基于层叠条件随机场的句法语义自动标注研究*

陈雪艳，吕国英，李茹，刘伟

山西大学计算机与信息技术学院，山西 太原 030006

E-mail:cxy6410@163.com

摘要：本文提出了一种基于层叠条件随机场的 CFN 句法语义自动标注方法。该方法在低层条件随机场模型中解决了框架元素的识别，将识别结果传递到上层短语类型识别的条件随场模型，再将识别结果传递到上层句法功能识别的条件随机场模型，其低层模型为上层模型提供决策支持，并且在每层自动标注完成后，增加后处理规则去识别那些没有被正确标注的词语。实验选用 CFN 中“陈述”框架下的句子库，实现了基于层叠条件随机场句法语义自动标注的原型系统。

关键字：层叠条件随机场, CFN, 框架元素, 句法语义

Syntax and Semantic Automatic Labeling Based On Cascaded Conditional Random Fields

Chen Xueyan, Lu Guoying, Li Ru, Liu Wei

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

E-mail:cxy6410@163.com

Abstract: This paper presents an approach of Syntactic and Semantic of CFN automatic tagging based on cascaded conditional random fields model. The frame elements are first recognized by the lower random fields model .The result then is passed to the upper phrase type identification with the random fields mode, and then transfer the results to identify the upper syntactic function identification with the random fields mode. The lower model supports the decision for high model and adds post-processing rules for the labeling blocks which is unrecognized after each automatic hierarchical tagging. The experiment selects the sentence of "Statement" frame in CFN, has realized Syntactic and Semantic automatic tagging based on cascaded conditional random fields model.

Keyword: Cascaded Conditional Random Fields, CFN, frame element, syntax and semantic.

1、引言

对句子进行正确的语义分析，一直是从事自然语言理解研究的学者们追求的主要目标。随着自然语言处理基础技术，如：中文分词、词性标注、句法分析、机器学习等的逐步成熟，以及语义分析在问答系统、信息抽取、机器翻译等领域的广泛应用，使得其越来越受到重视。所谓语义分析，指的是将自然语言句子转化为反映这个句子意义（即句义）的某种形式化表示。即将人类能够理解的自然语言转化为计算机能够理解的形式语言，做到人与机器的互相沟通。

我们选用汉语框架语义知识库(Chinese FrameNet，简称 CFN)来表示句子的语义信息。汉语框架语义知识库是一个以框架语义学为理论基础、以真实语料为事实依据的语义词典，其资源用语义 Web 标记语言描述，使其成为一部计算机可读、可理解的语义词典。框架语义学的根本特点

*基金项目：国家 863 高技术研究发展计划资助项目（2006AA01Z142）

是经验主义方法，即根据背景框架的不同，对于属于同一个框架的一类词语，明确其具体的框架元素，不同的框架在框架元素的类型和数量上有差别，而传统格语法的“语义格”是相对于所有词汇而言。框架语义学摆脱了格清单难以确定的问题，具有个性特征的框架元素更适合用来描述千变万化的自然语言语义。提供数量多、类型多的框架元素，并突出框架的个性，适合计算机处理语言的需要。用这种语义知识库表示句子的语义结构，表示结果更深入、语义信息更丰富[1]。

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。提供语义标注句子库，详尽描绘了词汇的框架语义在真实语料中的实际情况，这就使得该语义知识库可以直接应用于句法语义自动标注的研究。

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，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。

例如：句子“本书还集中介绍了现实的浏览器问题”的标注结果如下：

<medium-np-subj 本书 n > 还 d <manr-ap-adva 集中 aq > <tgt=陈述 介绍 v > <null 了 u > <msg-np-obj 现实 a 的 u 浏览器 n 问题 n > 。 w

tgt 是目标词标记，目标词“介绍”属于“陈述”框架； medium(媒介)、manr(修饰)、msg(信息)等是框架元素标记；np(名词短语)、ap(形容词短语)等是短语类型标记；subj(主语)、adva(状语)、obj(宾语)等是句法功能标记；其他标记依此类推。一个框架涉及多个词元，用同一个框架的框架元素集合进行标注。

