复杂名词短语中的语义角色自动标注研究*

李丽
中国传媒大学文学院
Email lilirosa@cuc.edu.cn

摘 要： 汉语句子中包含谓词的名词短语表述的事件关系能够为理解整个句子的意义提供更细致的信息。针对汉语中该类表述事件关系的名词短语，本文提出了一个利用汉语句法结构与语义结构对应关系知识进行语义角色自动标注的方法。本文研究中分析了真实语料中该类名词短语句法语义对应关系的特点，总结了一个分层的对应规则集合，并应用于相应语料上进行了语义角色自动标注实验，取得了77%的自动标注正确率。与同类研究的比较显示了本文所提方法具有的通用性和不依赖语料资源的优越性。

关键词： 语义角色自动标注，复杂名词短语，句法语义对应关系

Automatic Labeling of Semantic Roles on Multiple Nominal Phrase

LI Li
Literature School of Communication University of China
Email lilirosa@cuc.edu.cn

Abstract: Some nominal phrases in Chinese can also describe semantic relations that are useful to understand the meaning of a whole sentence. This paper gives a knowledge driven method of automatic semantic role labeling on Chinese nominal phrases which contain predicate-argument structures. We analyze the relations between syntactic structure and semantic structure of these phrases in real world texts and summarize a set of linking rules. The tests with the linking rules show a total precision of 77% and some superior qualities of our method.

Keywords: Automatic Semantic Role Labeling, Multiple Nominal Phrase, syntactic-semantic relation

1 概述

语义角色的自动标注（Automatic Semantic Role Labeling）是按照选定的目标动词和语义角色描述体系，识别输入语料中担当语义角色的成分并标注适合的语义角色的过程。目前几乎所有的语义角色自动标注研究都关注句子层面上的事件关系，在汉语中有些名词性短语也能表述事件关系，并为理解整个句子的意义提供更细致的事件信息。本文中将这类单元称为复杂名词短语。

在现有的汉语语义角色自动标注研究中还没有针对复杂名词短语进行的较全面的研究。Xue（2008）讨论了使用汉语名词短语树库（Chinese Norm Bank，CNB）作为语料数据，应用统计机器学习方法进行的汉语语义角色自动标注研究。该研究针对以名词化的谓词为中心词语的名词性短语，识别并标注该短语中担当语义角色的句法单元，取得了84%的正确率。本文研究的复

*本文工作得到了国家自然科学基金资助项目（编号：60573185, 60873173）和国家863计划资助课题（编号：2007AA01Z173）支持，特此表示感谢。

杂名词短语的范围更加宽泛，其中就包括了谓词名词化的形式。本文的研究通过人工分析和总结汉语真实语料中复杂名词短语的句法单元与语义角色的对应关系特点，提出一种由汉语句法语义对应知识驱动的语义角色自动标注方法。与统计机器学习方法相比，本文提出的知识驱动的方法几乎不需要训练语料，且具有一定的通用性和扩展性。

2 复杂名词短语语料

语义角色描述体系定义了事件和事件关系的描述方法，因此语义角色自动标注工作都是建立在确定的语义角色描述体系上，针对相应输入语料进行的。语义角色自动标注过程包括两个部分，首先需要识别出输入语料中能够担当目标动词事件框架中各语义角色的句法单元，即论元识别过程；然后按照确定的语义角色标注体系给这些句法单元标注语义角色，即标注角色过程。本文以词汇语义知识库中的复杂名词短语语料为主要研究对象。词汇语义知识库是一个能够用于汉语语义分析的语料资源，其中包括人工标注了语义标记的汉语句子类语料和复杂名词短语类语料。

词汇语义知识库中使用了一个汉语语块自动分析器对语料进行预处理加工，识别需要标注的句法单元。预处理过程的主要目的是提取、识别和描述有可能担当目标动词描述事件中主要语义角色的句法单元。对于复杂名词短语的输入，该识别程序能够标示出目标动词（tgt）、以目标动词为谓语（P）中心词的主语（S）、宾语（O）、部分状语（D）和其他修饰成分（A），以及整个名词短语的中心语（词）（H）。经过预处理的语料再由人工标注目标动词在多个词典中的词义以及预处理过程识别出的句法单元对应的语义角色。需要注意的是，与处理提取出的名词短语中心语（H）并不总是目标动词支配的论元。

