文章编号：1003-0077(2013)02-0086-05

基于篇章上下文的统计机器翻译方法

于 惠，谢 军，熊 皓，吕雅娟，刘 群，林守勋

（中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190）

摘 要：上下文信息对于统计机器翻译(Statistical Machine Translation，SMT)中的规则选择是很重要的，但是之前的 SMT 模型只利用了句子内部的上下文信息，没有利用到整个篇章的上下文信息。该文提出了一种利用篇章上下文信息的方法来提高规则选择的准确性，从而提高翻译的质量。首先利用向量空间模型获得训练语料的文档和测试集中文档的相似度，然后把相似度作为一个新的特征加入到短语模型中。实验结果表明，在英语到汉语的翻译工作中，该方法可以显著提高翻译质量。在 NIST-08 和 CWMT-08 两个测试集上 BLEU 值都有显著的提高。

关键词：统计机器翻译；上下文信息；向量空间模型

中图分类号：TP391　　**文献标识码**：A

Improving Statistical Machine Translation by Document Context

YU Hui, XIE Jun, XIONG Hao, LV Yajuan, LIU Qun, LIN Shouxun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The present statistical machine translation (SMT) models only exploit the context information in a sentence and neglect that in the document which is more useful to find the correct translation. In this paper, we propose a new method of using the context of the whole document to improve the quality of SMT. We obtain the similarities between the documents of the training corpus and the documents of the test set using Vector Space Model. The similarity is then considered as a new feature and integrated into a phrase-based model. Large scale experiments show that our approach improves more than one point for NIST-08 and CWMT-08 in term of BLEU in English to Chinese translation task.

Key words: Statistical Machine Translation; context information; Vector Space Model

1 引言

广泛意义上来讲，统计机器翻译[1]也可以看作是利用规则来翻译的，例如，基于短语的翻译模型[2-4]利用的是短语翻译规则，基于句法的翻译模型[5-7]利用的是句法翻译规则。一般的翻译规则包含源端和目标端，他们可能是词、短语或者是句法树，这依赖于它们所属的模型。通常，一个源端可能对应着多个目标端，由于对齐错误或其他原因其中的一些规则可能是错误的。统计机器翻译的一个主要任务就是对于给定的规则源端选择出正确的目标端，这会直接影响翻译模型的质量。

传统的方法是利用在训练语料中估计的翻译概率来做规则选择，这种方法没有充分利用上下文信息。对于不同的上下文一个词可能有不同的含义。例如，mouse，根据不同的上下文环境可以翻译成“老鼠”或“鼠标”。如表 1 所示，在文档 1 的句子 1 中，mouse 的含义是“鼠标”，在文档 2 的句子 1 中 mouse 的含义是“老鼠”，这是根据 mouse 所在句子的上下文信息判断出来的。在文档 3 中，仅根据 mouse 所在的句子 2，不能判断出它的含义，但是再

收稿日期：2011-09-13　**定稿日期**：2012-03-16

基金项目：国家自然科学基金资助项目(61202216)；国家 863 计划资助项目(2011AA01A207)

作者简介：于惠(1983—)，女，博士研究生，主要研究方向为自然语言处理；谢军(1978—)，男，博士研究生，主要研究方向为自然语言处理；熊皓(1985—)，男，博士研究生，主要研究方向为自然语言处理。

加上句子 1 和句子 3 的信息，我们就可以判断出它的含义了。同样，要得到正确的翻译结果，解码器也需要当前句子或者周围句子的上下文信息。

表 1　mouse 在不同上下文的不同含义

文档 1	句子 1	The camp commander was directing a drill with the click of a ***mouse*** on the small screen of a computer.
文档 2	句子 1	There is a ***mouse*** running in the yard.
文档 3	句子 1	I use a mouse to choose documents from my files.
	句子 2	I like the ***mouse***.
	句子 3	The mouse and the keyboard are two parts of my computer.

