

汉语框架语义角色的自动标注*

李济洪¹⁺, 王瑞波¹, 王蔚林², 李国臣³

¹(山西大学 计算中心,山西 太原 030006)

²(山西大学 数学科学学院,山西 太原 030006)

³(山西大学 计算机与信息技术学院,山西 太原 030006)

Automatic Labeling of Semantic Roles on Chinese FrameNet

LI Ji-Hong¹⁺, WANG Rui-Bo¹, WANG Wei-Lin², LI Guo-Chen³

¹(Computer Center, Shanxi University, Taiyuan 030006, China)

²(School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China)

³(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

+ Corresponding author: E-mail: lijh@sxu.edu.cn

Li JH, Wang RB, Wang WL, Li GC. Automatic labeling of semantic roles on Chinese FrameNet. Journal of Software, 2010,21(4):597-611. <http://www.jos.org.cn/1000-9825/3756.htm>

Abstract: Based on the semantic knowledge base of Chinese FrameNet (CFN) self-developed by Shanxi University, automatic labeling of the semantic roles of Chinese FrameNet is turned into a sequential tagging problem at word-level by applying IOB (inside/outside/begin) strategies to the exemplified sentences in CFN corpus, and the Conditional Random Fields (CRF) model is adopted. The basic unit of tagging is word. The word, its part of speech, its relative position to the target word, the target word, and their combination are chosen as the features. Various model templates are formed through optional size windows in each feature, and the orthogonal array within statistics is employed for screening of the better template. All experiments are based on the 692 exemplified sentences of 25 frames selected from CFN corpus. The separate model is trained for each frame on its exemplified sentences by 2-fold cross-validation, and the processing of identification and classification for the semantic roles are taken simultaneously. Finally, with the target word given in a sentence, as well as the frame name of the target word, the experimental results on all 25 frames data for the precision, the recall, and *F1*-measure are 74.16%, 52.70%, 61.62%, respectively.

Key words: Chinese FrameNet; semantic role labeling; orthogonal array; feature selection; conditional random fields

摘要: 基于山西大学自主开发的汉语框架语义知识库(CFN),将语义角色标注问题通过 IOB 策略转化为词序列标注问题,采用条件随机场模型,研究了汉语框架语义角色的自动标注模型以词为基本标注单元,选择词、词性、词相对于目标词的位置、目标词及其组合为特征,针对每个特征设定若干可选的窗口,组合构成模型的各种特征模板,

* Supported by the National Natural Science Foundation of China under Grant No.60873128 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z142 (国家高技术研究发展计划(863))

Received 2008-11-22; Revised 2009-06-01; Accepted 2009-10-14

基于统计学中的正交表,给出一种较优模板选择方法.全部实验在选出的 25 个框架的 6 692 个例句的语料上进行.对每一个框架,分别按照其例句训练一个模型,同时进行语义角色的边界识别与分类,进行 2-fold 交叉验证.在给定句子中的目标词以及目标词所属的框架情况下,25 个框架交叉验证的实验结果的准确率、召回率、F1-值分别达到 74.16%,52.70%和 61.62%.

关键词: 汉语框架语义知识库;语义角色标注;正交表;特征选择;条件随机场

中图法分类号: TP391 **文献标识码:** A

1 引言

近年来,自然语言处理开始逐步进入语义分析阶段,其研究重点是句义分析.所谓句义分析是指根据句子中每个实词的词义和关系推导出能够反映这个句子意义(即句义)的某种形式化表示.语义角色的自动标注是对句子中谓词所支配的语义角色进行自动标注,是对句子进行浅层语义分析的一种方法.近几年来,国际上举行过 6 次语义角色自动标注的评测,吸引了众多的研究者参与其中.为此,2008 年自然语言研究的国际上权威学术期刊《Computational Linguistics》出了一期“语义角色标注研究”的专辑.国内近年来有关语义角色标注的研究也越来越得到同行的重视,在重要的学术会议上成果渐增.语义角色标注技术在大规模语义知识库的构建、机器翻译、信息提取、问答系统、信息检索等应用领域都有着广泛的应用,其深入的研究对自然语言处理技术的整体发展有着重要意义.

1.1 汉语框架网知识库

汉语框架网络(Chinese FrameNet,简称 CFN)^[1]工程是以 Fillmore 的框架语义学为理论基础、以加州大学伯克利分校的 FrameNet^[2]为参照、以汉语真实语料为依据的供计算机使用的汉语词汇语义知识库.框架语义学认为,“框架(frame)”是一个与激活性语境相一致的结构化范畴系统,是存储在人类认知经验中的图示化情境.框架中的各种参与者称为框架元素(frame element),它们在使用中与语义角色、格角色等概念相当.框架语义学认为,框架是词语理解的背景和动因,因此可以根据背景框架的不同,对词语,主要是动词、形容词和事件名词进行分类描述.与传统的语义角色或者格角色相比,语义角色或者格角色是相对于词汇而言的,而框架元素仅适用于具有相同框架背景的一组词语,它摆脱了格角色个数无法确定的问题,因此用其来描述自然语言的语义更为恰当.

参照英文 FrameNet 并针对汉语自身所具有的特点,山西大学从 2006 年开始构建汉语框架网络.截至 2007 年 12 月,汉语框架网络中已构建框架 130 个,15 000 条例句,覆盖词元 1 760 个,其中,动词词元 1 428 个,名词词元 192 个,形容词词元 140 个.语料库中所有例句均来自北京大学汉语语言学研究现代汉语语料库 CCL 语料库,每条例句均具有了正确的分词及词性标注信息(遵循山西大学分词及词性标注规范),并且针对每条句子人工标注了目标词(每个例句只标注了一个目标词)及其框架语义角色,以“陈述”框架为例:

陈述(statement):此框架表达的是说话者用语言向听话者传达信息的行为.

核心框架元素(3 个):媒介 Medium[Medium]、信息 Message[Msg]、说话者 Speaker[Spkr]

非核心框架元素(7 个):听话者 Addressee[Add]、关涉 Topic[Top]、致因 Cause[Cau]、程度 Degree[Degr]、形容 Depictive[Depic]、修饰 Manner[Manr]、时间 Time[Time]

父框架:[信息交流[Communication]]

词元:动词词元:演说 v、供认 v、宣称 v、宣布 v 等,名词词元:断言 n、宣告 n、声明 n、评论 n 等

例句:1 603 条

例如:(spkr-np-subj 英 jn 方面 n) 作为 v 报复 v 措施 n 也 d (tgt 宣布 v) (msg-dj-obj 4 m 名 q 俄罗斯 nsy 大使馆 n 的 u 外交官 n 为 v 不 d 受 v 欢迎 v 的 u 人 n).w.

这里,“宣布”为“陈述”框架的词元,为标注的目标词(tgt).角色的标注信息包含 3 层内容,如(spkr-np-subj 英方面)担任“宣布”的“说话者[spkr]”角色,为名词短语[np],且是目标词<tgt 宣布>的主语[subj].