2、基于层叠条件随机场的句法语义自动标注

自动标注的基本单元可以是句法成分（Constituent）、短语（Phrase）、词（Word）或者依存关系（Dependency Relation）等等。一般认为每个语义角色是与某一句法成分相对应的。也就是说一个语义角色必然对应着一个句法成分，反之未必。因此，现在多数的语义标注系统通常都是以句法成分为基本的标注单元的。例如：基于最大熵分类器的语义角色标注[2]中以句法成分为基本单元，使用最大熵分类模型基于 Prop Bank 语料，研究了语义角色的自动标注问题。这种策略，在句法分析比较成熟的语言（如英文等）上表现较好。然而，在其它语言上，很难自动获得这种深层句法分析的结果，而且现有的句法分析系统，在通用领域表现欠佳。尤其是汉语，很难获得真实语料的句法分析结果，为此以词作为自动标注的基本单元。

2.1、条件随机场模型

条件随机场模型特别擅长处理序列标记问题。它与最大熵模型有相同的特征指数加权形式。对于输入序列x 和输出序列 $y=(y_1,y_2,...,y_n)$，可以定义一个CRF 模型，形式如下：

$$p(y\mid x)=\frac{1}{Z(x)}\exp(\sum_k\sum_i\lambda_k f_k(y_{i-1},y_i,x)+\sum_k\sum_i\mu_k g_k(y_i,x))$$

其中，$f_k(y_{i-1},y_i,x)$ 是观察序列x 中位置为i和i-1 的输出节点的特征，$g_k(y_i,x)$ 是位置为 i 的输入节点和输出节点的特征，λ_k 和 μ_k 是特征函数的权重，Z(x) 是归一化因子[3]。

隐马尔科夫模型（HMM）一个最大的缺点是由于其输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择。最大熵隐马尔科夫模型(MEMM)解决了这一问题，可以任意的选择特征，但由于其是在单个节点的归一化，所以只能找到局部的最优值，同时也带来了标记偏记的问题。条件随机场模型则很好的解决了这一问题，它并不在每一个节点进行归一化，而是所有特征进行全局归一化，因此可以求得全局的最优解。

2.2、层叠条件随机场模型

由于CFN句子标注，是以框架库为基础，针对一个句子先确定要标注的词元和该词元所属的框架，然后再确定标注的范围，最后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。用条件随机场模型直接进行框架元素、短语类型和句法功能的标记，发现存在严重的数据稀疏问题而且消耗的时间和空间特别多，因此需要引入多个层次的条件随机场模型用于识别这类问题。当前建立多层模型主要有两种不同的方法：一种方法是按层叠加建立模型，多个模型之间呈线性组合，称之为层叠模型；另一种方法采用递归方式建立模型，低层模型被嵌入为高层模型的一个子模型，称之为层次模型。相对于层叠模型，层次模型是更复杂的数学模型，其训练复杂度和解码复杂度也远大于层叠模型。而在层叠模型中，是一种松（弱）耦合的关系，各层模型可以独立地建立，整个模型的复杂度与句子的长度成线性关系。而且在层叠模型中，低层模型所产生的错误可以经过适当的过滤和调整后，再将结果传递到高层模型，从而可以避免错误的传播和扩散。基于以上考虑，本文提出了一种基于层叠条件随机场（CCRFs）模型的句法语义自动标注方法。在 CCRFs 模型中，低层的条件随机场模型仅以观察值为条件，识别的结果再传递到高层模型，这样高层模型的输入变量将不仅包含观察值，而且包含了来自低层模型的识别结果，从而为高层条件随机场模型识别提供了决策支持。

2.3、句法语义自动标注步骤

在汉语框架语义知识库中每个框架涉及多个词元，用同一个框架的框架元素集合进行标注，并且每个框架下有相应的句子库，因此首先选择一个框架下的句子进行自动标注实验。“陈述”框架中的框架元素的类型比较丰富，并且句子库中的句子数达到1393句。以“陈述”框架下句子库中句子的单号为测试集，双号为训练集，则训练集为696句，测试集为697句。本实验的自动分词、词性标注软件采用“山西大学分词2000”软件，本软件具有分词、词性、命名体识别功能。以下进行的实验均是在已经识别出目标词的基础上进行的。

