CWMT2011 哈工大机器智能与翻译研究室技术报告

梁华参 刘乐茂 于墨 刘宇鹏 胡鹏龙 李婷婷 张春越 曹海龙 赵铁军
哈尔滨工业大学 机器智能与翻译研究室 哈尔滨 150001
E-mail: {lmliu, tjzhao}@mtlab.hit.edu.cn

摘 要: 本文详细介绍了哈尔滨工业大学机器智能与翻译研究室（HIT-MITLA）参加 2011 年全国机器翻译研讨会(CWMT2011)翻译评测任务的情况。在本次评测中，HIT-MITLAB 共使用了 5 个翻译系统,它们分别是 DINO, MUSES, KIMI, MOSES 和 MOSES_H, 它们共参与了评测中 5 个子项目—汉英、英汉新闻领域，英汉科技领域，日汉新闻领域和藏汉政府文献的翻译。在这 5 个子项目中 HIT-MITLAB 提交了共 18 个翻译结果。本文对参评的各个系统及其配置，数据的使用和处理进行了全面的描述，同时对各翻译的效果进行了比较和分析。

关键字: 机器翻译、短语翻译、层次短语翻译

HIT-MITLAB Technical Report for the 2011 China Workshop on Machine Translation

Huashen Liang, Lemao Liu, Mo Yu, Yupeng Liu, Penglong Hu, Tingting Li, Chunyue Zhang, Hailong Cao, Tiejun Zhao

Machine Intelligence and Translation Laboratory, Harbin Institute of Technology,

Harbin 150001, P. R. China

E-mail: {hsliang,lmliu,gflfof,ypliu,plhu,ttli,cyzhang,hailong,tjzhao}@mtlab.hit.edu.cn

Abstract: *This paper describes the details of machine translation and evaluation task for HIT-MITLAB's joining in the China workshop on Machine Translation in 2011(CWMT2011). In this task, we submit 5 translation systems, which are DINO, MUSES, KIMI, MOSES and MOSES_H respectively, and take part in 5 translation sub-tasks, which contain the translation domains of Chinese-to-English and English-to-Chinese news, English-to-Chinese science and technology, Japanese-to-Chinese news and Tibetan-to-Chinese government reference. HIT-MITLAB submitted 18 translation results in the sub-tasks. This paper gives a detailed description of the translation systems, their configurations, the used data and its processing methods, and gives the comparisons and analysis of the translation results for all the 5 systems.*

Keywords: *Machine translation, phrase-based translation, hierarchical phrase-based translation*

1 引言

哈工大机器智能与翻译研究室参加了 2011 年全国机器翻译评测（CWMT2011）中 5 个子项目：英汉新闻领域翻译评测；汉英新闻领域翻译评测；英汉科技领域翻译评测；日汉新闻领域翻译评测；藏汉政府文献翻译评测。本文对 HIT-MITLAB 所有 5 个参评系统、相关数据和实验结果进行详细介绍。

2 机器翻译项目参评系统介绍

本次评测中，共使用了 5 个不同的统计机器翻译系统——DINO，MUSES，KIMI，MOSES 和 MOSES_H，其中两个是基于短语的翻译模型，另外三个是基于形式语法的翻译模型。所

有的翻译模型都是由各自翻译规则的一组特征函数经过 log-linear 模型组合而成。这些系统参加了共 5 个项目的评测，各项目中使用的主系统和对比系统如表 1 所示。下面分别介绍这 5 个系统。

2.1 DINO 系统

DINO 系统是一个基于短语模型(koehn et al., 2003)的机器翻译系统。它的最小翻译单元为短语，即连续的词序列。给定一个源语言的句子，它的翻译过程可以描述成：将句子切分成一系列的短语，然后对每个短语进行翻译，最后对短语的翻译进行调序并输出候选翻译。该系统的翻译模型包含共 14 特征：词汇化调序特征(msd-bidirectional-fe)，语言模型特征，短语表特征，词惩罚特征。它采用 beam-search 的栈式搜索方法进行解码，栈的数量为源语言句子长度加上 1，每一个栈用来存储覆盖同样数量的源语言单词的假设，第一个栈中只有一个假设，表示当前覆盖了 0 个源语言单词，最后一个栈中存储覆盖了所有源语言单词的假设，即候选翻译结果。栈式搜索算法通过 beam-size 来控制栈中存储的翻译假设数量，以此来保证搜索的速度。在 DINO 系统提交的结果中，beam-size 设为 100。系统采用 lazy-pruning 的策略，当栈中的假设数量达到 200 时，保留前 100 个得分较高的假设（也即 k-best 解码时 k-best-size = 100），删除得分较低的 100 个假设。

