文章编号: 1003-0077(2009)05-0053-09

基于语义组块分析的汉语语义角色标注

丁伟伟, 常宝宝

(北京大学 计算语言学研究所, 北京 100871)

摘 要: 近些年来, 中文语义角色标注得到了大家的关注, 不过大多是传统的基于句法树的系统,即对句法树上的节点进行语义角色识别和分类。该文提出了一种与传统方法不同的处理策略, 我们称之为基于语义组块分析的语义角色标注。在新的方法中, 语义角色标注的流程不再是传统的"句法分析——语义角色识别——语义角色分类", 而是一种简化的"语义组块识别——语义组块分类"流程。这一方法将汉语语义角色标注从一个节点的分类问题转化为序列标注问题, 我们使用了条件随机域这一模型, 取得了较好的结果。同时由于避开了句法分析这个阶段, 使得语义角色标注摆脱了对句法分析的依赖, 从而突破了汉语语法分析器的时间和性能限制。通过实验我们可以看出, 新的方法可以取得较高的准确率, 并且大大节省了分析的时间。通过对比, 我们可以发现在自动切分和词性标注上的结果与在完全正确的切分和词性标注上的结果相比, 还有较大差距。

关键词: 计算机应用; 中文信息处理; 语义角色标注; 语义组块分析; 条件随机域; 序列标注
中图分类号: TP391 文献标识码: A

Chinese Semantic Role Labeling Based on Semantic Chunking

DING Weiwei, CHANG Baobao

(Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: In recent years, the Chinese SRL (semantic role labeling) has aroused the intensive attention. Many SRL systems have been built on the parsing trees, in which the constituents of the sentence structure are identified and then classified. In contrast, this paper establishes a semantic chunking based method which changes the SRL task from the traditional "parsing-semantic role identification-semantic role classification" process into a simple "semantic chunk identification-semantic chunk classification" pipeline. The semantic chunking, which is named after the syntactic chunking, is used to identify the semantic chunk, namely the arguments of the verbs. Based on the semantic chunking result, the Chinese SRL can be changed into a sequence labeling problem instead of the classification problem. We apply the conditional random fields to the problem and get better performance. Along with the removal of the parsing stage, the SRL task avoids the dependence on parsing, which is always the bottleneck both of speed and precision. The experiments have shown that the outperforms of our approach previously best-reported methods on Chinese SRL with an impressive time reduction. We also show that the proposed method works much better on gold word segmentation and POS tagging than on the automatic results.

Key words: computer application; Chinese information processing; semantic role labeling; semantic chunking; conditional random fields; sequence labeling

1 引言

语义角色标注(Semantic Role Labeling), 其主要任务是分析句子的"谓词—论元"结构, 即标记出句子中某个动词的所有论元。语义角色标注对自然语言处理领域的很多任务都有帮助, 比如问答系统(Narayanan[1]等)、信息抽取(Surdeanu[2]等)和机器

收稿日期: 2008-08-28 定稿日期: 2008-10-09
基金项目: 国家自然科学基金资助项目(60303003); 国家社会科学基金资助项目(06BYY048)

作者简介: 丁伟伟(1985—), 男, 硕士生, 主要研究方向为自然语言处理; 常宝宝(1971—), 男, 副教授, 主要研究方向为自然语言处理。

翻译(Boas[3])等。

语义角色标注的研究最早开始于 Dan Gildea 和 Dan Jurafsky[4],他们的实验所用语料是 Berkeley 大学开发的 FrameNet[5]。在 FrameNet 之后,宾州大学在树库的基础上完成了英文 PropBank[6]。之后,语义角色标注这个任务逐渐得到了国际的关注,众多的经验主义方法被应用到语义角色标注之中,并且取得了很好的结果,例如 Carreras[7-8] 等,Moschitti[9], Pradhan[10] 等, Zhang[11] 等。

中文语义角色标注的工作开展较晚,研究得也不是很充分。最早进行研究的是 Sun[12] 等,由于在当时还没有中文方面的专门语料,所以他们只是人工标记了包含某些动词的一些语料,并在这些语料上进行研究。虽不成系统,但是毕竟是一个有意义的开端。后来,伴随着中文 PropBank[13] 的构建,Xue Nianwen 开始了比较系统的中文语义角色标注的工作(Xue Nianewen[14-15] 等),并得出了一些很有意思的结论,比如:语义角色识别和语义角色分类所采用的特征是有区别的。这些工作不仅对中文的语义角色标注很有意义,也对英文的语义角色标注有所启发。

中国国内对语义角色标注的关注最早起始于刘挺[16] 等,于江德[17] 等,这些研究的重点仍然集中在提升英文的语义角色标注的性能,实验的语料是 CoNLL-2005 的评测语料。汉语方面的研究有刘怀军[18] 等,他们对汉语语义分类也进行了系统的研究。此外,还有一些语料建设方面的成果,例如袁毓林[19]。不过目前,国内对于汉语语义角色标注的研究还主要局限在语义角色分类方面,完整的语义角色标注研究还不多见。总的来说,与英文方面的工作相比,汉语语义角色标注方面的研究仍处在开始阶段。