2.3.1、使用“BIO”标记例句库

对“陈述”框架下的句子，使用“BIO”进行标记，B表示某个标记块的开始、I表示该标记块的内部、O表示不属于标记块，便于统计机器学习方法使用。

例如：“陈述”框架的一个例句为：

句子库中原句为：

<medium-np-subj 本书 n > 还 d <manr-ap-adva集中 aq > <tgt=陈述 介绍 v > <null 了 u > <msg-np-obj 现实 a 的 u 浏览器 n 问题 n > 。 w

使用BIO标记后变为：本书 n q B-medium-np-subj 还 d q O 集中 aq q B-manr-dp-adva 介绍 v tgt B-tgt 了 u h B-null 现实 a h B-msg-np-obj 的 u h I-msg-np-obj 浏览器 n h

I-msg-np-obj 问题 n h I-msg-np-obj 。 w h O

其中，每个词语后面标记的第1个符号为相应的词性标记，第2符号为相对于目标词的位置（q表示此词位于目标词前，tgt表示此词就是要标注的目标词，h表示此词位于目标词后），第3个符号为标注块的标记（B表示此词是框架标注块的左边界，I表示此词是在标注块内的部分，O 表示此词不是标注的范围）。

具体实验中，标记符号可以根据需要做相应变化，比如上面例句可以变化为：

本书 n q B-medium np subj　还 d q 0 0 0　集中 aq q B-manr dp adva

介绍 v tgt B-tgt tgt tgt　了 u h B-null null null　现实 a h B-msg np obj

的 u h I-msg np obj　浏览器 n h I-msg np obj　问题 n h I-msg np obj 。 w h 0 0 0

实验中第一层标注是直接以句子中的词、词性和相对于目标词的位置作为输入，确定词语相对标记块的边界信息和框架元素（即相对每个词语根据上述的前三列信息确定第四列的信息）；第二层在第一层标注的基础上进行的短语类型标注（即相对每个词语根据上述的前四列信息确定第五列的信息）；第三层在前两层标注的基础上进行句法功能标注（即相对每个词语根据上述的前五列信息确定第六列的信息）。图1 显示了基于层叠条件随机场模型的句法语义自动标注流程：

