基于汉语框架网的问句语义角色自动标注研究

彭洪保[1] 李茹[1,2] 段建勇[3]

1. 山西大学 计算机与信息技术学院 山西 太原 030006;

2. 山西大学 计算智能与中文信息处理教育部重点实验室 山西 太原 030006;

3. 北方工业大学 信息工程学院 北京 100144

E-mail: hongbao.peng@gmail.com

摘 要: 语义角色标注是近些年来自然语言处理领域的一个新的研究热点。本文针对中文问句的自身语言特点，基于汉语框架网(Chinese FrameNet, CFN)，提出了一种基于词性筛选和层叠条件随机场模型的汉语问句语义角色自动标注方法。该方法根据所需标注问句与语料库原有句子相似程度选择不同标注模型，并对 2011 条中文问句进行了自动标注实验。其中，框架识别准确率为 86.7%，语义角色自动标注结果准确率为 81.1%，召回率为 74.9%。与 D. Gildea 等对英语框架元素语义角色自动标注结果准确率和召回率相比，取得了较好实验效果。

关键词: 汉语框架网，层叠条件随机场，语义角色，自动标注

Automatic Semantic Role Labeling for Questions Based on Chinese FrameNet

PENG Hongbao , LI Ru, DUAN Jianyong

1. School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China;

2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan Shanxi 030006, China

3. School of information engineering North China university of technology, Beijing 100144, China

E-mail: hongbao.peng@gmail.com

Abstract: Semantic Role Labeling is a new rising research in natural language processing field in recent years. According to the characteristics of Chinese questions, a method based on the selection of POS and Cascading conditional random field (CRF) model is proposed to research the automatic labeling of CFN Semantic Role for Chinese Questions. This method selects different labeling model in line with the similarity between the input question and the source sentences in corpus. The automatic labeling experiment for 2011 Chinese questions, in which shows that 86.7% precision for frame identification, 81.1% precision for Automatic Labeling results and 74.9% recall respectively , which performs better than automatic labeling of Semantic Role for English Frame elements studied by D. Gildea.

Key Word: Chinese FrameNet, cascading conditional random field, semantic role, automatic labeling

1 引言

语义角色标注(Semantic Role Labeling)，又称浅层语义分析(Shallow Semantic Parsing)，指的是根据一个句子中的谓词所激起的框架与相关的句子成分之间的语义关系而赋予这些句子成分的语义角色信息。因该任务并不涉及深层次的语义分析和计算，所以称为浅层语义分析。

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：彭洪保（1982—），男，硕士研究生，研究方向为自然语言处理；李茹（1963—），女，教授，研究方向为自然语言处理。

语义角色标注起始于Dan Gildea和DanJurafsky[1]，他们的实验所用语料是Berkeley大学开发的FrameNet[2]。在此之后，语义角色标注逐渐得到了国际的关注，标注语料得到不断地丰富。在FrameNet之后,宾州大学在树库的基础上完成了英文PropBank[3]，并有与之相关的VerbNet等配套语义词典的构建。近年来，且出现了一些相关的国际评测，CoNLL-2004 和CoNLL -2005[4]都包含了语义角色标注的任务。目前，人们大多采用统计学习的方法解决语义角色标注问题。

相对于英语语义角色标注的研究，中文语义角色标注的工作起步较晚。国内对汉语语义角色标注的研究最早起始于刘挺等[5]，不过他们的研究重点主要集中在英文的语义角色标注之上，实验语料来自CoNLL-2005 的评测语料。刘怀军等[6]针对汉语进行了语义角色标注的研究，但主要局限于语义角色分类方面的研究，没有一个公开完整的语义角色标注系统。此外还有吕德新等[7]，他们的研究集中在特定句式方面的研究。不过总的来说，与英文上的工作相比，汉语语义角色标注方面的研究比较少，相关的文章也不是很多。

