CWMT 2017 Machine Translation Evaluation Guidelines

CWMT 2017 Machine Translation Evaluation Committee

I. Introduction

The 13th China Workshop on Machine Translation (CWMT 2017) will be held at Dalian, China on September 27-29, 2017. CWMT 2017 will continue the ongoing series of machine translation (MT) evaluation campaigns. Compared with the previous evaluation (CWMT 2015), there are several changes in the evaluation plan:

1. The Chinese-English and English-Chinese news translation tasks are co-organized by CWMT and WMT2017; participants for WMT 2017 are welcome to submit their results to CWMT, and to participate the CWMT event, as well.

2. A new Japanese-Chinese patent domain translation task is added, which is co-organized by CWMT and Beijing Lingosail Technology Co. Ltd.; we also welcome other contributors from industry to participate in the organization of the evaluation in the future.

3.The training period starts immediately after the release of this guideline. Participants could get the corresponding data and tools and start the training process right after their registration. We would like to encourage potential participants to finish the registration as soon as possible.

4.The "Double Blind Evaluation" task is not held this year; the organizer will not provide the Baseline System, corresponding key steps or intermediate result files for the evaluation tasks.

We sincerely hope that this campaign will enhance the cooperation and connections of machine translation research and technology among domestic and overseas research sites, and promote the cooperation between academia and industry.

Information on CWMT 2017 is provided below:

The sponsor of CWMT 2017 machine translation evaluation is:

Chinese Information Processing Society of China

The organizers of this evaluation are:

Nanjing University

Institute of Computing Technology, Chinese Academy of Sciences

The resource providers of this evaluation are:

Beijing Lingosail Technology Co. Ltd.

Datum Data Co., Ltd.

Harbin Institute of Technology

Inner Mongolia University

Institute of Automation, Chinese Academy of Sciences

Institute of Computing Technology, Chinese Academy of Sciences

Institute of Intelligent Machines, Chinese Academy of Sciences

Nanjing University

Northeastern University

Northwest University of Nationalities

Peking University

Qinghai Normal University

Tibet University

Xiamen University

Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences

Xinjiang University

The chair of this evaluation is:

HUANG Shujian (Nanjing University)

The committee members of this evaluation are:

Aishan Wumaier (Xinjiang University)

WEI Yongpeng (Beijing Lingosail Technology Co. Ltd)

XIAO Tong (Northeastern University)

YANG YaTing (Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences)

Yiliyaer Jiaermuhamaiti (Nanjing University)

ZHANG Jiajun (Institute of Automation, Chinese Academy of Sciences)

ZHAO Hongmei (Institute of Computing Technology, Chinese Academy of Sciences)

For more information about CWMT 2017 and the MT evaluation tasks, please visit:

http://nlp.nju.edu.cn/cwmt2017/evaluation.html

II. Evaluation Tasks

CWMT 2017 MT evaluation campaign consists of six tasks involving 4 domains and 5 language pairs overall , as listed in Table 1.

No	Task ID	Task Name	Domain
1	CE	Chinese-to-English News Translation	News
2	EC	English-to-Chinese News Translation	News
3	MC	Mongolian-to-Chinese Daily Expression Translation	Daily Expressions
4	TC	Tibetan-to-Chinese Government Document Translation	Government Documents
5	UC	Uyghur-to-Chinese News Translation	News
6	JC	Japanese -to Chinese Patent Domain Translation	Patent

Table 1 CWMT2017 MT evaluation tasks

For every evaluation task, participants can freely choose the MT technology they wish to use, such as techniques based on rules or cases, or statistical machine translation or neural network machine translation, etc.

Participants can also use system combination technology, but should provide explicit indication in the system description and describe the performance of each single system in the technical report. Here system combination technology means using translations from two or more single systems for the reconstruction or selection of translation results at character, word, phrase or sentence level. Techniques which do not generate results from two or more single systems explicitly will not be considered as system combination technologies this time. Examples of such techniques are collaborative decoding in statistic machine translation, output ensemble in neural machine translation and reranking of nbest results from a single system. Systems that use system combination technology will be indicated as such in the report of the evaluation result.

III. Evaluation Methods

1. Evaluation Metrics

Automatic evaluation

Automatic evaluation tools will be used to evaluate the system performance in automatic metrics including: BLEU-SBP, BLEU-NIST, TER, METEOR, NIST, GTM, mWER, mPER, and ICT.

The organizers will use the following setting in the automatic evaluation:

(1) All the scores of these metrics will be case-sensitive; case-insensitive scores for some metrics may also be listed as reference;

(2) BLEU-SBP will be the primary metric;

(3) The evaluation of EC, TC, UC, MC and JC will be based on characters instead of words.

(4) In the evaluation of EC, TC, UC, MC and JC, the organizer will convert the full-width Chinese characters in the A3 area of GB2312 in the Chinese translations to half-width characters;

(5) The evaluation of CE will be based on English words.

2. Evaluation Procedure

CWMT 2017 MT evaluation will take the following four stages:

(1) Registration: The potential participant sends the registration form and evaluation agreement to the organizer. The organizer sends the training and development data to the confirmed participants (by FTP).

(2) Training stage: The participants train and develop their systems based on the data released or additional data (data conditions are listed in Section IV).

(3) Test stage: The organizer releases the source file of the test set. The participants run their systems and submit the translation results and system descriptions in the required format (Appendix B) by the deadline.