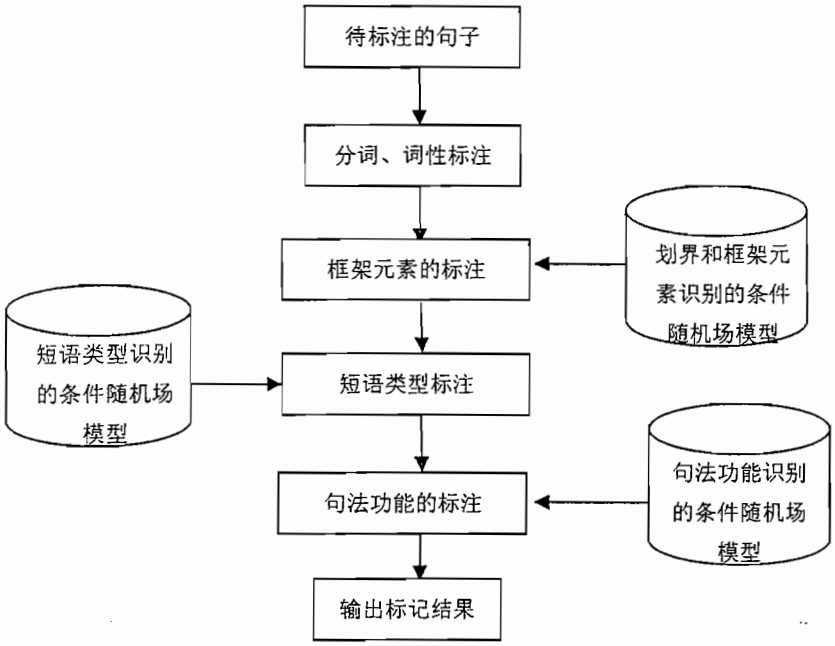

图 1 基于 CCRFs 模型的句法语义自动标注流程图

2.3.2 特征模板的设置

条件随机场模型中一个非常重要的因素是如何针对特定的任务为模型选取合适的特征集合，用简单的特征表示复杂的语言现象. 在进行实验时，最重要的往往是如何选取合适的特征，以达到最好的实验效果。在进行特征选取的过程中，我们遵循如下 2 个原则：一是尽量使标注的结果最优，二是保证特征的数目不会太大，使得算法能够在合理的时间内完成训练。表 1 中，C 表示

当前的词，S 表示当前词的词性，L 表示当前词相对于目标词元的位置，CL 表示当前词和相对目标词位置的组合特征，下标-3，-2,-1,1,2,3 表示特征位置的偏移，另外所有特征模板都使用了标记的二元特征 B。

表 1：特征模板

模版	特征
T1	C，S，L，CL
T2	C，Sn，Ln，CL，（n=-1，0，1）
T3	Cn，Sn，Ln，CL（n=-1，0，1）
T4	Cn，Sn，Ln，CL（n=-1，0，1），$S_{-1}S_0$，S_0S_1
T5	Cn，Sn，Ln，CL（n=-1，0，1），$S_{-1}S_0$，S_0S_1，$L_{-1}L_0$，L_0L_1
T6	Cn，Sn，Ln，CL（n=-2，-1，0，1，2），$S_{-1}S_0$，S_0S_1，$L_{-1}L_0$，L_0L_1
T7	Cn，Sn，Ln，CL（n=-2，-1，0，1，2），$S_{-1}S_0$，S_0S_1，$L_{-1}L_0$，L_0L_1，$S_{-2}S_{-1S0}$，$S_0S_1S_2$，$L_{-2}L_{-1}L_0$，$L_0L_1L_2$
T8	Cn，Sn，Ln，CL（n=-3，-2，-1，0，1，2，3），$S_{-1}S_0$，S_0S_1，$L_{-1}L_0$，L_0L_1，$S_{-2}S_{-1}S_0$，$S_0S_1S_2$，$L_{-2}L_{-1}L_0$，$L_0L_1L_2$

以上是在第一层自动标注时所使用的模版，第二层标注时所使用的模版是在上述各个特征模版的基础上增加了第一层标注的特征，第三层标注时所使用的模版是在上述各个特征模版的基础上增加了第一层和第二层标注的特征。

3、实验结果与分析

由于本次实验共分三层，因此我们对各层的实验结果分别进行了统计，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)。

第一层句中词语相对标记块的边界信息和框架元素自动标注的测试结果如下表2所示，从中可以看出模版T8的标注效果比较理想，因此我们选用模版T8的自动标注结果作为下一层短语类型自动标注的输入；第二层短语类型自动标注的测试结果如下表3所示，从中可以看出模版T7的标注效果比较理想，因此我们选用模版T7的自动标注结果作为下一层句法功能自动标注的输入；第三层句法功能自动标注的测试结果如下表4所示，从中可看出模版T8的标注效果最为理想。

表2 ：框架元素的自动标注结果

模版	准确率	召回率	F 值
T1	81.1%	65.6%	72.5%
T2	83.2%	69.6%	75.8%
T3	84.8%	70.1%	76.8%
T4	84.6%	70.2%	76.8%
T5	85%	69.5%	76.4%
T6	84.8%	70.3%	76.9%
T7	80%	72.8%	76%

T8	84.7%	70.5%	76.9%

表3 ：短语类型的自动标注结果

模版	准确率	召回率	F值
T1	79.1%	65.8%	71.8%
T2	79.4%	66.1%	72.1%
T3	79.4%	66.1%	72.1%
T4	79.4%	66%	72.1%
T5	79.6%	66.2%	72.3%
T6	79.4%	66%	72.1%
T7	79.6%	66.3%	72.3%
T8	79.3%	66%	72%

