基于多词块的框架元素语义核心词自动识别研究

李双红　李茹　钟立军
山西大学计算机与信息技术学院 太原 030006
E-mail lishuanghong09@gmail.com

摘　要： 抽取一个句子的核心依存图是对句子进行语义理解的有效途径。本文基于汉语框架网给出了表示汉语句子语义骨架的框架核心依存图模型。为了把框架依存图转换成框架核心依存图需要提取每个框架元素的语义核心词。本文提出了基于多词块标注的框架元素语义核心词识别和提取方法，通过对比分析，给出了多词块和框架元素的融合策略，并建立了在多词块标注基础上提取框架元素语义核心词的规则集。基于 6771 个框架元素上的实验结果显示，采用本文的方法和规则集提取框架元素核心词的平均准确率和覆盖率分别为 95.58%和 82.91%。

关键词： 核心依存图，框架元素，语义核心词，多词块，规则库

The Automatic Identification of the Semantic Core Words for Frame Elements Based on Multi-Word Chunk

LI Shuanghong, LI Ru, ZHONG Lijun
School of Computer & Information Technology, Shanxi University, Taiyuan , China　030006
E-mail lishuanghong09@gmail.com

Abstract: It is an effective way to understand the semantic information of a sentence that extracting the frame kernel dependency graph (FKDG) from the sentence. This paper presents the frame kernel dependency graph model that represents the semantic framework of a sentence based on CFN (Chinese FrameNet). It is necessary to extract semantic core words for each frame element (FE) to convert FDG into FKDG. This paper proposes a method to identify and extract core words of frame elements by multi-word chunk technology. By comparative analyzing, we establish the strategy that integrates the multi-word chunk and FE, then the rules system for extracting core words of frame elements based on multi-word chunk labeling is established. The experimental results from 6771 FEs show that the average precision and average coverage are 95.58% and 82.91%.

Key words: Kernel dependency graph, frame elements, semantic core words, multi-word chunk, rule base.

1　引言

对一个句子进行完全语义分析是自然语言处理领域追求的目标，但是目前进行完全的语义分析还不现实。基于汉语框架网，对一个句子只针对一个目标词进行语义角色标注，并在此基础上建立框架语义依存图是进行浅层语义分析的一种有效途径。在框架语义依存图中的框架元素中，不同的词对理解这个语义角色的重要性是不同的。提取一个框架元素的语义核心词对基于核心依存图的语义计算具有十分重要的意义。

关于自动提取短语型框架元素语义核心词的研究在国内外尚未见到，但是已经有一些对短语中心词和短语结构等的相关基础研究。吴云芳在文献[1]中基于中文概念词典CCD, 提出了并列结

基金项目：国家 863 高技术研究发展计划资助项目（2006AA01Z142）；山西省高等学校拔尖人才基金项目。
作者简介：李双红（1984-），男，硕士生，研究方向为计算语言学；李茹（1963-），女，教授，研究方向为自然语言处理。

构中心语的概念，并对并列成分中心语的语义相似性进行了定量考察。周强和俞士汶在文献[2]中对短语的结构和功能做了详细的划分说明。除此之外，对短语的研究主要集中在短语的自动识别上，其中主要研究有张昱琪、周强应用基于实例的方法，对汉语中较常见的9种基本短语的边界及类别进行识别，并利用短语内部构成结构和词汇信息对预测中出现的边界歧义和短语类型歧义进行了排歧处理[3]。赵军、黄昌宁从语言学的角度提出了汉语基本名词短语的概念，并在此基础上设计了一种基于转换的基本名词短语识别模型[4]。周雅倩、郭以昆等使用了基于最大熵的方法识别中文基本名词短语，在Chinese TreeBank 上得到了较高的查全率和准确率[5]。干俊伟、黄德根运用规则和统计相结合的方法构造了一个介词短语识别算法，准确率和召回率都达到了80%以上[6]。