本文基于汉语框架网(Chinese FrameNet，简称CFN)进行中文问句语义角色自动标注研究。汉语框架网是一个以Fillmore的框架语义学[8][9]为理论基础，以加州大学伯克利分校的FrameNet为参照，根据不同语种的个性加以改造，以汉语真实语料为事实依据的汉语语义词典[10]。汉语框架网络工程主要用于语言学、计算语言学研究及自然语言处理研究。汉语框架语义知识库(CFN)由框架库、句子库和词元库三部分组成。目前，CFN课题组针对汉语 2100 个词元构建了 300 个框架，标注了 21600 条句子；涉及认知领域用词、科普文章常用谓词以及部分中国法律用词。

本文主要研究了中文问句的汉语框架语义角色标注问题。汉语问句具有句子较短、用词固定、结构特征较明显等特点。在此基础上，本文提出了规则方法和统计方法相结合的语义角色标注模型。实验结果证明，该模型可以明显地改善问句中汉语框架语义角色的标注性能。

2 中文问句语料预处理

2.1 针对词性筛选方法的预处理

此处的预处理输入经过 CFN 标注好的中文问句，输出结果为：词序列、词性序列、去虚词后的标注、框架元素信息。

如对已标注好的问句“有/v <thm-np-subj 哪些/r 名人/n><tgt=到达 到/v >过/u <goal-sp-obj 乔家大院/ns>” 经过处理后理想的数据结果为下面四方面信息：

1) 去掉虚词之后的问句词语序列“有 哪些 名人 到 乔家大院”。

2) 去掉虚词之后的问句词性序列“v+r+n+v+ +ns”。

3) 标注结果“有/v <thm-np-subj 哪些/r 名人/n><tgt=到达 到/v > <goal-sp-obj 乔家大院/ns>”

4) 得到 CFN 标注句子中中的框架元素信息<thm-np-subj 哪些/r 名人/n>、<goal-sp-obj 乔家大院/ns>及其所属框架信息。

因为“过”这个词是虚词，所以在语料预处理阶段相应“过”的位置进行删除处理。

2.2 针对 CRFs 模型方法的预处理

为了便于机器学习方法的使用，对句子库中的句子使用“BIO”策略进行标记。如例句：<src-pp-adva 从/p 北京/ns ><tgt=到达 到/v ><goal-sp-obj 五台山/ns >坐/v 火车/n 怎么/r 走/v ？/w 经过处理后，每个词语后面依次的标记序列为词性标记、BIO 标记、框架元素标记、短语类型标记、句法功能标记、相对于目标词的位置（L 表示此词位于目标词左，T 表示此词就是要标注的目标词，R 表示此词位于目标词右）、相对于疑问词的位置。

3 CFN 语义角色自动标注

3.1 CFN 语义角色自动标注流程

本文所采用的 CFN 语义角色标注系统采用两层标注的体系，第一层先用词性筛选的方法进行处理，如果此种处理过程中遇到输入处理的生句子与库中原有句子的最大相似度小于 0.8（当相似度小于 0.8 时，第一层效果不够理想）时转入第二层 CRFs 模型进行自动标注。其流程如下图：

