Introduction to China's HTRDP("863") Machine Translation Evaluation

Qun Liu, Hongxu Hou
Shouxun Lin, Yueliang Qian

ICT, CAS, China

Yujie Zhang
Hitoshi Isahara

NICT, Japan

MT Summit X, Phuket, Thailand, 2005.9.14

ICT

INSTITUTE OF COMPUTING TECHNOLOGY

中科院计算所

Outline

- Preface
- Origination and History
- Organizer
- Time cycle
- Evaluation Tracks
- Participants
- Evaluation Metrics
- Data
- Results
- Comparison with NIST MT evaluation
- Conclusion and future work

Preface

- Evaluation is recognized as an important drive for machine translation research.
- Other MT Evaluations
  - NIST (Supported by DARPA Tides Project)
  - IWSLT (Organized by CSTAR)
  - TC-STAR (Organized by EU’s TC-STAR Project)
- China’s HTRDP MT Evaluation
  - Supported by China’s HTRDP (“863” Programme)

Origination

- HTRDP:
  - China's national High-Tech Research and Development Programme
- "863" Programme: another name of HTRDP
  - In 1986, four famous Chinese scientists submitted a proposal to Chinese government for founding a high technology research and development programme
  - China's previous leader Deng Xiaoping approved this suggestion in March of 1986
  - The nick name "863" Programme is to commemorate the month when Deng Xiaoping approved the proposal

HTRDP Evaluation

- An abbreviation of “the HTRDP Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface Technology”
- Also called “863” Evaluation
- It is a series of evaluation activities which is sponsored by HTRDP on the research area of natural language processing and human-machine interation
- Seven HTRDP evaluations had been conducted from 1991 to 2004.

History

- 1990: preparative evaluation
- 1991: 1st
- 1992: 2nd
- 1994: 3rd
- 1995: 4th
- 1998: 5th
- 2003: 6th
- 2004: 7th
- 2005: 8th

Technologies covered by HTRDP Evaluation

- 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)

Technologies covered by 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							●	

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.

Time cycle

- The evaluation time cycle is a calendar year, normally:
  - Guidelines Releasing: in spring
  - Result Submission: in autumn
  - Workshop: in winter
- Time Table of 2005 HTRDP Evaluation:
  - 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 28: Evaluation workshop

Evaluation Tracks (1)

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

Definition of evaluation tracks

Evaluation Tracks (2)

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

Participants

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

Evaluation Metrics

- Human Evaluations
  - Intelligible measurement (before 2004)
  - Adequacy and Fluency (2005)
- Automatic Evaluations
  - Test Point Methods (1995,1998)
  - N-gram Metrics and Edit distance Metrics (2003~2005)
  - Entropy Metric (2005)

Human Evaluation

- Four human experts are invited to evaluation the results
- Each expert is asked to evaluate all the translations, using a score ranged from 0 to 10, with at most one decimal
- For human experts, the results of the same source sentences are evaluated in the same time, however, for different source sentences, the order of the results of are given randomly.

Guidelines of Intelligible measurement (used before 2004)

Score	Description	Intelligibility
0	The translation is completely unintelligible.	0%
1	Readers cannot understand 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%

Test Point Method (1)

- Proposed by:

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

- A automatic MT evaluation system MTE-94 was developed based on this method

Test Point Method (2)

- Professor YU Shiwen was in charge of the 1994, 1995 and 1998 HTRDP machine translation evaluation.
- His later publications introduced the experiments of MTE-94 on the 1995 and 1998 HTRDP MT evaluation.
- Unfortunately, Prof. Yu did not give the real evaluation results in his publications and in official report of HTRDP MT evaluation.

