基于深度学习的地址信息自动标注研究

凌广明[1],徐爱萍[1],王 伟[2]
(1. 武汉大学计算机学院,湖北武汉 430072; 2. 武汉大学测绘遥感信息工程国家重点实验室,湖北武汉 430079)

摘 要: 文本序列的自动标注能够解决深度学习普遍面临的人工标注成本过高的问题. 本文针对地址信息的实体表述特征,构建基于实体边界矩阵(Entity Boundary Matrix,EBM)的表示模型,在此基础上提出了一种基于深度学习和 KNN 标签修正算法(K-Nearest Neighbours Correction Algorithm,KNN-CA)的不需要任何人工标注训练集的自动标注算法. 首先获取预置小区数据集并构建离线特征库和初始化在线特征库;接着通过匹配算法求解 EBM 并利用 KNN-CA 进行优化,再通过数据增广得到自动标注的训练集;然后训练 BiLSTM-CRF 深度学习模型并预测所有未曾标注的地址信息的序列标注;最后再次利用 KNN-CA 优化可求解 EBM 的序列标注,由此构建适用于中文地理命名实体(Chinese Geospatial Named Entities,CGSNE)识别及相关研究的序列标注语料库. 实验表明,标注数据的 *F*1 值达到了 95.35%.

关键词: 深度学习; 自动标注; 地址信息; K 近邻; 语料库
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2081-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.001

Research of Address Information Automatic Annotation Based on Deep Learning

LING Guang-ming[1], XU Ai-ping[1], WANG Wei[2]
(1. *School of Computer Science, Wuhan University, Wuhan, Hubei* 430072, *China*;
2. *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, Hubei* 430079, *China*)

Abstract: Automatic annotation of text sequence can address the common issue of high manual annotation labor cost in deep learning. In this paper, a representation model based on the entity boundary matrix (EBM) is constructed. On the basis, we propose an automatic annotation algorithm combining deep learning with KNN annotation correction algorithm (KNN-CA) where the manual labeling training set is not required. Firstly, the offline feature library and online feature library is built and initialized respectively with the utilization of collecting estate dataset. In addition, EBM is solved by matching algorithm and optimized via KNN-CA technique. After the data augmentation process, a training dataset of automatic annotation is obtained. Then the BiLSTM-CRF deep learning model is trained and all unlabeled annotation sequence is predicted. Eventually, the annotation sequence of solvable EBM is optimized via KNN-CA again so as to construct a sequence annotatied corpus dataset which is suitable for the identification of Chinese Geospatial Named Entities (CGSNE) and related researches. The experiment demonstrates that *F*1 score of labeled data reaches 95.35%.

Key words: deep learning; automatic annotation; address information; KNN(K-Nearest Neighbours); corpus

1 引言

对地址信息进行精准而快捷的分析,识别其中姓名、电话、位置、小区名称和小区详情在社会安全、商业管理等领域起着越来越重要的作用. 地理命名实体(Geospatial Named Entities,GSNE)是具有地理属性的命名实体[1],识别 GSNE 是分析地址信息的重要方法. 针对中文地理命名实体(CGSNE)的特征,沈达阳等人[2]通过统计地名用字在地名位置的概率分布构建了地名规则库;郑家恒等人[3]构建地名字典并结合下上文信息规则,采用了基于变换的错误驱动的机器学习方法有效地提升了地名识别效果;张雪英等人[4]分析了中

收稿日期:2019-12-30;修回日期:2020-06-04;责任编辑:马兰英
基金项目:国家重点研发计划资助(No. 2017YFC0803700)

文文本和地理实体描述和表达机制的差异,结合 GSNE 语言特点,制定了中文文本的 GSNE 标注体系和标注规范;官登水[5]提出了基于拆分策略的 GSNE 的识别方法.这些研究成果有效地解决了 CGSNE 问题,构建了规则特征库和标注资源,具有重要的研究和应用价值,但是却需要大量的人工特征和专业领域知识,存在不同程度的"特征工程"问题.

随着深度学习的发展,Graves A 等人[6]将正反两个方向的 LSTM[7]模型结合起来,提出 BiLSTM 模型,有效地使用过去和未来的输入信息;Lample G 等人[8]进一步将 CRF 结合进来使得该模型能够使用句子级标签信息,形成 BiLSTM-CRF 模型;Huang Z H 等人[9]和 Dong C H 等人[10]提出并有力推动了 BiLSTM-CRF 深度学习模型在中文命名实体识别(Chinese Named Entity Recognition,CNER)任务中的广泛应用,同时还具有对字向量依赖性较小的健壮性;Shen Y Y 等人[11]证明了 LSTM 解码器优于 CRF 解码器,并且当实体类型的数量较大时,LSTM 解码器的训练速度更快.利用字向量 BiLSTM-CRF 模型不需要人工构建特征,可以有效地解决"特征工程"问题,但是需要大量的人工标注数据,不仅效率低,而且人力成本昂贵.

人工标注成本在一定程度上制约了深度学习的发展,因此自动标注算法的研究日益重要.图像自动标注技术起步较早[12~15];在文本领域,徐飞等人[16]通过大量实验证实了在未加入任何人工特征的情况下,BiLSTM-CRF 模型比 CRF、RNN 和 BiLSTM 等模型对词性自动标注具有显著优势;王姬卜等人[17]提出基于回标技术的地理实体关系语料库构建方法以实现语料的自动构建;朱珠等人[18]对样本置信度的评估做了较为系统和深入的研究,提出了结合主动学习和自动标注的评价对象抽取方法;Schulz C 等人[19]对话语级序列标注任务进行深入研究,提出自动为领域专家给出注释建议的基于 BiLSTM-CRF 的建议模型,对注释的速度和性能有积极的影响,同时不会引入明显的偏差.这些研究在不同任务中对自动标注进行了积极有效的探索,为解决深度学习的技术瓶颈提供重要参考.

自动标注文本地理信息能够有效解决用深度学习处理 GSNE 时面临的人工标注成本过高的问题.本文结合 GSNE 的优秀研究成果,针对相关语料匮乏和人工标注成本昂贵两个问题,通过分析地址信息中地理实体的语言特点,提出基于实体边界矩阵(***EBM***)的表示模型;利用实体边界特征构建词级实体边界特征库,并结合 KNN 提出了 KNN-CA 算法;在自动标注训练集和优化预测序列标注两个阶段将词的上下文信息融入到基于字向量的 BiLSTM-CRF 深度学习模型中,实现了无需人工标注训练集的自动标注.本文主要贡献在于:(1)提出 ***EBM*** 以形式化数据特征并定义相关操作,构建 ***EBM*** 数据表示模型;(2)利用实体表述特征构建轻量级的实体边界特征库,并结合实体置信度实现了基于 KNN 的 KNN-CA 算法,实现了标签自动修正;(3)基于深度学习实现了地址信息中姓名、电话、位置、小区名称和小区详情五种实体类型无需任何人工标注训练集的自动标注,构建了地址信息语料库.

2 数据特征分析

地址信息形如"李@@ ~15 ******* 60 ~ ~ ~桐柏南路帝湖花园 *** 东 * 单元"(@ 和 * 是为了保护用户隐私而引入的屏蔽字符,其中@ 代表一个汉字,* 代表一个数字,下同),所有数据均由统一的管理平台导出,因此具有统一的结构.引入 PER、TEL、LOC、EST 和 INF 分别表示姓名、电话、位置、小区名称和小区详情,则有"李@@"→PER、"15 ***** 60"→TEL、"桐柏南路"→LOC、"帝湖花园"→EST、" *** 东 * 单元"→INF("→"表示属于某一种类型).LOC、EST 和 INF 共同构成了完整的地址信息,记为 ADD;~ 作为分隔符,将 PER、TEL 和 ADD 分隔并依次排列.由于地址信息只与 LOC、EST 和 INF 有关,而且 PER 和 TEL 通过规则匹配的方法即可完成高质量的自动标注,因此本节的讨论仅限于 LOC、EST 和 INF 三类实体.