本文中我们提出了一种利用整个文档的上下文信息来帮助规则选择的方法，首先我们利用向量空间模型[8-9]建立两个矩阵，一个是训练集的，一个是测试集的。矩阵中的每一行代表一个文档中每一个单词的出现次数(停用词表中的单词和出现次数太少的单词被过滤掉)。然后，利用这两个矩阵生成一个训练集和测试集的相似度矩阵。我们把相似度作为一个新的特征加入到 BTG 模型[4]中。实验表明，在英语到汉语的翻译工作中，我们的方法可以显著提高翻译质量。

2　相关工作

近年来，很多研究者利用上下文信息来提高机器翻译的质量。文献[4]提出了一种利用边界词信息预测短语重排序的方法，他的工作中把重排序看做分类问题，利用最大熵模型实现。文献[10-11]用上下文信息来帮助目标端规则的选择。他们利用非终结符的边界词信息建立最大熵模型。文献[12]提出了利用功能词选择源端规则的方法。文献[13]利用了相邻词的信息来计算语言模型的值。文献[14]提出了用上下文信息来调序的方法，所用的上下文信息包括源端的词汇化特征(边界词和相邻词)，词性标注特征和目标端的词汇化特征(边界词)。也有研究者利用上下文信息来做领域自适应[15]。

这些研究者的方法都利用了一部分上下文信息，但大都利用的是当前跨度内的信息或周围词的信息，并没有利用到整个篇章中的信息。本文中我们提出了一种利用整个篇章信息的方法，实验表明，我们的方法可以显著提高翻译质量。

3　基于篇章上下文的统计机器翻译

3.1　基于 VSM 的相似度矩阵

VSM 是信息检索中运用很广泛的一个模型，近年来，一些研究者也把它用到了统计机器翻译中[16]。在 VSM 中，每一篇文档表示为一个向量，向量中的每一个元素对应文档中的一个词，如果一个词在文档中出现，那么在向量中对应的就是一个非零的值，否则它对应的值就是零。第 j 个文档可以用向量 D_j 表示，$D_j = < W_{1j}, W_{2j}, \cdots, W_{nj} >$。其中，$W_{kj}$ 表示第 k 个词的权重。在计算权重时我们用了 TF-IDF (Term Frequency-Inverse Document Frequency)[9]。

TF-IDF 是一种统计方法，TF 表示一个词在一个文档中的出现次数，当然经常被归一化。第 i 个词的 TF 可以表示为：$tf_{i,j} = \dfrac{n_{i,j}}{\sum_k n_{k,j}}$，$n_{i,j}$ 是第 i 个词在第 j 个文档中的出现次数，分母就表示在第 j 个文档中所有所有词的出现次数之和。IDF 是评价一个词在一个语料中重要性的指标。第 i 个词的 IDF 可以表示为：$idf_i = \log \dfrac{|D|}{|d|}$，$|D|$ 表示一个语料中文档的个数，$|d|$ 表示第 i 个词出现过的文档的个数。这样，$tfidf_{i,j} = tf_{i,j} \cdot idf_i$，利用 TF-IDF 可以过滤掉常见的词，并且保留重要的词。

在计算两个文档间的相似度时，我们利用了 cosine angle[9]。

$$sim(D_1, D_2) = \frac{\sum_k (W_{1k} \times W_{2k})}{\sqrt{\sum_k (W_{1k}^2 \times \sum_k W_{2k}^2)}} \quad (1)$$

其中，W_{1k} 和 W_{2k} 分别表示第 k 个词在文档向量 D_1 和 D_2 中的权重。

本文中，我们用了源端语言的篇章上下文信息，我们用的所有语料都是按照篇章信息分好的文档，每个文档就是一个篇章。首先，扫描训练语料并且记录下其中出现的所有词，去掉停用词后保存在一个向量中，我们称为 words-vector。然后建立一个 $M \times H$ 的矩阵 A，每个元素的初始值都为 0。M 是训练语料中文档的个数，H 是 words-vector 的长度。A 中的每一行表示一个文档。然后扫描训练语料中每一个文档，如果一个词在此文档中出现过，并且也在 words-vector 中(即这个词不是停用词)，矩

阵 A 中对应元素的值加 1，最后我们可以得到包含所有文档上下文信息的矩阵 A。

对于测试集，建立一个 $P\times Q$ 的矩阵 B。P 是测试集中文档的个数，Q 是在训练语料或测试语料中出现但不在停用词中出现的词的个数。用与得到 A 相同的方法得到 B。

