

文章编号: 1003-0077(2024)02-0015-10

基于短语及依存的标注规则和短语识别算法研究

刘 广, 涂 刚, 李 政, 刘泽键

(华中科技大学 计算机科学与技术学院, 湖北 武汉 430074)

摘 要: 目前, 自然语言处理大多是借助于分词结果进行句法依存分析, 主要采用基于监督学习的端对端模型。该方法主要存在两个问题, 一是标注体系繁多, 相对比较复杂; 二是无法识别语言嵌套结构。为了解决以上问题, 该文提出了基于短语窗口的依存句法标注规则, 并标注了中文短语窗口数据集(CPWD), 同时引入短语窗口模型。该标注规则以短语为最小单位, 把句子划分为 7 类可嵌套的短语类型, 同时标示出短语间的句法依存关系; 短语窗口模型借鉴了计算机视觉领域目标检测的思想, 检测短语的起始位置和结束位置, 实现了对嵌套短语及句法依存关系的同步识别。实验结果表明, 在 CPWD 数据集上, 短语窗口模型比传统端对端模型 F_1 值提升超过 1 个百分点。相应的方法应用到了 CCL2018 的中文隐喻情感分析比赛中, 在原有基础上 F_1 值提升了 1 个百分点以上, 取得第一名成绩。

关键词: 自然语言处理; 标注体系; 短语识别; 依存分析

中图分类号: TP391

文献标识码: A

Research on Annotation Rules and Phrase Recognition Algorithm Based on Phrase and Dependency

LIU Guang, TU Gang, LI Zheng, LIU Yijian

(School of Computer Science and Technology, Huazhong University of
Science and Technology, Wuhan, Hubei 430074, China)

Abstract: At present, most syntactic dependency analysis is conducted via supervised learning with the help of word segmentation results. This practice is challenged by complex label schemes and the nesting structure which is difficult to parse. This paper proposes a phrase window model together with a dependency syntax labeling rule based on the phrase window. The labeling rule divides sentences into 7 types of nestable phrases, with annotation for the syntactic dependence between phrases. Inspired by the idea of target detection in the computer vision field, the phrase window model detects the beginning and end positions of phrases and realizes the synchronous recognition of nested phrases and syntactic dependencies. Experimental results show that on the self-built Chinese Phrase Window Dataset (CPWD), the phrase window model is more than 1 point better than the traditional end-to-end model. The corresponding method won the champion in the CCL2018 Chinese Metaphor Sentiment Analysis Competition, which improved more than 1 point than the baseline.

Keywords: natural language processing; tagging system; phrase extraction; dependency parsing

0 引言

随着即时通信、微博、论坛、朋友圈等的快速流行, 人们在网络上发布的文字信息也越来越多。这些文字信息蕴含极大的价值, 它们是分析社会整体及公众个体的观点、喜好、情绪、趋势等的入口。快

速准确地分析文字信息, 是自然语言处理研究的目标。

目前, 自然语言处理研究领域存在两个问题: 一是标注规则复杂。比如哈工大等词性类别多达二十多种, 依存类别多达十几种, 如表 1 所示; 这种标注规则, 语言学家可以轻松掌握, 但是一般的标注人员, 非常难以掌握。即便掌握了, 在标注数据的过程

中,由于语言的嵌套情况以及熟练程度等原因,会出现各种各样问题,这就造成了标注数据慢而且错误较多,使得监督学习的模型预测准确性难以提高。二是算法不能识别语言成分的嵌套情况。比如,“敌人的进攻”是个名词,但是“进攻”却是动词,传统深度端到端模型只会预测出一个类别,无法预测这种短语嵌套情况。

表 1 依存标注类别举例

施事关系, 当事关系	比较角色, 属事角色	并列关系, 选择关系	连词标记, 的字标记
...	...	...	...
成事关系, 源事关系	结局角色, 方式角色	转折关系, 原因关系	程度标记, 根节点
...	...	...	...

为了解决这两个问题,本文提出了基于短语窗口的标注规则,同时设计了相应的短语识别算法。该标注规则以短语为最小单位,把句子分成动词短语、名词短语等 7 类可嵌套的短语类型,同时标示出短语之间的句法依存关系。我们使用该规则,标注了各种类型的句子数据,把这个数据集称为中文短语窗口数据集(Chinese Phrase Window Dataset, CPWD)。对应的短语识别算法,借鉴了图像领域识别目标区域的 Faster RCNN 算法思想,可发现句子中各种短语的起始位置与结束位置,实现对嵌套短语及句法依存关系的同步识别,对应模型称为语法窗口模型(Syntax Window Model, SWM)。实验的结果表明,SWM 模型比端到端模型更加适用于短语嵌套特征,短语识别准确性有明显提高。

1 相关工作

语块分析体系最早是由 Abney 在 1991 年提出的语块描述体系^[1],之后 Kudoh 等^[2]提出了一种基于支持向量机的语块自动分析方法;同时,Shen 等^[3]提出了一种投票分类策略,将多种不同的数据表示和多种训练模型结合在一起,根据投票分类策略确定最终结果;此外,Mancev 等^[4]提出了一种处理支持向量机非凸结构的斜率损失的最小化问题的序列双向方法。在汉语的语块分析方面,周强等^[5-6]构造了基于规则的汉语基本块分析器,并设计了相应的基本块规则,给出了一整套解决方案,提高了基于规则的基本块分析器的性能。此外,李超等^[7]应用最大熵模型和马尔科夫模型构建了一套汉语基本

