第四届全国机器翻译研讨会（CWMT2008）评测大纲

一、 引言

第四届全国机器翻译研讨会（CWMT2008）即将于2008年11月27-28日举行。该研讨会是前三届统计机器翻译研讨会（SSMT2005、SSMT2006、SSMT2007）的延续。

根据惯例，本次研讨会将继续组织统一的机器翻译评测，以推进参评单位的实质性交流和机器翻译技术的发展。

本次评测的主办机构为：

中国中文信息学会

本次评测的组织单位是：

中国科学院计算技术研究所

本次评测的合作单位包括：

中国科学技术信息研究所
微软亚洲研究院

本次评测的资源提供单位包括：

北京大学
哈尔滨工业大学
厦门大学
万方数据公司
中国科学技术信息研究所
中国科学院自动化研究所
中国科学院计算技术研究所

本次评测的评测委员会主席是：

刘群（中国科学院计算技术研究所）

本次评测的评测委员会委员有：

黄德根（大连理工大学）
林钦佑（微软亚洲研究院）
吕雅娟（中国科学院计算技术研究所）
王惠临（中国科学技术信息研究所）
杨沐昀（哈尔滨工业大学）
张玉洁（日本情报通信研究机构，NICT）
宗成庆（中国科学院自动化研究所）

有关会议和评测的更多信息请参见以下网址：

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

二、 评测项目

本次评测的项目设置如下：

语种与领域	评测项目	项目代号
汉英新闻领域	机器翻译	ZH-EN-NEWS-TRANS
汉英新闻领域	系统融合	ZH-EN-NEWS-COMBI
英汉新闻领域	机器翻译	EN-ZH-NEWS-TRANS
英汉科技领域	机器翻译	EN-ZH-SCIE-TRANS

1. “机器翻译”项目

“机器翻译”项目的评测采用目前国际上普遍采用的评测方式。由评测组织方提供测试数据，参评单位在给定时间内返回翻译结果，由评测组织方进行评价。主要采用自动评测指标。

“机器翻译”项目允许采用评测组织方提供的数据之外的数据进行训练。在最后公布的评测结果中，对使用了外部数据的系统将加以明确的标记。

“机器翻译”项目的每个参评单位必须提交一个正式结果，最多可以提交两个对比结果。

对于汉英新闻领域“机器翻译”项目，由于“机器翻译”项目的输出文件将用于后面的“系统融合”项目评测，因此，每个“机器翻译”项目参评单位除了按照上述要求提交翻译结果以外，还应完成以下工作：

1) 除了在给定的测试数据上提交评测结果以外，参评单位还应提供同一个参评系统在SSMT2007汉英机器翻译测试数据上的翻译结果，用作“系统融合”项目的训练数据，为此本次评测要求在汉英新闻领域的“机器翻译”项目中，不得使用SSMT2007的测试数据作为训练数据。有关SSMT2007机器翻译评测数据的说明，请参加本大纲附录五（ChineseLDC资源编号：2007-863-001）。

2) 如果参评系统所采用的方法允许，请对所有的测试数据（包括本次评测的测试数据和SSMT2007的测试数据）提供N-best的翻译结果，要求给出的N-best结果按照评分从高到低排序。N最多不超过200。“机器翻译”项目评测只使用评分最高的结果进行评价，其他结果将提供给“系统融合”项目的参评单位作为输入数据。如果参评系统所采用的方法无法提供N-best的翻译结果，则可以不提供。

2. “系统融合”项目

1) 在“机器翻译”项目评测结束后，将所有参评单位的N-best翻译结果发给“系统融合”项目的所有参评单位；

2) “系统融合”项目的参评单位在上述的多个评测结果基础上进行系统融合，并给出系统融合的结果；在系统融合的过程中，参评单位可以利用各系统在

SSMT2007评测数据上的输出结果以及SSMT2007评测数据的参考译文获得各参评系统性能的先验知识。

3）评测组织方对所有系统融合的结果进行自动评价并给出评价结果。

4) “系统融合”项目都采用训练语料受限的方式，也就是说，仅允许采用评测组织者提供的语言数据资源进行训练。

三、 评测指标

1. “机器翻译”项目的评测指标

“机器翻译”项目的自动评测采用多种自动评价标准，包括：BLEU、NIST 、GTM 、mWER 、mPER、ICT和Woodpecker。自动评测的算法是大小写敏感的。中文的评测是基于字的，而不是词。要求中文译文中所有非汉字字符（比如数字、字母和标点符号）都转换成半角字符。