词汇语义知识库采用了两种语义角色描述体系对语料标注语义角色，原型角色描述体系和情境知识库[4]。原型角色描述体系是一个由概念概括化、类型化的角色组成的语义角色描述体系，其中的角色由描述事件（场景）意义必需的原型角色和非必需的外围角色组成，对所有的事件采用通用的标记，包括5类角色：原型施事、原型受事、第三方角色和两个外围角色。情境知识库中对各角色的描述可以看作对原型角色描述体系中各角色的具体化定义，因此自动标注过程依照原型角色描述体系中的角色定义和标记进行标注。

I. 原型施事角色，标记x，表示事件中符合施事特征最多的角色；

II. 原型受事角色，标记y，表示事件中符合受事特征最多的角色；

III. 第三方，标记z，描述在事件中已经确定了原型施事和原型受事角色之后，事件对象的提供者或接收方；

IV. 外围角色：其他角色，标记O，表示事件中除主要参与者外的（原型角色和第三方之外的）其他角色，如事件中涉及的工具、方式、原因等；广义时空角色，标记L，描述事件发生的时间、空间关系。

由于在自动识别的结果中有些句法单元不充当事件框架的语义角色，人工标注过程中，将这些单元标记为修饰角色（Q）。

修饰角色，标记Q，描述与事件框架无关的句法单元，可以使用二级标记进行扩展，标示修饰与被修饰关系差异。

我们从词汇语义知识库中选取经过语块自动识别处理的语料和目标动词词义信息作为本文研究中自动标注的输入，依据原型角色描述体系和情景语义描述体系自动标注各句法单元相应的

语义角色。研究中定义复杂名词短语类语义角色自动标注研究中的输入和输出如下：

- 输入：以复杂名词短语为区分单位的汉语真实文本句子，对目标短语进行了正确的词语切分、词性标注、确定了目标动词和词义，并根据该目标动词划分了相关的句法单元、标注了正确的语法功能信息；
- 输出：对自动识别出的各句法单元自动标注语义角色标记和修饰角色标记，输出自动标注结果。

3 自动标注方法

语言的句法结构与语义结构之间存在一定的对应关系。Dowty（1991）提出的原型角色概念和原型角色的论元选择原则反映了英语的一部分句法语义对应关系。在汉语的句子层面，句法语义对应关系表现两方面特点。一方面，句法语义对应关系具有一般性，即大部分汉语句子中主语位置由具有施事特征最多的语义角色占据，而宾语位置由具有受事特征最多的语义角色占据。另一方面，句法语义对应关系受到目标动词和句式的影响，有区别于一般对应关系的特殊情况。从分析动词的角度来看，动词词义表述的事件决定了事件中参与角色的数量和类型，即语义角色的数量和类型；这些参与角色中一部分通过人们的主观透视方式实现在句子当中，成为动词支配的句法论元；动词的价态就是从语言事实中总结出的动词能够支配的句法论元数量；这些句法论元依据语言的语法规则占据句子中特定的位置，表现为句子中具有不同语法功能的句法单元。如此，句法单元与语义角色通过动词的词义、动词的价态联系起来，体现了不同的句法语义对应关系。

复杂名词短语中围绕目标动词的语法关系可以看作是句子层面语法关系的变形，因而可以参考汉语句子层面的句法语义对应关系考察复杂名词短语。表述事件关系的复杂名词短语句法结构变化比较灵活，因而句法语义的对应关系比句子层面的对应关系更加复杂。

表 3.1 语料中句法单元与语义角色的对应关系

	x	y	z	L	O	Q
S	**84.62%**	4.19%	0.08%	9.54%	0.66%	0.90%
O	12.55%	**74.84%**	4.13%	6.65%	1.18%	0.64%
A	15.87%	**68.94%**	0.15%	5.23%	1.17%	8.65%
H	17.30%	23.10%	0.15%	7.57%	0.80%	**51.08%**
D	11.36%	2.24%	2.07%	**54.39%**	26.85%	3.10%

我们首先考察了词汇语义知识库中复杂名词短语标注语料的句法单元和语义角色的对应情况，分析对应关系的一般性。表 3.1 给出了在不区分事件框架的情况下，人工标注语料中各句法单元对应各语义角色的比例。表中显示：状语单元（D）对应到语义角色的规律与句子层面状语单元对应到语义角色的规律非常相似，即大部分对应外围角色（L 和 O）；主语单元（S）的对应情况最为单一，其次为宾语单元（O）和修饰单元（A）；短语的中心语（H）有半数以上标记为修饰角色（Q），即不担当语义角色，而另外半数则较均匀地分布于受事角色、施事角色和外围角色上。对于复杂名词短语的语义角色自动标注研究，重点和难点都在于判定短语中心语单元是否充当事件框架的语义角色，且应充当哪类语义角色上。

