南京大学第五届全国机器翻译研讨会系统评测报告[1]

黄书剑 赵迎功 李博渊 吴秋锋 戴新宇 陈家骏
南京大学 计算机科学与技术系 自然语言处理研究组 南京 210093
E-mail: {huangsj, zhaoyg, liby, wuqf, daixy, chenjj}@nlp.nju.edu.cn

摘 要: 本文介绍了南京大学自然语言处理研究组(NJU-NLP)参加第五届全国机器翻译研讨会机器翻译评测（汉英新闻领域单一系统）的情况。在本次评测中，NJU-NLP 主要采用了一个基于短语的统计机器翻译系统；并通过添加短语特征，利用词性标注等方法提高了系统的性能。

关键字: 统计机器翻译、基于短语的翻译模型、词性标注

NJU-NLP's Technique Report for the 5th China Workshop on Machine Translation

Shujian Huang, Yinggong Zhao, Boyuan Li, Qiufeng Wu, Xinyu Dai, Jiajun Chen
NLP Research Group, Department of Computer Science and Technology
Nanjing University, Nanjing 210093, China
E-mail: { huangsj, zhaoyg, liby, wuqf, daixy, chenjj}@nlp.nju.edu.cn

Abstract: *This paper describes our (NJU-NLP) participation for evaluation of the 5th China Workshop on Machine Translation. We submit results for the Chinese to English Single System Task (News Area). During the evaluation, a phrase-based translation system is employed. In addition, we improve the system performance by adding features for phrase scoring and making use of POS-tag information.*

Keywords: *statistical machine translation, phrase-based translation model, POS-tag*

1 引言

南京大学计算机科学与技术系自然语言处理实验室（NJU-NLP）参加了第五届全国机器翻译研讨会（CWMT2009）组织的汉英新闻领域单一系统的机器翻译评测。本文主要描述 NJU-NLP 在本次评测中所使用的系统情况以及相应的实验结果。

NJU-NLP 在本次评测中利用开源工具 Moses[Koehn et al., 2003]构建了一个基于短语的统计机器翻译系统。与默认的 Moses 系统相比，我们从以下两个方面改进了基线系统的性能。第一，为了更好的估计短语翻译概率从而区分不同质量的短语，我们在短语翻译模型中添加了三个描述性的特征，并通过最小错误率训练(MERT)[Och, 2003]来调节这三个特征的权重。第二，将源语言的词性标注的信息引入到翻译模型中。实验表明，上述两个方法都能在一定程度上改善机器翻译的效果。此外，我们在评测过程中还对基于句法的预调序[Wang et al., 2007]等问题进行了一些实验和研究，不过由于结果并不理想，在本文中不予详述。

[1]本文的工作受到 863 国家高科技项目（编号 2006AA010109），国家自然科学基金（编号 60673043）以及南京大学研究生科研创新基金（编号 2008CL08）的资助。

本文的后续部分结构如下：第二部分简述基于短语的统计机器翻译模型；第三部分描述我们用于改进短语评分所加入的三个特征；第四部分介绍我们引入词性标注信息的情况；第五部分记录了系统所使用的数据和实验结果；第六部分为全文的小结。

2 基于短语的统计机器翻译模型

2.1 对数线性翻译模型

本文采用了[Koehn et al., 2003]提出的基于短语的统计机器翻译模型，通过对数线性模型将句子的得分描述为若干特征的线性组合（见公式一，其中 **e**，**f** 为目标语言和源语言的句子，$h_m(\mathbf{e},\mathbf{f})$为第 m 个特征函数，$\lambda_m$为第 m 个特征函数所对应的权重）。模型的解$\hat{\mathbf{e}}$通过 beam search 算法来获得。

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})\right)}{\sum_{e'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}',\mathbf{f})\right)} \quad (1)$$

这一系统采用的基本特征为如下 14 个[2]：

- 正向短语翻译概率（phrase translation probability）
- 反向短语翻译概率（inverse phrase translation probability）
- 正向词汇翻译概率（lexical weighting）
- 反向词汇翻译概率（inverse weighting）
- 短语惩罚（phrase penalty）
- 基于距离的调序惩罚（distance-based reordering penalty）
- 6 个词汇化调序模型特征（lexicalized reordering model）：
- 语言模型（language model）
- 词惩罚（word penalty）

