中文语义角色标注研究概述

南京师范大学文学院　陈菜芳[1]

摘　要：语义角色标注是实现浅层语义分析的一种方式，在问答系统、机器翻译和信息抽取等方面得到了成功地应用，是目前自然语言理解领域中比较热门的一个研究方向。本文介绍了中文语义角色标注语料资源、中文语义角色标注发展现状以及对中文语义角色标注未来工作进行了展望。

关键词：浅层语义分析　语义角色标注资源　语义角色标注

0 引言

语义角色的自动标注是对句子中谓词所支配的语义角色进行自动标注,是对句子进行浅层语义分析的一种方法。语义角色标注技术在大规模语义知识库的构建、问答系统、机器翻译和信息抽取等领域都有着广泛的应用,其深入的研究对自然语言处理技术的整体发展有着重要意义。下面主要从三个方面来介绍中文语义角色标注研究状况：首先，介绍相关的中文语义角色标注语料资源；其次，描述了中文语义角色标注的发展现状；最后，对中文语义角色标注未来的工作进行展望。

1 中文语义角色标注语料资源

语义角色标注离不开语料资源的支持。英语较为知名的语义角色标注资源有 FrameNet、PropBank 和 NomBank 等。中文语义角色标注语料资源主要是从英语语义角色标注语料资源的基础上发展起来或参照其建设的。

Chinese Proposition Bank(CPB)同英文 PropBank 基本类似。在 CPB 中，总共定义了 20 多个角色，只对每个句子中的核心动词进行了标注，所有动词的主要角色最多有 6 个，均以 Arg0～Arg5 和 ArgM 为标记，其中核心的语义角色为 Arg0～5 六种，其余为附加语义角色，用前缀 ArgM 表示，后面跟一些附加标记来表示这些参数的语义类别。它几乎对 Penn Chinese Treebank 中的每个动词及其语义角色进行了标注，国内大多数语义角色标注研究都是基于此资源。

中文 Nombank 是在英文命题库（Proposition Bank）和 Nombank 的标注框架上进行扩展，对中文名词性谓词的标注。中文 Nombank 加入了语义角色层的标注信息，与 CPB 一样，也标注了核心语义角色和附加语义角色这两类语义角色。中文 NomBank 中的角色位置有两类情况：一是角色在以名词性谓词为核心词的名词短语中；二是当以名词性谓词为核心词的名词

[1]陈菜芳，女，南京师范大学 2010 级硕士研究生，研究方向计算语言学

短语作支持动词的宾语时，允许语义角色在名词短语外。

山西大学构建的 Chinese FrameNet 是基于框架语义理论，类似 FrameNet 风格的中文词典。它描述了框架元素的详细句法信息和词汇单元以及参与者框架元素之间的关系。Chinese FrameNet 的架构和英文 FrameNet 相似，并且有许多只是稍作修改直接对英文 FrameNet 进行翻译，但也有一些创新，增加了相应语义角色的汉语名称。目前 Chinese FrameNet 已经有 130 多个汉语框架，还在不断补充。

台湾中研院陈凤仪建立的中文句结构树资料库(Sinica Treebank)。Sinica Treebank 是一个包含语义标记和句法标记的混合语料库。它的基本框架是以讯息为本的格位语法，主要是对小句进行标注。目前已标注了 61 087 个句子，包含了 361 834 个词语。语义角色标记共有 50 多个，基本沿袭了格语法的标记体系，如：受益格、感受格等。

北京大学袁毓林教授组织建设的中文网库，是在北大汉语句法分析树库的基础上进行语义标注的。有着更为细致的语义角色设置，尤其是核心论元，分别在主体论元和客体论元内部各划分出五个子类。具体如下：(一)必有论元：A 主体论元：施事、感事、经事、致事、主事；B 客体论元：受事、与事、对象、系事。 (二)非必有论元：A 凭借论元：工具、材料、方式、原因、目的；B 环境论元：时间、处所、源点、终点、路径、范围、量幅。