(4) Evaluation and reporting stage: The organizer evaluate the submitted translations and report back the final evaluation results. The participants prepare their system technological reports and attend CWMT 2017 workshop.

The exact time schedule is listed in Section VII.

IV. Evaluation Data and Training Conditions

The organizer will offer several language resources including the training corpora, development sets and test sets (source files).

1. Training Sets

Please refer to Appendix D for the resource list released by CWMT2017 and Appendix B for the document structure description.

CWMT 2017 MT evaluation campaign will update and add the following training data:

New datasets for the CE, EC tasks:

- NEU English-Chinese Parallel Corpus (2017)
- Datum English-Chinese Parallel Corpus (2017)

The EC and CE tasks are co-organized by WMT17, so the data provided by WMT17 can also be used for the evaluation of the corresponding EC and CE translation. In addition to the training, development and test set data provided by CWMT2017, WMT17 also allows the following data to be used[1]:

Parallel data in English and Chinese (News Commentary V12 and UN Parallel Corpus V1.0)

Chinese and English monolingual training data (Europarl, News Commentary, Common Crawl, News Crawl, News Discussions etc.); LDC Gigaword in English and Chinese (LDC2011T07, LDC2009T13, LDC2007T07, LDC2009T27)

New datasets for the MC task:

- IMU Mongolian-Chinese Parallel Corpus (2017)
- ICTCAS Mongolian-Chinese Parallel Corpus (2017)

New datasets for the TC task:

- ICTCAS Tibetan-Chinese Parallel Corpus (2017)

New datasets for the UC task:

- XJU Chinese- Uyghur Parallel Corpus (2017)
- ICTCAS Uyghur-Chinese Parallel Corpus (2017)
- XJIPC-CAS Uyghur-Chinese Parallel Corpus (2017)

New datasets for the JC task:

- Lingosail Chinese-Japanese Parallel Corpus (2017)

Participants can obtain the training data of the tasks to which they register.

2. Training Conditions

For statistical machine translation systems, two kinds of training conditions are allowed in CWMT 2017: Constrained training and Unconstrained training.

(1) Constrained training

Under this condition, only data provided by the evaluation organizer can be used for system development. System development must follow the following restrictions:

- The primary systems must be in the Constrained training condition in order to be evaluated under similar conditions.

[1] Please refer to the following webpage for details: http://www.statmt.org/wmt17/translation-task.html

- Rule-based MT modules or systems can use hand-crafted translation knowledge sources such as rules, templates, and dictionaries. Participants using rule-based MT system are required to describe the size of the knowledge sources and the ways to construct and use the knowledge sources in the system description and the technical reports.
- Tools for monolingual processing (such as lexical analyzers, parsers and named entity recognizers) are not subject to the training data restrictions.
- Tools for bilingual translation (such as named entity translator, syllable-to-character converter) must not use any additional resources. The exceptions here are tools for translating numerals and time words.
- For any evaluation task, systems can only use the corpora related to this task. Usage of the corpora of any other evaluation task is not acceptable, even the participant takes participation in more than one task.
- Constrained training corpora for Chinese-English and English-Chinese evaluation task (co-organized by CWMT and WMT) are composed of data provided by CWMT (listed in appendix D) and data provided by WMT. The participant should state clearly, in the system description and technical report, which parts of the data are being used: WMT data, CWMT data or both. The submissions with different training data conditions will be indicated in the report of the evaluation result.

(2) Unconstrained training

Under this condition, the participant is allowed to used data from other resources to assist the training of their systems. System development must follow the following restrictions:

- The contrast systems of participants can be developed under the unconstrained training condition.
- If participants use additional data, they should declare whether the data can be accessed publicly in the system description and technological report. If the data can be accessed publicly, the participants should provide the origin of the data.
- Participants are welcome to use their own online translation systems, but these systems should also be described in the system description and technological report briefly. The translation results of online systems will only be used as reference, which will be excluded from the ranking of unconstrained training systems.

3. The Development Sets

Information on the development sets is provided in Table 2.

Task ID	Size	Provider	Note

CE	2,002 sentences	Nanjing University	single reference
EC	2,002 sentences	Nanjing University	single reference
MC	1,000 sentences	Inner Mongolia University	4 references
TC	650 sentences	Qinghai Normal University	4 references
UC	700 sentences	Xinjiang University	4 references
JC	3,000 sentences	Beijing Lingosail Technology Co. Ltd.	single reference

Table 2 The development sets for CWMT2017

The CE and EC tasks share the same development data, which is the combination of 1002 sentences translated from English to Chinese, and 1000 sentence translated from Chinese to English.

4. The Test Sets

Information on the test sets is provided in Table 3.

Task ID	Size	Provider	Note
CE	1,000 sentences	Nanjing University	single reference
EC	1,000 sentences	Nanjing University	single reference
MC	1,001 sentences	Inner Mongolia University	4 references
TC	729 sentences	Qinghai Normal University	4 references
UC	1,000 sentences	Xinjiang Technical Institute of Physics & Chemistry, CAS	4 references
JC	1,000 sentences	Beijing Lingosail Technology Co. Ltd.	single reference

Table 3 The test sets for CWMT2017

The MC, TC and UC tasks use the same evaluation data from the last evaluation (CWMT2015).

Please refer to Appendix B for instructions regarding the format of the development and test sets.

