中文语义依存树库构建及自动分析技术*

邵艳秋[1]，邱立坤[2]，梁春霞[1]，毛 宁[1]
[1]北京城市学院 人工智能研究所，北京 100083；
[2]北京大学 计算语言学教育部重点实验室，北京 100871
E-mail: {yqshao, nanyanglcx, maoning}@ bcu.edu.cn; qlk@pku.edu.cn

摘 要：语义依存分析是一种对句子进行深层语义分析的技术。语义依存树库是依存分析的基础。本文综合了不同学者定义的汉语语义关系体系，面向语义分析的实际应用，设计了一套语义关系体系，该体系中除了常规的语义关系定义，对定语加中心语的短语内部涉及到的语义关系进行了更详细的定义。同时，依据此关系体系，采用自动和手工相结合的方式建立了大规模的汉语语义依存关系树库，并在此树库的基础上构造了语义依存标注模型。

关键词：语义依存；语义关系；语义依存树库；语义依存分析

Construction and Analysis Technology of Chinese Semantic Dependency TreeBank

Shao Yanqiu[1], Qiu Likun[2], Liang ChunXia[1], Mao Ning[1]
[1] Institute of Artificial Intelligence, Beijing City University, Beijing 100083
[2] Key Laboratory of Computational Linguistics, (Peking University) Ministry of Education, Beijing 100871
E-mail: {yqshao, nanyanglcx, maoning}@ bcu.edu.cn; qlk@pku.edu.cn

Abstract: Semantic dependency parsing is a kind of deep semantic parsing technology. The semantic dependency resource is the basis of dependency parsing. This paper integrates some Chinese semantic relation system given by different scholars, and presents a more comprehensive system for semantic dependency parsing. Besides the regular semantic relations, the system defines attribute relations in details. According to the relation system, a large scale Chinese semantic dependency relation treebank is constructed by the combination of automatic and manual means. On the basis of the treebank, one semantic dependency relation labeling model is built and the relation system is tested by the model.

Keywords: semantic dependency; semantic relation; semantic dependency tree bank; semantic dependency analysis

1 引言

语义分析是理解句义的必经之路，是句法分析不能替代的。对句子进行语义分析可以透过多变的句法形式抓住句子的本质，例如句子“他把玻璃杯打碎了。”、“他打碎了玻璃杯。”、“玻璃杯被他打碎了”，这些句子虽然在句法表示上各不相同，但是却可以统一为同一种语义表示形式：“打碎（他，玻璃杯）”，这里“他”是“打碎”这个动作的发出者，“玻璃杯”是动作的承受者。从这个例子可以看出，相对于句法关系而言，语义关系更为稳定，是真正理解句子意义的关键所在。

2 语义依存分析相关研究工作

2.1 语义依存分析的概念

目前对句义方面的研究主要集中在语义角色标注(Semantic Role Labeling，SRL)这种浅层语义分析的任务上，作为向深层语义分析进军的一个过渡阶段，浅层语义分析发挥了一定的作用，但是浅层语义分析对句子意义的理解不够深入，存在一定的局限性，语义依存分析（Semantic Dependency Parsing, SDP）就是一种深层的语义分析。

* 本文受北京大学计算语言学教育部重点实验室开放课题基金(KLCL-1003)，国家自然科学基金项目 NSFC(60873156)，国家社科基金项目（09BYY032）支持。

语义依存分析的理论基础是依存句法理论，它是一种更深层的中文语义表示方式。它融合了依存结构和语义信息，更好地表达了句子的结构与语义关系。语义依存分析提取句子中所有的修饰词与核心词对间的语义关系，句子中的每一个词都有其核心节点（除了整个句子的核心节点外）。语义依存分析是面向整个句子的，而不像 SRL 那样只是处理句子中主要谓词与其论元之间的语义关系。语义依存分析还含有非主要谓词包含的语义信息，如数量（quantity）、属性（attribute）和频率（frequency）等。图 1 给出一个经过语义依存分析的句子实例。图中的每一条弧都连接一对（核心-修饰词）词对，连接弧从核心词出发，指向修饰词。每条弧上都标注有语义依存关系。每一个词都有唯一一个核心词作为其父节点（指向谓语动词的核心词为全句的核心标记 EOS）。