为了分析 PER、TEL、LOC、EST 和 INF 五种实体类型的分布情况,对经过预处理的 6751 条数据(详见 4.1 节)进行系统抽样,共进行 10 次不放回抽样,每次样本容量为 20.按照 LOC、EST 和 INF 三类实体出现的次数以及次序,分为 16 种情形进行统计,表 1 显示了第一次抽样样本的实体分布情况.在 10 次抽样中,infNAEst、locMiddle、ests、loc2、locs、inf2 和 infs 等 7 种情形都未曾出现,其他 9 种分布情况如图 1 所示.

图 1 中的(a)和(b)分别显示了数据量较多和较少的类型分布.在分析样本整体分布的同时,也发现了一些重要的局部特征:(1)**实体连续性** LOC、EST 和 INF 三种实体是无缝衔接的.6751 条用户数据中,只有 84 条数据含有中英文逗号、括号等分隔符,而且这些分隔符对实体连续性影响甚微;(2)**实体有序性** LOC、EST 和 INF 的位序比较固定,INF 基本上都在 EST 之后,而 LOC 一般在 EST 之前;(3)**实体有界性** 中文地名具有天然的结构和规律[2,3,20,5],LOC、EST 和 INF 具有较为明显的边界特征,这与中文语言现象有关;其中 LOC 和 EST 具有明显的尾部特征,INF 具有明显的头部特征.LOC 大多以"路、街、交叉口"等词汇结尾,EST 多以"小区、社区、家属院"等词汇结尾,而 INF 多以阿拉伯数字或者英文字母开头,如"19 号楼、C 座"等;(4)**易变性** 随着社会的飞速发展,地理信息随之发生日新月异的变

表 1　第一次抽样的 20 条地址信息实体分布表

分布类型	数量	百分比	说明	示例
normal	13	65	正常蕴含	赵@ @ ~18 ******* 75 ~ ~ ~长江路亚星悦都 * 号楼 * 单元 ****
infNAEst	0	0	INF 不紧随 EST	
locEnd	0	0	位置在最后	
locMiddle	0	0	位置在中间	
est0	1	5	不含小区	赵@ ~15 ******* 33 ~ ~ ~郑州市文劳路 * 号院 * 号楼 * 单元 * 楼 ***
est1	19	95	蕴含 1 个小区	刘@ @ ~13 ******* 88 ~ ~电务小区 ** # -*-***
est2	0	0	蕴含 2 个小区	
ests	0	0	蕴含多个小区	
loc0	5	25	不含位置	常@ @ ~13 ******* 51 ~ ~大唐明村 * 号楼 * 单元 ** 中
loc1	15	75	蕴含 1 个位置	范@ @ ~15 ******* 06 ~ ~ ~周口市新区碧桂园小区 ** 栋 * 单元 *** 室
loc2	0	0	蕴含 2 个位置	
locs	0	0	蕴含多个位置	
inf0	2	10	不含小区详情	郭@ ~13 ******* 63 ~ ~ ~青屏大街大鸿路西 *** 米意利宝橱柜
inf1	18	90	蕴含 1 个小区详情	王@ @ ~15 ******* 81 ~ ~烟草局后小院第 * 家
inf2	0	0	蕴含 2 个小区详情	
infs	0	0	蕴含多个小区详情	

正常蕴含是指完整包含 PER、TEL、LOC、EST 和 INF,且 PER、LOC、EST 和 INF 个数都为 1,而且依次排列

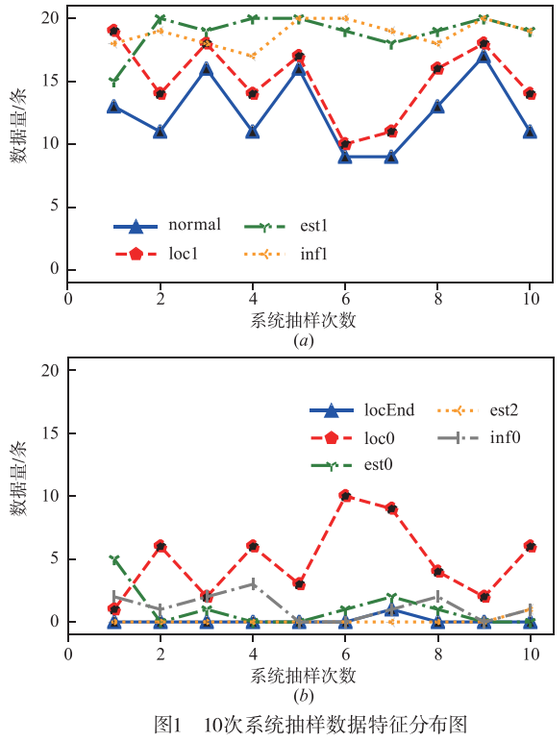

图1　10次系统抽样数据特征分布图

化,位置、小区甚至行政区划分都会发生意想不到的变化[21];(5)**缺乏规范性和完整性**　中文地址信息普遍存在结构混乱、信息错误、信息不完整、信息丢失以及信息冗余等问题[22],比如"绿云小区平湖里 *-*-* 西户"缺失位置信息、"中原西路欧凯龙"误写为"中原西欧欧凯龙".

其中,实体的连续性、有序性和有界性是自动标注算法的重要前提,下面将详细讨论;而易变性、缺乏规范性和完整性是采用深度学习模型的重要原因.

3　基于 *EBM* 的数据表示模型

3.1　实体边界特征库

本文利用实体边界词(即开头和结尾)的概率分布构建以词为基本单位的实体边界特征库[2],为了提升系统性能,以字典(数据结构)组织和管理数据. 实体边界特征库按照 LOC_ES,EST_ES,INF_BS(BS 和 ES 后缀分别表示开始和结尾)分别存储,并根据运行时状态分为离线特征库和在线特征库. 其主键为尾部(头部)词,值为实体置信度或者尾部(头部)词以对应类型出现的词频.

3.1.1　离线特征库

将最显著和需要特殊处理的特征库进行固化存储,以提升系统的性能,该特征库在运行时不发生任何变化,故称为离线特征库(Offline library,OffLib). 根据专业知识库[1,3-5]和实验分析,本文通过边界词构建了如表 2 所示的轻量级 OffLib.

表 2 离线特征库(OffLib)

类型	数量	主键	值	说明
LOC_ES	75	郑州,市,县,镇,乡,村,区,*号,*号院,号院……	0.99	LOC 的尾部特征
EST_ES	5	*期,新村,社区,城市,都市	0.99	EST 的尾部特征
INF_BS	18	*号院,*号,*座,*区,*号楼,*楼,西门,东门……	0.99	INF 的头部特征

其中 * 为通配符,代表了阿拉伯数字和英文字母,离线特征库的置信度为 0.99,LOC_ES 和 INF_BS 仅显示部分内容

3.1.2 在线特征库

当实体的置信度达到阈值 ϖ 后,将实体的边界特征信息加入特征库. 该过程持续发生在整个标注过程,因此称为在线特征库(Online library,OnLib). 第 5.4 节的实验表明,选取适当的 ϖ 能够扩充 OnLib,从而提升自动标注精度,该过程记为 $\mathrm{OnLib}\leftarrow\Lambda(\boldsymbol{EBM},u,\varpi)$.