V. Results Submission

The translation result(s) must be returned to the organizer before the deadline. Each participant should submit one final translation result as primary result and at most three other translation result(s) as contrast results for each task that they registered to. Each system submission should be accompanied by its system description. Please refer to Appendix B for the format of MT evaluation data and submission files.

Participants in CE and EC news translation tasks can submit results to CWMT2017 or WMT2017, or both. However, the submission should follow the requirements of that particular event.

VI. Technical Report Submission

After the evaluation, participants should submit a detailed technical report to CWMT 2017, which describes the architecture, major technology and the use of data. Each team ought to send at least one person to attend CWMT2017 and exchange related technology. Please see the reporting requirements in Appendix C.

VII. Evaluation Calendar

1	March 15, 2017	Registration starts. The organizer send training and development data, and scripts for scoring and format checking to the participants, according to their registration.
2	March 31, 2017	Deadline for registration. The data and scripts will not be provided to other organizations. (Please contact the organizer for later registrations.)
3	May 15, 2017 10:00 AM GMT+8	Release of test data for CE and EC tasks to participants
4	May 22, 2017 17:30 PM GMT+8	Deadline for submitting results for the CE and EC tasks
5	May 20, 2017 10:00 AM GMT+8	Release of test data for JC, UC, MC and TC tasks to participants
6	May 27, 2017 17:30 PM GMT+8	Deadline for submitting results for the JC, UC, MC and TC tasks
7	June 15, 2017	Preliminary release of evaluation results to participants
8	June 30, 2017	Deadline for submitting technical report
9	July 10, 2017	Reviews of technical reports sent to participants, who should modify reports accordingly
10	July 30, 2017	Deadline for submitting technical report camera-ready
11	September 27 to Sep-	CWMT 2017 workshop. Official public release of results.

	tember 29, 2017	

VIII. Appendix

This document includes the following appendices:

Appendix A: Registration Form and Evaluation Agreement

Appendix B: Format of MT Evaluation Data

Appendix C: Requirements of Technical Report

Appendix D: List of Resources Released by the Organizer

Appendix A: Registration Form and Evaluation Agreement

Any organization engaged in MT research or development can register for the CWMT 2017 evaluation. The participating sites of CWMT 2017 evaluation should fill the following form and agreement and send it to the organizer by either email or post. The registration form and the agreement should be signed by the person in charge of the participating team/organization, or stamped with the official seal of the team/organization.

The evaluation does not charge any registration fees. Each participant should send at least one person to attend the workshop (CWMT 2017).

Please send the registration form and the agreement to:

Name: HUANG Shujian

Email: huangsj@nju.edu.cn

Address: Department of Computer Science and Technology, Nanjing University, 163 Xianlin Avenue, Nanjing 210023, China

Post Code: 210023

Telephone: 025-89680690

Registration form for CWMT2017 Machine Translation Evaluation

Organization Name			
Address			
Contact		Telephone	
Post code		Email	
Evaluation Tasks	☐ Chinese-to-English News Translation ☐ English-to-Chinese News Translation ☐ Japanese-to-Chinese Patent Translation ☐ Mongolian-to-Chinese Daily Expression Translation ☐ Tibetan-to-Chinese Government Document Translation ☐ Uyghur-to-Chinese News Translation		

The participating site agrees to comply with the following terms:

1. After receiving the evaluation data, the participating site should submit the results including system description and primary system's results to the evaluation organizer before the submission deadline.

2. The participating site agrees to submit a formal technical report, attend the CWMT2017 workshop, present their systems and communicate with other participants.

3. The participating site confirms that it has the intellectual property of the participating system. If any technology in the participating system is licensed from other person or organization, it should be clearly described in the system description.

4. The registrant guarantees that the data obtained in the evaluation, including the training set, the development set, the test set, reference translations, results from other participants' primary system and evaluation tools, will only be used in research related to this evaluation. No other usage is permitted.

5. The participating site guarantee that the evaluation data will only be used within the research group that takes part in the evaluation, and neither will it be distributed to other sites in any way (written, electronically, or by network), nor will it be used by any partner or affiliated organizations of the participating site.

6. The participating site guarantee to credit the resource providers by referring to the resources being used (e.g., training data, development data, test data, reference translations, and evaluation tools) in their publications and other research accomplishments.

7. If a participating site violates terms 4-6, the evaluation sponsor and resource providers have the right to request the participating site and the cooperators and/or affiliation organizations using the resources without granted licenses to compensate 3-5 times the cost of the distributed resources. If insufficient, the compensation fee should be increased to be equal to the actual loss of related resource providers.

Signature of the person in charge or the official seal of the participating site:

Date:

CWMT 2017 MACHINE TRANSLATION EVALUATION PARTICIPATING SITE AGREEMENT

(Non-profit Agreement)

This agreement is made by and between:

Name of The Participating Site (hereinafter called "the participating site"), participating site of the CWMT 2017 Machine Translation Evaluation, having its principal place of business at:

Address of the Participating Site

AND

Chinese Information Processing Society of China (hereinafter called "the sponsor"), the sponsor of CWMT 2017 machine translation evaluation, having its principal place at:

No.4, Forth Southern Street, Zhongguancun, Beijing, China.

Whereby it is agreed as follows:

1. The sponsor provides the participating site with **evaluation data** including the training set, the development set, the test set, reference translations, and evaluation tools.

