编者按：中国中文信息学会于 2010 年 10 月在华中师范大学成功地召开了“第五届全国青年计算语言学研讨会(YWCL2010)”。会议的程序委员会向本刊推荐了 19 篇论文，并经作者仔细修改，编辑部得到授权，在 2011 年第 2 期发表，以飨读者。

文章编号：1003-0077(2011)02-0003-06

基于北京大学中文网库的语义角色分类

杨 敏，常宝宝

(北京大学 计算语言所，北京 100871；
北京大学 计算语言学教育部重点实验室，北京 100871)

摘 要：语义角色标注的研究方法中使用最频繁的一类是基于特征工程，将任务转化成分类问题使用机器学习的方法来解决，几乎所有的有指导语义角色标注采用的标注语料都是宾州大学命题库标注体系。近年来，北京大学开发出一套新的标注语料—北京大学中文网库，该文的目的在于测试这类研究方法在新语料的效果，验证之前所使用的特征是否对标注语料具有依赖性。通过实验发现前人方法中的一些不足，尤其个别特征在北大网库上作用更关键。

关键词：语义角色标注；北京大学中文网库；序列标注
中图分类号：TP391 文献标识码：A

Semantic Role Classification Based on Peking University Chinese NetBank

YANG Min, CHANG Baobao

(Institute of Computational Linguistics, Peking University, Beijing 100871, China;
Key Laboratory of Computational Linguistics, Ministry of Education, Beijing 100871, China)

Abstract: Among all the researches on semantic role labeling(SRL), one important method which has been carried out by many researchers is to convert the task into a classification problem by selecting features, and thenapplying different kinds of classifiers. While almost all the researches based on this kind of supervised learning have been done on the same corpus-Penn Proposition Bank, here we test the same method on a new corpus—Peking University Chinese NetBank, with the goal to figure out whether the wildly used features have a strong dependence on corpus. The experiments have shown that the method and the features have good performance on the new corpus. And compared to the PropBank, some features play crucial roles in classification on the new corpus.

Key words: semantic role labeling; Peking University Chinese NetBank; sequence labeling

1 引言

语义角色标注是当前浅层语义分析的一种主要的实现方式，主要任务是找出给定句中每个谓词的动词——论元结构。语义角色标注意义广泛，在许多复杂的自然语言处理中，都有很大的用处，它对信息抽取、机器翻译等研究都会产生巨大的帮助。

语义角色标注的研究最早关注于英文，最早研究开始于 Dan Gildea 和 Dan Jurafsky[1]，随着宾州大学命题库的建立，语义角色标注任务得到广泛的国际关注，并取得了许多很好的结果，例如 Carreras

收稿日期：2010-08-15 定稿日期：2010-11-30
基金项目：国家自然科学基金资助项目(60975054)；社会科学基金资助项目(06BYY048)
作者简介：杨敏(1987—)，女，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。

等[2-3]，Moschitti[4]等。另一方面出现了一些相关的国际评测：CoNLL 2004[2]、CoNLL 2005[3]、EMNLP-CoNLL 2007和CoNLL 2008都包含了语义角色标注的任务也促进了语义角色标注研究的蓬勃发展。国内对语义角色标注的关注最早起始于刘挺等[5]，他们主要关注的依然是英文语义角色性能的提升。而关注于中文的语义角色标注工作较晚，最开始研究的是Sun等[6]。后来伴随着中文PropBank的构建，Xue Nianwen开始了比较系统的中文语义角色标注的工作[7-8]。国内还有刘怀军等[9]，丁伟伟等[10-11]对汉语的语义角色研究进行了系统的研究。

纵观以前的有指导的语义角色标注任务，无论是对英文还是中文的研究工作，大都是基于宾州大学命题库的语义角色标注体系进行的，CoNLL 2004[2]、CoNLL 2005[3]更是推动了所有研究都基于宾州大学命题库的研究这一趋势，因此研究的一大类方法便是在宾州大学命题库的基础上，基于特征的研究方法。由于北京大学中文网库(以下简称北大网库)的建立，网库的标注方法与宾州命题库的标注方法有所区别，本文的主要任务是将之前的研究方法使用到新的标注语料中，考察之前的研究方法在新标注体系中的作用，进而讨论是否以前的特征选择会有对标注体系的依赖性问题。

本文以下部分是这样组织的：第2节介绍中文Proposition Bank和pku网库标注语料；第3节是具体介绍实验的相关设置；实验的相关结果在第4节；第5节主要介绍两个改进实验。最后一节是结论与展望。