框架元素分为核心元素和非核心元素两大类,并且它们都是与所给定的框架密切相关的。框架不同,其核心元素和非核心元素也不同。本文以下所提到的“CFN 框架语义角色”指的是“CFN 框架元素”。

1.2 研究现状

国际上,2004 年以来对于语义角色标注举行的 6 次评测分别为 Senseval-3^[3],SemEval2007^[4],CoNLL (Conference on Computational Linguistics Learning)会议主办的 SRL(semantic role labeling) Shared Task 2004^[5], 2005^[6]以及 CoNLL Shared Task 2008^[7],2009^[8]。其中,CoNLL SRL Shared Task 是以 PropBank,NomBank 为语料,而 Senseval-3 的 SRL Task 和 SemEval 2007 Task-19 是以英文 FrameNet 为语料。

以英文 PropBank,NomBank 为语料的语义角色自动标注研究主要有 Chen^[9],Thompson^[10],Kadri Hacıoglu^[11],Pradhan^[12,13],Cohn^[14],Surdeanu^[15],董静等人^[16],于江德等人^[17]的工作。刘挺等人^[18]基于英文 PropBank,以句法成分为基本标注单元,使用最大熵模型研究了语义角色的自动标注问题,得到了较好的结果,在开发集和测试集上分别获得了 75.49%和 75.60%的 $F1$ -值。CoNLL-2008 Shared Task 除要求角色标注之外,还增加了谓词的识别和词义排歧,以及论元中心词与谓词的句法依赖关系(syntactic dependency)和语义依赖关系(semantic dependency)的分析,最好的评测结果 $F1$ -值达到 84.86%。国内有 5 支队伍参加了评测,Che 等人^[19]的评测结果最好,达到 82.66%的 $F1$ -值,在所有 19 支参加评测的队伍中排名第二。

CoNLL-2009 Shared Task 与 2008 年的任务基本相同,但在包含中英文的 7 种语言上评测时,最终以 7 种语言评测的平均 $F1$ -值作为排名依据,在其 Joint Task 中,Che 等人^[20]结果排名第一,7 种语言的平均 $F1$ -值达到 82.64%,同时,Che 等人的结果也是中文评测的第一名,中文的 $F1$ -值达到 76.38%。在语义角色标注子任务中,Zhao 等人^[21]的结果最好,7 种语言的平均 $F1$ -值达到 80.47%,中文评测的最好结果为 Nugues 系统^[22],其 $F1$ -值达到 78.60%。

基于英文 FrameNet 语料的语义角色标注的研究最早是 Gildea 等人^[23]的工作。他们使用概率插值方法建立模型,在句子完全句法分析树上构建特征,在 FrameNet 测试集上取得了 65%的准确率和 61%的召回率。Senseval-3 的 SRL Task 有 8 个队参加了评测,最好的结果达到 89.9%的准确率和 77.2%的召回率。SemEval 2007 Task-19 的任务除语义角色标注之外,增加了句子中框架的识别(包括目标词识别、目标词所属框架名的排歧(类似于 WSD(word sense disambiguation)))以及句子中所激活的框架的语义依赖关系(semantic dependency)的分析。只有 3 个队提交了评测结果,在 Task-19 给定的测试集上,只进行框架识别的 $F1$ -值在 49%~75%之间;在自动识别框架的基础上,再进行语义角色标注的 $F1$ -值在 36%~49%之间。

汉语的语义角色自动标注的较早研究是 Sun^[24],Xue^[25]基于中文 PropBank 的自动标注研究工作。Xue^[26]基于中文 PropBank 语料库,在使用手工标记良好的句法分析上得到了 94.1%的 $F1$ -值。但若采用自动的句法分析,则只有 71.9%的 $F1$ -值。PropBank 只对每个句子的核心动词进行了标注,语义角色的定义总共有 50 多个,所有动词的主要角色最多有 6 个,均以 Arg0~Arg5 为标记。因此,相对于框架语义角色来说,PropBank 的语义角色标注所有例句可以在一起训练和测试,有较为丰富的资源。Xue^[26]的实验结果表明,Arg0,Arg1 自动标注的精度要好于其他角色的标注,最主要的原因是 Arg0,Arg1 角色的涵义具有同质性,如 Arg0 在绝大多数情况下是指 Agent,有利于统计模型的训练;而其他角色虽然标记相同,但其涵义却差别很大。正如袁毓林^[27]所指出的:“中文 PropBank 中,论元标记 Arg2 至 Arg5 对于语义角色的严重超载,使用这种标注语料来训练角色自动标注系统,其性能势必会受到影响”。事实上,PropBank 中语义角色类型忽略了语言表达中的细节,词汇义项的描述显得不够精确。

尽管框架网络看起来像一部词典,但却不同于 PropBank,框架网络的基本单元不是个别的词(如传统词典的基本单元),而是被称作框架的概念结构,它由框架元素配置而成。框架与框架相互联系构成语义网络,通过这些框架网,语言单元相互关联,从而形式化出句子、句群、乃至篇章的语义结构,为自动语义理解提供有用的线索。因此,框架语义角色的自动标注研究有着十分重要的意义。

近年来,山西大学在手工构建的 CFN 知识库的基础上,开始着手研究汉语框架语义角色的自动标注系统,并得到国家自然科学基金、社会科学基金以及国家“863”项目的支持。文献[28]基于规则的方法研究了“自主感知”和“非自主感知”两个框架的语义角色的标注,但每个框架构建规则工作量大,适应性差。文献[29]中对若干框

架作了初步的自动标注研究.本文总结这两年来标注实验工作,通过对“陈述”框架自动标注实验的详细分析,试图探索基于机器学习的汉语框架语义角色自动标注模型,为下一步大规模语义知识库的构建提供一个实用的标注工具.

本文从 CFN 的 130 个框架中挑选出例句个数相对较多的框架(25 个),将其所有例句构成训练和测试集,用于建模分析.事实上,利用统计学习方法进行语义角色的标注,其关键环节是特征的选择.本文使用条件随机场(conditional random field,简称 CRF)模型建立汉语框架语义角色标注模型,并给出了一种使用统计正交表的特征模板优选方法,很大程度上避免了传统模板选择中“试”着选的办法的盲目性,实验结果也验证了这一方法的可行性.

本文第 2 节给出汉语框架语义角色标注的任务描述.第 3 节说明相应的语义角色标注模型.第 4 节介绍使用正交表进行最优特征模板选择的方法.第 5 节和第 6 节给出语料的来源以及评价指标.第 7 节给出后处理方法.第 8 节给出实验结果及分析.最后对全文进行总结,并给出下一步的研究方向.

2 汉语框架语义角色标注任务描述

考虑到汉语框架 CFN 的建设仍然处于初始阶段,可用的语料规模相对较少,本文将 CFN 语义角色标注的任务规定为:对于一个汉语句子,在给定目标词及其所属框架的前提下,自动识别出目标词所支配的语义角色的边界,并标注出该目标词所支配的语义角色(框架元素,包括核心和非核心元素).这个任务与 Senseval-3 中针对英文 FrameNet 的语义角色标注任务是相同的.

本文的语义角色标注均在句子已分好词和有词性标注的基础上(山西大学的分词 2000 规范).

2.1 标注单元

英文的语义角色标注研究中使用的标注单元主要有句法成分、短语、词或者依存关系等,其中以句法成分为标注单元的研究较多,标注精度也比较理想.

本文选择以词作为基本标注单元,主要是考虑到:(1) 汉语的分词和词性标注技术相对来说比较成熟,选择词为标注的基本单元可以减少标注错误向下一处理环节的累积;(2) 目前,CFN 语料库中只标注了词、词性(包含命名体标注)、目标词、语义角色边界和名称,虽然每个语义角色有短语类型和句法功能标注(相对于目标词的功能性成分),但却没有整个句子完整的句法分析或浅层的句法分析的标注信息.

