中文语义角色标注的特征工程

刘怀军，车万翔，刘挺

（哈尔滨工业大学计算机学院，哈尔滨 150001）

摘 要： 基于统计机器学习的语义角色标注在自然语言处理领域越来越受到重视，丰富多样的特征直接决定语义角色标注系统的性能。本文针对中文的特点，在英文语义角色标注特征的基础上，提出了一些更有效的新特征和组合特征：例如，句法成分后一个词、谓语动词和短语类型的组合、谓语动词类别信息和路径的组合等，并在 Chinese Proposition Bank(CPB)语料数据上，使用最大熵分类器进行了实验，系统 F-Score 由 89.76%增加到 91.31%。结果表明，这些新特征和组合特征显著提高了系统的性能。因此，目前进行语义角色标注应集中精力寻找丰富有效的特征。

关键词： 语义分析；语义角色标注；特征工程；最大熵分类器

Feature Engineering for Chinese Semantic Role Labeling

Huaijun Liu, Wanxiang Che, Ting Liu

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract: In the natural language processing field, researchers have experienced a growth of interest in semantic role labeling by applying statistical and machine-learning methods. Using rich features is the most important part of semantic parsing system. In this paper, some new effective features and combination features are proposed, such as next word of the constituent, predicate and phrase type combination, predicate class and path combination, and so on. And then we report the experiments on the dataset from Chinese Proposition Bank (CPB). After these new features used, the final system improves the F-Score from 89.76% to 91.31%. The results show that the performance of the system has a statistically significant increase. Therefore it is very important to find better features for semantic role labeling.

Keywords: Semantic Parsing; Semantic Role Labeling; Feature Engineering; Maximum Entropy Classifier

1 引言

语义分析就是根据句子的句法结构和句中每个实词的词义，推导出能够反映句子意义的某种形式化表示。对句子进行正确的语义分析，一直是从事自然语言理解研究的学者们追求的主要目标。随着自然语言处理基础技术，如：中文分词、词性标注、句法分析、机器学习等的逐步成熟，以及语义分析在问答系统、信息抽取、机器翻译等领域的广泛应用，使得其越来越受到重视。

语义角色标注（Semantic Role Labeling, SRL）是目前语义分析的一种主要实现方式，它采用“谓语动词-角

基金资助：基金资助：自然科学基金 60435020, 60575042, 60503072

作者简介：刘怀军（1982-），男，山西人，硕士研究生，hjliu@ir.hit.edu.cn

色”的结构形式，标注句法成分为给定谓语动词的语义角色，每个语义角色被赋予一定的语义含义。例如“[委员会 Agent][明天 Tmp]将要[通过 V][此议案 Passive]。”其中，“通过”是谓语动词，“委员会”、“此议案”和“明天”分别是其施事、受事和动作发生的时间。语义角色标注通常被看作分类问题，目前的研究大多基于有指导的机器学习方法，比如支持向量机（SVM）[1]，最大熵（Maximum Entropy）[2]，SNoW（Sparse Network of Winnows）[3]等。由于各种机器学习方法都已经比较成熟，仅依靠单纯机器学习算法的改进，在性能上很难有质的提高。所以，丰富有效的特征对语义角色标注来说更加重要。

文章第2部分简单介绍了中文语义角色标注的语料库资源。第3部分介绍了中文语义角色标注系统，重点描述其基础特征、扩展特征和一些组合特征。接下来第4部分给出了系统的分析和实验结果的讨论。最后第5部分对本文进行了总结并作了后期工作的展望。

2 语料资源

我们实验中使用来自Chinese Proposition Bank(CPB)的数据。CPB是Upenn基于Penn Chinese Treebank(PCT)标注的汉语浅层语义标注资源，在PCT句法分析树的对应句法成分中加入了语义信息。PCT的标注数据主要来自新华新闻专线、Sinorama新闻杂志和香港新闻

