汉语虚词相关的短语边界在句法分析中的应用研究

冯晓波，穆玲玲*，昝红英, 张坤丽

（郑州大学信息工程学院 河南 郑州 450001）

摘要：本文通过在构建汉英双语树库的工作中发现，包含虚词成分的短语边界错误能够影响到句法分析结果，因此本文使用基于规则、CRF 模型和 CNN 模型进行虚词相关的短语边界识别研究。针对 CTB8.0 的实验结果表明，基于 CNN 模型的短语边界识别效果最好，平均准确率达到了 75.63%。本文另外提出了一种基于虚词相关的短语边界的句法分析模型 Phrase_Based Parser，在 CTB8.0 上的句法分析结果的平均准确率达到了 80.78%。

关键词：CNN；短语边界识别；短语结构句法分析；双语树库

Studies on Boundary Identification of Phrases with Function Words and Its Application in Syntactic Parsing

Abstract: In the construction of a Chinese English bilingual treebank found that the error of phrase boundary identification that contains functional words can affect the result of syntactic analysis . Therefore, In this paper, we use the rule based approach, CRF model and CNN model for the identification of phrase boundaries. Experimental results on CTB8.0 data showed that the CNN model performs best, the average accuracy reached 74.97%.In addition, this paper presents a parser named Phrase_Based Parser that based on phrase boundary identification on functional words.The average accuracy rate of syntactic analysis on the CTB8.0 dataset has reached 80.78%.

Key words:CNN;BoundaryRecognition;Syntactic Parsing

1 引言

句法分析是根据给定的语法，自动地推导出句子语法结构，即句子所包含的句法单元和这些句法单元之间的关系。句法分析是自然语言处理的基本问题之一，它是机器翻译、舆情分析和自动文摘等应用系统的基础，句法分析的准确性直接影响应用系统的效果。

在构建汉语双语短语树库的工作中发现自动汉语句法分析错误主要集中于与虚词相关的并列结构、介词短语和“的”字短语边界错误。本文提出一种基于短语边界的句法分析模型 Phrase_Based Parser。首先利用CNN等方法自动识别虚词相关的短语边界，然后利用边界知识进行结构短语句法分析。

本文其余部分组织如下：第 2 节介绍相关工作，主要介绍双语树库的句法树错误分析和虚词相关的短语边界识别，第 3 节提出了基于短语边界的句法分析模型 Phrase_Based Parser，第 4 节介绍 Phrase_Based Parser 在 CTB8.0 和双语树库的实验，第 5 节是总结和展望。

2. 相关工作

2.1 汉英双语树库

汉英双语树库[1]面向机器翻译，其中原始语料是从全国机器翻译研讨会（CWMT）汉英新闻训练语料选取的 4000 句汉英句对，其中汉语句长最长 140 个字（96 个词），对应英文句子 79 个词；最短 18 个字（10 个词），对应英文句子 12 个词。构建双语树库时先对原始语料进行预处理，然后对预处理后的语料进行分词，分词结果按照宾州树库[2]的分词准则进行人工校对，将校对后的分词结果进行自动句法分析得到句法树对，对句法

*基金项目：本文承国家自然科学基金项目（61402419）、国家社会科学基金项目（14BYY096）、国家重点基础研究发展计划（2014CB340504）、河南省科技厅基础研究项目（142300410231，142300410308）、河南省高等学校重点科研项目计划（15A520098）支持资助。

*通讯作者：穆玲玲，副教授，iellmu@zzu.edu.cn

树对按照宾州树库标注体系进行人工校对后得到符合宾州树库标注体系的双语句法树库。

本文对双语句法树库中的汉语树库进行了统计分析，发现汉语树库中共有 101882 个词，其中包含介词，连词和助词“的”等虚词 15733 个，包含上述虚词的句子 3857 句，其中部分高频词相关的短语边界准确率如表 2.1、表 2.2 和表 2.3 所示。介词短语，连词相关短语和“的”短语等虚词相关的短语边界错误导致的错误句法树一共为 3053 句，占 76.32%。

