基于词向量语义分类的微博实体链接方法

冯 冲 [1] 石 戈 [1] 郭宇航 [1] 龚 静 [1] 黄河燕 [1,2]

摘 要 微博实体链接是把微博中给定的指称链接到知识库的过程, 广泛应用于信息抽取、自动问答等自然语言处理任务 (Natural language processing, NLP). 由于微博内容简短, 传统长文本实体链接的算法并不能很好地用于微博实体链接任务. 以往研究大都基于实体指称及其上下文构建模型进行消歧, 难以识别具有相似词汇和句法特征的候选实体. 本文充分利用指称和候选实体本身所含有的语义信息, 提出在词向量层面对任务进行抽象建模, 并设计一种基于词向量语义分类的微博实体链接方法. 首先通过神经网络训练词向量模板, 然后通过实体聚类获得类别标签作为特征, 再通过多分类模型预测目标实体的主题类别来完成实体消歧. 在 NLPCC2014 公开评测数据集上的实验结果表明, 本文方法的准确率和召回率均高于此前已报道的最佳结果, 特别是实体链接准确率有显著提升.

关键词 词向量, 实体链接, 社会媒体处理, 神经网络, 多分类

引用格式 冯冲, 石戈, 郭宇航, 龚静, 黄河燕. 基于词向量语义分类的微博实体链接方法. 自动化学报, 2016, **42**(6): 915−922

DOI 10.16383/j.aas.2016.c150715

An Entity Linking Method for Microblog Based on Semantic Categorization by Word Embeddings

FENG Chong[1] SHI Ge[1] GUO Yu-Hang[1] GONG Jing[1] HUANG He-Yan[1,2]

Abstract As a widely applied task in natural language processing (NLP), named entity linking (NEL) is to link a given mention to an unambiguous entity in knowledge base. NEL plays an important role in information extraction and question answering. Since contents of microblog are short, traditional algorithms for long texts linking do not fit the microblog linking task well. Precious studies mostly constructed models based on mentions and its context to disambiguate entities, which are difficult to identify candidates with similar lexical and syntactic features. In this paper, we propose a novel NEL method based on semantic categorization through abstracting in terms of word embeddings, which can make full use of semantic involved in mentions and candidates. Initially, we get the word embeddings through neural network and cluster the entities as features. Then, the candidates are disambiguated through predicting the categories of entities by multiple classifiers. Lastly, we test the method on dataset of NLPCC2014, and draw the conclusion that the proposed method gets a better result than the best known work, especially on accurancy.

Key words Word embedding, entity linking, social media processing, neural network, multiple classifiers

Citation Feng Chong, Shi Ge, Guo Yu-Hang, Gong Jing, Huang He-Yan. An entity linking method for microblog based on semantic categorization by word embeddings. *Acta Automatica Sinica*, 2016, **42**(6): 915−922

微博是一种通过关注机制分享简短实时信息的广播式的社交网络平台, 已成为目前最流行的社交平台之一. 截至 2014 年 9 月 30 日, 微博的月活跃用户已经达到 1.67 亿, 用户每天产生的微博数目达到 2 亿[1]. 如何从海量微博中自动地及时分析、获得信息已成为研究和应用热点问题, 微博实体链接是其中关键任务之一.

微博实体链接是指将微博中已经识别出的实体指称链接到知识库中的一个具体真实实体的过程[2−3]. 例如, 微博 “在我眼中, 科比还是比乔丹棒的” 中, “乔丹” 作为实体指称, 在知识库中有 6 个实体义项. 实体链接的目标就是要确定, 这里的 “乔丹”, 指代的是知识库中哪个实体义项.

以往实体链接研究主要集中在新闻等长文, 对于微博等短文本的研究工作刚起步. 微博具有两个特点[4]: 1) 内容非常简短, 通常每篇至多包含 140 个字符; 2) 格式不规范, 经常出现口语和缩写等灵活

收稿日期 2015-10-29 录用日期 2016-05-03
Manuscript received October 29, 2015; accepted May 3, 2016
国家重点基础研究发展计划 (973 计划) (2013CB329303), 国家高技术研究发展计划 (863 计划) (2015AA015404), 国家自然科学基金 (61502035), 高等学校博士学科点专项科研基金 (20121101120026) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329303), National High Technology Research and Development Program of China (863 Program) (2015AA015404), National Natural Science Foundation of China (61502035), and Specialized Research Fund for the Doctoral Program of Higher Education (20121101120026)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 北京理工大学计算机学院 北京 100081 2. 北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081
1. College of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 2. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing 100081

的非正式表达. 传统的长文本实体链接方法主要从实体指称的上下文中抽取特征用于实体消歧, 但是因为微博内容简短, 传统方法难以抽取有效特征.

