Introduction to China's HTRDP Machine Translation Evaluation

LIU Qun, HOU Hongxu,
LIN Shouxun, QIAN Yueliang
Instititue of Computing Technology,
Chinese Academy of Sciences
No.6, Kexueyuan Nanlu, Zhongguancun
Beijing 100080, China
liuqun@ict.ac.cn

ZHANG Yujie
ISAHARA Hitoshi,
National Institute of Information
and Communications Technology
3-5 Hikaridai, Seika-Cho, Soraku-gun,
Kyoto, 619-0289 Japan
yujie@nict.go.jp

Abstract

Since 1994, China's HTRDP machine translation evaluation has been conducted for five times. Systems of various translation directions between Chinese, English, Japanese and French have been tested. Both human evaluation and automatic evaluation are conducted in HTRDP evaluation. In recent years, the evaluation was organized jointly with NICT of Japan. This paper introduces some details of this evaluation.

1 Introduction

Evaluation is recognized as an important drive for machine translation research. DARPA have a long history to organize machine translation evaluation. Recent years, inspired by the efficient work of the DARPA-supported NIST evaluation, some other machine translation evaluations are supposed, such as ISWLT and TC-STAR MT evaluation.

In China, the HTRDP machine translation evaluation also has a long history. Since 1994, HTRDP MT evaluations have been conducted for five times. We will give a detailed introduction to China's HTRDP MT evaluation.

2 Origination and history

HTRDP means China's national High-Tech Research and Development Programme. The full name of HTRDP evaluation is "HTRDP Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface Technology"[1]. It is a series of evaluation activities which is sponsored by HTRDP. HTRDP evaluation covers a wide range of technologies, which include:

- Machine translation (MT)
- Automatic speech recognition (ASR)
- Speech to text (TTS)
- Chinese character recognition (CR)
- Information retrieval (IR)
- Chinese word segmentation (CWS, includes part of speech tagging and named entity recognition)
- Text classification (TC)
- Text summarization (TS)
- Human face detection and recognition (FR)

Table 1 gives the year of each HTRDP evaluation and technology categories which were tested in that year. Please note that the 2005 HTRDP evaluation is ongoing. We can see that machine translation is firstly test in the 3rd HTRDP evaluation in 1994, and five HTRDP MT evaluations has been conducted up to now.

	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							●	

Table 1 Technology categories tested in each HTRDP evaluation

3 Organizer

HTRDP evaluation is organized by Institute of Computing Technology (ICT), Chinese Academy of Sciences. Since 2004, ICT started its cooperation with the National Institute of Information and Communications Technology (NICT) of Japan on the organization on HTRDP MT evaluation.

[1] HTRDP is also called "863" Programme, in order to commemorate the time "March 1986" when China's previous leader Deng Xiaoping approved the suggestion of four famous Chinese scientists to found such a programme. So the HTRDP evaluation is also called "863" evaluation. Please refer to the website: http://www.863data.org.cn

4 Time table

Normally, in HTRDP evaluation, the guideline is released in spring, and the evaluation is conducted in autumn. For an instance, the time schedule of 2005 HTRDP evaluation is given as follows[2]:

- March-April: Discussion of the guidelines
- April 29: Release of the evaluation guidelines
- July 29: Deadline of registration
- August 1: Releasing the training data
- August 22: Releasing the development data
- September 20: Releasing the test data
- September 22: Deadline of result submission
- October 21: Notification of evaluation results
- November: Evaluation workshop

5 Tracks

HTRDP MT Evaluation concerns machine translation technologies between Chinese, English, Japanese, French, and etc. Table 2 gives the definition of evaluation tracks ever defined in HTRDP MT Evaluation. Table 3 gives the tracks in each evaluation.

CEMT	Chinese→English	Machine Translation
ECMT	English→Chinese	
CJMT	Chinese→Japanese	
JCMT	Japanese→Chinese	
JEMT	Japanese→English	
EJMT	English→Japanese	
CFMT	Chinese→French	
CEWA	Chinese↔English	Word Alignment

Table 2 Definition of evaluation tracks

6 Participants

The participants of HTRDP MT Evaluation mainly come from China mainland. From 2004, through the cooperation with NICT, some Japanese companies joined the evaluation. All the participants are listed below (including those who has registered in 2005 evaluation):

- Beijing University of Technology
- CCID Cooperation
- Futsuji Cooperation (Japan)
- Huajian Cooperation
- Harbin Institute of Technology

[2] Please refer to 2005 HTRDP Evaluation Workplan, available at: http://www.863data.org.cn/english/2005plandown_en.php

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

Table 3 Tracks set in each HTRDP MT Evaluation

- Institute of Automation, Chinese Academy of Sciences
- Institute of Computing Technology, Chinese Academy of Sciences
- Kodensha Cooperation (Japan)
- Multran Cooperation
- National University of Defense Technology
- Nanjing University
- Sharp Cooperation (Japan)
- Transtar Cooperation
- Xiamen University

Up to now, most of the participating systems adopt rule-based approach or example-based approach.