本文首先研究了框架元素语义核心词的内在机理，并给出了形式化的定义。在考察了多词块技术在提取短语型框架元素核心成分上的独特优势后，给出了处理多词块和框架元素这两种体系的融合策略，并建立了在CFN标注和多词块序列标注的基础上提取短语型框架元素语义核心词的规则集，最后通过实验分析了可行性、提取效率及存在的问题。

2 框架核心依存图

在汉语框架网（Chinese FrameNet，简称CFN）中的每个框架包括核心框架元素和非核心框架元素以及若干词元[7]。CFN句子标注，是以框架库为基础，针对一个句子，确定一个词元和该词元所属框架，并给框架元素所在的成分标记框架元素、短语类型和句法功能三种信息。

例1：<reason-pp-par 在/p 村民/n 们/k 的/u 强烈/a 要求/v 下/f > ，/w <agent-np-subj 东门/n 村/n 村委会/j > <man-pp-adv 以/p 村委会/j 选举/v 的/u 方式/n > <tgt=替换 换/v > 掉/v 了/u <old-np-obj 那个/r 腐败/a 的/u 村长/n > 。/w

tgt 是目标词标记，所谓目标词就是属于一个框架的词元。目标词“换”属于“替换”框架；reason (原因)等是框架元素标记；np (名词短语)等是短语类型标记；subj (主语)等是句法功能标记。

框架核心依存图（Frame Kernel Dependency Graph，FKDG）来源于一个句子，它是对这个句子基于一个目标词和依存于这个目标词的各个框架元素的语义依存关系的图形化表示。它由目标词、依存于目标词的框架元素的语义核心成分组成。由给定句子中抽取的核心依存图，可以看作是这个句子深层语义及其各个体现它的语义的核心成分关系的图形化表示[8]。

框架依存图（Frame Dependency Graphs，FDG）与框架核心依存图唯一不同的是它的每个依存项是一个没有提取语义核心词的框架元素。

例1的依存图和核心依存图分别如图1(a)和图1(b)所示。

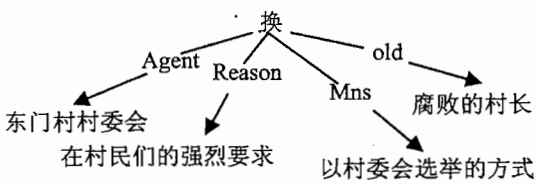

图1(a) 框架依存图

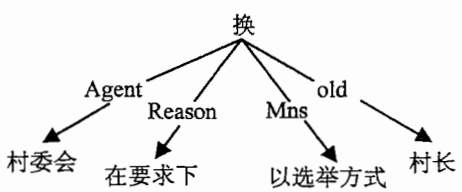

图1(b) 框架核心依存图

本文所做的工作就是要提取框架依存图中每个框架元素的语义核心词从而把框架依存图转换为框架核心依存图，为以后的基于框架核心依存图的语义计算奠定基础。

3 框架元素语义核心词

3.1 短语语法核心与语义核心的区别和联系

对一个短语来说，无论是英语中的HPSG理论[9]还是汉语的多词块理论[10]都认为语法上的核心词只有一个。但我们认为一个短语语义上的核心词可以有多个。在框架元素中，语义上的核心词汇指的是通过这些词可以理解一个框架元素所承担的语义角色的实在内容，并且没有冗余词汇。换句话说，一个框架元素中的词汇可以分成两部分：一部分是核心词汇，这些词汇对理解这个框架元素的语义是必要的；另一部分是修饰核心词汇的词以及各种功能词,如：叹词、语气词、助词、标点符号等。

3.2 框架元素语义核心词形式化定义

（1）名词短语框架元素

定中结构 x+“的”+n 和链式结构 n+ n+…+n ：取最后面的名词；并列结构 n + {n}* +[c] + n：除连词外的所有名词；c 是如“和，与，同”等连词；同位结构 n + n/r 或 r + m/n/r：当名词和代词组合在一起时，取名词为核心词，如果其它词与代词组合在一起则取代词为核心词；准名词性短语 n+“们”：只提取前面的名词。

