第十三届全国机器翻译研讨会（CWMT 2017）评测大纲

CWMT 2017 评测委员会

一. 引言

第十三届全国机器翻译研讨会（CWMT 2017）将于 2017 年 9 月 27 日至 29 日在中国大连举行。根据惯例，本次研讨会将继续组织统一的机器翻译评测。

CWMT 2017机器翻译评测方案与上届评测（CWMT 2015）相比有如下变化：

1、汉英、英汉新闻领域的评测项目，由CWMT与WMT2017合作组织，欢迎WMT汉英、英汉项目的参评单位同时向CWMT提交系统结果、评测报告，并参加CWMT进行会议交流；

2、新增日汉专利领域的评测项目，由CWMT与北京语智云帆科技有限公司联合组织，希望能以该项目作为起点，加强科研单位与产业界的合作和联系；

3、本次评测不再设置统一发放数据的时间，各参评单位报名之后即可获取数据并进行系统训练，请有意向参与的单位尽快报名；

4、此外，本次评测暂停双盲评测（Double Blind Evaluation）项目，对其他评测项目评测组织方也不再提供评测项目的“基线系统（Baseline System）”及相应的关键步骤中间结果文件。

希望本次评测能够促进国内外科研单位、产业界相关单位之间的学术交流和联系，共同推动机器翻译研究和技术的发展。

本次评测的组织信息如下（三个以上的并列项以汉语拼音为序）：

评测主办机构：

中国中文信息学会

评测组织单位：

南京大学
中国科学院计算技术研究所

评测资源提供单位：

北京大学
北京语智云帆科技有限公司
点通数据有限公司
东北大学
哈尔滨工业大学
南京大学
内蒙古大学
青海师范大学
西北民族大学
西藏大学
厦门大学
新疆大学
中国科学院合肥智能机械研究所
中国科学院计算技术研究所

中国科学院新疆理化技术研究所
中国科学院自动化研究所

评测委员会主席：
黄书剑（南京大学）

评测委员会委员：
艾山·吾买尔（新疆大学）
魏永鹏（北京语智云帆科技有限公司）
肖　桐（东北大学）
杨雅婷（中国科学院新疆理化技术研究所）
伊力亚尔·加尔木哈买提（南京大学）
张家俊（中国科学院自动化研究所）
赵红梅（中国科学院计算技术研究所）

有关评测的更多信息请参见以下网址：
http://nlp.nju.edu.cn/cwmt2017/evaluation.ch.html

二. 评测项目

本次评测包含 6 个评测项目，我们将为各个评测项目的参评单位提供相应的训练语料和开发语料。具体项目设置如表 1 所示。

表1 CWMT 2017 评测项目表

序号	项目代号	评测项目名称	语种	领域
1	CE	汉英新闻领域机器翻译	汉语→英语	新闻领域
2	EC	英汉新闻领域机器翻译	英语→汉语	新闻领域
3	MC	蒙汉日常用语机器翻译	蒙古语→汉语	日常用语
4	TC	藏汉政府文献机器翻译	藏语→汉语	政府文献
5	UC	维汉新闻领域机器翻译	维吾尔语→汉语	新闻领域
6	JC	日汉专利领域机器翻译	日语→汉语	专利领域

对于每个评测项目，参评单位可以自由选择所采用的机器翻译技术（如：基于规则的机器翻译技术、基于实例的机器翻译技术、统计机器翻译技术及神经网络机器翻译技术等）。参评单位也可以使用系统融合技术，但在系统描述中必须做出明确说明，并在技术报告中说明进行系统融合的各个单系统的性能。此处，系统融合技术指使用两个及两个以上单系统的翻译结果进行字、词、短语、句子级别的重构或选择的技术。没有明确产生两个或两个以上单系统翻译结果的技术，如统计机器翻译中的协同解码、神经网络机器翻译的输出层ensemble、单个系统结果的重排序等，本次评测不认定为系统融合技术。评测组织方在发布评测结果时，将对使用系统融合技术的系统进行标注说明。

三. 评测方法

1. 评测指标

自动评测：自动评测是指利用自动评价工具对参评单位提交的最终翻译结果文件进行评价。本次评测

中的自动评测采用多种自动评价标准，包括：BLEU-SBP、BLEU-NIST、TER、METEOR、NIST、GTM、mWER、mPER 以及 ICT。

评测组织方进行自动评价时将采用如下设置：

(1) 所有自动评测将采用大小写敏感（case-sensitive）的方式，评测结果中也包含部分大小写不敏感的评价作为参考；
(2) BLEU-SBP作为主要的自动评价指标；
(3) 英汉、藏汉、维汉、蒙汉和日汉五个方向将采用基于字符（character-based）的评价方式；
(4) 英汉、藏汉、维汉、蒙汉和日汉五个方向中，评测组织方将对GB2312编码的A3区字符进行全角到半角的转换；
(5) 汉英项目则采用基于词（word-based）的评价方式。