针对微博文本上下文不足的问题, 部分工作借助微博的结构特点扩充微博的上下文. Jiang 等[5] 利用 Twitter 中的转发、回复以及同一用户的其他帖子扩充上下文进行情感分类. Shen 等[6] 利用同一个 Twitter 用户的数据对其兴趣建模, 提高与用户兴趣模型一致性高的候选实体的权重. Guo 等[2] 利用类似主题的微博建模来对候选实体进行消歧. Liu 等[7] 利用指称上下文–实体上下文、指称上下文–指称上下文、实体上下文–实体上下文的文本相似度来对实体消歧.

以上方法虽然能够改善微博实体链接中上下文特征匮乏的状况, 但本质上受限于对更多微博数据资源 (用户的转发、回复和其他微博等内容) 的获取, 增加了处理开销. 如果缺乏符合建模要求的数据, 仍难建立有效模型[8].

本文从充分利用指称和候选实体本身所含有的语义信息入手, 提出假设 "一条微博中的名词, 包括实体指称, 位于相近的语义空间", 从而把微博实体链接问题转化为语义空间中的分类问题. 以 NLPCC 2014[9] 评测数据集中的微博样本 "好怀念当时的那支队伍啊! 弗朗西斯、麦迪、巴蒂尔、大姚、斯科拉、穆托姆博、诺瓦克" 为例, "大姚" 是 "姚明" 和 "姚晨" 的别名. 两人都是媒体热点人物, 实体指称具有类似的词汇和语法特征, 传统方法难以识别. 而考察指称与上下文中其他名词的语义距离则可进行有效区分. 统计 577 条微博训练数据, 得出结果如表 1 所示.

表 1 训练集数据统计

Table 1 Statistics in training data

平均每条微博中名词个数	7.91
同一语义类别名词个数超过 7 的微博	34
同一语义类别名词个数超过 6 的微博	81
同一语义类别名词个数超过 5 的微博	207
同一语义类别名词个数超过 4 的微博	416
同一语义类别名词个数超过 3 的微博	502

从统计数据可以看出, 平均每条微博中含有 7.91 个名词, 同一语义类别名词个数超过 3 的微博占训练数据的 87 %, 验证了假设 "一条微博中的名词位于相近的语义空间" 的合理性.

基于以上假设, 利用知识库中实体的深层语义信息, 基于词向量对微博进行建模和实体消歧. 传统的方法已经验证, 足够多的语义特征可以提高实体链接的准确率[10], 但由于微博是短文本, 从微博本身很难加入更多的特征, 因此从实体链接的另一方面入手, 将知识库中的实体表征为含有语义、语法信息的分布式向量, 从语义分类层面对微博进行建模和实体消歧.

本文的主要贡献是提出了一种基于神经网络和多分类回归模型的命名实体链接方法, 将微博中上下文名词与对应的待链接实体映射到同一个语义主题空间, 并以此训练分类模型对实体进行语义消歧. 其创新之处在于, 从神经网络语言模型的角度, 以分类器分类预测的方式提出了实体消歧方法, 不仅能够充分地利用上下文语义信息, 也能够利用实体的语义分类信息来进行消歧, 并降低了获取训练语料的难度.

本文结构如下: 第 1 节介绍本文提出的方法; 第 2 节是实体链接部分; 第 3 节是实验部分; 最后是结论和展望.

1 词向量语义分类方法

1.1 任务描述

本文用 $\boldsymbol{M} = (m_1, m_2, \cdots, m_n)$ 表示微博中给定的一些指称, 用 $\boldsymbol{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$ 表示与指称对应的候选实体的集合. 则实体链接的任务是将给定的指称 m_i 链接到候选实体集合 $\boldsymbol{e}_i$ 中某个无歧义实体 $\boldsymbol{e}_{ij}$ 的过程. 基于前述假设, 词向量语义分类模型的目标是获得语义分类特征.

$$SCWE_{m_i} = f(m_i, \boldsymbol{n}_i) \tag{1}$$

其中, $\boldsymbol{n}_i$ 是指称 m_i 对应微博中的名词集合.