[1]。CPB包含20多个语义角色，相同语义角色对于不同谓语动词有不同的语义含义。其中核心的语义角色为Arg0-5六种，Arg0通常表示动作的施事，Arg1通常表示动作的影响等等。其余的语义角色为附加语义角色，用前缀ArgM表示，后面跟一些附加标记（Secondary Tags）来表示这些参数的语义类别，如ArgM-LOC表示地点，ArgM-TMP表示时间等等[4]。图1是CPB中一个句子的标注实例。我们实验中选取了共760个文档，10,384个句子。其中9,288个句子作训练语料，剩余1,096个句子作测试语料。

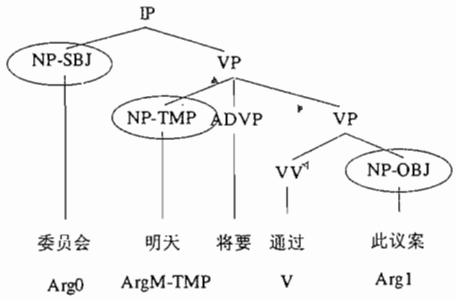

图1 Chinese Proposition Bank中一个句子的标注实例

Fig.1 Syntax tree for a sentence illustrating in Chinese Proposition Bank

3 中文语义角色标注系统

3.1 标注步骤和分类器

语义标注的基本单元可以是句法成分（Constituent）、短语（Phrase）、词（Word）或者依存关系（Dependency Relation）等等，现在多数语义角色标注系统都以句法成分为基本标注单元。句法成分就是句法分析树中非终结节点，比如图1的句法分析树中，NP-SBJ，VP等都是句法成分。因为语义角色多与句法成分对应，所以我们采用句法成分作为标注单元可获得较高的性能。

语义角色标注系统一般通过三个阶段实现[5]：首先，使用一些启发式规则把多数不可能是语义角色的句法成分过滤掉；其次进行语义角色识别，用二元分类器把角色候选分为语义角色和非语义角色；最后使用多类分类器把第二阶段识别的语义角色分到对应的类别。也有系统会加入基于启发式规则的后处理阶段。为了提高系统召回率，避免过滤过程中语义角色的丢失，我们系统没有使用过滤。并且由于最大熵分类器的效率很高，因此我们把角色识别和分类一步实现，属于语义角色的句法成分被分到对应类别，不属于任何角色的句法成分被赋予空类别。

[1] http://www.cis.upenn.edu/~chinese/

3.2 基本特征

特征一直是决定统计自然语言处理系统性能的重要因素。相比特征空间较小的底层自然语言处理任务，比如分词、词性标注和命名实体（NE）识别，语义角色标注任务的一个显著特性就是特征空间很大。在 Xue 等人[6][7]的语义角色标注工作中使用了许多有效的特征，我们实验中也采用了这些特征，并且引入了更多有效的特征。下面我们简要介绍部分基本特征并分析其有效性。

1. 短语类型
2. 中心词及其词性：在中心词提取中，我们使用 Sun 等人[8]的中心词规则（Head rules for Chinese）
3. 子类框架：谓语动词父节点及其子节点。如图 1 中，“通过”的子类框架是 VP→VV-NP-OBJ
4. 谓语动词的类别信息：目前的中文语义角色标注任务中还没有统一规范的动词分类，文章使用 Xue 等人[7]的方法来对动词分类
5. 路径：句法分析树中从当前句法成分到谓语动词的句法路径。如图 1 中，NP-TMP 的路径是 NP-TMP↑VP↓VP↓VV
6. 位置：句法成分在谓语动词前面还是后面，这是一个二值特征

我们实验中，训练和测试数据不是按动词来划分，因此总有一些仅在测试数据中出现的动词。通过统计，测试语料中 1,211 个动词有 227 个动词在训练语料中没有出现过，从训练数据中学习的最大熵模型就不能很好的对这些动词进行预测。CPB 中许多动词有相似的语义结构，比如动词“显现”和“显示”都带两个核心语义角色，主语指描述的实体，宾语指所描述实体的特性。这样，动词类别信息就可以在动词稀疏的情况下正确预测角色类别。