2.The participating site confirms that the evaluation data obtained from the sponsor will only be used in research related to this evaluation. No other usage is permitted.

3. The participating site agrees that the evaluation data will only be used within the research group that takes part in the evaluation, and neither will it be distributed by any way (written, electronically, or by network), nor will it be used by any partner or affiliated organization of the participating site.

4. The participating site agrees to credit the resource providers by referring to the resources being used in their publications and other research accomplishments.

In witness whereof, intending to be bound, the parties hereto have executed this AGREEMENT by their duly authorized officers.

AUTHORISED BINDING SIGNATURES:

_____	_____
On behalf of Chinese Information Processing Society of China	On behalf of Name of the Participating Site
Name:	Name:
Title:	Title:
Date:	Date:

Appendix B: Format of MT Evaluation Data

This appendix describes the format of the data released by the organizer and the result files that the participants should submit.

All the files should be encoded in UTF-8 format. Among them, the development set (including its reference), the test set and the final translation result files must be strict XML files (whose formats are defined by the XML DTD described in section III) encoded in UTF-8 (with BOM), and all the others are plain text files encoded in UTF-8 (without BOM).

I. Data released by the organizer

The organizer will release three kinds of data: training sets, development sets and test sets. Here we take the "Chinese-to-English Translation" task as an example for illustration purposes.

1. Training Set

The training data contains one sentence per line. The parallel corpus of each language pair is made of a source file and a target file, which contain the source and target sentences respectively.

Figure 1 illustrates the data format of the parallel corpus.

Chinese File:	English File:
战法训练有了新的突破	Tactical training made new breakthrough
第一章总则	Chapter I general rules
人民币依其面额支付	The renminbi is paid by denomination
……	……

Figure 1 Example of the parallel corpus

2. Development Set

There are source files and reference files in the development set and the test set.

(1) Source File

A source file contains one single *srcset* element, which has the following attributes:

- *setid*: the dataset
- *srclang*: the source language. One element of this set:{en, zh, mn, uy, ti, jp}
- *trglang*: the target language. One element of this set:{en, zh, mn, uy, ti, jp}

A *srcset* element contains one or more DOC element(s), and each DOC element contains one single attribute *docid*, which indicates the genre of the DOC.

Each DOC element contains several *seg* elements with attribute *id*.

One or more segments may be encapsulated inside other elements, such as *p*. Only the text surrounded by *seg* elements is to be translated.

Figure 2 shows an example of the source file.

```
<?xml version="1.0" encoding="UTF-8"?>
<srcset setid="zh_en_news_trans" srclang="zh" trglang="en">
<DOC docid="news">
<p>
<seg id="1">sentence 1 </seg>.
<seg id="2">sentence 2</seg>
……
</p>
……
</DOC>
</srcset>
```

Figure 2 Example of the source file

(2) **Reference file**

A reference file contains a *refset* element. Each *refset* element contains the following attributes:

- *setid*: the dataset
- *srclang*: the source language. One element of this set:{en, zh, mn, uy, ti, jp}
- *trglang*: the target language. One element of this set:{en, zh, mn, uy, ti, jp}

Each *refset* element contains several DOC elements. Each DOC has two attributes:

- *docid*: the genre of the DOC
- *site*: the indicator for different references. One element of this set:{1, 2, 3, 4}

Figure 3 shows an example of the reference file.

```
<?xml version="1.0" encoding="UTF-8"?>
<refset setid="zh_en_news_trans" srclang="zh" trglang="en">
<DOC docid="news" sysid="ref" site="1">
<p>
<seg id="1">reference 11 </seg>
```

```
<seg id="2">reference 12</seg>
……
</p>
……
</DOC>
<DOC docid="news" sysid="ref" site="2">
<p>
<seg id="1">reference 21 </seg>
<seg id="2">reference 22</seg>
……
</p>
……
</DOC>
<DOC docid="news" sysid="ref" site="3">
<p>
<seg id="1">reference 31</seg>
<seg id="2">reference 32</seg>
……
</p>
……
</DOC>
<DOC docid="news" sysid="ref" site="4">
<p>
<seg id="1">reference 41 </seg>
<seg id="2">reference 42</seg>
……
</p>
……
</DOC>
</refset>
```

Figure 3 Example of the reference file

3. Test set

For convenience of evaluation, organizer will only release the test set source files in the same format as the development set.

Ⅱ. Data Format in the Submission

Participants need to submit translations and system descriptions in the format below.

1. File Naming

Please name the submitted files according to the naming mode in the following table (we use "ce", "ict" and "2017" here as examples of Task ID, Participant ID and year of the test data respectively).

File	Naming mode	Example
final translation result	Task ID – year of the test data – Participant ID – Primary vs. contrast system – System ID.xml	ce-2017-ict-primary-a.xml ce-2017-ict-contrast-c.xml

2. Final translation result

The final translation result is the translation result from the participant's translation systems, with proper post-processings such as recase, detokenize, etc.

The submission of each system should be an xml file which has a format similar to the source file of the test set.

