

文章编号: 1003-0077(2019)06-0080-08

多场景文本的细粒度命名实体识别

盛剑¹, 向政鹏¹, 秦兵¹, 刘铭¹, 王莉峰²

(1. 哈尔滨工业大学 社会技术与信息检索研究中心, 黑龙江 哈尔滨 150001;

2. 腾讯科技(深圳)有限公司, 广东 深圳 518000)

摘要: 命名实体识别一直是数据挖掘领域的经典问题之一, 尤其随着网络数据的剧增, 如果能对多来源的文本数据进行多领域、细粒度的命名实体识别, 显然能够为很多的数据挖掘应用提供支持。该文提出一种多领域、细粒度的命名实体识别方法, 利用网络词典回标文本数据获得了大量的粗糙训练文本。为防止训练文本中的噪声干扰命名实体识别的结果, 该算法将命名实体识别的过程划分为两个阶段, 第一个阶段先获得命名实体的领域标签, 之后利用命名实体的上下文确定命名实体的细粒度标签。实验结果显示, 该文提出的方法使 F_1 值在全领域上平均值达到了 80% 左右。

关键词: 命名实体识别; 细粒度类别划分; 语料回标

中图分类号: TP391

文献标识码: A

Fine-grained Named Entity Recognition for Multi-scenario

SHENG Jian¹, XIANG Zhengpeng¹, QIN Bing¹, LIU Ming¹, WANG Lifeng²

(1. Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China; 2. Tencent Technology(Shenzhen) CO., Ltd. Shenzhen, Guangdong 518000, China)

Abstract: Name entity recognition is a classical research issue in data mining community. To recognize the entities in multi-domain with fine-grained labels, we propose a method of utilizes web thesaurus to annotate web data automatically to acquire large-scale training corpus. To minimize the influence of the noises in training corpus, we design a two-phase entity recognition method. First, the entity's domain label is obtained. After that, the context of each recognized entity is used to determine the fine-grained label for one entity. Experimental results demonstrate that the proposed method can obtain high accuracy on entity recognition in multiple domains.

Keywords: named entity recognition; fine-grained category annotation; corpus annotation

0 引言

命名实体识别(named entity recognition, NER)是自然语言处理领域(natural language processing, NLP)的基础任务之一, 也是信息抽取中最为重要的一个子任务, 并且可以对后续的抽取任务提供帮助。命名实体识别任务意在识别文本中的事物的名称, 例如人名、地名和机构名。本文主要在多场景多领域下研究命名实体识别, 以 LSTM-CRF 为基础并引入 CNN(卷积神经网络)从文本中进一步提取有用的语义特征。

早期的命名实体识别大多是基于规则的方法, 但是由于语言结构本身具有不确定性, 制订出统一完整的规则难度较大。基于规则的方法需要构造特定的规则模板, 采用的特征包括统计信息、标点符号、关键字、位置词、中心词等, 以模式和字符串相匹配为主要手段, 尤其依赖于知识库和词典的建立。针对不同领域, 需要专家重新书写规则, 代价较大, 存在规则建立周期长、移植性差且需要建立不同领域知识库作为辅助以提高系统识别能力等问题。

传统的命名实体识别方法大多采用有监督的机器学习模型, 如隐马尔可夫模型^[1]、最大熵^[2-3]、支持

收稿日期: 2019-01-28 定稿日期: 2019-02-28

基金项目: 国家自然科学基金(61632011, 61772156, 61472107)

向量机和条件随机场^[4]等。最大熵模型具有结构严谨、通用性良好的特点,但训练时间复杂度高,需要明确的归一化计算^[5],导致计算上的开销比较大。条件随机场在分词和命名实体识别上表现出色,提供了一个特征灵活、全局最优的标注框架,但同时存在收敛速度慢、训练时间长的问題。通常来讲,最大熵和支持向量机在正确率上比隐马尔可夫模型略高,但是隐马尔可夫模型在训练和识别时的速度要更快一些,主要是由于 Viterbi 算法在解码时具有较高的效率。基于统计的方法对特征的选取依赖性较高,需要从文本中分析选择对于此项任务影响因子较大的特征,并将这些特征加入到特征模板中,通过对训练语料所包含的语言语义信息进行统计和分析,进行有效的特征选择,从训练语料中不断发现强特征。有关特征可以分为具体的停用词特征、上下文特征、词典及词性特征、单词特征、核心词特征以及语义特征等。与基于规则的方法相似,基于统计的方法对于语料库的依赖性也较大,而建立较大的领域语料库又是一大难点。