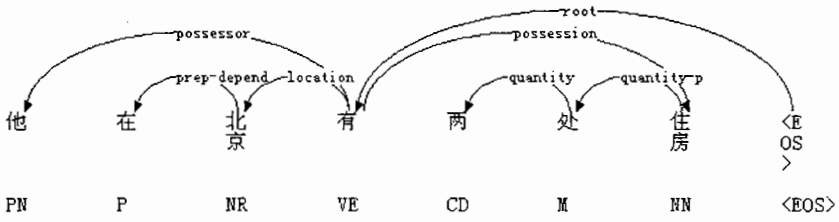

图 1　语义依存树实例

2.2　国内外相关研究

对于语义分析的研究涉及到对语义分析理论、语义关系的定义、语料资源建设、语义分析模型等很多方面。语义分析理论很多，除了语义依存分析外，还包括论元结构、语义角色标注和格语法等，此处不详细赘述。

而对汉语语义关系的定义，不同的语言学家也给出了不同划分，如袁毓林先生提出语义关系标注体系包括论旨角色标记集、逻辑关系标记集和语篇关系标记集，总共有 40 种关系标记[1]；冯志伟先生在 20 世纪 70 年代末和 80 年代初根据依存语法，对汉语动词、形容词和部分名词的论元结构进行了研究，提出了 30 种论元关系[2]；鲁川先生提出的“意合网络”中归纳出了 6 大类，共计 26 种关系[3]；董振东先生在知网中提出事件内部语义关系总计 83 类，分为主语义角色和辅语义角色两大类[4]。

在资源建设方面，目前，国内外尚无公开发表的大规模语义依存分析语料库，而与之相关的语料库主要包括两部分，分别为句法依存语料库和语义角色标注语料库。Penn TreeBank[5]是目前使用最为广泛的英文短语结构句法树库，其具有较高的一致性和标注准确性，已经成为当前研究英语句法分析所公认的训练集和测试集。汉语方面短语结构树库标注方面，较早的是台湾中央研究院标注的 Sinica 树库（繁体）、美国宾夕法尼亚大学 Penn Chinese Treebank 树库，以及清华大学的汉语树库 TCT，党政法、周强等利用核心节点映射表，然后利用规则确定弧的依存关系类型，将 TCT 转化为依存结构[6]，以及哈工大信息检索研究中心自动转换的句法依存树库[7]。PropBank 是宾夕法尼亚大学在 Penn TreeBank 句法分析语料库的基础上标注的语义角色标注语料库。PropBank 只对谓语动词（不包括系动词）的论元进行角色标注，只包含 20 多个语义角色。其中核心的语义角色为 Arg0~5 六种，而相同的核心角色对于不同的谓语动词可能会有不同的语义含义。对于汉语也有相应的 Chinese PropBank 角色标记库。

目前尚不存在针对语义依存分析设计的实用算法，与其最相关的算法是依存句法分析以及在句法分析基础上进行的语义角色标注的相关算法。依存句法分析器，比较经典的方法主要有基于图的方法和基于转移的方法[8]，也有学者试图将二者相结合。而对于语义角色标注的研究，无论是特征选取还是机器学习都有比较深入的研究，如研究者除了词法、句法层面的特征之外也考虑词

汇意义等更深层的特征，以及应用各种不同的机器学习方法。

3 语义依存关系体系构建

通过对比目前各学者提出的语义关系体系，我们认为 HowNet 的语义角色是比较丰富且细致的。但是 HowNet 中定义的主语义角色都是针对动词的语义角色，即针对谓语动词的论元角色所标注的语义关系。而且 HowNet 中没有定义句法关系，对定语类修饰关系定义的语义关系亦不够全面。因此，本文同时参考了鲁川先生与袁毓林先生的语义体系，对 HowNet 进行了扩充、整理，形成了一套新的语义关系体系。