2. 评测流程

本次评测的具体流程如下：

(1) 参评单位向主办方发送报名表和评测协议，主办方据此向参评单位发送训练、开发数据获取方法（ftp形式）；
(2) 在训练阶段，参评单位使用主办方发放的数据或其他数据进行系统训练、调优；
(3) 在测试阶段，评测组织方将发放测试数据，参评单位在规定时间内提交最终翻译结果文件和系统描述；
(4) 测试阶段结束后，评测组织方将对参评单位提交的最终翻译结果文件进行评测，并为参评单位提供各参评系统的评测结果；参评单位撰写技术报告并参加CWMT2017进行讨论。

具体评测日程安排请参见第七节。

四. 评测数据和训练条件

评测组织方将提供包括训练数据、开发数据和测试数据（源语言文件）在内的评测数据。

1. 训练数据

评测组织方提供的训练语料资源的清单请参见附件四，语料资源的格式说明见附件二。

其中，今年新增或更新的训练语料有:

汉英英汉新闻翻译项目：

- 东北大学英汉平行语料库（2017）（200 万句对）
- 点通公司英汉平行语料库（2017）（100 万句对）

汉英和英汉评测项目与 WMT17 联合组织，因此 WMT17 提供的数据也可以作为本次评测对应的汉英和英汉项目数据使用[1]。除了 CWMT2017 组织提供的训练集、开发集和测试集数据外，WMT17 还允许使用下列数据：

1.英语和汉语的平行数据（News Commentary v12 和 UN Parallel Corpus V1.0）

2.英语和汉语的单语训练数据（Europarl，News Commentary，Common Crawl，News Crawl，News Discussions 等）；LDC 的英语和汉语的 Gigaword（LDC2011T07，LDC2009T13，LDC2007T07，LDC2009T27）

蒙汉日常用语翻译项目：

- 内蒙古大学蒙汉平行语料库（2017）（新增 100001 句对）
- 中国科学院计算技术研究所蒙汉双语语料库（2017）（新增 30007 句对）

藏汉政府文献翻译项目：

[1] 请查阅下列网站以获取关于 WMT2017 的准确信息 http://www.statmt.org/wmt17/translation-task.html

- 中国科学院计算技术研究所藏汉双语语料库（2017）（新增 30004 句对）

维汉新闻翻译项目：

- 新疆大学维汉平行语料库（2017）（新增 152527 句对）
- 中国科学院计算技术研究所维汉双语语料库（2017）（新增 30071 句对）
- 中国科学院新疆理化技术研究所维汉双语语料库（2017）（新增 30000 句对）

日汉专利翻译项目：

- 北京语智云帆科技有限公司日汉专利平行语料库（2017）（300 万句对）

参评单位将获得参评子项相关的语料资源。

2. 训练条件

对于以基于平行数据进行训练的机器翻译技术（如统计机器翻译、神经网络机器翻译等）为主的参评系统，可以以“受限”和“非受限”两种方式参与评测。

受限训练：受限训练是指只可以使用评测组织方指定范围的数据进行训练。具体说明如下：

- 参评单位提交的“主系统”必须采用受限训练，以便于在尽可能一致的条件下对不同参评系统所采用的技术进行比较；
- 对于以基于规则的机器翻译技术为主的参评系统，允许采用通过人工方式构造的翻译知识（如规则、模板、词典等），但要在系统描述和技术报告中对所使用的翻译知识的规模、构造和使用方式等给出清晰的说明。
- 单语分析工具可以使用外部数据，如词法分析、句法分析及命名实体识别工具等可以使用外部数据；
- 涉及双语翻译的工具不能使用外部数据，包括命名实体翻译、音字转换工具等（数词和时间词翻译不受此约束）；
- 每个评测项目只允许使用评测组织方发布的该项目相关的训练数据，不可以使用其他评测项目的数据。即对于参加多个评测项目的单位，不同项目提供的数据不可以联合使用。
- 与 WMT 联合组织的汉英、英汉领域评测项目的受限训练语料包括附件四列表中的 CWMT 方提供的数据；也包括由 WMT 组织提供的数据。为便于比较，请参评单位提交汉英、英汉领域系统的评测报告时说明是使用 CWMT 数据还是 WMT 数据还是两者皆有，评测组织方将在发布评测报告时对相应的系统结果予以标识。

非受限训练：非受限训练是指可以使用评测组织方指定范围的数据之外的数据进行训练。具体说明如下：

- 参评单位提交的“对比系统”可以采用非受限训练。
- 采用非受限训练方式的系统，需要在系统描述和技术报告中对使用的数据进行说明（如数据规模和领域类型、是否为可公开获取的数据等。若为可公开获取的数据，则需说明数据出处）。
- 欢迎参评单位使用自有的在线系统参与评测。在线系统一般认定为非受限系统，需要在系统描述和技术报告中对技术做简要说明。在线系统的结果仅作为参考，不参与非受限训练排名。

3. 开发数据

本次评测开发数据的情况请见表 2。

表2 CWMT 2017评测开发数据情况

评测项目名称	规模（单位：句）	提供单位	说明
汉英新闻领域机器翻译	2,002	南京大学	单参考译文

英汉新闻领域机器翻译	2,002	南京大学	单参考译文
蒙汉日常用语机器翻译	1,000	内蒙古大学	4个参考译文
藏汉政府文献机器翻译	650	青海师范大学	4个参考译文
维汉新闻领域机器翻译	700	新疆大学	4个参考译文
日汉专利领域机器翻译	3,000	北京语智云帆科技有限公司	单参考译文

