Guidelines for CWMT2008 Machine Translation Evaluation

I. Introduction

The 4th China Workshop on Machine Translation (CWMT) will be held in November 27-28, 2008. It is the continuation of the first three Symposiums of Statistical Machine Translation of China (SSMT), which were held in 2005, 2006 and 2007 respectively.

As a convention, a machine translation evaluation will be organized along with CWMT2008, to promote the substantial interaction among the participants and the advances of machine translation research in China.

The sponsor of CWMT2008 machine translation evaluation is:

Chinese Information Processing Society of China

The organizer of this evaluation is:

Institute of Computing Technology, Chinese Academy of Sciences

The cooperators of this evaluation include:

Institute of Scientific and Technological Information of China
Microsoft Research Asia

The resource providers of this evaluation include:

Peking University
Harbin Institute of Technology
Xiamen University
Wanfang Data Corporation
Institute of Scientific and Technological Information of China
Institute of Automata, Chinese Academy of Sciences
Institute of Computing Technology, Chinese Academy of Sciences

The president of this evaluation is:

Qun LIU (Institute of Computing Technology, CAS)

The committee members of this evaluation include:

Degen HUANG (Dalian University of Technology)
Chin-yew LIN (Microsoft Research Asia)
Yajuan LU(Institute of Computing Technology, CAS)
Huilin WANG (Institute of Scientific and Technological Information of China)
Muyun YANG (Harbin Institute of Technology)
Yujie ZHANG (National Institute of Information and Communication Technology)
Chengqing ZONG (Institute of Automata, CAS)

For more information about CWMT2008 and the machine translation evaluation, please visit:

http://www.cipsc.org.cn/cwmt-2008.html
http://www.nlpr.ia.ac.cn/cwmt-2008.html

II. Evaluation Tasks

The evaluation tasks in this evaluation include:

Language and domain	Evaluation Task	Task ID
Chinese->English News	Machine Translation	ZH-EN-NEWS-TRANS
Chinese->English News	System Combination	ZH-EN-NEWS-COMBI
English->Chinese News	Machine Translation	EN-ZH-NEWS-TRANS
English->Chinese Science & Technology	Machine Translation	EN-ZH-SCIE-TRANS

1. “Machine Translation” Task

The “machine translation” evaluation task will adopt similar methods which are used in other international machine translation evaluations. Firstly, the evaluation organizer will provide the test data to the evaluation participants, then the participants will return the translation results to the organizer, and finally the organizer will assess the results. Automatic evaluation metrics will be used in this task.

In the “machine translation” task, the participants are allowed to use the data which are not in the resource list provided by the organizer. However, the systems using out-of-list data will be marked in the evaluation result reports.

The participants of “machine translation” task should submit one primary result, and at most two contrast results.

For the Chinese-English “machine translation” task in news domain, the results of “machine translation” task will be used as the input of the “system combination” task. So the participants of the “machine translation” task are asked to complete some additional work:

1) Besides submitting the results on the CWMT2008 evaluation data, the participants are asked to submit the results of the same participating system on the SSMT2007 evaluation data, which will be used as the input data for “system combination” task. So for the Chinese-English “machine translation” task in news domain, the SSMT2007 evaluation data are not allowed to be used as training data. For the introduction of SSMT2007 evaluation data, please refer to the Appendix X (ChineseLDC resource ID: 2007-863-001).

2) If possible, please submit N-best translations, which are sorted by the scores in descending order, both for the CWMT2008 evaluation data and the SSMT2007 evaluation data. Here N is no more than 200. In the “machine translation” task, only the translation with the highest score will be used in the assessment, while other

translations will be used as the input data in the "system combination" task. This requirement can be ignored if the participating system adopts a method which is unable to generate N-best results.

2. "System Combination" Task

1) When the "machine translation" task is finished, the organizer will collect all the N-best translations submitted by the participants of "machine translation" task, and send them to the participants of "system combination" task.

2) The participants of "system combination" perform combinations on these results, and submit the combined result to the organizer. On the system combination procedure, the participating sites may make use of the results of the given systems on the SSMT2007 evaluation data and the reference translations to obtain the prior knowledge of the performance of these systems.

3) The organizer will assess all these combined results and give the evaluation reports.