3.3 扩展特征

上节介绍了中文语义角色标注中一些基本特征，这节将描述我们引入的一些新特征。

1. 句法成分的句法功能：CPB 手工标注的句法分析中，短语类型后缀有功能标记，比如-OBJ 表示直接宾语，-SBJ 表示主语等。这些功能标记作为特征能够有效暗示语义角色的类型
2. 句法成分前一个词和后一个词
3. 从句层数：在 Xue 等人[9]有关 Penn Chinese Treebank 的句法标注文章中，对汉语句子提出了几种类型：带补语的子句（CP）、简单子句（IP）、不带疑问词的疑问句（IP-Q）等。我们把句法成分到谓语动词的路径上经历的子句 IP、CP、IP-Q 等的个数作为特征
4. 句法成分到谓语动词的路径上出现的名词短语个数
5. 句法成分和谓语动词的相对位置：我们从三方面来考察他们的相对位置：它们是否兄弟节点关系，是否属于相同动词短语（VP）的儿子节点，是否属于相同子句 IP 或 CP 短语的儿子节点
6. 句法成分和谓语动词的共同最近父节点
7. 谓语动词的搭配模式：CPB 语料数据中，Arg2 大多情况在含有下面 5 种结构的句子中出现：介词-动词结构、使-动词结构、把-动词结构、被-动词结构、动词-数量词结构五种搭配结构。这种搭配模式能够提高对 Arg2 的预测效果，比如对于动词“修到”，Arg2 表示修建的地点，那么在语句“把公路修到山顶上”中“把-动词结构”就暗示句法成分“公路”属于角色 Arg2

许多单一特征对语义角色分类已经非常有效，把这些单一特征组合在一起时，能更有效的增强分类能力。由于最大熵分类器不能够自动地对特征进行组合，因此我们通过基础和扩展特征构造了一些组合特征，比如谓语动词和短语类型、谓语动词和中心词、谓语动词类别信息和路径等。

3.4 后处理阶段

我们系统中会出现两个标注的句法成分位置包含的情况，但在 CPB 标注语料中是不允许的。最大熵分类器能够预测每一个类别的概率，所以当发生包含时，保留了概率最大的那个句法成分。CPB 标注语料中允许标注为相同语义角色的多个句法成分在句子中同时出现。处理这种情况时，我们设置一个阈值，当预测概率都大于阈值时全部保留，否则仅保留概率最大的那个句法成分。语义角色 Arg0-PSR 和 Arg0-PSE 表示持有和被持有关系，在句子中往往成对出现。我们系统的标注结果中，可能出现一个句子只有 Arg0-PSR 或 Arg0-PSE 的情况，当预测概率高于某个阈值时，我们保留旧的标注，否则，把标注更新为最大熵预测的概率次高的那个语义角色类型。

4 实验结果及讨论

4.1 系统分析

我们实验中，首先建立一个基于基础特征的系统，称为基础系统；然后把扩展特征和组合特征逐个加入基础系统中，表 1 列出了加入这些特征后系统性能的变化。在 F-Score 列中，粗体表示性能提高，前面加星号表示性能显著提高。