2.2 MUSES 系统

本系统使用基于形式语法——括号的转录文法 BTG（Xiong et al.,2006）。BTG 在解释翻译现象时遵循一个重要的假设：同步规则源语言和目标语断的语序只存在两种可能性：正序或者逆序。正序是指源语言和目标语言的语序完全一致；逆序则是指它们的语序恰恰相反。同 DINO 系统一样，本系统的最小翻译单元也是短语；但在调序时，它将短语之间的调序限制在相邻的短语之间。训练调序模型时，它从双语训练语料中抽取了正序和反序的相邻短语对，并采用最大熵分类器来训练之。本系统采用开源工具包(Zhang,2006)来训练最大熵分类器，在训练的参数设置上，使用了 L-BFGS 算法选项并选择了高斯先验来抑制模型的过拟合现象。为了减少最大熵的训练空间，本系统仅仅对训练数据(汉英/英汉新闻语料)的前 20 万句抽取调序模型训练实例。模型使用的特征包含基本的短语特征（即翻译模型的特征），正序/反序的调序模型特征，词惩罚，短语惩罚和语言模型。解码上，它采用标准的 CKY 的柱搜索算法，为了加速解码过程，本系统将 Beam-size 设为 30, k-best 解码时的 k-bese-list-size 也设为 30。

2.3 KIMI 系统

本系统实现的是层次短语的翻译模型(Chiang 2005)，它基于另外一种形式文法同步上下文无关文法。同步上下文无关文法由一些层次短语翻译规则组成，这些层次短语规则可以从双语对齐句对中抽取。为了对系统性能和效率进行平衡，在规则抽取中，限制每条层次短语规则在句子中最多覆盖 10 个词，且抽取的层次短语规则其源语言端包含的终结符和非终结符数不超过 5 个。系统抽取出的规则均为二元规则，即每条规则只能生成不多于两个的非终结符。

该系统使用了如下 8 个特征：翻译概率（含源端和目标端 2 个方向的概率）、词汇化翻译概率（2 个方向的）、短语惩罚、glue 翻译规则惩罚、词惩罚和语言模型。系统的解码方法是 CKY+风格的柱搜索。在进行 k-best 和 1-best 搜索时，使用了 cube-pruning 技术，从而加速了解码过程。在加载翻译模型时，对于源语言端相同的规则至多有 10 个（目标端不同

的 10 个规则）载入内存，进行柱搜索时，默认设定的 beam-size 为 200，k-best-list-size 为 100。

表 1. 参评系统在各项目中的分布

翻译评测项目	系统编号	Primary/Contrast
英汉新闻领域	DINO	Primary-systemA
	MOSES	Contrast-systemB
	MUSES	Contrast-systemC
	KIMI	Contrast-systemD
汉英新闻领域	DINO	Primary-systemA
	MOSES	Contrast-systemB
	KIMI	Contrast-systemC
	MUSES	Contrast-systemD
英汉科技领域	MOSES_H	Primary-systemA
	DINO	Contrast-systemB
	KIMI	Contrast-systemC
日汉新闻领域	MOSES	Primary-systemA
	DINO	Contrast-systemB
	KIMI	Contrast-systemC
藏汉政府文献	MOSES_H	Primary-systemA
	MOSES	Contrast-systemB
	DINO	Contrast-systemC
	KIMI	Contrast-systemD

2.4 MOSES 系统

MOSES 使用的特征和解码方式同上面介绍的 DINO 相同，这里不作过多的介绍，更详细的内容可以参见 MOSES 工具包（Koehn et al.，2007）。在评测过程中，系统的 beam-size=200，k-best-list-size=100.