表格 2.1 双语树库中介词（频次>50）分布和介词短语边界准确率

介词	频率	准确率	介词	频率	准确率
在/P	1302	86.25%	通过/P	131	71.76%
对/P	377	81.70%	由/P	127	86.61%
为/P	188	81.38%	由于/P	78	75.64%
从/P	167	84.43%	因为/P	73	75.34%
于/P	152	82.89%	到/P	71	64.79%
向/P	147	89.11%	同/P	63	58.73%
以/P	143	84.61%	根据/P	59	72.88%
与/P	136	78.68%	比/P	54	75.93%
关于/P	135	73.33%	按/P	53	86.92%

表格 2.2 双语树库连词（频次>50）分布和连词相关短语边界准确率

连词	频率	准确率
和/CC	2588	0.02%
或/CC	522	2.68%
并/CC	204	0.98%
及/CC	191	0.52%
以及/CC	152	1.32%
至/CC	120	0.00%
与/CC	119	1.68%
而/CC	69	0.00%

表格 2.3 双语树库助词“的”（频次>50）分布和“的”字短语边界准确率

助词	频率	准确率
的/DEC	3650	76.85%
的/DEG	3445	76.20%

2.2 虚词相关的短语边界识别

从 2.1 的分析可知，虚词相关短语的边界识别的准确性将会影响到句法分析结果的准确率。因此在进行句法分析之前，提高汉语虚词相关的短语边界自动识别的准确性具有重要的意义。

文献[3]利用虚词用法知识库[4]的连词用法规则进行连词相关短语边界的识别，在 2000 年 1 月《人民日报》分词与词性标注语料上的实验结果表明，准确率达 48.67%。

文献[5]提出利用 CRF 模型进行介词短语边界识别，在语料 2000 年 1 月《人民日报》分词与词性标注语料上实验结果表明，准确率达 80.46%。

2.2.1 基于 CNN 的虚词相关短语识别

卷积神经网络[6]（Convolutional Neural Network,CNN）模型在语音分析和图像识别领域取得了较好的结果。本文使用基于 CNN 模型的方法来进行短语边界识别。该方法构建一个五层的 CNN 模型，如图 2.1 所示。包括：输入层，卷积层，池化层，Tanh 层和输出层。

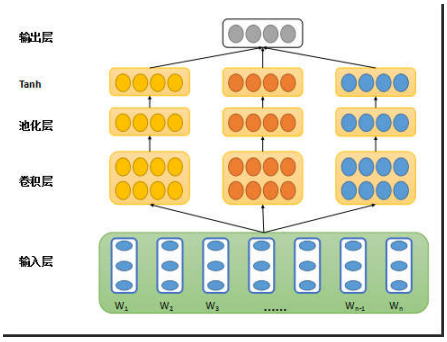

图 2.1CNN 模型

（1） 输入层：将训练语料中经过分词和词性标注后的句子用 Google 开源的 word2vec[7][8]的 Skip-gram 模型训练出 50 维的词向量，若句子的长度为 n，输入层为 n×50 的矩阵的词向量矩阵。

（2） 卷积层：根据词向量和滑动窗口计

算输入层的特征矩阵。考虑到词语的上下文关系，用公式（1）[9]计算滑动窗口 k。本文训练语料的滑动窗口大小如表 2.4 所示。滑动窗口为 k 即意味着每个词向量向下滑动 k 行，得到一个 k×50 的特征矩阵。

$$L=\frac{1}{N}\sum_{3}^{K}iN_i \quad (1)$$

表格 2.4 部分虚词相关短语的滑动窗口

短语类型	滑动窗口 k
介词短语	5
连词相关短语	7
助词“的”相关短语	6

（3） 池化层：将卷积层的特征矩阵分解拼接成的一个列向量。本文选择将 k×50 的矩阵分解，行向量拼接成一个 k×50 维的列向量。

（4） Tanh 层：为激活层。采用激活函数（公式 2）处理池化层的列向量：

$$y_i=\tanh(Wx_i+b) \quad (2)$$

（5） 输出层：对 Tanh 层的结果采用标注集{B，I，E，O}来进行标注，其中 B 为短语边界的开始标记，I 为短语边界中的标记，E 为短语边界结束的标记。O 为短语边界外部的标记。

