第七届机器翻译研讨会厦门大学技术报告

胡金铭，甘兴超，陈毅东，史晓东
厦门大学信息科学与技术学院智能科学与技术系 厦门 361005
hjm946637@gmail.com

摘要：本文主要介绍了厦门大学参加2011年第七届全国机器翻译研讨会（CWMT2011）评测的情况。本单位参加了全部9个评测项目中的3个子项：汉英新闻领域机器翻译，英汉新闻领域机器翻译和藏汉政府文献机器翻译三个评测项目。文章将简要介绍本单位统计机器翻译系统的实现框架、模型以及阐述它们在评测数据上的性能表现，并针对结果加以适当分析。
关键字：人工智能、自然语言处理、基于短语的统计机器翻译

XMU Technical Report for the 7th China Workshop on Machine Translation

Jinming Hu,Xingchao Gan, Yidong Chen,Xiaodong Shi
Department of Cognitive Science, School of Information Sciences and Technologies, Xiamen University, Xiamen 361005
hjm946637@gmail.com

Abstract: This is an overview of XMU technical report for the 7th China Workshop on Machine Translation.XMU participated in three machine translation evaluation tasks: ZH-EN-NEWS, EN-ZH-NEWS, TI-ZH-GOVE. This paper will briefly introduce the implement framework and models of our statistical machine translation systems, present and analyze the experimental results over the evaluation data.
Keywords:artificial intelligence, natural language processing, phrase-based statistical machine translation

1 引言

作为参评单位之一，厦门大学参加了三个评测任务：汉英新闻领域机器翻译任务；英汉新闻领域机器翻译任务；藏汉政府文献机器翻译任务。在本次评测中利用开源工具Moses来构建基于短语[1]和基于层次短语[2]的统计机器翻译系统。对于这三个子项，我们分别提交了汉英新闻领域（3个系统），英汉新闻领域（4个系统）和藏汉政府文献（4个系统）。

在后文中第二部分会简要的介绍各个参评系统的描述及原理，第三部分会介绍实验中前期数据的处理和整个实验的流程。最后会给出评测系统的结果，并进行若干分析，提出本次实验的诸多问题以及影响评测结果的几个因素。

2 系统介绍

由上文可知，XMU 在本次评测的三个机器翻译子项中，一共提交了 11 个结果，每一个系统所采用的翻译模型可由系统的命名清晰的看出。如在汉英新闻领域机器翻译任务当中，XMU 提交的的翻译结果来自下述三个系统的输出：

zh_en_news_trans-xmu-primary-moses_phrase1;

zh_en_news_trans-xmu-contrast-moses_phrase2;

zh_en_news_trans-xmu-contrast-xmu_hpb，

前两个就是使用开源工具包 moses 中的短语模型，而后一个则是使用了层次短语模型。类似的英汉和藏汉当中的翻译模型也可由命名清晰可见。

2.1 短语模型

短语模型使用开源的工具包 moses 中的基本的短语模型，Moses 是当前比较流行也很稳定的基于短语的统计机器翻译系统。该系统利用 log-linear 模型将多个翻译特征融合，它采用了 MSD (Monotone, Swap, Discontinuous) 词汇化的调序模型。其中使用 msd-bidirectional-fe[3]的调序模型，此调序模型可解释如下：

1. 调序类别 MSD，定义当前源语言 f 的定义当前源语言短语 f 的翻译结果 e 与前一个源语言短语 f' 的翻译结果 e' 为如下三种关系：
   a) Monotone 即翻译结果连续，且顺序与源语言中的顺序相同（e，e' 相邻，且 e' 在前）；
   b) Swap 即翻译结果连续，但顺序与源语言中的顺序相反（e，e' 相邻，且 e 在前）；
   c) Discontinuous 即翻译结果不联系（e，e' 不相邻）。
2. 双向调序模型（Bidirectional）

   同时考虑两个调序模型，即由前一个短语决定的当前短语的调序模型和由当前短语决定的下一个短语的调序模型。
3. 词汇化调序条件（fe）

   调序模型的概率同时由目标语言短语和源语言短语决定。

2.2 层次短语模型

层次短语模型使用开源的工具包 moses 中的层次短语模型。层次短语，顾名思义，就是短语本身包含了更小的短语，从而利用层次短语本身的规则来解决短语之间的长距离调序问题。该系统同样也是利用 log-linear 模型将多个翻译特征融合，下面简单介绍一下规则抽取过程：

