DOI：10.3969/j.issn.1001-8972.2013.11.052

基金项目：苏州大学大学生创新训练计划项目(11cxxj048)

主题模型在统计机器翻译中的应用

巢佳媛 贡正仙
苏州大学计算机科学与技术学院，215006

Research on Topic Model-based SMT

摘 要
在基于短语的统计机器翻译中，短语表是解码器工作的核心部分，它主要包含了源端和目标端短语的翻译概率和词汇互译概率；但传统的短语表数量庞大，并且领域自适应能力差。为了在解码过程中挑选出更高质量的短语对，适当减少内存消耗和缩小解码空间，本文介绍了一个简单易行的基于主题模型的统计机器翻译系统的实现：首先使用LDA工具建立主题模型，然后把主题模型信息嵌入到短语表中，最后为解码器设计一个预处理模块可以使解码器不需要改变就可以在翻译过程中根据主题选择合适的短语对，从而提高了翻译的效率和质量。

关键词
统计机器翻译；主题模型；LDA

Abstract
Phrase table is a key part for a phrase-based statistical machine translation, which contains probabilities of bi-lingual phrases and words. However, the traditional phrase table is very huge and its capacity of domain adaptation is very poor. In order to choose better phrase pairs and reduce the memory cost and the searching scope for the decoder, this paper introduces a simple yet effective topic model-based SMT. The detailed procedure includes：building topic model by LDA tool, embedding topic information to a phrase table and designing a pre-processing module to implement topic model-based SMT without modifying the decoder.

Keywords
statistical machine translation；topic model；LDA

引言

机器翻译是利用计算机把一种自然语言转换成另一种自然语言的过程。20世纪80年代以来，随着语料库语言学的兴起，基于统计方法的机器翻译取得了长足的进步，统计机器翻译（Statistical Machine translation，SMT）已经成为自然语言处理领域备受关注的研究热点。自20世纪90年代以来，SMT涌现出了诸多引人关注的方法和技术，经历了基于词的机器翻译、基于短语的翻译和基于句法的翻译等多个阶段，目前基于短语的统计机器翻译（Phrase-based SMT，简称PBSMT）[2,4,5,6]仍然是主流的统计机器翻译方法。

PBSMT在进行翻译时，源语言句子以“短语”（不是语法意义上的短语，只是相邻的词组）为单位进行切分，每一个短语被翻译成相对应的目标语言短语，目标语言短语经过重排序后生成较符合语法的目标语言句子。源语言短语和目标语言短语组成了翻译对，在解码之前大量的翻译对已经从词语对齐的双语语料库中自动抽取出来，所有抽取的双语短语组成了短语表，短语表中包含了PBSMT所依赖的核心概率：短语翻译概率和词汇互译概率。给定一个源语言句子，对这个句子进行短语划分（源语言短语）有很多种可能，每个不同的源语言短语所对应的目标短语通常也会对应几个甚至几十个目标短语，我们把与源语言短语相对应的所有目标语言短语称之翻译候选项。

为了得到高质量的双语对齐，往往需要大规模的平行语料库，因此语料中往往会包括多个领域的数据，解码器在面临“at the bank|||在银行”和“at the bank|||在堤岸”的选择时，哪个翻译对会被采用显然跟当前翻译句子的上下文信息有关，若是翻译财经类文章，显然使用第一个翻译对的可能性要大于第二个翻译对。传统的PBSMT在翻译过程中一般采用使整个目标端句子得分最高的候选项，而不考虑特殊的语义信息。为了有效地提高短语选择的准确性，本文将主题模型融入到翻译模型中，在传统的短语表的基础上，利用主题模型（Topic Model）[1]为短语表添加主题信息，改善基于短语的统计机器翻译的质量。与聚类等方法不同，基于主题模型的方法具有的识别大规模文本集中潜藏的主题信息的能力，它认为文本是由多个主题随机混合而成的，并且每个主题又可被看作是语义相关的词的多项式分布，也即它能够利用词分布将文本信息转化为易于建模的数字信息，所以主题可以用于信息检索、聚类、摘要提取以及信息间相关性判断等一系列应用。

本文首先介绍基于主题模型的SMT的系统框架，然后分别介绍主题模型的建立、带主题信息的短语表的建立以及如何通过预处理把主题模型嵌入到SMT系统的详细过程，最后给出系统的实验结果。