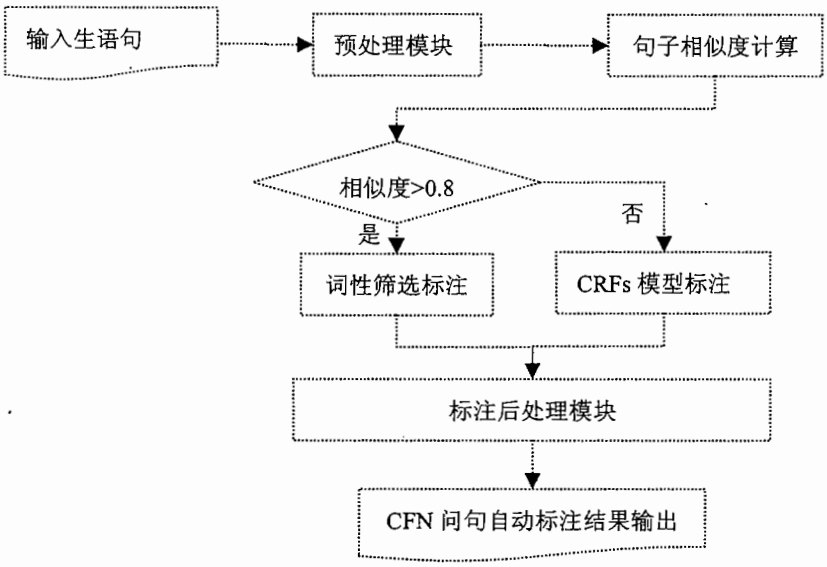

图1　CFN语义角色标注流程图

3.2 基于词性筛选的 CFN 语义角色自动标注

模型建立在已处理过的中文问句标注语料库之上，共分两层：语料库层为 CFN 的自动标注提供标注语料，是 CFN 自动标注的基础和前提，语料库的大小和语料范围直接影响着标注效果；自动标注层对输入的中文问句进行 CFN 语义角色自动标注。这一阶段共分四个步骤：

Step 1 全句词汇匹配，找出语料库中与输入语句词汇完全匹配的句子。如果能找到，则直接取出语料库中句子标注结果作为输出，否者进入 Step 2。

Step 2 全句词性匹配，找出语料库中与输入语句词性序列相同的句子。如能找到，则将语料库中标注问句输出，用输入问句的词汇替换掉语料库中标注问句的词汇，否者进入 Step 3。

Step 3 使用编辑距离方法计算深入问句与训练集问句的相识度，并找出最相似问句，如果能找到，则根据最相似的问句结构对输入问句进行 CFN 语义角色标注。

Step 3 具体方法如下。例如：

Source：　驾车/v 从/p 太原/ns 到/v 五台山/ns 路/n 怎样/r 走/v

找到语料库中与之最相似的目标语句为：

Target：　<src-pp-adva 从/p 运城/ns ><tgt=到达 到/v ><goal-sp-obj 五台山/ns >坐/v 火车/n 怎么/r 走/v ？/w

1）先对两个句子进行去除虚词处理，因虚词在计算词性串相似度时起到的效果很小，还会影响到计算结果。

2）针对 Source 的词性序列从前往后依次与 Target 中的词性进行比较。如果相似度大于某一阈值（此处取阈值为 0.6），则认为词性匹配成功。

其中词性相识度比较因单独某一词性进行比较时所表达的语义现象过于单一和分散，所以

在此处把某一词性左右近邻的词汇加入进行比较（左右近邻个数可以设定，实验结果显示当近邻为 2 时效果较好）。

如：Source 驾车/v 中的 v 与 Target 到/v 中的 v 进行比较时不是 v 和 v 的比较，而是 v p ns（驾车/v 从/p 太原/ns）和 p ns v ns v(从/p 运城/ns 到/v 五台山/ns 坐/v)进行比较，此处得到相似度小于 0.6，所以此处两个动词 v 不匹配。

3）如果 2）中匹配成功则把 Target 中的相应局部结构加入结果集，加入前用 Source 中的相应词汇替换掉 Target 中的词汇。如匹配失败，则把 Source 中的词汇/词性加入结果集。如所有词汇匹配则跳入 4）否则跳入 2）。

4）对 3）处理后的结果进行修正，如括号匹配，虚词添加等工作。

Step 4 对 Step 3 中结果可以进行人工修正，然后存入数据库，扩大语料库规模，提高标注的准确率。此处是一个人机交互的中文问句 CFN 语义角色自动标注辅助系统。

如果以上四个阶段都没有处理输入问句，则用下一节将要介绍的“基于 CRFs 模型的问句 CFN 语义角色自动标注”来进行处理。

3.3 基于 CRFs 模型的 CFN 语义角色自动标注

基于 CFN 的中文问句语义角色标注模块又分为下面四步进行处理：目标词及所属框架的判定、框架元素标注、短语类型标注、句法功能标注。每一步都对应有一个独立的 CRFs 训练模型。