与 OffLib 不同,OnLib 的值不是置信度,而是对应主键出现的次数,这样便于统计. 其置信度由式(1)给出:

$$\mathrm{OnLibCon}_i=\begin{cases}\dfrac{\log(\mathrm{OnLib}[k_i]+1)}{\log(\max(\mathrm{OnLib})+1)}\times(1-\eta)+\eta, & (k_i\in\mathrm{OnLib})\\ \delta, & (\mathrm{else})\end{cases}\tag{1}$$

其中,$\mathrm{OnLib}[k_i]$是实体的尾部(头部)词以对应类型出现的次数,$\max(\mathrm{OnLib})$是 OnLib 中出现最多的次数. 如果 k_i 在特征库中,则置信度映射到区间$[\eta,1]$,其中 η 是 OnLib 中置信度的下限,本文通过实验确定为 0.9. 需要指出的是,统计 estates(详见 4.1 节)中结尾词频,可以有效初始化 EST_ES 特征库.

3.2 实体边界矩阵(*EBM*)

官登水[5]利用式(2)给出地理位置实体的形式化表达:

$$\mathrm{GLNE}=\{\mathrm{TG}_m+\mathrm{BG}_n\}+\cdots+\{\mathrm{TG}_p+\mathrm{BG}_q\}\tag{2}$$

其中,$m\in N^+$,$(n,p,q)\in N$,TG 是传统的地理位置实体,比如“通泰路”“执法局”等;BG 是基本的地理位置实体,主要指的是 TG 的扩展和一些通名、饰名和连接词,比如“大厦”“小区”“与”“交叉口”等. 本文在此基础上并结合第 1 节中分析的实体连续性、有序性和有界性,提出基于实体边界矩阵(***EBM***)的地址信息形式化表示模型. 对于一条包含 n 个字(中文、英文、数字等字符)的地址信息,即 $u=(c_1,c_2,\cdots,c_n)$,令$\mathbb{T}=\{1,2,3\}$表示实体类型集合,分别对应 LOC,EST,INF;t 表示实体类型,则有 $t\in\mathbb{T}$;相应地用 α,ω 分别表示实体在 u 中开始和结束的下标;l 表示实体长度;c 表示实体的置信度,其定义由式(3)给出:

$$c_i=\max(\mathrm{OnLibCon}_i,\mathrm{OffLibCon}_i)\tag{3}$$

特别地,如果实体不在实体边界特征库中,则其置信度很低,记为 δ,本文取值 0.0002,因此,$c_i\in[0.0002,1]$.

所以可以用列向量 $\boldsymbol{g}_i=(\alpha_i,\omega_i,l_i,t_i,c_i)^{\mathrm{T}}$ 表示一个实体信息,由此得到实体边界矩阵 ***EBM*** 的定义,如式(4)所示:

$$\boldsymbol{EBM}=[g_1,g_2,\cdots,g_N]=\begin{bmatrix}\alpha_1&\alpha_2&\cdots&\alpha_N\\ \omega_1&\omega_2&\cdots&\omega_N\\ l_1&l_2&\cdots&l_N\\ t_1&t_2&\cdots&t_N\\ c_1&c_2&\cdots&c_N\end{bmatrix}\tag{4}$$

其中,$t_i\in\mathbb{T}(i=1,2,\cdots,N)$,$N$ 为实体个数,由 g_i 的定义可知,***EBM*** 为 $5\times N$ 阶矩阵,令 e_i 表示 g_i 对应的实体,则 e_i 由式(5)唯一确定:

$$e_i=\mathrm{substring}(u,\alpha_i,\omega_i)\tag{5}$$

其中 substring 是获取 u 中从 α_i 到 ω_i 的子字符串函数. 特别地,如果某一种实体不存在,其对应的列向量为 $\boldsymbol{g}_i=(0,-1,0,t_i,0)^{\mathrm{T}}$,简记为 $\boldsymbol{g}_i=\phi$,此时 e_i 为空串,为了方便获取实体整体信息以及优化 ***EBM*** 相关的操作,在 ***EBM*** 中仍然保留其信息;如果是一般情况,即全部实体都存在,则简记为 ϕ_0. 由式(2)和表 1 可知 $N\geqslant 3$ 且一般情况下为 3. 以表 1 中 normal 的示例“赵@@ ~18 ******* 75 ~ ~ ~ 长江路亚星悦都 * 号楼 * 单元 **** ”为例做一个具体的说明:LOC、EST 和 INF 分别为“长江路”、“亚星悦都”和“ * 号楼 * 单元 **** ”,且“长”、“亚”和“ * 号楼”中的“ * ”所在的位置分别为 19、22 和 26(位置是从 1 开始的). 因此得到如式(6)所示的 ***EBM***:

$$\boldsymbol{EBM}_{\mathrm{zhao}}=\begin{bmatrix}19&22&26\\ 21&25&35\\ 3&4&10\\ 1&2&3\\ 0.99&\delta&0.99\end{bmatrix}\tag{6}$$

其中,“长江路”以“路”结尾而“ * 号楼 * 单元 ***** ”以“ * 号楼”开头,根据 OffLib 其置信度取值 0.99. “亚星悦都”的置信度无法通过 OffLib 获取,需要由 OnLib 来确定,如果也不在 OnLib 中,则取值 δ.

3.3 *EBM* 基本运算定义

3.3.1 求解原始 *EBM*

由于 ADD 的核心是 EST,同时,获取 estates 相对比较容易. 而且实体具有连续性(详见第 2 节),如式(7)所示:

$$\begin{cases} L(\mathrm{ADD}) = \sum_{i=1}^{N} l_i, & (l_i \geqslant 0) \\ \alpha_i \leqslant \alpha_{i+1}, & (i = 1,2,\cdots,N-1) \\ \omega_i + 1 = \alpha_{i+1}, & (i = 1,2,\cdots,N-1) \end{cases} \tag{7}$$

由式(7)可知,通过匹配算法可以方便地获取 g_2,即 EST 对应的列向量,然后由式(4)和式(7)求解 ***EBM***. 但是需要满足以下三个条件:(1)有且仅有一个 estates 中的小区;(2)ϕ_0 且 $N=3$,即三类实体都存在且每类只有一个实体;(3)保证式(8)能够成立,即满足有序性.

在 ***EBM*** 中,N 和类型并不确定,因此实体的排列比较复杂. 然而由表 1 和图 1 可以发现,同类型实体几乎没有出现多次的情况,而且 LOC 基本上出现在最前面(对于出现在最后面的少数情况,本文将通过数据增广加以解决,详见第 4.3 节),所以实体是有序的. 如式(8)所示:

$$t_i \leqslant t_{i+1} (i = M, M+1, \cdots, N-1 \quad \text{且 } g_i \neq \phi) \tag{8}$$

其中 M 由式(9)给出:

$$M = \begin{cases} 1 & (\phi_0) \\ \arg\max\limits_{i \in [1,N]} (g_i = \phi) & (\text{else}) \end{cases} \tag{9}$$

其中第一个条件 ϕ_0 表示三种实体全部出现;第二个条件为至少有一个实体缺失,此时由式(7)可知,所有缺失的实体都对应 ***EBM*** 中最左边的列向量.