块的分布识别系统。

深度学习方法出现后,短语识别研究迎来了快速发展。Chiu 等^[8]使用双向 LSTM 提取文本全局特征,同时,使用 CNN 提取单词的特征,进行名词短语实体的识别;Kuru 等^[9]使用 Stacked Bidirectional LSTMs 提取文本全局特征进行名词短语实体识别,取得了较大进展;侯潇琪等^[10]利用深度模型,将词的分布表征作为模型的输入特征维度,用于基本短语识别任务中,性能比使用传统的词特征表示方法提高明显;李国臣等^[11]以字作为标注单元和输入特征,基于深层模型研究短语的识别问题,并将基于 C&W 和 Word2Vec 两种方法训练得到的字分布表征作为模型的特征参数,避免了对分词及词性标注结果的依赖;徐菁^[12]利用知识图谱,提出基于主题模型和语义分析的无监督的名词短语实体指称识别方法,同时具备短语边界检测和短语分类功能;程钟慧等^[13]提出了一种基于强化学习的协同训练框架,在少量标注数据的情况下,无须人工参与,利用大量无标注数据自动提升模型性能,从非结构化大数据集中抽取有意义的名词短语。

句法依存最早由著名的法国语言学家特思尼耶尔提出,我国学者徐烈炯^[14]认为,语义角色是一个“句法—语义”接口概念,而不是单纯的语义概念;刘宇红^[15]提出语义和语法双向互动的观点;孙道功^[16]基于词汇义征和范畴义征的分析,研究了词汇与句法的衔接机制;亢世勇等^[17]通过构建“现代汉语句法语义信息语料库”,研究了义类不同的体词在施事(主语、宾语、状语)和受事(主语、宾语、状语)六个语块的分布特点。这其中还包括哈工大、腾讯、百度、清华等团队的语法分类贡献。

在语法分析方面,McDonald 等^[18]提出了基于图模型的依存句法分析器 MSTParser;Nivre 等^[19]提出了基于转移模型的依存句法分析器 MaltParser;Ren 等^[20]对 MaltParser 依存句法分析器的 Nivre 算法进行了优化,有效地改进了在汉语中难以解决的长距离依存等问题;Che 等^[21]对 MSTParser 依存句法分析器进行了改进,使用了图模型中的高阶特征,提高了依存句法分析的精度;Dyer 等^[22]在基于转移模型的依存句法分析框架上运用长短时记忆神经网络,将传统的栈、队列、转移动作序列看作 3 个 LSTM 细胞单元,将所有转移的历史均记录在 LSTM 中,改进了长距离依存问题;Ji 等^[23]开发了一种依赖树节点表示形式,可以捕获高阶信息,通过使用图神经网络(GNN),解析器可以在 PTB 上实现最佳的 UAS 和

LAS; Wang 等^[24]提出了一种基于神经过渡的解析器,通过使用基于列表的弧跃迁过渡算法的一种变体,进行依赖图解析,获得了较好的效果;Fried 等^[25]通过强化学习来训练基于过渡的解析器,提出了将策略梯度训练应用于几个解析器的实验,包括基于 RNN 的解析器。

在语义分析方面,丁伟伟等^[26]利用 CRF 在英文语料上利用论元之间的相互关系以及标注准确率高的特点,将其运用到中文命题库,使用 CRF 对中文语义组块分类,取得好的效果;王丽杰^[27]提出了基于图的自动汉语语义分析方法,使用哈工大构建的汉语语义依存树库完成了依存弧和语义关系的分析;王倩等^[28]基于谓词和句义类型块,使用支持向量机的语义角色对句子的句义类型进行识别,也有一定的启发意义。

综上,各种方法存在两个主要问题:首先是标注规则复杂。如哈工大等依存关系多达几十种。这就造成了标注数据慢而且错误较多,在进行监督学习的时候,模型预测准确性难以提高;其次是算法不能识别语言成分的嵌套。基于深度端到端的模型,无法对嵌套词语进行预测,无法预测嵌套多类别。

本文提出了基于短语窗口的标注规则,发布了短语窗口数据集 CPWD,同时设计了相应的短语识别算法。该标注规则以短语为最小单位,把句子分成动词短语、名词短语等 7 类可嵌套的短语类型,同时标示出短语之间的依存关系。对应的算法,借鉴了图像领域识别目标区域的 Faster RCNN 算法思想,可以发现句子中各种短语的起始位置与结束位置,实现对嵌套短语及句法依存关系的同步识别,对应模型称为语法窗口模型 SWM。

2 短语标注规范

为了实现句子的短语识别与句法依存分析,制定了一套完整的短语标注规范。该短语标注规范不仅可以对嵌套短语进行标注,而且可以反映短语之间的依存关系。例如,嵌套词语“中国海军”,传统端对端模型只可以识别出“中国”“海军”或者“中国海军”,不可以同时识别 3 个短语,本文提出的短语标注规范和 SWM 模型解决了词语嵌套问题。相对于哈工大 LTP 提出的 28 种词性和 15 种依存分析^[29],百度 AI 开放平台提出的 24 种词性和 34 种依存关系^[30],本文提出的短语标注规则相对简单,

容易掌握。

2.1 句子句法依存关系

标注规范将句子中的短语分成名词短语、动词短语、数量词短语、介词短语、连词短语、语气词、从句共 7 类基本类型。句子由短语组成,7 类基本短语类型通过树状结构组成句子,即句法依存关系,如图 1 所示。