为了保证从训练语料中得到的文档向量和测试语料中得到的文档向量长度相同，我们扩展矩阵 A 的列的个数由 H 到 Q，新扩展出的元素的值都为 0。在矩阵 A 和 B 中分别利用 TF-IDF，得到每个词对应的权重。然后用 cosine angle 对矩阵 B 的每一行和矩阵 A 的每一行计算相似度，可以得到相似度矩阵 C，矩阵中的每一个元素 (i,j) 是由矩阵 B 中的第 i 行和矩阵 A 的第 j 行得到的。在解码过程中，我们会用到这个相似度矩阵。

3.2 带有文档 ID 的规则

我们所用的语料都是带有文档标记的，即每个文档都有一个 ID 标记，比如 1,2,3…在抽规则的同时，记录下每个规则所在的文档的 ID。有的规则可能同时出现在几个文档中，这样的规则就有多个 ID。比如一条规则表示为：

mouse ||| 鼠标 0.1 0.1 0.1 0.1 ||| 1 3

这个规则中最后的 1 和 3 表示它出现在第 1 个和第 3 个文档中。

3.3 解码过程

对数线性模型的框架使得我们很容易在解码器中加入新特征，如式(2)所示。

$$e=\arg\max_{e_1^I}\{\Pr(e_1^I\mid f_1^J)\}$$
$$=\arg\max_{e_1^I}\left\{\sum_{m=1}^{M}\lambda_m h_m(e_1^I,f_1^J)+\lambda_{simi}h_{simi}(e_1^I,f_1^J)\right\} \quad (2)$$

我们的目的是找到概率最大的 e，λ 代表特征函数，h 是特征函数对应的权重，$\lambda_m(m=1,2,\cdots M)$ 为原有的特征，λ_{simi} 为新加入的特征。

我们用的基准系统是基于 BTG 的解码器。该解码器用的是 CKY 形式的解码算法。为了利用整个文档的上下文信息，我们把相似度作为一个新的特征加入到基于 BTG 的解码器中。

对于给定的一个句子 S，它所在文档的 ID 为 d。首先，用规则表中的规则初始化 chart 图。对每个 span(i,i)，在规则表中可以找到有用的规则 P_k $(k=1,2,\cdots K)$，有用规则的个数为 K，有用规则指的是源端和 span(i,i)一致的规则。我们用 ID_{kt} 表示规则 R_k 所在的文档 ID，$t=1,2,\cdots,T$，T 是规则 R_k 所在的文档 ID 的个数。相似度特征的分数可以在相似度矩阵 $C(d,ID_{kt})$ 中找到，代表矩阵 C 中第 d 行，第 ID_{kt} 列的元素的值。如果 T 大于 1，需要得到一个最终的分数，实验中分别用了它们的最大值和平均值，所以最终的相似度分数可以表示为：

$$\max(C(d,ID_{kt}))\text{ 和 }\text{average}(C(d,ID_{kt}))=\frac{\sum_{t=1}^{T}C(d,ID_{kt})}{T},\text{ 其中 } t=1,2,\cdots,T。 \quad (3)$$

对于 span(i,j)，$i\neq j$，如果一条规则不能覆盖整个 span，为了得到 span(i,j)的候选翻译，需要把 i,j 间的子 span 进行组合，这时 span(i,j)的相似度分数不能直接在相似度矩阵中得到，因为短语表中可能没有对应的规则。我们采用了比较简单的方法，例如，取两个子 span 相似度特征分数的平均值或最大值。

得到 span(i,i)和 span(i,j)相似度特征分数都有平均值和最大值两种方式，所以有四种不同的组合方式，可以得到四个解码器，如表 2 所示。表 2 中“init”表示解码的初始化过程，此过程中相似度特征的值可以取最大值或平均值；“cat”表示解码时两个 span 的拼接过程，此过程中相似度特征的值也可以取最大值或平均值。解码器中其他特征的计算和基准系统相同。在 CKY 解码的最后，我们可以得到整个句子的候选翻译。

表 2 不同组合产生的四种解码器

init \ cat	average	maximum
maximum	BTG_sim_1	BTG_sim_4
average	BTG_sim_2	BTG_sim_3