2.2 语义角色的标注步骤

语言学家一般认为,谓词的每个语义角色分别对应于完全句法树中的某一个句法成分.基于完全句法分析树,Marquez 等人(2008 年)^[30]将语义角色标注的一般步骤描述为:1) filtering or pruning:在句子完全句法分析之后,过滤掉句子中明显不可能是语义角色的成分;2) local scoring:对过滤后剩余的每个句法成分进行标注;3) joint scoring:这一步考虑整个句子层面的所有角色标注序列,寻求概率最大的标注序列;但由于这些语义角色标注步骤建立在完全句法分析树的基础上,因而语义角色标注的性能也完全依赖于句法分析的性能.

本文对所有例句使用 IOB^[31]策略,记标注集合为 {B-X,I-X,O}(其中,X 为语义角色标记),示例如下:

英|B-spkr 方面|I-spkr 作为|O 报复|O 措施|O 也|O 宣布|O 4|B-msg 名|I-msg 俄罗斯|I-msg 大使馆|I-msg 的|I-msg 外交官|I-msg 为|I-msg 不|I-msg 受|I-msg 欢迎|I-msg 的|I-msg 人|I-msg .|O.

其中:“宣布”是目标词,属于“陈述”框架;B-spkr 代表“陈述”框架中角色“说话者[spkr]”的开始;I-spkr 表示“陈述”框架中“说话者[spkr]”角色的延续;O 表示块外词,即不是语义角色.

这样,就将 CFN 语义角色标注转化为词层面的线性序列标注.为此,CFN 语义角色自动标注步骤为:1) 边界识别:自动识别出句子中语义角色块边界;2) 角色分类:标注语义角色的类型;3) 后处理:针对不合理标注串等明显标注错误进行后处理,输出合理的角色标注序列.

3 标注模型

条件随机场模型^[32]目前被广泛应用于序列标注任务中.许多学者已使用条件随机场模型进行语义角色标注的尝试,Cohn^[14]将树条件随机场模型直接使用到 PropBank 的句子的完全句法分析树上,建立标注模型.实验结果表明,树条件随机场的标注性能要显著好于最大熵模型.董静等人的工作^[16]以 PropBank 为实验语料,将句法分析树“压平”,并考虑句法树中水平层次上的角色标签之间的马尔可夫依赖关系,以线性链 CRF 为标注模型进行了语义角色标注实验,提高了模型的精度.于江德等人^[17]使用条件随机场模型研究了英文 PropBank 的语义角色标注问题,他将完全句法分析树转换成浅层短语块序列,并使用浅层短语块和命名实体块作为标注单位,也取得了不错的结果.所有这些结果都表明,条件随机场模型在语义角色标注中有不俗的表现.然而,文献中使用条件随机场进行框架语义角色自动标注研究得很少.

框架知识库中不同框架所拥有的语义角色不同,标注模型可以每个词元建一个模型^[23],也可以每个框架建一个模型^[29].本文认为,不同词性词元的框架语义角色分布及句法搭配模式的可能有很大的不同,例如“陈述”框架名词词元的例句:

我们常常遇到(spkr-np-atta 某些同志) <tp-np-atta 对中央几个主要负责同志的> <manr-np-atta 不正确的> (tgt=“陈述” 言论 n),常常是不经过组织,也不合乎组织原则的.

与动词词元最明显的区别是,名词词元“言论”的框架元素绝大部分在其所处的最大短语(np)中,角色多数是“的”字结构.就现有的文献来看,基本上也是将动词(语料 PropBank)和名词(语料 NomBank)分别建模的.为此,本文建议对同一个框架,区分不同词性词元来分别建模.

在具体实验中,考虑到每个框架的例句偏少,本文首先使用如下近似卡方统计量来检验同一个框架下不同词性词元的在例句库中的角色分布是否存在显著差别.若显著不同,则要分别建模;否则,只需一个框架建一个模型.若检验为显著不同,但只是框架的一种词性的例句较多,其他词性类的词元的例句数较少(例句数与该框架语义角色个数之比小于 20),则也不再区分,只将同一个框架下所有例句放在一起训练建模.

检验所用卡方统计量如下:

设某框架有 K 个语义角色,其词元有 M 种词性,记第 m 种词性的第 k 个语义角色在例句库中的频次为 C_{km} , $m=1, \dots, M, k=1, \dots, K$, 则用如下的卡方检验来判断该框架的第 m 种词性的例句的语义角色分布是否显著不同于其他词性的语义角色分布:

$$\chi_m^2 = \sum_{k=1}^K \frac{(C_{km} - P_{km})^2}{P_{km}},$$

其中, $P_{km} = r \sum_{k=1}^K C_{km}$, $r = \left(\sum_{m=1}^M C_{km} \right) / \left(\sum_{k=1}^K \sum_{m=1}^M C_{km} \right)$.

若 $\chi_m^2 \geq \chi_{0.05}^2(K-1)$, 则需将第 m 种词性的例句单独训练一个模型,即将其看作一个不同的框架.否则,不区分不同类型词性的词元,将所有例句放在一起训练和测试.这里, r 为第 k 个语义角色在所有 M 种词性例句中的比例, P_{km} 是将第 m 种词性的角色总个数 $\left(\sum_{k=1}^K C_{km} \right)$ 按比例 r 分配到第 k 个语义角色的个数.特别地,若不同词性词元的语义角色分布相同(“理想”状态),即,对任一给定的 $k, C_{k1} = C_{k2} = \dots = C_{kM}$ (或频率相等), 则有 $P_{km} = C_{km}$.

4 特征选择及优化

事实上,模型特征是影响机器学习性能的重要因素.构建良好的特征以及特征信息的有效利用是提高机器学习性能的关键.

条件随机场模型的特征定义比较灵活,特别地,可以通过定义各特征的窗口来有效地描述标注单元与其上下文的某种依赖关系,较大的窗口还能反映标注单元与其上下文之间长距离的特征信息的依赖关系.当然,并不是将窗口定义得越大越好,大窗口会导致数据稀疏,使得模型的泛化能力显著下降.因此,每种特征(包括组合特

征)均需要调整相应的窗口到适当大小以有效反映依赖性,这是 CRF 模型中特征选择的重要环节.一般地,在使用 CRF 模型的文献中,对特征窗口的选择都没有作深入的研究,本文拟对此作初步的探讨.

条件随机场模型中的特征选择,以各特征窗口大小的组合而构成的各种特征模板的选择来体现.因此,特征选择事实上是特征模板的选择.

4.1 候选特征及窗口

根据目前语料库的状况,候选的特征有:词、词性、词相对于目标词的位置、目标词,还有这些特征的二元组合、三元组合特征、这些特征的两两组合特征.本文将这些特征的窗口大小限定在 3 以内,并人为地给出几个可选窗口,见表 1.