Woodpecker是微软亚洲研究院研制的新的机器翻译评测指标。该指标的评测使用Woodpecker系统平台，自动评测翻译系统对各种语言学知识（称为检测点）的翻译能力。依据测试数据集和参考翻译数据集之间的词对齐结果和它们的句法分析树，Woodpecker系统首先自动从源语言和目标语言中抽取各种检测点类型，包括名词短语，动宾搭配，介词短语，新词等几十种语言学类型。然后，Woodpecker系统通过计算测试语料中检测点的参考翻译结果与翻译系统翻译结果之间的匹配程度来评估翻译系统在特定语言学现象方面的翻译能力。有关Woodpecker评测系统的相关介绍参见附录六。

2. “系统融合”项目的评测指标

本次“系统融合”项目采用对融合后的结果进行评估的方法。评测指标将采用唯一的BLEU指标。

四、 评测数据

1. 评测数据说明

a） 训练数据

评测组织者将提供一些语言资源，供参评单位用作系统训练之用。资源的清单请参见附件。

b） 测试数据

本次评测的测试数据分为新闻类和科技类两种类型。

c） 分割日期

为了确保训练数据和测试数据不会重叠，评测组织方定义了一个训练数据和测试数据的分割日期（Cut-off Date）。本次评测定义的分隔日期是2008年1月1日。

所有的训练数据和开发数据，包括评测组织方提供的数据和参评单位自己收集的数据，都必须是在分割日期之前（不含分割日期）产生的数据。

评测组织方提供的测试数据将是在截止日期之后（含分割日期）产生的数据。

2. 训练数据使用规定

a) “机器翻译”项目

“机器翻译”项目可以使用任何数据进行训练。

如果使用评测组织者提供的数据之外的数据（以下简称外部数据），那么应该说明这些数据是否是可以公开获得的数据。如果是可以公开获得的数据，参评单位应该说明所采用的外部数据的出处；如果不是可以公开获得的数据，参评单位应该说明该数据的内容和规模。

在评测组织方公布的评测结果报告中，对于使用外部数据的系统，将会加以专门的标记。

特别要说明的是，对于汉英新闻领域“机器翻译”项目，不允许使用SSMT2007的测试数据（ChineseLDC资源编号：2007-863-001）作为训练数据。

b) “系统融合”项目

“系统融合”项目的参评系统，仅允许使用评测组织者提供的数据，不允许使用任何外部数据进行训练。

五、 评测日历

1.	报名截止日期	2008年8月31日
2.	**评测组织方**发放训练数据	2008年8月31日
3.	**评测组织方**发放所有翻译方向“机器翻译”项目测试数据	2008年10月8日
4.	**各翻译方向“机器翻译”项目参评单位**提交运行结果和系统描述	2008年10月15日
5.	**评测组织方**发放汉英“系统融合”项目的测试数据（即“机器翻译”项目参评单位提交的运行结果的汇总）	2008年10月23日
6.	**汉英“系统融合”项目参评单位**提交的运行结果和系统描述	2008年10月30日
7.	**评测组织方**进行结果评估，向**所有参评单位**通知评测结果	2008年11月8日
8.	**所有参评单位**提交参加评测的技术报告	2008年11月15日
9.	在研讨会上进行研讨	2008年11月27-28日