2.5 MOSES_H 系统

MOSES_H 是开源 MOSES 工具包（Koehn et al.，2007）中自带的层次短语翻译。同 KIMI 系统一样，它是基于同步上下文无关文法，翻译模型含有 8 个特征，它的解码也是 CKY 风格的柱搜索过程。在层次短语规则抽取中，限制每条层次短语规则在句子中最多覆盖 15 个词，且抽取的层次短语规则其源语言端包含的终结符和非终结符数不超过 5 个。在解码时，对于每个子串，只保留得分最高的前 1000（beam-size=1000）个推导，同时在系统 K-best 解码时，设定 k-best-list-size =100。系统对层次短语的调序距离进行了限制，当两个非终结符所对应源语言子串的长度和大于 10 时，系统在合并两个子串时不再使用规则表中的调序规则，而是使用 glue 规则直接对两个非终结符进行顺序拼接。

3 实验

3.1. 数据及其预处理

训练数据

以上系统的训练数据包含数据 A, B 两部分，A 部分数据用于训练翻译模型也即抽取翻译规则，B 部分数据用于训练语言模型。数据 A 为 CWMT2011 官方提供的双语平行语料；数据 B 为双语平行语料的目标语言部分，同时还包含 sogou 中文语料。训练数据的基本信息及其在 5 个评测子项目中的分布如表 2，3 所述。

表 2. 训练翻译模型数据的基本信息

翻译评测项目	句对数	源/目标语言词数
英汉新闻领域	5830855	117314113/112338130
汉英新闻领域	5830855	112338130/117314113
英汉科技领域	911527	25291336/25093501
日汉新闻领域	282486	3238177/2477812
藏汉政府文献	101629	1461695/995179

表 3. 训练语言模型数据的基本信息

翻译评测项目	语料名称	词数
英汉新闻领域	目标语+Sogou	112338130+390196660
汉英新闻领域	目标语	117314113
英汉科技领域	目标语+Sogou	25093501+390196660
日汉新闻领域	目标语+Sogou	2477812+390196660
藏汉政府文献	目标语+Sogou	995179+390196660

数据前处理

在处理双语训练语料时，依次进行了如下操作。在中一英双向翻译任务中，中英文的训练语料都对标点符号进行了统一的规范化处理，其中有一些标点符号需要根据上下文进行规范化。如中文语料中的半角逗号，如果出现在数字中间，则保留半角形式，否则转换为全角。在日一中和藏一中翻译任务中，我们只对中文训练语料的标点符号进行了统一的规范化处理。在处理英文语料时，Moses 的 tokenization 工具不能 tokenize 某些情况如数字间的逗号和小数点两边不加格，www 网址是全部连起来的等。为此，我们重新实现了一个 tokenization 工具。在处理 sogou 语料(汉语)时，我们也对其中的全半角做了统一处理。

中文分词处理采用的是 stanford 中文分词工具（Tseng et al.,2005）。日语分词采用的是分词工具 mecab（Kudo,2009），由于 mecab 工具对数字的处理如“33.5”、“25%”的结果是“33 . 5”、“25 %”，后续又将其合并回去。由于没有相关的训练语料，我们没有对原始的藏语进行分词，我们直接使用主办方提供的、已经做过分词处理的藏语语料，但是移除了词与词之间出现的分字符。

翻译模型的训练

我们采用 GIZA++(Och and Ney,2000)对双语训练语料进行词对齐，词对齐方式是 grow-diag-final-and。在完成词对齐后，我们采用启发式规则合并两个方向的词对齐结果，作为最终的词对齐结果。然后按照各翻译系统的要求，从双语词对齐的语料上获得各自的翻译模型。

语言模型的训练及工具

语言模型的训练数据如表 3 所示，除了汉英新闻领域项目的翻译外，其他的项目都训练了 2 个语言模型：目标语端的语言模型和 sogou 语言模型。

由于 sogou 语言模型训练数据存在着一些噪音，直接使用所有的 sogou 数据训练的语言模型对翻译效果的正作用不大，一个证据就是 tuning 后该语言模型对应的权重很小甚至为负数。我们认为这种噪音产生的一个可能的原因就是 sogou 语料中含有一些和训练数据相差较大的数据，甚至是所处项目领域外的句子。为了处理这个问题，我们采用了一种基于迷惑度——perplexity（Brew et al.,2000）的方法来改变 sogou 语言模型的分布。其主要思想就是和增大和训练数据目标语言端相近的句子的分布，同时降低和目标语言端相差较远的句子的分布。对 sogou 语料中的每个句子 s，它的迷惑度定义成

$$\mathrm{p}(s) = 2^{-1/N_w} \sum_{i=1}^{N_w} \log_2 p(w_i \mid w_{i-k+1}, \ldots, w_{i-1})$$

