影响统计翻译系统性能的因素分析

柴春光 宗成庆

（中国科学院自动化研究所模式识别国家重点实验室，北京 100080）

摘 要： 统计翻译方法已经成为目前国际上机器翻译研究的主流方法，但对于一个统计翻译系统来说，哪些因素是影响系统性能的关键因素，它们对系统性能的影响有多大，并没有相关的文献对此做详细的调研和分析。本文以基于短语的（phrase-based）统计翻译系统为例，针对影响系统性能的几个因素做了一系列实验，并对其进行了详细地分析。实验结果表明：影响基于短语的统计翻译系统性能的主要因素依次为系统模型选择的特征、训练语料的规模和预处理。

关键词： 统计机器翻译；基于短语的翻译模型；系统性能；实验分析

Investigation into the Performance of SMT Systems

Chunguang CHAI, Chengqing ZONG

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academic of Sciences, Beijing, 100080, China)

Abstract: Nowadays, statistical machine translation(SMT) has become the most popular method in the research field of machine translation. However, there is no literature that analyzes what factors mainly affect the performance of SMT systems and how important they are in detail. In this paper, a series of experiments about many factors are made on the phrase-based Chinese-to-English SMT system and detailed analysis about the experimental results are given. We then can conclude that three main factors are contributing to the performance of the SMT system. They are the features in the system model, the size of corpus and pre-processing.

Keywords: Statistical Machine Translation; Phrase-based Translation Model; Performance of Systems; Experiment Analysis

1 引言

自从 IBM 提出了五个经典的统计翻译模型[1]，统计翻译方法越来越受到重视，基于统计的方法已经逐渐成为国际上机器翻译研究的主流方法。IBM 的统计翻译模型是基于词的翻译模型，这样的模型只考虑了词与词之间的线性对位关系，没有考虑源语言和目标语言多个词语之间的对齐关系，不利于处理短语内部词语的关系和短语的翻译。针对这个问题，一些学者提出了基于短语的统计翻译方法[2][3]。

在 C-STAR 组织的国际口语翻译系统评测（International Workshop on Spoken Language Translation, IWSLT）[4][5]和 NIST[1]最近两年的机器翻译评测中，翻译系统都是以基于短语的统计翻译方法为主，而且基于短语的统计翻译系统取得了最好的成绩。虽然近年来机器翻译取得了很大的进步，但是目前的翻译系统仍然不能令人满意。很多学者对基于短语的统计翻译系统提出了自己的改进方法，但是没有全面地分析影响系统性能的因素。在中文到英文的翻译中，由于两种语言的差异，需要针对其自身的特点进行改进。本文将根据实验分析影响从中文到英文的

作者简介：柴春光（1981- ），男，北京人，硕士研究生，E-mail: cgchai@nlpr.ia.ac.cn.

[1] http://www.nist.gov/speech/tests/mt/

基于短语的统计翻译系统性能的因素，指出主要影响因素，有助于针对性地对翻译系统进行改进。

本文第二部分介绍基于短语的统计机器翻译模型的基本原理；第三部分介绍翻译过程中使用的搜索策略；第四部分通过实验分析影响基于短语的统计翻译系统性能的因素；最后给出了相关结论和下一步工作计划。

2 基于短语的统计机器翻译模型

基于词的统计翻译模型只考虑了词与词的线性对位关系，没有考虑短语内部词的关系和翻译，针对这个问题，很多学者提出了基于短语的统计翻译方法[2][3]。基于短语的方法不是将每个源语言词单独的翻译为目标语言词，而是将源语言句子 f 切分为由 I 个短语（一个连续的单词序列，所以称之为短语）组成的短语序列 $\bar{f}_1^I$，然后将每个短语翻译为目标语言短语，从而生成目标语言句子 $\bar{e}_1^I$。

在最大熵模型[6]下，给定的 f，其最佳译文 e 可以用以下公式表示：

$$\hat{e}_1^I = \arg\max_{e_1^I}\{\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)\} \tag{1}$$