对数线性模型的参数通过在开发集上进行最小错误率训练（MERT）方法[Och, 2003]对 BLEU4 进行优化获得。

2.2 调序模型

本文采用 msd-bidirectional-fe 的调序模型，具体解释如下：

2.2.1 调序类别（MSD）

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：

1. 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）；
2. 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）；
3. Discontinuous 即翻译结果不联系（e，e’不相邻）。

2.2.2 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。

2.2.3 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]本文在实验过程中添加了部分特征，这些特征将在第三部分进行描述。

3 改进短语评分

在使用上述模型进行翻译的过程中我们注意到，系统会自动抽取出大量的短语，其中很大一部分都并不具有实际的意义。我们观察发现，这些无意义的短语大多都具有以下两个特征之一：一，含有较多未对齐的词，二，源语言和目标语言部分的长度相差较大。

为了有效的区分开质量参差不齐的短语，使得短语翻译模型能够更准确的选择候选翻译，我们向模型中添加了如下特征：

3.1 实对齐词比例

该特征用于描述短语翻译项中源语言或目标语言部分有对齐的词占全部词的比例：

$$h_{\text{aligned\_ratio\_e}} = 1 - \frac{c(e_{\text{null}})}{|\mathbf{e}|} \tag{2}$$

$$h_{\text{aligned\_ratio\_f}} = 1 - \frac{c(f_{\text{null}})}{|\mathbf{f}|} \tag{3}$$

其中 $c(e_{null})$和 $c(f_{null})$分别为目标语言和源语言中未对齐的词数。

3.2 短语长度比例

该特征用于描述短语翻译项中源语言和目标语言部分的长度比例关系，两者相差越大该特征的值就越大。

$$h_{\text{length\_ratio}} = \frac{1}{\frac{|e|}{|f|} + \frac{|f|}{|e|}} = \frac{|e| * |f|}{|e| + |f|} \tag{4}$$

4 利用词性标注信息

词性标注是自然语言处理研究中进行语法语义分析的重要步骤。为了使得翻译结果在语法和语义方面更加合理，我们考虑向目前的翻译模型中添加词性标注的信息以改进其性能。

考虑到由于中文常常有一词多义的情况出现，且很多情况下动词和名词的形式相同，中文词的翻译较为容易出现歧义，我们将中文（即源语言）的词性标注的信息引入到翻译过程之中。在本文中，我们采用了一个相对简单的策略，即将词性作为词的一部分合并入词的字符串中，这与[王博等, 2008]采取的策略是相似的。经过处理之后的双语数据和抽出的短语如表 1 所示。

表 1. 加入源语言词性标注后的数据

源语言句子	我|PN 想|VV 要|VV 烤|VV 牛肉|NN 三明治|NN 和|CC 咖啡|NN
目标语言句子	i would like a roast beef sandwich and some coffee
短语表中的短语	我|PN 想|VV 要|VV ||| i would like 我|PN 想|VV ||| i would like …

用于进行翻译模型训练的双语语料的中文部分以及开发集和测试集语料的中文部分都需要进行词性标注；语言模型保持不变；翻译模型的训练仍然可以按照第二部分所述方法进行。

5 实验

5.1 数据

本文所使用的训练数据来自评测组委会的双语训练数据，共计近 380 万句对（见表 2）。

本文所使用的语言模型采用 SRILM[Stolcke, 2002]训练，训练数据为路透社提供的单语数据和上述双语训练数据的英文部分，共计 1600 万句。

本文内部使用的开发数据集是 863 评测语料 2005 年中英篇章翻译的数据（2005zeWrit）下文实验中提到的系统如非特殊说明，最小错误率训练均是在这一数据上进行；本文内部使用的测试数据集是第四届全国机器翻译研讨会（CWMT2008）中英翻译评测的数据。

表 2. 双语训练数据列表

数据名称	双语句对数（万句）
中信所英汉科技文献句子级对齐语料库	61.15
计算所汉英平行语料库	107.63
点通汉英平行语料库	100
CLDC-LAC-2003-004 中科院计算所和自动化所中英句子级对齐双语语料库（扩充版）	25.23
CLDC-LAC-2003-006 北京大学汉英/汉日双语语料库（汉英部分）	20
厦门大学英汉电影字幕平行语料库	17.6
哈工大信息检索组英汉句子级对齐语料库	10
哈工大机器翻译组英汉句子级对齐语料库	5.22
万方汉英中文科技期刊论文摘要语料库	32.1
总计	378.93

