两类中文特殊句式的语义角色标注

刘娜　王小捷

北京邮电大学 信息工程学院 北京 100876

E-mail:liuna19832002@gmail..com

摘要： 语义角色标注为谓语动词的论元及附属成分分派语义角色，从而得到句子的浅层语义结构。本文针对两类中文特殊句式“把”字句和“被”字句的句式特点，提出了一种基于规则的语义角色标注方法。实验表明，这种方法使“把”字句和“被”字句的语义角色标注正确率分别达到88.4%和93%。

关键字： 语义角色标注　“把”字句　“被”字句　基于规则

Semantic Role Labeling for two special Chinese clauses

LiuNa　Wang Xiaojie

Department of Information and Engineering , Beijing University of posts and telecommunications , Beijing　100876

E-mail:liuna19832002@gmail.com

Abstract: Semantic role labelling is to assign semantic role for the predicate verb and its ancillary components, in order to acquire the shallow semantic structure.According to the characters of ba-construction and bei-construction----two types of Chinese special sentences, an approach of rule-based for semantic role labelling is proposed. As the experiments shows,the accuracy of semantic role labeling for ba-construction and bei-construction reaches　88.4% and 93%.

Keywords: semantic role labelling　ba-construction　bei-construction　rule-based

1 引言

语义角色标注（Semantic Role Labeling，SRL）是指为谓语动词的论元及附属成分分派其在句子中承担的语义角色，从而得到句子的浅层语义结构。例如“[委员会Agent][明天Tmp]将要[通过V][此议案Passive]。”其中，“通过”为谓语动词，名词短语“委员会”和“此议案”是该动词的论元，分别标记为该动词所表示的动作的施事和受事，而谓语动词的附属成分“明天”标记为动作发生的时间。这里，施事、受事和时间就是语义格系统规定的动词论元和附属成分可能承担的语义角色。在进行这种标注后，句子所表示的事件语义结构就比较清晰了：句子描述的是一个“通过”的动作事件，动作发出者、承受者和动作发生时间分别是“委员会”、“此议案”和“明天”。这种事件语义结构的获得对于进一步的语言理解和语言处理应用，如问答系统、信息抽取、机器翻译等具有重要的价值[16][17]。

目前，人们大多采用有监督的机器学习方法来解决语义角色标注问题[1][2]。Chen等人[5]使用决策树C4.5算法进行语义角色标注；近年出现的随机森林算法是对决策树算法的一种改进，Nielsen等人将其应用于语义角色标注任务[6]。基于支持向量机((Support Vector Machine)的语义角色标注系统获得了较好的结果[7]。Winnow和Perceptron及其各种变形算法等是常用的在线学习算法，也被成功地应用于语义角色标注中[8,9]。在[10]中有其他一些方法运用于语义角色标注。但是这些有监督的方法非常依赖于手工标注的训练集，例如propbank[3]和framenet[4]。而这种人工标记是非常费时费力的，同时，随着语料领域和语体的改变，利用原训练集训练的标注器不能较好地涵盖一些新的语言现象，这将导致标注性能的下降。因此，出现了半监督标注方法，只利用较少的甚至只用无标语料来进行标注器的构造。Robert S. Swier, Suzanne Stevenson[15]采用bootstrapping半监督技术构建了一个语义角色标注器，蔡洁等[18]在此基础上进行改造设计了一个中文的语义角色标注系统，但是标注系统只是对于中文的普通简单句具有较好的性能，而对

于中文中大量存在的特殊句式，如“把”字句、“被”字句的标注性能较低，这是由于这些句式的特殊构造导致的，目前还没有关于此类特殊句式的较好的标注器。本文将在[15]的基础上，针对特殊句式提出了一种基于规则的语义角色标注方法，在提高这些特殊句式标注性能的同时，整体提高半监督标注系统的性能。

本文的安排如下，在第二节介绍“把”字句和“被”字句，并结合语料提出了对这些特殊句式进行分类的方案；基于这种分类，在第三节介绍了“把”字句和“被”字句的语义角色标注方法；第四节是实验评估；最后给出结论，并对后续工作进行了讨论。