记 $\boldsymbol{EBM}_0$ 为原始 ***EBM***,利用 estates 通过匹配算法求解 $\boldsymbol{EBM}_0$ 的过程记为 $\Gamma(u \mid \text{estates}) \Rightarrow \boldsymbol{EBM}_0$,如果不满足条件而无法求解时有 $\boldsymbol{EBM}_0 = \phi$.

3.3.2　合并近邻同类实体

考虑 $\boldsymbol{g}_i$、$\boldsymbol{g}_{i+1}$ 和 $\boldsymbol{g}_{i+2}$ 三个实体,其中 $t_i = t_{i+2}$,$t_{i+1} \neq t_i$,而且 $l_{i+1} \leqslant 2$,则修正 t_{i+1} 为 t_i. 特别地,对于任意 $t_i = t_{i+1}$ 即相邻的同类实体将全部合并为一个实体,该运算过程记为 $\Xi(\boldsymbol{EBM})$.

3.3.3　词位移

令 $\boldsymbol{g}_i$ 对应的实体 $e_i = (w_1, w_2, \cdots, w_n)$,其中 w_i 为 e_i 的一个分词,令 $\overleftarrow{\Psi_i}(\boldsymbol{EBM} \mid u)$ 表示将 w_1 取出并追加到 g_{i-1} 对应的实体 e_{i-1} 的尾部,同样地,$\overrightarrow{\Psi_i}(\boldsymbol{EBM} \mid u)$ 表示将 w_n 取出并追加到 $\boldsymbol{g}_{i+1}$ 对应的实体 e_{i+1} 的头部. $\overleftarrow{\Psi_i}(\boldsymbol{EBM} \mid u)$ 和 $\overrightarrow{\Psi_i}(\boldsymbol{EBM} \mid u)$ 统称为词位移(因为移动具有方向性)运算. 用 $\tilde{N}$ 表示改变实体的存在性,即发生以下两种情况之一:(1)$\boldsymbol{g}_i$ 只包含一个词,经过词位移运算后,$\boldsymbol{g}_i = \phi$;(2)$i = M$(详见式(9))或者 $i = N$,此时将产生新的实体,其类型由 KNN-CA 算法根据实体边界特征库和 $\boldsymbol{g}_i$ 决定. 特别地,词位移是以 EST 为中心的,因此在相同条件下,优先向 EST 的外侧移动.

3.3.4　由序列标注生成 *EBM*

如果有完整的序列标注 $y = (y_1, y_2, \cdots, y_n)$,就唯一确定了一个对应的 ***EBM***(如果存在的话). 首先获取所蕴含的实体,然后按照自然顺序依次得到所有的 $\boldsymbol{g}_i$(包含不存在的实体类型),按照 α_i 升序排列全部 $\boldsymbol{g}_i$ 并组成矩阵 $\boldsymbol{M}$. 如果 $\boldsymbol{M}$ 满足式(7)和式(8)则通过 $\Xi(\boldsymbol{M})$ 运算得到最终的 ***EBM***,该运算记为 $\Upsilon(y) \Rightarrow \boldsymbol{EBM}$. 如果无法生成 ***EBM***,记为 $\boldsymbol{EBM} = \phi$.

3.3.5　求解 *EBM* 整体置信度

EBM 中每一个列向量都有一个置信度 c_i,式(10)给出了 ***EBM*** 整体置信度的定义:

$$Con_{\boldsymbol{EBM}} = \frac{\sum_{i=M}^{N} c_i}{N - M + 1} + \partial \tag{10}$$

其中 M 和 c_i 分别由式(9)和式(3)给出;∂ 是 EST 和 INF 的联合置信度增益,即如果相邻的两个实体同时达到了 EST 和 INF 的置信度阈值 ϖ,则增大 $Con_{\boldsymbol{EBM}}$,具体定义见式(11):

$$\partial = \begin{cases} \lambda_2 * c_i + \lambda_3 * c_{i+1}, & t_i = 2, t_{i+1} = 3, \min(c_i, c_{i+1}) \geqslant \varpi \\ 0, & \text{else} \end{cases} \tag{11}$$

其中,λ_2 和 λ_3 分别为 EST 和 INF 的权重,本文取 0.6 和 0.4.

3.3.6　由 *EBM* 生成序列标注

对于给定的 u 和对应的 ***EBM***,就可以唯一确定其序列标注 $y = (y_1, y_2, \cdots, y_n)$,其过程实际上是 $\Upsilon(y) \Rightarrow \boldsymbol{EBM}$ 的逆运算,记为 $\overline{\Upsilon}(\boldsymbol{EBM}) \Rightarrow y$.

4　自动标注算法

4.1　数据预处理

自动标注算法(详见 4.5 节)需要 users、estates、OffLib、Tests(测试集)和 Devs(开发集),其中 OffLib 已经在第 3.1.1 节中讨论过了,本节讨论其他数据的获取过程. 本文数据全部是来自商家的真实数据. 随机选取两个月份的数据并直接由管理平台导出,过滤重复数据后得到 6751 条数据,记为 users;由于绝大部分地址信息是郑州市的,所以通过网络爬虫获取 5386 个郑州市小区,清洗包含非法字符的 138 条数据,并过滤名称长度不足 3 和超过 15 的小区,得到 5153 条数据,即为 estates.

BiLSTM-CRF 深度学习模型不仅需要标注良好的测试集和开发集,还需要保证数据的多样性以有效地避免过拟合问题[23]. 为此,在系统抽样的基础上,附加了小区检测功能以确保覆盖尽可能多的小区,而且尽可能多地包含表 1 中的分布情形,由此得到各包含 300 条数据的 Tests 和 Devs.

4.2　KNN-CA 算法

He G L 等人[24]利用 KNN 算法自动标注数据,其核

心思想是邻近标签具有相关性. 本文在此基础上,基于 ***EBM*** 实现了 KNN-CA 算法,该算法以 ***EBM*** 的词位移运算为基本操作,利用近邻标签的相关性对 ***EBM*** 进行优化,记为$\overline{\boldsymbol{EBM}}$←KNN-CA(***EBM***),算法 1 描述了具体实现.

算法 1 KNN-CA 算法

输入:estates,OffLib,OnLib,u,***EBM***,K
输出:$\overline{\boldsymbol{EBM}}$
1: redo←True
2: while redo = True do
3: redo←False
4: for each $i \in [M,N]$ do
5: $\boldsymbol{EBM}' \leftarrow \boldsymbol{EBM}$
6: if t_i! = 1 then
7: 连续进行 K 次$\underleftarrow{\Psi_i}(\boldsymbol{EBM}' \mid u)$,$\overline{\boldsymbol{EBM}}$←整体置信度最高的 ***EBM***′
8: if $\widetilde{N}$ then
9: redo←True, $\Xi(\overline{\boldsymbol{EBM}})$,break
10: end if
11: end if
12: if t_i! =3 then
13: 连续进行 K 次 $\Psi_i(\overline{\boldsymbol{EBM}}' \mid u)$,EBM←整体置信度最高的 ***EBM***
14: if $\widetilde{N}$ then
15: redo←True, $\Xi(\overline{\boldsymbol{EBM}})$,break
16: end if
17: end if
18: end for
19: end while
20: OnLib← $\Lambda(\overline{\boldsymbol{EBM}},u,\varpi)$

当 ***EBM*** 中的实体数量发生改变即 $\widetilde{N}$ 时,由于当前的下标失效,整体的 $\boldsymbol{g}_i$ 都将发生变化,因此需要合并近邻同类实体,并对新的 ***EBM*** 重新进行求解,但先前的 $\overline{\boldsymbol{EBM}}$ 仍然有效. 因为 $\overline{\boldsymbol{EBM}}$ 总能够保证是置信度最大的 ***EBM***,在算法最后利用它扩充 OnLib.