The final submission file contains a **tstset** element with the following attributes:

- *setid*: the dataset
- *srclang*: the source language. One element of this set:{en, zh, mn, uy, ti, jp}
- *trglang*: the target language. One element of this set: {en, zh, mn, uy, ti, jp}

The tstset element contain a **system** element with the following attributes:

- *site*: the label of the participant
- *sysid*: the identification of the MT system

The value of the **system** element is the description of the participating system including the following information:

- *Hardware and software environment*: the operating system and its version, number of CPUs, CPU type and frequency, system memory size, etc.
- *Execution Time*: the time from accepting the input to generating the output.
- *Technology outline*: an outline of the main technology and important parameters of the participating system. If the system uses system combination techniques, single systems being combined and the combination techniques should be described here.
- *Training Data*: a description of the training data and development data used for system training, with indication of the training condition (i.e. constrained or unconstrained). For CE and EC tasks, please also indicate the source of the training data (i.e. WMT, CWMT or both).

- *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 code, free software, shareware, and commercial software.

The content of each DOC element is exactly the same as that of the test set's source file, which is described before.

Here is an example of the final submission file in Figure 4.

```
<?xml version="1.0" encoding="UTF-8"?>
<tstset setid="zh_en_news_trans" srclang="zh" trglang="en">
<system site="unit name " sysid="system identification">
description information of the participating system
.....
</system>
<DOC docid="document name " sysid="system identification">
<p>
<seg id="1">submit translation 1</seg>
<seg id="2">submit translation 2</seg>
……
</p>
……
</DOC>
</tstset >
```

Figure 4 Illustration of the final submission file

3. Document details

please pay attention to the following points when generating the submission file:.

- Please note that CWMT 2017 evaluation adopts strict XML file format. The main difference between the XML file format and the NIST evaluation file format lies in the following: in an XML file, if the following five characters occur in the text outside tags, they should be replaced by escape sequences:

Character	Escape sequence
&	&amp;
<	<
>	>

"	&quot;
'	&apos;

- As for Chinese encoding, the middle point in a foreign person name should be written as "E2 80 A2" in UTF-8, for example, "托德·西蒙斯" ;
- As for English tokenization, the tokenization should be consistent with the "normalizeText" function of the Perl script "mteval-v11b.pl" released by NIST. The main part of the script is listed in Figure 5.

```
# language-dependent part (assuming Western languages):
$norm_text = " $norm_text ";
# Add a space before the beginning and after the end of the text respectively (and then delete)
$norm_text =~ tr/[A-Z]/[a-z]/ unless $preserve_case;
# Uppercase letters are generally converted to lowercase, unless the user specifies the reserved case
$norm_text =~ s/([\{-\~\[-\` -\&\(-\+\:-\@\/])/ $1 /g;   # tokenize punctuation
# Add a space to both sides of the following symbols (corresponding the ASCII character of hexadec-
imal for-mats marked behind)
#{ | } ~   (0x7b-0x7e)
#[ \ ] ^ - ` (0x5b-0x60)
#( Space )! " # $ % & (0x20-0x26)
#( ) * + (0x28-0x2b)
# : ; < = > ? @ (0x3a-0x40)
# / (0x2f)
$norm_text =~ s/([^0-9])([\.,])/$1 $2 /g; # tokenize period and comma unless preceded by a digit
#When non-numeric characters are follow by a period '.' or a comma ',', a space character should be add-
ed to both sides of the period of comma. (No space character will be added if a number is followed by
period or comma.)
$norm_text =~ s/([\.,])([^0-9])/ $1 $2/g; # tokenize period and comma unless followed by a digit
#When periods '.' or commas ',' aren't followed by numeric character 0-9, a space character should be
added to both sides of the period or comma
$norm_text =~ s/([0-9])(-)/$1 $2 /g; # tokenize dash when preceded by a digit
#When numeric characters 0-9 are followed by a hyphen, a space character should be added to both sides
of the hyphen
$norm_text =~ s/\s+/ /g; # one space only between words
#Replace continuous space characters with one single space character
$norm_text =~ s/^\s+//;  # no leading space
#Remove space characters at the beginning of the text
$norm_text =~ s/\s+$//;  # no trailing space
#Remove space characters at the end of the text
```

Figure 5 the main code of the tokenizeation script

III. Description of CWMT 2017 XML files' document structure

```
<?xml version=”1.0” encoding = “UTF-8”?>
<!ELEMENT srcset (DOC+)>
<!ATTLIST srcset setid CDATA #REQUIRED>
<!ATTLIST srcset srclang (en | zh | mn | uy | ti | jp ) #REQUIRED>
<!ATTLIST srcset trglang (en | zh | mn | uy | ti | jp) #REQUIRED>
<!ELEMENT refset (DOC+)>
<!ATTLIST refset setid CDATA #REQUIRED>
<!ATTLIST refset srclang (en | zh | mn | uy | ti| jp ) #REQUIRED>
<!ATTLIST refset trglang (en | zh | mn | uy | ti | jp) #REQUIRED>
<!ELEMENT tstset (DOC+)>
<!ELEMENT tstset (system+)>
<!ATTLIST tstset setid CDATA #REQUIRED>
<!ATTLIST tstset srclang (en | zh | mn | uy | ti | jp) #REQUIRED>
<!ATTLIST tstset trglang (en | zh | mn | uy | ti | jp) #REQUIRED>
<!ELEMENT system (#PCDATA) >
<!ATTLIST system site CDATA #REQUIRED >
<!ATTLIST system sysid CDATA #REQUIRED >
<!ELEMENT DOC ( p* )>
<!ATTLIST DOC docid CDATA #REQUIRED>
<!ATTLIST DOC site CDATA #IMPLIED>
<!ELEMENT p(seg*)>
<!ELEMENT seg (#PCDATA)>
<!ATTLIST seg id CDATA #REQUIRED>
```

Appendix C: Requirement of Technical Report

All participating sites should submit a technical report to the 13th China Workshop of Machine Translation (CWMT 2017). The technological report should describe the technologies used in the participating system(s) in detail, in order to inform the reader about how the reported results are obtained. A good technological report should be detailed enough so that the reader could replicate the work which is described in the report. The report should be no shorter than 5,000 Chinese characters or 3,000 English words.