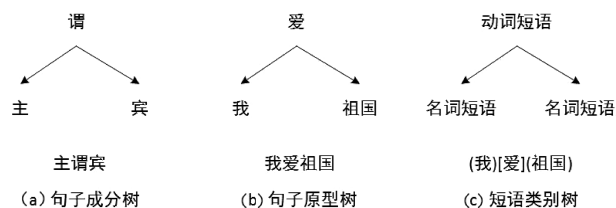


图 1 句子语法树结构图

通常,句子的树状结构由主、谓、宾关系组成,[图 1(a)]是句子成分树,句子“我爱祖国”按照句子语法可以分为主语“我”、谓语“爱”、宾语“祖国”,对应的句子原型树如图 1(b)所示,也就是把“我”“爱”“祖国”放到对应的主谓宾位置;对应的短语类别树如图 1(c)所示,其中“我”是名词短语,“爱”是动词短语,“祖国”是名词短语。

对于复杂的句子,同样可以采用这种方法进行短语识别和句法依存的分析。图 2 是复杂句子的语义单元划分过程。为了方便介绍,我们使用“()”表示名词短语,“[]”表示动词短语,“{ }”表示数量词短语,“< >”表示介词短语,“# #”表示连词短语,“@ @”表示语气词短语,“/ \”表示从句。

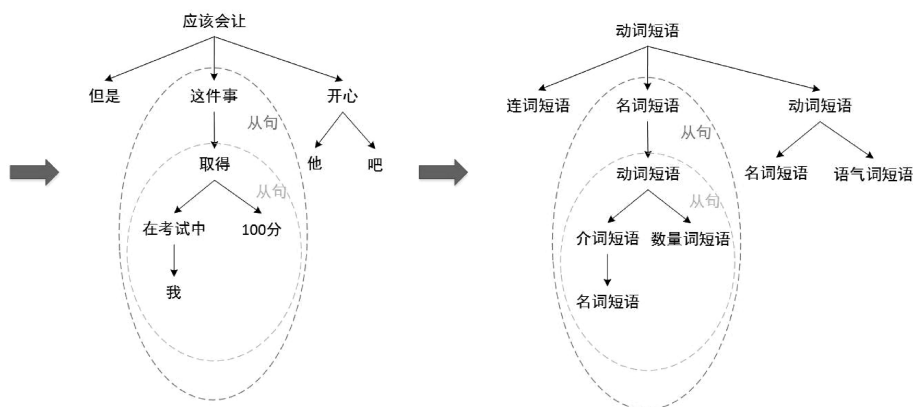
2.2 短语标注规范

根据中文的基本语法分析,常用词性包括名词、代词、动词、介词、形容词、数量词、语气词、连词和副词,本文通过组合不同词性形成模板,引入 7 种短语,基本囊括常用真实句子的所有情况。短语类别有名词短语“()”、动词短语 “[]”、数量词短语 “{ }”、介词短语 “< >”、连词短语 “# #”、语气词短语 “@ @”、从句 “/ \”。标注允许嵌套结构的存在,每种短语类别的标注规则如下。

(1) 连词短语

连词短语是用来连接词与词、词组与词组或句子与句子,表示某种逻辑关系的虚词。连词短语可以表并列、承接、转折、因果、选择、假设、比较、让步等关系。连词短语主要成分就是连词,如“但是”表转折,“因为”“所以”表因果等。

原句：但是我在考试中取得100分这件事应该会让开心吧



短语划分后：

#但是#/(我)<在(考试中)>[取得][100分]\([这件事])\[应该会让](他)[开心]@吧@

图2 语义单元划分过程

在标注体系中,连词短语一般无嵌套关系。

(2) 语气词短语

语气词短语是表示语气的虚词,常用在句尾或句中停顿处表示种种语气。语气词短语主要成分就是语气词,如“吗”“吧”“呢”等在词语、句子末表示语气。

在标注体系中,语气词短语一般无嵌套关系。

(3) 名词短语

名词短语表示人或事物以及时间、方位等,在句子中主要充当主语、宾语、定语。它包括:以名词为中心词的偏正短语(如“伟大祖国”“这些孩子”);用名词构成的联合短语(如“工人农民”);复指短语(如“首都北京”);方位短语(如“桌面上”“大楼前面”);“的”字短语(如“打更的老汉”)等。名词短语有多种形式的主要成分,模板种类较多,主要包括以动词、形容词为中心词,以代词、名词和其他名词短语为定语而组成的名词短语。

(4) 动词短语

动词短语代表动作,包括起修饰作用的状态与补语。动词短语主要成分包括:动词以及对应的状语和补语修饰词。如“马上开始了”包括状语“马上”、中心动词“开始”、补语“了”。

(5) 介词短语

介词短语又称为介宾短语,是介词和其他代词或名词或从句搭配形成的短语,形式如下:

- “在这次考试中”为介词+名词短语,标注为<在({这次}考试中)>。

- “被”“把”字句。如<被(他)>, <把(他)>。

(6) 数量词短语

数量词短语指由数词和量词组合构成的短语,形式如下:

- 数量词和名词搭配,如({一首}动听的曲子)。
- 作为状语,如[一蹦一跳]地走着]。
- 作为补语,如[看了](他){一眼}。

(7) 从句

为了标注一个完整的语义单元,需要使用从句结构体现短语间的层次关系。从句的主要成分比较多,其成分包括有名词、动词、介词等,从句允许存在嵌套现象。

- 兼语句标为从句。如“我命令他去外面”,这里“他”既是前面的宾语,又是后面的主语,标注:(我)[命令]/(他)[去](外面)\。
- 连动句标为从句。如“我出去骑车打球”,“骑车”“打球”为连动,标注为:(我)//[出去]\/[骑](车)\/[打](球)\。
- 主语从句、宾语从句。如(他)[说]/(计算机)[正在改变](世界)\。

