基于大规模语料库的英语从句识别*

黄玉[1] 李生[1] 孟遥[1] 丁华福[2]

[1](哈尔滨工业大学科学与技术学院, 哈尔滨 150001);

[2](黑龙江省信息中心, 哈尔滨 150001)

E-mail:{hy, lisheng, meng}@mtlab.hit.edu.cn

摘要: 英语从句识别对于英语复合句的分析至关重要。本文基于 Penn tree bank 语料库，通过分析从句的组成规律，利用统计规则，通过分析从句的结构，从句在主句中的位置以及与主动词的关系来识别从句的左右边界，在识别的过程引入了关键词，并考虑到从句嵌套的问题。测试结果表明，句首的封闭测试精确率和召回率分别为 91.06%和 94.07%，开放测试精确率和召回率分别为 82.13%和 85.05%。

关键词: 从句识别，语料库，规则

引言

在英语句法分析中，复合句的分析精度还不够理想。可以考虑先把复杂的句子简单化，然后再进行句法分析。对于含有从句的句子，这个简单化的过程，可以通过先识别出英语从句，再分别对主句和从句进行分析来实现。从句识别的目的是通过界定从句的边界，降低对句子进行深层次分析的复杂性。

对于英语从句的识别，国内外的研究者提出了一些研究方法。90 年代中期美国三所大学—New Mexico State Unive(简称 NMSU)、Univ of Southern California(简称 USC)、Carnegie Mellon Univ(简称 CMU)联合实现的机译系统 PANGLOSS MARK III 使用 DCG 规则识别 4 种类型的从句[1]。2001 年国外的一些研究者们提出了采用机器学习的算法在文本中确定从句的边界。Antonio Molina 采用隐马尔可夫模型的方法来识别从句的边界[2]，Xavier Carrearas 采用 Adaboost Decision Trees 的方法[3]来界定从句的边界。国内一些研究者利用从句的连接词等功能词，建立词典库或语法规则库来切分长句[4]；或者基于语料库，利用词性信息，将规则和统计相结合识别从句的边界[6]。

本文通过对大规模语料库的分析，总结出英语从句的组成规律，采用规则的方法识别英语从句。识别的过程分为三步：首先识别从句的句首（从句开始的第一个词或者短语），然后识别从句的句尾（从句结束的最后一个词或者短语），最后根据前两步的结果，输出完整的从句。在识别过程中，综合词汇和词性的信息，基于语料库，利用规则来确定从句的句首；利用从句动词在主句中的位置，以及从句和主句间的关系，来划定从句的句尾。并且对递归从句的识别加以考虑，取得了比较好的效果。

本文的组织结构如下：第一部分，介绍从句的结构和组成规律；第二部分，介绍识别从句边界所需规则的获取；第三部分，介绍从句的识别算法；第四部分，试验结果的分析；第五部分，本文的结论和今后工作展望。

* 本项目受到国家 "863" 项目基金(项目编号 2001AA114101)的资助。

1 英语从句的特征分析

从构成方式上看，复合句分为如下两种：

(1)用连接词把从句与主句连接起来（在一些情况下连接词可以省略）

If you' re not good at figures,(从句)it is pointless to apply for a job in a bank.(主句)

(2)用动词不定式或分词结构，它们构成复合句（而不是简单句）的一部分，

To get into university you have to pass a number of examinations.

在这里，将第二种情况当作一般短语处理，主要考虑第一种情况的从句的识别问题。

本文对英语从句的识别，主要是对从句边界的划定。这里，把从句开始的第一个词或者短语称为句首，把从句结束的最后一个词或者短语称为句尾。

本文基于的语料库是Penn Tree Bank语料库，Penn Tree Bank是一个拥有各层次标注的英语语料库，例如词性级，短语级以及从句级等等。抽取其中21115个从句，其中递归从句有4936句，通过分析这些句子的结构发现，不管是递归从句还是非递归从句，从句的构成都可以分为以下3种情况：

（1）由一定的功能词引导，这些功能词包括who, which, when 等WH词，连词例如before, after, if 等等；

（2）特殊的动词，后面跟有宾语从句，例如say, think,等等，在许多描述个人情感的形容词（如afraid，glad等）或者表示确信无疑的形容词（如certain，sure等）后跟有从句；