Table 1 Candidate features and their optional sizes of windows
表 1 候选特征及其可选的窗口大小

No.	Candidate features	Optional sizes of window
1	Word	[0,0] [-1,1] [-2,2] [-3,3]
2	Bigram of word	- [-1,1] [-2,2] [-3,3]
3	Part of speech	[0,0] [-1,1] [-2,2] [-3,3]
4	Bigram of POS	- [-1,1] [-2,2] [-3,3]
5	Position (relative to target word)	[0,0] [-1,1] [-2,2] [-3,3]
6	Bigram of position	- [-1,1] [-2,2] [-3,3]
7	Word/POS	- [0,0] [-1,1] [-2,2]
8	Word/Position	- [0,0] [-1,1] [-2,2]
9	POS/Position	- [0,0] [-1,1] [-2,2]
10	Trigram of position	- [-2,0] [-1,1] [0,2]
11	Word/target word	- [0,0]
12	Target word (compulsive)	[0,0]

在候选特征集中,确定了窗口的所有特征就构成了一个特征模板.表 1 中,需要注意以下几点:

- (1) 在每个模板中,每种特征只能选取一种窗口大小,“-”表示不选择该特征;
- (2) $[-m,n]$ (m 和 n 为正整数)表示窗口的大小,其中, $-m$ 代表当前标注单位的左边开大小为 m 的窗口, n 代表当前标注单位的右边开大小为 n 的窗口,一般选择对称大小的窗口.例如,以词特征为例,窗口大小 $[-1,1]$ 代表选取词及其前一个词、后一个词作为特征;其他类推;
- (3) 当前词、当前词的词性和当前词相对于目标词的位置为基本特征,窗口大小的选择有 4 种,窗口最小为 $[0,0]$,所以这 3 个为必选特征,只是在不同的模板中,窗口大小有所不同;
- (4) 目标词为必选特征,在每个训练模板中都要有,窗口只需取 $[0,0]$;
- (5) 对词的二元组合特征,由于窗口 $[0,0]$ 中只容纳一个标注单位,无法构成二元组合,因此,词的二元组合特征窗口没有 $[0,0]$ 的情形.其他特征情况类似;
- (6) “位置”特征是指被标注的当前词是在目标词(谓词)的左面(前面)还是右面(后面)或就是目标词,反映了当前词与目标词搭配的顺序关系,取值可以有 L(左面),T(目标词),R(右面).比如,例句“两 m 位 q 领导人 n 介绍 v 了 u ...”中,若标注的当前词为“领导人”,则窗口 $[-1,1]$ 的位置特征取值应为:L L T; 当前词的位置与词性的组合特征的窗口为 $[-1,1]$ 的取值为 L_q, L_n, T_v ,其他取值类推.根据当前词的相邻词位置的三元组合特征的特点,文中采用了非对称窗口.

另外,本文还将语义角色标记之间的一阶转移特征作为必选特征加入到每个模板之中.

4.2 基于正交表的特征模板的选优

上述特征的所有组合可以构成 $4^{10} \times 2$ 种特征模板.在所有模板上进行训练、测试,挑出最好的模板,计算量很大,显然不现实.

传统的特征模板选取方法有两种:(1) 通过主观经验为每种特征选定一个固定窗口,形成一个个固定窗口的特征模板^[12];或通过主观经验给定几个候选特征模板,并从中挑选出最优的模板^[29];(2) 以一个模型为 Baseline 模型,顺序地加入每种特征,并使用贪心算法逐步调整该特征的窗口大小,确定每种特征的最优窗口,形

成相应的特征模板^[33]。第1种方法过于主观;第2种方法实验训练时间较长,而且很难保证选到的模板较优。

事实上,每个特征的候选窗口有如下性质:1) 窗口大的特征信息包含窗口小的特征信息;2) 一个模板中每种特征只能具有一个窗口值;3) 由于各种特征信息之间相关性较大,特征窗口大小会交互影响模型的性能。因此,不能在选定特征后再调整窗口大小,将特征和窗口大小一起选择较为妥当。

本文建议,在一批适量的候选特征模板中,挑出最好的模板。这批适量的候选模板应满足:1) 数量不能太多,因为计算量的原因;2) 要有“代表性”,也即在 $4^{10} \times 2$ 种全部特征模板中具有一定“代表性”的模板。如果将表1中的候选特征看作因子,窗口大小看作水平,把各种组合构成的特征模板看作全体“实验点”空间,这个问题转化为统计实验设计中的在全部实验空间中“均匀”选取实验点的优化问题。“均匀”选出的点所对应的模板就是具有“代表性”的模板。

本文使用统计学实验设计的正交表选优法来选择模板。正交表具有均匀分散性,常用于优化选点问题,其原理和方法在实验设计文献[34]中有详细的描述。用统计学实验设计的术语,这里对应的就是11个因子(特征),前10个因子每个因子4个水平(窗口大小),最后一个因子为2水平(窗口大小),文献[34]中可以选用的正交表有 $L_{32}(4^9 \times 2^4)$ 。为了使用该正交表,本文将“相邻词的位置的三元组”特征拆成3个2水平的特征(列),3列的2水平分别对应不选取(-)和[-2,0],[-1,1],[0,2](窗口大小),因子排列分别对应表中的前9列和最后4列。使用这个正交表,只需按表里列出的各组合(每种组合对应一个模板)形成相应的32个(编号0~31)候选模板,并从中选择一个最优模板。最优模板选取的标准完全以每个框架测试集上全部语义角色标注的F1-值的大小来确定,F1-值大者为优。

5 语料来源

实验所用语料从130个框架中选择例句个数较多的25个框架的例句库(详见表2)。本文将所选语料例句拆分为4份,其中,每个框架的例句库按照词元均匀拆分到每一份中。由于目前语料规模不大,本文采用2-fold交叉验证进行实验。即,任取2份作为训练集,其他2份作为测试集,这样共可以作3组2-fold交叉验证。最终的评价指标以3组交叉验证实验的F1-值的平均值来评价标注模型的性能。

Table 2 Data from 25 frames selected from the corpus

表2 语料库所选的25个框架数据表

Names of frame	Number of frame elements	Number of exemplified sentences	Distribution of frame element counts in 4-folder data sets	Number of lexical units	Names of frame	Number of frame elements	Number of exemplified sentences	Distribution of frame element counts in 4-folder data sets	Number of lexical units
感受	6	569	169, 176, 161, 164	61	因果	7	140	63, 64, 63, 71	14
知觉特征	5	345	128, 125, 130, 123	36	陈述	10	1603	988, 1 011, 992, 963	87
思想	3	141	29, 32, 30, 32	15	拥有	4	170	87, 83, 86, 81	15
联想	5	185	76, 76, 77, 70	19	适宜性	4	70	40, 41, 42, 39	7
自主感知	14	499	278, 291, 279, 282	27	发明	12	198	109, 111, 123, 126	13
查看	9	320	157, 158, 152, 155	31	计划	6	90	51, 53, 47, 48	9
思考	8	283	111, 120, 111, 116	33	代表	7	80	50, 46, 46, 46	8
非自主感知	13	379	205, 200, 202, 201	28	范畴化	11	125	83, 84, 74, 84	10
获知	9	258	160, 156, 158, 164	15	证明	9	101	50, 48, 46, 46	11
相信	8	218	103, 90, 97, 100	26	鲜明性	9	260	110, 106, 106, 102	30
记忆	12	298	163, 169, 162, 181	17	外观	10	106	47, 47, 43, 45	13
包含	6	126	66, 64, 65, 65	24	属于某类	8	74	39, 38, 39, 37	7
宗教信仰	5	54	31, 23, 31, 27	4	Total	200	6 692	3 393, 3 412, 3 362, 3 368	560

注:框架元素定义个数是指在CFN中相应框架的定义中核心元素与非核心元素个数之和。

6 评价方法

框架语义角色自动标注正确指的是语义角色的边界和类型均完全识别正确.在测试集上对每个框架的所有语义角色,采用信息检索中的准确率(precision)、召回率(recall)和 $F1$ -值来评价标注模型的性能.它们的定义分别为

$$\begin{aligned} \text{准确率} &= \text{正确标注为语义角色的个数} / \text{自动标注为语义角色总数}, \\ \text{召回率} &= \text{正确标注为语义角色的个数} / \text{测试集中语义角色总数}, \\ \text{F1-值} &= 2 \times \text{准确率} \times \text{召回率} / (\text{准确率} + \text{召回率}). \end{aligned}$$

若重复使用交叉验证,则最终的评价指标以几组交叉验证实验的准确率、召回率的平均值而计算出的 $F1$ -值来评价标注模型的性能.

