-04

Id	System1	System3	System4	System5
FR	67.7059%	85.2059%	91.1471%	20.3824%
FR10	81.7647%	93%	96.2941%	36.8235%
AT(ms/f)	1831.765	2029.706	1480.588	562.3529

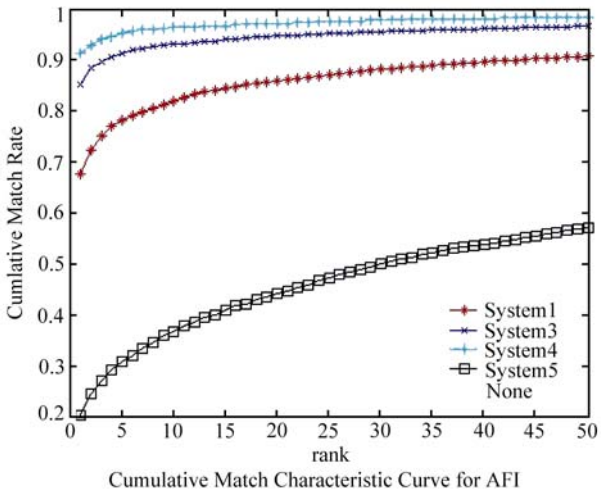

Fig. 13 Automatic face identification results in FD&FR-04

Table 74 Partially automatic face identification results in FD&FR-04

Id	System1	System2	System3	System4	System5
FR	74.3529%	71.6176%	86.5%	91.9118%	40.5588%
FR10	88.4118%	83.3529%	94.7059%	96.8235%	60.0882%
AT(ms/f)	815.2941	152.9412	660.5882	610.5882	36.47059

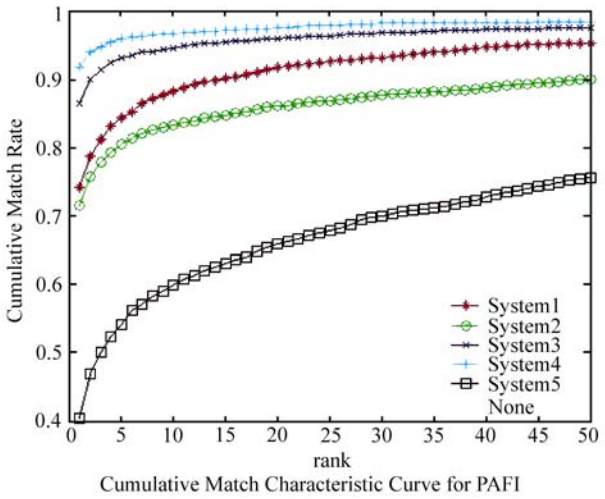

Fig. 14 Partially automatic face identification results in FD&FR-04

Table 75 Automatic face verification results in FD&FR-04

Id	System1	System3	System4	System5
ERR	9.38315%	8.0468%	1.48325%	31.1682%
FAR100	90.9412%	92.764%	99.7941%	51%
FAR1000	78.0882%	83.147%	97.9118%	26.3235%
AT(ms/f)	1811.765	2034.706	1499.118	563.2353

For face identification task, system4 has the highest correct recognition rate on first selection with 91.1471% for AFI and 91.9118% for PAFI. For face verification task, system4 also is the top performer with the lowest ERR 1.48325% for AFV task and ERR 1.27105% for PAFV. From Table 73 to Table 76, we observe that the results of PAFI and PAFV are similar to the results of AFI and AFV. That is to say whether providing the coordinate of the centers of the eyes does not seriously affect face recognition performance for most of the test systems. Therefore, we can only test automatic face recognition task in future evaluation.