其中，汉英、英汉新闻领域机器翻译项目使用相同的开发集，分别包含由英语翻译成汉语的 1002 句和汉语翻译成英语的 1000 句。

4. 测试数据

本次评测中，各子项的测试数据规模如表 3 所示。测试数据的格式说明请参见附件二。

表3 CWMT 2017评测测试数据规模

评测项目名称	规模（单位：句）	提供单位	说明
汉英新闻领域机器翻译	1,000	南京大学	单参考译文
英汉新闻领域机器翻译	1,000	南京大学	单参考译文
蒙汉日常用语机器翻译	1,001	内蒙古大学	4个参考译文
藏汉政府文献机器翻译	729	青海师范大学	4个参考译文
维汉新闻领域机器翻译	1,000	中国科学院新疆理化技术研究所	4个参考译文
日汉专利领域机器翻译	1,000	北京语智云帆科技有限公司	单参考译文

其中，蒙汉日常用语、藏汉政府文献、维汉新闻领域机器翻译项目使用与 CWMT2015 相同的评测数据。

五. 提交结果

参评单位收到测试数据后，应在规定时间内提交最终翻译结果文件。具体数据格式请参见附件二。对于每个评测子项，参评单位可以提交一个主系统翻译结果（Primary Result）及最多三个对比系统的翻译结果（Contrast Result）。提交的每个结果文件都应包含详细的系统描述，具体要求请参见附件二。

汉英、英汉新闻领域机器翻译项目的参评单位可以选择向 CWMT2017，或 WMT17，或同时向 WMT17 和 CWMT2017，提交测试数据的翻译结果。向 WMT17 提交的结果应满足 WMT17 对结果提交的要求，向 CWMT2017 提交的结果应满足 CWMT2017 对结果提交的要求。

六. 提交技术报告并参加评测研讨会

评测结束后，参评单位应向CWMT 2017研讨会提交一份详细的技术报告，说明系统的架构、原理，使用的主要技术，以及数据使用的情况。参评单位应派至少一人参加CWMT 2017研讨会进行相应技术交流。技术报告的要求请见附件三。

七. 评测日程

1	2017年3月16日	评测报名开始。评测组织方向报名单位提供训练集、开发集数据，以及BLEU-SBP打分程序、格式检查程序（通过ftp方式发放）
2	2017年3月31日	报名截止，停止发放训练集、开发集数据（如错过报名时间请联络组织方）

3	2017年5月15日上午10:00	评测组织方发放汉英、英汉新闻领域机器翻译两个项目的测试数据
4	2017年5月22日下午17:30	参评单位提交汉英、英汉新闻领域机器翻译两个项目的翻译结果
5	2017年5月20日上午10:00	评测组织方发放日汉专利领域、维汉新闻领域、蒙汉日常用语、藏汉政府文献机器翻译四个项目的测试数据
6	2017年5月27日下午17:30	参评单位提交日汉专利领域、维汉新闻领域、蒙汉日常用语、藏汉政府文献机器翻译四个项目的翻译结果
7	2017年6月15日	评测组织方向参评单位通知初步评测结果
8	2017年6月30日	参评单位提交评测技术报告
9	2017年7月10日	评测组织方返回评测技术报告，供参评单位修改
10	2017年7月30日	评测技术报告终稿
11	2017年9月27日- 9月29日	研讨会召开，会上正式报告评测结果并进行研讨

八. 附件

附件一：报名表及评测协议

附件二：机器翻译评测数据文件格式

附件三：技术报告要求

附件四：评测组织方发布的资源清单

附件一：报名表及评测协议

任何从事机器翻译研究或者开发的组织都可以报名参加 CWMT 2017 评测。CWMT 2017 的参评单位必须填写评测报名表和评测协议，通过邮寄或者电子邮件的方式将报名表和评测协议发送给评测组织方。报名表需要有负责人正式签字或者单位盖章。

本次评测不收取注册费用，请所有参评单位至少派一人参加第十三届全国机器翻译研讨会（CWMT 2017）并进行会议交流。

报名表请寄至：

联 系 人： 黄书剑　　电子邮件：huangsj@nju.edu.cn

通信地址：江苏省南京市栖霞区仙林街道仙林大道 163 号南京大学仙林校区计算机科学与技术楼 902 房间

邮政编码：210023　　电话：025-89680690

第十三届全国机器翻译研讨会（CWMT 2017）评测报名表

单位名称			
通信地址			
联系人		联系电话	
邮政编码		电子邮件	
评测项目	☐ 汉英新闻领域 ☐ 英汉新闻领域 ☐ 日汉专利领域 ☐ 维汉新闻领域 ☐ 蒙汉日常用语 ☐ 藏汉政府文献		