Natural language processing (almost) from scratch^[6]是使用神经网络进行命名实体识别较早的工作,文中,作者提出使用窗口方法和句子方法两种网络结构来进行命名实体识别。窗口方法的输入为预测词的上下文窗口,使用传统的神经网络进行求解。句子方法将整个句子作为当前预测词的输入,加入句子中相对位置特征来区分句子中的每个词,然后使用一层卷积神经网络求解。训练时作者提出两种目标函数,一是词级别的对数似然函数,使用 softmax 来预测标签概率,当成一个传统的分类问题来做;二是句子级别的似然函数,即考虑到 CRF 模型在序列标注任务中天然的优势,将标签之间的转移分数加入到目标函数中。这也是后来神经网络——CRF 模型的先驱工作。

当前最好的实体识别模型是 LSTM-CRF 模型^[7-8],该网络由两个长短时记忆网络组成,一个前向记忆网络和一个后向记忆网络,前者用于学习前向的序列信息,后者用于学习后向的序列信息,得到每个隐层的表示,将隐层映射到所需分类的特征维度,之后选取概率最大的一维作为其实体类别,该方法也称之为 softmax^[9]。尽管该模型在独立的序列标注任务中取得了成功,例如词性标注,但是该模型忽略了标签间的依赖关系,这一缺点导致了部分精度的损失。实体识别任务存在某些内在限制,例如 I-PER 标签并不能接在 B-LOC 标签的后边。因此,

用条件随机场模型(CRF)来学习标签之间的关系,而不是进行独立的标注。目前将神经网络和 CRF 模型相结合的模型已经成为命名实体识别的主流模型。由于 RNN 有天然的序列结构,所以 RNN-CRF 使用更加广泛。使用神经网络有天然的无需大量人工特征的优势,只需要词向量或者字符向量就可以达到主流的水平,加入高质量的词典特征可以进一步提升效果。

本文提出的细粒度命名实体识别算法以 LSTM-CRF 模型为基础,并引入 CNN 从文本中进一步提取有用的语义特征做实体边界划分,之后交给细粒度划分模块做小类别分类。第一步通过使用 RNN 对命名实体的上下文进行表示,并使用 softmax 分类确定该命名实体所属的大类别(即领域);第二步利用每个领域下的语料构建模型以确定命名实体所属的细粒度类别(小类别)。实验结果显示,命名实体识别的 F_1 值在全领域上平均值达到 80% 左右,能在一定程度上说明实现的分阶段方案是有效的。

1 方法描述

本文实现的模型共包含三个模块:语料回标模块、命名实体识别模块、命名实体细粒度划分模块。在获取某个词条为命名实体后,则将该命名实体和命名实体的上下文交由细粒度的命名实体类别划分模块,先由模块确定该命名实体的大类别标签(所属领域),然后再确定该实体的小类别标签。对句子的识别提供了有交叉多标签的命名实体识别结果,即有可能出现输入“ABCDE”,识别结果为“AB”和“BC”为不同领域下的命名实体。另外,在对实体打标签时也会出现一个实体有多个标签的情况,类似于“马龙”可能是一个人名,也是一个体育明星,这时该方案会对命名实体提供属于多个领域的细粒度类别标签。

举例来说,输入文本为:

赵丽颖,1987 年 10 月 16 日出生于河北省廊坊市,华语影视女演员,毕业于廊坊市电子信息工程学校。

输出文本为:

[赵丽颖][影视明星,人名],1987 年 10 月 16 日出生于[河北省][地名][廊坊市][地名],华语影视女演员,毕业于[廊坊市电子信息工程学校][机构名,教育机构]。