1 基于主题的PBSMT系统框架

对于给定的源语言句子S来说，P(S)是个非随机量，可以忽略不计，那么最终的翻译结果可以用这样一个式子[2]来表示：

$$T = \arg\max_{T} P(T \mid S) = \arg\max_{T} P(T)P(S \mid T) \quad (1)$$

在公式（1）中，P(T)是目标语言T的概率，即语言模型。语言模型反映了一个句子在目标语言出现的可能性，即该句子在目标语言规则上的合理性。语言模型只与目标语言有关，与源语言无关。P(S|T)是目标语言T翻译成源语言S的概率，即翻译模型。翻译模型与源语言和目标语言都有关，反映的是它们互相翻译的概率。在基于短语的统计机器翻译中，翻译模型不仅考虑词对齐，更加注重源语言与目标语言的短语之间的翻译概率。

一般的SMT系统都包括训练和解码两个阶段[4,5]，翻译模型和语言模型一般在训练阶段完成，开发者分别通过对并行语料库和目标单语语料库的处理可以获得它们。解码[6]是真正进行翻译的过程，因为一个句子可能会存在多个翻译，解码就是求最优的翻译结果的搜索过程。

本文描述的基于主题模型的统计机器翻译系统的总体框架如图1所示，也包括训练和解码两个阶段。但与传统的PBSMT的训练过程不同，基于主题的PBSMT在此基础上需要额外进行主题模型的训练，并且要改造传统的短语表使其包含主题信息。此外，另一个不同的地方在于—本文为传统的解码器设计了一个预处理模块，以实现不改变解码器就能在翻译过程中进行主题的区分。

2 带主题信息的短语表

在众多的主题建模中，D.M.Blei在2003年提出潜在狄利克雷分配模型（Latent Dirichlet Allocation，LDA）[1]，引起了自然语言处理等领域的广泛关注，并成功运用于主题检测。LDA模型是一个三层贝叶斯具有文本主题表示能力的产生式概率模型(一元模型)，三层是指：文本，主题和词，如图2所示。

假设一个文本语料库中包括D个文本，N个唯一词汇w_i，i∈(1，N)和K个主题，那

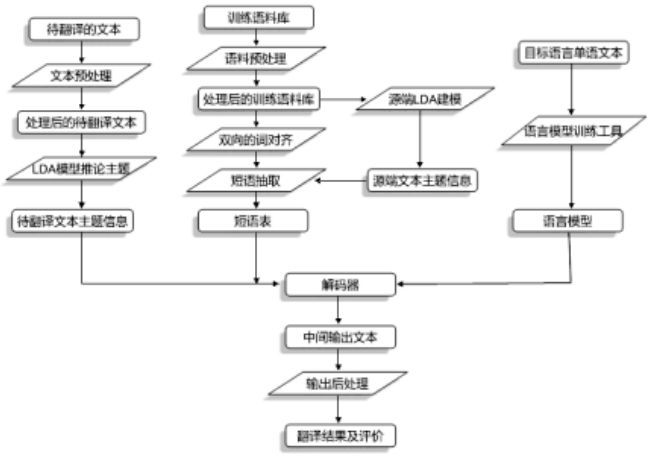

图1 主题相关的统计机器翻译流程图

么某个单词w_i的概率可用公式（2）表示:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_j)P(z_j) \quad (2)$$

其中，z_j表示第j个主题，$p(z_j)$表示第j个主题的概率，$j\in(1，K)$。w_i表示取自主题z_i的单词，$p(w_i|z_j)$ 表示单词w_i属于主题z_j的概率。

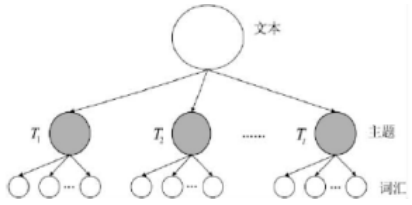

图2 LDA主题模型三层结构

上述主题模型需要通过训练才能获得，本文作者把翻译系统依赖的双语并行语料FBIS的中文部分独立出来进行如下训练：把语料等分成5份，然后取其中一份作为测试，多次计算主题模型的困惑度[1]，最后选择使困惑度达到最低的K值（即主题数目）为120。