在以前的研究中,一个完整的语义角色标注系统通常由两个阶段组成:前一个阶段是挑选出句法树上可能充当动词论元成分的节点,这是语义角色识别。后一个阶段的任务是对识别出来的节点进行分类,具体判断出是指定动词的哪类论元,这是语义角色分类。语法树可以是人工标注的,也可以是句法分析器自动分析的结果。语义角色标注可以看作是句法树上的节点的分类问题。CoNLL-2004 评测选择了另一个思路,将语义角色标注的工作建立在浅层语法分析之上,不再对树上节点进行分类,而是利用分析出来的语法组块进行语义角色标注,希望利用相对更准确些的组块分析结果提升语义角色标注准确率,绕过分析准确率相对较差的完全句法分析。

Hacioglu[20] 等曾经进行过类似的工作,他们在英文上进行了基于词的语义角色标注,他们称之为语义论元组块分析。不过他们的工作是在 FrameNet 基础上,并且使用的特征是直接从浅层语法分析借鉴而来,没有体现出来语义角色标注是语义分析问题的特点。语法组块和语义角色之间不存在对应关系。相对于语法组块来说,能够充当语义角色的成分长度变化很大,并且非常依赖于句子中的主要动词。所以,该系统的性能即使在语义标注任务产生的早期也并不算高,他们的工作只应该被看作是一个较为简单的起点。

本文在前人工作的基础上,将语义角色标注的任务分为两个阶段:语义组块的识别和语义组块分类。直接在词的基础上进行语义角色的识别和分类。在每个部分都充分提取反映中文语义角色标注任务特点的独特特征,希望借此提高系统的准确率。本文以下部分是这样组织的:第二部分是介绍中文 PropBank,第三部分是具体介绍语义组块分析,第四部分介绍了实验的设置和特征模板的选择,数据和实验在第五部分。最后是结论与展望。

2 中文 Proposition Bank

中文 Proposition Bank(以下简称中文 PropBank)是宾州大学建设的中文语义角色标注语料库。它主要由两个资源构成:1) 语义角色标注语料。2) 动词框架。其中资源 1 是 PropBank 的主要内容,具体标记了动词和其论元成分在中文 TreeBank 中的位置;资源 2 是一个支持性的内容,类似于词典,标记了所有出现在 PropBank 中的动词的子语类框架。

中文 PropBank 是在中文 TreeBank 的基础上添加了一个语义角色标注层,标记出动词和对应论元在 TreeBank 中的位置。图 1 是 PropBank 中的一个例子(chtb_433.fid 第 1 句)。

在这个例子中,核心动词是“提供”。“提供”只有一个子语类框架,这个子语类框架包含三个论元成分:“提供者”,“被提供物”,分别对应原型施事和原型受事,在 PropBank 中标记为 arg0 和 arg1。此外还有一个与事成分,在 PropBank 中标记为 arg2。在图 1 中,“保险公司”是“提供者”,“保险服务”是“被提供物”。“三峡工程”则是与事成分,是服务的

接受者。

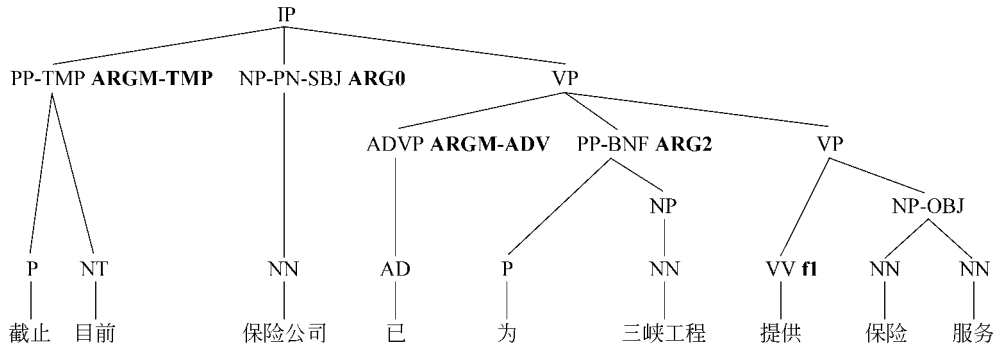

图 1　PropBank 的一个例子

除了原型施事和受事, 在这个例子中还有另一类论元成分。“截至目前”表示“提供”的时间信息, 标记为“argM-TMP”, 其中“argM”是论元标记, “TMP”是一个二级的功能标记, 同类的论元成分在该句中还有“已”, 它被标记为“argM-ADV”, 表示了一个与时间有关的成分。此外, 二级标记在 PropBank 中还有很多, 比如“MNR”(方式), “TMP”(时间) 等。

PropBank 中出现的语义角色可以分为两大类, 核心论元和非核心论元。前一个又可以分为施事、受事、与事等多种论元, 由于 PropBank 中的论元划分依据的是 Dowty 的原型理论, 所以施事、受事等角色包括的范围都是很广的。非核心论语又可以按照功能分出小类, 比如上面的 ADV、MNR、TMP 等就是其中的小类。