2 “把”字句和“被”字句的分类

2.1 “把”字句和“被”字句介绍

“把”字句 “被”字句是现代汉语的常见句式。在语言学上有着广泛的研究。

“把”字句是用介词“把”将动词支配或关涉的对象置于谓语动词之前的句子。基本结构如下：

1. 主语＋把＋宾语＋动词＋其他成分(宾语、补语等)

S＋把＋O＋V＋Other elements(O / C)

结构中的宾语是指“把”所涉及的事物，即“把”的宾语，而不是动词的宾语，且这个宾语必须是有定的，也就是说是交际双方所共知的；动词后必须要有其他成分，包括宾语、补语、动词重叠式(VV)、动态助词“了”、“着”等。例如：

(刮风了)，你把窗户关上吧。(“窗户”是你我共知的，“关”后带补语“上”)

我们把房间打扫打扫吧。(“房间”是有定的，动词用重叠式)

我把那个盒子给她了。(动词“给”后带宾语“她”)

2. 主语＋想 / 要 / 能 / 已 / 没＋把＋宾语＋动词＋其他成分

S＋想 / 要 / 能 / 已 / 没＋把＋O＋V＋Other elements

其中“想”、“要”“能”是能愿动词，还有“会”、“可以”“应该”等也可以用于此结构；“已”还可以用“已经”替代；“没”还可以用“没有”替代。也就是说，能愿动词、表示已然的副词以及否定副词等都应放在“把”字之前，而不能放在动词的前边。举例如下：

我想把剩下的胶卷都照完。

你要 / 应该把口袋里的东西都掏出来。

他能把这些馒头吃光。

我已经把借她的书还回去了。

“被”字句是指用介词“被”、“叫”或“让”等引进动作施事的一种句式。句子的主语是动作的受事。基本结构是：

名（受事）+被+名（施事）+动+其它

例如：

（1）敌人被我们打败了。（我们打败敌人）

（2）他被公司开除了。（公司开除他）

“被”的书面语色彩较浓，口语中常用“叫”“让”；“被”的宾语有时可以不出现，“叫”“让”的宾语则必须出现。例如：

（3）他被老师批评了一顿。（他被批评了一顿）

（4）困难终于被大家克服了。（困难终于被克服了）

2.2 “把”字句和“被”字句分类

我们抽取了CTB5.1的所有“把”字句 、“被”字句，依据“把”字句的句法结构特点，将“把”字句分为3类：

(1) NP+把+NP+VP

例如：他们把市场打乱。

(2) NP+把+NP+VP+NP

例如：中国人把上海称作东方大学城。

(3) NP+把+NP+VP+PP

例如：你把书放在桌子上。

依据语料中“被”字句的句法结构特点，将其分为4类：

（1）NP+被+NP+VP+NP

例如：帛琉被潜水爱好者誉为世界七大海底奇观之首。

（2）NP+被+VP+NP

例如：三百多公顷的田地被划入高铁车站特定区。

（3）NP+被+VP

例如：台湾关系不怎莫被谈论。

（4）NP+被+NP+VP

例如：澳门未被二次大战波及。

针对语料得到的句式分类基本上包括了先前介绍的“把”字句和“被”字句的基本结构，并根据动词后面的成分进一步进行了分类。此处进行分类的目的是在进行语义角色标注时基于不同的类别进行规则设计。

3 “把”字句和“被”字句的语义角色标注

本文针对“把”字句和“被”字句不同类别的句式特点，提出了一种基于规则的方法对其进行语义角色标注。“把”字句和“被”字句的语义角色标注可以分为3步进行，首先是句式分类，进而是组块，最后是语义角色标注。下面逐步进行介绍。

3.1 句式分类

句式分类是指把特殊句式依据其结构分到上一节指定的各种结构类别中，本文采用最大熵分类器，利用句子的词汇和词性信息作为特征进行分类。

3.1.1 最大熵分类器

最大熵模型是最大熵分类器的理论基础，其基本思想是为所有已知的因素建立模型，而把所有未知的因素排除在外。也就是说，要找到这样一个概率分布，它满足所有已知的事实，且不受任何未知因素的影响。