7 后处理方法

由于是以词为基本标注单位,模型的概率最大的自动标记序列可能出现不合理的序列,例如:…O I-X…,或…B-X I-Y…,或…B-X I-X I-Y…,或目标词对应标记不为“O”等.从而无法按照 IOB 的标记策略还原它们所对应的语义角色块.

在测试时,文本以概率最大的合理标注序列(以 $C(Y)=\text{TURE}$ 判断)作为最优标注序列输出,即

$$\begin{cases} Y^* = \arg \max_Y P(Y | X) \\ \text{s.t. } C(Y) = \text{TRUE} \end{cases},$$

其中:当序列 Y 为合理序列时,函数 $C(Y)=\text{TURE}$;否则, $C(Y)=\text{FALSE}$.在具体实现时,将条件随机场模型输出的标注序列,依序列的概率值从大到小顺序排列,逐个判断 $C(Y)$,直至找到第 1 条合理的标记序列输出即可.判断一个标注序列是否为合理的标注序列的算法如下:

算法 1. 判断标注序列合理性的算法.

输入:(1) 当前句子的标注序列: $S=(s[1],s[2],\dots,s[n])$;

其中, $s[i]$ 属于集合 $\{B-X,I-X,O\}$, X 为语义角色类型, $1 \leq i \leq n$;

(2) 目标词在序列中的位置: $pos_tgt, 1 \leq pos_tgt \leq n$;

1: 初始化:

当前语义角色标记 **chunkTag** 赋值为空;

序列状态 **state** 赋值为“合理”;

2: 检查目标词的预测标记 $s[pos_tgt]$:

如果 $s[pos_tgt]$ 为 O,则跳至 3;否则,则将 **state** 设置为“不合理”,跳转到 4;

3: 循环开始 $s[i]: 1 \leq i \leq n$ //依次判定标注序列 S 中每个词语的标记

(1) 如果 $s[i]$ 为 B-X,则将当前语义角色标记 **chunkTag** 设置为 X ;

(2) 如果 $s[i]$ 为 I-X,判断 X 是否与 **chunkTag** 匹配.如果不匹配,则将 **state** 设置为“不合理”,跳转到 4;

(3) 如果 $s[i]$ 为 O,则将当前语义角色标记 **chunkTag** 设置为空;

循环结束

4: 结束

输出:序列状态 **state**.

标注的序列可能很多,若在有限步(比如 200 步)内还无法找到合理序列,则将最大概率序列直接修改为合理序列,具体的修改算法如下:

算法 2. 将不合理序列直接修改为合理序列的算法.

输入:(1) 当前句子的标注序列: $S=(s[1],s[2],\dots,s[n])$;

其中, $s[i]$ 属于集合 $\{B-X,I-X,O\}$, X 为语义角色类型, $1 \leq i \leq n$;

(2) 目标词在序列中的位置: $pos_tgt, 1 \leq pos_tgt \leq n$;

1: 初始化:

当前语义角色标记 `chunkTag` 为空;

2: 检查目标词的预测标记 $s[pos_tgt]$:

如果 $s[pos_tgt]$ 不为 O, 则将标记 $s[pos_tag]$ 设置为 O;

3: 循环开始 $s[i]: 1 \leq i \leq n$ //依次判定标注序列 S 中每个词语的标记

(1) 如果 $s[i]$ 为 B-X, 则将当前语义角色标记 `chunkTag` 设置为 X;

(2) 如果 $s[i]$ 为 I-X, 则判断 X 是否与当前语义角色 `chunkTag` 匹配. 如果不匹配, 则将标记 $s[i]$ 设置为 (“I-”+`chunkTag`);

(3) 如果 tag 为 O, 则将当前语义角色标记 `chunkTag` 设置为空;

循环结束

输出: $S'=(s[1], s[2], \dots, s[n])$; //修改后的标注序列.

8 实验结果及分析

实验所用 CRF++ 工具包来自于 Sourceforge^[35], 使用其中的 CRFL2 算法, 并选取 $C=1$ 进行参数平滑.

按照第 3 节所述方法对 25 个框架分别进行检验, 得到只有“陈述”框架需要分开训练.“陈述”框架的词元具有动词和名词两种词性, 有 $K=10$ 个角色. 根据第 3 节的 χ_m^2 检验方法, 自由度应为 9, 经过计算得到动词的卡方值为 $\chi^2 = 144.2 > \chi_{0.05}^2(9) = 16.919$ (“陈述”框架), 名词的卡方值为 $\chi^2 = 967.24 > \chi_{0.05}^2(9) = 16.919$, 并且两类的例句个数都较多. 因此, 将“陈述”框架的语义角色标注模型分为动词模型和名词模型两种.

下面先给出“陈述”框架动词模型的详细实验结果和实验分析, 最后给出 25 个框架的自动标注结果和整体标注性能.“陈述”框架动词词元有例句 1 350 条, 角色个数 3 441. 四份例句数据集的分布为: 第 1 份(D1): 例句数 438, 角色个数 846 个; 第 2 份(D2): 例句数 438, 角色个数 890 个; 第 3 份(D3): 例句数 438, 角色个数 875 个; 第 4 份(D4): 例句数 437, 角色个数 858 个.

本文的所有实验均是标注步骤 1) 的边界识别与步骤 2) 的角色分类合并进行的.

8.1 “陈述”框架动词模型的语义角色的标注结果

依 $F1$ -值的大小, 从正交表的 32 个特征模板中挑选出了“陈述”框架动词模型的最好模板(#12). 表 3 给出了“陈述”框架动词模型#12 模板的 3 组交叉验证的实验结果, 最好模板的 $F1$ -值为 67.75%.

Table 3 Results of 2-fold cross-validation on SRL (semantic role labeling) for the verbal model of “statement” frame

表 3 “陈述”框架动词模型语义角色标注的 2-fold 交叉验证结果

	2-fold CV Group 1		2-fold CV Group 2		2-fold CV Group 3	
Component parts of training set and testing set	D1, D2↔D3, D4 training and testing each other		D1, D3↔D2, D4 training and testing each other		D1, D4↔D2, D3 training and testing each other	
Best template (#12)	68.21%	65.68%	70.14%	66.67%	69.04%	66.75%
Average of each group	66.95%		68.41%		67.90%	
Average in all	67.75%					

一般文献中都是基于经验给出若干个模板, 并从中选出最好的模板. 如使用文献[29]选出的最好模板, 用同样数据得到的“陈述”框架动词模型的 $F1$ -值为 64.13%, 在 $\chi^2(1)$ 检验下, 两者有显著差异 ($p < 0.05$). 表 4 中列出了最优模板(#12)的含义.

表 4 中, “目标词”特征为给定前提条件(必选), 词、词性、位置 3 个特征也为必选特征(加粗体).