该文本在经过 12 个领域的命名实体识别模块

后,在传统、娱乐和教育领域均有命名实体被识别出来,将结果全部保留送入细粒度类别划分模型得到最终的分类。此例中,赵丽颖被识别成影视明星和

人名,河北省和廊坊市被识别为地名,廊坊市电子信息工程学校被识别成机构名和教育机构名。

模型的整体结构如图 1 所示。

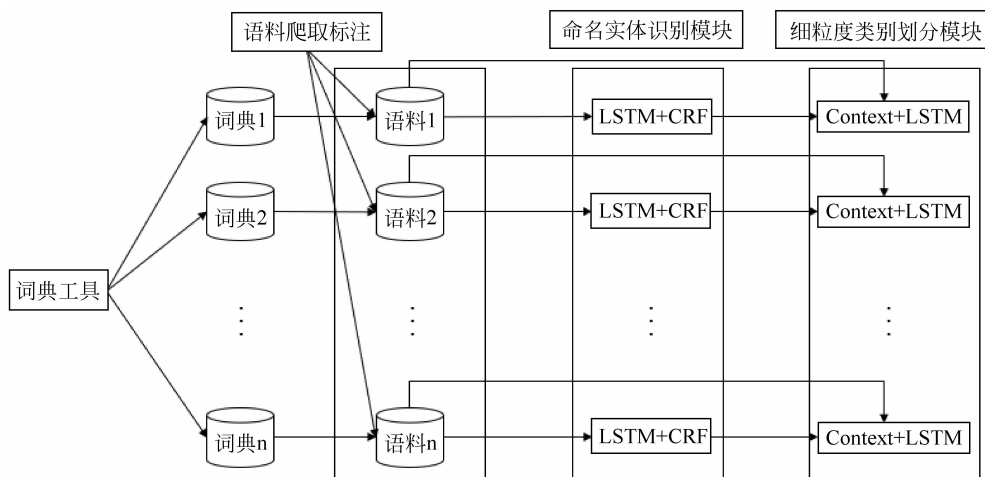


图 1 模型整体结构

1.1 词典回标模块

在词典回标模块中,从网络中爬取了除人名、地名、机构名外其他类别下的细粒度命名实体词典。并从网络中爬取文本数据,利用词典回标数据形成大规模的训练集、开发集和测试集。

实现阶段发现,自动标注的数据里面有很多误标注的数据(噪声)存在,例如把“你的名字是什么”这句话标注为“[你的名字][影视剧]是什么”,即认为“你的名字”是个影视剧,但是显然该词条不是一个命名实体。

通过抽样分析,发现语料回标的质量不是很高,回标结果举例,如表 1 所示。

表 1 回标结果举例

领域	示例	回标结果是否正确
餐饮	清蒸[黄花鱼][食品名],可以说是“一鱼两吃”吧	正确
餐饮	[苹果][食品名]申请智能保护套专利以防 ipad 跌落	错误
旅游	【超值套餐】[千岛湖][景点]梅地亚君澜豪华 2 天 1 晚美食套餐自由行	正确
旅游	雅马哈鬼火天蝎[黄龙][景点]600 摩托车改装配件	错误
电商	[七匹狼][品牌]家居裤男士睡裤纯棉长裤单件装纯色款 xxl	正确
电商	[中华][品牌]成语故事小学生语文新课标必读书系	错误

基于此类问题,人工对粗糙的机器标注的语料

进行二次标注,得到了小规模的较为准确的标注语料。

1.2 命名实体识别模块

本文以 LSTM+CRF 构建基准的命名实体识别模块^[10-11],引入 CNN 特征提取模块^[12],用于识别输入文档中的命名实体。

卷积神经网络最早应用在计算机视觉任务上,如今也广泛应用在自然语言处理任务上。

卷积神经网络其实就是多层卷积运算,然后对每层的卷积输出用非线性激活函数做转换。卷积过程中每块局部的输入区域与输出的一个神经元相连接。对每一层应用不同的卷积核,每一种卷积核可以理解为是对一种特征进行提取,然后将多种特征进行汇总。