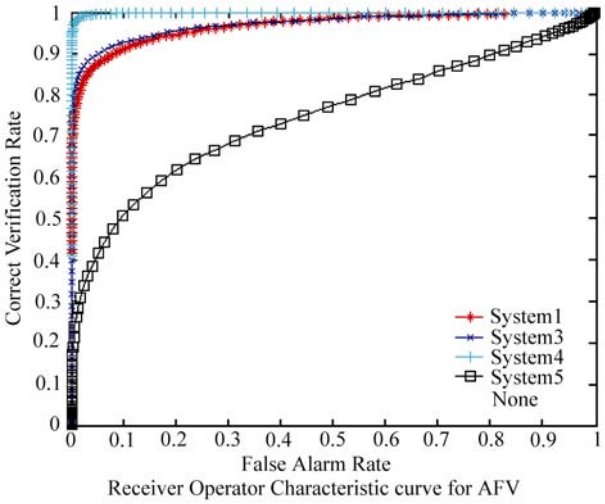

Fig. 15 Automatic face verification results in FD&FR-04

Table 76 Partially automatic face verification results in FD&FR-04

id	System1	System3	System4	System5
ERR	6.64315%	7.0652%	1.27105%	24.2973%
FAR100	95.2353%	94.1176%	99.8824%	66.9118%
FAR1000	84.7941%	85%	98.3824%	43.5882%
AT(ms/f)	851.1765	673.8235	617.6471	61.47059

Another goal of the FD&FR-04 evaluations is to identify areas of strengths and weaknesses in the field of face detection and recognition. So in addition to testing the result on the whole probe set, we also test on some sub probe sets to find the detailed performance of the algorithm. These sub probe sets include varying illumination, pose, expression, and background, with glasses or cap accessories, etc. Fig. 17 to Fig. 19 show the result of FD&FR-04 on the sub probe set.

We observed variation in performance due to changing the probe set. Despite the overall variation in performance, definite conclusions about algorithm performance can be made from the above results. In face detection task, as Table and show, system 2 and system4 also have better performance than other systems on sub probe sets. shows that almost all of the systems have lower CD rate on sub probe set such as pose, illumination and multi face than on the frontal face probe set. Also Fig.18 and Fig. 19 show, in the face recognition task, that system4 has better results than the other systems on the sub probe set. The results on the sub probe set of pose, illumination and with accessories are less than the results on the probe set of different background and expression.