4) In the "system combination" task, the training data are limited to the resource list given by the organizer. In other words, the participants cannot use the data out of the list to train their combination systems.

III. Evaluation metrics

1. Metrics for "Machine Translation" Task

The automatic evaluation metrics of "machine translation" task include: BLEU, NIST, GTM, mWER, mPER, ICT and Woodpecker. All these metrics are case-sensitive. The evaluation of Chinese translation is based on Chinese characters instead of words. All the non-Chinese characters in the Chinese translations such as digits, alphabets and punctuations, are required to be transformed to half-width characters.

The Woodpecker toolkit is a linguistic evaluation platform to evaluate the linguistic translation capability of machine translation (MT) systems. Instead of assigning a general score to an MT system, Woodpecker evaluates the capability of the system in handling various important linguistic test cases called Check-Points. A Check-Point is a linguistically motivated unit, (e.g. an ambiguous word, a noun phrase, a verb-object collocation, a prepositional phrase, a new word etc.) which is pre-defined in a linguistic taxonomy for diagnostic evaluation. The reference of a check-point is its corresponding part in the target sentence. On the basis of syntax parse trees and word alignments over source testing data and target reference data, Woodpecker can automatically extract various check-points from source and target sentences and evaluate the translation performance of an MT system from the point of view of linguistics by computing the matching degree between its translation result and the references of check-points. Please refer to Appendix VI for detailed information about WoodPecker system.

2. Metrics for "System Combination" Task

In the "system combination" task, the translation quality of the generated results will be assessed. BLEU will be the only metric used in the "system combination" task.

IV. Evaluation data

1. Description of evaluation data

a) Training Data

The organizer will provide some language resources, for the participants training their system. The list of resources provided by the organizer is given as an appendix of this document.

b) Test Data

The test data are collected from two domains: News and Science & Technology.

c) Cut-off Date

To ensure that there is no overlap between the training data and test data, a Cut-off Date is specified. The cut-off date for this evaluation is January 1st, 2008.

All the training data and development data, including the data provided by the organizer and the data collected by the participants themselves, should be generated before (not including) the cut-off date.

The test data provided by the organizer are generated after (including) the cut-off date.

2. Limitation of the usage of training data

a) "Machine Translation" Task

In the "machine translation" task, any data can be used in training.

If the participants use data out of the list provided by the organizer (hereafter referred as out-of-list data), they should declare if this out-of-list data can be accessed publicly. If it can be accessed publicly, the participants should give the origin of the data; or if it can not be publicly accessed, the participants should describe the content and amount of the data in detail.

In the final report of the evaluation results, the systems which use out-of-list data for training will be marked explicitly.

Please note that, for Chinese to English "machine translation" task in news domain, the SSMT2007 evaluation data (ChineseLDC resource ID: 2007-863-001) cannot be used as training data.

b) "System Combination" Task

In the "system combination" task, only the data in the resource list provided by the organizer can be used. No out-of-list data can be used for training.

V. The Evaluation Calendar

1.	Registration deadline	August 31st, 2008
2.	Release of the training data	August 31st, 2008
3.	Release of the test data for the "Machine Translation" task	October 8th, 2008
4.	Submission deadline for the results on the test data and the system descriptions of the "Machine Translation" task	October 15th, 2008
5.	Release of the test data of "System Combination" task (i.e. the collected translation results of "Machine Translation" task) to participants of "System Combination" task	October 23rd, 2008
6.	Submission deadline for the results and system descriptions of the "system combinations" task	October 30th, 2008
7.	Notification of the evaluation results of all the tasks to all the participants	November 8th, 2008
8.	Submission deadline for the technical reports of all the participants	November 15th, 2008
9.	Workshop	November 27th-28th, 2008

VI. Appendixes

This document includes the following appendixes:
Appendix A: The Registration Form
Appendix B: the Format of Input and Output Files
Appendix C: Requirements of System Description
Appendix D: Requirements of Technical Report
Appendix E: Resource List Released by the Organizer
Appendix F: Introduction of the Woodpecker Machine Translation Evaluation System

Appendix A: Registration Form

Any organization who is engaged on research or development of machine translation can register for CWMT2008 machine translation evaluation, whatever the technology is used, for example, the rule-based, the example-based, or the statistical-based technologies. The participating site of CWMT2008 evaluation should fill the following form, and send it to the organizer by both email and post (or fax). In the post (or fax), there should be a signature of the person in charge or a stamp of the participating organization.