（2）动词短语框架元素

述宾结构 v+x：取动词和后面宾语核心词；述补结构 v + pp，v+“得”+a/v：取动词和补语的核心词；状中结构 dp/pp/tp/sp+v：取最后的动词；连动结构 v + v：同时提取两个动词；重叠式 v +“了”+ v/v +“一”+ v：提取第一个动词；附加结构 v +“了、着、过”：提取其中的动词。

（3）介词短语框架元素

介词短语分两种情况：对于 p + n /r 和 p + vp，提取介词后面的名词或代词或短语的核心词，对于结构 p+np+f(方位词)/比较词，提取方位词以及 np 短语的核心词。

（4）处所短语框架元素

定中结构 d + f/ n + s：提取后面的方位词或处所词；方位结构 n /r/np + f：提取前面的名词或代词或名词短语的核心词和后面的方位词。

（5）形容词短语框架元素

状中结构 d/dp+a/ap:提取后面的形容词；述补结构和述宾结构，提取前面的形容词和后面的补语或宾语；并列结构 a + {a}* [+ c + a]：按顺序提取所有并列的形容词；附加结构 a+“着，了，过”：提取前面的一个形容词；重叠结构 abab：只提前面的第一个重叠成分。

（6）副词短语框架元素

'地'字结构 ap/mp/vp +“地”：省略“地”；状中结构 d + d/r：提取后面的 d/r；并列结构 dp + {dp}* [+ c + dp]：提取连词以外的短语核心词。

4 多词块与框架元素的融合

汉语多词块（MWC）[10]作为组块体系的重要组成部分，是由两个或两个以上的词语按照一定的关联关系组合形成的信息描述单位。其描述核心是以下三种基本拓扑结构：左角中心结构(LCC)、右角中心结构(RCC)和链式关联结构(CHC)[10]。这三种结构又被细化为八种关系：ZX(右角中心结构)，LN(链式关联结构)，LH（并列关系CHC），PO（述宾关系LCC），SB（述补关系LCC），AD（附加关系LCC），JB（介宾关系LCC），CD（重叠关系）。关于多词块的序列标记等其它的详细介绍请参考文献[10]。

4.1 多词块对提取框架元素语义核心成分的作用

当一个框架元素标注了词性和短语类型后，我们能够知道它的整体语法结构和内部各个词的

独立语言信息，但是不知道内部各个词之间的依存关系。多词块的关系标注给出了这个短语结构整体的依存特性，序列标注具体给出了内部各个成分在这个依存结构中所承担的角色[11]。要找到一个短语中的核心词本质上就是要分析这个短语的语义依存关系，同时充分考虑不同词对理解这个短语语义的重要性，从而提取非修饰性的词汇。

4.2 多词块短语与框架元素短语的融合

在 CFN 标注中“短语”是广义的，既指由两个或两个以上词语组成的结构，也包括由一个词语构成的句法单位。多词块也没有严格的定义，它也是一种广义上的句法结构，所以从内容上讲，短语型框架元素可以认为是多词块。另外，本文讨论的这六种框架元素的短语类型与多词块句法标记的形式和内容也是一致的。

经过 CFN 语料库统计，长度在 1-3 的框架元素占 71.64%，长度在 1-5 的占 85.38%。而在清华大学的多词块语料库中，多词块长度为 1-3 的占 91.87%，长度为 1-5 的占 98.47%。从上面看出，CFN 中框架元素所形成的多词块普遍比较长。实验表明，对于复杂的长短语，从多词块标注中提取语义核心词的难度将会增加。为了解决较长框架元素的多词块标注问题，我们可以先对长框架元素识别出其中的每个基本短语，然后对每个单独的基本短语进行多词块标注，并提取其核心词，最后按先后顺序组合起来。但是这样会增加算法的复杂性，更重要的是，基本短语的识别效率并不是很高。我们制定如下策略来代替基本短语的自动识别：