其中，N_w 为 s 的长度，$p(w_i \mid w_{i-k+1}, \ldots, w_{i-1})$ 为双语训练语料的目标语端的 k 元语言模型。我们的具体做法是根据迷惑度对所有句子进行排序，剔除迷惑度较高的 10%句子后训练语言模型。

我们使用了开源的 SRILM 工具(Stolcke,2002)来训练语言模型，并采用 Modified Kneser-Ney(Chen and Goodman,1998)进行平滑。各项目使用的语言模型情况为：汉英新闻领域含有一个 5 语言模型，英汉新闻、科技均含有 2 个 5 元语言模型，日汉新闻和藏汉政府文献均含有一个 4 元和 5 元语言模型(其中 sogou 为 5 元，目标语端的为 4 元)。

翻译后处理

对于翻译结果，我们采用了如下的后处理：目标语言是中文的翻译结果，去掉词与词之间的空格；目标语言是英文的翻译结果，利用 recaser 工具进行英文大小写自动恢复(http://www.statmt.org/wmt07/baseline.html)，同时对句子进行 detokenize；在中一英新闻藏汉政府文献翻译项目中，去掉了翻译结果中的未登录词。

3.2. 实验环境

硬件:

CPU 品牌:Intel

CPU 型号: Xeon

CPU 主频:2.13GHz

CPU 数量: 16

内存容量:128GB

软件:

操作系统类型:Linux

操作系统版本:CentOS release 5.6 (Final)

33. 实验结果及分析

5 个评测项目上的开发集都采用官方提供的数据：汉英新闻的开发集共 1006 句；英汉新闻的开发集 1000 句；英汉科技的开发集有 1116 句；日汉新闻的开发集有 500 句；藏汉

政府文献有 650 句。所有系统的训练均采用 MERT(Och,2003)进行参数训练，并使用 MOSES 工具包中的训练的脚本。在 tuning 时，除了 MUSES 的 k-best-list-size 为 30 外，其他系统的均为 100.

表 4 各参评系统在汉英新闻领域项目上的评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
Primary-systemA	0.2191	0. 2278	6.9503	0.6939	0.6891	0.5059	0.3557
Contrast-systemB	0.2021	0.2093	6.505	0.6664	0.7039	0.5331	0.3337
Contrast-systemC	0.2023	0.2123	7.0335	0.664	0.7138	0.5227	0.3122
Contrast-systemD	0.2065	0.2191	7.0053	0.6721	0.7212	0.5222	0.3068

所有 4 个系统在汉英新闻领域项目上均使用了一个语言模型—目标语言端训练出的，它们的翻译效果见下表 4。从表 4 中可以看出，我们的主系统—DINO 性能最好，比对比系统至少提高 1.3BLEU5-SBP；而且，我们的 3 个对比系统之间势均力敌，表现都较为稳定。

在剩下的所有以汉语作为目标语的4个项目中，所有系统都是使用了2个语言模型，见表 3. 从表5中可以发现，在英汉新闻领域项目上，主系统--DINO在两个测试集Progress和Current上都比对比系统效果更优，而KIMI系统表现不好。

表 5 各参评系统在英汉新闻领域项目的 2 个测试集上的评测结果

测试集	System	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
Progress	Primary-systemA	0.3319	0.3405	0.2777	9.2015	9.2124	0.7733	0.6216	0.3876	0.4333
	Contrast-systemB	0.3195	0.3305	0.2689	9.1078	9.1178	0.7601	0.6381	0.4015	0.4027
	Contrast-systemC	0.3153	0.3239	0.2626	8.872	8.882	0.7715	0.6221	0.3978	0.438
	Contrast-systemD	0.2979	0.3083	0.2452	9.1352	9.1429	0.7669	0.648	0.392	0.3961
Current	Primary-systemA	0.3191	0.3252	0.2668	8.7607	8.7717	0.755	0.6094	0.3938	0.4069
	Contrast-systemB	0.3089	0.3166	0.2588	8.7203	8.7313	0.746	0.6266	0.406	0.38
	Contrast-systemC	0.3056	0.3122	0.2541	8.4843	8.4945	0.7549	0.6174	0.4034	0.4153
	Contrast-systemD	0.2915	0.3001	0.2412	8.8581	8.8672	0.7523	0.6381	0.3954	0.3735