The participating site should pay registration fee for the evaluation. The registration fee includes the license fee of the training data resources, the registration fee of CWMT2008 for one person from the participating site, and part of the cost of the organization of the evaluation. The registration fee of CWMT2008 machine translation evaluation is: 3000RMB for participants from China mainland, or 1000USD for participants from Hong Kong, Macau, Taiwan and foreign countries.

The deadline of registration is: August 31st, 2008

Contact Information：

Name: Ms. ZHAO Hong-mei

Email: zhaohongmei@ict.ac.cn

Address: Institute of Computing Technology, CAS
No.6 Kexueyuan South Road, Zhongguancun, Haidian
P. O. Box 2704, Beijing 100190, P. R. China

Post Code: 100190

Telephone: +86-10-62600667

Fax: +86-10-82611846

Registration form for CWMT2008 Machine Translation Evaluation

Evaluation Task	● Machine Translation □ Chinese→English, News Domain □ English→Chinese, News Domain □ English→Chinese, Science & Technology Domain ● System Combination □ Chinese to English, News Domain		
Affiliation			
Address			
Contact person		Telephone	
Post code		Email	

The participating site agrees to commit to the following terms:

1. After receiving the evaluation data, the participating site will process the entire test set following the evaluation guidelines, and submit the results to the evaluation organizer before the submission deadline.
2. The participating site will submit a technical report of the participating systems and attend the 4th China Workshop on Machine Translation (please refer to the "call for papers" of CWMT2008)
3. The participating site confirms that it have the intelligent property of the participating system. If any technology in the participating system is licensed from other person or organization, it will be clearly d

escribed in the system description.

4. The participating site agrees that all the evaluation data (including training data, development data, test data, reference translation, and evaluation tools) acquired from the organizer during the evaluation process will only be used for the research purpose of participating the evaluation. No other purpose of usage is permitted.
5. The participating site agrees that the evaluation data will only be used within the research group who participates the evaluation, and it will neither be distributed by any way (written, electronically, or by network), nor will it be used in the cooperators or affiliation organizations of the participating site.
6. The participating site agrees that if any result which is yielded by using the evaluation data (including training data, development data, test data, reference translation, and evaluation tools) is published, the fact of using these data will be publicly acknowledged.

Signature of the person in charge or stamp of the participating site:

Date:

Appendix B: The Format of Input and Output Files

1、 Naming Rule of Submitted Result Files

The participating sites should name the submitted result files in the following form:

Evaluation_Task – Participating_Site – “primary/constract” – system_name . xml

For example, in the Chinese to English “machine translation” in news domain, the Institute of Computing Technology (ICT) will submit one result of primary system (systema) and two results of contrast systems (systemb and systemc), these three files should be named as:

```
zh_en_news_trans-ict-primary-systema.xml
zh_en_news_trans-ict-contrast-systemb.xml
zh_en_news_trans-ict-contrast-systemc.xml
```

2、 File Format for “Machine Translation” Task

The test data of source language is stored in a “source file”. The participating sites are required to generate a corresponding “target file” for the “source file”. The “source file” and the “target file” are all in standard XML format.

(1) Format of Source File

The “source file” is a XML file, encoded in UTF-8.

The “source file” includes a <*srcset*> element, which have three attributes: *setid*, *srclang*, and *tgtlang*. The <*srcset*> element includes several <*doc*> elements (the part between tag <*doc*> and </*doc*>), where each <doc> element is correspondent to an article to be translated, and the attributes of <*doc*> elements describe the information of the article. The *docid* attribute gives the title of the article, whose value should be quoted in double quotation marks. The *srclang* attribute depicts the source language code, while the *tgtlang* attribute depicts the target language code. The language code “*en*” is for English, “*ja*” is for Japanese, and “*zh*” is for Chinese.