Generally, a technology report should provide the following information:

Introduction: Give the background information; introduce the evaluation tasks participated, 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, that should be described in detail. If the participating site uses system combination techniques, the single systems(including results from those systems) as well as the combination technique should be described. If the participating site uses handcrafted translation knowledge sources such as rules, templates, and dictionaries, the size of the knowledge sources and the ways to construct and use the knowledge sources should be described.

Data: Give detailed description of 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 obtained on the evaluation set. Analyze the results.

Conclusion: (open)

Appendix D: Resource List Released by the Organizer

Without special indication, the data file provided by the organizer is encoded in UTF-8 (without BOM).

1. The Chinese-English news resources provided by the organizer

1.1 Training data

ChineseLDC resource ID	Resource description	
Datum2015	Name	Datum English-Chinese Parallel Corpus (2015) (Part)
	Providers	Datum Data Co., Ltd.
	Languages	Chinese-English
	Domain	Multi-domain, including: textbooks for language education, bilingual books, technological documents, bilingual news, government white books, government documents, bilingual resources on web, etc.
	Size	1,000,004sentence pairs
	Description	It is a part of the "Bilingual / Multi-lingual Parallel Corpus" developed by Datum Data Co., Ltd under the support of the National High Technology Research and Development Program of China (863 Program).
CASICT2011 (CLDC-2010-001) (CLDC-2012-001)	Name	ICT Web Chinese-English Parallel Corpus (2013)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	Multi-domain
	Size	1,936,633 sentence pairs
	Description	The parallel corpus is automatically acquired from web. All the processes, including parallel web page discovering and verification, parallel text extraction, sentence alignment, etc., are entirely automatic. The accuracy of the corpus is about 95%. This work was supported by the National Natural Science Foundation of China (Grant No. 60603095).
CASICT2015	Name	ICT Web Chinese-English Parallel Corpus (2015)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	Multi-domain

	Size	2,036,834 sentence pairs
	Description	The parallel corpus is automatically acquired from web. All the processes, including parallel web page discoverin g and verificati-on, parallel text extraction, sentence align ment, etc., are entirely automatic. The Institute of Computing Technology has corrected this corpus roughly. The accuracy of the corpus is greater than 99%. Three sources of sentences were selected to provide this cor-pus: 60% from the web, 20% from movie subtitles, and the rest 20% from the English-to-Chinese or Chinese-to-English dictionari-es.
CASIA2015	Name	CASIA Web Chinese-English Parallel Corpus (2015)
	Providers	Institute of Automation, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	Multi-domain
	Size	1,050,000 sentence pairs
	Description	The parallel corpus is automatically acquired from web. All the processes, including parallel web page discoverin g and verificati-on, parallel text extraction, sentence align ment, etc., are entirely automatic.
Datum2017	Name	Datum English-Chinese Parallel Corpus (2017)
	Providers	Datum Data Co., Ltd.
	Languages	Chinese-English
	Domain	
	Size	1,000,000 sentence pairs,divided into 20 parts
	Description	
NEU2017	Name	NEU Chinese-English Parallel Corpus (2017)
	Providers	Natural Language Processing Group, Northeastern University
	Languages	Chinese-to-English, English-to-Chinese
	Domain	
	Size	2,000,000 sentence pairs
	Description	
SSMT2007 MT	Name	SSMT2007 Machine Translation Evaluation Data (a part of Chi-

Evaluation Data (2007-863-001)		nese-English & English-Chinese MT evaluation data)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	News
	Size	This is the test data of SSMT 2007 MT Evaluation, which contain data of 2 translation directions (Chinese-English and English-Chinese) in news domain. The source file of Chinese-English data contains 1,002 Chinese sentences. The source file of English-Chinese data contains 955 English sentences. There are 4 reference translations made by human experts for each test sentence.
	Description	
HTRDP(863)2005 MT Evaluation Data (2005-863-001)	Name	HTRDP("863 Program") 2005 Machine Translation Evaluation Data (a part of Chinese-English & English-Chinese MT evaluation data)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	The data contains two genres: one is dialog data from Olympics-related domains, which includes game reports, weather forecasts, traffic and hotels, travel, foods, etc, and the other one is text data from news domain.
	Size	The source files of dialog and text data in Chinese-to-English and English-to-Chinese directions contain about 460 sentence pairs respectively.
	Description	
HTRDP(863)2004 MT Evaluation Data (2004-863-001)	Name	HTRDP ("863 Program") 2004 Machine Translation Evaluation Data (a part of Chinese-English & English-Chinese MT evaluation data)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	Two data genres: one is text data, the other is dialog data. The data covers general domain and Olympic-related domains which include game reports, weather forecasts, traffic and hotels, travel, foods, etc.
	Size	The source files of Chinese-to-English direction contain dialog data of 400 sentences and text data of 308 sentences. The source files of English-to-Chinese direction contain dialog data of 400 sentences and text data of 310 sentences. There are 4 reference translations