最大熵模型的一个最显著的特点是其不要求具有条件独立的特征，因此，人们加入丰富的对最终分类有用的特征，而不用顾及它们之间的相互影响。另外，相对SVM等基于空间距离的分类方法，最大熵模型能够较为容易地对多类分类问题进行建模，并且给各个类别输出一个概率值结果，便于后续推理步骤使用。

在预测句子类别时会涉及各种因素，假设X就是一个由这些因素构成的向量，变量y的值为句

子类别，则最大概率输出为：

$$Px(y|X)=\frac{1}{Z(X)}\exp(\sum_{i}\lambda i fi(X,y))$$

$$Z(X)=\sum_{y}\exp(\sum_{i}\lambda i fi(X,y))$$

其中，$f_i(X,y)$为最大熵模型的特征，λi是每个特征的权重.

我们使用张乐的最大熵模型工具包和带有高斯先验平滑的L-BFGS参数估计算法[12].

3.1.2 特征选择

“把”字句分类时采用的特征包含词和词性信息,基本特征组成为：把+的+NN+VP+PP+NN，其中，“把”指的是句中“把”这个字，句式的一个标志；“的+NN”代表“把”字和动词之间的成分（名词短语NP），NN表示该名词短语的头名词，如果句中没有“的”字，则其特征为空；VP指句中的主动词；PP指示动词后面是否有介词短语，如果有，则取其头名词作为特征，如果没有，则记为空，此特征为“把”字句第三类的独有特征；NN指示动词后面是否有名词短语，如果有，则取其头名词作为特征，如果没有，则记为空，此特征能很好地表征“把”字句第二类。

“被”字句的分类特征同样采用词和词性信息,基本特征结构为：被+的+NN+VP+NN,其中，“被”指的是句中的“被”字，句式的一个标志；“的+NN”代表“被”字和动词之间的成分(名词短语NP)，NN表示该名词短语的头名词，如果句中没有“的”字，则其特征为空；VP指句中的主动词；NN指示动词后面是否有名词短语，如果有，则取其头名词作为特征，如果没有，则记为空，此特征能很好地对“被”字句的第一二类和第三四类进行区分。

3.2 基于规则组块方法

为进行语义角色标注，先要进行基于功能块的组块。由于“把”字句和“被”字句属于特殊句式，句式中有两个动词，当使用蔡洁[18]的自动组块的方法（C_chunk）进行组块时，得到的效果较差，本文采用基于规则的方法(R_chunk)进行。该方法建立在特殊句式分类的基础上，以“把”字句为例，对于已判定属于一个特定类别(上述三类)的“把”字句，其组块有固定的格式，并分别为不同的类别制定了如下确定各个组块的规则：

(1) NP+把+NP+VP　　BA+DEG+N+VV　　BA+N+VV　　BA+N

(2) NP+把+NP+VP+NP　　BA+DEG+N+VV+N　　BA+N+VV+N

(3) NP+把+NP+VP+PP　　BA+DEG+N+VV+P+N　　BA+N+VV+P+N　　BA+DEG+N+VV+N　　BA+N+VV+N

在对一个输入句子进行组块时，首先要依据上一步对其类别的判断，找到该类别对应的规则。例如，对于“把”字句的第一类：NP+把+NP+VP，首先识别出把字和动词，在把字和动词之间的名词短语构成一个组块,如果在把字前面仍存在一个名词短语，则又构成一个组块。

例如：并/AD 毫不/AD 吝惜/VA 地/DEV 继续/VV 把/BA 它/PN 发扬光大/VV 。/PU

对其判断属于第一类，然后与第一类的3个规则相匹配。首先找到把字（把/BA）和动词(发扬光大/VV)，进行标注，在动词和把字之间查找DEG（“的字”），由于此句话中没有“的字”，所以与第一条规则BA+DEG+N+VV没有匹配上，则进入第二个规则BA+N+VV，可以匹配上，则它/PN就作为一个组块，并进行标注，最后标注结果如下：