4.3 自动标注训练集

根据 users、estates 和 OffLib,通过 $\Gamma(u \mid \text{estates}) \Rightarrow \boldsymbol{EBM}_0$ 运算和 KNN-CA 算法自动构建训练集,同时初始化 OnLib. 主要流程如图 2 所示.

图 2 中,“+”号表示信息融合,$\overline{\boldsymbol{EBM}}$ 是 KNN-CA 算法优化后的 ***EBM***,最后由 $\overline{Y}(\overline{\boldsymbol{EBM}}) \Rightarrow y$ 运算求解 u 的序列标注. 由于 ***EBM*** 的有序性,y 中如果存在 LOC,则一定在最前面. 对于 LOC 在后面的情况本文采用数据增广的方法加以解决,即如果 y 中包含 LOC 的序列标注,则直接后置即可,将增广得到的序列标注记为 $\preccurlyeq(y)$.

4.4 利用 BiLSTM-CRF + KNN-CA 自动标注其他数据

在 4.3 节中,满足第 3.3.1 节中三个条件的数据全部自动标注,其他数据可以通过合适的模型进行预测,然后通过 KNN-CA 算法进行优化. 刘章勋[25] 通过研究发现,人名、地名因其颗粒度较小,用基于字的方法能取得更好的结果;同时基于字向量的 BiLSTM-CRF 模型被广泛用于中文命名实体识别(CNER)任务中并被证实是理想的模型[9,10,11,26,16]. 因此本文结合 ***EBM*** 自身特征,采用 BIEOS 标注方案,采用基于字符的 BiLSTM-CRF 模型,利用已经获取的开发集、测试集和训练集得到 BiLSTM-CRF 深度学习模型,记为 Model($\boldsymbol{\Theta}$). 由此预测未标注的用户数据得到序列标注 y,再通过 $\boldsymbol{Y}(y) \Rightarrow \boldsymbol{EBM}$ 运算(详见第 3.3.4 节)得到对应的 ***EBM***,如果 ***EBM***! $=\phi$,还可以利用 KNN-CA 算法对其优化,其具体的性能分析详见 5.4 节.

4.5 实现自动标注算法

根据上面的讨论,可以得到自动标注算法的具体实现,如算法 2 所示.

算法 2 自动标注算法

输入:estates,users,OffLib,Tests,Devs
输出:labels(序列标注结果)
1: Trains,OnLib←ANNoTRAIN(estates,users,OffLib)
2: Model($\boldsymbol{\Theta}$)←MODEL. FIT(Trains)
3: users←users-users0
4: labels←$\boldsymbol{\Phi}$
5: for each $u \in$ users do
6: y←Model($\boldsymbol{\Theta}$). predict(u)
7: $\boldsymbol{Y}(y) \Rightarrow \boldsymbol{EBM}$
8: if ***EBM*** ! $=\phi$ then
9: $\overline{\boldsymbol{EBM}}$←KNN-CA (***EBM***)
10: $\overline{Y}(\overline{\boldsymbol{EBM}}) \Rightarrow y$
11: end if
12: labels←labels ∪ Labels(y,u)
13: end for

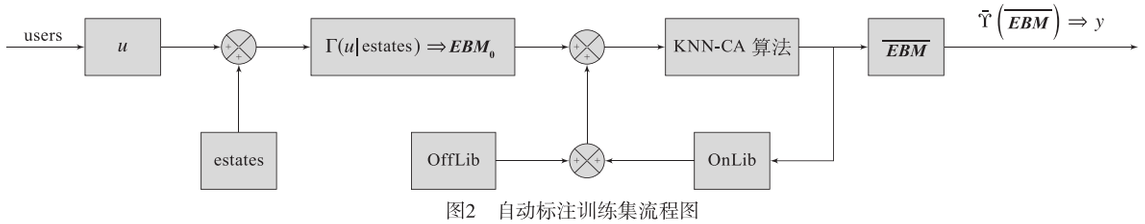

图2 自动标注训练集流程图

其中,第 1 行实现了自动标注训练集(详见 4.3 节);第 2 行通过训练得到 BiLSTM-CRF 深度学习模型;第 3 到 13 行实现了其他数据的自动标注(详见 4.4 节),特别地,users0 是自动标注训练集过程中被标注的数据,KNN-CA 算法(算法 1)详见第 4.2 节.

5　实验及分析

5.1　自动标注数据

本文处理的地址信息以及最后的标注结果如图 3 所示. 整个自动标注过程包括数据预处理、自动标注训练集、训练模型、预测标注和优化标注五个过程,如表 3 所示.

其中,具体的过程在第 4 节中进行了详细的讨论. 为了实现自动标注任务,本文硬件配置采用 3.4GHZ 的 i5-7500CPU,16G 内存,8G GeForce GTX 1070 GPU;操作系统使用 64 位的 Windows10,程序开发语言使用 python3.7 和 C#7.3,并由 tensorflow 实现 BiLSTM-CRF 模型. 第5.2 节将在此环境下对标注耗时做进一步分析.

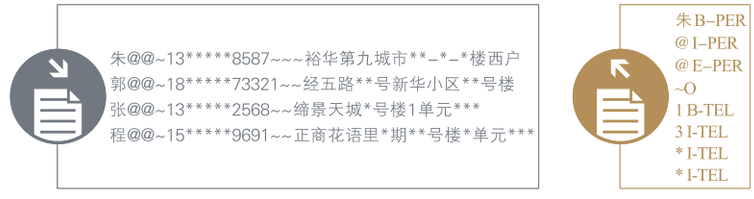

图3　地址信息和标注结果示意图

表 3　自动标注过程明细表

过程	输入	输出	耗时(h)	说明
数据预处理	商家真实数据	user,estates,OffLib,Tests,Devs	T_1	标注测试集和开发集较为耗时
自动标注训练集	user,estates,OffLib	OnLib,Train	T_2	KNN-CA 较为耗时
训练模型	Train,Tests,Devs	Model(Θ)	T_3	训练深度学习模型较为耗时
预测标注	users	labels	T_4	加载模型较为耗时
优化标注	labels,estates,OffLib,OnLib	最终序列标注	T_5	评估较为耗时

5.2　标注数据的耗时分析

本文核心工作几乎不需要任何人工标注,但是在构建开发集和测试集时,由系统抽样得到 800 条数据,共计 30801 个字,该过程需要人工标注,记为 Task,其耗时记为 T_0,下面给出 T_0 的估算过程. 标注一条用户数据如图 3 所示,其中左侧的数据要处理成右侧的标注数据,需要整体分析和逐字逐句标注. 以单位时间内标注数量(以字为单位)作为估算基线值[14],记为 count_per_unit,式(12)给出了人工标注的耗时估算:

$$T_{\text{human}} = \frac{|\text{users}|}{\text{count\_per\_unit}} \tag{12}$$

其中|users|是标注语料的总字数. 本文对 50 名志愿者进行的评测显示,完成五条用户数据(共计 191 个字)的标注任务,除去用时最短和最长的 10 人,人均每字用时 6.9s. 按此估算出 T_0 为 59h. 为了提升人工标注的效率,本文专门设计并开发了人工标注辅助工具(简称辅助工具),如图 4 所示.