（1）短语主题的确定

假设句子s由单词序列$w_1w_2\cdots\cdots w_n$组成，可以通过公式（3）来计算句子的主题分布，

$$P(s=w_1w_2\ldots w_n|z_j)=\frac{1}{n}\sum_{i=1}^{n}P(w_i|z_j) \quad (3)$$

一般大家公认为文档会包括多个主题，但一个句子却只属于一个主题T_s[]，因此本文建议选择使$P(S|Z_j)$值最大的Z_j作为该句子的主题，因为短语从属于句子，所以短语的主题就可以用句子的主题来表示，这里标记为T_{ph}。

（2）包含主题的短语对的构成

在通用的PBSMT中，短语对的翻译概率按照公式（4）进行计算：

$$P(\hat{e}|\hat{f}) = \frac{count(\hat{e},\hat{f})}{count(\hat{f})} \quad (4)$$

$\hat{f}$表示源端短语，$\hat{e}$表示目标端短语，count（$\hat{e}$，$\hat{f}$）表示在语料库中$\hat{e}$和$\hat{f}$共现的次数，$\hat{f}$表示源端短语总共出现的次数。通过以上计算我们可以获得通用的短语表。通用的短语表的一般格式如下：“源短语|||目标短语|||概率值”。

与此类似，当短语确定了主题之后，我们应该按照公式（5）来重新计算带主题的各个短语对的概率。

$$P(\hat{e}|\hat{f},t_{ph}) = \frac{count(\hat{e},\hat{f},t_{ph})}{count(\hat{f},t_{ph})} \quad (5)$$

虽然训练语料库有20万对的平行句对，但按150个主题来分，在某些主题上存在严重的数据稀疏问题，为了缓和这种状况，本文采用了一个折中的方案，假定：$P(\hat{e}|\hat{f},t_{ph})\approx P(\hat{e}|\hat{f})$，即让短语对的主题概率约等于不区分主题的短语对概率，但是短语对的源端需要区分不同的主题。按照此约定，建立的新短语表（标记为PT_t），其存储格式变更为“源短语，主题标号|||目标短语|||概率值”。以短语“很香”为例，一般的气味芳香可以翻译成“very fragrant”，但在表示“菜很香”时，却翻译成“very appetizing”。因此在PT_t中，将区分为“很香，t_0|||very appetizing|||概率值”和“很香，t_1|||very fragrant|||概率值”。若待翻译文本属于主题T_0（餐饮），那么在解码器加载翻译候选项时，只加载主题属于T_0的翻译对，那么即可实现翻译过程中区分主题的目的。

3 解码器的预处理模块

与传统的解码器不同，基于主题的SMT的实现需要额外的两个处理：（1）自动推断出待翻译句子的主题信息；（2）对解码器进行修改，使得解码器能根据输入的源语言句子的主题从短语表中筛选出具有相同主题的翻译候选项，然后在此新的空间中进行搜索，得到符合主题的特定译文。

第一步借助LDA已训练好的模型，再联合公式（3）可以很方便地推断出测试句子的主题信息，然后该句子包含的所有可能的源短语都沿袭这一主题即可。但第二步的实施会有些困难。本文的实验以一个著名的开源的PBSMT系统—Moses[1]作为基准平台。目前的Moses为了兼顾多种翻译模型，其代码类越来越复杂，为避免破坏Moses的整体性和可靠性，本文建议用一种更简单的方法，不用修改解码器，同样可以实现基于主题的翻译。具体的实现包括短语表特殊处理和解码器预处理两个部分。

（1）短语表特殊处理

本文的翻译任务是针对汉英翻译，为了根据主题来区分短语表里短语对的中文端短语，作者对短语表进行如下特殊处理：

扫描新短语表PT_t中每个短语对，假设中文端短语包括n个词汇，表示成$W_{c1}W_{c2}\ldots W_{cn}$，对每个词汇进行唯一编号，在这过程中确保相同主题的同一词汇具有相同的词编号，这个处理将产生新的短语表PT_{t_new}；