根据上述分析，我们首先区分复杂名词短语的句法结构模式，然后结合目标动词的价态信息

和词义信息（确定的事件框架）标注除短语中心语之外的句法单元，最后根据短语模式、目标动词信息、中心语语义类以及其他句法单元的角色标注情况综合判定短语中心语单元应标注的角色标记。表 3.2 给出了复杂名词短语的句法结构变化模式以及在语料中各模式的分布情况。

表 3.2 复杂名词短语结构模式和分布

句法单元序列	形式	待标注单元	比例
A Tgt-H	目标动词名词化为名词短语的中心词	A	27.19%
A Tgt H	目标动词前有修饰单元	A，H	2.24%
Tgt-A H	目标动词被标注为短语中心语的修饰单元	H	37.09%
S （D）Tgt H	短语中有主语（和状语）单元	S，（D），H	10.98%
（D）Tgt O H	短语中有宾语（和状语）单元	O，（D），H	19.09%
S （D）Tgt O H	短语中有主语、宾语（和状语）单元	S，（D），O，H	3.41%

我们提出了一个层级化的对应规则集合应用于复杂名词短语的语义角色自动标注。

1). 模式匹配层面：针对一个复杂名词短语输入，根据识别出的句法单元序列选择模式匹配规则，例如，区分“A Tgt H”模式输入和“A Tgt-H”模式输入。

2). 事件框架层面：需要合理搭配如下四种类型的条件和规则。

i. 事件框架条件：针对一类模式下的输入，获取句子表述的事件框架描述信息，确定应用原型角色描述体系还是情境语义描述体系标注角色，同时获取动词（词义）分类信息。

ii. 动词价态条件：针对一类模式下的输入，获取目标动词的价态信息，区分不同价态的规则。

iii. 一般对应规则：对于输入中的非短语中心语单元，配合事件框架规则和动词价态信息自动标注语义角色。

iv. 短语中心语规则：在标注完其它句法单元之后，根据事件框架规则、动词价态信息和其他句法单元的标注结果判别短语的中心语是否担当语义角色。

3). 句法单元内部信息层面：应用句法单元的中心词代表整个单元，考查各句法单元中心词的词性信息、语义类信息等方面，考查句法单元间的连接词信息，对标注结果进行进一步调整。例如，中心词为“方位词”时，该单元不论处于什么位置，通常都标注为时空角色（L）。再如，对于上面 3e 和 3f 的例子中短语中心词是否担当语义角色需要通过分析短语中心语的中心词才能确定。

总结上述各层面和各类型的知识规则，我们提出了约 40 条在不同模式下，相互搭配使用的规则，构建了自动标注系统。

4 标注实验与结果分析

研究中从词汇语义知识库中获取了 7738 例复杂名词短语输入，这些输入语料共覆盖 835 个目标动词。我们从语料中均匀地抽取了约二分之一的语料用于总结和改进对应知识规则，获得自动标注实验的封闭测试结果。这些语料覆盖 633 个动词，共 3825 例输入。剩下的二分之一语料用于开放测试，测试自动标注方法的通用性。

表 4.1 封闭测试结果

	x	y	L	z	O	Q	Total
正确率	82.18%	84.80%	71.34%	96.23%	84.35%	74.17%	79.84%
召回率	77.65%	77.83%	78.21%	80.39%	54.84%	67.35%	74.57%
F 值	79.85%	81.17%	74.62%	87.60%	66.47%	70.60%	77.12%

表 4.1 给出了封闭测试中各角色的标注正确率。其中，语义角色自动标注 F 值能够达到 77%。系统对各角色的标注正确率都高于召回率，说明目前规则集提取出的角色特征知识能够有效判断个角色的自动标注情况，但是还需要加强从输入中识别和提取更多语义角色特征的能力。由于各角色在预料中所占比例不同，这里重点讨论施事角色（x）、受事角色（y）和修饰角色（Q）的自动标注效果。其中修饰角色的标注正确率最低，说明规则集未能很好地捕捉这类角色表现出来的句法特征，而目前的规则集对施事角色和受事角色的把握更好一些。另外，第三方角色（z）的标注正确率较高，说明担当该角色的句法单元特点较鲜明，易于识别。