董振东主持建立的知网（HowNet）是一个常识知识库，描述对象为汉语和英语的词语所代表的概念，揭示了概念与概念之间以及概念所具有的属性之间的关系。《知网》描述了多种类型的词汇语义关系，涉及了词汇语义的各个方面，着重描述了不同词性的词语所代表的概念之间的语义关系，其中特别重视名词所代表的概念与动词所代表的概念之间的语义关系，也即我们通常称作实体与事件之间的语义关系即语义角色关系，例如作为实体的“医生”和作为事件的“医治”，两者有着“事件”与“施事”的关系。在知网中，800 个事件主要特征中的每一个都标识有一个角色框架。

2 中文语义角色标注的发展现状

2.1 语义角色标注的研究最早关注的是英文，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。出现了一些相关的国际评测：如 CoNLL2004、CoNLL2005、EMNLP-CoNLL2007 和 CoNLL2008 都包含了语义角色标注的任务，同时也促进了语义角色标注研究的蓬勃发展。

2.2 中文语义角色标注的工作开展较晚，最早进行研究的是 Sun 等人，当时因为还没有中文方面的专门语料，所以他们只能先人工标记了包含某些动词的语料然后在此基础上进行

研究。后来，伴随着 Chinese Proposition Bank(CPB)的构建，就有了一些比较系统的中文语义角色标注的工作。国内最早关注语义角色标注是刘挺、于江德等人，不过他们研究的重点是提升英文的语义角色标注的性能。

2.3 语料资源和中文自动句法分析的不理想等因素使得国内中文语义角色标注的研究还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。虽然与英文方面的工作相比，中文语义角色标注方面的研究仍处在开始阶段，但该项工作已引起了许多研究人员的重视。国内的研究工作主要集中在以下四大高校。

北京大学关于语义角色标注的工作主要集中在两个方面：一是基于语义组块分析和词汇语义特征的语义角色标注；二是利用北大网库的标注语料进行语义角色标注的研究。丁伟伟[1]提出了一种基于语义组块分析的语义角色标注的处理策略。该方法将中文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，是一种简化的“语义组块识别——语义组块分类”流程，而不是传统的“句法分析——语义角色识别——语义角色分类”的流程。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性能限制。北大网库构建了一种全新的语义角色标注资源，改变了以往无论中英文研究都基于宾州大学命题库的标注体系的局面。文献[2]的主要目的是将之前的各种研究方法在北大网库的标注语料中进行验证，考察它们在北大网库标注体系中的作用，进而讨论特征的选择对标注体系的依赖性问题，这种在北大网库基础上建立的语义角色分类系统，在语义角色分类阶段取得与在 PropBank 上相当的实验结果。

哈尔滨工业大学主要贡献是在不断优化特征和特征组合的基础上，进行不同方法的实验。文献[3]把汉语的特点与英文语义角色标注特征相结合，构建出一些新的特征和组合特征，如：谓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 CPB 语料数据上使用最大熵分类器进行了实验。文献[4]以 CPB 为实验数据，首次将核方法应用于汉语语义角色标注中，通过对已有特征进行组合或分解，提取了更适用于汉语的新特征，得到了接近英文语义角色标注的性能。文献[5]提出一种基于特征组合和支持向量机的语义角色标注方法。该方法的基本标注单元是句法成分，基本特征集合是从当前基于句法分析的语义角色标注系统中选出高效特征，然后选择基于统计的特征组合方法，利用支持向量机在 CPB 语料上进行分类实验。

苏州大学的研究重点在两个方面：一是名词性谓词语义角色标注，二是以依存关系为标注单元进行语义角色标注。文献[6]和[7]讨论了汉语名词性谓词的语义角色标注特征问题。通过对名词性谓词语义角色标注的研究，探索了新的词汇、句法特征，选取了适合名词性谓词