Test Point Method (3)

- In test point method, detailed guidelines were given before the evaluation, which described all the test points for each MT direction.
- Some test points in Chinese-English machine translation:
  - Chinese word segmentation
    - Combinational disambiguation (马上 or 马 / 上 ?)
    - Overlapping disambiguation (的 / 确切 or 的确 / 切 ?)
  - Chinese POS tagging
    - N-V disambiguation (工作 work n. or v. ?)
    - N-Q disambiguation (头 “ head” or a quantifier?)
    - …

Test Point Method (4)

- (cont.):
  - Chinese parsing
    - N-N structural disambiguation (a modificative NP such as 木头椅子 , a coordinative NP such as 苹果香蕉 , or a subject-predicate clause such as 老王山东人)
    - ……
  - Chinese word sense disambiguation ……
  - Syntax structure transfer ……
  - English structure generation …… (e.g. position of aux. v.)
  - English word generation …… (e.g. form of irregular v.)

Test Point Method (5)

- Hundreds of test points were given by linguistics in the guidelines of each translation direction
- For each direction, a set of test sentences is given
- Each test sentence can be used to test more than one test points
- For each test sentence, simple substring matching is used to determine if the specific test point has been corrected processed, e.g. for the Chinese sentence, 我马上回来 , if the word “immediately” or “as soon as possible” occurs in the English translation, the test point “ 马上” is regarded to be correctly processed
- More then 3300 sentences is collected in MTE-94

Test Point Method (6)

- One of the earliest automatic MT evaluation
- Similar to the human’s standard test, such as TOFEL
- The idea is quite clever, however, the problem is, it is hard to define the test points and to construct the test set.

N-Gram Metrics and Edit Distance Metrics (1)

- N-Gram metrics is firstly proposed by:

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

N-Gram Metrics and Edit distance Metrics (2)

- A problem in using n-gram method to evaluation Chinese and Japanese translations: The n-gram cannot be clearly defined because of word segmentation ambiguities in Chinese and Japanese.
- Solution: character-based n-gram is used instead of word-based n-gram

Entropy Metric (1)

- A new method proposed by our group, which will be used in 2005 HTRDP MT evaluation
- Basic idea:
  - The MT system translation is firstly compared against the reference translations.　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 words (or characters), or an unmatched word (or character).
  - We assume that the more distributive the sentence is segmented, the poor the translation quality is. Thus we use a “distribution score” to evaluation the translation quality.
  - The distribution score can be well defined by the entropy, so we use the entropy to measure the translation quality.
  - Besides, some other factors, such as matching weight and length penalty, should also be taken into consideration.

Entropy Metric (2)

●	●	●		■	■	■	■		●	■	■	■	■	■

- Example
  - A MT system translation with 15 words
  - Matching all substrings in the translation against the reference translations, we get the above segmentations
  - The sizes of segmentations are: 3+1+4+1+1+5
  - The entropy of this segmentations is (without matching weight):

$$H = \sum -p \log_2 p$$

  - However, matched segmentations and unmatched segmentations should have different matching weight.
  - Considering the weights, we will get a weighted entropy.
  - The score of the translation is defined based on the weighted entropy, where length penalty is also considered.

Entropy Metric (3)

- In the n-gram metrics, it is quite subjective or experiential to determine the order of n-gram.
- Specifically, when we used character-based n-gram method to evaluate Chinese or Japanese translations, should we use a higher order of n-gram? Why? Which?
- Advantage: we do not need to select the order of n-gram in entropy method.
- In our experiments, entropy metric correlate with human evaluation quite well
- More details will be described in a future paper

Evaluation of Word Alignment

- The metrics include: Precision, Recall, F1-measure and Error Rate
- The metrics proposed by:

  Franz Josef Och, Hermann Ney. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.
- In the gold alignments, there are two kinds of alignment links: sure links and possible links.