1.2 词向量语义分类模型构建 (SCWE)

1.2.1 词向量语义模板构建

图 1 是本文的词向量语义分类模型. 其中神经网络部分采用的是 CBOW 模型[11]. CBOW 是一个三层神经网络模型, 从左至右依次是输入层、隐含层和输出层. 其基本思想是通过训练将每个词映射成含有语义、语法信息的 K 维实数向量 (K 是可选参数, 一般为 $50 \sim 200$), 通过向量之间的距离 (例如欧氏距离、cosine 相似度等) 来判断它们之间的语义相似度. 该模型是对语言模型进行建模, 在建模的同时获得词语在分布式向量空间上的表示.

假设语料库是由 S 个句子组成的一个句子序列, 整个语料库有 V 个词, T_j 表示第 j 个句子的词个数, 则对整个语料库来说, 该模型的目标函数可以表示为

$$l(\theta) = \log L(\theta) =$$

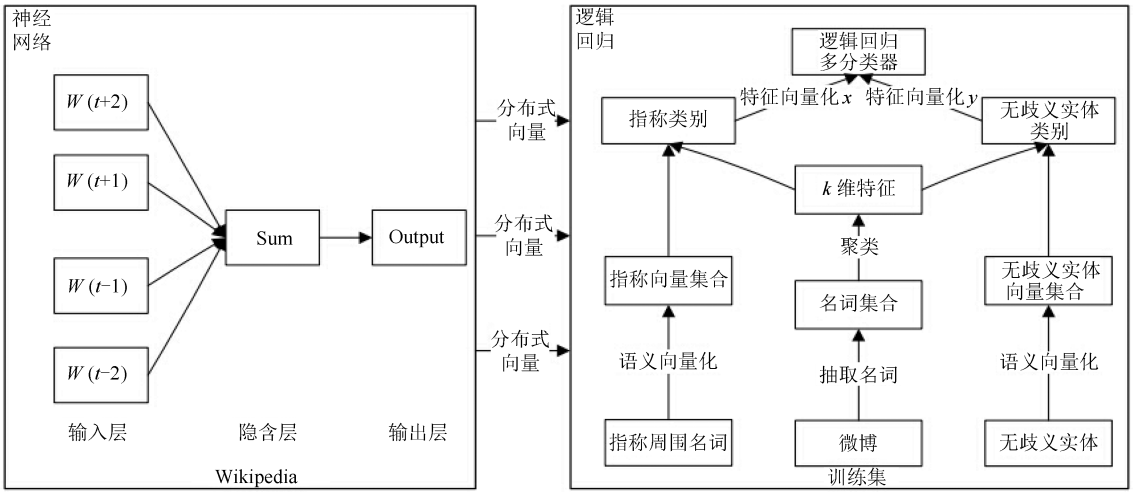


图 1　词向量语义分类模型

Fig. 1　Model of semantical categorization by word embeddings

$$\arg\max \frac{1}{v}\sum_{j=1}^{s}\left(\sum_{ij=1}^{T_j}\log p(w_{ij}|context_{ij})\right) \tag{2}$$

通过随机梯度下降对目标函数求解, 即可将语料库中 V 个词表示为含有深层语义特征的分布式向量.

1.2.2　特征选择

对于给定的训练数据集, 用 $T=(t_1,t_2,\cdots,t_n)$ 表示训练数据集中的每条微博, $S=(s_1,s_2,\cdots,s_n)$ 表示与微博相对应的, 已经链接到知识库的无歧义实体的集合. 基于假设 "一条微博中的所有名词, 包括实体指称, 位于相近的语义空间", 抽取训练集中的名词, 通过第 1.2.1 节中方法获得词向量模板, 将抽取的名词表示为分布式向量, 得到名词向量集合 N. 分布式向量中含有深层语义信息, 对集合 N 用 k-means[12] 进行聚类, 获得 k 个中心点 $C=(c_1,c_2,\cdots,c_k)$ 作为 k 个特征 (其中 k 为 k-means 聚类核心个数). 同时, 通过计算每个词到 k 个中心点的距离, 获得集合 N 中每个词的类别标签.