传统的语义角色标注系统包括中文 PropBank 的构建, 它们的理论基础是连接理论。这个理论集中阐述了语义层面的角色如何实现为语法层面的句子成分, 依据这一理论, 很自然地会让人想到如何使用一种类似求逆的过程依据句子结构得出论元结构。但是, 如果我们跳出这一理论, 我们会发现语义角色标注的目标还是去确定句子中的哪些部分(语义组块) 是论元成分, 语法分析只是一种手段, 而不是目的。我们完全可以抛开句法分析的步骤, 直接去句子中定位论元的位置, 据此本文提出了一种新的方法, 即语义组块分析的方法。

3　语义组块分析

语义组块分析得名于与语法组块分析的相似性。论元可以是词, 可以是词组。我们可以将论元的识别, 分类看作是一种特殊的组块的发现与分类。为了表示与传统的语法组块之间存在的区别, 我们称之为语义组块。相应地, 我们将语义组块的识别和定性称为语义组块分析。

基于语义组块分析的语义角色标注系统可以分为三个步骤:

1) 将一句话按照其中包含的动词数量复制相同多的拷贝, 每一个拷贝中的目标动词都只有一个。

2) 识别拷贝的句子中的语义组块, 即论元。

3) 对识别出来的语义组块进行分类, 即分配一个语义角色的类标记。

之所以要先将句子按照动词数量进行复制, 是因为句子中不同动词的论元可能存在着重叠, 例如:

(1) 中国建筑业对外开放呈现新局面[1]

这个句子中包含两个动词, 一个是“开放”, 一个是“呈现”。这两个动词分别有其不同的论元结构, 如(1a) 和(1b):

(1a) [arg1 中国建筑业] [arg2 对外] [f1 开放] 呈现新局面。

(1b) [arg0 中国建筑业对外开放] [f1 呈现] [arg1 新局面]。

从这两个例子中, 我们可以发现动词“开放”的受事“中国建筑业”与“呈现”的施事“中国建筑业对外开放”之间是重叠的。如果不将不同动词分开, 语义组块的分析会比较困难。

下面仍然以句 1 为例, 具体说明一下语义组块分析的步骤。我们首先将这个句子复制为两个, 如下:

(1a') 中国建筑业对外开放呈现新局面

(1b') 中国建筑业对外开放呈现新局面

其中(1a') 中的目标谓词是“开放”, (1b') 中的

[1] 这个句子摘自中文 TreeBank ctb_004. fid.

核心谓词是“呈现”。之后我们对两个句子分别进行语义组块分析,以(1a')为例,词语“呈现”、“新”、“局面”都不是“开放”的论元,“中国建筑业”、“对外”分别是目标动词“开放”的 arg1, arg2。故而我们的语义组块分析的流程应该是:在复制的句子(1a')中,识别出来“中国建筑业”、“对外”分别是一个组块。其他的词语不是。然后对识别出来的组块进行分类,看这个组块是动词的何种性质的论元。

语义组块分析包含语义组块的识别和分类,下面将分别加以介绍。

3.1 语义组块识别

从实现的功能上来看,语义组块识别相当于传统方法中句法分析和语义角色识别两步。语义组块识别的目的是确定句子中的哪些部分是该句中目标谓词的可能的语义角色。

语义组块识别是一个序列标注的问题。首先我们需要确定应该采用何种序列的表示法。常用的序列表示法有 IOB 和 start/end。本文将会对各种表示法进行比较,看在语义组块识别这个问题上,哪种表示法能够取得更好的效果。

IOB 表示法可以分为 IOB1, IOB2, IOE1, IOE2 四种。IOB1 最早出现于 Ramshaw[21] 等,后来 Sang[22] 等在其之上进行改造,提出了其余三种不同的表示法。四种表示法大同小异,相同点是:其中的“I”代表当前词在一个组块中,“O”代表的是当前的词不在任意一个组块中。不同点是四种表示法对组块的开始或者(不是并且)结束的表达方式不同。具体如下:

IOB1 B 代表当前词是紧跟前一个组块的新组块的开始。非紧邻的新组块开始标记为 I, IOE1 与之类似。

IOB2 B 代表当前词是一个组块的开始。

IOE1 E 代表当前词是一个组块的终结,当这个组块后门紧跟着另一个组块时。

IOE2 E 代表当前词是一个组块的终结。

start/end 是另一个类型的表示法,最早出现于 Uchimoto et al.[23][1]。该表示法表达得更为细致,共有五种符号:B, I, E, O 和 S。这五种符号表示的意义是:

B 当前词是一个组块的开始。

I 当前词在一个组块内部。

E 当前词是一个组块的终结。

O 当前词不在任意一个组块中。

S 当前词是一个组块,该组块只有一个词。

3.2 语义组块分类

语义组块的分类与传统方法中的语义角色分类基本一样。在这个阶段,前一个阶段识别出来的语义组块分别被分配不同的语义角色。但是由于没有了树结构,很多在之前的对语义角色分类的研究中被证实有用的特征无法获得,使得我们必须想办法提取更多的特征来获得较高的分类准确率,这在第四部分得到了体现。