选择翻译模型、语言模型和扭曲模型作为基本特征，则：

$$\hat{e} = \arg\max_{e}\{\lambda_T \log p_T(f \mid e) + \lambda_{LM} \log p_{LM}(e) + \lambda_D \log p_D(e, f)\} \tag{2}$$

其中 $p_T(f \mid e)$ 代表短语的翻译概率，可以通过公式（3）计算，$p(\bar{f}_i \mid \bar{e}_i)$ 为每个短语的翻译概率；

$$p_T(\bar{f}_1^I \mid \bar{e}_1^I) = \prod_{i=1}^{I} p(\bar{f}_i \mid \bar{e}_i) \tag{3}$$

$p_{LM}(e)$ 代表目标语言的语言模型概率；$p_D(e, f)$ 代表扭曲模型概率（位变概率），用于调整目标语言短语的次序，每个短语的位变概率可以使用公式（5）计算：

$$p_D(\bar{e}_1^I, \bar{f}_1^I) = \prod_{i=1}^{I} d(a_i - b_{i-1}) \tag{4}$$

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|} \tag{5}$$

a_i 表示翻译为第 i 个目标语言短语的源语言短语的起始位置，b_{i-1} 表示翻译为第 i-1 个目标语言短语的源语言短语的结束位置。

3 解码算法

解码过程就是搜索使得公式（1）的概率最大的目标语言句子。解码过程使用基于柱搜索（Beam Search）的搜索算法[7]，按从左到右的方向生成目标语言句子。下面针对解码部分可能影响翻译系统性能的因素进行说明。

(1) 加入翻译对空的候选项

考虑到中文和英文之间的差异，一些英文词不与任何中文词对应，例如“the”、“of”等，这些词可以看作与中文句子中的空位置对应，即可以由中文的 NULL 翻译得到。文献[8]中提出了每扩展一个英文短语，可以在其后面添加一个对空的英文词。

(2) 启发式信息

在对翻译假设进行扩展时，还需要对当前假设的未来代价（启发式信息）进行估计。启发式信息需要覆盖所有的未翻译的源语言单词。计算启发概率时只考虑翻译概率和语言模型概率。

(3) 重组假设

重组假设是一种没有风险的减小搜索空间的方法。如果两个假设的下列属性一致，则只保留打分最高的假设，其它的假设可以安全的丢弃。

- 到目前为止已覆盖（翻译）的源语言单词。
- 最后生成的两个目标语言单词（使用三元语言模型）。
- 最后覆盖的源语言短语。

(4) 剪枝操作

剪枝操作包括观察剪枝和柱容量剪枝。观察剪枝是在搜索过程中对每个源语言短语只选择最优的 n_0 个目标语言短语作为翻译候选项进行扩展。柱容量剪枝即为：对于假设集合 C，其假设的个数为 $N(C)$，设定一个最大容量值 n_c，如果 $N(C) > n_c$，则只保留最好的 n_C 个假设，其余的将被丢弃。

4 实验分析

为了分析影响基于短语的统计翻译系统的性能的因素，我们进行了一系列实验。主要分析预处理、扭曲模型、语料库规模、解码策略、解码器参数等因素对系统性能影响的大小以及哪些是主要因素。

4.1 训练语料和测试语料

我们使用 BTEC（Basic Travel Expression Corpus）130000 句和 IWSLT'05 机器翻译评测提供的 20000 句中英文双语对齐的句子作为训练语料。使用 IWSLT'05 的开发集 506 句用于评测系统的性能，每个中文句子有 16 个参考译文。训练语料的详细信息见表 1。

表 1 训练语料

语料		句子数	总词数	单词表大小	平均长度
BTEC	中文	130,000	861,333	15,008	6.63
	英文	130, 000	888, 098	14, 922	6.83
IWSLT'05	中文	20, 000	176,199	8, 687	8.81
	英文	20, 000	183, 452	6, 956	9.17