表 1 每个扩展特征和组合特征对系统性能的影响

Tab.1 The effect of new features on the system performance

特征	Precision（%）	Recall（%）	F-Score（%）
基础系统	90.94	88.62	89.76
逐个加入扩展特征			
+ 句法成分的功能	90.99	88.77	**89.87**
+ 句法成分前一个词	91.15	88.70	**89.91**
+ 句法成分后一个词	91.58	89.14	***90.34**
+ 句法成分前一个词的词性	90.97	88.71	**89.82**
+ 句法成分后一个词的词性	91.05	88.86	**89.94**
+ 从句层数	91.01	88.91	**89.95**
+ 谓词到句法成分的路径上名词短语个数	90.93	88.69	**89.80**
+ 句法成分和谓语动词的相对位置	90.97	88.81	**89.88**
+ 句法成分和谓语动词的共同最近父节点	90.87	88.68	89.76
+ 谓语动词的搭配模式	91.08	88.81	**89.93**
逐个加入组合特征（冒号前后为组合对应的单一特征）			
+ 谓语动词 ： 短语类型	91.64	89.19	***90.40**
+ 谓语动词 ： 中心词	91.48	88.75	**90.09**
+ 谓语动词 ： 位置	91.27	88.79	**90.01**
+ 谓语动词 ： 路径	91.44	88.98	**90.19**
+ 谓语动词类别信息 ： 路径	91.58	89.06	***90.30**
+ 谓语动词的搭配模式 ： 位置	91.01	88.45	89.71
+ 句法成分的句法框架 ： 短语类型	90.81	88.62	89.70
+ 谓语动词 ： 句法成分的句法框架	91.22	88.87	**90.03**
+ 中心词 ： 中心词词性 ： 路径	91.15	88.59	**89.85**
+ 短语类型 ： 路径	91.09	88.75	**89.90**

实验中我们采用 χ^2（自由度为 1）显著性检验[10]，测试数据中角色总数 $n=10,822$，χ^2 检验的上侧 α 分位数 $\alpha=0.10$，基础系统性能为 F_b，加入一个新特征后系统性能为 F_n，则

$$\chi^2=\frac{(nF_b-nF_n)^2}{nF_n}+\frac{(nF_b-nF_n)^2}{n(1-F_n)} \tag{1}$$

则仅当 $\chi^2 \geq \chi_\alpha^2(1)=2.706$ 时，性能 F_n 增加显著。

从表 1 可以看出，加入句法成分后一个词、谓语动词和短语类型的组合、谓语动词类别信息和路径的组合都显著提高了系统的性能 F-Score 值。其它特征或特征组合加入后，除了少数特征和特征组合的加入使得性能降低，

多数都使系统性能提高。

句法成分后一个词能够显著提高系统性能，一方面由于汉语语法的一个重要特点是十分重视词序，词序不同表达的意思就不同；另一方面，句法成分后一个词作为上下文特征，能够反映当前句法成分的特定语境意义。短语类型是一个非常有效的特征，不同句法类型的短语总是趋向于充当不同的语义角色，而当给定句子中谓语动词时，这种特性更加明显，所以谓语动词和短语类型的组合显著提高了系统的性能。比如例句“今年比去年同期增长九十点七亿美元”，对于动词“增长”，其后的数词短语往往趋向于做 Arg2。路径特征在谓语动词已知时非常有效，但两者组合会导致数据稀疏，所以我们采用谓语动词类别信息和路径组合，显著提高了系统的性能。

4.2 结果讨论

我们在基础系统上加入了使性能提高的扩展特征和组合特征，构成了新系统。表2列出了基础系统和新系统的性能。

表2 加入扩展特征和组合特征前后的系统性能

Tab.2 The performance comparison after adding the new features

实验	Precision（%）	Recall（%）	F-Score（%）
基础系统	90.94	88.62	89.76
新系统	92.68	89.97	91.31

从上表可以看出：尽管每个特征单独加入后，系统性能的增加幅度不是很大，但这些特征全部加入后，系统的性能就有了明显的改进，增加了1.55个百分点。

从实验的结果可以看出，在汉语手工标注句法语料上性能能够达到91.31%，主要有下面的一些因素：首先，动词类别信息能够有效提高系统性能。因为汉语中动词一词多义的现象比较少，并且汉语中形容词作谓语的情况很多，这样谓语动词的角色相对单一，句法实现也简单。其次，在中文语义角色标注的训练语料数据中，有更多附加角色 ArgMs，相对少的而又难分辨的核心角色 Arg3，Arg4，这样使得角色的识别和分类更加容易。并且，Penn Chinese Treebank 使用了更加层次化的结构方式，在完全句法分析树中，使用了许多空标记（-NONE-）来表示深层的含义。