根据该规则,我们标注了中文短语窗口数据集 CPWD,数据集包括 45 000 条从对话、新闻、法律、政策、小说中挑选的非文言文的中文句子。为了方便模型设计,句子最大长度限制在 50 个字以内。

3 算法

在短语标注方法中,一些标签存在嵌套现象,所以不能采用传统的端到端深度模型。为了适配新的短语标注方法,借鉴了图像领域 Faster RCNN 的算法思想,即找窗口的方法。不同的是,Faster RCNN 找目标区域窗口,本文的算法是找短语的位置窗口;Faster RCNN 会找对应的图像锚点 anchor 和 4 个偏置值,我们的算法找的是短语的锚点和两个偏置值。我们称这个算法对应的模型叫语法窗口模型(Syntax Window Model, SWM)。

3.1 算法流程

(1) **模型输入** 使用“我爱祖国”字向量作为输入。

(2) **短语特征提取** 通过特征提取网络进行特征提取,特征抽取模型可以采用 LSTM、BERT 等网络结构。

(3) **短语窗口定位** 对于特征提取结果,通过分类网络实现短语窗口定位。例如,窗口有“我”“爱”“祖国”三个,分类网络的输出是窗口在原句中的开始位置与结束位置。例如,窗口“我”的输出是(1,1),代表在句子中的开始位置是 1,结束位置是 1。

(4) **短语分类** 提取窗口对应的短语,通过分类网络识别出短语类别。例如,“我”是名词短语,“爱”是动词短语,“祖国”是名词短语。

(5) **得到结果** 将短语窗口定位结果与对应的短语分类结果综合,使用标注符号表示出来,得到结果,例如,“(我)[爱](祖国)”。

3.2 SWM 模型介绍

SWM 模型主要包括字向量输入层、特征提取层、短语窗口定位层、短语窗口提取层、分类网络组成。图 3 是短语识别算法模型的基本结构图。

模型主要包括四个组成部分:①特征提取网络:通过特征提取层实现特征的抽取,提取出的特征将被用于后续短语窗口定位网络和短语窗口提取。特征提取可以使用 LSTM 或者 BERT 模型。②短语窗口定位网络:短语窗口定位层用于定位短语窗口。③短语窗口提取层:收集输入的短语窗口位置 proposal、偏置值以及特征提取结果。④分类网络:使用输入的短语窗口位置,从句子中抽取对应的短语片段,对短语片段进行分类,同时会再次修正偏置值。

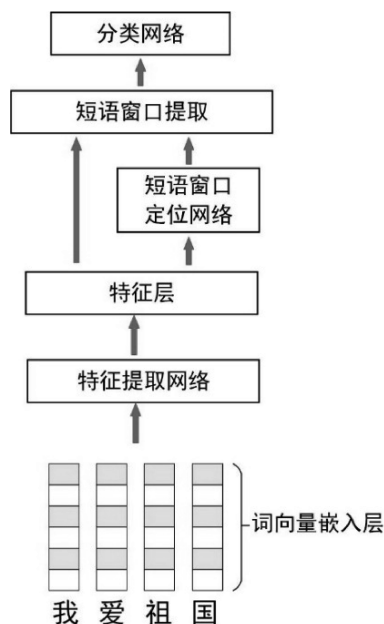


图 3 SWM 模型基本结构

3.2.1 短语窗口定位网络

短语窗口定位包括窗口穷举层、全连接层、判别和偏置层。

如图 4 所示,该层类似 Faster RCNN 的定位锚点 anchor,及推荐 proposal 的功能,可以从背景中定位锚点,也就是短语窗口,同时结合偏置值(Δx 与 Δy)修正短语窗口的位置,得到短语窗口的 proposal。

(1) 窗口穷举层

窗口穷举层用来生成候选窗口框,每个窗口由起始位置 x 和结束位置 y 组成的二元组决定,用 (x, y) 表示。图 5 是窗口穷举示意图,其中“我”是第一个字,它的窗口起始位置是 1,结束位置是 1,表示成(1,1),同理,“我爱”对应(1,2)。

采用所有的子窗口锚点 anchor 来生成候选窗口 proposal。当句子总长度为 N 时,第一个字作为起始位置的窗口有 N 个,第二个字作为起始位置的窗口有 $N-1$ 个,……,最后一个字作为起始位置对应的窗口数量有 1 个,所以总的窗口数量是 $N(N+1)/2$,对应同样数量的锚点 anchor。

(2) 背景和偏置层

所有 $N(N+1)/2$ 个子窗口的特征,经过全连接层后,输出是 $N(N+1)$ 个分类及对应的 $N(N+1)$ 个偏置值,如图 6 所示。也就是说,每个子窗口特征,经过全连接层后,输出两个分类的 one-hot 值,用于确定哪些短语窗口 anchor 是短语 proposal,以及 x 与 y 的偏置值 Δx 与 Δy 两个值,每个子窗口特征得到 4 个输出值。

