西藏大学 CWMT'2015 技术报告

尼玛扎西[1]、拥措[1]、仁增旺堆[1]
(1.西藏大学 信息化研究所,西藏拉萨，850012)

摘要：本文简要介绍了西藏大学信息化研究所参加第 11 届全国机器翻译研讨会机器翻译评测的情况。本单位参加了藏汉领域的机器翻译评测项目。本文简述了本单位机器翻译系统的实现框架以及实施细节。

关键词：机器翻译；自然语言处理；藏文分词

Technical Report of MT System of Tibet University for CWMT'2015

Nyimatrashi[1],Yongtso[1],Renzengwangdui[1]
（1.Information Research Institute,Tibet University,Lhasa,Tibet,850012）

Abstract: This is an overview of the Tibetan-Chinese MT system of Tibet University Information Research Institute for the 11th China workshop on machine translation. Tibet University Information Research Institute participated in the Tibetan-to-Chinese machine translation task. This paper describes the framework and technical details of our machine translation system.

Keywords: machine translation; natural language processing; Tibetan word segmentation

1 引言

西藏大学信息化研究所参加了 2015 年第11届全国机器翻译研讨会(CWMT' 2015)组织机器翻译评测活动。在所有的评测项目中，我们只参加了藏汉机器翻译项目。在这个项目的评测中，我们采用了经典的基于短语的统计机器翻译。

本文的结构安排如下：第二节给出西藏大学信息化研究所阳光藏汉机器翻译系统的总体框架和实现细节；第三节介绍数据的使用、处理和评测结果；最后在第四节给出结论和展望。

2 系统描述

我们提交的藏汉翻译结果是基于经典的短语统计机器翻译。系统运行的软硬件参数如下：

Windows 7 64位

CPU 8

Intel(R)Core （ TM ） i7-4790 CPU @3.60GHz

Memory 32G

系统运行时间约 125.92 秒；

训练语料和开发集均为 CWMT2015 组织方提供的语料；

语言模型训练语料为由《搜狗全网新闻语料库》08 版和训练语料中汉语部分训练而成的 5 元统计模型；

藏文分词系统采用西藏大学阳光藏文分词系统，汉语分词采用厦门大学开发的分词系统 segtag。

采用基于短语的统计机器翻译模型，子模型参数手动设定，没有采用开发集；其中调序模型采用厦门大学提出的一种调序模型,可较好地解决藏汉文之间因动词语序的不同而产生的句法差异。提出的调序需要用到（待调序的短语和作参考的短语）两组短语的所有可用信息，包括词汇本身、词性、甚至其他的词法语义信息。另外这两组短语可以是相邻的，也可以是不相邻的。调序操作也做了扩充，包括 M, S, L, R, F, I 等，分别表示单调、交换、左不连续、右不连续、句末、句首等几种情况。由如下公式表示：

$$p(o|e,f)=\prod_{i=1}^{k} p(o|<\bar{e}_{a_{i-n}},\overline{f_{i-n}}>,<\bar{e}_{a_i},\bar{f}_i>)$$

采用的非开源软件有：厦门大学的词组抽取工具,这是一种从 giza 训练的结果中抽取短语概率表的工具。

采用的开源软件有：语言模型训练工具 SRILM，语言模型计算工具 kenlm，词语对齐

工具 GIZA++等；

要说明的是，我们的系统并没有采用多线程。

2．1 总体框架

给定源语言句子 *f*，基于对数显示模型的统计机器翻译过程就是搜索具有最大概率的目标翻译 *e*：

$$e^* = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f)$$

其中，$h_m(e, f)$为特征函数，λ_m为特征权重。在基于短语的模型中，使用的特征函数有：

1： 短语翻译概率（2 个方向）；
2： 词翻译概率（2 个方向）；
3： 语言模型；
4： 基于源短语的重排序；
5： 长度惩罚；

一共 7 个特征。解码的搜索策略为柱状搜索算法，最后生成 1-Best 结果。目前，每个特征函数的权重由经验设定。

2.2 实现细节

采用 GIZA++工具包进行训练，采用 Grow-Diag-Final 式启发函数进行词对齐扩展生成短语。构建短语表时，融合了 GIZA++的单词翻译概率，以避免某些单词的翻译不在短语表中的情况。

调序模型采用厦门大学提出的基于源短语的重排序模型，用以处理藏文语序和汉语差异的情况，但是尚不能很好处理长距离调序的情形。

3 数据的使用、处理和评测结果

训练语料采用了发布的藏汉语料，但是对过长的句子进行了过滤。语言模型训练数据采用了训练语料的中文部分和搜狗语料，采用 srilm[2]工具训练为 5 元模型。

对中文数据没有进行任何处理，仅采用厦门大学的 Segtag 系统进行中文分词；对藏文数据没有进行任何处理，仅采用阳光藏文分词系统进行藏文分词。词对齐工具采用了 GIZA++，并对该对齐结果进行扩展对齐（grow-diag-final）[1]。

整个训练过程约为 2 小时。

整个测试语料翻译约 125.92 秒。对藏文未登录词的翻译采用了删除处理。系统评测的成绩为 BLEU 0.80，在 4 个参赛主系统中排名第三。

4 结论

西藏大学信息化研究所参加了 CWMT’2015 藏汉机器翻译评测项目，效果一般。分析以及整个参评过程，我们有 2 点经验：

1）数据的预处理问题。我们后来发现训练数据中有很多错误。没有提前对训练数据进行处理，浪费了宝贵时间。在提交最终翻译结果后，我们利用自己研发的“基于藏文拼写形式语言的拼写检查软件”对藏文训练语料进行了拼写检查预处理，纠正了拼写错误。使用预处理过的训练数据之后，翻译效果有所提高。

2）机器翻译系统还没得到优化。不管是分词系统，还是各个子模型的参数，都可以得到优化，从而提高成绩。

致谢

本论文受西藏自治区重点科研项目《基于短语的藏汉统计机器翻译关键技术研究》和国家自然科学基金(61262086)资助。

参考文献

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation[C]. In Proceedings of the Human Language Technology conference / North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003), 2003, pages 127-133.

[2]Andreas Stolcke, 2002. SRILM-An extensible language modeling toolkit. In Proceedings of International Conference on spoken language processing, volumn 2, pages 901-904.