表4：句法功能的自动标注结果

模版	准确率	召回率	F值
T1	75.1%	62.5%	68.2%
T2	75.2%	62.6%	68.3%
T3	75.1%	62.5%	68.2%
T4	75.2%	62.6%	68.3%
T5	75.3%	62.6%	68.4%
T6	75.3%	62.6%	68.4%
T7	75.3%	62.7%	68.4%
T8	75.4%	62.8%	68.5%

从以上的实验结果发现造成第二层和第三层标注效果下降的主要原因是由于错误的累积，因此在每层自动标注完成后加入一些后处理规则可以减少错误的累积。

对第一层模版T8自动标注结果进行详细的统计分析发现其中的核心框架元素“medium”“msg”“spkr”标注结果的F值在70%左右，而非核心框架元素的标注效果却相差很大，“add”“time”标注结果的F值在70% 左右，但“manr”“top”“depic”的标注效果很不理想，为此我们提出了一些后处理规则来提高“manr”“top”的标注效果，后处理规则如下：

（1）在目标词前，如果有“ap”+“的”、“ap”+“地”、“ap”或“a”的结构且第一层标注时识别为“O”（块外成分时）则识别为“manr”。

（2）如果识别出的标注块是由“对 p”开始的，则把此标注块的框架元素标为“top”。

对第二层模版T7的自动标注结果提出了如下的一些后处理规则：

（1）如果识别出的标注块是由“m+p”或“m” 构成的则把此标注块的短语类型识别为 “mp”。

（2）如果识别出的标注块是“……中 nd”的结构则把此标注块的短语类型识别为“sp”。

对第三层模版T8的自动标注结果提出了如下的一些后处理规则：

（1）如果一个句子中有标注块被标为“B-supp”，则把此句中目标词前的所有句法功能（除“ext”外）加上后缀“_s”，例如“subj”变为“subj_s”。

在每层自动标注结果都加入后处理规则，其第三层的模版T8的标注结果如下表5，它的F值提高了1.5%。

表5：加入后处理规则后的第三层模版T8标注结果

模版	精确率	召回率	F值
T8	64.1%	77%	70%

4、 结论

汉语的语义分析一直是一个比较困难和有挑战性的工作。我们使用层叠条件随机场模型，对CFN语料库中的“陈述”框架下句子库中的句子进行了句法语义自动标注的实验。实验中除了使用层叠条件随机场模型外，还使用了一些后处理规则去识别那些识别效果不好的标注单元。由于CFN的框架库和句子库正在构建中，而且其它框架下的句子库中的句子数目有限，不能代表大多数的语言现象。因此我们要不段的扩大其他框架下句子库中的句子的数目，使其能够用于统计方法的自动标注研究。然后将设计完成一个自动的句法语义标注器，为进行大规模真实文本的语义信息标注提供有力支持。

参考文献：

[1] 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．中文信息学报，2007.5，21（5）．

[2] 刘 挺，车万翔，李 生．基于最大熵分类器的语义角色标注．软件学报，2005.

[3] Lafferty, J, McCallum, A, & Pereira, F (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. ICML.

[4] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. Computational linguistics, 2001, volume 99, number 9, pp 1-42.

[5] McCallum, A. (2003). Efficiently inducing features of conditional random fields. Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence.

[6] 刘怀军，车万翔，刘挺．中文语义角色标注的特征工程．第三届学生计算语言学研讨会论文集，2006，沈阳.

[7] Fleischman, M. and E. Hovy: 2003. ‘A Maximum Entropy Approach to FrameNet Tagging’. In: Proceedings of the Human Language Technology Conference. Edmonton, Canada.

[8] N. Xue, M. Palmer. Automatic semantic role labeling for Chinese verbs. In Proc. IJCAI2005，2005.

[9] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. Comput. Linguist, 28(3):245-288, 2002.

[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.