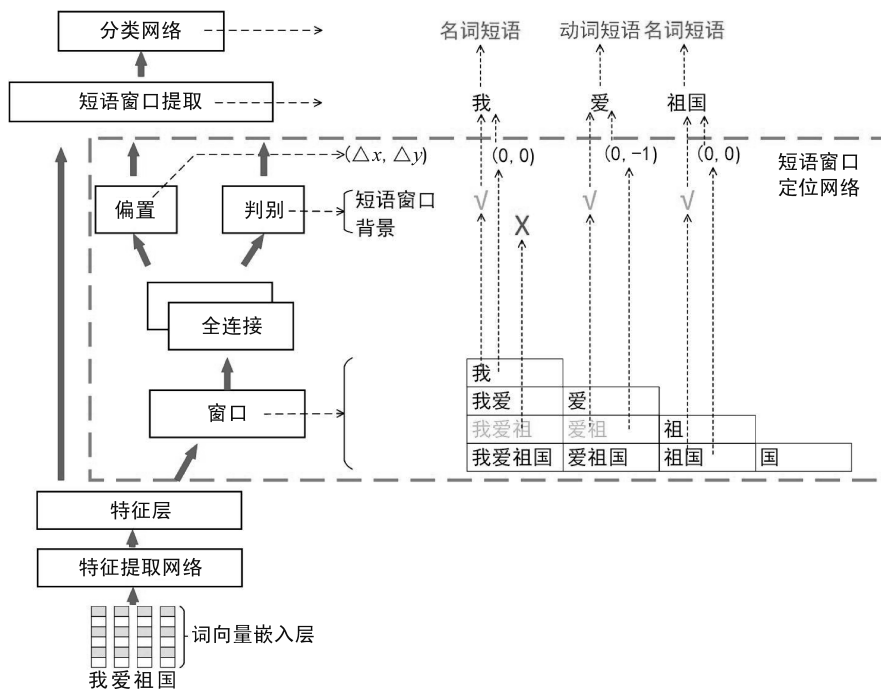
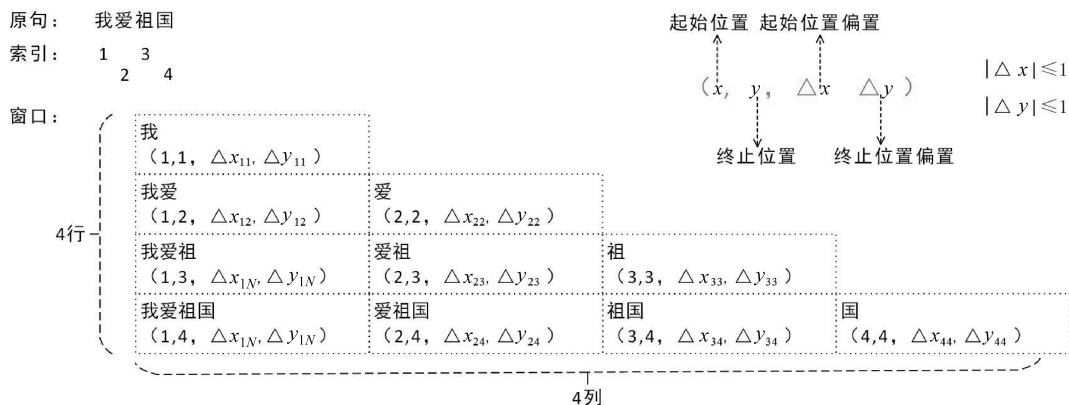


图4 短语窗口定位网络示意图



对于4个字的句子窗口数量: $4+3+2+1=10$

对于N个字的句子窗口数量: $N+(N-1)+(N-2)+\dots+2+1=N(N+1)/2$

图5 窗口穷举示意图

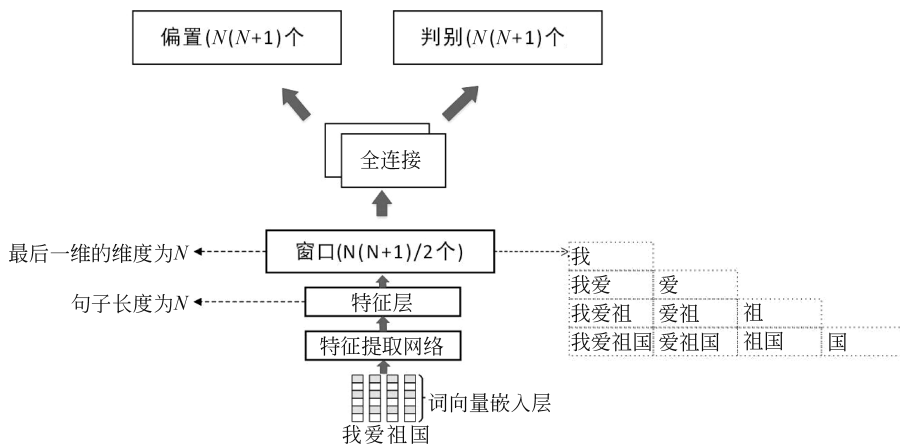


图6 窗口判别与偏置

为了使定位的窗口位置更加精确,增加两个偏置值 Δx 与 Δy ,相当于一个回归准确性的过程。

图 7 是偏置示意图,窗口“爱祖”

通过全连接层后,被识别为短

我 爱 祖 国 ...

