ZZX_MT 系统 CWMT2013 评测报告*

巢文涵[1]，李舟军[1]

（1.北京航空航天大学，北京市 100191）

摘要：本文介绍了我们的机器翻译系统 ZZX_MT 参与 CWMT2013 的评测情况。本次评测中我们参与了三个子任务，分别是：藏-汉翻译、维-汉翻译及蒙-汉翻译。我们将简单介绍系统的基本组成及其参与 CWMT2013 各个子任务的翻译评测情况。

关键词：统计机器翻译；ZZX_MT；CWMT2013

中图分类号：TP391　　**文献标识码：**A

1 引言

本文将对我们的 ZZX_MT 机器翻译系统及其在 CWMT2013 中的评测情况进行详细描述。在本次评测中，ZZX_MT 系统参加了三项翻译评测，分别是藏-汉机器翻译评测、维-汉科技机器翻译评测。由于 ZZX_MT 参与了 CWMT2008[1]、CWMT2009[2]、CWMT2011[3]机器翻译评测任务，并在评测报告中给出了详细的系统介绍，本文将重点关注系统在 CWMT2013 评测中的实验设计及相应的结果。

论文结构如下：第 2 节简单介绍在评测中采用的 ZZX_MT 系统及其基本框架；第 3 介绍我们参与的三项机器翻译评测的实验设计情况，最后是结论。

2 ZZX_MT 系统概述

ZZX_MT 是一个基于对数线性（Log-Linear）模型的统计机器翻译系统，它主要关注于解决词对齐问题以及重定序问题。同时，我们试图结合各种语言学知识以提高翻译质量，在本次评测中，采用的句法知识主要是 Wu 提出的反向转录文法（Inversion Transduction Grammar，ITG）[4]，以及中文句子与英文句子的句法分析树。

ZZX_MT 主要由三部分组成：

a) 词对齐模块：以 Giza++的词对齐结果为输入，加入启发式规则，输出质量更高的词对齐语料库；

b) 模型训练模块：以词对齐语料库为输入，提取短语对及其相关信息，建立翻译模型 $P(e|c)$ 和 $P(c|e)$、$P_w(e|c)$ 和 $P_w(c|e)$，以及重定序模型；

c) 解码模块：输入源句子，使用上述的翻译模型、重定序模型以及语言模型等，搜索最佳的翻译结果。

在评测中，训练阶段的流程如下：

* **收稿日期：**　　**定稿日期：**

基金项目：

作者简介：巢文涵（1979—），男，博士、讲师，主要研究方向为机器翻译、文本挖掘；李舟军（1965—），男，博士、教授，主要研究方向信息安全、数据挖掘。

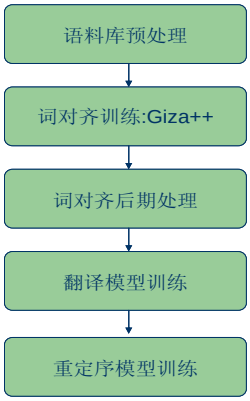

图 1 ZZX_MT 系统训练流程

其中词对齐后期处理是在合成两个方向的 Giza++词对齐之后，使得每个句对的词对齐均满足 ITG 约束，从而可以获得相应的重定序模型。

ZZX_MT 的 SMT 模型采用对数线性模型，其中包括以下特征（共 6 类 10 个特征）：

1. 翻译模型：$P(e|c)$，$P(c|e)$，$P_w(e|c)$ ，$P_w(c|e)$
2. 重定序模型：$r(o|b_k,b_{k+1})$
3. 语言模型：$p_{lm}(E)$
4. 词惩罚：I
5. 短语惩罚：K
6. 树的同构模型：$\Pr(T_G|T_1^K,T_C)$

我们把解码过程看作是 ITG 规则的应用序列，因此最终形成的目标句子会形成一棵 ITG 树，而树的同构模型则是为了调整该 ITG，使得其与源句子的句法分析树的结构基本相似。

由于模型中结合了 ITG 约束，因此，ZZX_MT 采用带有束搜索的 CYK 类型的解码器，可以在所有的 ITG 树空间内搜寻最佳的 ITG 树。

在本次评测中，由于我们参与的评测任务资源所限，无法对藏、维及蒙语进行句法分析，因此，只采用了前 5 种类型的 9 个特征。

3 实验

在 CWMT2013 中，训练及测试环境如下：

表 1. 训练及测试环境

	机器 1	机器 2
操作系统	windows xp (32bit) Cygwin (32bit)	windows 7 (64bit)
CPU 数量	4	4
CPU 类型 及其频率	Intel Core(TM)i3 M380 @2.53GH	Intel i5-3320m 2.60GHz
系统内存大小	4G	4G

其中， Cygwin 用于运行 Giza++获得初始词对齐。

3.1 藏-汉翻译

我们采用 CWMT2013 提供的语料作为训练集、开发集和测试集。其中，训练集是由基线系统提供的，并且所有的数据集都是采用了分词后的数据。

由于我们不熟悉藏文，为了完成评测，我们对语料进行了相应的预处理工作，主要包括：

- 中文全角处理；
- 分词：采用 ICTCLAS 分词系统进行中文分词，藏文分词直接采用评测方提供的分词数据；
- 数字化编码：由于不识别藏文，为了方便处理，对所有藏文进行了数字化处理。

最后得到的训练语料的句对数目为：109,356。

系统的语言模型采用 Sogou 单语语料库：

表 2. 语言模型所采用的训练语料库

序号	语料库描述	备注
1.	搜狗全网新闻语料库（SogouCA）	经过全角处理和分词后约 25M 个句子

我们采用 SRILM[5]训练语言模型，其中取 Order=3.