Each <*doc*> element consists of several <*p*> elements (the part between tag <*p*> and </*p*>). Each <*p*> element consists of several <*s*> elements (the part between tag <*s*> and </*s*>). The value of attribute *id* of <*s*> element is a positive integer. The values of *id* attributes of <*s*> elements should be unique in a <*doc*> element; however, these values should not be consecutive values. Each <*s*> element may contain one or more sentences. <*s*> elements may also directly belong to a <*doc*> element without a <*p*> element.

```
<?xml version="1.0" encoding="UTF-8"?>

<srcset setid="zh_en_news_trans" srclang="zh" tgtlang="en">
```

```
<doc docid="文档名称">

<p>

<s id=”1”> 玻利维亚举行总统与国会选举 </s>

</p>

<p>

<s id=”2”> (法新社玻利维亚拉巴斯电)玻利维亚今天举行总统与国会选举，投票率
比预期更高，选民希望选出的新领导阶层能够振兴经济，改善人民的生活水准，抑
制这个南美洲最贫穷国家的劳工骚动。 </s>

</p>

<p>

<s id=”3”> 投票所于下午四时(台北时间七月一日清晨四时)关闭，选务人员说，选
举结果将于两小时之后开始发布。 </s>

</p>

<p>

<s id=”4”> 稍早，玻利维亚总统与参与选举的候选人援引巴西赢得世足赛冠军为
例，鼓励民众踊跃投票，虽然联邦法律规定，凡达投票年龄的玻利维亚人都必须投
票。 </s>

</p>

</doc>

</srcset>
```

(2) Format of Target File

The “target file” is also an XML format file. The format of a “target file” is similar with the format of a the “source file”, and the contents of a “target file” are correspondent to a “source file”.

For each <*srcset*> element in the “source file”, there should be a <*tgtset*> element in the correspondent “target file”, whose attributes are the same.

A <*tgtset*> element should contain a <*system*> element, which includes the description of the system who generates the “target file”. A <*system*> element should have two attributes: the *site* attribute depicts the name of the participating site, and the *sysid* attribute depicts the id of the participating system. A participating site can submit several results which are generated by different participating systems. The descriptions of these different participating systems should be included in the <*system*> elements.

A <tgtset> element should also include several <doc> elements which is correspondent to the <doc> elements in the “source file”. A <doc> element and the <p> elements and <s> elements in it should be correspondent to the same elements in the “source file”. And the value of the docid attribute of the <doc> element and the id attribute of the <s> element should also be the same with those in the “source file”. A <s> element should also include one segment of text and several <cand> elements. The text in the <s> element is the best translation of the text in the correspondent <s> element in the “source file”, which will be used for the evaluation of the “machine translation” task. A <s> element should also have a score attribute, whose value will be the score of the best translation. The multiple <cand> elements provide n-best candidate translations for the “system combination” task. Each <cand> element contains one candidate translation. Each <cand> element should also have a score attribute, whose value is the score of the candidate translation. The value of score attribute of a <s> element and the <cand> element should be a real value between 0 and 1. The high the score is, the better the translation is. The candidate translation should be sorted by the score from high to low.

```
<?xml version="1.0" encoding="UTF-8"?>
<tgtset setid="zh_en_news_trans" srclang="zh" tgtlang="en">
<system site="单位名称" sysid="系统标识">
这里给出参评系统的描述信息
...........
...........
...........
</system>
<doc docid="文档名称">
<p>
<s id="1" score=0.03>
Bolivia Holds Presidential and Parliament Elections
<cand score=0.005> Bolivia Holds Elections for Presidential and Parliament </cand>
<cand score=0.0007> ... </cand>
……
</s>
</p>
<p>
<s id="2" score=0.5>
 (AFP, La Paz, Bolivia)    Bolivia held its presidential and parliament elections today.
```

With a higher than expected turn-out rate, voters hope the newly elected leadership can revitalize the economy, improve the people's living standards and control the labor unrest in this poorest country in South America.

<cand score=0.2> … </cand>

<cand score=0.003> … </cand>

……

</s>

</p>

<p>

<s id="3" score=0.01>

The polling stations closed at 4 p.m. (4 a.m. on July 1, Taipei time).   The polling staff said that the results of the elections will be released within two hours.

<cand score=0.0003> … </cand>

<cand score=0.000009> …. </cand>

<cand score=0.000001>…. </cand>

…….

</s>

</p>

<p>

<s id="4" score=0.0002>

Earlier, the Bolivian president and candidates in the elections, citing Brazil's championship at the World Cup soccer tournament, encouraged the public to actively participate in the elections even though every Bolivian who has reached the voting age is required by the federal law to vote.

