句法信息与短语模型相融合的统计机器翻译

宋鼎新，黄德根
大连理工大学　计算机科学与技术学院　辽宁大连　116024
E-mail：songdx@mail.dlut.edu.cn　huangdg@dlut.edu.cn

摘要：基于短语的统计机器翻译系统中使用的短语并不限制句法边界，而基于句法的统计机器翻译则是根据句法分析利用语言学意义上的短语进行翻译。本文首先分别从中英文句法分析树中获取句法短语，然后使用短语表作为媒介，以一致性短语原则作为辅助形成双语句法短语对。最后采用加入短语表和增加句法特征两种方法将其与短语翻译模型的原有短语表相结合，以改善机器翻译质量。实验结果表明：双语句法短语能够提高基于短语的统计机器翻译质量，其中增加句法特征方法效果较为明显，BLEU 值提高 0.56。

关键字：统计机器翻译、双语句法短语、短语翻译模型

Integrating Syntactic Information into Phrase-Based Statistical Machine Translation

Song Dingxin and Huang Degen
School of Computer Science and Technology,
Dalian University of Technology, Dalian, Liaoning 116024, China
E-mail：songdx@mail.dlut.edu.cn　huangdg@dlut.edu.cn

Abstract: *The phrases in the phrase-based statistical machine translation are not grammatical ones, while the phrases in syntax-based statistical machine translation are grammatical phrases getting from the syntactic analysis. this paper proposes to integrate the grammatical phrases in the phrase-based machine translation in order to improve the performance of the Chinese-English machine translation system. First, the grammatical phrases are extracted from the syntactic trees in the two languages. Based on the phrase translation table and the principle of consistent phrase bilingual phrase translation pairs are formed. Finally, add the pairs and syntactic features in the phrase table of phrase-based machine translation system. The experiment result shows that bilingual grammatical phrases can be used to improve the statistical machine translation system and integrating the syntactic features in the system can improve the performance of the system significantly(0.56 BLEU over the baseline system)*

Keywords: *Statistical Machine Translation, Bilingual Syntactic Phrases, Phrase Translation Model*

1　引言

目前，统计机器翻译较为流行的两个发展方向分别是基于短语[Zens，2002][Koehn，2003]或层次短语[Chiang 2005，2007]的机器翻译和基于句法的机器翻译（包括树到串[Liu，2006]、串到树[Yamada，2001]、树到树[Quirk，2005]等方法）。前者更多的利用了形式化句法的信息，其短语表中的短语不要求必须符合句法边界；而后者则是利用句法分析，从而产生不同的句法规则来进行翻译，这种方式得到的是符合语法边界的短语，也就是语言学意义上的短语。两种方式都取得了不错的效果，但同时又有各自的弱点，短语翻译系统受到分词（中文）错误、词对齐错误、短语表规模等多种因素影响导致最后的翻译结果难以达到实际应用的标准，而句法翻译系统受制于句法分析的准确性和自然语言的不规范性等等因素，也存在着很多缺陷。因此，近些年来很多学者都在尝试将两种翻译模型的优点相互结合，从而改善统计机器翻译的性能。结合方法一般分为两种：一种是将短语翻译系统中得到的短语表经过过滤等处理后，加入到句法翻译系统的过程中去；另一种思路则是恰恰相反，在句法

翻译系统过程中获取有用的句法信息后对短语翻译过程加以约束，从而改善翻译质量。

John Tinsley[Tinsley，2007，2009]等提出了一种独立于语言对的句法树子树对齐工具，能够自动对齐短语级句法结构；Greg Hanneman[Hanneman，2009]认为，基于句法的机器翻译的一个关键问题就是如何将短语翻译模型中的更多信息融合进来，尤其是不严格对应句法成分的短语，从而引入更多的上下文信息来辅助调序工作，同时在John Tinsley工作的基础上提出了新的融合方法并进行了实验证明；Alon Lavie[Lavie，2008]使用平行语料、词对齐和短语句法分析树等资源首先将树形结构进行节点对齐，然后抽取所有对齐子树形成句法短语表，最后通过这些短语在句中的位置确定同步上下文语法规则；[丁鹏，2013]中提出了句法短语的概念，使用基于EM算法的双语句法短语抽取算法从句法树中获得双语句法短语，并在基于短语的统计机器翻译系统中证明其对统计机器翻译性能有显著的改善作用。