CNN 在句子建模上有着广泛的应用,CNN 强大的特征捕捉能力,使得在句子建模过程中,先组合底层邻近的词语信息,逐步向上传递,上层又组合新的短语信息,从而使得相距较远的句子也存在联系,这种联系通常是语义上的联系。

本文卷积的时候是整行进行的,卷积核的高为词向量的维度,宽为 2、3 不等。如图 2 最下方 CNN 提取句子特征部分所示,宽为 2 的卷积核对整个序列进行卷积得到维度为 4 的向量。对每个卷积核卷积得到的向量通过 max pooling 操作,将结果拼接在一起,得到的向量作为整个句子的表示。将此句子表示加入到 LSTM 获取到的每个词的上下文表

示,从而使每个词可以既具有句子表示又可以具有词级别表示。

LSTM+CNN+CRF 的基本结构,如图 2 所示。

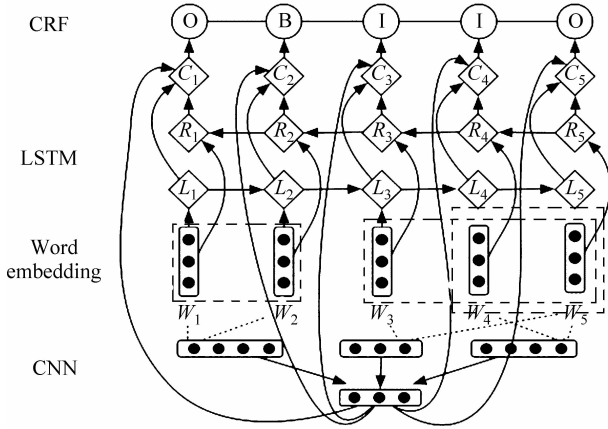


图 2 LSTM+CNN+CRF 基本结构

LSTM+CRF 为两层结构,其中输入表示层由双向 LSTM 组成,通过 LSTM 中的每个 cell 单元对输入文档进行编码,将隐层得到的输出与 CNN 提取的类似 n-gram 的特征进行拼接^[13],通过全连接层进行 tag 分类,得到的矩阵为 CRF 层的发射矩阵^[14]。

在 CRF 中,给定一个观察序列 x ,目的是希望找到一个概率最大的标记序列 y 。在命名实体识别任务中, x 表示词序列, y 表示命名实体标记。那么给定的概率可由式(1~3)计算:

$$p(\vec{y} | \vec{x}) = \frac{1}{Z(\vec{x})} \prod_{j=1}^n \phi_j(\vec{x}, \vec{y}) \quad (1)$$

$$Z(\vec{x}) = \sum_{\vec{y}} \prod_{j=1}^n \phi_j(\vec{x}, \vec{y}) \quad (2)$$

$$\phi_j(\vec{x}, \vec{y}) = \exp\left(\sum_{i=1}^m \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad (3)$$

在式(1~3)中, j 代表词序列 \vec{x} 中的第 j 个位置, n 表示 \vec{x} 的长度, m 表示特征的数量, $f_i(y_{j-1}, y_j, \vec{x}, j)$ 是 CRF 的特征函数, λ_i 则是对应特征函数的权重。假设 \vec{x} 中的一个子串 $x_k x_{k+1} \cdots x_{k+l}$ 被标注为一个命名实体,记为 N_e ,则 N_e 的标记序列为 $y_k^* \cdot y_{k+1}^* \cdots y_{k+l}^*$,记做 \vec{y}_{N_e} 。使用式(4)计算边缘概率 $\varphi(N_e)$,作为识别到的命名实体的置信度。计算时,可采用前向后向算法^[15]。

$$\varphi(N_e) = \frac{1}{Z(\vec{x})} \sum_{\vec{y}': y'_k \cdots y'_{k+l} = \vec{y}_{N_e}} p(\vec{y}' | \vec{x}) \quad (4)$$

根据每个被识别出的命名实体在语料中出现的频率和命名实体的置信度($\varphi(N_e)$)设定阈值,然后

选择大于阈值的命名实体作为可信的标注结果。

条件随机场的目标函数在考虑了状态特征函数的同时,还包含有标签之间转移特征函数。使用 SGD 学习模型参数,在给定已训练好的模型时,给定输入序列,预测输出序列使目标函数最大化的最优序列,是一个动态规划问题,使用维特比算法进行解码。