语,而正确的短语是窗口“爱”,

图 7 偏置示意图

这就需要通过偏置值修正,把原

来的窗口坐标加上偏置值,原来的窗口 (x, y) 变为 $(x + \Delta x, y + \Delta y)$ 。其中, Δx 表示的是窗口起始位置偏置值, Δy 表示的是窗口结束位置偏置值。

偏置值计算举例:

- $x=6, \Delta x=1$, 修正结果: $x=7$
- $x=6, \Delta x=0$, 修正结果: $x=6$

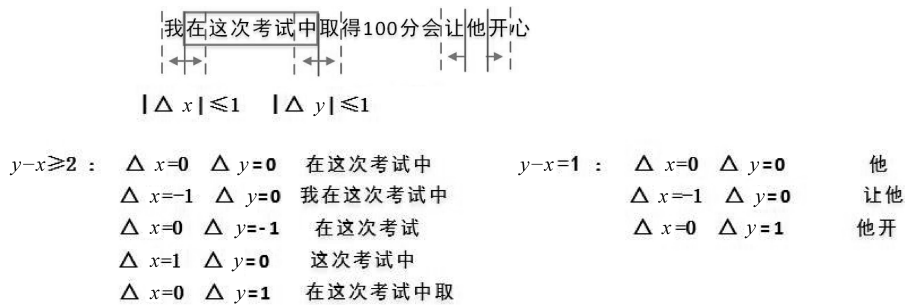


图 8 有效偏置示意图

(3) 损失函数

模型对应的损失函数有两个,分别是式(1)的判别损失函数和式(2)的偏置损失函数。式(1)是判别损失函数,采用的是交叉熵损失函数,其中 M 是短语类别数量,即 $M=7$,是分类(one-hot 向量)的损失函数,即判断是否为短语窗口的判别损失函数。式(2)是偏置损失函数,是回归偏置值 Δx 与 Δy 的损失函数,即短语窗口和真实窗口之间的误差,由于是多个值,所以采用的是均方根误差 RMSE。可以看出,窗口偏置值越大,损失越大,此处仅适用于训练过程。测试过程中,规定当预测窗口坐标和真实短语窗口坐标偏差小于 1 时,窗口的偏置才是有效的。

$$\text{Loss}_{\text{判别}} = - \sum_{c=1}^M y_c \log(p_c) \quad (1)$$

$$\text{Loss}_{\text{偏置}} = \sqrt{\frac{1}{M} \sum_{c=1}^M (y_c - \hat{y}_c)^2} \quad (2)$$

3.2.2 分类网络

分类网络利用已经获得的短语窗口 proposal 和偏置,通过全连接层与 Softmax 层推算每个短语窗口具体属于哪个类别(如介词短语、名词短语、数量词短语等),输出每个短语类别的概率向量,得到

- $x=6, \Delta x=-1$, 修正结果: $x=5$
- ...

为了防止预测窗口 proposal 过多,规定当预测窗口坐标和真实短语窗口坐标偏差小于 1 时,窗口的偏置才是有效的,其中,这里的坐标偏差指的是窗口边界的索引值差值,按字计算,规定预测标签在真实标签右边为正值,左边为负值。图 8 是有效偏置示意图。短语起始位置可以向左或右最多偏移一个字,终止位置也可以向左或右最多偏移一个字。当窗口长度大于或等于 2 时,规定最多有 5 种偏移情况(proposal),当窗口长度等于 1 时,最多只有 3 种偏移情况(proposal)。

单一分类标签;在预测阶段,同时再次利用回归偏置获得每个短语的位置偏移量、用于微调的短语窗口位置;在训练阶段,不需要根据位置偏移值再微调短语窗口位置,因为训练阶段引入式(2)偏置损失函数的作用,就是为了保证位置偏移值尽可能小,即不需要微调。

4 实验

4.1 数据集与评估标准

实验数据集使用标注的中文短语窗口数据集 CPWD,包括 45 000 条从对话、新闻、法律、政策、小说中挑选的中文句子。其中文言文只占不到 5%,而且多是一些成语与谚语组成的句子。为了方便模型设计,句子最大长度限制在 50 个字以内。训练集、验证集和测试集的数据随机选用了 35 000 条、5 000 条和 5 000 条,在训练集和验证集上采用了交叉验证方法。

4.2 实验环境

实验采用 Python 语言实现,Python 版本为

3.6.1.使用的框架为 TensorFlow,版本为 1.12.0。使用的计算机配置为内存: 32GB,处理器: Intel Xeon(R) CPU E5-2623 v3 @3.00GHz * 8,显卡: TITAN XP,操作系统类型: Ubuntu 14.04 64-bit。

4.3 实验分析

首先,SWM 在多种网络结构的情况下进行对比优化。优化方法包括采用双层 BiLSTM,采用 BERT 代替 BiLSTM,使用 CRF 层,在 BiLSTM 层之前加入 CNN 层进行特征抽取,选取不同比例的反例。在本文提出的 CPWD 数据集上,利用 SWM 网络在不同的特征抽取网络下进行测试,在结果如表 2 所示。

表 2 中,ACC 指的是平均准确率,即 $ACC = \frac{1}{N} \sum_{i=1}^N \frac{x_i}{y_i}$,其中 N 为测试集样本总数, x_i 为第 i 个预测样本预测正确的短语个数, y_i 为第 i 个预测样本的短语总数。 F_1 指的是平均 Macro- F_1 ,即 $F_1 = \frac{1}{N} \sum_{i=1}^N Macro F_1$,其中 $Macro F_1 = \sum_{j=1}^K \frac{TP_j}{TP_j + FP_j}$,其中,TP 为预测正确的正例数,FP 为预测错误的正例数, K 为短语种类数, N 为测试样本总数。前向传播时间,指的是测试过程中,平均每个测试样本从输入的词向量层到输出的短语类别标签结果的整个过程的耗时,单位为 ms。