短语内部往往蕴藏着丰富的语义关系表达，是深层语义分析必须要好好解决的问题。所以本文除了常规的主辅语义关系外，还详细定义了定语类的语义关系，包括当动词作为修饰词和动词性名词充当中心词的情况。动词作为名词的修饰节点，如“去世的爷爷”、“被打伤的群众”，如果将这种情况简单地标注为一种修饰关系，就会掩盖这种短语内部词和词之间实质上存在的语义关系，因为这两个短语并不是动词短语，不能表示成诸如“经验者”和“受事”这样的关系，在本文中，针对这种情况，定义使用相应的该名词与动词的语义关系加 r 来表示（代表“反”），譬如“去世的爷爷”，将修饰词“去世”和中心词“爷爷”之间标注成 r-experiencer（反经验者），也就是说是一种动词中心形式（“爷爷去世”）的反面形式。此时弧的箭头指向名词性动词。

当动词性名词或形容词充当名词短语的中心词时，如“企业管理”、“对重点建设项目的支持”，其中“管理”和“支持”均为动词性名词，他们有动词形式，而且作为名词谓词与其动词形式具有相同的语义角色，这时为了区别于动词带名词（如“管理企业”）的情况，本文规定使用该修饰词与其核心词为动词形式的语义关系表示加 j 表达（代表“间接”）语义关系，譬如“企业管理”标注成“j-patient”。

同时，由于 HowNet 中有一些标记出现频率过低或是在实际语料标注中不存在，本文对 HowNet 的语义关系体系进行了部分修改，譬如将 coagent 合并到 agent 中、将 DurationBeforeEvent 和 DurationAfterEvent 合并到 duration 中、增加了 causer 等标记。另外，如果仅定义语义依存关系是无法对句子中某些具有句法功能的成分进行描述的，比如“不但，而且”这样的词。因此我们还定义了如原因、让步、假设等句法语义关系。

总的来说，本文所定义的语义依存关系体系中，共包含主语义角色 29 个，如施事、经验者等主体角色和受事、内容成品、物质成品等客体角色；辅助语义角色 44 个，包含如空间、时间、方式等角色；定语语义角色中包含 19 个直接修饰类的角色，以及前面提到的动词修饰名词的情况和动词性名词为中心词的情况，分别由“r+主语义角色”和“j+主语义角色”来表示反关系和间接关系；除此之外还有 16 个表句法关系的角色。按照我们所定义的语义关系体系，理论上应当有 150 个语义关系，在实际标注语料中出现的语义关系共有 122 个。

4 语义依存树库构建

语义依存树库是研究语义依存分析的资源基础，本文采取两个途径来构造树库，分别为：转化现有依存句法树库以及语义角色标注语料库；人工标注新的语料库。在人工标注语料的过程中，我们采用了主动学习的方法来进行辅助标注，以提高标注效率。

4.1 原有语料库转化

4.1.1 将 Penn Chinese TreeBank 的功能标记转化为语义关系

Penn Chinese Treebank（PCT）是短语结构句法树库，作为本文使用的源语料，应用头节点查

找规则将其转换成依存句法树库。在转化过程中，为了减少标注者的工作量，本文应用了短语结构的功能标记作为参考，编写规则完成了部分语义关系的自动标注。如 PCT 中的“SBJ，OBJ，TMP”等表示“主体、客体、时间”等的功能标记，以及介词短语 PP 所后缀的功能标记如“LOC，DIR，MNR”分别表示“地点、方向、方式”等。

4.1.2 将 Chinese PropBank 中动词的角色框架转化为依存框架

Chinese PropBank (CPB) 是在 PCT 句法分析树的对应句法成分中加入了语义角色信息，核心角色用 Arg0~5 来表示，而它们的具体含义是由 PropBank 中的 Frames（框架）文件给出，本文根据 PropBank 中谓语与其论元的语义关系解释，手工将其转化成语义依存关系，这样就可以根据 PropBank 的语义角色框架建立谓词的语义依存框架，而语义依存框架中的关系是统一且具体的。