7 Evaluation Method

In HTRDP MT evaluation, both human evaluation and automatic evaluation are conducted.

7.1 Human evaluation

Human evaluation is used in each HTRDP MT evaluation. In previous evaluation, human evaluation is based on a single metric measurement. The metric is called Intelligibility which is defined in table 4.

Usually, four human experts are invited to evaluate all the results made by the participating systems sentence by sentence. The human experts will score all the results of the same source sentence in the same time. However, the results of each source sentence are shuffled, so the human experts cannot know which results are made by the same participating system according to the order of the results.

Human experts will give each result sentence a score from 0 to 100. Then the intelligible score of each participating systems will be calculated by the average over all the sentences and experts.

Score	Description	Intelli-gibility
0	The translation is completely unintelligible.	0%
1	You cannot figure out what the translation wants to express. But some phrases are properly translated	20%
2	Parts of the source text are properly translated. Keywords are properly translated.	40%
3	The translation conveys the meaning of the source text fairly well. You can guess the meaning of source text from the translation. There are some errors.	60%
4	The translation conveys the meaning of the source text quite well. You can figure out the meaning of source text from the translation. There are several errors.	80%
5	The translation exactly conveys the meaning of the source text. The structure of sentence is properly chosen. There are only one or two trivial errors.	100%

Table 4 Intelligible measurement for human evaluation

In the 2005 evaluation, we will adopt a new measurement based on two metrics: adequacy and fluency, which is more commonly used in other MT evaluations.

7.2 Automatic evaluation

In the history of HTRDP MT evaluation, different kind of automatic evaluation method has been tried.

7.2.1 Test Point Method

In early 1990s, Prof. YU Shiwen of Peking University has proposed a method for automatically evaluating machine translation quality, and developed an automatic machine translation evaluation system named as MTE-94 based on this method [YU 1993]. This method is somewhat like the "standardized test" for human foreign language learners. The key of this method is the idea of "test point". Firstly, a set of "test points" is defined for each machine translation directions. A "test point" is a difficult problem which a MT system has to resolve. For example, "Chinese word segmentation" is one test point for Chinese-English MT system, and "translation word selection" is another test point. For each test point, tens of source sentences are given. MTE-94 will judge if the system can resolve the problem in the "test point" according the translation of these source sentences. For example, for the test point of "Chinese word segmentation", MTE-94 gives tens of Chinese sentences which containing word segmentation ambiguities (such as "和服务", which has a ambiguity of "和服+务" or "和+服务"). Then MTE-94 will test how many result translations contain the translation words ("service") of the correct segmented source words ("服务"). In MTE-94 system, hundreds of test points are defined in a test set containing 3300 source sentences. The participating systems are asked to translate these sentences to target language. However, MTE-94 system does not try to give a score to each result sentences. For each source sentences, MTE-94 will give a judgement if the translation is correct in the test points defined in the source sentence. This judgement can be made automatically by character string comparing. The overall score of the system will be calculated according the ratio of corrected processed sentences in each test points.

The test point method had been used in the early HTRDP MT evaluation experimentally (before 1998), in order to improve the MTE system. Prof. YU had published several papers to introduce these experiments. Unfortunately, no formal automatic evaluation results can be found in the HTRDP MT evaluation reports.

7.2.2 N-gram method

In the test point method, it is a very hard work to define the test points, to select source sentences containing these test points, and to give all the answers for each selected sentences. So this method was no longer used in later HTRDP MT evaluations.

From 2003, we adopt the automatic evaluation method based n-gram which is first proposed by IBM [Papineni 2001] [NIST 2001] and has been used in NIST MT evaluations. In 2003 evaluation, the NIST metric is used. In later evaluations, multiple metrics are used, which including: BLEU, NIST, GTM [Turian 2003], mPER, mWER.

A problem in using such metrics on Chinese and Japanese translations is that there are no word boundary in Chinese and Japanese. The word segmentation is ambiguous in Chinese and Japanese. Our approach is to use the n-gram based on Characters.