本文提出的基于 CNN 的虚词相关短语边界识别方法以及文献[2]和文献[4]的方法在 CTB8.0 上的实验结果如表 2.5 所示。

表格 2.5 多种方法短语边界识别结果

短语类型	准确率			召回率			F 值		
	CNN	CRF	规则	CNN	CRF	规则	CNN	CRF	规则
介词短语	80.09%	78.34%	50.43%	81.41%	80.56%	55.64%	80.74%	79.43%	52.60%
连词相关短语	74.63%	69.87%	43.72%	71.58%	71.67%	27.93%	73.08%	70.76%	34.09%
包含助词“的”的短语	72.18%	70.50%		74.49%	71.22%		72.80%	70.87%	

根据表格 2.5 可知，基于 CNN 模型和基于 CRF 模型的短语边界识别结果的准确率要高于基于规则的短语边界识别的结果，在基于 CRF 模型的短语边界识别结果中，介词短语边界识别的结果准确率比助词“的”短语边界识别的结果更高，因为介词短语边界识别只需要确定后边界。基于 CNN 模型的短语边界识别结果的准确率要比基于 CRF 模型的短语边界识别结果的准确率高，介词短语边界识别结果的准确率提升了 1.44%，连词相关短语边界识别结果的准确率提升了 3.02%%，包含助词“的”的短语边界识别结果的准确率提升了 1.21%。

2.3 基于虚词用法的句法分析

文献[10]将虚词用法应用到依存句法分析的依存关系识别中，在并列关系识别过程中加入连词的用法信息，提高了并列关系的识别效果，实验结果表明，包含连词的并列关系的 LAS 及 UAS 分布提高了 3.43%和 2.29%。

文献[11]使用基于介词“在”用法属性的边界结果对 Stanford Parser 进行了后处理，实验结果表明，融入用法属性特征的句法分析结果比之前结果有了一定提高，但是文献[11]没有考虑边界交叉的错误情况。

文献[12]将句法树中介词“在”的短语用特定标记替换，将替换后的句法树和短语内的子句使用 PCFG 模型进行句法分析，实验结果表明，准确率达到 70.8%，但是文献[12]没有考虑其他虚词相关短语边界在句法分析中的应用。

3. Phrase_Based Parser

本文提出一种基于短语边界的句法分析模型 Phrase_Based Parser，其中 Phrase 特指介词短语，连词相关短语和包含助词“的”的短语。该句法分析模型的主要思想是：第一步对训练语料预处理，确定树库语料中的短语标记，将短语标记内的子句法树用固定

标记替换，构建短语标记树库。同时将短语标记内的子句法树也作为语法模型的训练语料。第二步将经第一步得到训练语料用 PCFG 模型统计句法规则，训练出新的语法模型。新的句法模型可用于自动生成句法树。Phrase_Based Parser 模型如图 3.1 所示。

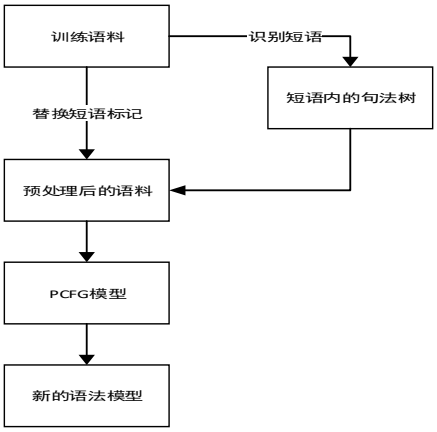

图 3.2Phrase_Based Parser

3.1 构建短语标记树库

本文先将训练语料划分成包含不同虚词成分的树库，再对树库进行预处理。在获得包含不同虚词成分的树库后，对树库进行预处理，预处理语料过程为：（1）去除不需要的功能性标记；（2）获得新的短语树库；（3）替换短语标记。在步骤（3）中由于连词相关短语的特殊性所以需要在替换短语标记时另外增加短语的标记。例如 CTB8.0 标准树库预处理语料过程中的句法树如图 3.2、图 3.3、图 3.4 和图 3.5 所示