六、 附件

本大纲包括以下附件：

附件一：报名表

附件二；输入输出文件格式

附件三：系统描述要求

附件四：技术报告要求

附件五；评测组织方发布的资源清单

附件六：Woodpecker机器翻译评测系统简介

附件一：报名表

任何从事机器翻译研究或者开发的组织都可以报名参加CWMT2008评测。参评系统所采用的方法不限，可以是基于规则的、基于实例的、或者基于统计的。CWMT2008的参评单位请填写以下表格，并同时通过电子邮件和信件（或传真）两种方式发送给组织者。后者需要有负责人正式签字或者单位盖章。

本次评测注册费为：中国大陆地区单位：3000元人民币，中国港澳台及国外单位：1000美元。

报名截止日期为：2008年8月31日

联系方式：

联 系 人：赵红梅

电子邮件：zhaohongmei@ict.ac.cn

通信地址：北京市中关村科学院南路6号中科院计算所

邮政编码：100190

电　　话：+86-10-62600667

传　　真：+86-10-82611846

第四届全国机器翻译研讨会（CWMT2008）
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	● 机器翻译 □ 汉英新闻领域　　□ 英汉新闻领域 □ 英汉科技领域 ● 系统融合 □ 汉英新闻领域		

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和系统描述。

2. 参评者同意提交正式的技术报告，并参加第四届全国统计机器翻译研讨会（见 CWMT2008 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；也不在评测者的下属单位或合资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（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。

负责人签字或单位盖章：

2008 年　月　日

附件二：输入输出文件格式

1、 提交结果文件的命名方式

参评单位提交的测试结果文件请按照以下方式命名：

参评项目 - 参评单位 -“Primary/Contrast”- 系统名称 . xml

举例来说，中科院计算所（ICT）参加汉英新闻领域系统“机器翻译”项目评测，提交了一个正式系统 systema 的结果，两个对比系统 systemb 和 systemc 的结果，那么这三个结果文件应该分别命名为：

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、 “机器翻译”项目文件格式

源语言测试数据存放在一个**源语言文件**中，要求参评系统对**源语言文件**产生一个对应的**目标语言文件**，**源语言文件**和**目标语言文件**都采用标准的xml格式。

(1) 源语言文件格式

源语言文件为一个XML文件，字符编码为UTF-8。

每个**源语言文件**包含一个<srcset>元素，这个元素包含setid、srclang、tgtlang属性，分别说明测试集的id、源语言和目标语言代码。<srcset>包含若干<doc>元素（由<doc>和</doc>括起来的部分），其中每个<doc>元素对应于一篇被翻译的文章，<doc>元素的属性说明该文章的相关信息。docid属性给出文档名称，属性值用双引号引起。Srclang属性为源语言代码，tgtlang属性为目标语言代码。语言代码中，英语用“en”表示，日语用“ja”表示，汉语用“zh”表示。

每个<doc>元素由若干个 元素（由 和 括起来的部分）组成。每个 元素由若干个<s>元素（由<s …>和</s>括起来的部分）组成，其中<s>元素的属性id的值是正整数。同一个<doc>中的每个<s>元素的id各不相同，但不一定是连续的数值。每个<s>元素可能包含一个或多个句子。也可能没有 元素，而<s>元素直接包含在<doc>元素中。

```
<?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>
```

<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>

<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>

<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>

</doc>

</srcset>

(2) 目标语言文件格式

目标语言文件也采用xml格式，**目标语言文件**与**源语言文件**格式基本相同，内容一一对应。

源语言文件中的每个<srcset>元素在**目标语言文件**中应该有一个对应的<tgtset>元素，其属性不变。

<tgtset>元素中应包含一个<system>元素，包含对产生该目标语言文件的系统的描述。<system>元素有两个属性：site表示参评单位名称，sysid表示参评系统标识符。一个参评单位可以提供多个参评系统的结果。<system>元素中应该有对参评系统的描述信息。