4.2 预处理对系统性能的影响

对训练语料的预处理：IWSLT'05 的训练语料英文句子中存在很多的缩写词（"I'll"、"I'd"）,并且标点符号和单词之间没有空格。在进行训练之前对英文句子进行预处理，将缩写词和标点符号进行简单地预处理，即将"I'll"拆分为"I 'll"，在标点和单词之间添加空格，而对于"Mr."等不处理。在得到翻译结果后进行反向处理，即将"I 'll"还原为"I'll"。

对数字的预处理：在统计翻译中，训练语料很难包含测试语料中的所有命名实体，而且命名实体的翻译又有其自身的特点。可以在解码之前根据命名实体的特点使用不同的翻译方法对其进行翻译，然后将翻译结果直接加入到翻译候选项表。在本文的实验中只针对数字利用规则的方法进行了预处理。

经统计 IWSLT'05 训练语料英文句子（20000 句）中"I'll"等形式的缩写共有 6954 个，测试语料的参考译文（8096 句）中共有 2067 个。测试语料（506 句）中的预处理的数字共有 78 个。

由表 2 可以看出，进行训练预处理后，系统的 BLEU 打分提高了 14.7%。对于数字这样规律性很强而且训练语料中很难全部覆盖到的语言现象，通过规则的方法进行翻译可以取得更好的效果。

表 2 训练语料预处理

	未进行预处理	加入训练语料预处理	加入训练语料预处理及数字预处理
BLEU	0. 1522	0.1745	0.1932

4.3 扭曲模型等对系统性能的影响

表 3 模型、预处理和训练语料的影响

	IWSLT'05		BTEC	
	BLEU	NIST	BLEU	NIST
Base	0.2043	5.0402	0.2374	4.1175
Base+R	0.2212	5.3823	0.2623	4.9462
Base+D	0.1508	3.9764	0.2211	4.3189
Base+F	0.2183	5.2938	0.2548	4.6760
Base+R+F	0.2165	5.3533	0.2647	5.0408
Base+R+D+F	0.1550	3.9472	0.2170	4.1802

注：Base：只包含语言模型和翻译模型；R（Recombine）：重组假设；D（Distortion）：扭曲模型；F（Future）：启发概率。

根据第二组实验可以分析得到（见表 3）：按照公式（5）计算的扭曲模型偏向于源语言和目标语言语序一致的情况，而中文和英文之间语序的差异性很大，加入扭曲模型后反而降低了系统的性能，表明该扭曲模型并不适合中文到英文的翻译；由于假设重组可以将一些不可能产生最优结果的假设安全地删除，加入启发信息可以优先扩展最有可能产生最优结果的假设，这两个策略可以使一些好的结果优先扩展并保留在容量有限的栈内；训练语料的规模和覆盖度的增大可以训练出更加合理的翻译模型和语言模型参数。

4.4 加入翻译对空对系统性能的影响

从表 4 中可以看出加入中文对 NULL 和英文对 NULL 的短语翻译候选项都可以提高系统的 BLEU 打分。这是因为中英文之间的词并不是线形对应的，例如中文词“的”、“了”、“吧”等词可能不与英文句子中的任何词相对应，英文中的“of”、“the”等词可能不翻译为任何中文词，这时加入这些词对空的翻译候选项可以使翻译结果更加合理。

表 4 翻译对空的影响

	Base	Base+EC0	Base+CE0	Base+EC0+CE0
BLEU	0.1300	0.1347	0.1465	0.1550

注：Base：不包含翻译对空的词；EC0：英文对应中文的 NULL；CE0：中文翻译为 NULL。

4.5 解码器参数对系统性能的影响

解码器的参数主要为翻译候选项表的大小和存储翻译假设的栈的大小。如表 5 中所示，将翻译候选项表和翻译假设栈的大小分别扩大 5 倍和 10 倍，系统的 BLEU 打分虽然有所提高，但并不明显。经过分析发现，当翻译候选项表的大小超过 10 以后，每个短语对应的翻译候选项最大的概率值与最小的概率值之间的差异一般已经达到 100 倍以上；同样，翻译假设栈的大小超过 300 以后，最优假设和最差假设之间的差异一般可以达到 1000 倍或者更高，这时加大这两个参数只会产生一些概率很小的翻译假设，而由这些假设扩展得到最优结果的概率很小。