1. 短语抽取是从具有词对齐的双语语料开始的，可以用三元组 <f,e,~>的集合表示，这里 f 表示源语言句子，e 表示目标语言句子，~是词在源语言中的位置与词在目标语言中的位置的多对多的二元关系。对于双语语料的词对齐，本文使用 GIZA++跑出的对齐。层次短语模型的抽取方法比短语模型的抽取方法更复杂一些，层次短语的抽取过程可以分为两步：
   a) 使用短语模型的方法抽取出初始短语对，一个初始短语对要满足至少有一个词对齐，并且两个短语之间的所有词对齐不能超过两个短语的范围；

b) 为了从短语中获取规则，我们找到那些包含其他初始短语的短语，将被包含的子短语用非终结符号代替；

2. 层次短语模型只要在短语模型训练的脚本上加入三个选项：
   a) -hierarchical 表示使用层次短语模型
   b) -glue-grammar 表示产生层次短语模型的胶水规则
   c) -max-phrase-length 表示抽取规则时对短语长度的限制

2.3 子项中系统差异

在三项评测任务中，我们提交的系统中都使用了短语模型和层次短语模型。

1. 汉英新闻领域机器翻译任务：
   a) zh_en_news_trans-xmu-primary-moses_phrase1
      zh_en_news_trans-xmu-contrast-moses_phrase2
      i. 语言模型：phrase1 中语言模型使用 Srilm 开源工具包利用所给语料 1（见后文数据部分）生成四元语言模型，由于语料较小，所以使用 wbdiscount 即 Witten-Bell 折扣算法，而 phrase2 中语言模型使用 Srilm 开源工具包利用语料 2（见后文数据部分）生成五元语言模型，由于语料较大，所以使用 kndiscount 即 Kneser-Ney 折扣算法，该参数对应于 modified Kneser-Ney 折扣。
      ii. 翻译模型：利用训练集 1（见后文数据部分），使用 moses 开源工具中的短语模型。
      iii. 解码：也是 moses 中的，默认的解码器是使用 beam-search 算法进行搜索获取翻译结果。
   b) zh_en_news_trans-xmu-contrast-xmu_hpb
      i. 语言模型：使用 Srilm 开源工具包利用语料 2 生成五元语言模型，使用 kndiscount 即 Kneser-Ney 折扣算法，该参数对应于 modified Kneser-Ney 折扣
      ii. 翻译模型：利用训练集 1，使用的是层次短语模型
      iii. 解码：该系统使用了带有 cube pruning 的改进的 cky 算法来进行解码

2. 英汉新闻领域机器翻译任务和藏汉政府文献机器翻译任务：

   因为这两个任务所使用的模型类似，而英汉中 current 和 progress 所提交的结果使用的模型是相同的，所以在此只介绍 progress 的系统，所以，在这里一并介绍：
   a) en_zh_news_progress_trans-xmu-contrast-moses_phrase1
      en_zh_news_progress_trans-xmu-contrast-moses_phrase2
      ti_zh_gove_trans-xmu-contrast-moses_phrase1
      i. 此处的两个英汉短语系统，和汉英中的类似，phrase1 和 phrase2 的特点都与汉英中的 phrase1 和 phrase2 相同。在此不再赘述。
      ii. 藏汉的短语系统使用 Srilm 工具利用语料 4（见后文数据部分）生成的五元语言模型，使用 wbdiscount 即 Witten-Bell 折扣算法。翻译模型使用的训练集是训练集 3（见后文数据部分）。
   b) en_zh_news_progress_trans-xmu-primary-moses_hier1
      en_zh_news_progress_trans-xmu-contrast-moses_hier2
      ti_zh_gove_trans-xmu-primary-moses_hier1
      ti_zh_gove_trans-xmu-contrast-moses_hier2
      ti_zh_gove_trans-xmu-contrast-moses_hier3
      i. 语言模型：英汉的两个系统与英汉的 phrase 相对应的使用的同一语言模型。藏汉中 hier1 使用 Srilm 工具利用语料 4 用 wbdiscount 生成的五元语言模型，hier2

使用 Srilm 开源工具包利用语料 3（见后文数据部分）用 kndiscount 生成五元语言模型，hie3 使用 Srilm 开源工具包利用语料 4 用 kndiscount 生成五元语言模型。

ii. 翻译模型：开源工具 moses 的基于层次短语模型，获取泛化规则集，其中的参数设置都利用 moses 工具包的默认设置。英汉所使用的训练集是训练集 1，藏汉中 hier1 和 hier2 使用的训练集是训练集 3，hier3 使用的训练集是训练集 2

iii. 解码：解码算法使用 moses 自身的 chart-decoder，即 cky 解码算法，这可以处理任意的对于终结或非终结符数量不受限的上下文无关文法。