<tgtset>元素中还应包含与源语言文件对应的<doc>元素。<doc>元素及其内部的 元素、<s>元素应与源语言文件一一对应。对应的<doc>元素的docid属性和<s>元素的id属性应与**源语言文件**相同。<s>元素应包含一段译文文本和若干个<cand>元素。其中，译文文本对应于**源语言文件**中对应的<s>元素的最优翻译结果，用于“机器翻译”项目评测。<s>元素应该有一个score属性，用于记录最优翻译结果的评分。多个<cand>元素为“系统融合”项目提供其他n-best候选翻译结果。每个<cand>元素对应于一个候选翻译结果。每个<cand>元素应该有一个score属性，表示该候选翻译结果的评分。<s>元素和<cand>元素的score属性值（即译文评分）应该是一个0到1之间的实数，评分越高表示该候选译文越好。默认所有的候选翻译结果按照评分从高到低的顺序排列。

<?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">

* *

<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>

* *

* *

<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>

* *

* *

<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>

* *

* *

<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>

* *

</doc>

*</**tgtset**>*

3、“系统融合”项目文件格式

“系统融合”项目的输入文件就是“机器翻译”项目中各参评单位提交的输出文件。

“系统融合”项目的输出文件格式类似于“机器翻译”项目的输出格式文件，但只需要输出 1-best 结果，而且不需要输出译文评分。

附件三：系统描述要求

参评单位在提交运行结果时，除按规定提交相关的输出数据文件外，对每一个项目，还应提交一个系统描述，系统描述嵌入到提交的XML格式的结果文件中。系统描述对以下问题给出说明：

- 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；
- 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；
- 外部技术说明：说明除了参评单位自己的技术外，还采用了那些外部技术，包括各种开源代码、自由软件、共享软件或商业软件。

附件四：技术报告要求

所有参评单位应向第四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，目的是使读者知道你的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。不少于5000汉字或3000英文词。

一篇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；

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。

附件五：评测组织方发布的资源清单

ChineseLDC 资源编号	资源描述	
CLDC-LAC-2003-004	提供单位	中国科学院计算技术研究所 中国科学院自动化研究所
	语种	汉语-英语
	领域	综合
	规模	原语料库包括双语对齐文本 3384 个，共计 209486 个中英双语句子对，其中自动化所加工 3098 个文件，10,7436 个汉英句对，计算所加工 250 个文件，共 102050 个汉英句对。后经自动化所和计算所补充部分语料,现语料库共有 337301 个汉英句对。
	说明	本资源是在国家 973 子课题支持下建立的、大规模、具有统一标准和规范、多领域、多体裁的句子级对齐的双语语言信息和知识库。
CLDC-LAC-2003-006	提供单位	北京大学计算语言学研究所
	语种	汉英，汉日
	领域	综合
	规模	汉英句子级对齐语料 20 万句对；汉日句子级对齐语料 2 万句对；汉英词汇级对齐语料 1 万句对。
	说明	在 863 课题《中文平台总体技术研究与基础数据库建设》子课题《汉英/汉日多语语料库》（编号：2001AA114019）资助下开发而成。 CWMT2008 评测仅提供汉英部分句子级对齐语料。
	提供单位	厦门大学
	语种	英语→汉语
	领域	对话
	规模	176148 句子对
	说明	电影字幕
	提供单位	哈尔滨工业大学信息检索实验室
	语种	英语-汉语
	领域	综合
	规模	100000 句子对
	说明	
	提供单位	哈尔滨工业大学机器翻译课题组
	语种	英语-汉语
	领域	综合