并/AD 毫不/AD 吝惜/VA 地/DEV 继续/VV <把/BA><BA> [它/PN_head](objt) [发扬光大/VV_main](VP) 。/PU

3.3 语义角色标注方法

在得到了句子组块信息后，进一步依据句式特点按规则进行语义角色的指派。语义角色标注需要标到各个组块的头词上。

4 实验

4.1 语料

在CTB5.1的子句中，抽取到246句把字句，217句被字句。在因特网上收集一些影评语料（eg. http://www.allmov.com etc），这些语料根据标点符号分成子句的形式，对这些子句进行切分和词性标注[13]。在完成这些预处理的子句中抽取到235句把字句，332句被字句。

4.2 实验设计

采用上述特征用最大熵模型进行句式分类，分类后的句子采用基于规则的方法进行组块，识别出句子中的句法成分，进而进行语义角色标注。

所有的实验均采用5倍交叉验证。

在系统中正确率，accuracy，inaccuracy的定义如下：

$$\text{正确率} = \frac{\text{已分类并分类正确的数目}}{\text{已分类的数目}}$$

$$\text{accuracy} = \frac{\text{已标注语义角色并标注正确的数目}}{\text{已标注语义角色的数目}}$$

$$\text{inaccuracy} = \frac{\text{已标注角色但标注不正确的数目}}{\text{已标注语义角色的数目}}$$

4.3 实验结果

如下是对“把”字句进行分类过程中，选择不同的特征对应的正确率：

	特征	正确率
A	基本特征+相应的词性（无序列）	80.2083%
B	基本特征+相应的词性（有序列）	86.4583%
C	基本特征相应的词性（无序列）	79.1667%
D	基本特征相应的词性（有序列）	82.2917%

Table1：特征选择及其相应的正确率

有序列是指对每一个特征标上序列号，以此对相同的特征进行区分。无序列只是标出特征，并没有顺序。

如下是对“被”字句进行分类过程中，选择不同的特征对应的正确率：

	特征	正确率
A	基本特征+相应的词性（无序列）	71.5596%
B	基本特征+相应的词性（有序列）	87.156%

C	基本特征相应的词性（无序列）	84.4037%
D	基本特征相应的词性（有序列）	93.578%

Table2：特征选择及其相应的正确率

对于已经完成切分和词性标注的语料，如果分别用C_chunk和本文R_chunk进行组块分析，之后分别用蔡洁等[18]以及本文的方法进行语义角色标注，结果如下：

		Accuracy	Inaccuracy
“把”字句	R_chunk	0.884(153-173)	0.116(20-173)
	C_chunk	0.15 (20-132)	0.73 (96-132)
“被”字句	R_chunk	0.93(140-150)	0.07(10-150)
	C_chunk	0.06(8-135)	0.69(92-135)

Table3:基于规则的方法组块和自动组块

括号内的数字，例如(153-173)中的153表示已标注语义角色并标注正确的数目，173表示已标注语义角色的数目，也指示共有多少个待标句法槽。

从上图中可以看出，使用基于规则的方法进行组块得到的结果远远好于自动组块。当使用自动组块时，在一个子句中，可能会出现多种错误，例如识别出多个主语、宾语或谓词，一些功能块过大等等。如上表所示，对“把”字句自动组块后进行语义角色标注，总共有173个待标句法成分，但是系统只能对132个进行标注，只标正确了20个。

如果对于特殊句式，都利用基于规则的方法来进行组块，之后分别用蔡洁等[18]的方法(C_srl)和本文的方法(R_srl)进行语义角色标注，结果如下：

		Accuracy	Inaccuracy
“把”字句	R_srl	0.884(153-173)	0.116(20-173)
	C_srl	0.79(136-173)	0.18(32-173)
“被”字句	R_srl	0.93(140-150)	0.07(10-150)
	C_srl	0.16(24-150)	0.80(120-150)