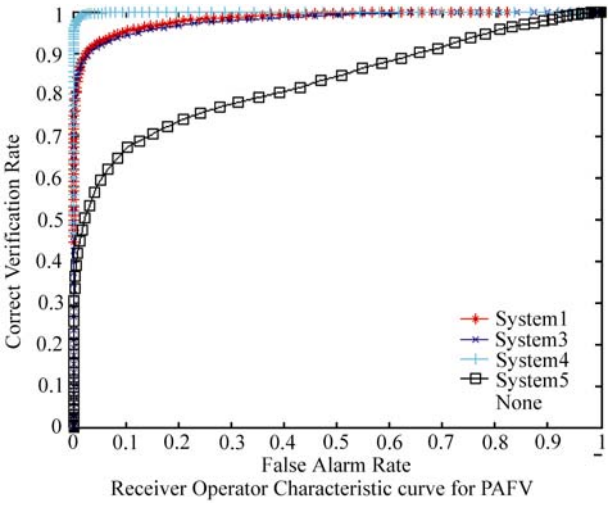

Fig. 16 Partially automatic face verification results in FD&FR-04

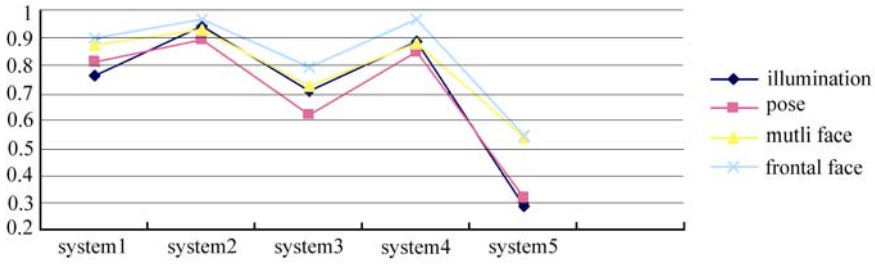

Fig. 17 Face detection tested on the sub probe set in FD&FR-04

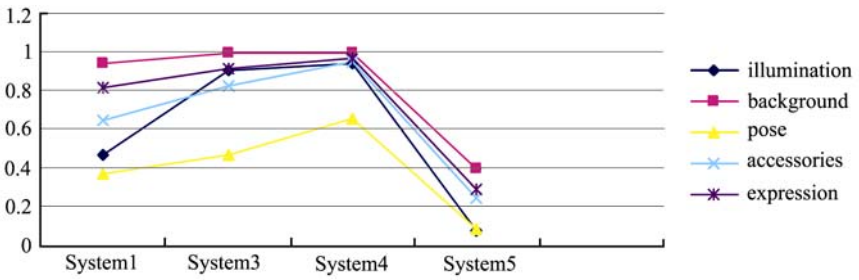

Fig. 18 Automatic face identification tested on sub probe set in FD&FR-04

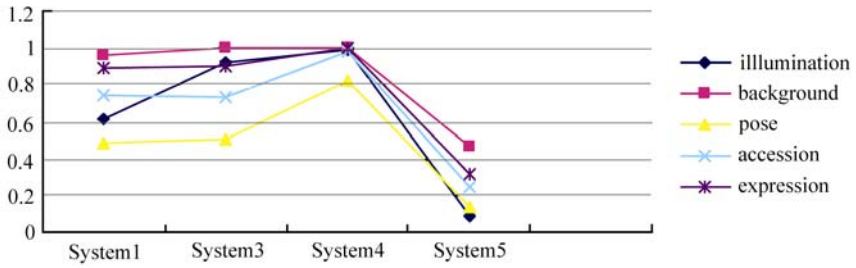

Fig. 19 Automatic face verification tested on sub probe set in FD&FR-04

11.5 Conclusion

In this section, we presented the FD&FR-04 evaluation method for face detection and recognition tasks. The evaluation method was designed so that performance can be measured on different probe sets. The comparative analysis shows that future areas of face detection and face recognition evaluation research include changing the gallery set and probe set, increasing the number of individuals in the gallery and probe, attempting the open a universal method for face recognition, which shows some probes not in the gallery, as well as the detail and depth of analysis performed. The research on algorithms in face detection and face recognition is to develop algorithms to compensate for changes in illumination, to research pose problems and with different accessories. Achieving this objective requires an evaluation on a much larger and broader scale than any previous biometric evaluations.

12 Overall conclusion and future work

In this paper, details of the HTRDP evaluations are presented. The general information of HTRDP evaluations such as the history, the concerned technology categories, the organizer, the participants, the procedure, etc., is introduced. This was followed by details of the evaluations on all technology categories, covering Chinese word segmentation, machine translation, acoustic speech recognition, text to speech, text summarization, text categorization, information retrieval, character recognition, and face detection and recognition. For the evaluations on each technology categories, the history, the evaluation tasks, the data, the evaluation method, and the results are given.

The HTRDP evaluations cover a wide range of fields and tasks in the domain of Chinese information processing and intelligent human-machine interface. As the most famous evaluations in China, it has played a very important role in providing comparison and communication for researchers and in boosting the technique in all related fields. In the past decades, the HTRDP evaluations have contributed to the boosting of such fields as Chinese character recognition, speech synthesis, and machine translation in China, and nowadays the researchers of China still lead in those areas.

In recent years, with participation from all over the world, the HTRDP evaluations are gaining international reputation. In the future, the HTRDP evaluations will continue to be conducted, while adopting new tasks and new evaluation methods according to the state-of-the-art of the techniques.

Acknowledgements The research was sponsored by the National Hi-Tech Program of China (No. 2004AA114010，2003AA111010).