	规模	52227 句子对
	说明	
	提供单位	中国科学技术信息研究所
	语种	英语→汉语
	领域	科技
	规模	
	说明	英文科技文献摘要及其汉语翻译
	提供单位	中国科学院计算技术研究所 万方数据公司
	语种	汉语→英语
	领域	科技
	规模	10 万篇双语论文摘要
	说明	从中文期刊中提取的中英文双语论文摘要
2007-863-001	提供单位	中国科学院计算技术研究所
	语种	汉语→英语，英语→汉语
	领域	新闻
	规模	本次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 1002 个汉语句子，42256 汉字。英汉机器翻译测试语料含 995 个英语句子，23627 个词。每个测试句子包括四个人工翻译的参考译文。 本次词语对齐评测的测试语料为汉英双语对齐的句子，其中包括汉语翻译成英语的句子 251 对，含 7030 个汉字和 8109 个英语词，英语翻译成汉语的句子 253 对，含 6872 个汉字和 6981 个英语词。语料也来自新闻领域。每个汉英双语句对由两个不同的人独立进行人工标注。
	说明	SSMT2007 机器翻译评测的测试数据及相关文档、软件。 CWMT2008 评测仅提供机器翻译测试数据，不提供词语对齐测试数据。 特别需要说明的是，本项数据不得用于 CWMT2008 评测中汉英新闻领域“机器翻译”项目。
2005-863-001	提供单位	中国科学院计算技术研究所 日本情报通信研究机构
	语种	机器翻译评测语料含 6 个翻译方向：汉语→英语，英语→汉语，汉语→日语，日语→汉语、英语→日语，日语→英语。 词语对齐评测语料仅含一个语言对：汉语—英语。
	领域	机器翻译评测包括两种语料的评测，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。 词语对齐评测仅采用篇章语料，领域为新闻领域。

	规模	机器翻译评测数据： 汉英、汉日采用相同的测试数据，含对话和篇章各约 460 句。 英汉、英日采用相同的测试数据，含对话和篇章各约 460 句。 日英、日汉采用相同的测试数据，含对话和篇章各约 460 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。 词语对齐评测数据：含 505 个句子对，每个句子对提供两个人工标注的词语对齐结果。
	说明	2005 年 863 机器翻译评测的测试数据及相关文档、软件。 CWMT2008 评测仅提供汉英、英汉部分的机器翻译测试数据。不提供词语对齐的测试数据。
2004-863-001	提供单位	中国科学院计算技术研究所 日本情报通信研究机构
	语种	机器翻译评测语料含 5 个翻译方向：汉英、汉日、汉法、英汉、日汉。
	领域	本次评测包括两种语料的评测，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	机器翻译评测数据： 汉英、汉日、汉法采用相同的测试数据，含对话语料 2 个文件 400 句，篇章语料 24 个文件 308 句。 英汉评测数据含对话语料 2 个文件 400 句，篇章语料 29 个文件 310 句。 日汉评测数据含对话语料 2 个文件 400 句，篇章语料 16 个文件 309 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	2004 年 863 机器翻译评测的测试数据及相关文档、软件。 CWMT2008 评测仅提供汉英、英汉部分测试数据。
2003-863-001	提供单位	中国科学院计算技术研究所
	语种	机器翻译评测语料含 4 个翻译方向：汉英、英汉、汉日、日汉五个方向。
	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	机器翻译评测数据： 汉英、汉日采用相同的测试数据，含对话语料 437 句，篇章语料 15 个文件。 英汉评测数据含对话语料 496 句，篇章语料 13 个文件。 日汉评测数据含对话语料 410 句，篇章语料 30 个文件。 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 机器翻译评测的测试数据及相关文档、软件。 CWMT2008 评测仅提供汉英、英汉部分测试数据。

附件六：Woodpecker机器翻译评测系统简介

1、 系统概况

Woodpecker评测系统是由微软亚洲研究院自然语言计算组研发的一个用于自动评测机器翻译系统翻译性能的平台工具。与其他评测方法（如Bleu）不同，Woodpecker系统的评测方法不是对翻译系统在测试数据集上的翻译结果计算出一个总体得分，而是计算评估翻译系统对包含在测试数据中各种具体语言学知识的翻译能力,其中各种语言学知识称为检测点(Check-Point)。

Woodpecker系统具有两种功能。一方面，Woodpecker系统可以借助于测试数据集和参考翻译数据集上的词对齐结果和句法分析树来自动完成检测点的抽取；另一方面，Woodpecker系统可以根据特定算法计算出在给定测试数集上检测点参考翻译结果与翻译系统翻译结果之间的匹配程度，并以此作为评估翻译系统在某种特定语言学知识方面翻译能力的依据。