		made by human experts for each source sentence.
	Description	The test data for the 2004 “863 Program” machine translation evaluation.
HTRDP(863)2003 MT Evaluation Data (2003-863-004)	Name	HTRDP (“863 Program”) 2003 Machine Translation Evaluation Data (A part of Chinese-English & English-Chinese MT evaluation data)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	The data covers Olympic-related domains which include game reports, weather forecasts, traffic and hotels, travel, foods, etc.
	Size	The source files of Chinese-to-English direction contain dialog data of 437 sentences and text data of 169 sentences, and the source files of English-to-Chinese direction contain dialog data of 496 sentences and text data of 322 sentences. There are 4 reference translations made by human experts for each source sentence.
	Description	The test data for the 2003 “863 Program” machine translation evaluation.
CWMT2008 Machine Translation Evaluation Data (CLDC-2009-001) (CLDC-2009-002)	Name	CWMT2008 Machine Translation Evaluation Data
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	News
	Size	The source files of dialog and text data in Chinese-to-English and English-to-Chinese directions contain about 1000 sentence pairs respectively. There are 4 reference translations made by human experts for each source sentence.
	Description	
CWMT2009 Machine Translation Evaluation Data	Name	CWMT2009 Machine Translation Evaluation Data
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Chinese-English
	Domain	News
	Size	The source files of dialog and text data in Chinese-to-English and English-to-Chinese directions contain about 1000 sentence pairs respectively. There are 4 reference translations made by human experts for each source sentence.

	Description	
CWMT2011 Machine Translation Evaluation Data	Name	CWMT2011 Machine Translation Evaluation Data
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	English→Chinese
	Domain	News
	Size	The source files of dialog and text data in English-to-Chinese directions contain 3187 sentence pairs. There are 4 reference translations made by human experts for each source sentence.
	Description	

1.2.The Chinese monolingual resources

XMU-CWMT2017	Name	XMU Natural Language Processing Group XINHUA News Corpus (2017)
	Providers	Xiamen University
	Languages	Chinese
	Domain	News
	Size	662,904 articles ,almost 11,000,000 words.
	Description	The corpus is collected by Xiamen University, including all topic channels news of XINHUA in 2011, such as domestic news, international news, financial news, forums, education etc..

1.3. development resources

newsdev2017-enzh (newsdev2017-zhen)	Name	NJU Chinese-English news corpus (2017)
	Providers	Nanjing University
	Languages	Chinese-to-English, English-to-Chinese
	Domain	News
	Size	2,002 sentence pairs.
	Description	Chinese English / English news corpus

2. The Mongolian-Chinese daily expression resources provided by the organizer

2.1 Training data

IMU-	Name	IMU Mongolian-Chinese Parallel Corpus (2013)

CWMT2013 (CLDC-2010-005)	Providers	Inner Mongolia University
	Languages	Chinese-Mongolian
	Domain	Government documents, laws, rules, daily conversation, literature
	Size	104,975 sentence pairs. Including: 1) 67,274 sentence pairs for CWMT 2011 MT evaluation, covering domains such as daily conversation, literature, government documents, laws and rules; 2) 37,701 newly added sentence pairs for CWMT 2013 MT evaluation, including 17,516 sentence pairs from news domain, 10,394 sentence pairs from government documents, 5,052 sentence pairs from text books and 4,739 sentence pairs from a Mongolian-to-Chinese dictionary.
	Description	
IMU-CWMT2015	Name	IMU Mongolian-Chinese Parallel Corpus (2015)
	Providers	Inner Mongolia University
	Languages	Chinese-Mongolian
	Domain	Government documents, laws, rules, daily conversation, literature
	Size	24,978 sentence pairs
	Description	
IIM-CWMT2015	Name	IIM Mongolian-Chinese Parallel Corpus (2015)
	Providers	Institute of Intelligent Machines, Chinese Academy of Sciences
	Languages	Mongolian→Chinese
	Domain	News
	Size	1,682 sentence pairs.
	Description	
ICT-MC-corpus-CWMT2017	Name	ICT Mongolian-Chinese Parallel Corpus (2017)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Mongolian→Chinese
	Domain	News
	Size	30,007 sentence pairs

	Description	
IMU-corpus-CWMT2017	Name	IMU Mongolian-Chinese Parallel Corpus (2017)
	Providers	Inner Mongolia University
	Languages	Mongolian→Chinese
	Domain	Multi-domain,include: government documents, government reports, the State Council documents, laws and regulations, etc.
	Size	100,001 sentence pairs
	Description	

2.2. development resources

IMU-dev-mnzh-CWMT2017	Name	IMU development Mongolian-Chinese Parallel Corpus
	Providers	Inner Mongolia University
	Languages	Mongolian→Chinese
	Domain	government documents, government reports, daily conversation and literature
	Size	1,000 sentence pairs, 4 reference**s**
	Description	CWMT2017 Mongolian→Chinese development is the same for CWMT2015,CWMT2013,CWMT2011

3. The Tibetan-Chinese government document resources provided by the organizer

3.1 Training data

QHNU-CWMT2013	Name	QHNU Tibetan-Chinese Parallel Corpus (2013)
	Providers	Qinghai Normal University
	Languages	Tibetan-Chinese
	Domain	Government document
	Size	33,145 sentence pairs
	Description	The sentence alignment accuracy of the corpus is over 99%. The construction of the corpus was supported by NSFC (Grant No. 61063033) and 973 Program (Grant No. 2010CB334708).
QHNU-CWMT2015	Name	QHNU Tibetan-Chinese Parallel Corpus (2015)
	Providers	Qinghai Normal University
	Languages	Tibetan-Chinese
	Domain	Government document