3 实验介绍

3.1 操作系统性能

CPU： Intel（R）-Xeon（R）11 个 2.93GHZ
内存：96GB
操作系统：Fedora14-64bit

3.2 机器翻译评测

3.2.1 数据准备

分为两部分，翻译模型的训练集和语言模型的训练集

1. 翻译模型训练集：

a) 训练集 1（英汉/汉英新闻）：

利用官方发布的平行双语语料，构建翻译系统的训练集。首先对语料进行筛选，选择的语料包括：

CLDC-LAC-2003-004.zip
Datum English-Chinese Parallel Corpus (Part).zip
HIT-IR English-Chinese Sentence-Aligned Corpus.zip
HIT-MT English-Chinese Sentence-Aligned Corpus.zip
HTRDP("863") 2003 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data).zip
HTRDP("863") 2004 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data).zip
HTRDP("863") 2005 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data).zip
ICT Web Chinese-English Parallel Corpus（Version 2011）.zip
NEU Chinese-English Parallel Corpus.zip
PKU Chinese-EnglishChinese-Japanese parallel corpora (Chinese-English part).zip
SSMT2007 Machine Translation Evaluation Data.zip

在这里去除掉了电影字幕语料，上述选中语料进行人工的筛选，筛选原则大致如下：

i. 对平行语料的对齐问题进行排查，每一千句抽取，再层叠抽取，避免大规模的对齐错误。

ii. 对一些直接从网络中抓取的语料如 ICT Web Chinese-English Parallel Corpus（Version 2011）.zip 进行人工校验，对其中的质量较差，文不成句或者对应翻译并不符合常规的进行大规模删减。

iii. 因为是参与新闻类的翻译评测，所以对内容也进行选择，如 HTRDP("863") 2003-2005 中都有对话的语料，与新闻关系不大，不进入训练集的挑选。还有一些关于文学艺术等的内容也进行了一定程度上的删除。

最终，我们筛选出了符合这几条原则的汉英/英汉新闻的翻译模型训练集，筛选结果如表 1 所示

语料名称	原大小	筛选大小
新闻训练语料 ICT Web Chinese-English Parallel Corpus（Version 2011）	817MB	567MB
新闻训练语料 Datum English-Chinese Parallel Corpus (Part)	279MB	279MB
HTRDP("863") 2003 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data)	0.97MB	472KB
HTRDP("863") 2004 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data)	1.18MB	663KB
HTRDP("863") 2005 Machine Translation Evaluation Data (The part of Chinese-English & English-Chinese MT data)	1.46MB	602KB
HIT-MT English-Chinese Sentence-Aligned Corpus.zip	4.69MB	1.90MB
SSMT2007 Machine Translation Evaluation Data.zip	1.64MB	1.1MB
PKU Chinese-EnglishChinese-Japanese parallel corpora (Chinese-English part).zip	29.63MB	10.1MB
CLDC-LAC-2003-004.zip	44.8MB	13.2MB

表 1

在选取语料之后，就要对平行句对进行处理，首先就是分词。对汉语的分词使用 ICTCLAS 的开源分词系统，英文分词使用 tokenizer.perl 脚本。其次是过滤，利用已有的平行句对过滤工具对筛选出的分好词的语料进行过滤，主要过滤较长英文单词，非法字符，中英文句子长度差过大等等。并使用 clean-corpus-n.perl 来进行长短语句过滤。再将其中的英文部分小写化，这样就生成了最后真实的训练集：train.ch 242.2MB train.en 238.1MB。平行句对是 2541026 行，这就是汉英/英汉新闻领域的翻译模型训练集，训练集 1

b) 训练集 2（藏汉政府文献）

使用官方提供的藏汉训练集，和英汉/汉英不同，对于藏汉的语料，由于对于藏语的不熟悉，所以并未做人工的筛选工作。其中使用史晓东老师的藏语分词系统对未切分的藏语语料进行分词，汉语分词仍然使用 ICTCLAS 的开源分词工具。使用 clean-corpus-n.perl 来进行长短语句过滤。得到的训练集：ict.ti 20.3MB ict.ch 5.5MB 101626 行平行句对

c) 训练集 3（藏汉政府文献）

使用官方提供的藏汉训练集，使用中科院分词的藏语语料，汉语部分使用 ICTCLAS 的开源分词工具。在用 clean-corpus-n.perl 过滤之后得到的训练集：shi.ti 19.6MB shi.ch 5.3MB 100698 行平行句对。