3.4 时间复杂度

计算训练语料中文档和测试语料中文档相似度的时间复杂度是 $O(M\times P\times Q)$。M 是训练语料中文档的个数，P 是测试语料中文档的个数，实验中它们的值如表 3 所示，Q 是过滤后训练语料或测试语料中出现的词的个数，实验中其值为 40 000。

4 实验

4.1 实验数据

我们的实验是在英语到汉语的翻译工作中做

的。双语平行语料包括 LDC2003E14，LDC2005T06，还有一部分 LDC2004T08，共 100 万平行句对，用 GIZA++做词语对齐，并且使用“grow-diag-final”的启发式方法。语言模型是 5 元 giga 语言模型。

我们在 HTRDP-MT2005 上做的最小错误率训练，测试集使用 CWMT2008 和 NIST2008。翻译质量的评价指标是 BLEU-4，测试工具是 mteval-v11b.pl。实验中用的语料都是新闻领域的，详细信息如表 3 所示。

表 3　实验中所用语料信息

	句子个数	文档个数	句子的平均长度
Training corpus	1 039 140	54 637	27
HTRDP MT2005	494	29	28
CWMT2008	1 000	34	25
NIST2008	1 859	129	25

4.2　实验结果

可以通过 BLEU 值证明我们方法的有效性，各个系统的结果见表 4。在表 4 中，N 表示阈值，即如果一个词在训练语料的每个文档中出现次数不超过 N 次，就过滤掉这个词。在 N 为 1 时，只有 BTG_sim_4 的结果比较好，对两个测试集分别提高了 0.95 和 1 个点。N 为 2 时，只有 BTG_sim_3 低于基准系统的结果，其他的都高于基准系统。N 为 3 时，几乎所有的结果都低于基准系统的结果。

表 4　两个测试集的 BLEU 值

N	decoder	NIST08	CWMT08
	baseline	32.22	31.19
1	BTG_sim_1	27.49	29.28
	BTG_sim_2	29.28	31.51
	BTG_sim_3	30.17	32.52
	BTG_sim_4	33.17	32.19
2	BTG_sim_1	32.82	31.78
	BTG_sim_2	33.66	32.42
	BTG_sim_3	31.49	30.26
	BTG_sim_4	33.43	32.49
3	BTG_sim_1	31.13	30.01
	BTG_sim_2	32.02	31.06
	BTG_sim_3	32.31	30.90
	BTG_sim_4	30.74	29.74

从这些结果我们可以得出一个结论：如果一个词仅出现了一次或两次，它可能是一个干扰，应该过滤掉，所以 N 为 2 时效果最好；如果一个词出现了 3 次，这个词是有用信息的可能性比较大，应该保留。在 N 为 2 时（最好的阈值），BTG_sim_2 和 BTG_sim_4 的效果最好，即解码的过程中初始化和连接相邻 span 时都取平均值或都取最大值时效果最好。

我们通过两个例子比较一下新解码器和基准系统的翻译结果，见表 5 和表 6，其中的两个句子来自 NTST08。

表 5　单词 orders 的翻译结果

Sentence	In London, rolls-royce soared 3.45 percent to 554 pence after the engine maker said it had won **orders** across all areas of its businesses and increased its global customer base significantly during the second quarter.
context 1	pence, businesses, customer
context 2	price, dollars, shares, euros, markets
Reference	在伦敦市场，引擎制造商劳斯莱斯宣布其各项业务均获得**订单**，并且二季度全球客户数量增加，其股票狂飙 3.45%，报 554 便士。
BTG-based model	在伦敦，劳斯莱斯劲升百分之三点四五五百五十四 pence 后发动机制造者说，它得到了**令**其业务涉及所有领域的全球的客户基础，二季度大幅增加。
BTG+similarity	伦敦劳斯莱斯劲升百分之三点四五五百五十四 pence 后发动机制造者说，它已获得**订单**，其业务涉及所有领域的二季度大幅增加其在全球的客户基础。

表 6　单词 fans 的翻译结果

Sentence	a wonderland for ' simpsons ' **fans**
Context 1	—
Context 2	character, movie, actor, episodes, tv
Reference	“辛普森一家”**迷**的乐园
BTG-based model	的一个“simpsons 仙境”**电风扇**的
BTG+similarity	“simpsons 仙境”**迷**的