4 基于语义组块分析的汉语语义角色标注

4.1 数据

本文使用的数据是中文 PropBank,版本 1.0 (LDC 序列号:),该数据库标记了中文 TreeBank 5.1 中的标号为 001 到 931 的文件,共标记了 4 865 个动词的 37 183 个论元结构。

经过我们的分析,中文 PropBank1.0 还存在不少的问题。大概有以下两种:

1) PropBank 与 TreeBank 之间存在不一致。由于这两个语料库不是同时构建,两者之间存在着一些不对应。例如在 PropBank 中记录着 TreeBank 中的某个节点是主要谓词,可是 TreeBank 中的相应位置却根本不是谓词。

2) 同一句话论元结构有问题。比如 PropBank1.0 中的第一个文件的第一个句子“上海浦东开发与法制建设同步”,就存在着问题。“同步”既是 arg1,同时是主要动词。这显然是不合理的。

对有上述两种问题的句子,我们一律删除。

PropBank 中的文件,我们进行如下划分进行本文的实验。训练集语料 648 个文件(chtb_081—chtb_899),测试集语料 72 个文件(chtb_001— chtb_041, chtb_900— chtb_931),开发集语料 40 个文件(chtb_041— chtb_080)。这个语料的设置与文献[15]相同,与文献[14]不是完全相同。

4.2 分类器

由于在序列标注上具有优势,本文选择了条件随机场作为分类器。CRF++[④]开源工具包实现了

[1] Kudo and Matsumoto (2001)[24] 后来对这种表示法进行了一些细小的改造,本文采用的是他们改造后的表示法。

④ http://sourceforge.net/projects/crfpp/

LBFSG 参数估计办法。

CRF++ 工具包区别了两个类型的特征。一个称作 Unigram 特征, 一个是 Bigram 特征, 区别是构建特征时是否包含前一个输出。Bigram 可以产生更多的特征但是效率较低。在本文中, 我们综合利用了两种特征。

对于语义组块分类, 我们使用了 CRF++ 作为分类器, 这是因为动词的论元之间存在着一定的依赖关系。这种关系体现为:

同一类型的论元在一个动词的论元结构中一般只能出现一次。

论元的出现是有顺序的, 例如一般而言, 在非被动句中, 主语位置上是原型施事比是原型受事的可能性大得多。

在英文上的实验已经证明了这一点[25]。所以本文也使用 CRF++ 用于语义组块分类, 希望可以利用论元之间的相互关系, 提高标注的准确率。

4.3　语义组块识别使用的特征模板

与传统的方法不同, 基于语义组块分析的语义角色标注系统不需要句法分析。这在提高了系统的效率同时, 也给特征的选择带来了一定困难。传统方法中句法相关的特征, 比如路径, 动词子语类框架等都无法使用。这使得我们必须多利用词一级的特征。对于语义组块识别, 我们采用了以下的特征模板。

Unigram 特征模板:

目标谓词　每个句子中的目标谓词。

距离　当前词和目标谓词之间相隔词语数量。包含正负号来区别当前词是在谓词之前(+)还是之后(-)。

词-1(0, +1)　前一个(当前, 后一个)词。

词性-1(0, +1)　前一个(当前, 后一个)词的词性。

是否目标谓词-1(0, +1)　前一个(当前, 后一个)词是否是目标谓词。如果是, 则特征为"Y", 否则为"N"。

词性序列前部 0, 词性序列中部 0, 词性序列后部 0　当前词和动词将一个句子中所有的词的词性构成的序列分为三个部分, 我们称其为词性序列前部, 中部, 后部, 0 指的是当前词, 下面出现的 1 和 -1分别指当前词的后一个词和前一个词。对于词性序列的每一个部分, 我们不是直接使用, 而是加以简化处理, 策略是将重复出现的模式省略, 比如词性序列的一个部分是"N-V-N-V-N-V-N", 其中"N-V"组合重复出现了 3 次, 我们只保留一个, 最后的形式是"N-V-N"。通过这样的处理, 我们将序列进行简化, 避免数据稀疏, 同时有助于利用序列之间的相似性。

词性序列前部+1, 词性序列中部+1, 词性序列后部+1　与上面的相似, 这里的前部, 中部, 后部是后一个词与目标谓词把一个句子分隔出来的三个部分。

词性序列前部-1, 词性序列中部-1, 词性序列后部-1　这里的三个部分是前一个词与目标谓词把一个句子分隔出来的三个部分。

间隔标点数量　当前词和目标谓词之间相隔的标点符号的数量, 包括逗号, 分号, 冒号等。

间隔动词数量　当前词和目标谓词之间相隔的动词的数量。

Bigram Feature Templates:

词-1/词 0/词+1　前一个词、当前词、后一个词的组合特征。

动词-1/动词 0/动词+1　动词-1、动词 0、动词+1 构成的复合特征。动词-1(0, +1)这一模板提取特征是前一个(当前, 后一个)词, 当这个词是动词。否则提取的特征是 NULL。