2. 语言模型训练集：

语言模型分值的大小在一定程度上反映译文的流畅程度，所以语言模型常常在翻译过程中起重要作用。但是，由于经筛选后得到的语料过小，生成的语言模型并不能很好的。那么扩大语言模型的训练规模、提高语言模型的阶数，是改善统计机器翻译系统的一个重要做法。因此，我们选择了更大的语料来训练语言模型。

a) 语料 1（英汉/汉英新闻）：

使用 srilm 开源工具利用上文中得到的 train.ch train.en 分别训练出英文和中文的 4 元语言模型。

b) 语料 2（汉英新闻）：

选用 giga-word3 中的新华部分约 1.6GB 的英语语料，使用 srilm 开源工具训练出英文的五元语言模型。

c) 语料 3（英汉新闻，藏汉政府文献）

同上利用 giga-word3 中的新华部分约 1.4Gb 的汉语语料，使用 srilm 开源工具训练出中文的五元语言模型。

d) 语料 4（藏汉政府文献）：

选用官方发布的训练语料中的汉语语料，使用 srilm 开源工具训练出汉语的五元语言模型。

Giga-word3 新华部分语料具体如表 2 所示：

语料名称	行数	大小
xinhua_lowecase.tok	9685593	1.4GB
Xinhua_chn	9789421	1.6GB

表 2

3.2.2 预处理

事实上在 3.2.1 中，已经提到了预处理的问题，在这里系统的总括一下。

1. 对于英汉/汉英的预处理：
   a) 中文分词，使用 ICTCLAS2011 版进行中文分词。使用 toknizer.perl，对于标点符号的分离处理。
   b) 过滤与对应英文句子长度差较大的句子
   c) 过滤英文中过长单词
   d) 英文大写转换成小写
   e) 过滤非法字符
   f) 过滤过长或较短句子
2. 对于藏语的预处理：
   a) 使用中科院分词语料
   b) 使用史晓东老师的分词系统做藏语分词处理。

3.2.3 翻译模型训练

本次参评的系统主要只有两个，一个是短语系统，一个是层次短语系统。由于使用开源工具包 moses 来运行两个系统，所以短语表和规则集都是通过 moses 来进行训练得来的。其中的参数设置都采用 moses 的默认设置，汉英的 hpb 系统是依据参考文献 2 所实现的层次短

语系统。

3.2.4 调整参数

开发集的获取

评测任务	英汉	汉英	藏汉
开发集规模	1000 个英语句子，4 个参考答案	1006 个汉语句子，4 个参考答案	650 个藏语句子，4 个参考答案

表 3

对于调参的过程，使用 moses 中的调参过程，将开发集与测试集合并过滤短语表或规则集，再使用开发集进行调参。

3.2.5 翻译解码

两类系统：短语系统和层次短语系统

短语系统使用 beam-search 算法进行解码。

层次短语使用 cky 解码。

汉英的 hpb 使用修改过的 cky 算法，即带有 cube pruning 的算法

3.2.6 后处理

删除汉英译文中的未翻译词汇，删除部分英汉译文中的未翻译词汇。

3.2.7 评测结果

汉英新闻评测结果

ZH-EN-NEWS (progress)			
评测系统	BLEU4-SBP	BLEU4	NIST5
zh_en_news_trans-xmu-primary-moses_phrase1.result	0.1579	0.1699	6.4643
zh_en_news_trans-xmu-contrast-moses_phrase2.result	0.1819	0.196	6.7755
zh_en_news_trans-xmu-contrast-xmu_hpb.result	0.1701	0.1778	5.985