（3）特殊的句子结构，如果一个谓语动词前面有连续两个BNP（基本名词短语），这个谓语动词极有可能是从句动词，或者，两个谓语动词在一起，其中一个肯定属于从句动词。

经过统计得出，第一类情况所占比例较大，约有75%，第三类的从句所占比例很小，只有710个。由训练语料统计出这三种情况的优先级是递减的，同时，在每一种情况的内部，也都按照概率从大到小对其进行排列，例如WH词的优先级大于其他连词。这样，在一个复杂的句子中，如果以上情况综合出现，可以根据优先级进行判定。在从句句首识别中，根据从句所属的不同情况，分别进行处理。通过对语料库中从句的分析，可以得出，从句句尾的位置跟以下几个因素有关：从句在句子中的位置，从句与主句的关系，从句的谓语动词在句子中的位置，以及一些标点符号，等等。在识别句尾时，根据以上信息加以处理。

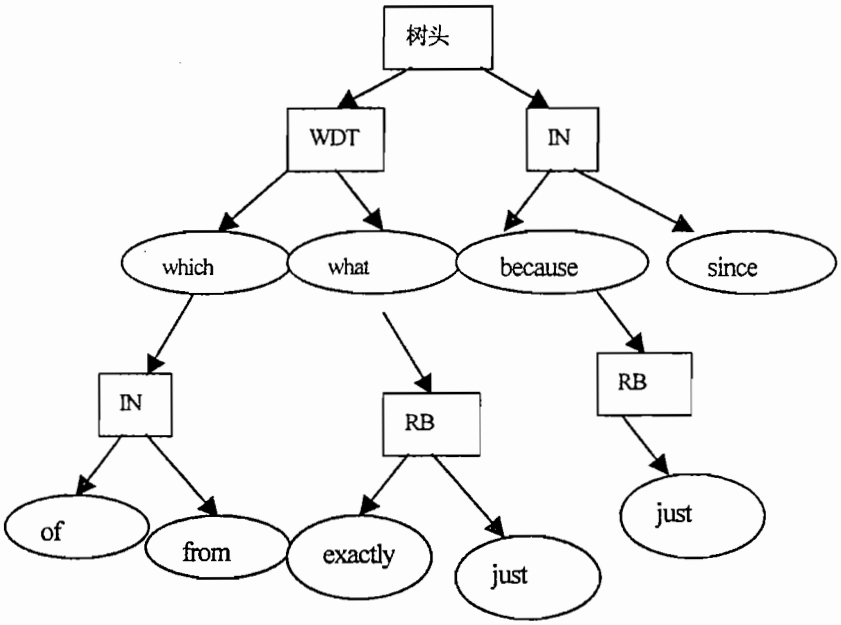
2 规则的获取

2.1 规则的形式

由从句的以上分类，在从句的句首识别中，引入了关键词词表。关键词定义如下：如果一个动词后面带有宾语从句，则把这个动词当作一个关键词；如果一个形容词后面带有从句，则把这个形容词当作关键词。关键词词表的形式如：accept / accepted/accepting/afraid；虽然非限定动词在句子中不能充当谓语，但是这样一些特殊的动词有可能引导从句，所以把这些动词的形态变化都加入到词表中去。对于功能词，采用的规则形式是综合词性和词汇特征的规则匹配树，在从句识别过程中采取最大长度匹配。规则匹配树的形式如图1所示。图1中方形符号表示的是词性，椭圆形符号表示的是词汇，从第二层节点到叶子节点的每一条路径代表了一条规则。另外，在复杂的英语句子中，以上所说的从句的三类情况可能会综合出现，这里通过概率来确定它们的优先级，概率定义如下：

$$P = \frac{\text{关键词或功能词在语料中引导从句的频度}}{\text{关键词或功能词在语料中出现的频度}} \quad (1)$$

图1　规则匹配树的一般形式



2.2　从句识别关键词和规则的自动获取

正文基于Penn Tree Bank语料库，采用其WSJ目录下除第21个子目录外的一共20324个句子，进行规则的获取。获取规则的算法如下：

1. 关键词集，规则集置空
2. 取训练集中一个句子，执行下列操作：
3. if 该从句由功能词引导，则转向4；if 该从句由关键词引导，则转向5
4. 取得该从句句首到此功能词的词性和词汇序列，如果规则集中没有此序列，则加入；转向6
5. 取得该关键词，如果关键词集合中没有重复，则加入；转向6
6. 训练集非空，转向2
7. 结束