Table4：使用本文语义角色标注方法和蔡洁等[18]的系统分别进行标注

如上表所示，对“把”字句采用C_srl的方法进行语义角色标注，对173个句法成分，可以标正确136个。但是对“被”字句，效果较差，对150个句法成分只能标正确24个。由于蔡洁等[18]的语义角色标注系统主要针对简单句，识别主动词以及与主动词相关成分，利用映射关系进行语义角色标注。但是由于特殊句式的特殊性，如果仍使用蔡洁等[18]的系统进行标注，正确率大大降低，于是本文采用一种基于规则的方法对其进行标注。

5 总结

对两类中文特殊句式“把”字句和“被”字句采用基于规则的方法进行语义角色标注，最终使“把”字句和“被”字句的标注正确率分别达到88.4%和93%。

由于对句子进行分类后，不同的类别结合规则进行组块，然后进行语义角色标注。句子分类的正确性大大提高了语义标注的正确率。要提高句子分类的正确率，特征选择是关键的一步。

由于本文采用的是基于规则的方法，并没有更多的考虑动词的信息，在处理过程中，可以适当的加入动词的信息。

本文中使用的语料很小，总结出3类“把”字句和4类“被”字句，在今后的工作中，可以扩充语料的规模，看是否包含在这几类中，以及所选用的规则是否适合所有的“把”字句和“被”字句。

参 考 文 献

[1] D.Gildea and M.Palmer. The necessity of syntactic parsing for predicate argument recognition. Proc. of the 40th Annual Conf. of the Assoc. for Computational Linguistics, p. 239–246.

[2] D. Gildea and D. Jurafsky. Automatic labeling for semantic roles. Computational Linguistics, 28(3):245–288, 2002.

[3] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. Submitted to Computational Linguistics.

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In Proceedings of COLING/ACL, pages 86–90, Montreal, Canada,1998.

[5] Chen J, Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments. In: Hinrichs EW, Roth D, eds. Proc. of the EMNLP 2003. Sapporo: ACL, 2003. 41–48.

[6] Nielsen RD, Pradhan S. Mixing weak learners in semantic parsing. In: Lin D, Wu D, eds. Proc. of the EMNLP 2004. Barcelona: ACL, 2004. 80–87.

[7] Pradhan S, Hacioglu K, Krugler V, Ward W, Martin JH, Jurafsky D. Support vector learning for semantic argument classification. Machine Learning Journal, 2005,60(3):11–39.

[8] Carreras X, Màrques L, Chrupala G. Hierarchical recognition of propositional arguments with perceptrons. In: Ng HT, Riloff E, eds. Proc. of the CoNLL 2004. Boston: ACL, 2004.106–109.

[9] Punyakanok V, Koomen P, Roth D, Yih W. Generalized inference with multiple semantic role labeling systems. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL, 2005. 181–184.

[10] Xavier Carreras and Lluis Màrquez, 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling, Proceedings of CoNLL-2005.

[11] Berger AL, Della Pietra SA, Della Pietra VJ. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39–71.

[12] Chen SF, Rosenfeld R. A Gaussian prior for smoothing maximum entropy models. Technical Report, CMU-CS-99-108, 1999.

[13] Zhang Suxiang, Qin Ying, Wen Juan, Wang Xiaojie. 2006. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 158–161, Sydney, July 2006. c 2006 Association for Computational Linguistics.

[14] 王力. 中国现代语法[M]. 中华书局. 1954.

[15] Robert S. Swier, Suzanne Stevenson. Unsupervised Semantic Role Labelling. Proceedings of EMNLP 2004, 95-102.

[16] M. Surdeanu, S. Harabagiu, J. Williams, et al. Using Predicate-Argument Structures for Information Extraction. Proceedings of ACL 2003, 2003

[17] E. Voorhees, D. Tice. The TREC-8 question answering track evaluation. 1999

[18] 蔡洁,张祎挺,罗思明.基于半监督方法的汉语语义角色标注. 中国人工智能学会2007年全国学术大会 CAAI-12. 2007.

[19] 张伯江. 论“把”字句的句式语义[A]. 上海第二届全国汉语配价语法研讨会论文, 1999.

[20] 崔希亮. “把”字句的若干句法语义问题[J]. 世界汉语教学, 1995, 3.

[21] 张济卿. 有关“把”字句的若干验证与探索[J]. 语文研究,2000, 1, 35.