1.3 细粒度类别划分模块

本文将命名实体的细粒度类别划分切分为两个步骤:①根据命名实体及命名实体的上下文进行领域(大类别标签)的划分;②根据第一步得到的领域信息(大类别标签)进行类别的细分,以得到命名实体的小类别标签。

本文模型采用 Tang^[16]提出的对句子中目标词的建模方法对命名实体的表示进行建模,这个模型在中文零指代消解任务^[17]中也发挥了较大的作用。具体来说,在第一步通过使用两个单向 LSTM 对命名实体的上文和下文分别进行表示,再将上文的表示和下文的表示拼接在一起作为命名实体的表示,最后使用 softmax 分类确定该命名实体所属的大类别(即领域);第二步利用每个领域下的语料构建同样的模型,以同样的方式确定命名实体所属的细粒度类别(小类别)。在两个步骤中使用同样的模型,区别在于:第一个步骤分类的类别为 12 个领域,第二个步骤将训练 12 个分类器,每个分类器只对一个领域下的小类别标签进行分类。模型结构如图 3 所示。

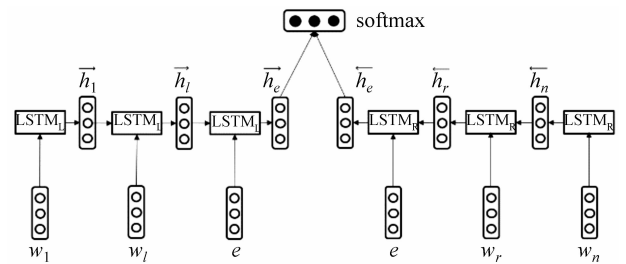


图 3 命名实体建模模型

依据模型,我们对每个识别出的命名实体采用两个单向 LSTM 分别表示其上下文。上下文的表示分为上文表示和下文表示两部分。其中,中间的两个 e 表示的是同一个命名实体, w_1 到 w_l 表示句子中 e 实体之前的词, w_r 到 w_n 表示句子中 e 实体之后的词。将句子中的第一个词 w_1 开始到命名实体 e 为止的所有词,从前往后依次读入 $LSTM_L$ 中,得到命名实体在上文中的表示,记为 \vec{h}_e 。将句子中

的最后一词 w_n 开始到实体 e 为止的所有词,从后往前依次读入 $LSTM_R$ 中,得到命名实体的下文表示,记为 \vec{h}_e 。将这两个表示拼接在一起之后,得到命名实体在句中的表示,然后使用 softmax 进行分类。

在建模命名实体的上下文时,本文对实体的表示按以下规则进行处理:

- (1) 每次仅对一个命名实体进行分类;
- (2) 句中如果有多个命名实体,则每次仅处理一个命名实体,将其余的部分都看作是上下文;
- (3) 命名实体如果由多个词组成,直接将多个词看作是一个词。

实验过程中,第一个步骤使用 12 个领域语料的训练集进行训练(在训练的时候会对训练语料随机打乱顺序),然后在 12 个领域语料的开发集上进行模型的选择,最后保留在开发集上效果最好的模型。在第 2 个步骤中,对 12 个领域中每一个领域,分别按上述方法利用该领域的训练语料训练分类器,最终得到 12 个个分类器。每一个分类器都做一个领域下的命名实体的细粒度划分。

表 2 标签体系

领域(大类别)	细粒度标签(小类别)	领域(大类别)	细粒度标签(小类别)
传统	人名、地名、机构名	旅游	景点、酒店、航空公司、车站
教育	学校/培训机构、课程名、书籍名、作者	餐饮	品牌、菜名、食品名、餐馆名、商场/购物中心
游戏	游戏名、虚拟角色、道具、技能	房产	小区、开发商、物业公司、家居/家具品牌
娱乐	明星、影视剧、歌曲/专辑、综艺	电商	品牌、店铺名、快递公司
金融	股票、股票代码、理财产品	体育	明星、体育项目、赛事
汽车	品牌、车系、车型、零部件	医疗	医院、疾病、科室、药品、症状