表 6 各参评系统在英汉科技领域项目上的评测结果

System	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
Primary-systemA	0.3787	0.3922	0.3247	10.379	10.401	0.8433	0.6217	0.3046	0.3875
Contrast-systemB	0.3852	0.4059	0.3386	10.303	10.325	0.8207	0.5964	0.31	0.4178
Contrast-systemC	0.3779	0.3914	0.3242	10.363	10.387	0.8378	0.64	0.3053	0.3893

表 7 各参评系统在日汉新闻领域项目上的评测结果

System	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
Primary-systemA	0.4178	0.436	0.3716	10.299	10.319	0.8352	0.5048	0.311	0.4505
Contrast-systemB	0.4195	0.4342	0.3693	10.404	10.424	0.8351	0.4994	0.304	0.4494
Contrast-systemC	0.3807	0.3888	0.325	10.106	10.123	0.8307	0.5327	0.3064	0.4274

表 8 各参评系统在藏汉政府文献项目上的评测结果

System	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
Primary-systemA	0.4882	0.5183	0.4741	9.7424	9.7834	0.8091	0.4206	0.2672	0.5794
Contrast-systemB	0.4601	0.4847	0.4387	9.385	9.4192	0.7961	0.4451	0.2845	0.5613
Contrast-systemC	0.486	0.5146	0.4692	9.6773	9.7199	0.8057	0.4272	0.2704	0.5648
Contrast-systemD	0.4921	0.5226	0.4791	9.742	9.7817	0.8097	0.4158	0.2653	0.5738

从表 6 中可以看到，在英汉科技项目上，我们的主系统—MOSES_H 的效果不如对比系

统 Contrast-systemB (DINO)的好，它与 Contrast-systemC（KIMI）相当。表 7 给出了日汉新闻项目的评测结果，从中可以看出 Primary-systemA(MOSES)与 Contrast-systemB(DINO)翻译效果相当，它们比 Contrast-systemC(KIMI)高出达 3 个 BLEU5-SBP。在藏汉翻译项目上（表 8），表现最好的是 Contrast-systemD(KIMI)，我们的主系统(MOSES_H)与它的效果相当，而 Contrast-systemB （MOSES）翻译效果不佳。

4 结论

在 CWMT2011 机器翻译评测中，HIT-MITLAB 共构建了 5 个翻译系统，它们是 DINO，MUSES，KIMI，MOSES 和 MOSES_H。在我们参加的汉英、英汉、日汉新闻，英汉科技，藏汉共 5 个翻译项目中，我们的系统表现稳定、性能较好。从提交的结果来看，我们发现：在大规模的翻译任务上，基于短语的翻译性能更好；而在较小规模的翻译项目上，比如藏汉，基于层次短语的翻译性能更具优势；同时可以看到，在我们所参加的各个翻译项目中，我们开发的系统能够超越开源的 MOSES 系统（包括短语模型和层次短语模型），至少和它性能相当。

此次参与 CWMT2011 机器翻译评测的经历，特别是与国内外同行的交流，为我们进一步开展统计机器翻译的研究工作带来诸多帮助！

参考文献

Brew, Chris, and Moens, Mark. March. 2000. Data Intensive Linguistics. Manuscript in progress, www.ltg.ed.ac.uk/?/chrisbr/dilbook/.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In Proc. of ACL.

P. Koehn, F. Och, and D. Marcu. 2003. Statistical Phrase-based Translation. In Proc. of NAACL.

P. Koehn, H. Hoang, A. Birch, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of ACL.

Taku Kudo. 2009. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Available at http://mecab.sourceforge.net/

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proc. of ACL.

Franz Josef Och. 2003. Minimum Error Rate Training in StatisticalMachine Translation. In Proc. of ACL.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In Proc. of ICSLP.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter. In Fourth SIGHAN Workshop on Chinese Language Processing.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. Proc. of COLING-ACL.

Le Zhang. 2006. Maximum Entropy Modeling Toolkit for Python and C++. Available at http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.