考虑到目前基于短语的机器翻译效果要好于句法翻译模型，本文采用短语翻译模型作为基线系统，从句法树中获取句法意义上的短语，并利用短语表信息和一致性短语思想获得双语句法短语对，然后再使用加入短语表和新增句法特征方式将其融入短语表，最终通过评价机器翻译结果来证明句法短语的有效性。

2 获取双语句法短语

双语句法短语如何获取本身就是一个课题。目前基于短语的统计机器翻译的训练过程和参数调优都是在平行语料的基础上进行，也就是句子级的双语对齐。而为了获取更多的粒度更小的句法信息，就需要获取短语级别的对齐信息，也就是符合句法边界的双语短语对。比如名词短语，动宾结构，介宾结构等等。本文利用句法分析树的分析结果，采用一种较为便捷有效的方式来获取双语句法短语。其主要步骤如下：

1.在单语句法分析树中提取句法短语。

中英文句法树表示方式分别为：

((IP (NP (NP (NP (NN 科学) (NN 规划)) (NP-M (ADJP (JJ 既有)) (NP (NN 指导性)))) (NP (PU ,))) (VP (ADVP (AD 又)) (VP (VP (VE 有) (NP (NN 约束力))) (VP (PU .))))))

((S (NP (DT a) (NP-M (JJ scientific) (NN plan))) (S-M (VP (VBZ has) (NP (DT both) (NP-M (NP (PRP$ its) (NP-M (VBG guiding) (NN function))) (NP-M (CC and) (NP (JJ binding) (NN force)))))) (. .))))

每一组括号内的单词或者短语都作为句子的某个成分，也就是说符合句法边界，即我们需要获取的句法短语。但由于句法分析是对中英文分别进行，因此得到的句法树的内部结构并没有对应关系，我们通过后面的步骤进行短语对齐。

2.在短语翻译表中进行双语匹配。

基于短语的统计机器翻译的训练过程会产生短语翻译表，这是通过平行语料在词对齐的基础上经过短语抽取过程后的输出结果，其中记录了源短语、目标短语、正反向翻译概率、词对齐等信息。在步骤1中获得的单语句法短语可以利用短语表的信息进行对应。例如，对一个中文句子中的所有句法短语逐一在短语表中进行搜索，如果此短语存在于短语表中，并且在其对应的多个英文翻译中同时存在英文一侧的句法短语，那么此翻译对即为双语句法短语对。如果中文句法短语匹配失败，则进入步骤3。

3.利用一致性短语思想作为补充。

经过步骤2的搜索匹配后，剩余少量匹配失败的短语。如果存在符合以下条件的短语对，那么也将视为双语句法短语对：中文短语中的所有词汇在词对齐中所对应的词汇都存在于某个英文短语之中，并且反之亦然。如果不满足此条件，则放弃匹配。另外，在匹配过程中需要限制短语长度，一般设置为与短语表中默认短语词汇长度一致。

4.双语句法短语对的后处理。

通过上述几个步骤得到双语句法短语对后，需要对其进行一些约束。由于基于短语的统计机器翻译中将标点符号视为词汇，与普通词汇做等同处理。因此在短语表中存在了大量含有标点符号的短语。而句法分析过程中也会将逗号或句号视为NP（名词短语）、VP（动词短语）或其他句法成分的一部分。但是我们希望得到的句法短语对并不包含句末的标点（短语内部的标点符号有时是必要的，比如：铁木尔·达瓦买提翻译为Tomur Dawamat），而且经常会出现中英文单侧带标点的情况，因此本文对短语前后带有逗号、句号、冒号等与短语内部无关的标点时，直接将其过滤掉。