预处理前的句法树：

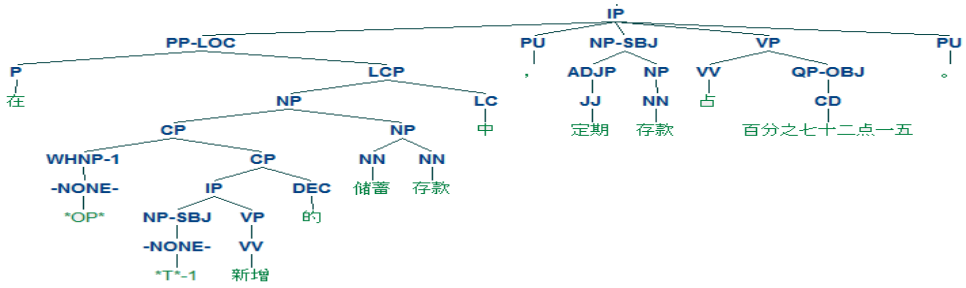

图 3.3 预处理前的句法树

去除不需要的功能性标记后的句法树：

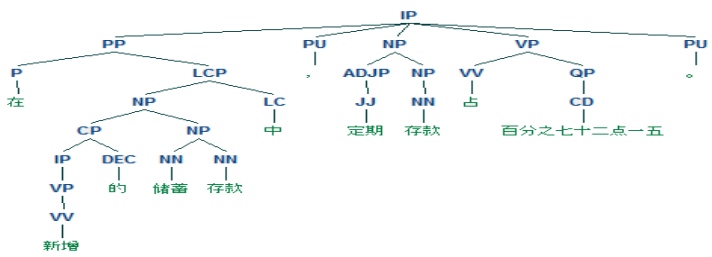

图 3.4 去除功能性标记后的句法树

新的短语句法树：

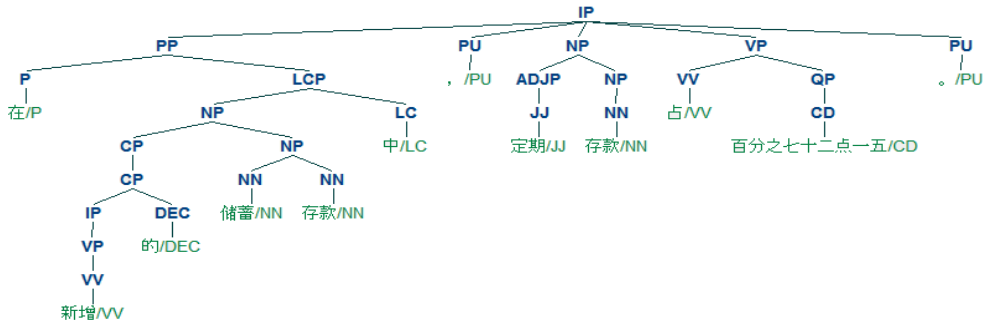

图 3.5 新的短语句法树

替换短语标记后的句法树：

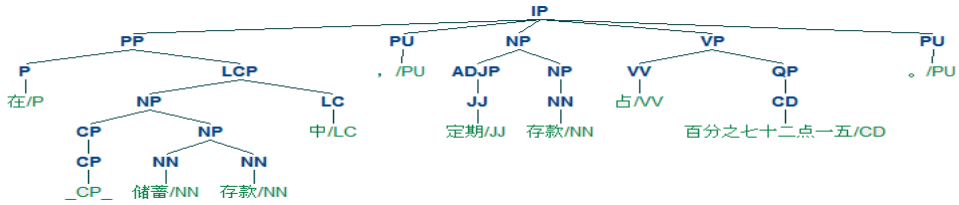

图 3.6 替换短语标记后的句法树

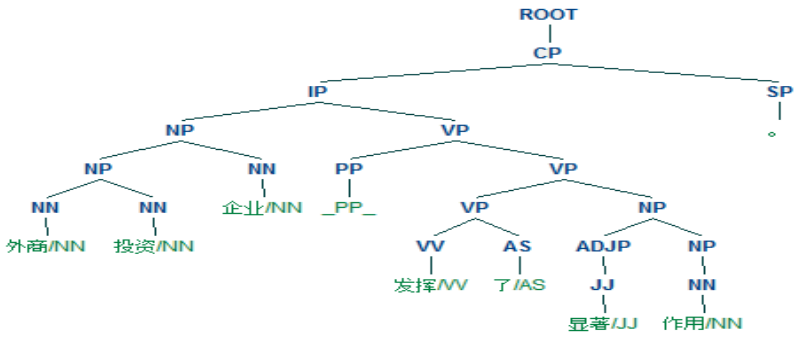

图 3.7 输入 I 使用 Phrase_Based Parser 句法分析结果

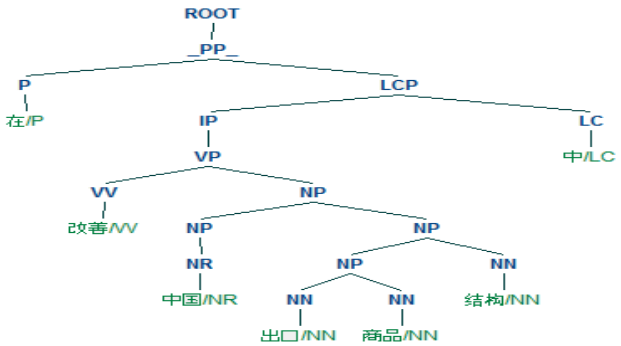

图 3.8 输入 II 使用 Phrase_Based Parser 句法分析结果

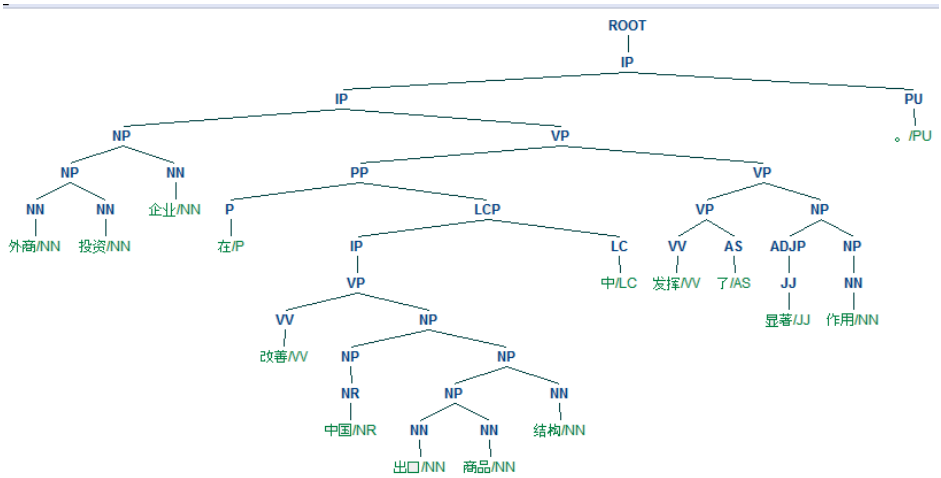

图 3.9 最终的句法分析结果

3.2 训练基于短语标记树库的 PCFG 模型

不同的树库经过训练后得到的语法模型是不同的。本文采用概率上下文无关文法（Probabilistic Context Free Grammar，PCFG）算法来自动提取句法规则，训练工具采用 Berkeley Parser[2]，训练的命令为：

java -cp BerkeleyParser1.7.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -path /data/train.txt -out X_MODEL.gr -treebank SINGLEFILE

其中 train.txt 是训练语料，要求是 UTF-8 编码格式，X_MODEL.gr 是新的语法模型。

3.3 生成句法树

使用 Phrase_Based Parser 进行句法分析的过程为：（1）将含有短语标记的分词序列使用 Phrase_Based Parser 进行句法分析得到带短语标记的句法树（2）将原短语标记内的子句的分词序列使用 Phrase_Based Parser 进行句法分析得到子句法树（3）将步骤（1）的句法树中的短语标记用步骤（2）的子句法树替换，得到最终的句法树结果。Phrase_Based Parser 的输入有两部分构成，输入 I 是句子对应的短语标记句法树的分词序列，输入 II 是虚词相关短语内的子句法树的分词序列。使用 Phrase_Based Parser 对“外商投资企业发挥了显著作用”进行句法分析的过程如下：