动词词性-1/动词词性 0/动词词性+1　动词词性-1、动词词性 0、动词词性+1 构成的复合特征。动词词性-1(0, +1)这一模板提取特征是前一个(当前, 后一个)词的词性, 当这个词是动词时。否则提取的特征是 NULL。

动词向量-1/动词向量 0/动词向量+1　动词向量-1、动词向量 0、动词向量+1 构成的复合特征。动词向量-1(0, +1)提取的特征是前一个(当前, 后一个)词与动词之间的词性序列中包含的各类动词的数量构成的向量。在中文 TreeBank 中动词共有四种: VV, VA, VC 和 VE。如果这四类动词出现的频次分别是 1, 2, 0, 3, 则该模板提取的特征是向量(1, 2, 0, 3)。这个特征在我们使用北京大学的标记体系的时候略有不同, 因为北京大学标记体系与中文 TreeBank 词性标记集不同。TreeBank 中的动词对应于北京大学标记体系中的动词、形容词和区别词。所以如果采取北京大学标记集的时候, 该模板提取出来的向量只有三维。例如词性序列中间出现的动词、形容词和区别词的次数分别是 2, 3, 0, 则提取的该特征是(2, 3, 0)。

音节数-1/音节数 0/音节数+1　汉语的韵律

往往会很有用, 在这里我们也使用了词的音节数这一韵律特点。这个特征是前一个、当前、后一个词的音节数的复合特征。

距离/目标谓词　距离和目标谓词的组合特征。

动词矩阵 **0**/目标谓词　动词矩阵 0 和目标谓词的组合特征。

间隔动词数量/目标谓词　间隔动词数量和目标谓词的组合特征。

当前词语义类/目标谓词　当前词的语义类与目标谓词的复合特征。

当前词/当前词词性/目标谓词　当前词、当前词的词性和目标谓词的复合特征。

当前词/当前词词性/距离/是否目标谓词 **0**　当前词、当前词的词性、距离、是否目标谓词 0 的复合特征。

前一个输出/当前输出　前一个输出和当前输出的复合特征。

4.4 语义组块分类使用的特征模板

Unigram 特征模板**:**

首词/尾词　首词和尾词的组合特征。

前词/前词词性　前词、前词词性的复合特征。

Bigram 特征模板**:**

首词　语义组块内部的第一个词。

首词词性　首词的词性。

尾词　语义组块内部的最后一个词。

尾词词性　尾词的词性。

前词　语义组块之前的一个词。

前词词性　前词的词性。

后词　语义组块之后的一个词。

后词词性　后词的词性。

长度　语义组块的长度, 即内部包含了多少个词。

距离　语义组块与目标谓词的距离, 即中间相隔多少个词。

前后关系　语义组块是在目标谓词前还是后。

目标谓词语义类　目标谓词的语义类, 提取自北京大学计算语言所的《现代汉语语义词典》, 如果没有则为 NULL。

尾词语义类　语义组块尾词的语义类。

目标谓词　句子中的目标动词。

中间词序列　识别出来的语义组块中间的词语构成的序列, 不包含语义组块的首词和尾词。如果语义组块只有两个词, 则该模板提取的特征是 NULL。

中间词性序列　与中间词序列相对应, 是语义组块中间的词语的词性构成的序列。如果语义组块只有两个词, 则该模板提取的特征是 NULL。

词性序列中部 **0**　这个特征模板与语义组块识别过程中的同名特征模板相同。

谓词框架简单表达式　动词子语类框架简单表达式与文献[15]中提到的 Frame 一致。如动词“保持”有两种子语类框架, 第一种包含三个核心论元, 是 arg0, arg1, arg2; 第二种包含两个论元, 是 arg0, arg1。那么其子语类框架简单表达式是: C3C2。

谓词框架复杂表达式　动词子语类框架复杂表达式是将动词所有子语类框架都连结起来构成的表达式。例如动词“保持”的子语类框架复杂表达式是: C0_1C0_1_2。

前一个输出/当前输出　前一个输出和当前输出的复合特征。

4.5 分词和词性标注

本文中的语义角色标注系统起点是已经完成了分词和词性标注的语料。这样的语料有以下两个来源。

第一个是我们采用了手工标注的结果, 这个结果提取自中文 TreeBank5.1。我们去除了中文 TreeBank 中的树结构信息, 只保留了切分和标注信息。采用手工标注的结果, 保证了切分和标注的正确性, 可以将精力集中在语义角色标注上, 因为汉语的分词和词性标注也是一个很有挑战的问题。

第二个为了说明我们的系统在实际情况下的运行情况, 我们也采用了一些分词和词性标记工具。我们实现了文献[16]中提到的系统进行分词和词性标注, 并且在实验的训练集上进行训练。该工具在测试集上的切分准确率是 95.90%, 召回率是 95.64%, F 值是95.77%; 标注的准确率是90.16%。