To using the n-gram based evaluation method, we make four reference translations by human translators. All the human translators are native speakers of the target language who know the source language very well. Fortunately, we can find people from all over the world in Beijing,

especially in the universities. And, all the Japanese reference translations are provided by our Japanese collaborator NICT.

7.2.3 Entropy method

In 2005 evaluation, we will try an additional automatic machine translation evaluation metric, which is proposed by our group, named as ICTMTE.

The ICTMTE metric for automatic MT evaluation is based on the idea of "entropy". We will introduce it in detail in the 2005 HTRDP evaluation workshop. Here we will give a brief introduction.

In this method, the translation sentence is firstly compared against the reference translations. We can find some continuous word (or character) sequences are matched. So the translation sentence is segmented into some pieces, where each piece is either a sequence of matched word (or character), or an unmatched word (or character). Thus we can give the translation sentence a "distribution score". We assume that the more distributive the sentence is segmented, the poor the translation quality is. Because the distribution score can be well defined by the entropy, we can use the entropy to measure the translation quality. Besides, some other factors should also be taken into consideration. More details of the method will be described in our future paper.

Compare with the n-gram method, one advantage of the entropy method is, we do not need to select the order of n-gram. In n-gram method, whether we use 4-gram or tri-gram is quite subjective or experiential. However, in entropy method, we do not need to make such a decision.

7.3 Evaluation for word alignment

For word alignment track, the evaluation is made automatically. In the gold alignments, there are two kinds of alignment links: sure links and possible links. The metrics include: Precision, Recall, F1-measure and Error Rate, which is the same as the definition in [Och 2003].

8 Data

8.1 Test data

In early HTRDP MT evaluations (before 1998), the test data is made by linguistics. Most of them are short sentences, which like the sample sentences in grammar textbooks. In each sentence there is at least one test point [Yu 1993].

In later evaluations (after 2003), the test data are mainly collected from real language. There are both dialog data and text data in test set. Usually, the size of the test set is about 800-1000 sentences.

In the 2003 evaluation, all the test data are in Olympic-related domains, including: sports news, whether forecast, travel, traffic, hotel, and catering information.

In the 2004 evaluation, test data came from both Olympic-related domains and general domains.

In the coming 2005 evaluation, the dialog data come from Olympic-related domains, and the text data come from general domains.

8.2 Development data and training data

In previous evaluation (before 2004), no training data and development data was provided. However, in the coming 2005 evaluation, we begin to provide the training data and development data.

The development data is just the collection of test data and there reference translations which was used in previous HTRDP MT evaluation. However, because the evaluation tracks between Japanese and English are newly added, we will make new development data for them, where the source sentences come from previous Japanese to Chinese evaluation or English to Chinese evaluation, and the reference translations are newly made by NICT. For word alignment track, a development data set containing about 1000 sentence pairs will be provided. Word alignments have been made manually in these sentence pairs.

The training data will only be provided for machine translation between Chinese and English, which containing about 700,000 sentence pairs. Most of the training data is provided by ChineseLDC[3].

Up to now, no limit is made to the participants on the training data they can use. That means, the participants can use any data to training their system. However, they should give a description to all the data they used to training their system in the workshop.

9 Conclusion

HTRDP ("863") MT evaluation is the official MT evaluation in China. Almost all the machine translation research institutes and corporations in China mainland are involved, and some participants are from overseas. Besides the translation evaluation between Chinese, English, Japanese and French, a new word alignment track is added in 2005 evaluation. Large training data set is provided to the participants freely from this year.

Participants from all over the world are welcome. For more information, please visit the evaluation

[3] Please refer to : http://www.chineseldc.org

website: http://www.863data.org.cn.

References

NIST, *Automatic evaluation of machine translation quality using n-ram co-occurrence statistics*. Technical report, NIST, 2001. Available at: http://www.nist.gov/speech/tests/mt/.

Franz Josef Och, Hermann Ney. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. *Bleu: a Method for Automatic Evaluation of Machine Translation*, IBM technical report, keyword: RC22176, 2001

Joseph Turian, Luke Shen, and I. Dan Melamed. *Evaluation of Machine Translation and its Evaluation*. In Proceedings of Machine Translation Summit IX Workshop "Machine Translation for Semitic Languages: Issues and Approaches", New Orleans, USA, 23-28 September 2003.

YU Shiwen, *Automatic Evaluation of Output Quality for Machine Translation Systems*, Machine Translation, 1993, 8:117-126, Kluwer Academic publisher, printed in the Netherlands