为了使关键词尽可能面向真实文本，从训练语料中获取关键词后，对该集合进行了一个后处理。对集合每一个关键词，通过语义词典，获取其同义词，判断这些同义词是否能够引导从句，如果可以，将其加入关键词集合。

3 从句边界识别算法描述

从句边界的识别分为三步：从句句首识别；从句句尾识别；输出完整的从句。

1. 从句句首的识别

首先判断输入的句子是否有从句，如果有，以谓语动词为触发条件，针对谓语动词，根据从句的三种分类，扫描句子是否满足这些情况，若满足则判定该谓语动词是从句动词，否则为主句动词，继续取下一个谓语动词。

对从句的第一种情况，则根据规则匹配树，最大长度匹配规则确定句首；第二种情况，如果关键词后面有“that”引导从句，则从句的句首为“that”，否则，从句的句首为关键词之后从句动词之前有主语性质的词或短语；第三种情况下，对这两个连续出现的 BNP，还需要进一步判定，例如，第一个 BNP 为[the only thing]，则从句的句首为第二个 BNP，具体的判定算法由于篇幅所限，不再详述。

上一步的输出是带有从句句首标记和从句动词标记的句子链，因为一个句子至少有一个主句动词，所以在上一步结束后，对句子进行二次筛选。如果整个句子没有主句动词，则根据概率确定每一个从句的优先级，释放优先级最小的从句，该从句动词确定为主句动词。

2. 从句句尾的识别

在第一步从句句首识别结束以后，句子中的每个谓语动词都已经打上了主句动词或者从句动词的标记。对于确定了句首和动词的从句，从句的句尾的位置，与从句句句首和从句动词在整个句子中的位置，从句与主句动词的关系等因素有关。对于是否有嵌套从句的存在，也是在从句的句尾识别过程中来判定的。对于从句句尾的界定，采取以下策略：

从句的位置在主句动词之前，首先判断此从句是否有做主语的可能或者其修饰的先行词可能做主语，若有，则从句的句尾界定在下一个主句动词之前，否则，从句的句尾在下一个主句动词的主语前；如果从句的动词与主句动词之间有逗号分离，则从句的句尾界定在逗号前。如果此从句充当状语成分（这可由从句的引导词和从句与主动词之间的关系来判断），则从句的句尾界定在下一个主句动词的主语前，如果从句的动词与主句动词之间有逗号分离，则从句的句尾界定在逗号前。

从句的位置在主句动词之后，则这个从句句尾的位置与下一个主句动词有关。如果从句之后没有主句动词，则从句的句尾界定在句末符号前；否则按照上一策略界定从句的句尾。

如果在从句句首和从句动词之间，还有从句句首和从句动词出现，则认为这个从句有嵌套；或者，在从句动词之后仍有从句句首和从句动词出现，并且第二个从句的句首与前一个从句在语法上没有明显的分隔，也认为这个从句有嵌套。对于嵌套从句句尾的界定，方法是把内部的从句当从句，把外部的从句当主句，按照上面的策略来解决。

在从句的识别过程中，对于从句在句子充当一些成分，从句和主句动词的关系等等也是需要一系列的算法来判定。由于篇幅有限，在此不再详述。

3. 识别完整的从句

在将从句的首尾都识别出来以后，还需要做一些后处理，比如删除不匹配的边界等等；另外，在有些复杂的句子中，还有可能会出现两个从句并列合成一个大从句的现象，这些工作在第三步来完成。