Table 4 Best template of SRL for the verbal model of “statement” frame**表 4** “陈述”框架动词模型语义角色标注的最优模板

No.	Features	Selected levels	Description
1	Word	[0,0]	Feature of current word
2	Bigram of word	—	Not selecting this feature
3	POS	[-3,3]	Feature of POS with [-3,3] window
4	Bigram of POS	—	Not selecting this feature
5	Position	[-1,1]	Feature of position with [-1,1] window
6	Bigram of position	[-3,3]	Bigram feature of position with [-3,3] window
7	Word/POS	[0,0]	Joint feature of current word and its POS
8	Word/Position	[-2,2]	Joint feature of current word and its position with [-2,2] window
9	POS/Position	[-1,1]	Joint feature of POS and position of current word with [-1,1] window
10	Trigram of position	—	Not selecting this feature
11	Word/target word	—	Not selecting this feature
12	Target word	[0,0]	Feature of target word be included in all templates

8.1.1 “陈述”框架动词模型语义角色标注结果的错误分析

“陈述”框架动词模型标注性能较低的原因主要有:

(1) 语义角色的边界识别正确率低

表 5 中详细列出了各语义块的错误分析结果.由表 5 可知,在标注错误的语义角色块中,大部分都是因边界识别错误造成的,边界正确但角色类别标注错误的语义角色块很少.

Table 5 Statistics on error labeling of semantic roles for the verbal model of “statement” frame**表 5** “陈述”框架动词模型下语义角色块的标注错误情况统计

2-fold cross validation		Total number of chunks in each test set	Chunks with wrong boundaries	Chunks with correct boundaries but wrong frame elements types	Both correct chunks
Group 1	D1, D2↔D3, D4	1 705	601	43	1 061
		1 736	622	66	1 048
Group 2	D1, D3↔D2, D4	1 720	565	50	1 105
		1 721	618	51	1 052
Group 3	D1, D4↔D2, D3	1 737	599	53	1 085
		1 704	600	44	1 060

如果仅对“陈述”框架进行边界识别,可得动词模型的 3 组交叉验证的平均边界识别准确率、召回率、F1-值分别为 74.97%,67.24%,70.90%;而在给定正确边界的基础上,“陈述”框架动词模型的语义角色分类精确率 (accuracy)为 89.64%.

(2) 词特征稀疏的影响

词特征是标注模型的关键特征.要使语义角色块标注正确,首先使得组成其语义块的词的标注正确.但是,从词的标注(而不是从语义角色块)情况来看,那些在测试集中出现而在训练集中未出现的词的标注正确率较其他词明显要低.表 6 给出了“陈述”框架动词模型中词的标注情况的统计信息.

Table 6 Labeling results w.r.t. words within the semantic roles for the verbal model of “statement” frame**表 6** “陈述”框架动词模型语义角色中按照词的标注结果

Experiment	Appeared in the training set					Not appeared in the training set				
	Word counts	Correct	Accuracy (%)	Wrong	Error rate (%)	Word counts	Correct	Accuracy (%)	Wrong	Error rate (%)
D1,D2	7 497	5 766	76.91	1 731	23.09	1 978	1 427	72.14	551	27.86
D1,D3	7 661	5 782	75.47	1 879	24.53	2 007	1 425	71.00	582	29.00
D1,D4	7 660	5 645	73.69	2 015	26.31	1 923	1 339	69.63	584	30.37
D2,D3	7 319	5 715	78.08	1 604	21.92	1 998	1 463	73.22	535	26.78
D2,D4	7 353	5 620	76.43	1 733	23.57	1 879	1 394	74.19	485	25.81
D3,D4	7 487	5 784	77.25	1 703	22.75	1 938	1 416	73.07	522	26.93
Average (%)			76.31		23.70			72.21		27.79

从表 6 中可以看出,那些在测试集中出现而在训练集中未出现的词的标注准确率会下降约 4%.因此,测试集

中出现而未在训练集中出现的词是性能下降的一个因素.如果加大训练集规模,相对缩小测试集规模,那么在测试集中未出现而在训练集中出现的词的数目将会变少.比如,当“陈述”框架动词模型的例句库训练、测试按 4-fold 交叉验证, $F1$ -值为 71.03%,模型的性能明显提升.

(3) 缺乏较好的句法分析器的支持

Xue^[26]的研究结果表明,基于人工标注的完全句法树,中文 PropBank 的语义角色的标注 $F1$ -值能够达到 94.1%,可与英文 PropBank 的性能相媲美.但是,如果采用自动的汉语句法分析器,则最终 F 值却只能达到 71.9%.

为了得到基于完全句法分析树的 CFN 语义角色标注结果,本文使用 Stanford 大学的汉语句法分析器 v1.6^[36](从文献[37]来看,它是性能较好的汉语完全句法分析器之一)对训练、测试集上所有例句进行了自动完全句法分析(在“正确”分词的基础上).本文尝试使用与 Xue^[26]相同的特征集和最大熵分类器进行 CFN 语义角色边界识别实验.在“陈述”框架的动词词元的测试集上,语义角色边界识别的 $F1$ -值只有 55.96%.

8.1.2 “陈述”框架动词模型语义角色标注的性能分析

随机抽取“陈述”框架动词词元训练集中例句数,使其例句个数 n 与框架定义的元素个数 e 的比值 $r=n/e$,以步长 5 从 5 取到 65,分别训练和测试,可以得到模型随着标注例句的规模变化的性能曲线,如图 1 所示.图 1 中给出了 3 组交叉验证的 $F1$ -值的变化曲线(共 6 条)以及 6 组实验的 $F1$ -值的平均值(macro- F).

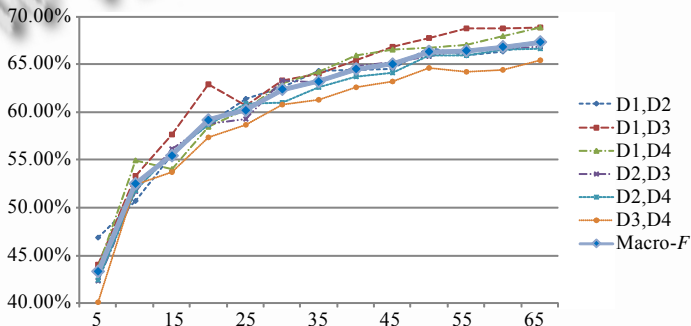


Fig.1 Learning curve of verbal model in “statement” frame

图 1 “陈述”框架下动词模型的性能曲线

由图 1 中可知,随着例句个数 n 与框架定义的元素个数 e 比值 $r=n/e$ 的增大,其标注的性能呈上升趋势.换句话说,如果继续增加语料库规模,则模型的性能还有增加的余地.

8.2 其他框架的自动标注结果

其他框架按照以上方法,分别训练并测试,可以得到如表 7 所示的实验结果.表 7 的最后一行给出了 25 个框架的总的实验结果,总 $F1$ -值为 61.62%.