3 双语句法短语与统计翻译模型的融合

在获取到双语句法短语后，本文采用两种方式将其与基于短语的统计机器翻译模型相融合。一种方法是直接将获取到的句法短语对加入到短语表中，而另一种方法则是利用对数线性模型增加一个句法短语特征，使这些更符合句法结构的短语在解码过程中发挥更大的作用。下面分别介绍这两种方法。

3.1 加入短语表

这种方式较为简单，将上一步中获得的双语句法短语对加入原系统中的短语表即可，引入句法分析方法中产生的新短语，弥补词对齐的不足和启发式短语抽取方式无法获得的短语对。

3.2 新增句法特征

基于短语的统计机器翻译系统中已存在正反向短语翻译概率特征、正反向词汇化概率特征、短语惩罚特征、调序特征等等。这些特征通过对数线性模型结合在一起，经过开发集的参数调优，最终共同决定解码器对翻译候选的选择。因此，从理论上分析这种方法要比直接加入短语表的方法更直接的影响最终的翻译结果。

在东北大学开发的开源机器翻译系统Niutrans[Xiao,2012]中提供了便捷的方法使开发者在对数线性模型中方便的增加新特征。只需要修改配置文件并在短语表中增加相应的信息即可。我们将短语表非句法短语的句法特征均置为0，所有句法短语对均置为1。包括原短语表中存在且能够与句法短语对匹配的短语和原短语表中不存在的新增短语。而后使用开发集对所有特征进行参数调优，以获得最后的翻译模型。

4 实验结果

本文实验使用的数据为Niutrans翻译系统提供的实验语料。其中训练语料为10000句对的平行语料，并为基于句法的翻译系统提供了相应的句法树库。开发集和测试集句子数量分别为400句和1000句。由于此语料为新闻领域语料，因此句子平均长度（单词数）较长是一个比较明显的特点。中文句子长度约为25个词，而英文则长达近33个词，这给句法分析带来了较大难度，尤其是目前正确率相对较低的中文句法分析。

表 1. 实验语料统计

	汉语	英语
训练集句子个数	100000	100000
训练集句子词数	24.9	32.7
开发集句子个数	400	400
开发集句子词数	25.3	33
测试集句子个数	1000	1000
测试集句子词数	24.7	32.8

本文采用的基线系统是东北大学提供的开源机器翻译系统Niutrans，支持多个翻译模型包括基于短语的模型、基于层次短语的模型和基于句法（树到串，串到树，树到树）的模型，自从2012年发布以来，得到了广泛的应用。本文采用Niutrans中基于短语的统计机器翻译模型，其良好的特征扩展功能使我们可以方便的引入新特征，在无需改动模型的情况下即可进行对比实验。最后采用目前通用的BLEU评分[Papineni，2002]标准对翻译效果进行评价。

第三部分介绍的两种融合方式的实验结果如表2所示。其中Baseline代表未做改动的Niutrans基线系统，Phrase代表将双语句法短语加入短语表后得到的翻译结果，Feature代表为短语翻译模型增加句法短语特征后得到的翻译结果。从表中结果可以看出，两种双语句法短语的应用方式都能够对短语翻译模型的效果有所改善，其中引入新特征方式的效果较为明显，与基线系统相比BLEU值提升达到0.56。

表 2. 实验结果对比

模型	BLEU 评分
Baseline	0.2153
Phrase	0.2186
Feature	0.2209

对比于Phrase方法，Feature方法能够有较好的性能。其原因主要是在基于短语的统计机器翻译系统中，双语句法短语起到了更明显的作用。双语句法短语与原短语表中的短语数量相距甚远，单纯的将其放在短语表中很难突出其优质短语对的优势，所有短语表中的短语都做等同处理，因此Phrase方法的性能略微提升仅仅是归结于短语数量上，而非短语质量。而Feature方法明确了句法特征这一短语属性，并且此属性在开发集的参数调优过程中发挥了自身作用，使之作为一个参考因素与其他特征共同决定解码过程中的模型对翻译候选的选择。同时也证明了符合句法边界的短语在短语翻译模型中能够改善翻译质量，虽然Koehn曾经论证过单纯使用语法意义上的短语在统计机器翻译中并不可行，但是使用这类短语配合现有的短语翻译模型，使翻译结果更符合人的语言表达习惯，相信句法短语还是大有可为的。