参评者保证遵守以下约定：

1. 收到测试数据之后，参评者应该按照评测规定的日期返回运行结果（含系统描述）。
2. 参评者同意提交正式的技术报告，参加第十三届全国机器翻译研讨会，并在会上宣读报告或粘贴评测报告海报，与参会人员交流（参见 CWMT 2017 征文通知）。
3. 参评者确认对参评的系统拥有自主知识产权，如果参评系统部分使用了他人的技术，请在所提交的系统描述中加以明确说明。
4. 参评者保证，对于在评测过程中得到的所有与评测相关的数据，包括训练集、开发集、测试集、参考答案及相关数据、其他单位主系统的结果数据和评测工具，参评者仅用于与本次评测项目相关的研究，不得用于其他任何用途。
5. 参评者保证，只在本单位使用上述的评测数据，不以任何形式（包括电子的、书面的或网络的形式）扩散到其他单位；也不在参评者的下属单位或合资、合作单位中使用该评测数据。
6. 参评者保证，在使用了上述评测数据（包括训练集、开发集、测试集、参考答案及相关数据、其他单位主系统的结果数据和评测工具）完成的科研成果对外发布时，应公开声明使用了上述评测数据。
7. 参评者如果违反 4-6 条款的约定，评测的主办方以及数据提供方有权利要求参评者和所有涉及有偿或无偿得到扩散数据的单位和个人按所使用数据成本价的 3-5 倍赔偿违约金。违约金不足以弥补数据提供方实际损失的，应当予以补偿。

负责人签字或单位盖章：

2017 年　月　日

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 be only 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 organizations of the participating site.

4. The participating site agrees to give credit to 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:

附件二：机器翻译评测数据文件格式

本文档对评测中的相关数据文件及格式进行说明，文件包括评测组织方发放的数据文件以及参评单位需要提交的结果文件。

所有文件均要求为 UTF-8 编码，其中评测组织方发放的开发集（含参考答案）、测试集以及参评单位最终提交的翻译结果文件均为 UTF-8（带 BOM）编码的 XML 文件（所使用的文档类型定义请参见本附件第二部分），其它文件均要求为 UTF-8（无 BOM）编码的纯文本文件。

一、评测组织方发放的数据格式说明

评测组织方发放的数据有三种，分别为：训练集、开发集和测试集。此处，以“汉-英翻译”子项为例说明这些数据文件的格式。

1、 训练集

子训练集由句逐行对应的源语言文件和目标语言文件组成，每行为一个句子。

图 1 和图 2 示例说明了“汉-英翻译”子项中源语言文件和目标语言文件的格式。

汉语
战法训练有了新的突破
第一章总则
人民币依其面额支付
...

图 1 训练语料源语言文件

英语
Tactical training made new breakthrough
Chapter I general rules
The renminbi is paid by denomination
...

图 2 训练语料目标语言文件

2、 开发集

开发集包括源语言文件和参考译文文件两种。

（1）源语言文件

源语言文件包含一个 srcset 元素，这个元素包含以下属性：

setid: 开发集的id
srclang: 源语言标识，值为：en（英语）、zh（汉语）、mn（蒙古语）、uy（维语）、ti（藏语）或jp（日语）
trglang: 目标语言标识，值为：en、zh、mn、uy、ti或jp

srcset 元素包含一个或多个 DOC 元素，每个 DOC 具有属性：

docid: 表示文档名称

DOC 元素还可以包含零个或多个 p 元素，不同的 p 元素对应着不同的篇章。每个 DOC 元素或 p 元素包含 1 个或多个 seg 元素，每个 seg 元素有一个属性 id。

图 3 给出了一个开发集源语言文件的示例。

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

```
...
</p>
...
</DOC>
</srcset>
```

图 3 开发集源语言文件示例

（2） 参考译文文件

参考译文文件包含一个 refset 元素，DOC 元素包含 docid 和 site 两个属性，其中 docid 表示文档名称，site 用以区分不同的参考译文。本次评测中，汉英、英汉和日汉开发集数据的每个句子将有 1 个不同的参考译文，蒙汉、藏汉和维汉开发集数据的每个句子将给出 4 个不同的参考译文。

图 4 给出了一个开发集参考译文的示例。

```
<?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">参考译文 11 </seg>
<seg id="2">参考译文 21</seg>
...
</p>
...
</DOC>
<DOC docid="news" sysid="ref" site="2">
<p>
<seg id="1">参考译文 21 </seg>
<seg id="2">参考译文 22</seg>
...
</p>
...
</DOC>
<DOC docid="news" sysid="ref" site="3">
<p>
<seg id="1">参考译文 31</seg>
<seg id="2">参考译文 32</seg>
...
</p>
...
</DOC>
<DOC docid="news" sysid="ref" site="4">
<p>
```

```
<seg id="1">参考译文 41 </seg>
<seg id="2">参考译文 42</seg>
...
</p>
...
</DOC>
</refset>
```

图 4 开发集参考文件示例

3、 测试集

为了方便组织评测，在评测阶段组织方将只发放测试集源语言文件，其格式与开发集源语言文件格式相同。

二、参评单位需要提交的数据格式说明

参评单位仅需要提供最终的结果文件及系统描述信息，其格式说明如下。

1、 文件命名

所有需要提交的文件的命名方式请遵循下表要求：

（其中：项目代号以 ce 为例，参评单位代号以 ict 为例，评测语料年份以 2017 为例）

文件	文件名模式	文件名举例
最终翻译结果	项目代号-评测语料年份-参评单位代号-主/对比系统-参评系统代号.xml	ce-2017-ict-primary-a.xml ce-2017-ict-contrast-c.xml