2 语料介绍

2.1 中文PropBank

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。它是在中文TreeBank的基础上添加了一个语义角色标注层，标记出来动词和对应论元在TreeBank中的位置。表1列出了PropBank中出现的所有论元。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。前一类又可以分为施事、受事、与事等多种论元，由于PropBank中的论元划分依据的是Dowty[12]的原型理论，所以施事、受事等角色包括的范围都是很广的。非核心论元又可以按照功能分出小类，比如ADV、MNR、TMP等就是其中的小类。结合图1可知，ARG0-ARG5是核心论元，其他都属于非核心论元。

表1 PropBank中的论元

1	ARG0	2	ARG1
3	ARG2	4	ARG3
5	ARG4	6	ARG5
7	ARGM-ASP	8	ARGM-BNF
9	ARGM-CND	10	ARGM-CRD
11	ARGM-DGR	12	ARGM-DIR
13	ARGM-DIS	14	ARGM-EXT
15	ARGM-FRQ	16	ARGM-LOC
17	ARGM-MNR	18	ARGM-PRD
19	ARGM-PRP	20	ARGM-TMP
21	ARGM-TPC	22	ARGM-ADV
23	TBERR		

2.2 北京大学中文网库

与宾州大学命题库相似，北大网库是在由詹卫东等开发的北大汉语句法分析树库的基础上进行语义标注的，由北京大学中文系袁毓林[13]教授组织完成，语义角色标签标注在句法树的节点上。在语义角色设置方面，与PropBank有些区别，尤其是核心论元的设置。具体论元设置如下[13]：

(一) 必有论元：

A. 主体论元：(1)施事A：自主性动作行为的施行者。(2)感事Se：非自主性的心理感觉的主体。(3)经事Ex：某种变化的具有感知性的主体。(4)致事Cau：某种致使性事件的引起者。(5)主事Th：性质、状态等无施动、感知性的主体。

B. 客体论元：(1)受事P：因施事的行为而受到影响的事物。(2)与事D：动作、行为的非主动的参与者。(3)结果R：动作、行为造成的结果。(4)对象Ta：感知性动作、行为的对象和目标。(5)系事Re：事件中跟主体论元相对的其他各种客体。

(二) 非必有论元

A. 凭借论元：(1)工具I：动作、行为所凭借的器具。(2)材料Ma：动作、行为所用的材料。(3)方式M：动作、行为所采取的方式、方法。(4)原因Rn：动作、行为、事件等发生的原因。(5)目的Ai：发生动作、行为、事件等的目的。

B. 环境论元：(1)时间 T：动作、行为、事件等发生的时间。(2)处所 L：动作、行为、事件等发生的处所。(3)源点 So：动作、行为、事件等开始的时间或处所。(4)终点 Go：动作、行为、事件等结束的时间、处所或状态。(5)路径 Pa：动作、行为、事件等中途经过的时间或处所。(6)范围 Ra：动作、行为、事件等所涉及的数量、频率、幅度、时间等事项。(7)量幅 EXT。

图 1 是北大网库中的一个例子。在这个例子中，出现了三个谓词，分别是：“毫不在意”、“抹去”、“当作”。对于“毫不在意”，句子中对应的论元成分有：感事“他”，对象“这一切”；对于谓词“抹去”，句中对应的论元有：施事“他”，受事“它们”和方式“当作蛛丝一样”；对于谓词“当作”，对应的论元有：施事“他”，受事“它们”，系事“蛛丝”。