<cand score=0.00001> … </cand>

<cand score=0.000005> … </cand>

……

</s>

</p>

</doc>

*</**tgtset**>*

3、 File Format for “System Combination” Task

The input files of “system combination” task are exactly the output files of the “machine translation” task.

The format of output files for “system combination” task is similar with the format of output file for “machine translation” task, except that only 1-best result should be generated and the scores of the translations are not required.

Appendix C: Requirements of System Description

The participating site should give a system description which is embedded in the result XML file when it is submitted. The system description should include the following information:

- The hardware and software environments, which include: operation system and its version, CPU number, CPU type and frequency, Memory size, and etc.
- Execution Time: The time from accepting the input to generating the output.
- Technology outline: A outline of the main technology and parameters of the participating system
- Training Data: A description of the training data and development data used for system training;
- External Technology: A declaration of the external technologies which are used in the participating system but not owned by the participating site, including: open source codes, free software, shareware, and commercial software.

Appendix D: Requirements of Technical Report

All participating site should submit a technical report to the 4th China Workshop of Machine Translation (CWMT2008). The technological report should describe the technologies which are used in the participating system in detail, in order to tell the reader how the results are obtained. A good technological report should be detailed enough so that the reader could repeat the work which is described in the report. The technology should be no short than 5000 Chinese characters or 3000 English words.

Generally, a technology report should provide the follow information:

Introduction: Give the background information; introduce the evaluation tasks, and the outline of the participating systems;

System: Describe the architecture and each module of the participating system in detail. The technologies used in the system should be focused. If there is any open technology adopted, it should be explicitly declared; if the technologies are developed by the participating site itself, it should be described in detail.

Date: Give detailed description to the data used in the system training and the processing of the data.

Experiment: Give detailed description to the experiment process, the parameters, and the results on the evaluation. Analyze the results.

Conclusion.

Appendix E: Resource List Released by the Organizer

ChineseLDC Resource ID	Resource Description	
CLDC-LAC-2003-004	Provider	Institute of Computing Technology, CAS Institute of Automata, CAS
	Languages	Chinese-English
	Domain	Multi-domain
	Size	The original corpus include 3384 bilingual text files, which contains 209486 Chinese-English sentence pairs, where 3098 files with 107436 sentence pairs were developed by Institute of Automata (IA), CAS, and the other 250 files with 102050 sentence pairs were developed by Institute of Computing Technology (ICT), CAS. The current version is an extended version with additional data which was also provided by IA and ICT.
	Description	The resource is developed under the support of the-973-project. It is a large-scale Chinese-English bilingual Corpus on multi-domain and multi-style which is sentence-aligned.
CLDC-LAC-2003-006	Provider	Institute of Computational Linguistics, Peking University
	Languages	Chinese-English, Chinese-Japanese
	Domain	Multi-domain
	Size	200K Chinese-English sentence pairs, where 10K sentence pairs are word-aligned; 20K Chinese-Japanese sentence pairs.
	Description	The corpus is developed by the Institute of Computational Linguistics, Peking University (ICL/PKU), supported by an 863 subproject with title "Chinese-English / Chinese - Japanese parallel corpora"(project no. 2001AA114019). Only Chinese-English sentence-aligned part will be provided in the CWMT2008 evaluation.
	Provider	Xiamen University
	Languages	English - Chinese
	Domain	Dialog
	Size	176148 sentence pairs
	Description	Subtitles of films
	Provider	IR laboratory of Harbin Institute of Technology
	Languages	English - Chinese
	Domain	Multi-domain
	Size	100000 sentence pairs
	Description	
	Provider	Machine Translation Group of Harbin Institute of Technology
	Languages	English - Chinese
	Domain	Multi-domain
	Size	52227 sentence pairs