在编号的同时，新建一个文本文件wordmap.txt记录下三项内容：词、主题号和词编号。

经过以上处理，每个短语对转变成了这样的格式—“源端词编号序列|||目标短语|||概率值”。

（2）解码器预处理模块

1: Algorithm 1：Pre-processing
2: Input：TopicModel *tm*， Sentence S_{test}
3: Output：S_{test_new}
4: T_x=Infer（tm，S_{test}）;
5: File *hm*=Open(“wordmap.txt”);
6: S_{test_new}=””;
7: For each word *w* in S_{test}
8: word_ id =find(*hm*, *w*, T_x);
9: if (word_id==null)
10: word_ id =find(*hm*, *w*);
11: S_{test_new}= S_{test_new}+ word_ id+” ”;
12:End For
13:Return S_{test_new}

解码器预处理模块的功能如Algorithm 1所示：输入的是已经训练好的主题模型tm和待翻译的句子S_{test}，输出的是经过特殊的处理的待翻译文本S_{test_new}。算法第4行首先获得使公式（4）取得最大值的主题Tx，然后打开存有词、主题号和词编号的文件wordmap.txt，用找到的词编号（第8行和11行）代替普通的词，因为词编号对应了主题信息，因此新产生的待翻译文本变成了待主题信息的词编号的序列，配合已处理好的短语表PT_{t_new}，S_{test_new}就能像一般的文本一样被解码器正常处理。值得注意的是，因为有些词组成的短语属于通用短语，比如“应该|||ought to”，应该不需要区分主题的，但这些短语在训练语料中仅在主题t_0中出现过，而包含了这些词的待翻译文本的主题却被断定为t_1，那么根据t1就找不到相应的词编号，因此也就翻译不出对应的部分，所以如果待翻译文本的给定某个词和主题号后，wordmap.txt里没有找到相应的编号，需要去掉主题的限制重新找词编号（对应算法9-10行），这个处理是为了尽可能多地翻译出通用短语。

4 实验

本文的实验使用的是FBIS语料作为训练数据（training data），2002 NIST MT 测评数据作为开发集（development data），2005 NIST MT测评数据作为测试集（test data）。表1显示的是语料的统计数据。

表1 语料统计表

语料		句子数	文档数
角色	名称		
Train	FBIS	239413	10354
Dev	NIST2002	580	100
Test	NIST2005	1082	100

本文以基于短语的统计机器翻译系统Moses作为基准系统，使用SRILM语言模型工具，英语新闻专线文本训练三元通用语言模型，并用GIZA++（Och and Ney，2000）对训练后的平行语料库进行双向的词对齐。本文汇报的4元BLEU值源自于自动评测工具Mteval的运行结果，评测时不进行大小写的严格匹配。

表2汇报了基准系统和本文建议的基于主题的PBSMT的翻译结果，从BLEU值的角度来看系统性能提升了0.44个百分点，这个提升虽然没有超出文献[2]的汇报的结果，但是本文通过了一种没有改变解码器的简单方法就达到了翻译过程区分主题的目的，证明了主题模型对SMT是有作用的，并且以后随着语料库规模的增大，相信由此方法获得的性能还有提升的空间。

表2 实验结果

System	BLEU on Test(%)
Moses	27.10
Ours	27.54

参考文献

[1]David M. Blei，Andrew Y. Ng，and Michael I. Jordan. Latent Dirichlet Allocation[J]. Journal of machine Learning Research，2003，pages 993 1022
[2]Peter. F. Brown，Stephen A. Della Pietra，Vincent J. Della Pietra，Robert L. Mercer，The Mathematics if Statistical Machine Translation：Parameter Estimation. Computational Linguistics[M]. 1993，19(2)：263-311
[3]Zhengxian Gong，Guodong Zhou，Liangyou Li. 2011. Improve SMT with Source-Side“Topic-Document”Distributions[pdf]. Machine Translation Summit XIII：496-501
[4]Koehn P，Och JF and Marcu D. Statistical Phrase-Based Translation. Proceedings of the HLT_NAACL：Human Languag Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003，127-133
[5]Zens R，Och F.J. Phrase-Based Statistical Machine Translation. 25th Annual German Conference on Artificial Intelligence，volume 2479 of Lecture Notes in Artificial Intelligence. 2002，2479:18-32
[6]Koehn P. Pharaoh：a beam search decoder for phrase-based statistical machine translation models. In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas. 2004，pages 115-124

注释

[1]http://www.statmt.org