Test Data (1)

- In early HTRDP MT evaluations (1994, 1995, and 1998)
  - The test sentences are selected by linguistics
  - Most of the sentences are short sentences covering specific test points, somewhat like sample sentences in grammar books

Test Data (2)

- In recent HTRDP MT evaluation (2003, 2004, and 2005)
  - The test data are mainly collected from real language
  - Both dialog data and text data are collected
  - Size: about 700-1000 sentences in each track
  - Domain:

	Dialog	Text
2003	Olympic	
2004	Olympic and general	
2005	Olympic	General

  - Where Olympic-related domain covers: weather, sports, travel, traffic, hotel, restaurant, and etc.

Test Data (3)

- Four reference translations are given to each test sentences
- All the reference translations are made by the native speakers of target language who are familiar with the source language
- The reference translations of C->J, E->J and J->E tracks are provided by our Japanese collaborator NICT

Test Data (4)

- For word alignment track, two people are asked to make the word alignment manually, according to a specification.
- The word links labeled by both labeler are regarded as sure links
- The links labeled by only one labeler are regarded as possible links

Training Data

- No training data were provided before 2004
- Training data are provided for only E->C and C->E tracks in HTRDP MT evaluation 2005
- Amount: 870,000 sentence pairs, which have been examined manually
- Up to now, no limit is made to the participants on the training data they can use. The participants can use any data to training their systems
- However, in the workshop, participants are asked to give a description to all the data used to training their systems.

Development data

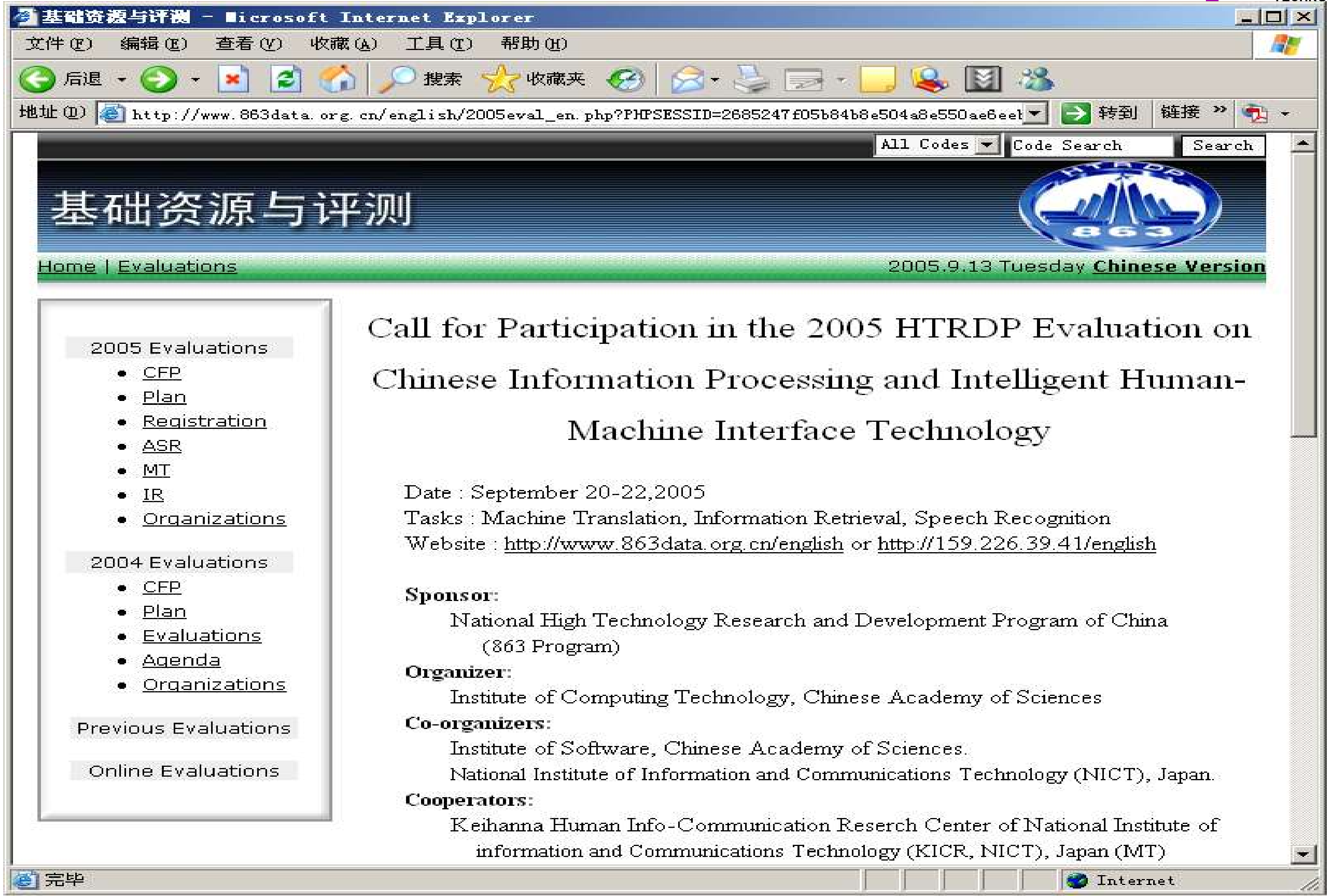
- No development data were provided before 2004
- Development data are provided for all tracks in 2005 evaluation
- For existing tracks before 2004, development data are just the test data and reference data used in 2003 and 2004 evaluations
- For new tracks (EJMT, JEMT and WACE), development data are newly created