通过对分类错误的语义角色进行分析，这些错误主要有如下几方面引起：首先，一般动词的主语（Subject）被标为 Arg0，宾语（Object）被标为 Arg1。但也有一些动词例外，比如“出现”。例如：这支新队伍以新面孔出现在世人面前。其中“这支新队伍”做主语却被标注为 Arg1。其次，占比例较高的角色 Arg2 的召回率较低。在汉语里中，Arg2，Arg3，Arg4 这几类角色非常灵活，对于不同的动词表示不同的含义，这种灵活性增加了分析的难度。下面是角色 Arg2 的例子：

1．他们都给我肯定的答复。 2．中国对外贸易合作部派驻澳门的直属企业。

3．外商独资企业增加了百分之四点一二，达八千四百八十四个。

在上面的三个句子中，例句 1 中 Arg2 表示“给”的接受者；例句 2 中 Arg2 表示“派驻”的地点;例句 3 中 Arg2 表示“增加”的数额。

5 结论

我们使用最大熵（Maximum Entropy）机器学习算法，在 Chinese Proposition Bank(CPB)的语料数据上进行了中文语义角色标注的实验。实验中，除了使用一些基础特征外，也引入了一些新的单一特征和组合特征。我们对引入的每个新特征进行了显著性验证，并且分析了对系统性能贡献的显著性。把这些新特征中有效提高系统性能的特征加入基础系统中后，系统性能明显提高，F-Score 从 89.76%提升到 91.31%。

尽管 CPB 的语料规模很小，中文语义角色标注系统还是达到了很高的性能。一方面取决于手工标注句法分析的高准确度，另一方面汉语有独特的语法特性，还有一个重要因素是我们使用了丰富有效的单一特征和组合特征。比如句法成分后一个词、谓语动词和短语类型的组合、谓语动词类别信息和路径的组合等。因此寻找合适有效的特征一直是语义分析领域研究的重点。

我们目前的工作基于手工标注的句法分析结果，下一步我们将使用自动句法分析进行中文语义角色标注，并考虑加入命名实体、依存关系等信息来进一步改善中文语义角色标注系统的性能。

参考文献：

[1] S. Pradhan, K. Hacioglu, V. Krugler, et al. Support vector learning for semantic argument classification. Machine Learning Journal, 2005.

[2] N. Kwon, M. Fleischman, E. Hovy. Senseval automatic labeling of semantic roles using Maximum Entropy models. R. Mihalcea, P. Edmonds, (Editors) Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain: Association for Computational Linguistics, 2004, 129–132

[3] P. Koomen, V. Punyakanok, D. Roth, et al. Generalized Inference with Multiple Semantic Role Labeling Systems. In Proceedings of CoNLL-2005, 2005, 181-184.

[4] M. Palmer, D. Gildea, P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. 2005, 31(1)

[5] V. Punyakanok, D. Roth, W. Yih. The Necessity of Syntactic Parsing for Semantic Role Labeling. In Proceedings of CoNLL-04, 2004.

[6] N. Xue, M. Palmer. Calibrating features for semantic role labeling. In Proc. of the EMNLP-2004, 2004.

[7] N. Xue, M. Palmer. Automatic semantic role labeling for Chinese verbs. In Proc. IJCAI2005，2005.

[8] H. Sun and D. Jurafsky. Shallow semantic parsing of Chinese. In Proceedings of NAACL 2004, Boston, USA, 2004.

[9] N. Xue, F. Xia. The Bracketing Guidelines for the Penn Chinese Treebank, IRCS Report 00-08 University of Pennsylvania, Oct 2000.

[10] 施雨，李耀武. 概率论与数理统计应用. 西安：西安交通大学出版社, 2000, 153-156.