Woodpecker系统评测方法是在俞士汶教授提出的基于检测点的机器翻译评测方法MTE基础（Yu, 1993）上做了若干扩展之后形成的。（Yu, 1993）主要基于专家建立语言学检测点体系，此方法在建立新的测试集的时候，需要花费很大代价手工建立检测点。Woodpecker系统利用句法分析器、自动词汇对齐工具等手段大批量自动地建立检测点，而且可以针对某一种检测点，专门制作测试集合。Woodpecker系统采用了一系列方法提高自动抽取的检测点的可信度，比如筛选测试简单句子，利用多个语法分析器融合决策选择可信度高的检测点，增加测试语料规模克服词对齐结果的噪声等。另外，在（Yu, 1993）中所采用的打分机制是硬性的二元化，即全匹配的时候得一分，部分匹配或不匹配的时候得零分，而Woodpecker系统的打分机制采用了通行的基于N-Gram的匹配计算公式，它可以更好地体现匹配程度，尤其是对部分匹配情况也可以打分。目前，Woodpecker系统不仅支持英汉翻译的检测点评测任务，还支持汉英翻译的检测点评测任务。Woodpecker系统由于采用了自动方法抽取检测点，因此当对一种新语言对的互译质量进行评测时，它也能体现出比较好的自适应能力。

2、 检测点的抽取

Woodpecker系统中定义的检测点是一个语言学单位（例如歧义词，名词短语，动宾搭配，介词短语等），可以建立在不同语法层次上。按照检测点的抽取来源，检测点可以分为源语言检测点和目标语言检测点两大类，任何检测点中的参考翻译是指该检测点中包含的目标语言内容，其中源语言检测点中的参考翻译是通过词对齐结果确定出的目标语言内容，检测点评测时只计算检测点的参考翻译与翻译系统翻译结果之间的匹配程度。若一个测试语句对应有多个参考翻译语句，则每个检测点中也会包含有多个参考翻译。

检测点类型按照翻译任务的不同可以形成汉英和英汉两个独立的分类体系，每个分类体系中包含词、短语和句子三个级别（Level）的检测点类型，每一个级别类型由

若干检测点组（Group）构成，每个组包含若干小组或检测点类(Category)，于是整个体系自顶向下形成一个树状结构，其中叶子节点为检测点类（如名词、动词、介词等）。检测点类在测试集中的每一次具体出现将被单独视为一个检测点，因此一个检测点类在一个测试语句中可能对应有多个检测点。为满足用户的评测需求，分类体系中任何的检测点组以及具体的检测点类都可以由用户自定义成一个新的检测点组，在检测点的评测过程中，只有用户自定义的检测点组所覆盖的所有检测点类作为评测对象，而且评测结果中不仅给出具体检测点类的得分，也给出每个检测点组的得分。

这里为简明起见，下表中只给出了部分的汉英检测点类型以供参考，详细的汉英和英汉检测点类型可参见Table 1和Table 2，其中词和短语级别的检测点类有可能进一步分为源语言和目标语言检测点类。

汉英检测点		
级别	**名称**	**例子**
词	歧义词	打(play)
	新词	朋克(Punk)
	介词	于(in), 在(at)
短语	短语搭配	油炸-食品(fired – food)
	词汇重叠用法	看看(have a look)
	主谓短语	他*说, (he*said)
句子	“把”字句	他把(BA)书拿走了. (He took away the book.)
	“被”字句	花瓶被(BEI)打碎了. (The vase was broken.)

基于以上检测点的分类体系，Woodpecker系统可以单独为每一种翻译任务自动抽取出所有检测点，抽取过程如下：

(1).准备测试集的双语语料，每一个源语言句子可以对应有多个参考翻译句子。

(2).对源语言和目标语言分别进行句法分析，得到句中每个词的词性，词与词之间的依存关系以及层次结构信息。使用的句法分析工具包括Stanford statistical parser和Berkeley statistical parser。