4.2 人工标注

4.2.1 手工标注

为了利于人工标注，我们采用了哈工大信息检索中心设计的可视化标注工具同时对该工具进行了功能上的扩充。语料标注人员共有 8 人，检查人员为 4 人。标注人员均为北京大学中文系的硕士研究生。语料的前 1000 句每句都由 2 位标注者同时标注，然后由检查人员进行比对，工程人员在这 1000 句的基础上训练自动标注模型。之后每个句子由一人在自动标注的初始结果上进行手工标注，每人在标注过程中都要对自己的标注结果利用一致性检查工具进行检查，标注结果由 4 位检查人员进行全体句子的一致性检查，标注人员队问题句子进行第二遍标注。最终手工标注句子的总量共为 10400 句。

4.2.2 一致性检查

不同的标注者，其标注思维和习惯会存在差异，比如对相同的一个修饰词-核心词词对，可能会标注得不一致。因此，需要对人工的标注进行一致性的检查，为此，我们设计了一致性检查的工具，其具体功能包括：对词对相同但语义关系不同的情况进行检查；相同语义关系但词对不同；另外，一致性检查还包含一些用词不同，但模式相同的情况应对应相同的语义关系。

4.2.3 自动辅助标注

相对于标弧结构而言，标注弧上关系要更加复杂，本文使用最大熵模型进行弧上关系的自动辅助初始标注。在模型训练中使用的特征，包括孩子节点词、词性，父亲节点词、词性，弧方向，孩子与父亲节点之间的距离，父亲节点左边词的词性，父亲节点右边词的词性，父亲节点的语义依存框架等等。开始先在 1000 个标注质量较高的句子基础上训练出一个最大熵标注模型，用此模型对其他句子进行初始标注，然后再配合人工标注进行修正。随着标注语料的增加，该模型被不断地完善，因为有了一个起始的标注，人工标注的过程被大大简化了，从而提高了标注效率。

5 语义依存分析实验结果

5.1 语义依存分析评价方法

这里语义依存分析的评价方法采用依存句法分析的评价方法，主要有两个评价指标，一个是对依存结构标注的评价指标 Unlabeled Attachment Score（UAS，UA），另一个是对依存关系标注的评价指标 Labeled Attachment Score （LAS，LA）。

$$UAS = \frac{\text{弧正确的词数}}{\text{所有词数}} \times 100\% \quad (1)$$

$$LAS = \frac{\text{依存弧及语义关系都正确的词数}}{\text{所有词数}} \times 100\% \quad (2)$$

5.2 实验结果及其分析

目前对语义依存分析尚没有独特的算法设计，最直接的方法就是将句法分析方法应用于语义依存分析问题上。本文采用开源的 MSTParser 句法分析器进行语义关系的自动标注，只是将其中的句法关系集替换为语义关系集。因为本文需要对前面所设计的语义关系体系进行一个初步的检验，所以对语义依存分析算法并未做更深的研究，对特征集的构造比较简单，只是采用了 MSTParser 默认的一些特征，包括：词、词性、语义关系、父节点等。本文采用了 9000 句进行训练，700 句做开发集，700 句做测试语料。测试的结果，UAS 为 80.18%，LAS 为 65.03%。

从实验结果看，LAS 比较低，我们对依存关系标注的结果进行了分析，由于依存关系数量比较多，表 1 只列出部分语义关系的错误统计结果，包括其 LAS 值，最易与某语义关系相混淆的关系，标注错误次数，及其在所有混淆关系中所占比例。