We would like to thank all institutes and individuals who have helped and supported the HTRDP evaluations since the early 1900's; some of them are listed as follows:

Prof. Shiwen Yu, Institute of Computational Linguistics, Peking University; Prof. Aijun Li, Institute of Linguistics, Chinese Academy of Social Sciences; Prof. Kaiying Liu and Dr. Erhong Yang, Computer Application Institute of Shanxi University; Prof. Xiaofeng Gu, Peking University; Prof. Kaizhu Wang, Harbin Institute of Technology; Prof. Jialu Zhang, Institute of Acoustics, Chinese Academy of Sciences; Prof. Renhua Wang, University of Science and Technology of China; Dr. Hitoshi Isahara and Dr. Yujie Zhang, National Institute of Information and Communications Technology, Japan; Prof. Benjamin Tsou and Dr. Olivia Kwong, Language Information Sciences Research Center,City University of Hong Kong; Dr. Hongfei Yan, Computer Networks and Distributed Systems Laboratory, Peking University; Dr. Ming Zhou and Dr. Ji-Rong Wen, Microsoft Research Asia.

References

1. Sproat R, Emerson T. The First International Chinese Word Segmentation Bakeoff. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. 2003, 133−143
2. Levow G -A, The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006, 108−117
3. Emerson T. The Second International Chinese Word Segmentation Bakeoff. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 2005. 123−133
4. ACE evaluations: http://www.nist.gov/speech/tests/ace/index.htm
5. YU S W. Automatic Evaluation of Output Quality for Machine Translation Systems, Machine Translation. Netherlands: Kluwer Academic publisher, 1993, 8: 117−126
6. Papineni K A, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109−022). IBM Research Division, Thomas J. Watson Research Center.2001
7. Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese Lexical Analyzer ICTCLAS, In: Proceedings of 2nd SigHan Workshop, 2003.184−187
8. http://www.nist.gov/speech/tests/mt
9. http://nlp.cs.nyu.edu/GTM/