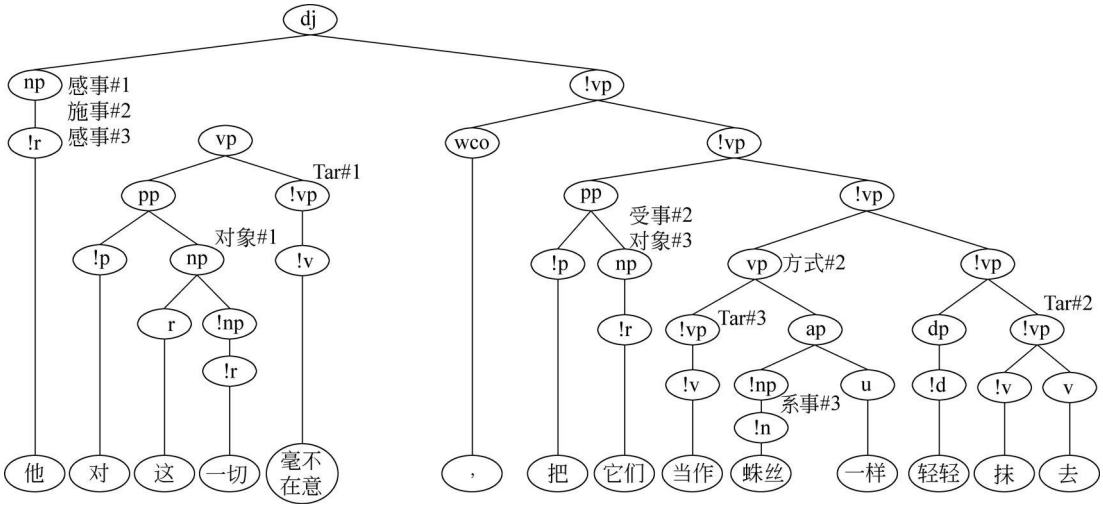

图 1　网库例句示例

2.3　PropBank 与网库的比较

直观从论元分类上看，相比 PropBank，北大网库的论元更细致，分别在主体论元和客体论元内部各划分出五个子类。从语义角色精细等级的理论上[1]看，两种语料库确实有所不同。

PropBank 的语义角色是编了号的原型角色，是中观层次上基于特定动词的角色，又借鉴了宏观层次上原型角色的抽象性地指派的做法，于是用了数目相对有限的带编号的论元，每一个具体动词的语义论元被编了号。对于一个特定的动词，ArgO 通常是表现出 Dowty[12] 中的原型施事的有关特征的论元，Arg1 则是原型受事和主事(Theme)。对于这种被编了号的高级论元，无法做出适合于不同动词的具有一致性的概括。而动词的特定用法相对应的一组角色叫角色集合，这组角色可以跟一组句法框架相联系，这组句法框架显示了那组角色的各种可能的句法变化。而中文网库的语义角色是属于所谓中观层级的语义角色，虽不是基于一个个具体的动词，而是基于具有句法、语义共性的一类动词。虽然北大网库也配套给出了动词的框架描述，但是针对每个动词，它的各类角色都标注在语料中，并不需要像 PropBank 一样从框架描述中才能确定具体的语义角色。

3　语义角色标注

一般的语义角色标注系统分为四个步骤，分别是剪枝 pruning、语义角色识别、语义角色分类以及后处理阶段。国内外很多学者对每个过程的研究也非常丰富，对于识别、分类阶段的特征挑选方面也进行了细致的研究。本文将只对论元分类部分进行研究。

3.1　实验数据

北大网库共 70 个文件，包括的句子总数为 12 434，论元总数为 65 967。我们在划分训练集、开发集以及测试集时采用了与文献[8]大概一致的比例。图 2 是网库中各类论元的分布图，由图可见，论元的分布很不均匀，不仅各大类(共四类)的论元总数相差很远，主体论元、客体论元、凭借论元和环境论元的比例大概为 16.5 ∶18 ∶6 ∶1，各类论元内部分布也不均匀。

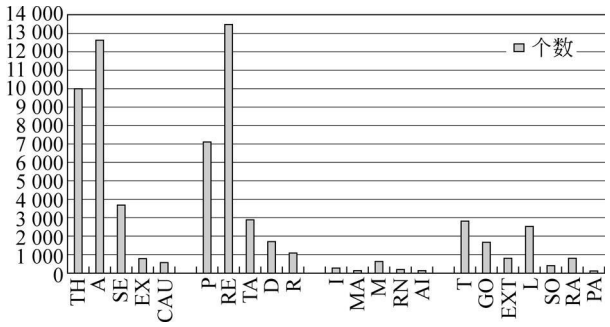

图 2 网库中各论元分布图

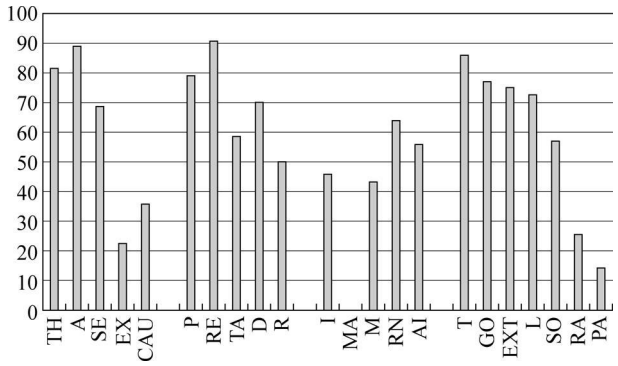

图 3 各论元分类的正确率图

3.2 分类器

本实验采用 Zhang Le 的最大熵分类器 MaxEnt①, 该分类器实现了包含高斯平滑的最大熵算法, 采用 LBFGS 参数估计方法, 可以很方便地处理多类划分的问题。