2、 最终翻译结果文件

最终翻译结果文件为解码器输出文件经过后处理的结果。

最终翻译结果文件格式与测试集源语言文件格式基本相同。

最终翻译结果文件包含一个 tstset 元素，tstset 元素包含一个 system 元素，system 元素包含 site（说明单位名称）和 sysid（说明系统标识）两个属性。

其中，system 元素应给出参评系统的描述信息，即对以下内容给出说明：

- 软硬件环境：包括操作系统及其版本、CPU 数量、CPU 类型及其频率、系统内存大小等等；
- 运行时间：参评系统从接受输入到产生全部输出所花费的时间；
- 技术概要：简要说明参评系统所采用的主要技术和重要参数，如果采用了系统融合技术，这里要进行说明；
- 训练数据说明：说明参评系统所使用的训练数据和开发数据，并注明是受限训练还是非受限训练，对于英汉、汉英项目，还应注明使用的数据为WMT数据，CWMT数据，还是两者皆有；
- 外部技术说明：说明除了参评单位自己的技术外，还采用了哪些外部技术，包括各种开源代码、自由软件、共享软件或商业软件。

tstset 元素中的 DOC 元素及其内部的 p 元素、seg 元素应与测试集源语言文件中相应的元素及其属性保持一一对应关系。

图 5 示例说明了最终翻译结果文件的格式。

```
<?xml version="1.0" encoding="UTF-8"?>
<tstset setid="zh_en_news_trans" srclang="zh" trglang="en">
```

```
<system site="单位名称" sysid="系统标识">
参评系统的描述信息
...
</system>
<DOC docid="文档名称" sysid="系统标识">
<p>
<seg id="1">提交译文 1</seg>
<seg id="2">提交译文 2</seg>
...
</p>
...
</DOC>
</tstset>
```

图 5 最终翻译结果文件示例

3、 文件细节

关于最终提交的结果文件，请参评单位注意以下细节：

- 本次评测的输入输出文件格式采用严格的 XML 格式。XML 格式与 NIST 评测的文件格式的最主要的区别在于，XML 格式的文件中，标签以外的文本如果出现以下五个字符，必须采用相应的转义序列来表示：

字符	转义序列
&	&amp;
<	<
>	>
"	"
'	'

- 中文译文中，外国人名中间的中圆点统一采用 UTF-8 编码中十六进制为“E2 80 A2”的样式，如“托德·西蒙斯”；
- 英文 token 采用的标准：对语料进行后处理时，按照 NIST 的 mteval-v11b.pl 中 normalizeText 函数的 tokenization 方式。如图 6 所示，我们对其中的主要语句增加了中文注释，供大家参考。

```
# language-dependent part (assuming Western languages):

    $norm_text = " $norm_text ";
    #把原文的最开头和最末尾各加上一个空格（最后再删去）

    $norm_text =~ tr/[A-Z]/[a-z]/ unless $preserve_case;
```

```
#除非用户指定保留大小写，否则一般把英文大写字母转化成小写

$norm_text =~ s/([\{-\~\[-\` -\&\(-\+\:-\@\/])/ $1 /g;      # tokenize punctuation
 #把下述符号两边各加上一个空格（对应 ASCII 字符集中的十六进制值标注在后）：
 #{ | } ~      (0x7b-0x7e)
 #[ \ ] ^ - ` (0x5b-0x60)
 #(空格）！" # $ % & (0x20-0x26)
 #( ) * + (0x28-0x2b)
 # : ; < = > ? @ (0x3a-0x40)
 # / (0x2f)

$norm_text =~ s/([^0-9])([\.,])/$1 $2 /g; # tokenize period and comma unless preceded by a digit
 #当非数字 0-9 后面紧跟着句号"."或者逗号","时,在句号或者逗号两边各加一个空格（即数字后面紧跟句号或逗号时，不在句号或逗号两边加空格）

$norm_text =~ s/([\.,])([^0-9])/ $1 $2/g; # tokenize period and comma unless followed by a digit
 #当句号"."或者逗号","后面没有紧跟数字 0-9 时，在句号或者逗号两边各加一个空格

$norm_text =~ s/([0-9])(-)/$1 $2 /g; # tokenize dash when preceded by a digit
 #当连字符号"-"前面紧跟数字 0-9 时，在"-"两边各加一个空格

$norm_text =~ s/\s+/ /g; # one space only between words
 #把多个空格替换成一个空格

$norm_text =~ s/^\s+//;   # no leading space
 #把文章打头的空格去掉

$norm_text =~ s/\s+$//;   # no trailing space
 #把文章末尾的空格去掉
```

图 6 英文 tokenization 语句及说明

三、CWMT 2017 XML 文件文档结构描述说明

```
<?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>
```

附件三：技术报告要求

所有参评单位应向第十三届全国机器翻译研讨会提交一篇技术报告。技术报告应该比较详细地介绍参评系统所使用的技术，目的是使读者知道该评测结果是如何得到的。一篇好的技术报告应该详细到使读者大致能够重复报告中描述的工作。技术报告应不少于 5000 汉字或 3000 英文词。