另外, 为了说明系统对标记集的依赖性, 我们还采用了文献[27]中提到的切分和标注工具, 并且采用了三种不同的词性划分标准, 包含的标记数量分别是 48, 105, 300。这三种标记系统依据的标准是北京大学的词性标注规范。我们希望借此考察不同的标记体系和标记精细程度对语义组块分析的影响。不过由于没有准确的性能数据, 因而两个分词和词性标注系统的结果缺乏严格的可比性, 仅具有

一定的参考价值。

5　实验结果

实验中 CRF++ 的参数设置是频率阈值为 2, 去掉了那些出现次数只有一次的特征, 以减少偶然性。C 值为拟合度, 取值为 5.7。

表 1 是在手工标注的语料上采用不同的表示法的标注准确率。从中我们可以看出, 使用 start/ end 表示法要比 IOB 表示法效果更好。对 IOB 的 4 种不同表示法来说, IOB2、IOE2 要比 IOB1、IOE1 的标记效果更好。这似乎说明了, 对标注节点分类越细致, 区别越明显, 对语义角色标注的帮助越大。

此外, 单就语义组块分类这个过程来说, 虽然相比较传统的语义角色分类, 少了很多可以利用的树结构的信息, 但是它的准确率却没有什么降低。对于这五种表示法识别出来的语义组块, 再进行语义组块分类的准确率都在 94.1%以上, 相比较 Xue[15] 中的 94.1% 的语义角色分类准确率, 两者差别不大。表 2 是在自动分词和词性标记进行语义组块分析的结果。尽管使用北京大学标记集和中文树库标记集两者使用的方法并不完全相同, 不过结果仍然是可比的。从中我们可以看出来, 使用中文 TreeBank 比使用北京大学标记集效果好。在北京大学 300 词性的标记集上取得的效果与中文 TreeBank 上的效果最接近, 也并未超过。对词性数量不同的北京大学标记集进行考察, 我们可以发现, 进行自动切分和词性标注时采用的词性数量越多, 标记的准确率越高。这也是可以理解的, 词性越细致, 揭示出词的语法、语义上的信息也就更充分, 可以给语义角色标注提供更多的有用信息。

一般来说, 对于序列标注问题, 长度增长往往意味着识别难度的加大。识别的准确率会下降很快。但是在语义组块识别中, 似乎并不是这样。图 2 展示了组块长度与标记效果之间的关系, 图中的 20 代表的是长度大于等于 20 的组块。我们可以从中看出, 在组块长度低于 9 的时候, 标记的 F 值从 80% 以上迅速下降到 60% 左右。而组块长度超过 9 之后, F 值在 0.6 左右波动。这种现象的出现在很大程度上是因为长度很长的组块在数量上比较有限, 而且其中一大部分的分布都集中于一些比较特殊的动词。比如动词“说”、“说明”、“表明”等。这些动词后跟随的很长的词串, 往往就是论元成分。这给语义组块的识别和分类带来了便利。

表 **1**　在手工切分标注的测试集上的实验结果

	语义组块识别			分　类	总　计
	召回率	准确率	F 值	准确率	F 值
IOB1	69.49%	78.09%	73.54%	94.17%	69.25%
IOB2	71.48%	78.45%	74.80%	94.25%	70.50%
IOE1	70.42%	77.53%	73.81%	94.19%	69.52%
IOE2	72.30%	79.53%	75.74%	94.10%	71.27%
start/ end	73.58%	81.00%	77.11%	94.20%	**72.64%**

表 **2**　在自动切分标注的测试集上的实验结果(使用 **start/ end** 表示法)

	语义组块识别			分　类	总　计
	召回率	准确率	F 值	准确率	F 值
北京大学 48 标记集	59.68%	75.84%	66.79%	89.16%	59.55%
北京大学 105 标记集	61.95%	75.81%	68.18%	92.41%	63.01%
北京大学 300 标记集	61.42%	76.10%	67.97%	92.98%	63.20%
中文 TreeBank 标记集	61.57%	74.71%	67.51%	94.05%	**63.49%**

另外, 为了说明我们的方法的有效性, 需要与别的系统进行对比。表 3 就显示了与相关系统的比较情况。本文的数据设置与文献[15] 是完全相同的, 与文献[14] 略有不同, 虽不具严格可比性, 但结果差异仍有参考价值。从表 3 中可以看出来, 在手工标记的切分和词性标注上, 本文系统的效果要略高些,