根据表 7 给出的实验结果以及前述结果分析,本文可以得到如下几点结论:

- (1) 最好模板 1 号最多,有 9 个;12 号模板次之,有 6 个.在其他框架的实验结果中,1 号模板均排在前 5 位.若全部框架都用 1 号模板,则可以得到 $F1$ -值 61.04%.这也是所有框架按同一个模板的最好的结果(全部使用 12 号模板得到 60.93%,为次好模板),与各自最优模板的 61.62%差异不大.因此,1 号模板可以作为所有框架的通用模板(见表 8).这样,在新框架的自动标注建模中,为节省时间,可以省掉模板选择的过程,直接用 1 号模板;
- (2) 本文用选出的通用模板 1 号,得到 25 个框架的 $F1$ -值为 61.04%;对 25 个框架全部采用文献[29]中的最优模板,只能得到 56.91%.在 $\chi^2(1)$ 检验下,两者有显著差异($p < 0.05$),说明正交表特征选择方法较传统的经验试探的模板选择方法可能更有效,并在一定程度上避免了模板选择的盲目性;
- (3) 本文曾尝试使用 Stanford 大学的汉语句法分析器 v1.6 对实验数据集上所有例句进行了自动分析,并使用 Xue^[26]的完全相同特征在 25 个框架的测试集上进行实验.角色的边界完全能够与句法分析树中

的句法成分对齐的有 76%,使用最大熵方法仅识别角色边界的结果也只有 52.43%.因此,在缺乏较好句法分析器的情况下,将语义角色标注问题看作词序列的序列标注问题,方法具有一定的可行性;

- (4) 所有框架的自动标注结果都是准确率显著大于相应的召回率,召回率是影响模型性能的主要指标.从“陈述”框架动词模型的错误分析来看,整个模型召回率低的主要原因是每个框架的角色边界识别时的召回率很低.因此,提高角色边界的识别的召回率应当是下一步主攻方向;
- (5) 同一个框架定义下虽然有相同的框架元素定义,但可以有不同的词性词元.我们分开来建模主要是考虑到不同的词性词元在句子中与其元素之间的组配方式明显不同,反映在条件随机场模型的所选特征上会有所不同,模板也有所不同.“陈述 n”和“陈述 v”标注的分开训练、测试结果明显不同,说明区分还是有必要的;
- (6) 在本文所选的 25 个框架中,标注结果的 $F1$ -值偏低的基本上都来自框架网的认知域,而认知域的框架即使在手工标注时也都很难把握,因此其自动标注有一定的难度,具有挑战性.

Table 7 Results on cross-validation for the selected 25 frames

表 7 选出的 25 个框架交叉验证的结果

Names of frame	No. of best template	P (%)	R (%)	F1 (%)	Names of frames	No. of best template	P (%)	R (%)	F1 (%)
感受	1	62.94	52.01	56.96	因果	12	72.94	40.51	52.09
知觉特征	16	81.37	62.71	70.83	陈述_v	12	74.53	62.11	67.75
思想	16	67.12	44.72	53.68	陈述_n	21	70.28	43.89	54.03
联想	1	50.23	24.30	32.75	拥有	1	77.57	59.35	67.25
自主感知	6	73.50	49.56	59.20	适宜性	2	69.14	49.42	57.64
查看	1	71.45	36.37	48.20	发明	12	69.91	42.20	52.63
思考	12	74.95	47.45	58.11	计划	1	72.53	52.01	60.57
非自主感知	9	72.60	48.27	57.99	代表	11	76.60	58.73	66.48
获知	1	74.71	52.68	61.79	范畴化	1	67.71	30.56	42.11
相信	12	83.39	51.06	63.34	证明	20	75.30	55.82	64.11
记忆	17	79.21	52.10	62.86	鲜明性	23	81.98	54.20	65.25
包含	1	87.78	84.87	86.30	外观	1	74.06	49.13	59.08
宗教信仰	12	85.10	47.28	60.79	属于某类	9	83.92	67.31	74.70
					Total		74.16	52.70	61.62

Table 8 Best template of SRL for all 25 frames

表 8 所有 25 个框架语义角色标注的最优模板

No.	Features	Selected levels	Description
1	Word	[-1,1]	Feature of current word with [-1,1] window
2	Bigram of word	—	Not selecting this feature
3	Part of speech	[0,0]	Feature of POS for current word
4	Bigram of POS	[-2,2]	Bigram feature of POS with [-2,2] window
5	Position	[0,0]	Position feature for current word
6	Bigram of position	[-2,2]	Bigram feature of position with [-2,2] window
7	Word/POS	[-1,1]	Joint feature of current word and its POS with [-1,1] window
8	Word/Position	[0,0]	Joint feature of current word and its position
9	POS/Position	[-2,2]	Joint feature of POS and position of current word with [-2,2] window
10	Trigram of position	—	Not selecting this feature
11	Word/target word	—	Not selecting this feature
12	Target word (compulsive)	[0,0]	Feature of target word be included in all templates

9 结论与展望

本文基于山西大学开发的汉语框架语义知识库(CFN),将语义角色标注问题看作以词为基本标注单元的序列标注问题,使用条件随机场模型,以每个框架相同词性的词元建一个模型,引入统计学正交表特征窗口的优选方法,研究了汉语框架语义角色的自动标注任务.在由 25 个框架例句所构成的语料上,以每个框架的最优模板得到的语义角色标注的 2-fold 交叉验证的 $F1$ -值达到 61.62%;若以选出的所有框架的统一的 最优模型模板来

看, $F1$ -值达到 61.04%。

本文基于 Stanford 句法分析器,使用 Xue 的方法得到的结果偏低,可能与我们的语料偏小、分类模型训练不够充分有很大关系。但我们一直在思考,语义角色是目标词的支配项,是词与目标词的语义依存关系。将词语级的语义依存分析基于完全句法分析之上,似乎有点将此依赖于一个更为复杂问题的解决。事实上,许多语言学家认为,汉语是一种意合语言,语义与句法形式结构相对应的特征并不明显,词语的语义结构多体现在词与某些词语经常性的组合。因此,本文将目标词的语义角色标注问题看作词序列的序列标注问题,在句子中抽取词与词之间的依赖关系特征,不失为一种有益的尝试。从本文的标注结果来看,这种方法具有一定的可行性。

下一步需要进行的研究工作还有许多,主要集中在如下 5 个方面:

- (1) 深入研究如何在模型中加入句法层面的特征,努力提高自动标注模型的性能;
- (2) 将语义角色边界识别和语义角色分类分别进行,并与本文中的模型进行对比分析;
- (3) 多种语义角色标注模型标注结果的比较和融合;
- (4) 增加语料规模,建立分词性模型的通用模板;
- (5) 设计完成一个框架语义自动标注器,为构建大规模语义知识库提供一个自动分析工具。

致谢 本文的实验分析和撰写过程,均得到刘开瑛老师的指导,在此表示感谢。并由衷地感谢审稿人为本文提出了非常具有建设性的意见。

References:

- [1] You L, Liu K. Building Chinese FrameNet database. In: Ren FJ, Zhong YX, eds. Proc. of the IEEE NLP-KE 2005. Wuhan: IEEE Press, 2005. 301–306.
- [2] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. In: Morgan K, ed. Proc. of the COLING-ACL'98. Montreal: ACL Press, 1998. 86–90.
- [3] Litkowski KC. Senseval-3 task automatic labeling of semantic roles. In: Mihalcea R, Edmonds P, eds. Proc. of the 3rd Int'l Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona: ACL Press, 2004. 9–12.
- [4] Baker CF, Ellsworth M, Erk K. SemEval 2007 task 19: Frame semantic structure extraction. In: Agirre E, Màrquez L, Wicentowski R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations. Prague: ACL Press, 2007. 99–104.
- [5] Carreras X, Màrquez L. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In: Ng HT, Riloff E, eds. Proc. of the CoNLL 2004. Boston: ACL Press, 2004. 89–97.
- [6] Carreras X, Màrquez L. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL Press, 2005. 152–164.
- [7] Surdeanu M, Johansson R, Meyers A, Màrquez L, Nivre J. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In: Clark A, Toutanova K, eds. Proc. of the CoNLL-2008. Manchester: ACL Press, 2008. 159–177.
- [8] Hajic J, Ciaramita M, Johansson R, Kawahara D, Marti MA, Màrquez L, Meyers A, Nivre J, Padó S, Stěpónek J, Stranak P, Surdeanu M, Xue NW, Zhang Y. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Stevenson S, Carreras X, eds. Proc. of the CoNLL-2009. Boulder: ACL Press, 2009. 1–18.
- [9] Chen J, Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments. In: Hinrichs EW, Roth D, eds. Proc. of the EMNLP-2003. Sapporo: ACL Press, 2003. 41–48.
- [10] Thompson CA, Levy R, Manning CD. A generative model for semantic role labeling. In: Lavrac N, Gamberger D, Todorovski L, Blockeel H, eds. Proc. of the ECML-2003. Croatia: Springer-Verlag, 2003. 235–238.
- [11] Hacıoglu K. Semantic role labeling using dependency trees. In: Nirenburg S, ed. Proc. of the COLING-2004. Geneva: COLING, 2004. 1273–1276.
- [12] Pradhan S, Hacıoglu K, Krugler V, Ward W, Martin J, Jurafsky D. Support vector learning for semantic argument classification. Machine Learning, 2005,60(1):11–39. [doi: 10.1007/s10994-005-0912-2]
- [13] Pradhan S, Ward W, Martin JH. Towards robust semantic role. Computational Linguistics, 2008,34(2):289–310. [doi: 10.1162/coli.2008.34.2.289]