技术报告大致应包括以下内容：

引言：介绍背景情况、所参加的评测项目、参评系统概述；

系统：详细介绍参评系统的总体结构和各个模块；要详细介绍所采用的技术。如果是采用公开的技术，应加以明确的说明；如果是自行开发的特有技术，应该详细说明；如果采用了系统融合技术，应对参与系统融合的单系统所采用的技术进行说明，并给出单系统的运行结果；如果采用了人工方式构造的翻译知识（如规则、模板、词典等），要对所使用的翻译知识的规模、构造和使用方式等进行说明；

数据：详细介绍所使用的数据及对数据所进行的处理；

实验：详细介绍参加评测的实验过程、实验参数和实验结果，并对结果进行分析；

总结：（略）

附件四：评测组织方发布的资源清单

如非特殊说明，评测提供的数据文件默认采用 UTF-8 无 BOM 编码

1 汉英/英汉新闻相关资源

1.1 训练数据

资源名称简写及 ChineseLDC 资源编号	资源描述	
Datum2015	名称	点通汉英平行语料库（2015）（部分）
	提供单位	点通数据有限公司
	语种	汉语—英语
	领域	综合领域，包括：语言教材、双语图书、技术文档、双语新闻、政府白皮书、政府公文和 Web 上双语资源等等
	规模	1000004 个句对
	说明	这是点通数据有限公司在 863 项目支持下开发的《双语/多语平行语料库》的部分内容。
CASICT2011 （CLDC-2010-001） （CLDC-2012-001）	名称	计算所 Web 汉英平行语料库（2013）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	综合领域
	规模	1936633 个句对
	说明	该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。语料库抽样评价的正确率在 95%以上。 该研究得到国家自然科学基金项目（编号：60603095）的支持。
CASICT2015	名称	计算所 Web 汉英平行语料库（2015）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	综合领域
	规模	2036834 个句对
	说明	该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。计算所在此基础上进行了大致的校对，语料库抽样评价的正确率在 99%以上。语料构成如下：网络语料占 60%，电影字幕语料占 20%，来自英汉辞书的例句语料占 20%。
CASIA2015	名称	中科院自动化所 Web 汉英平行语料库（2015）
	提供单位	中国科学院自动化研究所
	语种	汉语—英语

	领域	综合领域
	规模	1050000 句对
	说明	该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。
Datum2017	名称	点通公司英汉平行语料库（2017）
	提供单位	点通数据有限公司
	语种	汉语—英语
	领域	
	规模	100 万句对，分为 20 个文件
	说明	
NEU2017	名称	东北大学英汉平行语料库（2017）
	提供单位	东北大学 自然语言处理实验室
	语种	汉语—英语
	领域	
	规模	200 万句对
	说明	
SSMT2007 MT Evaluation Data （2007-863-001）	名称	SSMT2007 机器翻译评测数据（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	新闻
	规模	该机器翻译测试语料包含 2 个翻译方向（汉英、英汉），语料为新闻领域。其中汉英机器翻译测试语料含 1,002 个汉语句子。英汉机器翻译测试语料含 995 个英语句子。每个测试句子包括 4 个人工翻译的参考译文。
	说明	
HTRDP(863)2005 MT Evaluation Data （2005-863-001）	名称	2005 年 863 机器翻译评测数据（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	包括两种评测语料，一种是对话语料，领域为奥运相关领域，包括体育赛事、天气预报、交通住宿、旅游餐饮等；一种是篇章语料，领域为新闻领域。
	规模	汉英对话句对：467 句，汉英篇章句对：489 句。 英汉对话句对：459 句，英汉篇章句对：494 句。 每个翻译方向的每个测试句子各提供 4 个人工翻译的参考译文。
	说明	
HTRDP(863)2004 MT Evaluation Data	名称	2004 年 863 机器翻译评测数据（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所

（2004-863-001）	语种	汉语—英语
	领域	两种评测语料，一种是篇章语料，一种是对话语料。领域是通用领域和奥运的相关领域，其中奥运领域包括体育赛事、天气预报、交通住宿、旅游餐饮等。
	规模	汉英评测数据含 400 句对话语料，308 句篇章语料。英汉评测数据含 400 句对话语料，310 句篇章语料。每个翻译方向的每个测试句子各提供 4 个人工翻译的参考译文。
	说明	2004 年 863 机器翻译评测汉英、英汉部分测试数据。
HTRDP(863)2003 MT Evaluation Data （2003-863-004）	名称	2003 年 863 机器翻译评测数据（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	奥运相关领域，其中奥运领域包括体育赛事、天气预报、交通住宿、旅游餐饮等。
	规模	汉英评测数据含 437 句对话语料和 169 句篇章语料；英汉评测数据含 496 句对话语料和 322 句篇章语料。每个翻译方向的每个测试句子各提供 4 个人工翻译的参考译文。
	说明	2003 年 863 机器翻译评测汉英、英汉部分测试数据。
CWMT2008 Machine Translation Evaluation Data （CLDC-2009-001） （CLDC-2009-002）	名称	CWMT2008 机器翻译评测新闻语料（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	新闻
	规模	汉英评测数据含 1006 句对；英汉评测数据含 1000 句对。每个翻译方向的每个测试句子各提供 4 个人工翻译的参考译文。
	说明	
CWMT2009 Machine Translation Evaluation Data	名称	CWMT2009 机器翻译评测数据（英汉/汉英机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	汉语—英语
	领域	新闻
	规模	汉英评测数据含 1003 句对；英汉评测数据含 1002 句对。每个翻译方向的每个测试句子各提供 4 个人工翻译的参考译文。
	说明	
CWMT2011 Machine Translation Evaluation Data	名称	CWMT2011 机器翻译评测数据（英汉机器翻译部分）
	提供单位	中国科学院计算技术研究所
	语种	英语 → 汉语
	领域	新闻
	规模	英汉评测数据含 3187 句对。每个测试句子各提供 4 个人工翻译的参考译文。
	说明	