原句进行分词和词性标准后的输入 I 和输入 II 分别为：

（I）外商/NN 投资/NN 企业/NN _PP_ 发挥/VV 了/AS 显著/JJ 作用/NN 。/PU

（II）<PP> 在/P 改善/VV 中国/NR 出口/NN 商品/NN 结构/NN 中/LC </PP>

输入 I 使用 Phrase_Based Parser 进行句法分析的结果如图 3.6 所示；输入 II 使用 Phrase_Based Parser 进行句法分析结果如图 3.7 所示。将图 3.7 的子句法树替换掉图 3.6 中的标记_PP_，就得到了最终的的句法分析结果。如图 3.8 所示。

4. 实验

本文的测试语料为 CTB8.0，分别用 Berkeley Parser[12]和 Phrase_Based Parser 进行句法分析，考察虚词相关的短语边界对句法分析结果的影响。最后将基于 CNN 模型的短语边界识别方法与 Phrase_Based Parser 结合进行句法分析。

4.1 Phrase_based Parser 实验过程

本文分别用 CTB8.0 来验证 Phrase_based Parser。

对 CTB8.0 中包含介词短语、连词相关短语和助词“的”的句子按照 3.1 节的方法构造了基于 CTB8.0 的短语标记树库和子树结构，按 3.2 的方法训练语法模型，然后分别使用不同句法分析模型进行句法分析的结果如表 4.1、表 4.2 和表 4.3 所示：

[2]https://sourceforge.net/projects/crfpp/files/

表格 4.1 CTB8.0 包含介词短语句子的句法分析结果对比

句法分析模型	准确率(Precision)	召回率(Recall)	F 值
Berkeley_Parser	83.22%	93.25%	87.95%
Phrase_Base Parser	86.10%	95.77%	90.68%

表格 4.2 CTB8.0 包含连词相关短语的句子的句法分析结果对比

句法分析模型	准确率(Precision)	召回率(Recall)	F 值
BerkeleyParser	81.99%	92.58%	86.97%
Phrase_BaseParser	82.80%	94.56%	88.29%

表格 4.3 CTB8.0 包含助词“的”的短语句子的句法分析结果对比

句法分析模型	准确率(Precision)	召回率(Recall)	F 值
BerkeleyParser	83.36%	93.04%	87.94%
Phrase_BaseParser	73.43%	81.51%	77.26%

表格 4.4 短语标记内的子句用 Phrase_Base Parser 进行句法分析的结果对比

短语类型	准确率(Precision)	召回率(Recall)	F 值
介词短语	91.75%	88.60%	90.15%
连词相关短语	82.90%	97.53%	89.62%
包含助词“的”的短语	93.08%	88.42%	90.69%

其中短语标记内的子句使用 Phrase_Base Parser 进行句法分析的结果如表所示。

表 4.1、表 4.2 和表 4.3 是 Phrase_Based Parser 对包含不同虚词成分的短语的句子进行句法分析的结果。表 4.4 是短语内部的句子用 Phrase_Based Parser 句法分析的结果，由表 4.1、表 4.2 和表 4.3 可知，在含有介词短语的测试语料中，Phrase_Based Parser 的句法分析结果比 Berkeley Parser 准确率提高了 2.88%、召回率提升了 2.52%，F 值提高了 2.72%，在表 4.4 中介词短语内的句子使用进行句法分析的结果也相对较高，F 值达到了 90.23%。在含有连词相关短语的测试语料中，Phrase_Based Parser 进行句法分析的结果比 Berkeley Parser 准确率提高了 0.81%，召回率提升了 1.98%，F 值提高了 1.32%，连词边界内的分词序列进行句法分析的召回率是达到了 97.53%。包含助词“的”的短语的测试语料中，虽然包含助词“的”的短语内的分词序列进行句法分析的结果准确率达到了 93.08%，召回率和 F 值也较高，但是整句用 Phrase_Based Parser 进行句法分析的准确率下降了 9.93%，召回率下降了 11.53%，F 值下降了 10.68%。