3.2 维-汉翻译

我们采用 CWMT2013 提供的语料作为训练集、开发集和测试集。为了完成评测，对语料进行了相应的预处理工作，主要包括：

- 中文全角处理；
- 分词：采用 ICTCLAS 分词系统进行中文分词；
- 维语形态分析：采用 MeCab-uyghur 对维语进行形态分析；
- 数字化编码：由于不识别维文，为了方便处理，对所有维文进行了数字化处理。

最后得到的训练语料的句对数目为：109,895。

系统的语言模型采用 Sogou 单语语料库，见表 2，利用 SRILM 训练语言模型，取 Order=3。

3.3 蒙-汉翻译

我们采用 CWMT2013 提供的语料作为训练集、开发集和测试集。为了完成评测，对语料进行了相应的预处理工作，主要包括：

- 去掉空行
- 中文全角处理；
- 分词：采用 ICTCLAS 分词系统进行中文分词；
- 数字化编码：由于不识别蒙文，为了方便处理，对所有蒙文进行了数字化处理。

最后得到的训练语料的句对数目为：106,467。

系统的语言模型采用 Sogou 单语语料库，见表 2，利用 SRILM 训练语言模型，取 Order=3。

3.4 实验结果及分析

利用以上处理后的数据，经过 ZZX_MT 的训练之后，利用评测方提供的测试数据进行测试，得到的最终结果如下：

表 3. ZZX_MT CWMT2013 评测任务结果

评测任务	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
藏-汉	0.1607	0.1646	0.1227	6.255	6.2572	0.6792	0.7187	0.4374	0.217	0.4391	0.668
维-汉	0.2656	0.2663	0.215	8.147	8.156	0.782	0.6425	0.4282	0.223	0.3967	0.702
蒙-汉	0.0822	0.0951	0.0634	4.692	4.6926	0.5323	0.7386	0.6105	0.368	0.2422	0.806

从以上评测结果来看，系统对于各语种的翻译结果均不理想，可能的原因是：

1. 训练语料相对较小，各数据集均在 10 万左右；
2. 各源语言均具有丰富的形态变化，但是评测中除维语外，均未作形态分析；
3. 语言模型的利用：本次直接采用 Sogou 语料库，未利用训练集的单语语料训练语言模型。理论上，由于 Sogou 语料库较大，不会有太大影响，但是具体情况需要进一步实验分析。

另一个较为有趣的现象是，这三种语言在数据集规模相似，基本处理相似的情况下，翻译的效果相差较大，尤其是“藏-汉”语料属于政府数据，理论上应该会高于其他两种语种的翻译结果，但是结果显示“维-汉”翻译的效果最好。其中的原因可能是：

1. 各源语言差距较大：此次蒙语的翻译效果最差，可能是蒙语的形态变化更为复杂，也有可能是“蒙-汉”评测训练集和“测试集”的分布相差较大。
2. 维语进行了一定的形态分析，可能在一定程度上提高了翻译质量。

除了以上自动评测的结果，我们还参与了维-汉人工评测，结果如下：

表 4 CWMT2013 维汉新闻人工评测结果

参评系统	忠实度平均值	流利度平均值	总体平均值	评价句子总数
buaa	2.2486658	2.7059994	2.4773326	500

4 总结

本次评测是我们对资源稀缺、形态复杂语言进行机器翻译的一次尝试。我们利用 ZZX_MT 系统参与了 CWMT2013 的藏-汉、维-汉及蒙-汉机器翻译评测任务。从评测结果来看，效果不甚理想，其中主要的问题在于对各源语言不熟悉，无法对其进行有效的分析。

另外，从三种源语言横向对比分析，三种源语言的翻译效果质量相差较大，这说明以上三种语言的性质差别比较大，其中维语在进行一定的形态分析之后，效果有了较大提升。

由于条件限制，本次评测每种语言我们只完成了一个实验，接下来我们将：1）完成其他对比实验，尤其是在形态处理、语言模型、翻译模型等方面进行进一步实验；2）加深对各语言的理解，进行有针对性的处理。

参考文献

［1］巢文涵,李舟军. ZZX_MT 系统评测报告. 机器翻译研究进展——第四届全国机器翻译研讨会论文集, pp.93-100. 2008.

［2］巢文涵,李舟军. ZZX_MT 系统 CWMT09 评测报告. 机器翻译研究进展——第五届全国机器翻译研讨会论文集, pp.116-120. 2009.

［3］巢文涵;李舟军. ZZX_MT 系统 CWMT2011 评测报告[A];机器翻译研究进展——第七届全国机器翻译研讨会论文集[C]. 2011.

［4］Dekai Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics, 23(3):374. 1997.

［5］A. Stolcke. SRILM – An extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, pp. 901–904. 2002.

作者联系方式：巢文涵　北京学院路 37#北京航空航天大学计算机学院 100191

电话 13699223750　电子邮箱 chaowenhan@buaa.edu.cn