	Description	
	Provider	Institute of Scientific and Technological Information of China
	Languages	English - Chinese
	Domain	Science & Technology
	Size	
	Description	Abstracts of English science and technology papers and their Chinese translations
	Provider	Institute of Institute of Computing Technology, CAS Wangfang Data Corporation
	Languages	Chinese - English
	Domain	Science & Technology
	Size	100K bilingual abstracts
	Description	Chinese-English bilingual abstracts extracted from Chinese Journals
2007-863-001	Provider	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese – English, English - Chinese
	Domain	News
	Size	The test data of this evaluation contains data of 2 translation directions (Chinese-English and English-Chinese) in news domain. The Chinese-English data contain 1002 Chinese sentences with 42256 Chinese characters. The English-Chinese data contains 955 English sentences with 23627 English words. There are 4 reference translations made by human experts for each test sentence. The test data for word alignment task in this evaluation is Chinese-English sentence pairs, where 251 sentence pairs with 7030 Chinese characters and 8109 English words are translated from Chinese to English, and 253 sentence pairs with 6872 Chinese characters and 6981 English words are translated from English to Chinese. The data is in news domain also. Each Chinese-English sentence pair is labeled with word alignments by two persons independently.
	Description	The test data, documents and software for SSMT2007 machine translation evaluation. Only the evaluation data of machine translation task will be provided in the CWMT2008 evaluation, without the evaluation data of word alignment task. It should be noted that, this resource cannot be used as the training data in the Chinese to English "machine translation" task in news domain.
2005-863-001	Provider	Institute of Computing Technology, Chinese Academy of Sciences National Institute of Information and Communication Technology
	Languages	There are 6 translation directions for the machine translation tasks: Chinese to English, English to Chinese, C

		hinese to Japanese, Japanese to Chinese, English to Japanese, Japanese to English. There is only one sentence pair for word alignment task: Chinese – English.
	Domain	The data for machine translation task contains two types of data: one is dialog data in Olympic-related domains, which include game reports, weather forecasts, traffic and hotels, travel, foods, etc, the other one is text data in news domain. The data for word alignment test is text data on news domain.
	Size	Data for machine translation task: The Chinese to English and Chinese to Japanese tasks share the same test data with 460 sentences. The English to Chinese and English to Japanese tasks share the same test data with 460 sentences. The Japanese to Chinese and Japanese to English tasks share the same test data with 460 sentences. There are 4 reference translations made by human experts for each source sentence in each translation direction. The data for word alignment evaluation contain 505 sentence pairs, which is labeled with word alignments by two human experts independently.
	Description	The test data, documents and software for the 2005 "863" machine translation evaluation. Only the evaluation data of Chinese to English and English to Chinese "machine translation" tasks will be provided for the CWMT2008 evaluation, without the evaluation data for "word alignment" task.
2004-863-001	Provider	Institute of Computing Technology, Chinese Academy of Sciences National Institute of Information and Communication Technology
	Languages	The data for machine translation tasks is in 5 translation directions: Chinese to English, Chinese to Japanese, Chinese to French, English to Chinese, Japanese to Chinese.
	Domain	There are two data types for this evaluation: one is text data, the other is dialog data. The data cover general domain and Olympic-related domains which include game reports, weather forecasts, traffic and hotels, travel, foods, etc.
	Size	Data for machine translation task: The tasks of Chinese to English, Chinese to Japanese, Chinese to French share the same test data where the dialog data contains 2 files with 400 sentences, and the test data contains 24 files with 308 sentences. The data from English to Chinese contains dialog data of 2 files with 400 sentences and text data of 29 files with 310 sentences. The data from Japanese to Chinese contains dialog data of 2 files with 400 sentences and text data of 16 files with 309

		sentences. There are 4 reference translations made by human experts for each source sentence in each translation direction.
	Description	The test data, documents and software for the 2004 “863” machine translation evaluation. Only the part of Chinese to English and English to Chinese will be provided for the CWMT2008 evaluation.
2003-863-001	Provider	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	The data for machine translation tasks is in 4 translation directions: Chinese to English, English to Chinese, Chinese to Japanese, Japanese to Chinese.
	Domain	The data covers Olympic-related domains which include game reports, weather forecasts, traffic and hotels, travel, foods, etc.
	Size	Data for machine translation task: The Chinese to English and Chinese to Japanese tasks share the same test data, which contains dialog data of 437 sentences and text data of 15 files. The data of English to Chinese contains dialog data of 496 sentences and text data of 13 files. The data of Japanese to Chinese contains dialog data of 410 sentences and text data of 30 files. There are 4 reference translations made by human experts for each source sentence in each translation direction.
	Description	The test data, documents and software for the 2003 “863” machine translation evaluation. Only the part of Chinese to English and English to Chinese will be provided for the CWMT2008 evaluation.