表 5 解码参数的影响

n_0 \ n_c	300	1000	3000
10	0.1550	0.1549	0.1549
30	0.1557	0.1559	0.1559
50	0.1559	0.1561	0.1561

注：n_0：翻译候选项表的大小；n_c：翻译假设栈的大小；使用 BLEU 打分作为评价标准。

4.6 综合分析

表 6 主要因素对系统的影响（BLEU 打分的提高）

翻译对空	19.2%	语料库的规模	16. 2%
训练语料预处理	14.7%	数字预处理	10.7%
假设重组	8.3%	启发信息	6.9%
解码器参数	0. 7%	扭曲模型	-26.2%

从表 6 中可以看出：加入翻译对空对系统的影响最大，可以使系统的 BLEU 打分提高 19.2%；语料库的规模、预处理、搜索策略也使系统性能的性能有明显的提高；解码器的参数对系统性能的影响不大；只有扭曲模型的加入反而使得翻译结果更差。

表 7 表明系统加入预处理、翻译对空、假设重组、启发信息几个主要影响因素后，BLEU 打分由 0.1565 上升到 0.2321，提高了 48.3%。

表 7 主要因素的综合影响

	Base	Best
BLEU	0. 1565	0. 2321

注：base：使用 IWSLT'05 训练语料，考虑表 6 中的影响因素：Best：使用 IWSLT'05 训练语料，加入预处理、翻译对空、假设重组、启发信息。两个系统的解码参数都为 n_0=10，n_c=300；

5 结论和下一步工作

通过实验发现，使用基于短语的统计翻译系统进行中文到英文的翻译时，在搜索过程中加入重组假设和启发式信息可以使搜索更加有效，而语料库的大小、特征选择、预处理以及翻译概率表的计算是影响系统性能的主要因素并且可以进行改进。在实验中使用的扭曲模型并不适合中文到英文的翻译，下一步将研究如何根据中英文翻译的特点调整英文的语序，建立更有效的扭曲模型。由于在最大熵模型下可以很方便地加入新的特征，将进一步分析中英文翻译中存在的问题，并加入不同的特征针对性地解决这些问题。

参考文献：

[1] Brown, P. F., Stephen A. Della Pietra, Vincent J. Della Pietra, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics,1993, 19(2):263-309.

[2] Och, Franz Josef, Christoph Tillmann, and Hermann Ney. Improved Alignment Models for Statistical Machine Translation. In Proceedings of the Joint Conference of Empirical Method in Natural Language Processing and Very Large Corpora. University of Maryland, College Park, MD, June 1999. Pages 20-28.

[3] Marcu, Daniel, and William Wong. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, PA, USA. July 2002.

[4] Akiba, Yasuhiro, Marcello Federico, Noriko Kando, et al. Overview of the IWSLT04 Evaluation Campaign. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT). Sept. 30-Oct. 1, 2004. Kyoto, Japan. Pages 1-9.

[5] Eck, Matthias, and Chiori Hori.. Overview of the IWSLT 2005 Evaluation Campaign. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT). Pittsburgh, USA. Oct. 24-25, 2005. Pages 11-32.

[6] Och, Franz Josef, Hermann Ney. Discriminative Training and Mximum Entropy Models for Statistical Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. July 2002. Pages 295-302.

[7] Koehn, Philipp. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, 2004, pages 115-124.

[8] Pang, Wei, Zhendong Yang, Zhenbiao Chen, et al. The CASIA Phrase-based Machine Translation System. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT). Pittsburgh, USA. Oct. 24-25, 2005. Pages 114-121