2.2 数据集

本实验采集了与需求对应的 12 个类别下的命名实体构建词典。

在获取词典后,利用从网络爬取和腾讯方面给出的文本数据进行回标,获取每个类别下的数据集,句子级数据集规模(单位:条)如表 3 所示。

表 3 回标数据集统计

领域	句子条数	领域	句子条数
传统	14 060	旅游	20 500
教育	13 559	餐饮	14 466
游戏	20 716	房产	14 811
娱乐	9 552	电商	23 543
金融	5 797	体育	61 429
汽车	7 821	医疗	14 605

2 实验分析

本文将细粒度命名实体的识别分为两个阶段。第一个阶段利用序列标注算法确定某个词条是否为命名实体(利用 12 个领域下的命名实体识别器判断某一词条是否为该领域下的命名实体),第二个阶段根据第一阶段识别出的命名实体进行细粒度类别划分。因此实验部分分别对这两个阶段的效果进行检测。最后,检测两个阶段合二为一后的效果,即给定一个句子,模型识别出其命名实体以及命名实体的领域和细粒度标签信息的效果。

2.1 标签体系

本文从领域角度出发,定义一套命名实体(Named Entity, named entity)标签体系,研发面向互联网多场景文本数据的命名实体识别通用解决方案,提升文本词法分析质量。表 2 中展示了我们所定义的 12 个领域以及这 12 个领域下的总计 46 个细粒度标签。

在语料中,每个领域分别随机抽样 100 条语料,并计算出准确率(标注正确实体/所有实体)。抽样的结果如表 4 所示。

表 4 回标准确率抽样(%)

领域	准确率	领域	准确率
传统	98.91	旅游	96.67
教育	90.24	餐饮	93.07
游戏	86.96	房产	83.01
娱乐	70.05	电商	62.50
金融	82.79	体育	75.00
汽车	91.35	医疗	65.15

2.3 词向量

在模型的训练过程中,使用了预训练的中文词

向量,使用 Python 库 gensim 中的 Word2Vec 的接口训练中文维基百科的语料得到。

通过对训练集中的词进行统计。对于词表中的词 w ,如果有对应的预训练词向量,则使用对应的预训练的词向量作为词 w 的词向量的初始值;如果没有对应的预训练的词向量,则词 w 的词向量将随机产生。在模型的训练过程中,词向量会随着训练时得到的梯度进行微调。

2.4 实验结果

2.4.1 实体识别模块实验结果

此模块利用 12 个领域下的语料分别训练 12 个命名实体识别器,这些命名实体识别器仅进行 0-1 识别,即确定某词条是否为某领域下的命名实体。因此,本节给出了在各个领域下命名实体的 0-1 识别结果,并对比了与 LSTM+CRF 模型的结果,效

果有所提高。

实验结果如表 5 所示。由表 5 可见,加入 CNN 后,在大部分领域的命名实体识别效果均有所提高,说明 CNN 能够从文本中获得对命名实体识别有益的特征。

2.4.2 命名实体的细粒度类别划分模块实验结果

命名实体的细粒度类别划分模块分为两个阶段,因此实验部分也给出了两个阶段的划分结果。第一个阶段的输入为由命名实体识别模块识别出的命名实体及其上下文,输出为该命名实体的大类别标签,这个阶段记为领域划分阶段。在进行此部分实验时,假设输入的命名实体为正确的命名实体。此部分实验能够检测命名实体被划分到了正确的大类别下的准确程度。本部分的实验数据均为词典回标得到的数据,按照 7 : 1 : 1 的比例随机划分为训练集、开发集和测试集。

表 5 命名实体识别模块实验结果(%)