1.2 单语新闻数据

XMU-CWMT2017	名称	厦门大学 NLP 实验室新华网新闻汉语单语语料（2017）
	提供单位	厦门大学
	语种	汉语
	领域	新闻
	规模	现语料库共有 662,904 个文章，大约 1100 万词汇。
	说明	本资源由厦门大学 NLP 实验室收集，包括新华网 2011 年不同主题频道的新闻语料，例如：国内新闻，国际新闻，财经新闻，论坛，教育等。 每篇文章包含：标题，日期，URL 和内容。

1.3 开发集数据

newsdev2017-enzh（newsdev2017-zhen）	名称	南京大学 2017 汉英/英汉新闻语料开发集数据
	提供单位	南京大学
	语种	汉语—英语
	领域	新闻
	规模	共 2,002 句对
	说明	包含 1000 个汉语新闻句子及其英语翻译结果，以及 1002 个英语新闻句子及其汉语翻译结果。

2 蒙汉日常用语项目数据

2.1 训练数据

IMU-CWMT2013 （CLDC-2010-005）	名称	内蒙古大学汉蒙平行语料库（2013）
	提供单位	内蒙古大学
	语种	汉语—蒙古语
	领域	政府文献和法律法规、日常对话、文学
	规模	共 104,975 句对 其中：CWMT2011 评测训练语料 67274 句对,领域包括: 日常对话、文学、政府文献和法律法规; CWMT 2015 新增训练语料:包括新闻语料 17,516 句对,政府文献语料 10,394 句对,课本语料 5,052 句对,蒙汉字典语料 4,739 句对;
	说明	
IMU-CWMT2015	名称	内蒙古大学汉蒙平行语料库（2015）
	提供单位	内蒙古大学
	语种	汉语—蒙古语
	领域	政府文献和法律法规、日常对话、文学
	规模	共 24,978 句对

	说明	
IIM-CWMT2015	名称	中国科学院合肥智能机械研究所蒙汉双语语料库（2015）
	提供单位	中国科学院合肥智能机械研究所
	语种	蒙古语 → 汉语
	领域	新闻
	规模	1,682 句对
	说明	
ICT-MC-corpus-CWMT2017	名称	中国科学院计算技术研究所蒙汉双语语料库（2017）
	提供单位	中国科学院计算技术研究所
	语种	蒙古语 → 汉语
	领域	新闻
	规模	30,007 句对
	说明	
IMU-corpus-CWMT2017	名称	内蒙古大学蒙汉双语语料库（2017）
	提供单位	内蒙古大学
	语种	蒙古语 → 汉语
	领域	综合，包括：政府文件，政府工作报告，国务院文件，法律法规等
	规模	100,001 句对
	说明	

2.2 开发集数据

IMU-dev-mnzh - CWMT2017	名称	内蒙古大学蒙汉开发集数据
	提供单位	内蒙古大学
	语种	蒙古语 → 汉语
	领域	政府文献和法律法规、日常对话、文学
	规模	共 1,000 句蒙古语，每句 4 个汉语参考译文
	说明	CMWT2017 蒙汉开发集与 CWMT2011、CWMT2013、CWMT2015 蒙汉开发集相同

3 藏汉政府文献相关资源

3.1 训练数据

QHNU-CWMT2013	名称	青海师范大学藏汉平行语料库（2013）
	提供单位	青海师范大学
	语种	藏语—汉语
	领域	政府文献领域
	规模	33,145 句对