10. Och F J. Minimum error rate training in statistical machine translation. In: Proceedings of the 41st ACL, Sapporo, Japan, 2003. 160−167
11. Och F J. Statistical Machine Translation: From Single-Word Models to Alignment Templates. 38−39
12. Yasuhiro A, et al. Overview of the IWSLT04 Evaluation Campaign. 2004
13. Paul M, Nakaiwa H, Federico M. Towards innovative evaluation methodologies for speech translation. Working Notes of the NTCIR-4 2004 Meeting. 2004, 2(Suppl). 17−21
14. White J S, O'Connell T, O'Mara F. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: Proceedings of the AMTA, 1994, 193−205
15. Papineni K, Roukos S, Ward T. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th ACL, Philadelphia, USA, 2002. 311−318
16. Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the HLT 2002, San Diego, USA, 2002. 257−258
17. Turian J P, Shen L, Melamed I D. Evaluation of machine translation and its evaluation. In Proceedings of the MT Summmit IX, New Orleans, USA, 2003. 386−393
18. Liu Q, Liu Y. Machine translation automatic evaluating method and system thereof, Chinese Patent, CN1641631-A
19. http://www.nist.gov/speech/history/index.htm
20. http://www.tc-star.org/
21. Martin A, Doddington TKG, Ordowski M, et al. The DET curve in assessment of detection task performance. In: Proceedings of EuroSpeech'97. 1997, Vol4, 1895−1898
22. Van Santen J P H, Pols L C W, Abe M, et al. Report on the Third ESCA TTS Workshop Evaluation Procedure. Third ESCA TTS Workshop, 1998.
23. http://www.tc-star.org/
24. Benoit C, Grice M, Hazan V. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Commun. 1996, 18:381−392
25. Document Understanding Conference, http://duc.nist.gov/
26. He J, Tan A -H, Tan C -L. A Comparative Study on Chinese Text Categorization Methods. PRICAI 2000 Workshop on Text and Web Mining, 2000.
27. Lewis D D, Yang Y M, Rose T G., et al. RCV1 A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research. 2004, 5: 361−397
28. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002
29. Aas K, Eikvil L. Text Categorisation: A survey. Raport NR 941, 1999
30. Yang Y M. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1−2)
31. The Text Retrieval Conference, http://trec.nist.gov/
32. Harman D. The first text retrieval conference (TREC-1). Information Processing and Management, 1993, 29(4): 411−414
33. Voorhees E M. Overview of TREC 2005. In: Proceedings of the Text REtrieval Conference (TREC). Gaithersburg, Maryland, 2005
34. NII Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/
35. Cross Language Evaluation Forum, http://www.clef-campaign.org/
36. Zhang J L, et al. Research on the 863 Chinese Information Retrieval Evaluation. Journal of Chinese Information Processing, 2006, 20(suppl): 19−24(in Chinese)
37. Shah C, Croft W B. Evaluating High Accuracy Retrieval Techniques. In: Proceedings of SIGIR '04, 2004
38. Text Retrieval Conference, http://trec.nist.gov.
39. Chinese Web Information Retrieval Forum, http://www.cwirf.org/.
40. Cheng Y X, et al. 863 Web Track Experiments at ICST-PKU. Journal of Chinese Information Processing, 2006, 20(suppl): 102−106(in Chinese)
41. Zhao L, et al. 2005 THUIR Report for 863 Information Retrieval Evaluation. Journal of Chinese Information Processing, 2006, (suppl): 91−95 (in Chinese)
42. Zhang Z C, et al. Technology Report of HIT-IRLab for Evaluation 2005 of 863 Information Retrieval. Journal of Chinese Information Processing, 2006, 20(suppl): 83−90 (in Chinese)
43. Xu W R, et al. PRIS Information Retrieval System Report. Journal of Chinese Information Processing, 2006, 20(suppl): 96−101(in Chinese)
44. Lv B B, et al. 863 Information Retrieval Evaluation -Institute of Automation. Journal of Chinese Information Processing, 2006, 20(suppl): 78−82 (in Chinese)
45. Kanungo T, Marton G A, Bulbul O. Performance Evaluation of Two Arabic OCR Products. In: Proceedings of AIPR Workshop on Advances in Computer. Assist Recognition, SPIE. Vol 3584
46. Liu C L, Jaeger S, Nakagawa M. Online Recognition of Chinese Characters: The State-of-the-Art. IEEE Transaction on Patten Analysis and Machine Intelligence, 2004, 26(2): 198−213
47. Marti U -V, Bunke H. A full English sentence database for off-line handwriting recognition. In: Proceedings of the 5th Int. Conf. on Document Analysis and Recognition, Bangalore, India, 1999. 705−708
48. Liu C P, Qian Y L, et al. 863 Testing System on Handwritten Chinese Chararacter Recognition. Journal of Chinese Information Processing, 2000, 14(2): 2−7 (in Chinese)
49. Guo J, Lin Z Q, Zhang H G. A New Database Model of Off-line Handwritten Chinese Characters and Its Applications. Chinese Journal of Electronics, 2000, 28(5): 115−116 (in chinese)
50. Requirements and test procedure of on-line handwriting Chinese ideogram recognition. Chinese National Standard GB/T 18790-2002. July, 2002
51. Chinese Ideograms Coded Character Set for Information Interchange-Basic Set. Chinese National Standard GB 2312−1980, 1980
52. Information technology-Chinese ideograms coded character set for information interchange-Extension for the basic set. Chinese National Standard GB 18030−2000. March, 2000
53. Chellappa R，et al. Human and Machine Recognition of Faces：A Survey. In: Proceedings of the IEEE，1995，83(5): 705−741
54. Zhao W Y, Chellappa R, Rosenfeld A, et al. Face Recognition: A Literature Survey. ACM Computing Survey, 2003, 35(4): 399−458
55. Phillips P J, Moon H, Rizvi S, et al. The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1090−1104
56. Phillips P J, Grother P J, et al. Face Recognition Vendor Test 2002：Evaluation Report, Technical Report. NISTIR 6965, National Institute of Standards and Technology, 2003