- [14] Cohn T, Blunsom P. Semantic role labeling with tree conditional random fields. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL Press, 2005. 169–172.
- [15] Surdeanu M, Márquez L, Carreras X, Comas PR. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research*, 2007,29(1):105–151.
- [16] Dong J, Sun L, Lv Y, Feng Y. Semantic role labeling based on linear chain conditional random fields. In: Sun MS, Chen QX, eds. Proc. of the Progress and Forefront of Chinese Information Processing. Beijing: Tsinghua University Press, 2006. 32–37 (in Chinese with English abstract).
- [17] Yu JD, Fan X, Pang W, Yu Z. Semantic role labeling based on conditional random fields. *Journal of Southeast University (English Edition)*, 2007,23(3):361–364.
- [18] Liu T, Che WX, Li S. Semantic role labeling with maximum entropy classifier. *Journal of Software*, 2007,18(3):565–573 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/18/565.htm> [doi: 10.1360/jos180565]
- [19] Che WX, Li ZH, Hu YX, Li YQ, Qin B, Liu T, Li S. A cascaded syntactic and semantic dependency parsing system. In: Clark A, Toutanova K, eds. Proc. of the CoNLL-2008. Manchester: ACL Press, 2008. 238–242.
- [20] Che WX, Li ZH, Li YQ, Guo YH, Qin B, Liu T. Multilingual dependency-based syntactic and semantic parsing. In: Stevenson S, Carreras X, eds. Proc. of the CoNLL-2009. Boulder: ACL Press, 2009. 49–54.
- [21] Zhao H, Chen WL, Kit C, Zhou GD. Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In: Stevenson S, Carreras X, eds. Proc. of the CoNLL-2009. Boulder: ACL Press, 2009. 55–60.
- [22] Bjorkelund A, Hafdel L, Nugues P. Multilingual semantic role labeling. In: Stevenson S, Carreras X, eds. Proc. of the CoNLL-2009. Boulder: ACL Press, 2009. 43–48.
- [23] Gildea D, Jurafsky D. Automatic labeling of semantic roles. *Computational Linguistics*, 2002,28(3):245–288. [doi: 10.1162/089120102760275983]
- [24] Sun HL, Jurafsky D. Shallow semantic parsing of Chinese. In: Hirschberg JB, ed. Proc. of the NAACL-HLT 2004. Boston: ACL Press, 2004. 249–256.
- [25] Xue NW, Palmer M. Automatic semantic role labeling for Chinese verbs. In: Bramer M, ed. Proc. of the 19th Int'l Joint Conf. on Artificial Intelligence. Edinburgh: IJCAI, 2005. 1161–1165.
- [26] Xue NW. Labeling Chinese predicates with semantic roles. *Computational Linguistics*, 2008,34(2):225–255. [doi: 10.1162/coli.2008.34.2.225]
- [27] Yuan YL. The new trend and long term goal of constructing semantic resource. *Journal of Chinese Information Processing*, 2008, 22(3):3–15 (in Chinese with English abstract).
- [28] Liu M, You L. The preparatory study of semantic role labeling rule for Chinese perceptive word. In: Sun MS, Chen QX, eds. Proc. of the Frontiers of Content Computing: Research and Application. Beijing: Tsinghua University Press, 2007. 320–325 (in Chinese with English abstract).
- [29] Liu K, Chen XY, Li JH. Chinese frame elements automatic tagging. In: Bai S, Sun MS, eds. Proc. of the 4th National Conf. of Information Retrieval and Content Security. Beijing: NCIRCS Press, 2008. 48–55 (in Chinese with English abstract).
- [30] Márquez L, Carreras X, Litkowski KC, Stevenson S. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 2008,34(2):145–159. [doi: 10.1162/coli.2008.34.2.145]
- [31] Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Yarowsky D, Church K, eds. Proc. of the 3rd Workshop on Very Large Corpora. Cambridge: ACL Press, 1995. 88–94.
- [32] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th Int'l Conf. on Machine Learning. Williamstown: Morgan Kaufmann Publishers, 2001. 282–289.
- [33] Ding W, Chang B. Maximum entropy-based semantic role classification. *Journal of Chinese Information Processing*, 2008,22(6): 20–27 (in Chinese with English abstract).
- [34] Robust Design Team of CAAS. Robust Design of Computable Project. Beijing: Peking University Press, 1985. 165–165 (in Chinese).
- [35] Kudo T. CRF++ tools package. version: 5.0, 2007. <http://crfpp.sourceforge.net/>

- [36] MacCartney B. The Stanford parser. version: 1.6, 2007. <http://nlp.stanford.edu/software/lex-parser.shtml>
- [37] Levy R, Manning C. Is it harder to parse Chinese, or the Chinese treebank? In: Tsujii JI, ed. Proc. of the 41st Annual Meeting on Association for Computational Linguistics, Vol.1. Sapporo: ACL Press, 2003. 439-446.

附中文参考文献:

- [16] 董静,孙乐,吕元华,冯元勇.基于线性链条件随机场模型的语义角色标注.见:中国中文信息学会 25 周年学术会议论文集.北京:清华大学出版社,2006.
- [18] 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,2007,18(3):565-573. <http://www.jos.org.cn/1000-9825/18/565.htm> [doi: 10.1360/jos180565]
- [27] 袁毓林.语义资源建设的最新趋势和长远目标.中文信息学报,2008,22(3):3-15.
- [28] 刘鸣洋,由丽萍.汉语感知词语的语义角色标注规则初探.见:内容计算的研究与应用前沿会议论文集.北京:清华大学出版社,2007.320-325.
- [29] 刘开瑛,陈雪艳,李济洪.汉语框架元素自动标注实验报告.见:第 4 届全国信息检索与内容安全学术会议论文集.北京:清华大学出版社,2008.48-55.
- [33] 丁伟伟,常宝宝.基于最大熵原则的汉语语义角色分类.中文信息学报,2008,22(6):20-27.
- [34] 中国现场统计研究会三次设计组.可计算性项目的三次设计.北京:北京大学出版社,1985.165-165.



李济洪(1964—),男,山西长子人,副教授,主要研究领域为中文信息处理,统计学.



王蔚林(1984—),男,硕士,主要研究领域为统计学.



王瑞波(1985—),男,硕士,主要研究领域为中文信息处理.



李国臣(1963—),男,教授,主要研究领域为中文信息处理.