对于复合名词短语，以“的”、并列连词和顿号等表示并列结构的标志为分割，把复合短语型框架元素分割成简单的短语；对于复合动词和介词短语，以动词和介词为标志分割成基本短语；对于处所短语，以介词和后面的方位词为界进行分割成基本短语。然后对分割出来的每个基本短语进行多词块标注，而不是直接对较长的复合短语进行整体标注。

5 基于多词块序列标注的框架元素语义核心词提取规则

下面给出以关系标记为类别并结合 3.2 中的形式化定义，从大量语料库中抽取出的在多词块序列标注基础上提取不同短语型框架元素核心词的规则集。

1） 一般规则

(1) ZX：包括 np 和 sp 的定中结构，vp、ap 和 dp 的状中结构。提取序列标记中的 R。

(2) LN：包括 np 的链式结构，sp 的方位结构。提取标记为 H 的词，当序列中无 H 时提取所有标记为 J 的词。

(3) LH：包括 np、ap 和 dp 的并列结构，np 的同位结构，vp 的连动结构。按顺序提取所有标记为 J 的词。

(4) PO：包括 vp 和 ap 的述宾结构。①如果有两个或两个以上动词，则以每个动词为界对整个短语进行分割处理：除动词之外的每个成分依赖于它前面的一个动词，如果前面没有动词则依赖于它后面的动词；②动词后面有以序列标记 I 分割的 O 序列时，只考虑 I 后面的 O 序列；③当有两个或两个以上连续的 O 时，则先识别最后一个 O 的词性是否为“f/s/t/q”之一，如果是则提取最后的两个标记为 O 的词，否则只提取最后一个标记为 O 的词；④当没有 O 时，提取标记为 H 和 J 的词。

(5) SB：包括 vp 的述补结构，ap 的述补结构。不提取 I 和 M，别的按顺序提取。

(6) AD：包括准名词性短语，dp 的“地”结构，ap 的“的”结构，vp 和 ap 的附加结构。只提取标记为 H 的词。

(7) JB：包括 pp 的介宾结构。提取规则与 PO 类似，唯一不同的是介词不进入核心词集内。

(8) CD：包括 vp 和 ap 的重叠结构。如果里面含有诸如“不，了，一”的重叠结构，提取这些词前面的词；否则提取除过 I 和 M 的所有实词，然后删除里面重复的词。

2) 不同短语类型不相容的特殊规则

(1) 对于处所短语，如果关系标记为 ZX，并且序列标记为 R 的词的词性为 f 或 s，则提取标记为 R 和 R 前面离 R 最近的且序列标记不是 I 的一个词。

(2) 对于处所短语，如果关系标记为 LN，提取序列标记为 H 和 K 的词。

(3) 在 PO 和 JB 中，如果标记为 I 的词是诸如“和，与，或，连同，及其”等表示并列性质的词，则当作 P 处理，因为这样就防止在并列性质的宾语中只抽取最后的并列成分。

(4) 如果附加结构 AD 中没有 H，则提取除 I 之外的其它成分。

6 实验及结果分析

6.1 实验方案及评价体系

本文的实验语料来自于 CFN 句子库。对从中抽取的 6771 个框架元素按 2：1 分两部分：第一部分为规则抽取集，第二部分为规则测试集。传统的正确率不能细致地对实验做出评价，所以本文给出了一种改进的正确率计算方法。

设一个框架元素的语义核心词组成集合 K,大小为 n；在多词块基础上提取到的语义核心词组成集合 T，大小为 t，其中含有集合 K 中的 k 个词，那么规则集对这个框架元素的语义核心词提取正确率为：

$$\text{正确率}(precision) = \left[\frac{1}{2}\left(\frac{k}{n} - \frac{t-k}{t} - 1\right) + 1\right] \times 100\%$$

$$\text{平均正确率}(avg\text{-}precision) = \frac{1}{l}\sum \text{该类测试集中每个框架元素正确率} \times 100\%$$

其中 l 为测试集中属于该类别的框架元素个数。

$$\text{覆盖率}(coverage) = \frac{1}{l} \times \text{正确率等于100\%的框架元素总数} \times 100\%$$