表 4.2 开放测试结果

	x	y	L	z	O	Q	Total
正确率	81.25%	84.47%	72.50%	89.66%	84.85%	75.26%	79.91%
召回率	77.69%	78.60%	78.79%	65.38%	50.00%	67.61%	74.69%
F 值	79.43%	81.43%	75.52%	75.62%	62.92%	71.23%	77.21%

我们使用相同的规则集对未用于开发和提取规则的语料进行了开放测试，用以说明该规则集具有一定的通用性。开放测试语料包括 3913 例输入，覆盖了 660 个目标动词，因此其中至少有 27 个目标动词在提取规则所用的语料中未出现过。表 4.2 给出了开放测试的结果。开放测试的标注正确率、召回率和 F 值与封闭测试结果没有明显差异，说明使用现有规则集的自动标注系统在同类语料上自动标注效果无差别。

上述试验结果说明，由复杂名词短语中句法语义对应知识规则驱动的语义角色自动标注方法是可行的，且具有一定的通用性。在本研究中自动标注的难点在对短语中心语的标注。总结表 3.1 和封闭、开放测试的结果，目前的规则集已经能够很好的识别并标注核心语义角色（x， y，z）。由于修饰角色（Q）只可能用于标注短语中心语，因此，Q 角色的数据说明规则集还需提高对短语中心语的分析能力。在后续研究中，及时补充与句法单元中心词信息有关的知识规则将是一个重点改进方向。

5 类似研究比较与总结

我们本文的汉语复杂名词短语的语义角色标注研究和 Xue（2008）中论述的类似研究。Xue（2008）中主要论述了对 “短语中心词是名词化的谓词” 这一类别的语料的研究，这一类型的复杂短语，一部分对应于本文研究中 “A Tgt-H” 的模式，标注的主要任务是判定修饰单元 “A” 是否担当语义角色，且担当何种语义角色。另外一部分是有支撑动词的情况，如 “进行调查”，目标动词为 “调查”。这种情况不属于本文研究中定义的复杂名词短语范围，因此不作考察。Xue

（2008）研究中使用了用于句子层面的自动识别和标注语义角色的所有语料特征，使用机器学习方法进行自动标注，获得了 84.3%的标注正确率（F 值）。该研究中加入了专门用于识别和标注复杂名词短语的特征，如整个复杂短语与句子的关系、与句子中主要谓词的关系等。

本文的研究对象中“短语中心词是名词化的谓词”（“A Tgt-H”模式）这一类别的输入语料有2098条，占语料总量的27%，采用本文介绍的知识驱动的语义角色自动标注方法，获得了80.1%的标注正确率（F值）。尽管正确率低于 Xue（2008）的研究，但是本文描述的方法在两个方面显示了出优越性：首先，本文描述的方法具备基于规则方法的一切优点，基本不需要训练语料，在可用语料相当少的情况下，能够显示更好的稳定性；第二，自动标注系统主要依据原型角色描述体系进行标注，该体系对各语义角色的定义与本文提出的方法能够较好的契合，也有利于本文自动标注方法今后在应用上的扩展和对系统性能的改进。

本文介绍了利用汉语的句法结构与语义结构的对应关系知识，对复杂名词短语进行语义角色自动标注的研究思路和初步成果。表述事件关系的复杂名词短语可以看作是句子的变形，其语义结构与句法结构的对应关系比句子的更复杂。在本文的研究中，复杂名词短语语义角色自动标注问题的重点和难点在于识别和标注短语中心语的角色。我们对 7738 个输入语料中半数的语料进行了知识规则的人工抽取，获得了一个针对复杂名词短语的层级化规则集合。应用该规则集对开发语料进行封闭测试，获得了 77.12%的总体标注正确率。对另外的半数未用于开发的语料进行的开放测试说明了目前的自动标注系统具有一定的通用性。目前的自动标注系统仍需要提高识别短语中心语是否担当事件框架的语义角色的能力，这也是今后改进知识规则集的主要方向。

参 考 文 献

[1] Dong, Z. D, Dong Q. (2002). Hownet. http://www.keenage.com

[2] Dowty,D. 1991. Thematic Proto-roles and argument selection. Language,67:547-619.

[3] Xue, Nianwen.2008. Labeling Chinese Predicates with Semantic Roles. Computational Linguistics, 34(2):225-256

[4] Zhou, Q. 2007.Develop a Syntax Semantics Linking Knowledge Base for the Chinese Language. In Proceedings of 8th Chinese Lexical Semantics Workshop, Hong Kong, May 21-23, 2007

[5] 沈阳主编. 2000.《配价理论与汉语语法研究》. 语文出版社. 2000.