将准备标注的用户数据(如图 3 所示)导入系统,该工具能够通过规则匹配完成 PER 和 TEL 的自动标注,同时自动完成 $\Gamma(u\mid\text{estates})\Rightarrow \boldsymbol{EBM}_0$ 运算,然后通过 $\bar{Y}(\boldsymbol{EBM})\Rightarrow y$ 运算完成初始标注;对于 $\boldsymbol{EBM}_0=\phi$ 的数据全部按照 EST 进行初始化;最后可以方便地选择对应的实体,即按下图 4 中左侧对应的按钮即可将选中的文本标注为对应的类型,此时选中文本的背景色将设置为对应实体类型的颜色. 标注完所有数据,按下“确定”按钮将自动生成序列标注. 由此完成 Task 的时间即 T_0 为 40min 左右. 在表 3 中,利用图 4 所示的辅助工具,自动标注系统能够在 1h 之内完成数据预处理,即 T_1 为 1h;T_2 是 s 级的耗时;T_3 因 K 和置信度阈值 ϖ 不同略有不同(详见第 5.4 节),但是相差很小,耗时 7min 左右;T_4 和 T_5 耗时都在 2min 之内. 因此完成全部自动标注任务耗时 70min 左右. 综上所述,可以得到标注数据的具体耗时,如表 4 所示.

由表 4 可见,相比纯人工标注,自动标注的效率有显著的优势. 同时,辅助工具能够极大提升人工标注的效率,但是辅助工具并不是为了解决标注问题,只是为了降低人工标注成本. 实际上,自动标注最耗时的部分是预处理,但是用户地址信息数据量的增加,并不会明显增加预处理部分的耗时. 换句话说,当数据量增大后,自动标注的优势将更加明显.

表 4 标注数据耗时分析表

标注方法	标注数据量(条)	耗时(h)	说明
纯人工标注	6751	489	纯人工标注所有数据的耗时,同 T_0 的估算方法
借助辅助工具的人工标注	6751	5.6	借助辅助工具人工标注所有数据的耗时
自动标注	6751	1.2	利用自动标注系统标注所有数据的耗时

图4 人工标注辅助工具操作界面

5.3 评价指标

本文采用查准率(P)、召回率(R)和 $F1$ 三个指标对自动标注结果进行评价,不仅从整体上进行评估,而且按照 PER、TEL、LOC、EST 和 INF 五种类型分别进行评估. 正如第 2 节所述,PER 和 TEL 的高精度标注会提升整体的指标. 对本文而言,这种提升并不是想要的,因此引入加权调和平均值 $WF1$,式(13)给出了具体的定义:

$$WF1 = \sum_{i=1}^{5} \kappa_i \times F1_i \tag{13}$$

其中,κ_i 按照 PER、TEL、LOC、EST、INF 依次取 0.1、0.1、0.2、0.4、0.2. 图 5 完整地显示了以 $\varpi = 0.91$、$K = 4$ 作为参数时上述所有指标在模型训练过程的变化详情,同时也直观地表明了 $WF1$ 能更好地评估自动标注的性能. 其中图5(a) ~ 图5(d)分别对应整体、LOC、EST

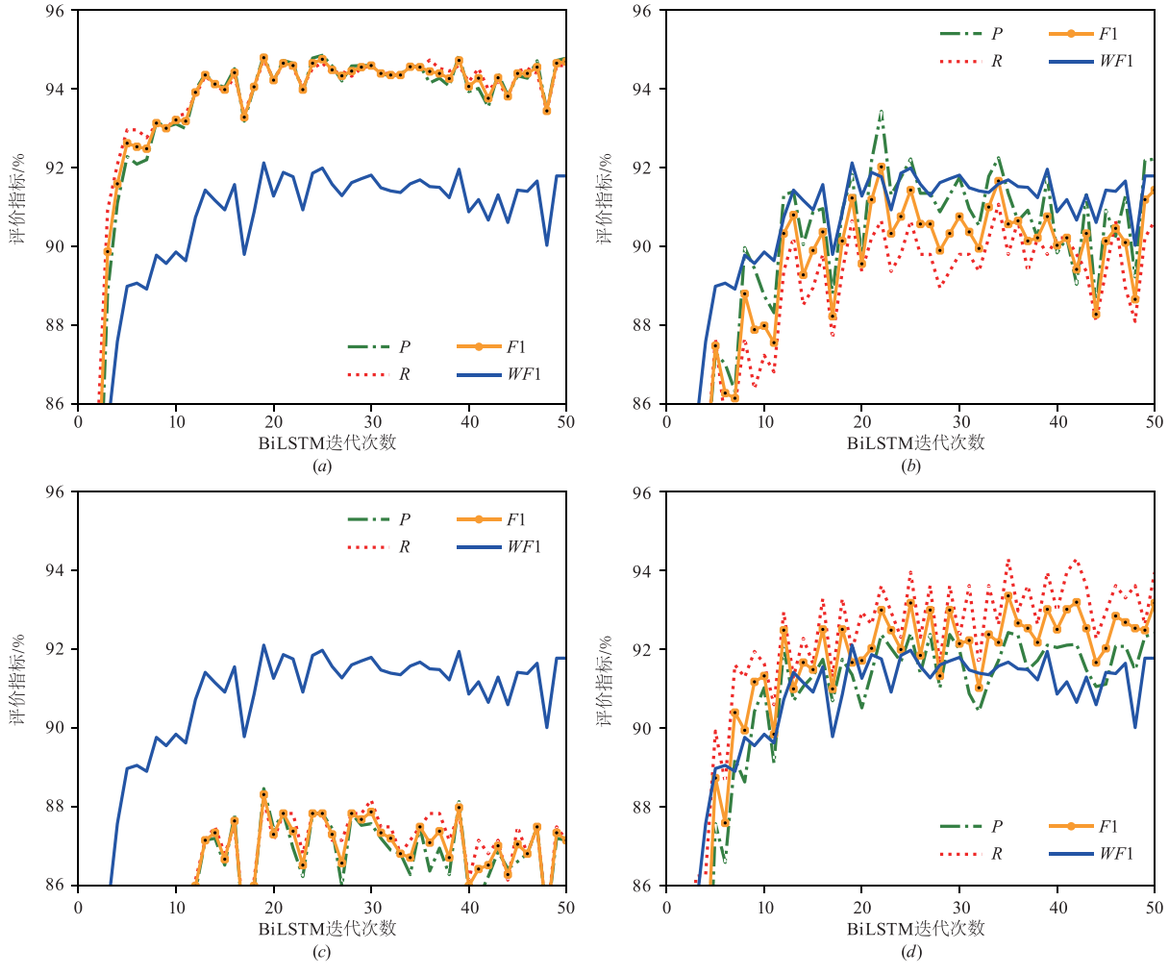

图5 BiLSTM-CRF模型训练过程指标分析图

和 INF 四种类型,由于 PER 和 TEL 所有的指标几乎都达到了 100%,图中不再显示.

5.4　KNN-CA 算法的参数敏感性分析

KNN-CA 算法有两个非常重要的参数,K 和置信度阈值 ϖ. 为了分析参数敏感性,首先自动构建训练集,分别设置 K 为 1、2、3 和 4,置信度阈值 ϖ 为 0.9、0.91、0.92、0.93、0.94、0.95 和 0.96;然后依次训练 BiLSTM-CRF 深度学习模型,共得到 28 个模型;最后使用统一的测试集对模型进行评估. 图 6 是以 $WF1$ 为指标所得的实验结果. ϖ 直接决定了训练集的大小,从深度学习模型考虑,训练集大一些会提升系统的稳定性;而 K 值决定了 KNN-CA 算法的复杂性. 由图 6 可知,将 ϖ 设定为 0.91,K 设定为 4 时,能够获得最佳的预测模型. 为了进一步提升标注的精度,本文利用 KNN-CA 算法对预测的序列标注再次进行优化,即自动标注算法(详见 4.5 节)的第 9 行和第 10 行,利用图 6 中性能最佳模型(即 $\varpi=0.91$,$K=4$),对 K 和 ϖ 取同样的 28 种组合,再次进行 28 次优化实验,图 7 显示了以 $WF1$ 为评估指标的最终实验结果.