5.2 数据的前后处理

我们对组织方提供的数据进行了全角到半角的转换，并根据长度比例进行了筛选（剔除句子长度比例较大的句对）。所有数据的中文部分都采用 ICTCLAS[3]进行了分词。此外，我们还用 Stanford POS Tagger[4]对中文数据进行了词性标注。

在后处理部分我们主要对系统不能识别的词语进行了规则翻译。主要分如下几步：将不能识别的中文数字转换成阿拉伯数字；将未识别的汉字字符串转换成对应的拼音序列；利用一个内部实现的 3 元 HMM 进行大小写恢复。

5.3 基线系统和参评系统

作为基线系统，我们采用了一个如第二部分所描述的基于短语的翻译系统。词对齐采用 Giza++[Och and Ney, 2003]得到；语言模型采用 SRILM 进行训练得到（5 元）；短语的抽取、评分、调序模型的训练以及最后的解码都采用了 Moses 工具包进行。

本次评测我们共提交了如下 4 个系统，都是在基线模型的基础上进行了一定的修改得到：

[3] http://www.ictclas.org
[4] http://nlp.stanford.edu/software/tagger.shtml

SystemA：本次评测的主系统，采用了添加短语翻译特征和使用词性标注信息的方法提高系统性能，并在 CWMT2008 的评测数据上进行 MERT 训练；

SystemB：本次评测的对比系统，采用了添加短语翻译特征和使用词性标注信息的方法提高系统性能；

SystemC：本次评测的对比系统，只利用了词性标注信息提高系统性能；

SystemD：本次评测的对比系统，只采用了添加短语翻译特征的方法提高系统性能。

5.4 实验结果及分析

表 3 依次列出了上述五个系统在内部使用的开发和测试数据集以及 CWMT2009 测试数据上的表现（Test 的得分是系统输出经过后处理之后的结果）。

表 3. 基线系统和参评系统在各数据集上的得分

	Dev	Test(CWMT2008)		CWMT2009
	CIBLEU	CIBLEU	CSBLEU	BLEU-SBP
Baseline	0.252091	0.2655	0.2444	-[5]
SystemA	**0.282507**	**0.2789**	**0.2598**	**0.2054**
SystemB	0.255759	0.2697	0.2503	0.2006
SystemC	0.253357	0.2577	0.2388	0.1934
SystemD	0.255772	0.2667	0.2473	0.1951

从结果中可以看出，本文所采用的两种方法都能给基线系统的性能带来一定的提高，其中加入短语评分特征的效果更加明显一些。将两种方法合并在一起使得最终的翻译效果有了进一步的提高。这与我们之前的预期是一致的。

我们在实验过程中发现，选择不同的开发数据集对实验结果有着巨大的影响。虽然同是新闻领域的语料，但是在 CWMT08 上进行参数训练的结果比在 2005zeWrit 上的结果有着明显的提高。我们推测随着时间的变化，新闻领域的语料的内容也发生了一些改变，这些改变虽然不容易显示的描述出来，但是对机器翻译的效果的影响不容忽视。

6 总结

本文系统的描述了南京大学计算机科学与技术系自然语言处理实验室（NJU-NLP）参加了第五届全国机器翻译研讨会（CWMT2009）组织的汉英新闻领域单一系统的机器翻译评测的情况。这是 NJU-NLP 首次参与全国机器翻译研讨会组织的评测，我们使用了开源的机器翻译工具 Moses，并在其基础上进行了一定的改进。实验表明，我们的系统相对基线系统而言有了明显的提高。

参考文献

Chao, Wang, Michael Collins and Philipp Koehn. 2007. Chinese Syntactic Reordering for Statistical Machine Translation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)

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

Och, J. Franz and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models",

[5] 基线系统的结果并未提交

Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Och, J. Franz. 2003. Minimum Error Rate Training in Statistical Machine Translation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. Proceedings of International Conference on Spoken Language Processing.

王博 蒋宏飞 梁华参 张春越 孙加东 赵铁军 刘树杰 马永亮 王欣欣. CWMT’2008 机器翻译评测技术报告. CWMT’2008 论文集.