领域	准确率		召回率		F ₁ 值	
	无 CNN	加入 CNN	无 CNN	加入 CNN	无 CNN	加入 CNN
传统	78.35	79.92	87.21	86.62	82.54	83.13
教育	91.21	91.90	89.03	91.01	90.10	91.45
游戏	95.03	97.23	95.33	96.72	95.17	96.98
娱乐	91.32	90.87	89.60	93.67	90.45	92.25
金融	94.78	94.22	92.23	94.59	93.48	94.40
汽车	92.34	93.15	95.23	96.62	93.76	94.86
旅游	89.98	89.67	92.54	94.49	91.24	92.01
餐饮	86.45	89.68	94.22	93.77	90.16	91.68
房产	78.30	81.75	88.93	87.10	83.27	84.34
电商	87.32	91.60	91.47	94.52	89.34	93.04
体育	79.21	80.30	78.17	77.28	78.68	78.75
医疗	92.11	93.86	92.40	91.16	92.25	92.49

本模块只对命名实体识别模块识别正确的实体进行细粒度类别划分,并不对实体的识别结果进行修改,因此本模块只对命名实体类别划分的准确率进行分析。领域划分阶段的实验结果,如表 6 所示。

表 6 领域划分阶段的实验结果(%)

训练集	开发集	测试集
97.93	95.22	95.32

由表 6 可见,领域划分的准确率在 95% 以上,说明在命名实体识别准确的情况下,依据上下文可

以获得较为准确的领域划分结果。因此我们首先对命名实体的领域进行确定,再对命名实体的细粒度标签进行划分。

细粒度类别划分模块的第二个阶段,对命名实体在第一个阶段识别出的领域下进行细粒度划分,记为类别划分阶段。本阶段同样假设输入为由命名实体识别模块识别出的命名实体及其上下文,输出则为该命名实体的细粒度类别标签。此部分实验能够检测命名实体被划分到细粒度的类别下的准确程度。表 7 给出了实验结果。

表 7 类别划分阶段的实验结果(%)

领域	训练集	开发集	测试集
金融	96.52	93.13	92.01
医疗	74.30	87.95	85.63
餐饮	65.27	73.29	75.87
教育	76.30	87.68	85.75
体育	95.15	98.58	98.82
旅游	86.96	87.70	88.16
传统	92.59	95.78	94.93
汽车	98.02	95.04	93.12
房产	86.45	93.71	93.62
游戏	79.31	88.68	86.06
电商	96.11	85.21	80.77
娱乐	64.49	69.71	72.05

表 7 中,医疗、娱乐、游戏等领域实验中训练集的效果不如开发集和测试集的原因是我们的训练数据中有回标误差。

2.4.3 命名实体识别整体实验结果

将命名实体识别和命名实体的细粒度类别划分两个步骤合二为一,测试了本文实现的命名实体细粒度划分的整体准确率,使用的语料是人工标注的语料,结果如表 8 所示。由于在细粒度类别划分模块中只对命名实体识别模型识别正确的实体进行细粒度类别划分,因此最后结果的准确率等于两个模块准确率的乘积,召回率等于命名实体识别模块的召回率。在此实验中,输入为未经处理的原始文本,而输出为带有细粒度标签的命名实体。

表 8 细粒度类别划分模块的实验结果(%)

领域	准确率 R/%	召回率 R/%	F_1 /%
金融	82.28	94.59	88.01
医疗	61.60	91.16	73.52
餐饮	58.48	93.77	72.04
教育	69.95	91.01	79.10
体育	80.70	94.49	87.05
旅游	87.65	96.62	91.92
传统	86.77	96.62	91.48
汽车	79.62	77.28	78.44
房产	87.12	87.10	79.30
游戏	83.19	94.52	88.49
电商	89.70	93.67	91.65
娱乐	61.66	86.62	72.04

从表 8 的实验结果可见,大部分领域的效果都

在 80%左右,小部分领域效果较差,这些领域效果较差的主要原因是数据在标注的过程中出现了错误,而在人工标注的语料中则不包含这些错误的信息,因而模型学到的更多的是错误的信息。

另外,由于训练语料是由词典回标产生的,所以模型往往可以正确识别常见的命名实体,而对于一些出现次数较少的命名实体就难以正确识别。部分实体的类别在自动标注的过程中被误标,原因是数据回标时将部分非命名实体标注为命名实体。例如,对于文本“losea 磨砂拼接小方包 2017 新款斜挎包迷你锁扣包韩版复古单肩包小”中的“迷你”,在汽车领域的标签是“汽车—品牌”,但是在当前语境下就是错误的标签。在该领域下实验效果较差的原因就是在数据回标时,将大量的“迷你”标注为“汽车—品牌”,所以在训练的时候,模型学习到的信息是将“迷你”的类别识别为“汽车—品牌”。