从图 7 中可以看出,K 并非越大越好,ϖ 也不是越小越好,K 和 ϖ 分别取 2 和 0.9 时取得最佳的优化结果. 可见,KNN-CA 算法在自动标注训练集和优化预测序列标注两个过程中,K 和 ϖ 最佳组合有所不同,在很大程度上与 OnLib 有关. 对比图 6 和图 7 可以明显看出利用 KNN-CA 算法优化预测序列的效果还是比较明显的.

5.5　实验结果分析

5.5.1　数据增广和 KNN-CA 实验分析

本文以 BiLSTM-CRF 模型为中心,在训练模型之前以获取训练集为主要任务. 综合起来,获取训练集有四种方法,即 estates1、aug-p1、KNN-CA 和 KM-aug. 利用每种方法获取的训练集训练模型,然后直接生成预测序列标注,表 5 展示了具体的评价指标. 其中 $F1(L)$、$F1(E)$ 和 $F1(I)$ 分别为 LOC、EST 和 INF 的 $F1$ 值. KAS-aug 为利用 KNN-CA 算法进一步优化 KM-aug 的结果,其中 K 和 ϖ 根据第 5.4 节的实验结果取值为 4 和 0.91.

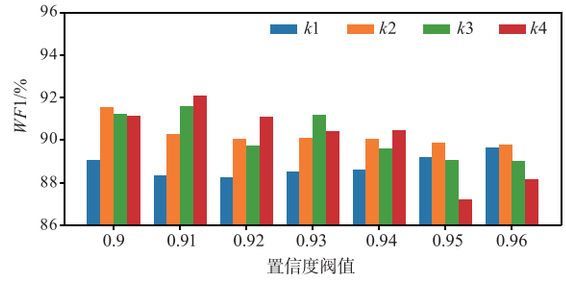

图6　KNN-CA算法在自动标注训练集过程中的参数敏感性分析

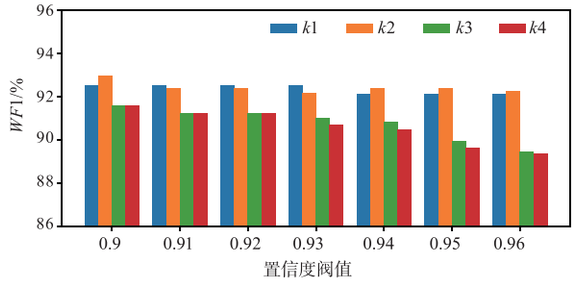

图7　KNN-CA算法在优化预测序列标注过程中的参数敏感性分析

从表 5 可见,estates1 即直接匹配预置小区集 estates 得到的训练集直接训练模型的性能较低,但是经过数据增广(aug-p1)能有效提升性能,尤其是 LOC 的提升最为明显;相比数据增广,KNN-CA 也取得了明显效果,尤其是 INF 的提升最为显著,这是因为 INF 和 EST 具有固定的位序,以及明显的边界特征(详见式(11));KM-aug 是将两者结合起来,整体的 $F1$ 值达到了 94.78%,取得了理想的标注结果. 经过深入分析,发现该过程中构建的 OnLib 也有显著的改善;KAS-aug 是利用 KNN-CA 对预测结果进行优化,得到了较好的标注指标,其 $F1$ 值达到了 95.35%,表明 KNN-CA 确实可以进一步优化预测序列标注.

5.5.2　和其他方法的对比分析

为了进一步分析 ***EBM*** 模型,本文基于文献[17,18]分别实现了基于自动回标和主动学习的自动标注,其结果如表 6 所示.

表 5　数据增广和 KNN-CA 在 $K=4$ 且 $\varpi=0.91$ 条件下性能对比表

方法	$F1$(整体)	$F1(L)$	$F1(E)$	$F1(I)$	$WF1$	说明
estates1	82.61	67.83	62.22	81.34	74.36	由单小区匹配构建训练集
aug-p1	87.15	87.50	69.67	76.67	80.70	数据增广
KNN-CA	87.13	77.29	66.44	89.11	79.81	KNN-CA 优化训练集但不增广数据
KM-aug	94.78	91.22	88.32	92.67	92.11	KNN-CA 优化训练集并数据增广
KAS-aug	**95.35**	**91.61**	**89.49**	**94.16**	**92.95**	KNN-CA 优化预测结果($K=2$,$\varpi=0.9$)

表 6　与其他方法性能对比表

方法	$F1$(整体)	$F1(L)$	$F1(E)$	$F1(I)$	$WF1$	说明
BA	87.21	86.74	70.35	76.79	80.85	基于文献[17]的自动回标
AL	94.44	**93.82**	86.78	90.48	91.57	基于文献[18]的主动学习
KAS-aug	**95.35**	91.61	**89.49**	**94.16**	**92.95**	基于 ***EBM*** 和 KNN-CA

自动回标[17]利用百度百科回标地理实体. 由于地址信息的自身特征(详见第 2 节),不需要考虑实体关系和实体歧义问题,而且百度百科中只有 EST,因此只需要判断存在性即可. 实验表明,300 条 Tests 中有 176 条数据进行了自动回标. 从表 6 可见,方法 BA 并不太理想,主要是无法处理特殊而固定的表达(比如小区分期)以及无法利用重要的边界特征.

主动学习[18]利用少量的已标注样本训练模型,获取待测样本的自动标注结果及其置信度,对置信度低的样本进行人工标注. 方法 AL 将 aug-p1 模型(详见 5.5.1 节)作为训练模型,Tests 作为待测样本,利用式(10)求解置信度,并将 ϖ 取为 0.9,对低于 ϖ 的 122 条数据进行人工标注,效果良好. 但是该方法需要一定量的人工标注,成本较高. 上述一系列实验表明,将轻量级且具有自动构建能力的人工特征库和基于字向量的 BiLSTM-CRF 深度学习模型结合起来,通过 KNN-CA 进行近邻标签修正,实现了比较理想的自动标注. 由此构建了在地址信息领域中能够快捷有效地解决 CGSNE 问题的中文地址信息语料库.

6 结论

本文将基于深度学习的字向量和基于 ***EBM*** 表示模型的词向量相结合,既利用了传统方法中人工特征的优势,又利用了深度学习模型的强大预测能力,有效地解决了文本地址信息的自动标注问题. 相对于主动学习而言,不需要任何人工标注成本,明显降低了深度学习解决 CGSNE 问题所需的人工成本. 由此构建了地址信息语料库,脱敏用户隐私信息后可以公开发布. 本文提出的基于 ***EBM*** 表示模型可以有效提升标注数据的置信度和优化序列标注预测结果,因此该模型能够在降低专家标注成本并提升主动学习框架迭代效率方面对半监督主动学习框架[24,27~29]产生积极的影响.

参考文献

[1] 何炎祥,罗楚威,胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件,2015,32(1):179 - 185 + 202.
He Yanxiang, Luo Chuwei, Hu Binyao. Geographic entity recognition method based on CRF model and rules combination[J]. Computer Applications & Software, 2015, 32(1):179 - 185 + 202. (in Chinese)

[2] 沈达阳,孙茂松. 中国地名的自动辨识[A]. 全国第三届计算语言学联合学术会议论文集[C]. 北京:清华大学出版社,1995. 68 - 74.