实验的参数设置如下: 迭代次数 500, 高斯平滑参数为 15。

在改进实验中采用了 CRF++分类器②。

3.3 特征模版

为了使实验结果与前人实验结果具有可比较性, 本实验中采用的特征集合与文献[8] 的 Baseline 一致。特征模版如下: 位置: 句法成分在谓词前面还是后面; 动词的框架: 动词的父节点及其所有子节点构成的框架; 短语类型: 该论元成分的短语类型; 首词: 句法成分的第一个词; 尾词: 该句法成分的尾词; 左兄弟的短语类型; 扩展的动词框架: 动词框架及围绕动词的 np ; 目标谓词; 路径: 句法分析树上句法成分到谓词的路径; 中心词: 该句法成分的中心词; 中心词词性; 复合特征: 谓词+中心词; 复合特征: 谓词+短语类型。

4 实验结果

在网库语料上, 论元分类的准确率为 78.86%。对比文献[8] 中的 93.1%的准确率, 可见该组特征在网库上的表现差很多。图 3 描述了 Baseline 各类论元的分类准确率。由图 3 可见, 各类论元中都有分类准确率比较高的论元, 也有准确率很低的论元。

为了确定被错误分类的论元是被误归类到所属大类的集合中, 还是被错误的分为别的大类中, 我们分别将各大类论元合并, 即采用各种不同颗粒的论元分类法, 同样适用上述特征集合, 具体实验结果如表 2, 由表 2 可知, 当将属于主体论元类的五种论元合并成一个大类、属于客体论元类的五中论元合并成另一大类时, 分类准确率明显提升至 89.18%, 由此可见, 这两大类论元在分类时的内部错误占了整个系统错误的很大一部分。同时, 如果将所有论元按最大粒度的分类方法, 分成四大类, 相比于第二种分类法, 分类的准确率提高了 1.7 个百分点, 由此可见, 依然有部分论元被错误地分到其他大类别中。

表 2 采用不同论元分类法的实验结果

论元分类法	baseline	主体、客体论元分别合并	四大类
准确率/%	78.86	89.18	90.87

5 实验改进及结果分析

由上述实验结果可知, 论元分类的主要错误来自各大类论元内部, 产生这个结果也是与语料标注有关的。网库的语义角色是基于特定谓词的各论元成分的论旨角色, 是属于所谓中观层级的语义角色, 同一动词虽然可以有多种义项, 但同一义项所带的论元框架是统一的, 因此动词框架信息对于论元分类, 尤其是判断主体、客体论元会有很多作用。同时, 由于同一谓词的论元配置具有相对固定性, 因此采用序列标注的思想对于论元分类也会有正面作用。以下两个改进实验就是分别基于上面两个思想进行。

5.1 动词相关特征

由上述分析可知, 在对主体、客体论元分类时,

① 下载地址 http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

② 下载地址 http://chasen.org/~taku/software/CRF++/#features.

谓词相关的特征非常重要。因此仿照前人在 PropBank 中使用的方法，提取每个句中每个谓词的论元框架，并添加以下三个特征 verbFrame，VerbFrame＋headword，verbFrame＋phraseType 后，总体分类准确率从 78.86%提高到 94.34%，提升幅度非常大，而在文献[8]试验中，加入谓词框架相关特征后的分类准确率也只是有一个百分点的提升，可见谓词框架信息对网库角色分类的至关重要性。

但是这种方法有一个很严重的弊端就是：它将所有语料中(包括测试语料中)的每个谓词—论元框架提取出来当做特征，而在真实情况中，是不可能预先知道测试语料中谓词的论元框架，因此这种提取特征的方法一定程度上夸大了分类的准确率。前人在针对 PropBank 的研究中，使用框架特征时普遍存在着这个问题。因此，我们提出一种更贴切现实的谓词框架提取方法，即只提取训练语料中的谓词—论元框架。实验结果如我们预期的一样，这种改进的方法使分类准确率较 Baseline 提升到 88.24%，但相比之前提取谓词—论元框架的方法，准确率降低 6 个百分点。这样验证了我们的观点。

图 4 给出了使用改进后框架特征与使用未改进框架特征的分类结果对比情况。可见，谓词框架的相关特征对提高论元分类准确率的效果很大。而且当去除测试语料中谓词—论元框架信息时的各类论元分类准确率都有所下降，尤其是一些本身数量就比较少的论元，如主体论元中的 CAU、EX。