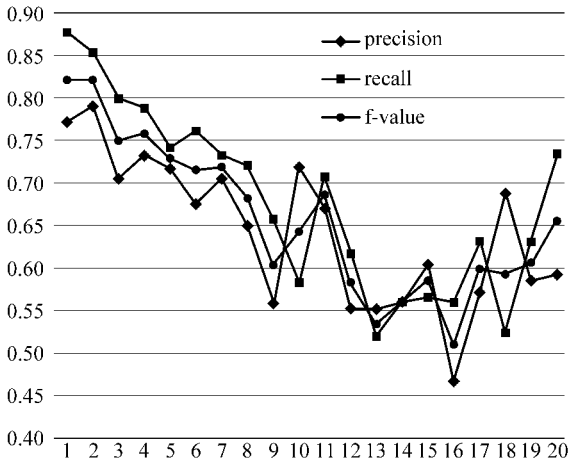

图 2 组块长度与标记效果相关关系

但是在自动切分和词性标注语料上,本文的系统与之前的系统仍然是可比的。而且这种差距某种程度上是由于分词和词性标注工具的不同造成的。Xue[14] 等采用了一个基于最大熵的分词、标注、句法分析一体化的分析器进行分词和词性标注; Xue[15] 则采用了两种方案,一个与 Xue[15] 等相同,另一个则是将基于最大熵的一体化分析器的分词结果拿出来,使用 Bikel Parser 进行词性标注和句法分析。表 4 中的 Xue[15] 的数据是后一个方案的数据,因为第一个方案与 Xue[14] 等相同,效果也不如第二个。这些系统都没有提供分词和词性标注的数据,所以无法判别本文的系统与这两个系统在分词和词性标注上的性能是否存在较大区别。

此外,还需要说明的是我们在基于语义组块的方法上产生了与基于句法分析的方法可比较的结果,这一点似乎是与前人关于语义标注中语法分析的重要性的结论(Carreras[8] 等, Punyakanok[27] 等)相矛盾的。不过这种情况的出现可以从以下两个方面来解释。首先, Carreras 等人的研究成果都是基于英文的而非中文。从分析准确率来看,中文的句法分析器相较于英文逊色许多,错误的分析对语义角色标注的负面影响是很大的,这使得使用基于句法分析的方法进行语义角色标注效果不好。其次,本文不同于基于语法组块的方法,提出了语义组块的概念,使组块分析直接面向语义角色标注。语义组块识别直接依赖于特定动词,有利于充分提取与语义角色标注相关的特征,这使得基于语义组块方法避免了传统的基于语法组块方法中由于句法组块分析和语义角色标注脱节(例如组块边界和语义角色边界不一致)带来的弊端,提高了标注的准确率。

表 3 与相似系统的比较

		文献[14]	文献[15]	本系统
手工标记	P	N/A	79.5%	76.31%
	R	N/A	65.6%	69.31%
	F	N/A	71.9%	**72.64%**
自动标记	P	67.0%	74.5%	70.26%
	R	56.4%	59.6%	57.90%
	F	61.3%	**66.2%**	63.49%

此外,我们对比了自己的方法与传统的基于句法分析的方法在时间消耗上的区别。我们修改了 Collins 句法分析器的源代码使之可以进行汉语的句法分析,采用的是 Collins 句法分析模型 1,实验环境是 Pentium Dual-Core 3.0GHz, 2G 内存,标注的语料是我们的测试集,耗费时间对比如表 4 所示。

表 4 时间消耗的比较

基于句法分析的方法	语义组块的方法
220 小时	17 秒

从表 4 的对比我们可以看出,由于基于语义组块分析的方法去除了语法分析的步骤,大大节省了分析的时间,使系统更为实用。

6 结论与展望

在本文中,我们构建了一个基于语义组块分析的中文语义角色的系统。和以前对语法树上的节点进行分类不同,我们把语义角色标注问题当作是一个序列标注问题,直接对语义角色进行识别和分类。这个新的系统获得了较高的准确率,同时由于去除了语法分析这个非常耗费时间的步骤,系统得以极大地节省了时间。而时间耗费的降低会使得语义角色标注系统更加实用,尤其是处理的语料非常大,比如是数以万计的网页的时候。

Carreras[7-8] 等揭示出来语法分析是语义角色标注中对系统标注准确率影响最大的一个因素。具体到汉语上的情况,语法分析的影响就更大了。文献[15] 揭示出在语法分析完全正确的情况下,英语和汉语上的语义角色标注准确率差不多,但是如果将自动句法分析的结果引入之后,汉语的语义角色标注性能就大大低于英语。本文提出的语义组块分析方法提供了一个去除句法分析影响的途径。另外对于汉语上来说,另一个比较现实的问题是可用的

句法分析器并不是很多。

基于语义组块分析的语义角色标注系统目前还没有表现出对基于句法分析的方法的优越性，未来的可能的一个发展方向是提取更多的特征以提高系统的性能，其他的自然语言处理任务比如命名实体识别，语法组块分析以及更多的语言资源可能会提供很大的帮助。

除了上面提到的发展方向，思考如何将基于语义组块分析的方法与基于句法分析的方法结合起来可能是很有意义的。这两个方法都各有自己的优缺点，基于语义组块分析的方法速度快，不需要语法分析的结果，但是能提取的特征比较有限。基于句法分析的方法可以提取更多的有意义的特征，但是极大地受制于语法分析的准确率。一个可能的结合点是利用语义组块分析去除不充当论语成分的词，从而让句子结构简单化。句法分析器在短句子上分析效果更好，而且更节省时间。利用这一策略，或者可以提高语义角色标注的准确率。

参考文献：

[1] S. Narayanan and S. Harabagiu. Question answering based on semantic structures[C]//Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland. 2004.

[2] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using predicate-argument structures for information extraction[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan. 2003.