4.2 融合基于 CNN 的短语边界识别与 Phrased_Based Parser 的实验过程

本文对 CTB8.0 中包含介词短语、连词相关短语和助词“的”的分词和词性标注语料，使用基于 CNN 的短语边界识别方法进行自动虚词相关的短语边界识别，对获得的短语边界进行短语标记，然后使用 Phrase_Based Parser 进行自动句法分析。结果如表 4.5 所示：

表格 4.5 基于 CNN 模型的短语边界识别句法分析的结果

短语类型	准确率(Precision)	召回率(Recall)	F 值
介词短语	82.74%	89.46%	85.97%
连词相关短语	80.29%	85.63%	82.87%
包含助词“的”的短语	66.78%	70.66%	68.67%

根据 4.1 节的句法分析结果和表 4.5 的对比可知，表 4.5 中句法分析的结果的平均准确率比 Berkeley Parser 句法分析的结果的平均准确率要低 6.26%，比 Phrase_Based Parser 句法分析的结果的平均准确率要低 4.12%，说明短语边界的准确率有进一步的提升空间。

5. 结论

本文先对汉英双语树库中错误进行分析，发现包含虚词成分的短语在句法分析中解析的准确性能够影响到句法分析的结果，因此提出了基于 CNN 的虚词相关短语边界识别方法和基于虚词相关短语边界的句法分析模型 Phrase_Based Parser。在 CTB8.0 上的实验结果表明，基于 CNN 的虚词相关短语边界识别的平均准确率为 75.63%，较基于规则和 CRF 模型的方法提高 25.82%和 2.73%；Phrase_Based Parser 在 CTB8.0 标准树库进行句法分析的结果相对 Berkeley Parser 进行句法分析的结果在包含介词短语和连词相关短语的句子上有提升，融合了基于 CNN 的短语边界识别方法和 Phrase_Based Parser 在 CTB8.0 上进行句法分析的结果的平均准确率为 76.60%。

Phrase_Based Parser 在 CTB8.0 上进行句法分析已经取得了较好的效果，下一步的工作是将 Phrase_Based Parser 应用于汉语双语树库，并进一步提升短语边界识别的准确率。

参考文献

[1] 汉英双语短语结构句法树库构建初探，第十五届汉语词汇语义学国际研讨会论文集，张坤丽；昝红英；韩英杰，2014.6: 430-435

[2] NianwenXue，Fei Xia，Fu-Dong Chiou，and Martha Palmer. The Penn Chinese TreeBank；Phrase Structure Annotation of a Large Corpus. Natural Language Engineering，2005；11(2)207~238.

[3] 昝红英，周丽娟，张坤丽.基于用法的现代汉语连词结构短语识别研究.中文信息学报，2012, 26 (6) : 72-78.

[4] 张坤丽, 昝红英, 柴玉梅,等. 现代汉语虚词用法知识库建设综述[J]. 中文信息学报, 2015, 29(3):1~8.

[5] 袁应成. 基于用法属性的现代汉语介词短语边界识别研究[D]. 郑州大学, 2011

[6] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4):193~202.

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111~3119.

[9] 王东波, 陈小荷, 年洪东. 基于条件随机场的有标记联合结构自动识别[J]. 中文信息学报, 2008, 22(6):3-7.

[10] 昝红英，张静杰.娄鑫坡.汉语虚词用法在依存句法分析中的应用研究.中文信息学报，2013, 2 7（5） : 35~42.

[11] 介词“在”用法在短语结构句法分析中的应用研究，穆玲玲；庞熠雅；昝红英，第十五届汉语词汇语义学国际研讨会论文集，2014-06，澳门，543~548

[12] Mu L, Feng X, Wang H, Studies on the Application of the Usages of Preposition “在[zai](in)” in Phrase Structure Syntactic Parsing[J].International Journal of Knowledge and Language Processing,2014,5(2):1~11.

[13] Klein D, Manning C D. Accurate unlexicalized parsing[C]Proceedings of the 41st Annual Meeting on Association for Computational

Linguistics-Volume 1. Association for Computational Linguistics, 2003: 423~430.