Data Availability

- All the data are provided to participants freely, with a limited usage license agreement
- Others can purchase the research usage license of these data through ChineseLDC after the evaluation

HTRDP Evaluation Website http://www.863data.org.cn

INSTITUTE OF COMPUTING TECHNOLOGY


基础资源与评测 - Microsoft Internet Explorer
文件 (F) 编辑 (E) 查看 (V) 收藏 (A) 工具 (T) 帮助 (H)
后退
搜索
收藏夹
地址 (D) http://www.863data.org.cn/english/2005eval_en.php?PHPSESSID=2685247f05b84b8e504a8e550ae6eel
转到
链接
All Codes
Code Search
Search
基础资源与评测
HTRDP
863
Home | Evaluations
2005.9.13 Tuesday Chinese Version
完毕
Internet

2005 Evaluations

- CFP
- Plan
- Registration
- ASR
- MT
- IR
- Organizations

2004 Evaluations

- CFP
- Plan
- Evaluations
- Agenda
- Organizations

Previous Evaluations

Online Evaluations

Call for Participation in the 2005 HTRDP Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface Technology

Date : September 20-22,2005
Tasks : Machine Translation, Information Retrieval, Speech Recognition
Website : http://www.863data.org.cn/english or http://159.226.39.41/english

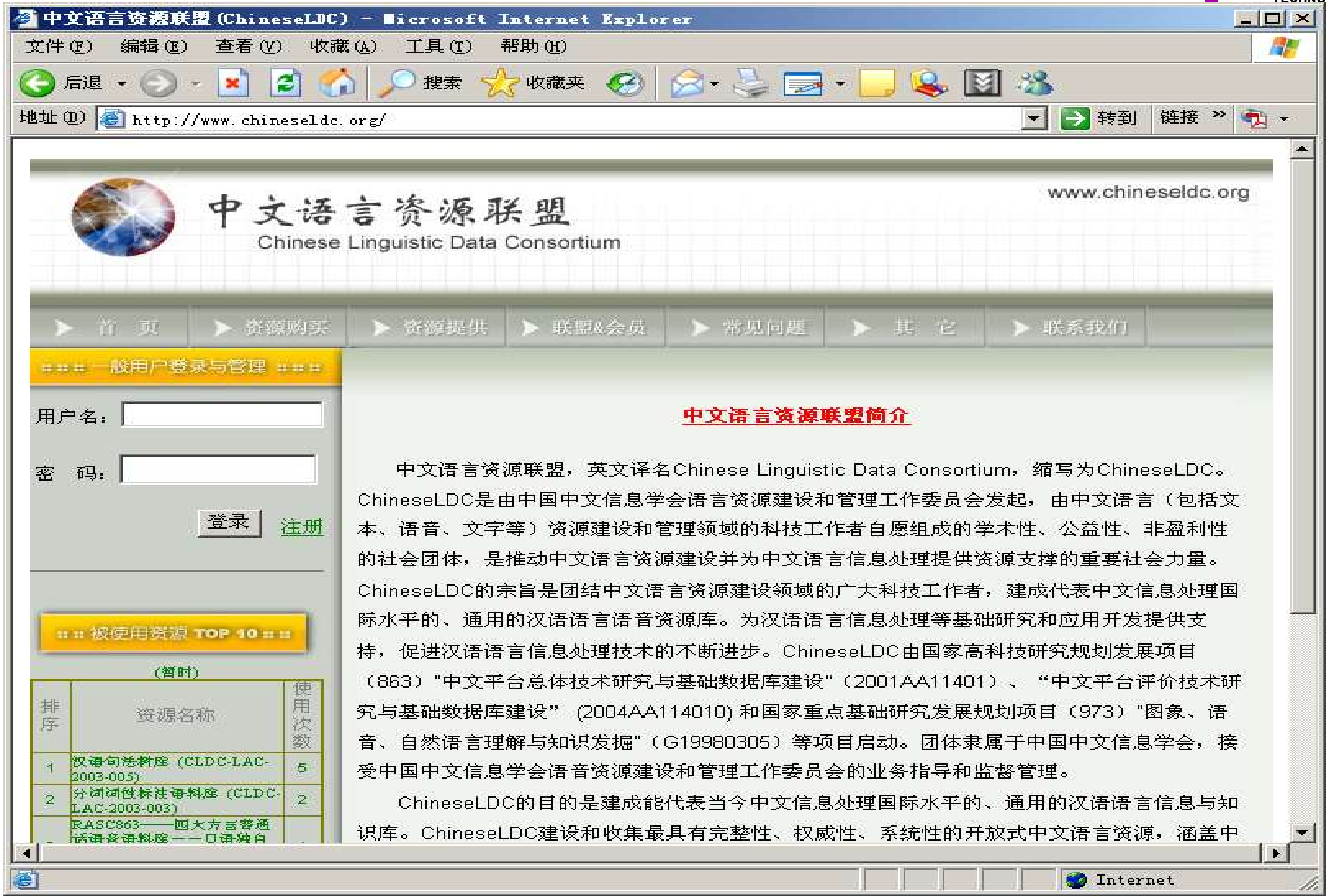
Sponsor:
National High Technology Research and Development Program of China (863 Program)

Organizer:
Institute of Computing Technology, Chinese Academy of Sciences

Co-organizers:
Institute of Software, Chinese Academy of Sciences.
National Institute of Information and Communications Technology (NICT), Japan.

Cooperators:
Keihanna Human Info-Communication Reserch Center of National Institute of information and Communications Technology (KICR, NICT), Japan (MT)

ChineseLDC Website http://www.chineseldc.org


中文语言资源联盟(ChineseLDC) - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退
搜索
收藏夹
地址(D) http://www.chineseldc.org/
转到
链接 »
www.chineseldc.org
中文语言资源联盟
Chinese Linguistic Data Consortium
首 页
资源购买
资源提供
联盟&会员
常见问题
其 它
联系我们
一般用户登录与管理
登录
注册
被使用资源 TOP 10
Internet

用户名：

密 码：

(暂时)

排序	资源名称	使用次数
1	汉语句法树库（CLDC-LAC-2003-005)	5
2	分词词性标注语料库（CLDC-LAC-2003-003)	2
	RASC863——四大方言普通话语音语料库——口语独白	