1） 目标词及所属框架的判定

本文对每层的自动标注都设置了多个模板，其中“目标词及所属框架的判定”仅用当前的词、词性和相对疑问词的位置这三个特征，然后逐步的扩大特征的窗口和增加组合特征，最后对自动标注有潜在帮助的特征被加进来，同时，本文去掉那些在实际实验中造成性能下降的特征。

2） 框架元素标注

框架元素识别是个分类的问题，把一个问句中能标出框架元素的词语块分出来进行框架元素标记，并为后续的短语类型标记和句法功能标记服务。不能标出框架元素的词语不进行处理。其选取模板规则是在“目标词及所属框架的判定”的基础上添加当前词的词性与目标词位置的搭配，当前词与左右各两个词的词性与目标词位置的搭配。

3） 短语类型标注

确定短语类型和句法功能自动标注特征模板的方法与确定框架元素自动标注特征模板的方法相同，由于短语类型自动标注是在框架元素自动标注的基础上进行的。最终确定的短语类型自动标注的特征是在选取框架元素自动标注的特征基础上增加了如下特征：当前词的框架元素标注标记结果和前后两个词的框架元素标注标记结果；相邻两个词的框架元素标注标记结果的二元组合特征；相邻三个词的框架元素标注标记结果的三元组合特征。

4） 句法功能标注

由于句法功能自动标注是在框架元素和短语类型自动标注的基础上进行的，其最终确定的句法功能自动标注的特征是在选取短语类型自动标注特征的基础上增加了如下特征：当前词的短语类型标记结果和前后三个词的短语类型标记结果；相邻两个词的短语类型标记结果的二元组合特征；相邻三个词的短语类型标记结果的三元组合特征。对以上各层的特征分别进行组合优化，并不断的调整各层的特征模版，使得标注结果最优。

4 实验结果及分析

4.1 实验工具介绍

实验中使用的 CRF++工具包实现了 CRFL1、CRFL2 以及 Perception 三种方法。在进行参数估计时，该工具包使用了 L-BFGS 算法来迭代寻找 MLE 的最大值。这里，仅使用了 CRFL2 算法，并选取 C=1 进行参数平滑。

4.2 实验结果

本文对 2011 条中文问句进行了自动标注实验，这些句子包含了“到达”、“穿越”、“出发”、“有”、“包含”五个框架。评价识别效果时采用了普遍使用的召回率(R)、准确率(P)和 F-值(F)，定义如下：假设模型标注出的块总数为 Cp，其中正确的块（必须保证左右边界正确，并且块的类型正确）数目为 Cc，在测试集中块的数目为 Co，那么：

准确率P=Cc/Cp；　**召回率**R=Cc/Co；　F**值**=2*P*R/(P+R)

实验对2011条标注问句按8：2分为训练集和测试集。本文发现相似度越高的句子其标注的精度越高，能够取得较好的效果，当相似度小于0.8时效果不理想，没有进行统计。改用CRFs模型进行标注。

本文对“到达”框架下的核心框架元素有目的地 Goal [Goal]、转移体 Theme [Thm]，非核心框架元素有修饰 Manner [Manr]、方法 Means [Mns]、传送模式 Mode_of_transportation [MoT]、轨道 Path [Path]、源点 Source [Src]等；“有”框架下的核心框架元素有所有者 Owner [Own]、拥有物 Possession [Pos]，非核心框架元素有形容 Depictive [Depic]、修饰 Manner [Manr]进行了统计。，其自动标注结果如下表 1 所示：

表 1　层叠条件随机场模型的自动标注结果

	标记层次	准确率P	召回率R	F值
层叠条件随机场	目标词及所属框架自动标注	86.7%	75.6%	80.77%
	第一层框架元素自动标注	83%	75.2%	78.91%
	第二层短语类型自动标注	76.4%	68.4%	72.18%
	第三层句法功能自动标注	72.3%	67.5%	69.82%