[3] 郑家恒,李鑫,谭红叶. 基于语料库的中文姓名识别方法研究[J]. 中文信息学报,2000,14(1):7 - 12.
Zheng Jiaheng, Li Xin, Tan Hongye. The research of Chinese names recognition method based on corpus[J]. Journal of Chinese Information Processing, 2000, 14(1):7 - 12. (in Chinese)

[4] 张雪英,朱少楠,张春菊. 中文文本的地理命名实体标注[J]. 测绘学报,2012,41(1):115 - 120.
Zhang Xueying, Zhu Shaonan, Zhang Chunju. Annotation of geographical named entities in Chinese text[J]. Acta Geodaetica et Cartographica Sinica, 2012, 41(1):115 - 120. (in Chinese)

[5] 官登水. 中文微博的地理位置命名实体识别研究[D]. 成都:西华大学,2016. 17 - 27.

[6] Graves A, Mohamed A, Hinton G E. Speech recognition with deep recurrent neural networks[A]. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Vancouver, BC, Canada: IEEE, 2013. 6645 - 6649.

[7] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735 - 1780.

[8] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[A]. Proceedings of NAACL-HLT'16[C]. Washington D. C., USA: IEEE Press, 2016. 260 - 270.

[9] Huang Z H, Xu W, Yu K, et al. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. https://arxiv.org/pdf/1508.01991. 2015-08-09.

[10] Dong C H, Zhang J J, Zong C Q, Hattori M, Di H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[A]. Natural Language Understanding and Intelligent Applications[C]. Cham: Springer International Publishing. 2016. 239 - 250.

[11] Shen Y Y, Yun H K, Lipton Z C, et al. Deep active learning for named entity recognition[A]. Proceedings of the 2nd Workshop on Representation Learning for NLP[C]. Vancouver, Canada: Association for Computational Linguistics. 2017. 252 - 256.

[12] 杨阳,张文生. 基于深度学习的图像自动标注算法[J]. 数据采集与处理,2015,30(1):88 - 98.
Yang Yang, Zhang Wensheng. Image auto-annotation based on deep learning[J]. Journal of Data Acquisition and Processing, 2015, 30(1):88 - 98. (in Chinese)

[13] 徐勇,张慧. 图像自动标注方法研究综述[J]. 现代情报,2016,36(3):144 - 150.
Xu Yong, Zhang Hui. Summary of automatic image annotation method[J]. Journal of modern information, 2016, 36(3):144 - 150. (in Chinese)

[14] 程冰. 基于卷积神经网络的自动标注技术的研究[J]. 电子世界,2019(16):124 - 126.

[15] 黄冬梅,许琼琼,贺琪,杜艳玲. 融合多特征的深度学习标注方法[J]. 计算机工程与应用,2018,54(1):224 - 228.

Huang Dongmei, Xu Qiongqiong, He Qi, Du Meiling. Multi-features fusion for image auto-annotation based on DBN model[J]. Computer Engineering and Applications, 2018,54(1):224 - 228. (in Chinese)

[16] 徐飞,叶文豪,宋英华. 基于 BiLSTM-CRF 模型的食品安全事件词性自动标注研究[J]. 情报学报,2018,37(12):1204 - 1211.
Xu Fei, Ye Wenhao, Song Yinghua. Part-of-speech automated annotation of food safety events based on BiLSTM-CRF [J]. Journal of the China Society for Scientific and Technical Information,2018,37(12):1204 - 1211. (in Chinese)

[17] 王姬卜,陆锋. 基于自动回标的地理实体关系语料库构建方法[J]. 地球信息科学学报,2018,20(7):871 - 879.
Wang Jibu, Lu Feng. Constructing the corpus of geographical entity relations based on automatic annotation [J]. Journal of Geo-Information Science, 2018, 20(7): 871 - 879. (in Chinese)

[18] 朱珠,李寿山,戴敏,周国栋. 结合主动学习和自动标注的评价对象抽取方法[J]. 山东大学学报(理学版),2015,50(7):38 - 44.
Zhu Zhu, Li Shoushan, Dai Min, Zhou Guodong. Opinion target extraction with active-learning and automatic annotation[J]. Journal of shandong university, 2015, 50(7): 38 - 44. (in Chinese)

[19] Schulz C, Meyer C M, Kiesewetter J, et al. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains[A]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics[C]. Florence, Italy: Association for Computational Linguistics, 2019. 2761 - 2772.

[20] 鞠久朋,张伟伟,宁建军,周国栋. CRF 与规则相结合的地理空间命名实体识别[J]. 计算机工程,2011,37(7):210 - 212 + 215.
Ju Jiupeng, Zhang Weiwei, Ning Jianjun, Zhou Guodong. Geospatial named entities recognition using combination of CRF and rules[J]. Computer Engineering, 2011, 37(7): 210 - 212 + 215. (in Chinese)

[21] 谢婷婷,严柯. 基于统计的中文地址位置语义解析方法研究[J]. 软件导刊,2017,16(10):19 - 21.
Xie Tingting, Yan Ke. The method of semantic resolution of Chinese addresses based on statistics [J]. Software Guide, 2017, 16(10): 19 - 21. (in Chinese)

[22] 黄爽. 中文地址位置语义解析方法的研究 [D]. 武汉:武汉工程大学,2017. 2 - 3.

[23] 邱锡鹏. 神经网络与深度学习[DB/OL]. https://nndl.github.io,2019.

[24] He G L, Li Y F, Zhao W, et al. An uncertainty and density based active semi-supervised learning scheme for positive unlabeled multivariate time series classification[J]. Knowledge Based Systems, 2017, 124(MAY15): 80 - 92.

[25] 刘章勋. 中文命名实体识别粒度和特征选择研究 [D]. 哈尔滨:哈尔滨工业大学,2010. 14 - 39.

[26] Zhang Y, Yang J. Chinese NER using lattice LSTM[A]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]. Melbourne, Australia: Association for Computational Linguistics, 2018. 1554 - 1564.

[27] He G L, Duan Y, Li Y F, et al. Active learning for multivariate time series classification with positive unlabeled data[A]. 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)[C]. Vietri sul Mare: IEEE, 2015. 178 - 185.

[28] Bota P J, Silva J, Folgado D, et al. A semi-automatic annotation approach for human activity recognition[J]. Sensors, 2019, 19(3): 501.

[29] 凌广明,徐武平,穆晓峰,徐爱萍. 一种基于深度学习的地理信息的自动标注方法及装置[P]. 中国专利:109614455A,2019-04-12.

作者简介

凌广明　男,1981 年生于河南南阳,武汉大学计算机学院博士研究生. 主要研究方向为自然语言处理、实用 GIS 系统设计与开发等.
E-mail:linggm_126@126.com

徐爱萍(通讯作者)　女,博士、武汉大学计算机学院教授、硕导博导. 1962 年 6 月出生,湖北武汉人. 主要研究方向为大数据分析和自然语言处理. 在从事教学工作的同时主持完成多项纵向和横向项目,在国内外学术期刊与会议上发表研究论文 50 余篇.
E-mail:xuaiping@whu.edu.cn

王　伟　男,1962 年出生,湖北英山人,博士,武汉大学测绘遥感信息工程国家重点实验室教授,博士生导师. 主要研究方向:智慧城市与时空大数据应用研究.
E-mail:wangwei8091@163.com