中文语言资源联盟简介

中文语言资源联盟，英文译名Chinese Linguistic Data Consortium，缩写为ChineseLDC。ChineseLDC是由中国中文信息学会语言资源建设和管理工作委员会发起，由中文语言（包括文本、语音、文字等）资源建设和管理领域的科技工作者自愿组成的学术性、公益性、非盈利性的社会团体，是推动中文语言资源建设并为中文语言信息处理提供资源支撑的重要社会力量。ChineseLDC的宗旨是团结中文语言资源建设领域的广大科技工作者，建成代表中文信息处理国际水平的、通用的汉语语言语音资源库。为汉语语言信息处理等基础研究和应用开发提供支持，促进汉语语言信息处理技术的不断进步。ChineseLDC由国家高科技研究规划发展项目（863）"中文平台总体技术研究与基础数据库建设"（2001AA11401）、“中文平台评价技术研究与基础数据库建设” (2004AA114010) 和国家重点基础研究发展规划项目（973）"图象、语音、自然语言理解与知识发掘"（G19980305）等项目启动。团体隶属于中国中文信息学会，接受中国中文信息学会语音资源建设和管理工作委员会的业务指导和监督管理。

ChineseLDC的目的是建成能代表当今中文信息处理国际水平的、通用的汉语语言信息与知识库。ChineseLDC建设和收集最具有完整性、权威性、系统性的开放式中文语言资源，涵盖中

Results: 2003 Dialog C→E

System	BLEU	NIST	Intelligibility
1	0.1747	5.9489	0.61575
2	0.1573	5.4694	0.438375
3	0.1099	5.5567	0.44625
4	0.3660	7.7722	0.731625
5	0.1823	6.0575	0.503875

Result: 2003 Text C→E

System	BLEU	NIST	Intelligibilit y
1	0.1186	5.3401	0.40325
2	0.0856	4.8462	0.319375
3	0.0556	4.6474	0.315875
4	0.1762	6.3113	0.464375
5	0.1095	5.5097	0.376

Results: 2004 Dialog C→E

ID	Automatic					Human
	NIST	BLEU	GTM	mWER	mPER	Intelligibility(%)
System1	5.8301	0.1896	0.6477	0.6165	0.4916	49.060
System4	4.5335	0.1279	0.5481	0.6909	0.5745	32.927
System6	6.1223	0.2094	0.6607	0.6202	0.4805	52.320
System7	4.4259	0.1009	0.5245	0.7392	0.6125	34.245
System9	5.4762	0.1540	0.5978	0.7225	0.5720	40.153
System10	5.7492	0.1697	0.6285	0.6830	0.5180	42.650

Results: 2004 Text C→E

ID	Aumotic					Human
	NIST	BLEU	GTM	mWER	mPER	Intelligibility(%)
System1	5.6075	0.1201	0.6569	0.7793	0.5750	52.720
System4	4.2326	0.0807	0.4813	0.8531	0.6868	32.768
System6	5.6274	0.1217	0.6331	0.7723	0.5639	52.110
System7	3.8949	0.0573	0.4904	0.8471	0.6874	36.258
System9	5.0503	0.0790	0.5475	0.8487	0.6428	39.452
System10	5.0898	0.0912	0.5696	0.8366	0.6347	39.437

MT evaluation: HTRDP vs. NIST

- HTRDP focus mainly on translations to and from Chinese and Japanese, while NIST focus on translations to English. There are much more translation directions in HTRDP than those in NIST
- New evaluation metric (entropy) will be used in HTRDP.
- The domain and genre of HTRDP test data is quite different from NIST test data.
- In our unofficial experiments, for some MT system, the HTRDP 2003 BLEU score is much lower than NIST 2005 BLEU score (about 0.06-0.10). Maybe it is because the diversity of the HTRDP test data.

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 and development data are provided to the participants freely from this year.

Future work

- In recently years, we will hold MT evaluation annually
- Provide more training data
- Research on better evaluation metrics

Participants from all over the world are welcome to HTRDP MT evaluation.

Thanks