表 4

英汉新闻评测结果（progress）

EN-ZH-NEWS (progress)			
评测系统	BLEU5-SBP	BLEU5	BLEU6
en_zh_news_progress_trans-xmu-primary-moses_hier1.result	0.3033	0.3223	0.2569
en_zh_news_progress_trans-xmu-contrast-moses_hier2.result	0.3076	0.3216	0.2583
en_zh_news_progress_trans-xmu-contrast-moses_phrase1.result	0.3043	0.322	0.2569

en_zh_news_progress_trans-xmu-contrast-moses_phrase2.result	0.3024	0.3144	0.2526

表 5

英汉新闻评测结果（current）

EN-ZH-NEWS (current)			
评测系统	BLEU5-SBP	BLEU5	BLEU6
en_zh_news_current_trans-xmu-primary-moses_hier1.result	0.289	0.3037	0.2432
en_zh_news_current_trans-xmu-contrast-moses_hier2.result	0.2974	0.3086	0.2486
en_zh_news_current_trans-xmu-contrast-moses_phrase2.result	0.2911	0.3007	0.242
en_zh_news_current_trans-xmu-contrast-moses_phrase1.result	0.2895	0.3026	0.2424

表 6

藏汉政府文献评测结果

TI-ZH-GOVE			
评测系统	BLEU5-SBP	BLEU5	BLEU6
ti_zh_gove_trans-xmu-primary-moses_hier1.result	0.3917	0.4031	0.3645
ti_zh_gove_trans-xmu-contrast-moses_hier2.result	0.4027	0.4162	0.3753
ti_zh_gove_trans-xmu-contrast-moses_hier3.result	0.4162	0.4279	0.3898
ti_zh_gove_trans-xmu-contrast-moses_phrase1.result	0.3713	0.3839	0.3438

表 7

3.2.8 分析及遇到的问题

通过基本系统和对比系统的对比结果发现几点问题：

1. 当语言模型使用大语料的时候，更多元的时候，翻译的结果会有明显的提高。
2. 层次短语模型对于相应任务的短语模型，评测结果也都有相应的提高。
3. 藏汉可以看到，使用中科院的分词的藏语语料时结果要好于史晓东老师分词的语料，但是在使用开发集评测的时候，史晓东老师分词的语料的结果要好于中科院分词的语料的结果。在提交结果后，我们又重新做过实验。如表 8 所示，这事在开发集上的情况。

TI-ZH-GOVE-dev-cmd			
评测系统	BLEU5-SBP	BLEU5	BLEU6
dev-cmd-hie.out.xml	0.4292	0.4432	0.4039
dev-cmd-hie-xinhua.out.xml	0.4438	0.4562	0.4158
dev-cmd-phrase.out.xml	0.3878	0.3974	0.3556
dev-cmd-phrase-xinhua.out.xml	0.3973	0.4064	0.3645
dev-ict-hie.out.xml	0.4203	0.4385	0.3984
dev-ict-phrase.out.xml	0.4050	0.4203	0.3784

表 8

由表中结果可见，dev-cmd-hie.out.xml 和 dev-ict-hie.out.xml 这两个的结果，相对于表 7 中的结果是 ti_zh_gove_trans-xmu-primary-moses_hier1.result 和 ti_zh_gove_trans-xmu-contrast-moses_hier3.result 可以发现，使用史老师分词的语料，在开发集上的结果要好于中科院分

词的语料，但是在测试集上的表现，却结果相反

4. 在使用 moses 做汉英，藏汉层次短语模型的时候，过滤规则集的时候，一直都出现过滤掉的规则过大。如汉英训练翻译模型的规则集有 40GB，但过滤出来的只有 1.5GB，藏汉也有类似问题。由于藏汉语料很小，所以索性没有进行过滤来进行调参和解码。但汉英由于语料较大，而做出的成绩非常差，所以就没有提交相应结果。但同样地问题并没有出现在英汉的层次短语模型上。英汉训练翻译模型的规则集有 38GB,过滤得到的规则集也有 8GB 左右。目前这一问题的原因还未查明。

4、总结

1. 语料筛选。由于经验不足，对语料的筛选工作并不是非常重视，所以前期的语料处理工作做得并不完善。对于不符合新闻类的语料没有完全剔除，这也导致了对最终的评测结果的影响。
2. 翻译模型。本次的参评系统只是利用开源工具包 moses 中的短语模型和层次短语模型，并未对已有模型进行改进。这是本次参评的最大不足之处。
3. 由于在模型改进方面所作工作较少，我们今年所使用的技术相对比较陈旧，因此评测成绩不理想。

参考文献

[1] Koehn, Philipp, Franz Josef Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. HLT-NAACL'2003

[2] David Chiang.Hierarchical Phrase-Based Translation.Computational Linguistics, 2007, 33(2): 201-228

[3] Koehn, Phillip, Franz Josef Och, and David Marcu. 2003. Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, 127-133.