在表 5 和表 6 中，第一行 sentence 表示词 orders 和 fans 所在的句子；第二行 context 1 表示当前句子中出现的有用的单词；第三行 context 2 表示当前文档中（除了当前句子）出现的有用的单词，一个文档中有用的单词可能有很多，这里只举了其中几个例子。有用单词指的是对我们关心的单词 or-

ders 和 fans 的翻译能起到帮助作用的词。第四行 reference 是第一行 sentence 的参考译文；第五行 BTG-based model 是基准系统的翻译结果；第六行 BTG+similarity 是加入相似度特征的解码器的翻译结果。

在第一个例子中，对于单词 orders，基准系统的翻译结果是“令”。在第二个例子中，对于单词 fans，基准系统的翻译结果是电风扇。我们可以看出这两个结果是错误的，但是加入相似度后的解码器找到了正确的结果“订单”和“迷”。在当前句子或当前文档中有很多上下文信息，相似度特征就是根据这些上下文信息得到的。通过这两个例子我们可以得出结论，在加入相似度特征后的解码器可以找到正确的规则，从而提高翻译质量。

5 总结和展望

本文中，我们提出了一种提高规则选择准确性的方法，利用整个篇章的上下文信息来提高翻译质量。首先我们利用向量空间模型获得训练语料的文档和测试集中文档的相似度，然后把相似度作为一个新的特征加入到短语模型中。实验表明，在英语到汉语的翻译工作中，我们的方法可以显著提高翻译质量。最后，我们对计算相似度的时间复杂度进行了分析。

下一步，我们会从两个方面继续这个工作。第一，在模型中加入相似度特征的方法是通用的，我们可以把它加入到层次短语模型或其他模型中；第二，由于很多语料是没有分好文档的，例如，WMT 的语料，不能直接使用，所以我们可以利用语料自动分类方法，把语料按照内容划分为多个文档。

参考文献

[1] 刘群. 统计机器翻译综述[J]. 中文信息学报，2003，17(4)：1-12.

[2] Och F J，Ney H. Improved statistical alignmentmodels [C]//Proceeding of ACL，2000：440-447.

[3] Koehn P，Och F J，Marcu D. Statistical phrase-based translation[C]//Proceeding of ACL，2003：48-54.

[4] Xiong D，Liu Q，Lin S. Maximum entropy based phrase reordering model for statistical machine translation[C]//Proceeding of ACL，2006：521-528.

[5] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceeding of ACL，2005：263-270.

[6] Liu Y，Liu Q，Lin S. Tree-to-string alignment template for statistical machine translation[C]//Proceeding of ACL，2006：609-616.

[7] Galley M，Graehl J，Knight K，et al. Scalable inference and training of context-rich syntactic translation models[C]//Proceeding of ACL，2006：961-968.

[8] Salton G，Wong A，Yang C S. A vector space model for automatic indexing [J]. Communications of the ACM，1975，18(11)：613-620.

[9] Salton G，McGil，M J. Introduction to Modern Information Retrieval [M]. McGraw-Hill New York 1983.

[10] He Z，Liu Q，Lin S. Improving statistical machine translation using lexicalized rule selection[C]//Proceeding of ACL，2008：321-328.

[11] Liu Q，He Z，Liu Y，et al. Maximum entropy based rule selection model for syntax based statistical machine translation[C]//Proceeding of ACL，2008：89-97.

[12] Setiawan H，Kan M Y，Li H. Topological ordering of function words in hierarchical phrase-based translation[C]//Proceeding of ACL，2009：324-332.

[13] Shen L，Xu J，Zhang B. Effective use of linguistic and contextual information for statistical machine translation[C]//Proceeding of ACL，2009：72-80.

[14] He Z，Meng Y，Yu H. Maximum entropy based phrase reordering for hierarchical phrase-based translation[C]//Proceeding of ACL，2010：555-563.

[15] 曹杰，吕雅娟，苏劲松，等. 利用上下文信息的统计机器翻译领域自适应[J]. 中文信息学报，2010，24(6)：50-56.

[16] Chen B，Foster G，Kuhn R. Bilingual sense similarity for statistical machine translation[C]//Proceeding of ACL，2010：1-10.