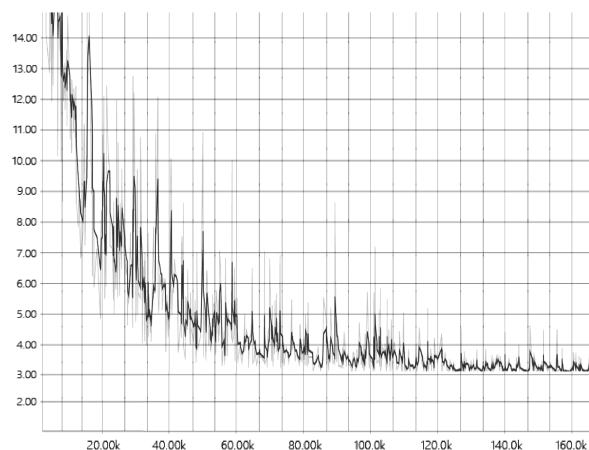
表 2 SWM 优化过程对比

	ACC/%	F_1 /%	前向传播时间/ms
1 层 BiLSTM	88.11	86.42	65
1 层 BiLSTM + CRF	88.41	86.62	78
2 层 BiLSTM	88.12	87.02	80
2 层 BiLSTM + CRF	88.52	87.37	94
BERT	88.79	88.65	106
BERT+CRF	88.94	88.91	117
CNN+BiLSTM	87.28	87.51	96
CNN+BiLSTM+CRF	87.51	87.84	104
1 : 1 正反例	86.84	87.23	87
1 : 2 正反例	87.87	87.67	89

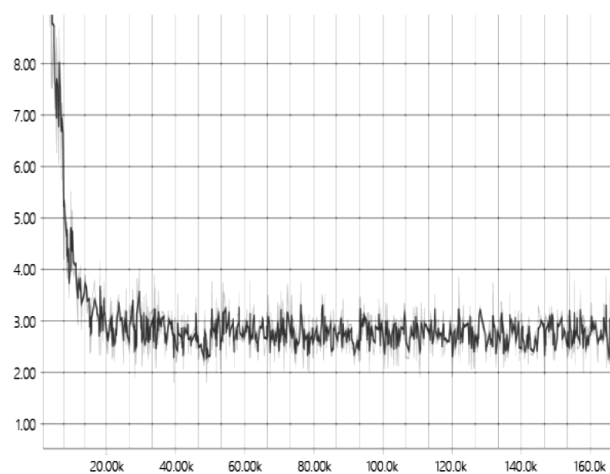
从表 2 可以看出,最优的模型结构是: 1 : 2 正反例、BERT+CRF。但是考虑到运行效率和资源占用情况,我们在做实验或者工程部署的时候,建议

采用的特征层模型是: 1 : 2 正反例,2 层 BiLSTM。

Loss 值收敛情况在 BERT、BiLSTM 两种最优模型之间进行比较,收敛情况如图 9 所示。可以看到,BiLSTM 收敛快些,BERT 收敛慢些,而且一个 epoch 的训练时间也长一些。



(a) BERT



(b) BiLSTM

图 9 Loss 收敛情况

4.4 对比实验结果

由于各种端到端算法输出与标签形式有所不同,所以将标签形式调整为统一的方式进行对比。SWM 模型输出标签比端到端多,附带有嵌套等信息,所以需要进行降维处理,之后与端到端模型进行对比。降维后可以形成命名实体标签、依存标签,分别与 BiLSTM、BERT 等端到端算法进行对比。表 3 为 CPWD 的测试集的实验结果,是不同模型基于 CPWD 的训练集进行训练的结果。

可以看到 SWM 模型比传统端到端模型有优势。这种优势的产生,我们分析是由于 SWM 更适合语言的嵌套特征,使得模型不用在多个嵌套的命

名实体之间做多选一的抉择,降低了模型的困惑度,解耦了文字与标签一对一的限制。同时,我们将 SWM 模型用于 CCL2018 的中文隐喻情感分析任务中,通过抽取名词短语、动词短语、介词短语和语气词短语,分析不同实体词对应的正负面情感词,并将此作为情感分类模型输入的一部分进行模型优化, F_1 在原有模型基础上提升了 1 个点以上,并在比赛中获得了第一名的成绩。

表 3 实验对比结果 (单位: %)

	命名实体 F_1	依存分析 F_1
BiLSTM	88.02	86.65
BiLSTM+CRF	88.95	87.84
BERT	90.78	89.12
BERT+CRF	90.22	89.89
CNN+CRF	88.34	87.75
SWM(BiLSTM)	90.20	89.14
SWM(BERT)	91.24	90.78

5 总结

本文针对传统自然语言处理存在的问题进行改进。包括两个方面的问题:标注方法过于复杂,对应的深度端到端算法无法解决短语嵌套问题。首先,定义了基于短语窗口的标注方法;然后,标注了中文短语窗口数据集 CPWD;最后,给出了对应的语法窗口识别算法模型 SWM。新的标注规则以短语为最小单位,把句子分成 7 类可嵌套的短语类型,同时标示出短语之间的句法依存关系,易于实施。SWM 模型借鉴了图像领域识别目标区域(Faster RCNN)的思想,可以发现句子中各种短语的起始位置与结束位置,实现了对短语及句法依存关系的同步识别,解决了嵌套短语问题。实验的结果表明,SWM 模型比端到端模型更加契合短语嵌套特征,提高了识别准确性。

参考文献

- [1] ABNEY S P. Parsing by chunks [M]. Principle-based Parsing, Springer Netherlands, 1991: 257-278.
- [2] KUDO T, MATSUMOTO Y. Chunking with support vector machines [C]//Proceedings of the 2nd Meeting of the North American Chapter of the Association for

Computational Linguistics, 2001: 1-8.