(3).对源语言句子和目标语言句子做词语对齐。可以使用自动词对齐工具，比如GIZA++或者类似的工具，也可接受人工的词对齐结果。

(4).依据源语言和目标语言的句法分析树抽取出各种类型的检测点，同时根据词对齐结果确定检测点中源语言所对应的目标参考翻译内容。

3、 基于检测点的评测

基于用户选定的检测点集合，Woodpecker系统可以计算出翻译系统的翻译结果与

检测点中参考翻译之间的匹配程度，给出评测结果。

给定测试数据集上的检测点类c和翻译系统的翻译结果t，则t相对于c的匹配得分的计算公式如(1)所示：

$$Score(c) = Recall(c) \times Penalty \tag{1}$$

其中Recall函数计算c的参考翻译与t之间n-gram匹配的召回率计算公式如(2)所示；Penalty函数用于惩罚翻译结果中出现冗余n-gram的现象，它对整个测试数据集的参考翻译R和翻译结果T中句子平均长度的比值进行惩罚，计算公式如(3)所示：

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

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

在公式(2)中，R*是由检测点类c中每个检测点的最佳参考翻译r*构成的集合，其中每个检测点的最佳参考翻译r*由公式(4)确定；G(r)表示由检测点中参考翻译r形成的n-gram集合；Count函数表示含有n-gram的总数；Match函数表示n-gram在翻译结果t中出现的次数，即匹配数目；DM函数用来评估检测点中参考翻译的质量，以降低由于检测点抽取过程中词对齐结果质量不良所导致的检测点评测误差，该函数的计算公式如(5)所示，其中Dic(c)表示检测点c的源语言内容在双语词典中的翻译结果，CoCnt(x,y)表示x和y中相同词的个数，WordCnt(x)表示x中词的个数。

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

$$DM(r) = \begin{cases} Max\{0.1, \dfrac{CoCnt(r, Dic(c))}{WordCnt(r)}\} & \text{当}c\text{的参考翻译由词对齐结果获得时} \\ 1 & \text{其他请况} \end{cases} \tag{5}$$

基于以上评测方法，Woodpecker系统对于检测点的具体评测过程如下:

(1).准备翻译系统在测试集上的翻译结果。

(2).用户指定检测点类别集合C以及翻译任务类型(汉英或英汉)。

(3).对C中的每一个检测点，计算它的目标参考翻译内容与翻译系统翻译结果之间的n-gram匹配数量并进行归一化。

(4).对C中同一类型中所有检测点的得分求和得到该类检测点的匹配得分；对C中所有检测点的得分求和得到翻译系统的总体匹配得分。

(5).详细列出翻译系统不同语法层次上各种类型检测点的匹配信息以及得分情况。

Table 1: 汉英检测点类型列表

词级(Word level)		
歧义词(Ambiguous word)	新词(New word)	成语(Idiom)
名词(Noun)	动词(Verb)	形容词(Adjective)
代词(Pronoun)	副词(Adverb)	介词(Preposition)
数量词(Quantifier)	叠词(Repetitive word)	搭配(Collocation)
短语级(Phrase level)		
主谓短语 (Subject-predicate phrase)	谓宾短语 (Predicate-object phrase)	介宾短语 (Preposition-object phrase)
量词短语 (Measure phrase)	方位短语 (Location phrase)	……
句子级(Sentence level)		
把字句(BA sentence)	被字句(BEI sentence)	是字句(SHI sentence)
有字句(YOU sentence)	N/A	

Table 2: 英汉检测点类型列表

词级(Word level)		
名词 (Noun)	动词（含时态分类） Verb (with Tense)	情态动词 (Modal verb)
形容词(Adjective)	副词(Adverb)	Pronoun 代词
Preposition(介词)	歧义词(Ambiguous word)	复数(Plurality)
被动语态 (Possessive)	比较级最高级 (Comparative & Superlative degree)	……
短语级(Phrase level)		
名词短语 (Noun phrase)	动词短语 (Verb phrase)	形容词短语 (Adjective phrase)
副词短语 (Adverb phrase)	介词短语 (Preposition phrase)	……
句子级(Sentence level)		
由引导词定义的多种从句 (various kind of clauses defined by leading words)		

参考文献：

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.