相关的特征集，用于名词性谓词语义角色标注，同时进一步利用动词性谓词已有的成果，极大地提高了名词性谓词语义角色标注的性能。文献[8]提出标注单元为依存关系的语义角色标注系统，经过依存关系分析、谓词标识、特征抽取、角色识别和角色分类，最终在 CoNLL2008 SRL Shared Task 自动依存分析的 WSJ 测试集取得了较好的结果，结果证明其性能明显好于基于句法分析的 SRL。

山西大学的工作主要是在汉语框架语义知识库(CFN)语料库上进行，文献[9]基于汉语框架语义知识库(CFN)，采用条件随机场模型，将语义角色标注问题通过 IOB 策略转化为以词为基本标注单元的线性序列标注问题，研究了汉语框架语义角色的自动标注。模型以词为基本标注单元，选择词、词性、词相对于目标词的位置、目标词及其组合为特征。从 CFN 的 219 个框架中，挑选那些例句个数相对较多的 25 个框架的 6 692 个例句的语料上进行。对每一个框架，分别按照其例句训练一个模型，同时进行语义角色的边界识别与分类，进行 2-fold 交叉验证。

其他还有南师大的陈丽江[10]利用清华大学的中文树库 (TCT)，通过梅家驹等人编纂的《同义词词林》对谓词、名词进行划分，建立了谓词词表、名词词表和介词词表等来区分语义角色。在标注过程中使用规则确定谓词论元，使用规则和词表判定成分的语义角色，基于决策树分类的算法，对汉语真实文本的语义角色标注进行了实验。

3 展望

可以说，对中文语义角色标注的研究还任重而道远，下一步需要进行的研究工作还很多，集中表现在如下三个大的方面：

3.1 成熟的语义理论。语义角色标注属于语义分析的范畴，离不开语义理论的支持。语义角色标注需要语义角色相关理论、语义分类体系、词汇语义等知识。目前，汉语语义这些相关理论都还不是很成熟。因此，建立合理有效的语义分类体系，系统地总结语法与语义之间的对应关系，是取得突破的关键。

3.2 资源库建设。语料库和知识库是自然语言处理的两大基础性工程，语料库是对真实语言现象的收集，知识库是对语言知识的系统性总结，它们对自然语言处理的质量起着关键性的作用。由于语言现象与语言知识的复杂性，语料库和知识库都十分庞大，一般都需要耗费十年乃至数十年的时间来构建。今后计算语言学工作开展的重点之一就是建立语义层次上的语料库和知识库。

3.3 改进分析方法。自然语言分析处理的方法包括基于规则的方法和基于统计的方法。

这两种方法同样也适用于语义角色标注。如何选择合适的方法，如何将这两种方法有机地结合起来，对语义角色标注任务是至关重要的。而且，无论是基于规则的方法，还是基于统计的方法，它们所采用的技术，以及得到的准确性和效率也同语义角色标注的准确性和实用性相关，这些也需要不断地研究与改进。

参考文献

[1]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，2009.9,VOL23(5).

[2]杨敏,常宝宝. 基于北京大学中文网库的语义角色分类[J].中文信息学报，2011.3,VOL25(2).

[3]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，2007.1,VOL21(1).

[4]车万翔. 基于核方法的语义角色标注研究[D].哈尔滨：哈尔滨工业大学,2008 年.

[5]李世奇,赵铁军,李晗静,刘鹏远,刘水. 基于特征组合的中文语义角色标注[J].软件学报,2011,22(2):222-232.

[6]李军辉,周国栋,朱巧明,钱培德. 中文名词性谓词语义角色标注[J]. 软件学报,2011, 22(8).

[7]徐靖,李军辉,朱巧明,李培峰. 中文名词性谓词语义角色标注的特征研究[J].计算机应用，2011.6,VOL31(6).

[8]汪红林,王红玲,周国栋. 基于依存关系的语义角色标注[J].计算机工程，2009.8,VOL35(15).

[9]李济洪,王瑞波,王蔚林,李国臣. 汉语框架语义角色的自动标注[J].软件学报，2010.4,VOL21(4).

[10]陈丽江. 汉语真实文本的语义角色标注[D].南京：南京师范大学,2007 年.