[3] H. C. Boas. Bilingual FrameNet dictionaries for machine translation[C]//Proceedings of LREC 2002, Las Palmas, Spain. 2002.

[4] D. Gildea, D. Jurafsky. Automatic labeling of semantic roles[J]. Computational Linguistics, 2002, 28(3): 245-288.

[5] F. C. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project[C]//Proceedings of the 17th international conference on Computational linguistics, Montreal, Canada. 1998: 86-90.

[6] P. Kingsbury and M. Palmer. From TreeBank to PropBank[C]//Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain. 2002.

[7] Carreras X, M rques L. Introduction to the conll-2004 shared task: Semantic role labeling[C]//Proceedings of CoNLL-2004, Boston, MA, USA, 2004: 89-97.

[8] Carreras X, M rques L. Introduction to the conll-2005 shared task: Semantic role labeling[C]//Proceedings of CoNLL-2005, 2005.

[9] A. Moschitti. A Study on Convolution Kernels for Shallow Statistic Parsing[C]//Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004: 335-342.

[10] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, D. Jurafsky. Support vector learning for semantic argument classification[J]. Machine Learning Journal, 2005, 60(1-3), 11-39.

[11] M. Zhang, W. Che, A. T. AW, C. L. Tan, G. Zhou, T. Liu, S. Li, A Grammar-driven Convolution Tree Kernel for Semantic Role Classification[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07), Prague, Czech Republic, 2007.

[12] H. Sun, D. Jurafsky. Shallow Semantic Parsing of Chinese[C]//Proceedings of the HLT/NAACL, 2004.

[13] N. Xue, M. Palmer. Annotating the Propositions in the Penn Chinese Treebank[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan. 2003.

[14] N. Xue, M. Palmer. Automatic semantic role labeling for Chinese verbs[C]//19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland. 2005: 1160-1165.

[15] N. Xue. Semantic Role Labeling of Chinese Predicates[J]. Computational Linguistics, 2008, 34(2): 225-255.

[16] 刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007，18(3)：565-573.

[17] 于江德，樊孝忠，庞文博，余正涛. 基于条件随机场的语义角色标注[J]. 东南大学学报，2007，23(3)：361-364.

[18] 刘怀军，车万翔，刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报，2007，21(1)：79-84.

[19] 袁毓林. 语义角色的精细等级及其在信息处理中的应用[J]. 中文信息学报，2007，21(4)：10-20.

[20] K. Hacioglu and W. Ward. Target word detection and semantic role chunking using support vector machines[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada. 2003.

[21] L. A. Ramshaw, M. P. Marcus. Text chunking using transformation-based learning[C]//Proceedings of the 3rd Workshop on Very Large Corpora. 1995.

[22] E. F. Sang, T. Kim, J. Veenstra. Representing text chunks[C]//Proceedings of the 38th Annual

(下转第 74 页)

[5] YUEN Raymond W. M., CHAN Terence Y. W., LAI Tom B. Y. et al. Morpheme-based derivation of bipolar semantic orientation of Chinese words [C]// Proc. Of the 20th International Conference on Computational Linguistics (COLING-2004), Geneva, Switzerland. 2004: 1008-1014.

[6] 朱嫣岚, 闵锦, 周雅倩, 等. 基于HowNet的词汇语义倾向计算[J]. 中文信息学报, 2006, 21(1): 14-20.

[7] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21[1]: 96-100.

[8] 王根, 赵军. 中文褒贬义词语倾向性的分析[C]//第三届学生计算语言学研讨会论文集. 沈阳. 2006: 81-85.

[9] 张伟, 刘缙, 郭先珍. 学生褒贬义词典[M]. 中国大百科全书出版社. 2004.

[10] 史继林, 朱英贵. 褒义词词典[M]. 四川: 四川辞书出版社. 2005.

[11] 杨玲, 朱英贵. 贬义词词典[M]. 四川: 四川辞书出版社. 2005.

[12] 王素格. 基于Web的评论文本的情感分类问题研究[D]. 博士论文. 上海: 上海大学. 2008.

(上接第61页)

Meeting of the Association for Computational Linguistics, Hong Kong, China. 1999.

[23] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, and H. Isahara. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules [C]//Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China. 2000.

[24] T. Kudo, and Y. Matsumoto. Chunking with Support Vector Machines [C]//Proceedings of Second Meeting of North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA. 2001.

[25] Z. P. Jiang, J. Li, H. T. Ng. Semantic Argument Classification Exploiting Argument Interdependence [C]//Proceedings of 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005: 1067-1072.

[26] H. T. Ng and J. K. Low. Chinese Part-Of-Speech Tagging: One-At-A-Time Or All-At-Once? Word-Based Or Character-Based? [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain. 2004.

[27] H. Duan, X. Bai, B. Chang, S. Yu. Chinese word segmentation at Peking University[C]//Proceedings of the second SIGHAN workshop on Chinese language processing. Sapporo, Japan, 2003: 152-155.

[28] V. Punyakanok, D. Roth, W. Yih. The importance of syntactic parsing and inference in semantic role labeling[J]. Computational Linguistics, 2008, 34(2): 257-287.