Appendix F: Introduction of the Woodpecker Machine Translation Evaluation System

1. What is WoodPecker?

WoodPecker is a linguistic diagnostic evaluation platform for machine translation developed by Natural Language Computing Group, Microsoft Research Asia. Different from other evaluation methods (e.g. Bleu) assigning a general score to an MT system, WoodPecker evaluates the system capability of translating various linguistic test cases contained in the test data, where each linguistic test case is called as a *Check-Point*.

WoodPecker platform has two functions. On the one hand, WoodPecker can automatically extract check-points with the help of word alignment and syntax parse trees over source test data and target reference data. On the other hand, WooodPecker computes the matching degree between candidate translation and the reference of check-points, which is used to indicate the capability of translating the specific linguistic cases.

The evaluation method in WoodPecker platform extends the work by (Yu, 1993) which proposed the MTE evaluation method based on check-points. (Yu, 1993) mainly suggested constructing linguistic check-points by human experts, which will be costly when manually building the check-points for the new test corpus. WoodPecker can automatically construct a large batch of check-points based on syntax parser and automatic word alignment tools. Meanwhile, it can produce test set for a specific check-point. WoodPecker adopted a series of methods to enhance the quality of extracted check-points, such as choosing simple test sentences, selecting high-quality check-points via the combination of multiple syntax parsing trees, augmenting test corpus to overcoming noisy from word alignments, and so on. Furthermore, the scoring method in (Yu, 1993) is binary where one is got for the whole-matching and zero for partial or no matching. However, WoodPecker uses popular n-gram based matching method which can capture a broader coverage of different levels of matching degree, especially for the case of partial matching. At present, WoodPecker can not only support English-to-Chinese evaluation task, but also support Chinese-to-English evaluation task. As it can extract check-points automatically, it has good adaptability to evaluate the translation task for a new pair of languages.

2. Check-point Extraction

A check-point defined in WoodPecker is a linguistically motivated unit, (e.g. an ambiguous word, a noun phrase, a verb-object collocation, a prepositional phrase etc.) which can be created at different syntactic levels. Based on the extracting source, check-points can be classified into two classes of source check-points and target check-points. The reference of a check-point refers to its corresponding target part contained in it, where the references of source check-points are the target contents determined by the word alignment. The check-point evaluation is conducted by computing the matching degree between the reference of check-points and system translation. Each check-point can contain multiple references while the test sentence corresponds to multiple reference sentences.

All check-point categories can form two independent taxonomies of Chinese-to-English and English-to-Chinese based on the translation task type. Each kind of taxonomy includes three levels of word, phrase and sentence, where each level consists of several groups and each group contains several sub-groups or categories. Thus, the whole taxonomy forms a top-down tree structure where leaf nodes correspond to check-point categories (e.g. Noun, Verb, Preposition, etc.). Each occurrence of a check-point category in test corpus denotes a concrete check-point. Therefore, a check-point category may correspond to more than one check-point in a test sentence. To meet the evaluation requirement, any check-point group and check-point category can be customized into a new check-point group by users. During the evaluation, only the check-point categories covered by the customized group are regarded as evaluation objects. Meanwhile, the evaluation result reports not only the score of each category, but also the sore of each group.

For simplicity, the following table gives some examples of check-points for Chinese-to-English translation task. The detailed list of check-point categories for Chinese-to-English and English-to-Chinese translation tasks can refer to Table 1 and Table 2.

Chinese-to-English Check-points		
Level	**Name**	**Example**
Word	Ambiguous word	打(play)
	New word	朋克(Punk)
	Preposition	于(in)，在(at)
Phrase	Collocation	油炸-食品(fired – food)
	Repetitive word combination	看看(have a look)
	Subject-predict phrase	他*说, (he*said)
Sentence	“BA” sentence	他把(BA)书拿走了. (He took away the book.)
	“BEI” sentence	花瓶被(BEI)打碎了. (The vase was broken.)

Based on above taxonomy, WoodPecker can automatically extract all check-points for each translation task with the following steps:

(1). Prepare bilingual test corpus where each source test sentence can correspond to multiple reference sentences.