6.2 实验结果及分析

表 1 不同关系标记的短语型框架元素实验结果

	ZX	LN	LH	PO	SB	JB	AD	CD
avg-precision (%)	93.77	97.12	95.30	95.44	92.17	94.03	95.14	98.53
Coverage (%)	82.42	94.23	94.74	77.22	83.75	71.03	88.89	95.72

表 2 不同短语类型的框架元素实验结果

	np	vp	pp	sp	dp	ap
avg-precision (%)	98.28	92.74	94.28	96.09	99	99.04
coverage (%)	96.08	69.81	70.83	87.5	96	96.15

表 3 相同短语类型的不同长度实验结果对比

	np>5	np<=5	pp>5	pp<=5	vp>5	vp<=5
avg-precision (%)	95.65	99.64	88.92	96.06	89.07	94.46
coverage (%)	95.37	96.43	45.83	79.17	58.54	76.92

从实验结果我们看到，名词短语的框架元素和副词及形容词短语的准确率比较高。其主要原因是：名词短语的框架元素比较规则，长度的影响不大；副词和形容词短语的框架元素大多数是由两个词构成的，并且以附加结构居多，从表 2 中看到，附加结构的正确率也是比较高的。

动词和介词短语框架元素，特别是长度大于 5 的动词和介词短语框架元素，许多含有两个或两个以上动词或介词，并且作宾语的短语结构较为复杂，从而降低了正确率。虽然针对这种情况，我们采取了如 4.2 中所述的分割处理策略，但是由于分割过程中并不一定会分割正确；并且对短语进行分割处理之后，分割出的基本短语的短语类型识别又会造成一定的错误积累。

7 小结与展望

本文立足于框架核心依存图，在研究了短语型框架元素的结构及其语义核心词的基础上，利用多词块标注体系对短语进行序列标注和关系标注，从而建立了一套根据短语的多词块序列标注提取框架元素核心成分的规则体系。最后通过实验验证了通过多词块标注提取短语核心词的可行性和有效性。

根据本文中出现的问题，下一步将重点展开以下方面的研究：首先进一步探讨框架元素语义核心词提取时的粒度选择问题，从而完善短语型框架元素语义核心词的形式化标准。二是通过对比实验，来分析多词块在提取框架元素语义核心词上的效率。最后将重点研究运用 CRF 模型、SVM 等学习算法进行短语型框架元素核心词的自动提取：在这里将探索框架元素的语义角色和句法功能对核心词提取的影响。

参 考 文 献

[1] 吴云芳. 并列成分中心语语义相似性考察[J]. 当代语言学, 2005, 7(4): 305-315.

[2] 周强, 俞士汶. 汉语短语标注标记集的确定[J]. 中文信息学报, 1996, 10(4): 1-11.

[3] 张昱琪, 周强. 汉语基本短语的自动识别[J]. 中文信息学报, 2002, 16(6): 1-8.

[4] 赵军, 黄昌宁. 基于转换的汉语基本名词短语识别模型[J]. 中文信息学报, 1998, 13(2)：1- 8.

[5] 周雅倩, 郭以昆, 黄萱菁, 吴立德. 基于最大熵方法的中英文基本名词短语识别[J]: 计算机研究与发展, 2003, 40(3): 440-446.

[6] 干俊伟, 黄德根. 汉语介词短语的自动识别[J]. 中文信息学报, 2005, 19(4): 17-23.

[7] 郝晓燕, 刘伟, 李茹, 刘开瑛. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报, 2007, 21(5): 96-100.

[8] 俞士汶, 黄居仁. 计算语言学前瞻[A]. Charles J. Fillmore, Josef Ruppenhofer, Collin F. 框架网络与语义、句法联系的表征[C]. 北京: 商务印书馆, 2005.

[9] Carl Pollard and Ivan A. Sag. Head-Driven Phrase Structure Grammar [M]. Chicago: University of Chicago Press, 1994.

[10] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007, 21(3): 21-27.

[11] 党政法, 周强. 短语树到依存树的自动转换研究[J]. 中文信息学报, 2005, 19(3):21-27.