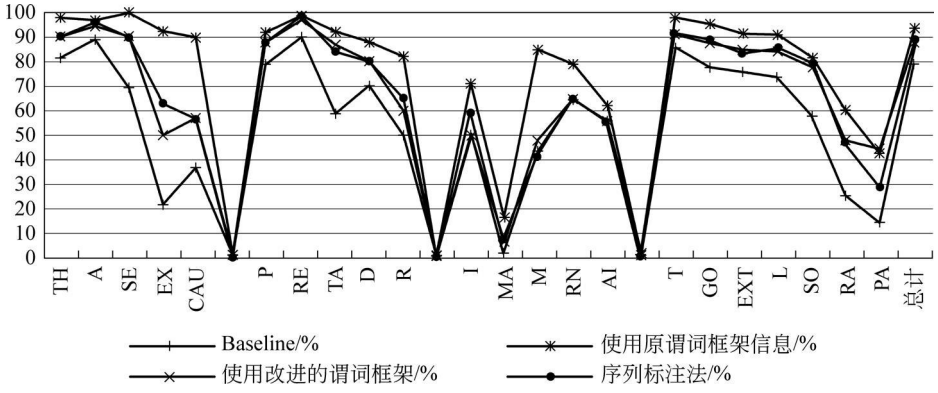

图 4　谓词框架信息修改前后的分类结果对比

5.2　序列标注的思想

在前面的所有实验中，我们都是将所有节点一个一个单独地提取特征，进行分类，各论元之间没有任何联系。但实际在一个句子中，某一谓词的论元之间具有相关性，特定动词的论元成对出现的可能性很大，例如：受事论元被定义为因施事的行为而受到影响的事物，因此受事常与施事论元成对出现，当前面论元已判定为施事时，后面很可能会出现受事论元。采用序列标注的思想，考虑论元之间的相关性。因此使用 CRF＋＋分类器进行分类，总体分类准确率为 88.50%。具体每一类论元的分类准确率如图 4。

将上面所有实验结果与前人结果综合起来，比较结果如表 3。从表 3 可以看到，相比较与在 PropBank 上的论元分类，Baseline 在网库上的效果差很多，也就是说 Baseline 中所使用的特征集合对网库论元分类的效果并不很明显，而谓词框架信息对网库中的角色分类的作用更加关键。然而，在使用修正后的谓词框架信息，分类准确率明显下降了不少，由此我们也可以看出前人在 PropBank 上中使用的谓词框架信息一定程度上夸大了分类的准确率。另外，采用序列标注的思想，将前一个论元的分类结果加入作为特征，对每种论元的分类准确率都有提升还是很大的，这里只是在 Baseline 的基础上使用序列标注，准确率比 Baseline 提高了近十个百分点，这也验证了我们对论元之间相关性的猜想。

表 3　实验结果比较

	Baseline	＋谓词框架信息	＋修正后的谓词框架信息	序列标注
Xue(2008)/%	93.1	94.1	n/a	n/a
本系统/%	78.86	94.34	88.24	88.50

6 结论与展望

本文中,我们全新的语料库上建立了一个中文语义角色分类系统,并将前人基于 PropBank 广泛使用的分类方法应用到新语料库中,在论元分类阶段取得与在 PropBank 上相当的实验结果。从实验结果可以看出,虽然之前的实验方法在网库中也能获得良好的效果,但是我们也验证了之前研究方法中的在提取谓词框架信息方面普遍存在的问题。另外论元框架信息在新语料中对提高正确率的重要作用,说明了此特征在不同语料上的良好扩展性,同时 Baseline 的低准确率也说明其他特征的作用比较弱,可见这些特征在不同语料上的重要性大有不同,因此我们认为特征对语料的依赖性是存在的,因此下一步工作是分别找出两种语料中的最佳特征组合,进行研究每个特征在两种语料上的重要性并找出真正不依赖于标注语料的特征集合。另外本文研究只是在北大网库上的语义角色分类,将来的工作可以继续关注语义角色标注的第一阶段——语义角色识别,并使其与现有的工作结合起来,从而构建一个完整的基于北大网库的汉语语义角色标注系统。

参考文献

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

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

[3] Carreras X, Màrques L. Introduction to the conll-2005 shared task: Semantic role labeling[C]//Proceedings of CoNLL-2005, stroudsburg, PA, USA, 2005: 152-164.

[4] 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.

[5] 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.

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

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

[8] N. Xue. Labeling Chinese Predicates with Semantic Roles[J]. Computational Linguistics, 2008, 34(2): 225-255.

[9] 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.

[10] 丁伟伟, 常宝宝. 基于最大熵原则的汉语语义角色分类[J]. 中文信息学报. 2008, 22(6): 20-26.

[11] 丁伟伟, 常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报, 2009, 23(5): 53-61, 74.

[12] Dowty, D. Thematic Proto-Role and Argument Selection[J]. Language, 1991, 67(3): 547-561.

[13] 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.