表 1　部分易混淆语义关系统计

	关系类型	出现次数	LAS（%）	关系错标次数	最易混关系	最易混关系混淆次数	最易混关系比例(%)
主体关系类	agent	572	61.71	72	experiencer	21	29.2
	experiencer	298	62.08	65	agent	42	64.6
	existent	76	28.95	33	possession	21	63.6
客体关系类	patient	350	66.00	60	content	29	48.3
	beneficiary	1	0	1	agent	1	100
直接修饰类	attribute	292	79.11	38	restrictive	27	71.1
	restrictive	1317	58.16	391	attribute	127	32.5
动词修饰名词类	r-agent	46	41.30	21	restrictive	9	42.9
	r-patient	30	46.67	6	attribute	3	50.0
句法关系类	abandonment	4	0	1	succession	1	100
	coordinate	965	45.39	226	succession	100	44.2

由表 1 可以看出，主体类语义关系中 agent 与 experiencer、existent 和 possession 都是易混关系；agent 和 experiencer 主要区别体现在事件主体是自主性还是非自主性，譬如“我们开始吧”和“会议开始了”，前者标为“agent”表示自主性，后者标为“experiencer”表示非自主性。existent 表示存在或消失的事件，多为时空关系，possession 表领属关系中的客体事件。例如动词“有”通常表领属，但在句子“天上有个不明飞行物”，这里“有”和“不明飞行物”间则不是领属关系而是存现体了，对于机器来说这种情况比较难判断，加之主体在句中后置更容易判为客体的领属关系。有些关系出现次数非常少，数据过于稀疏，难于判定，比如损益者 beneficiary。对于直接修饰类中的“attribute”和“restrictive”则互相成为最易混淆关系，restrictive 主要表示语义上一种区别性的分类性质，如：“大型”、“小型”；attribute 主要表示对实体的修饰，如“聪明”、“高大”。restrictive 限定关系比较庞杂，容易被当成直接修饰性关系的垃圾桶，如有些反关系和间接类关系很容易被收录到限定或修饰类关系当中。另外，由于句子中表句法关系的关联词可能不是很明显，导致很多句法关系标记标注错误，如 coordinate 并列关系被标成接续关系 succession。

通过对标注错误的分析，我们感觉在定义语义关系时，对有些语义关系的界限区别应当进一步予以明确，避免模糊不清，出现交叉或包含情况。另外，对于有些出现非常少甚至不出现的语

义关系，应当适当给予合并，使我们所定义的语义关系粒度可以再粗一些。

6 结论

定义合适的语义关系并建立一定规模的语义关系树库是进行深层语义依存分析的基础。本文参考了不同的汉语语义关系体系，对现有体系进行扩充、合并，定义了本文的语义关系体系，采用自动和手工相结合的方式建立了大规模的语义依存树库，在建立起来的语义依存树库基础上，进行了语义依存的自动标注实验，实验结果显示出了某些易混淆的语义关系。未来的工作计划根据实验结果对所定义的语义关系体系进行更加清晰的边界界定，将出现次数较少的语义关系进行适当合并、修正，同时还要研究如何建立更好的语义依存关系自动标注模型。

7 致谢

感谢哈尔滨工业大学信息检索研究室的车万祥老师，王丽杰同学在本文研究过程中给予的大力支持和帮助，感谢北京大学中文系的同学们在语料标注过程中认真、积极的合作。

参 考 文 献

[1] 袁毓林. 基于认知的汉语计算语言学研究. 北京大学出版社，2008.

[2] 冯志伟. 中文信息处理与汉语研究. 北京：商务出版社. 1992.

[3] 鲁川. 现代汉语的语义网络. 电子工业出版社. 1995.

[4] Qiang Dong and Zhendong Dong. Hownet and Computation of Meaning. World Scientific Publishing Company. 2006.

[5] M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In ARPA Human Language Technology Workshop.

[6] 党政法，周强. 短语树到依存树的自动转换研究. 中文信息学报, 2004, 19(3): 21-27.

[7] 李正华，车万翔，刘挺. 短语结构树库向依存结构树库转化研究. 中文信息学报 2008 年第 6 期.

[8] Covington, M. A. A fundamental algorithm for dependency parsing. In Proceedings of 39th Annual ACM Southeast Conference, 2001: 95-102.