本文对词性筛选和层叠条件随机场两种方法做了下面的比较分析。

表 2　词性筛选和层叠条件随机场两种方法结果对比

	语句相似度	准确率P	召回率R	F值
词性筛选	0.8-1	83.3%	86.5%	84.87%
	<0.8	56.3%	62.6%	59.28%
层叠条件随机场	0.8-1	75.4%	69.7%	72.44%
	<0.8	74.9%	69.2%	71.94%

从表 2 中可以看出本文用词性筛选的方法，当语句相似度低于 0.8 时准确率、召回率明显下降。而用层叠条件随机场方法则不会因语句相似度下降而导致准确率、召回率大幅下降。当语句相似度>0.8 时，本文实验模型采用词性筛选模型标注，否则采用层叠条件随机场模型标注。

4.3 实验结果比较分析

图 2 中给出了 5 个框架的标注结果对比。其中，“到达”框架的标注效果最为理想，并且各个框架的核心框架元素的识别效果比较理想，但有部分非核心框架元素和通用非核心框架元素的识别效果很不理想，其主要原因是由于数据稀疏造成的。与核心框架元素、非核心框架元素、通用框架元素关系不大。

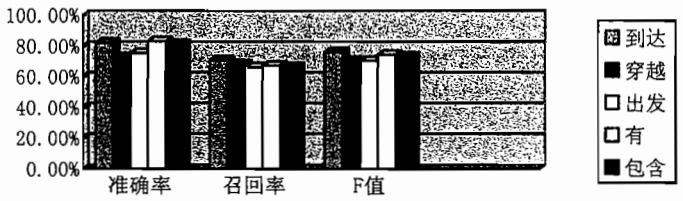

图2　五个框架的自动标注结果比较

5 总结及展望

目前，本文完成了“到达”、“穿越”、“出发”、“有”、“包含”五个框架的语义角色自动标注，这几个框架下所有的在语料库中出现 10 次以上的语义角色自动标注结果的准确率为 81.1%，召回率为 74.9%，其中框架识别准确率为 86.7%。从上述实验结果可以看出，中文问句汉语框架元素的自动标注结果还是比较理想的。

本文的实验能够以较高的准确率自动标注问句的语义角色，下一步计划通过在词性筛选层引进多词块技术进一步提高中文问句框架元素标注的准确率和召回率。相信中文问句的 CFN 语义角色自动标注在问答系统等领域能够取得重大的应用价值。

参 考 文 献

[1] D. Gildea, D. Jurafsky. Automatic labeling of semantic roles [J]. Computational Linguistics, 2002, 28 (3): 245-288.

[2] C. F. Baker, C. J. Fillmore, J. B. Lowe. The Berkeley FrameNet project [C]. In: Proceedings of the 17th international conference on Computational linguistics. Montreal, Canada: 1998, 86-90.

[3] P. Kingsbury, M. Palmer. From TreeBank to PropBank [C]. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas, Spain: 2002, 1989-1993.

[4] X. Carreras, L. Màrques. Introduction to the conll-2005 shared task: Semantic role labeling [C]. In: Proceedings of the 9th Conference on Computational Natural Language Learning. Ann Arbor, MI, USA: 2005, 152-164.

[5] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

[6] 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, 21(1): 79-84.

[7] 吕德新, 张桂, 蔡东风, 朱江涛. 沈阳航空工业学院学报[J]. 2006, 23(1): 44-46.

[8] C. J. Fillmore. Frame semantics and the nature of language [A]. In: Annals of the New York Academy of Sciences: Conference OD the Origin and Development of Language and Speech [C]. 1976, 280: 20-32.

[9] C. J. Fillmore, C. Wooters, C. F. Baker. Building a large lexical data bank which provides deep semantics [A]. In: Proceedings of the 15th Pacific Asia Conference On Language Information and Computation [C]. HongKong: 2001, 3-26.

[10] 刘开瑛, 由丽萍. 汉语框架语义知识库构建工程[A]. 中文信息处理前沿进展, 中国中文信息学会成立二十五周年学术会议论文集[C], 2006, 11: 64-71.