- [3] SHEN H, SARKAR A. Voting between multiple data representations for text chunking [C]//Proceedings of the Advances in Artificial Intelligence. Springer, Berlin Heidelberg, 2005: 389-400.
- [4] MANCEV D. A sequential dual method for the structured ramp loss minimization [J]. Facta Universitatis, Series: Mathematics and Informatics, 2015, 30(1): 13-27.
- [5] 周强. 基于规则的汉语基本块自动分析器 [C]//第七届中国信息处理国际会议论文集, 2007: 148-153.
- [6] 周强. 汉语基本块规则的自动学习和扩展进化 [J]. 清华大学学报(自然科学版), 2008, 48(01): 88-91.
- [7] 李超, 孙健, 关毅, 等. 基于最大熵模型的汉语基本块分析技术研巧 [R]. CIPS-ParsEval, 2009.
- [8] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.
- [9] KURU O, CAN O A, YURET D. CharNER: Character-level named entity recognition [C]//Proceedings of the 26th International Conference on Computational Linguistics. Association for Computational Linguistics, 2016: 911-921.
- [10] 侯潇琪, 王瑞波, 李济洪. 基于词的分布式实值表示的汉语基本块识别 [J]. 东北大学学报(自然科学版), 2013, 34(5): 582-585.
- [11] 李国臣, 党帅兵, 王瑞波, 等. 基于字的分布表征的汉语基本块识别 [J]. 中文信息学报, 2014, 28(06): 18-25.
- [12] 徐菁. 面向中文知识图谱的开放式文本信息抽取关键技术研究 [D]. 国防科技大学硕士学位论文, 2018.
- [13] 程钟慧, 陈珂, 陈刚, 等. 基于强化学习协同训练的命名实体识别方法 [J]. 软件工程, 2020, 23(01): 7-11.
- [14] 徐烈炯. 题元理论与汉语配价问题 [J]. 当代语言学, 1998(3): 1-21.
- [15] 刘宇红. 生成语法中词汇语义与句法的界面研究 [J]. 外语学刊, 2011(05): 56-60.
- [16] 孙道功. 基于大规模语义知识库的“词汇—句法语义”接口研究 [J]. 语言文字应用, 2016(02): 125-134.
- [17] 亢世勇, 许小星, 马永腾. 施事、受事句法实现的义类制约 [J]. 语文研究, 2011(04): 36-40.
- [18] MCDONALD R, LERMAN K, PEREIRA F. Multilingual dependency analysis with a two-stage discriminative parser [C]//Proceedings of the 10th Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2006: 216-220.
- [19] NIVRE J, HALL J, NILSSON J, et al. Labeled pseudo-projective dependency parsing with support vector machines [C]//Proceedings of the 10th Conference on Computational Natural Language Learning

- 2006: 221-225.
- [20] REN H, JI D, WAN J, et al. Parsing syntactic and semantic dependencies for multiple languages with a pipeline approach[C]//Proceedings of the 13th Conference on Computational Natural Language Learning, 2009: 97-102.
- [21] CHE W, LI Z, LI Y, et al. Multilingual dependency-based syntactic and semantic parsing[C]//Proceedings of the 13th Conference on Computational Natural Language Learning, 2009: 49-54.
- [22] DYER C, BALLESTEROS M, LING W, et al. Transition-based dependency parsing with stack long short-term memory[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing, 2015: 334-343.
- [23] JI T, WU Y, LAN M. Graph-based dependency parsing with graph neural networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019 (1): 2475-2485.
- [24] WANG Y X, CHEN W X, GUO J, et al. A neural transition-based approach for semantic dependency graph parsing[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018: 5561-5568.
- [25] FRIED D, KLEIN D. Policy gradient as a proxy for dynamic oracles in constituency parsing [C]//Proceedings of the ACL, 2018: 469-476.
- [26] 丁伟伟, 常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报, 2009, 23(5): 53-61.
- [27] 王丽杰. 汉语语义依存分析研究[D]. 哈尔滨: 哈尔滨工业大学硕士学位论文, 2010.
- [28] 王倩, 罗森林, 韩磊, 等. 基于谓词及句义类型块的汉语句义类型识别[J]. 中文信息学报, 2014, 28(2): 8-16.
- [29] 哈工大社会计算与信息检索研究中心. 语言云[OL]. <http://www.ltp-cloud.com/intro>[2020-9-22]
- [30] 百度 AI 开放平台. 语言处理技术 NLP 接口文档说明[OL]. <https://ai.baidu.com/ai-doc/NLP/>[2020-9-22]



刘广(1996—), 硕士研究生, 主要研究领域为自然语言处理、关系抽取、知识抽取。
E-mail: 960987802@qq.com



涂刚(1976—), 通信作者, 博士, 副教授, 主要研究领域为人工智能、机器学习、深度学习、无人机影像分析和自然语言处理。
E-mail: tugang@hust.edu.cn



李政(1998—), 硕士研究生, 主要研究领域为自然语言处理、关系抽取、知识抽取。
E-mail: jmxgodlz@hust.edu.cn

~~~~~  
(上接第 14 页)



崔洪振(1990—), 博士研究生, 主要研究领域为信息抽取、领域知识计算。  
E-mail: chz\_367@163.com



张龙豪(1996—), 硕士, 主要研究领域为实体抽取、知识融合。  
E-mail: zlh\_ustb2022@163.com



彭云峰(1971—), 通信作者, 博士, 教授, 主要研究领域为领域知识、自然语言处理、医学救援通信技术与智能信息系统。  
E-mail: pengyf@ustb.edu.cn