2.5 实验结果分析

部分抽样的错误样例见表 9。其中,由“【”和“】”括起来的为模型识别出的命名实体。

从表 9 中可以看出,错误样例中的错误是多样的,既有命名实体识别模块识别的错误,也有细粒度类别划分的错误。同时,来源于构建语料时词典回标造成的错误也是难以正确识别的。

表 9 错误样例示例

序号	例子	识别结果	错误原因
1	【渗出】的血珠,就像一腔热血无处挥洒而倾泻的情感。	渗出	渗出并不是一个命名实体
2	【长安 CS75】车型挡泥板汽车配件用品外饰改装专用挡泥胶带标	长安 CS75	细粒度类别划分错误,将“汽车—车型”识别为“汽车—车系”
3	【模式识别】机械工业出版社	模式识别	词典回标错误,将“教育—书籍名”识别为“教育—课程名”

3 总结

本文设计并实现的分阶段细粒度命名实体识别方案能够将大部分命名实体识别出来,并确定该命名实体的细粒度标签。对于某些细粒度的命名实体,比如药品、影视名、汽车车系等,由于在训练语料中分布稀疏,其中一些类别的命名实体的命名非常随意,比如车系有“唐”“A7”等,使得针对这类的命

名实体的细粒度划分变得不准确,但是如果不强调小类别,而是仅进行大类别的划分,本文实现的方法 F_1 值在全领域上平均值达到了 80% 左右,在一定程度上说明本论文实现的分阶段方案是有效的。

参考文献

- [1] Fine S, Singer Y, Tishby N. The hierarchical hidden Markov model: Analysis and applications[J]. Machine learning, 1998, 32(1): 41-62.
- [2] Borthwick A, Grishman R. A maximum entropy approach to named entity recognition[D]. New York University, Graduate School of Arts and Science, 1999.
- [3] McCallum A, Freitag D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of the 17th ICML, 2000: 591-598.
- [4] Lafferty J, Mc Callum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th ICML, 2001: 282-289.
- [5] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996: 133-142.
- [6] Collobert R, Weston J, Bootul L. Natural language processing(almost) from Scratch[J]. arXiv preprint arXiv: 1103.0398, 2011.
- [7] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 3111-3119.
- [8] Gers F A, Schraudolph N N, Schmidhuber J. Learning precise timing with LSTM recurrent networks[J]. Journal of Machine Learning Research, 2002, 3(Aug): 115-143.
- [9] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015:1-10.
- [10] Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data[J]. The Journal of Machine Learning Research, 2005(6): 1817-1853.
- [11] Jing H Y, Zhang T. Named entity recognition through classifier combination [C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL-2003, 2003: 168-171.
- [12] Kim Yoon. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1746-1751.
- [13] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1510.03820, 2015.
- [14] M Boden. A guide to recurrent neural networks and back-propagation[R]. SICS Technical Report J 2002: 03, SICS.
- [15] Hammerton J. Named entity recognition with Long short-term memory[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 172-175.
- [16] Tang D Qin B, Feng X. Effective LSTMs for target-dependent sentiment classification[C]//Proceedings of the COLING 2016, 2016: 3298-3307.
- [17] Yin Q, Zhang Y, Zhang W. Chinese zero pronoun resolution with deep memory network[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1309-1318.



盛剑(1992—),通信作者,硕士研究生,主要研究领域为自然语言处理、信息抽取。
E-mail: jsheng@ir.hit.edu.cn



秦兵(1968—),博士,教授,主要研究领域为自然语言处理、情感分析、信息抽取、篇章语义分析。
E-mail: bqin@ir.hit.edu.cn



向政鹏(1994—),硕士研究生,主要研究领域为自然语言处理、阅读理解、信息抽取。
E-mail: zpxiang@ir.hit.edu.cn