	说明	该平行语料库是通过录入、扫描、网页下载等方式获得的。双语平行的搜集、整理、确认、获取、句子对齐等过程是通过程序自动实现和人工干预实现的。语料库的正确率在 99%以上。 该研究得到国家自然科学基金项目（编号：61063033）和 973 前期研究专项（编号：2010CB334708）的支持。
QHNU-CWMT2015	名称	青海师范大学藏汉平行语料库（2015）
	提供单位	青海师范大学
	语种	藏语—汉语
	领域	政府文献领域
	规模	17,194 句对
	说明	该平行语料库是通过录入、扫描、网页下载等方式获得的。双语平行的搜集、整理、确认、获取、句子对齐等过程是通过程序自动实现和人工干预实现的。语料库的正确率在 99%以上。 该研究得到国家自然科学基金项目（编号：61063033）的支持。
XBMU-XMU	名称	央金藏汉平行语料库
	提供单位	厦门大学人工智能研究所 西北民族大学语言（技术）研究所
	语种	汉语 → 藏语
	领域	综合领域
	规模	52,078 句对
	说明	1)该藏汉平行语料库是用正式出版物、藏汉大词典和网络语料藏汉对照文本,经使用自主开发的“藏汉句子对齐工具”初步对齐之后,由人工逐句对齐。 2)5 万句对藏汉平行语料的对齐正确率为 100%。 3)该研究得到国家社科基金重点项目《藏语语料库建设研究》（批准号：05AYY001）和 863 重点项目《面向跨语言搜索的机器翻译关键技术研究》（批准号：2006AA010107）的支持。
XBMU-XMU-UTibet	名称	西北民族大学、西藏大学与厦门大学藏汉语料（2012）
	提供单位	西北民族大学语言（技术）研究所 西藏大学 厦门大学人工智能研究所
	语种	汉语 → 藏语
	领域	政论，法律
	规模	24,159 句对
	说明	语料来源：2008 年和 2009 年全国最新法律文件和十八大报告、2011、2012 年政府工作报告等，政论类与法律类语料各占一半。系西北民族大学、西藏大学与厦门大学于 2012 年通过对原材料进行扫描、识别、校对并独立加工完成。
ICT-TC-corpus-CWMT2017	名称	中国科学院计算技术研究所藏汉双语语料库（2017）
	提供单位	中国科学院计算技术研究所
	语种	藏语 → 汉语

	领域	新闻
	规模	30,004 句对
	说明	

3.2 开发集数据

QHNU-dev-tizh-CWMT2017	名称	青海师范大学藏汉开发集数据
	提供单位	青海师范大学
	语种	藏语 → 汉语
	领域	政府文献
	规模	共 650 句藏语，每句 4 个汉语参考译文
	说明	CMWT2017 藏汉开发集与 CWMT2011、CWMT2013、CWMT2015 藏汉开发集相同

4 维汉新闻相关资源

4.1 训练数据

XJU-CWMT2013 （CLDC-2013-002）	名称	新疆大学维汉双语句子对齐语料库（2013）
	提供单位	新疆大学
	语种	汉语 → 维吾尔语
	领域	新闻
	规模	79,935 句对
	说明	
XJIPC-CWMT2015	名称	中国科学院新疆理化技术研究所维汉双语语料库（2015）
	提供单位	中国科学院新疆理化技术研究所
	语种	汉语 → 维吾尔语
	领域	新闻
	规模	59,990 句对
	说明	此语料在 CWMT 2013 年的基础上新增加约 3 万句对 语料比例：2007 年-2014 年媒体新闻类语料比例约占 95%，2012-2014 年政府报告类语料和法律法规类语料合计比例约占 5%。 语料来源：系中国科学院新疆理化技术研究所在 2012 年-2014 年间采集、标注、校对完成。
ICT-UC-corpus-CWMT2017	名称	中国科学院计算技术研究所维汉双语语料库（2017）
	提供单位	中国科学院计算技术研究所
	语种	维吾尔语 → 汉语
	领域	新闻
	规模	30,071 句对
	说明	
	名称	新疆大学维汉双语平行语料库（2017）

XJU-corpus-CWMT2017	提供单位	新疆大学
	语种	维吾尔语 → 汉语
	领域	新闻
	规模	152,527 句对
	说明	
XJIPC-corpus-CWMT2017	名称	中国科学院新疆理化技术研究所维汉双语语料库（2017）
	提供单位	中国科学院新疆理化技术研究所
	语种	维吾尔语 → 汉语
	领域	新闻
	规模	30,000 句对
	说明	

4.2 开发集数据

XJU-dev-uyzh-CWMT2017	名称	新疆大学维汉开发集数据
	提供单位	新疆大学
	语种	维吾尔语 → 汉语
	领域	新闻
	规模	共 700 句维语，每句 4 个汉语参考译文
	说明	CMWT2017 维汉开发集与 CWMT2011、CWMT2013、CWMT2015 维汉开发集相同

5 日汉专利领域相关资源

5.1 训练数据

Lingosail-train-CWMT2017	名称	北京语智云帆科技有限公司日汉专利平行语料库（2017）
	提供单位	北京语智云帆科技有限公司
	语种	日语 → 汉语
	领域	综合
	规模	3,000,000 句对
	说明	

5.2 开发集数据

Lingosail-dev-CWMT2017	名称	北京语智云帆科技有限公司日汉双语开发集数据（2017）
	提供单位	北京语智云帆科技有限公司
	语种	日语 → 汉语
	领域	综合
	规模	3000 句日语，每句含一个汉语参考译文

	说明	

5.3 汉语专利数据

Lingosail-cn_for_lm-CWMT2017	名称	北京语智云帆科技有限公司汉语专利语料（2017）
	提供单位	北京语智云帆科技有限公司
	语种	汉语
	领域	综合
	规模	7,114,700 句对
	说明	