	Size	17,194 sentence pairs
	Description	The sentence alignment accuracy of the corpus is over 99%. The construction of the corpus was supported by NSFC (Grant No. 61063033).
XBMU-XMU	Name	Yang Jin Tibetan-Chinese Parallel Corpus
	Providers	Artificial Intelligence Institute, Xiamen University & Language Technology Institute, Northwest University of Nationalities
	Languages	Chinese→Tibetan
	Domain	Multi-domain
	Size	52,078 sentence pairs
	Description	1) The sources of the corpus include publications, a Tibetan-Chinese Dictionary, and Tibetan-Chinese Web Text. The corpus was automatically aligned and corrected manually. 2) The alignment accuracy is100% 3) The research was supported by NSSFC (Grant No. 05AYY001) and HTRDP (Grant No. 2006AA010107)
XBMU-XMU-UTibet	Name	NUN-TU-XMU Tibetan-Chinese Parallel Corpus (2012)
	Providers	Language Technology Institute, Northwest University of Nationalities & Tibet University & Artificial Intelligence Institute, Xiamen University
	Languages	Chinese→Tibetan
	Domain	Political writings, law
	Size	24,159 sentence pairs
	Description	It selects material from Chinese law and regulation files during 2008 to 2009 and government reports from 2011 to 2012. All the source materials have been scanned, recognized, checked and processed manually.
ICT-TC-corpus-CWMT2017	Name	ICT Tibetan-Chinese Parallel Corpus (2017)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Tibetan→Chinese
	Domain	News
	Size	30,004 sentence pairs
	Description	

3.2. development resources

QHNU-dev-tizh-CWMT2017	Name	Qinghai Normal University Tibetan-Chinese Parallel Corpus
	Providers	Qinghai Normal University
	Languages	Tibetan→Chinese
	Domain	Government document
	Size	650 sentence pairs, 4 references
	Description	CWMT2017 Tibetan→Chinese development is the same for CWMT2015, CWMT2013, CWMT2011

4. The Uyghur-Chinese news resources provided by the organizer

4.1 Training data

XJU-CWMT2013 (CLDC-2013-002)	Name	XJU Uyghur-Chinese Parallel Corpus (2013)
	Providers	Xinjiang University
	Languages	Chinese→Uyghur
	Domain	News
	Size	79,935 sentence pairs
	Description	
XJIPC-CWMT2015	Name	XTIPC Uyghur-Chinese Parallel Corpus (2015)
	Providers	Xinjiang Technical Institute of Physics & Chemistry, CAS
	Languages	Chinese→Uyghur
	Domain	News
	Size	59,990 sentence pairs
	Description	Newly added about 30,000 sentence pairs based on that of Version 2013. 95% of sentence pairs are from news (2007~2014) and the others are from laws and government reports.
ICT-UC-corpus-CWMT2017	Name	ICT Uyghur-Chinese Parallel Corpus (2017)
	Providers	Institute of Computing Technology, Chinese Academy of Sciences
	Languages	Uyghur→Chinese
	Domain	News
	Size	30,071 sentence pairs

	Description	
XJU-corpus-CWMT2017	Name	XJU Uyghur-Chinese Parallel Corpus (2017)
	Providers	Xinjiang University
	Languages	Uyghur→Chinese
	Domain	News
	Size	152,527 sentence pairs
	Description	
XJIPC-corpus-CWMT2017	Name	XJIPC Uyghur-Chinese Parallel Corpus (2017)
	Providers	Xinjiang Institute of physics and chemistry, Chinese Academy of Sciences
	Languages	Uyghur→Chinese
	Domain	News
	Size	30,000 sentence pairs
	Description	

4.2. development resources

XJU-dev-uyzh-CWMT2017	Name	XJU Uyghur-Chinese Development Parallel Corpus
	Providers	Xinjiang University
	Languages	Uyghur→Chinese
	Domain	News
	Size	700 sentence pairs, 4 references
	Description	CWMT2017 Uyghur→Chinese development is the same for CWMT2015,CWMT2013,CWMT2011

5. The Japanese-Chinese Patent Domain resources provided by the organizer

5.1 Training data

Lingosail-train-CWMT2017	Name	Lingosail Chinese-Japanese Parallel Corpus (2017)
	Providers	Beijing Lingosail Technology Co. Ltd.
	Languages	Japanese→Chinese
	Domain	Multi-domain
	Size	3,000,000 sentence pairs
	Description	

5.2. development resources

Lingosail-dev-CWMT2017	Name	Lingosail Development Japanese-Chinese Parallel Corpus (2017)
	Providers	Beijing Lingosail Technology Co. Ltd.
	Languages	Japanese→Chinese
	Domain	Multi-domain
	Size	3,000 sentence pairs
	Description	

5.3.The Chinese Language model data

Lingosail-cn_for_lm-CWMT2017	Name	Lingosail Chinese Parallel language model (2017)
	Providers	Beijing Lingosail Technology Co. Ltd.
	Languages	Chinese
	Domain	Multi-domain
	Size	7,114,700 sentence
	Description	