在实验过程中，如何确定短语表中不存在的句法短语的翻译概率是一个比较棘手的问题。经过限制长度和标点符号处理之后，虽然新短语比例较低，但是缺少相应翻译概率和调序概率使之很难起到本身应有的作用。究其原因，原本不存在于短语表中的短语是因为传统的启发式短语抽取方法不能将其抽取出来，有可能是词对齐错误引发，也有可能是句法分析方法与短语模型的差异性导致。既然不能在短语产生过程中确定其各项概率参数，我们也进行了下面的补充实验，将这部分短语直接加入训练语料，其内部对齐信息加入词对齐文件，从而得到表3中的实验结果。结果表明，将新短语加入训练语料后效果并不明显，几乎与原新特征方法效果一致。说明新短语的概率确定方法还需要进一步研究。

表 3. 补充实验结果对比

模型	BLEU 评分
Feature	0.2209
Feature+Training	0.2225

5 结论与展望

本文采用了一种简单有效的双语句法短语对的获取方法，将得到的句法短语对与基于短语的统计机器翻译系统相融合，使用加入短语表和使用新特征两种方式证明句法短语对短语翻译模型的改进作用，并探索了新短语引入短语表时确定概率参数的方法。实验结果表明，句法短语作为句法特征出现在短语翻译模型中，对翻译质量的改进较为明显，BLEU值提高

0.56。可见，在基于短语的统计机器翻译过程中加入辅助句法信息有利于翻译效果的提升。

本文进行的几个实验从理论上讲较为简单，引入了句法短语但方式比较粗糙，其概念相对笼统。下一步的努力方向可以考虑对句法短语进行细化分类，在短语模型中引入更多的句法特征，使参数优化过程可以更多的体现出句法信息，以弥补原模型中跨越短语边界的情况出现在最终的翻译结果中。

参考文献

Richard Zens, Franz Josef Och, Hermann Ney. 2002. Phrase-based statistical machine translation[C]. *Advances in Artificial Intelligence*. Springer Berlin Heidelberg. Pages 18-32.

Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. Statistical phrase-based translation[C]. In *HLT-NAACL: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada. Pages 48-54.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation[C]. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, University of Michigan, Ann Arbor. Pages 263-270.

David Chiang. 2007. Hierarchical phrase-based translation[C]. Computational Linguistics. Pages 202-228.

Yang Liu, Qun liu, and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation[C]. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA. Pages 609-616.

Kenji Yamada, Kevin Knight. 2001. A Syntax-Based Statistical Translation Model [C]. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France. Pages 523-530.

Chris Quirk and Arul Menezes and Colin Cherry. 2005. Dependency Treelet Translation : Syntactically Information Phrasal SMT[C]. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor. Pages 271-279.

John Tinsley, Mary Hearne and Andy Way. 2007. Exploiting parallel treebanks to improve phrase based statistical machine translation[C]. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, Bergen, Norway. Pages 175–187.

John Tinsley, Mary Hearne and Andy Way. 2009. Parallel Treebanks in Phrase-Based Statistical Machine Translation[C]. In *CICLING 2009: the 10th International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico.

Greg Hanneman, Alon Lavie. 2009. Decoding with syntactic and non-syntactic phrases in a syntax-based machine translation system[C]. In *Proceedings of the third workshop on syntax and structure in statistical translation at the 2009 meeting of the North-American chapter of the association for computational linguistics* (NAACL-HLT-2009). Boulder, CO.

Alon Lavie, Alok Parlikar and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora[C]. In *Proc. of the ACL-08: HLT SSST-2 Workshop. Association for Computational Linguistics*. Columbus, Ohio. Pages 87-95.

丁鹏. 基于双语句法短语的统计机器翻译研究.大连理工大学, 2013.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. 2012. NiuTrans : An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation[C]. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea. Pages 19-24.

Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002. BLEU: a method for automatic evaluation of machine translation[C]. In *Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Pages 311-318.