4 实验结果分析

本文的实验数据来自 Penn Tree Bank 语料库，训练语料来自 WSJ 目录下除第 21 个子目录外的一共 20324 个句子（含有从句 25436 个），开放测试语料来自 WSJ 目录下第 21 个子目录下的 791 个句子(含有从句 1127 个)。采用精确率和召回率来评测实验结果。

$$\text{精确率}=\frac{\text{正确识别从句的个数}}{\text{总共识别从句的个数}} \tag{2}$$

$$\text{召回率}=\frac{\text{正确识别从句的个数}}{\text{语料中从句的个数}} \tag{3}$$

如果只考虑使用词性信息，从句识别的精确率为 73.56%，召回率为 70%；如果将词汇和词性信息结合起来考虑，实验效果可以得到明显的改善。本文采用的算法，实验结果如表 1 所示：

表1 从句识别实验的精确率和召回率

	精确率（句首）	召回率（句首）	精确率（句尾）	召回率（句尾）
封闭语料	91.06%	94.07%	88.78%	86.15%
开放语料	82.13%	85.05%	78.35%	81.28%

通过分析上述实验结果可以看出，将词汇和词性信息结合起来，基于大规模语料库，分析从句的组成规律，来识别从句的左右边界，取得了较好的效果。

5 结论及工作展望

本文基于大规模语料库，针对从句的结构，将从句分成三类，对每一类从句分别利用关键词表和规则匹配树来确定从句的句首；分析从句在整个句子中的位置，从句与主句动词的关系等信息来界定从句句尾。通过对大规模语料的分析，发现从句的组成是相当有规律的，这些规律在词汇和词性信息上表现的尤为明显，因此将词汇和词性信息结合起来，利用规则来进行英语从句的识别，不失为一条有效的途径。

在从句句首的识别过程中，利用规则匹配树进行规则的最大长度匹配，会造成歧义，在以后的工作中，可以考虑通过加概率来进行消歧。在从句的句尾识别中，本文的算法是通过对大规模语料进行统计后提出的，对句子的内部分析得不很深入。自然语言纷繁复杂，通过分析实验结果，发现在复杂的英语句子中，有不少并列句存在，这对英语从句的句尾识别造成了很大的影响，要想提高递归从句的识别精度，在今后的工作中，还需要深入研究句子的内部结构，同时还要进一步解决并列现象。

参考文献：

[1] Sergei Nirenburg (editor). The PANGLOSS Mark III Machine Translation System. A Joint Technology Report, CMU-CMT-95-145, Apr. 1995

[2] Antonio Molina and Ferran Pla. 2001. Clause Detection Using HMM. In Proceedings of CoNLL-2001. Toulouse France.

[3] Xavier Carreras and Lluis Màrquez. 2001. Boosting Trees for Clause Splitting. In Proceedings of CoNLL-2001. Toulouse France.

[4] 刘志杰. 英汉机器翻译软件长句分析刍议，1999 年计算语言学全国联合学术会议论文集. 北京：清华大学出版社，1999

[5] 王虹. MT 系统的从句分析、设计与实现 [硕士学位论文]. 哈尔滨：哈尔滨工业大学计算机系，1992

[6] 张晶. 基于语料库的英语从句识别研究. 中文信息学报 2000，6，第十四卷，第六期

[7] L·G·亚历山大，郎文英语语法，外语教学与研究出版社

作者简介：　黄玉，女，硕士研究生，主要研究方向为英语复合句的识别；李生，男，博士生导师，主要研究方向为机器翻译；孟遥，女，主要研究方向机器翻译，自然语言处理

A Large Scale Corpus-based Approach to English Subordinate Clause Identification

HUANG Yu[1]　LI Sheng[1]　MENG Yao[1]　DING Fu Hua[2]

[1](school of computer science and technology, Harbin Institute of Technology, Harbin, 150001 China);

[2](Information Center of Heilongjiang, Harbin, 150001, China)

E-mail:{hy, lisheng, meng}@mtlab.hit.edu.cn

Abstract: The identification of English subordinate clause has an great influence on the analysis of complex sentences. This paper introduces a good stratagem to handle subordinate clauses. Based on a large scale corpus we analyze the structure of the clauses, extract rules to recognize the head of the clause, and at the same time check the tail of the clause by the position of clause and the relation of clause and main verb. The precision and recall ratios of subordinate clause identification are tested on both closed and open corpora. As far as the head of the clause is concerned, a result of 91.06% precision and 94.07% recall is obtained for the closed test and the open test result is 82.13% precision and 85.05% recall.

Key words: Subordinate Clause Identification, corpus , rule