(2). Parse the sentences from source and target languages and obtain the POS-tag of each word, dependencies and hierarchy information between words. The available parser includes Stanford statistical parser and Berkeley statistical parser.

(3). Perform the word alignment for each sentence pair. Any automatic word aligner can be used, such as GIZA++. Manually aligned results are also acceptable.

(4). Extract all kinds of check-points based on the parsed sentence pairs and determine the reference of source check-points via word alignments.

3. Check-point Evaluation

Based on the selected check-point set, WoodPecker can produce the evaluation result by computing the matching degree between the reference of check-points and system translation.

Given the check-point category c and the system translation t, the matching score between t and c can be computed by the formula (1):

$$Score(c) = Recall(c) \times Penalty \qquad (1)$$

where the function *Recall* computes the recall of n-gram matching between t and the reference of c according to the formula (2). The function *Penalty* is used to penalize the redundant n-grams in the system translation. Actually, it penalizes the ratio of average sentence length over the whole reference R and translation T as shown in formula (3):

$$Recall(c) = \frac{\sum_{r \in R^*} (DM(r) \times \sum_{n-gram \in G(r)} Match(n-gram))}{\sum_{r' \in R^*} (DM(r') \times \sum_{n-gram' \in G(r')} Count(n-gram'))} \qquad (2)$$

$$Penalty = \begin{cases} \dfrac{length(R)}{length(T)} & if\ length(T) > length(R) \\ 1 & Otherwise \end{cases} \tag{3}$$

In formula (2), R^* is the set of the best reference r^* of each check-point, where r^* is determined by formula (4); $G(r)$ represents the set of n-gram generated by the reference r of the check-point. The function *Count* denotes the total count of contained n-grams; The function *Match* captures the count of the n-gram occurring in the translation t, that is, compute the matching count. The function *DM*, defined by formula (5), is used to estimate the quality of the reference of the check-points which helps reduce the evaluation error caused by the noisy of word alignment during the check-point extraction. In formula (5), $Dic(c)$ gives the translation entries corresponding to the source part of c in the bilingual dictionaries, $CoCnt(x,y)$ denotes the count of common words in x and y, $WordCnt(x)$ represents the count of word in x.

$$r^* = \arg\max_{r \in R}(DM(r) \times \frac{\sum_{n-gram \in G(r)} Match(n-gram)}{\sum_{n-gram' \in G(r)} Count(n-gram')}) \tag{4}$$

$$DM(r) = \begin{cases} Max\{0.1, \dfrac{CoCnt(r, Dic(c))}{WordCnt(r)}\} & \text{if the reference of } c \text{ is got by word alignments} \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

Based on above formulas, Woodpecker conducts the check-point evaluation in terms of the following steps:

(1). Prepare the test sentences and the system translation.

(2). Specify the check-point set C and the evaluation task type (Chinese-to-English or English-to-Chinese).

(3). For each check-point in C, calculate the number of matched n-grams of the reference against the system translation and make necessary normalization.

(4). Compute the matching score of a check-point category by summing up the score of all check-point of this category. Compute the matching score of the MT system by summing up the score of all categories.

(5). Provide the detail information and matching score of all kinds of check-points for an MT system.

Table 1: Check-point category list for Chinese-to-English evaluation task

Word level		
Ambiguous word	New word	Idiom
Noun	Verb	Adjective
Pronoun	Adverb	Preposition
Quantifier	Repetitive word	Collocation
Phrase level)		
Subject-predicate phrase	Predicate-object phrase	Preposition-object phrase
Measure phrase	Location phrase	…
Sentence level		
BA sentence	BEI sentence	SHI sentence
YOU sentence	N/A	

Table 2: Check-point category list for English-to-Chinese evaluation task

Word level		
Noun	Verb (with Tense)	Modal verb
Adjective	Adverb	Pronoun
Preposition	Ambiguous word	Plurality
Possessive	Comparative & Superlative degree	…
Phrase level		
Noun phrase	Verb phrase	Adjective phrase
Adverb phrase	Preposition phrase	…
Sentence level		
various kind of clauses defined by leading words		

Reference:

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, Tiejun Zhao. *Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points*. Coling 2008.

Shiwen Yu. 1993. *Automatic evaluation of output quality for machine translation systems*, In Proceedings of the evaluators' forum, April 21-24, 1991, Les Rasses, Vaud, 1993.