1.2.3　训练数据特征化

由第 1.2.2 节中得到的每个名词的标签, 可以把集合 T 中的微博 t_i 表示成 k 维向量, 把 t_i 中的每个名词类别出现的频数作为该维特征上的权值.

如图 2 所示, 选取 $k=10$, 即选取 10 维特征. 乔丹、科比、奥尼尔、艾弗森对应的聚类标签为 3, 球员对应聚类标签为 1, 退役对应聚类标签为 5. 则可以将这条微博表示为 $(0,1,0,4,0,1,0,0,0,0)$. 与这条微博相对应的, 已经链接到知识库中的无歧义实体 s_i 为 "迈克尔 · 乔丹", 从向量模板中找出 "迈克尔 · 乔丹" 所对应的向量, 通过公式

$$\arg\max(\cos(m_i,C)) \tag{3}$$

计算与迈克尔 · 乔丹最接近的类别, 得出迈克尔 · 乔丹所属类别为 3, 于是可将 s_i 迈克尔 · 乔丹表示成向量 $(0,0,0,1,0,0,0,0,0,0)$.

虽然 乔丹/n 已经 退役/n 很久，但是 他 是 我 心目 中 最好 的 球员/n ， 科比/n 、奥尼尔/n 、艾弗森/n 都跟他 没什么可比性

图 2　训练数据示例

Fig. 2　Example of the training data

通过上述过程我们可以把训练集合中的微博和对应的待链接实体表示成 k (k 为所选取特征个数) 维的向量对.

1.2.4　多分类模型训练

相关工作表明[13−15], 在实际运用中, 逻辑回归分类器跟 SVM、随机森林等分类模型效果接近, 但逻辑回归分类器算法复杂度最低. 因此, 该部分我们采用逻辑回归分类器构建分类模型. 用 $T(t_1^*,t_2^*,\cdots,t_n^*)$ 表示特征化后的微博集合, 用 $S(s_1^*,s_2^*,\cdots,s_n^*)$ 表示特征化后与微博相对应的, 已经链接到知识库的无歧义实体的集合. 其中, t_i^* 和 s_i^* 均为 k 维向量. 在特征化训练数据并将之用向量表示后, 可以将问题转化为多分类问题. 这样做的意义在于, 既利用了微博待链接实体跟该条微博中的名词之间的关系, 又利用了实体的词向量语义特征.

令

$$\boldsymbol{x}=t_i^*=(x_0,x_1,\cdots,x_{k-1}) \tag{4}$$

$$\boldsymbol{y}=s_i^*=(y_0,y_1,\cdots,y_{k-1}) \tag{5}$$

则 $(\boldsymbol{x}, \boldsymbol{y})$ 为一个观测样本. 可以得出 $y = c_i$ (c_i 是第 1.2.3 节选取的特征, 取值范围为 $[0, 1, 2, \cdots, k-1]$) 的条件概率:

$$P(y = c_i|x) = \frac{\mathrm{e}^{g_{e_i}(x)}}{1 + \sum\limits_{j=1}^{k-1} \mathrm{e}^{g_{e_i}(x)}} \tag{6}$$

由此可以得出相应的多分类逻辑回归[16] 模型:

$$g_{c_i}(x) = \log \frac{P(y = c_i|x)}{1 - P(y = c_i|x)} = \beta_{m0} + \beta_{m1}x_1 + \cdots + \beta_{m(k-1)}x_{k-1} \tag{7}$$

通过构造似然函数对模型求解. 把 n 个独立的观测样本记作 (X_i, Y_{ji}), $i = 1, 2, \cdots, n$. 利用上面规定, 得出如下似然函数:

$$l(\beta) = \prod_{i=1}^{n} (\pi_0(X_i)^{y_{0i}} \pi_1(X_i)^{y_{1i}} \cdots \pi_{k-1}(X_i)^{y_{k-1i}}) \tag{8}$$

其中, $\pi_j(X_i) = P(y = j|X_i)$. 对等式两端取对数整理可以得到如下的对数似然函数:

$$L(\beta) = \sum_{j=0}^{k-1} \sum_{i=1}^{n} y_{ji} \log(\pi_j(X_i)) \tag{9}$$

通过梯度下降法对似然函数求解, 至此得到训练好的词向量语义分类模型.

2 实体链接过程

2.1 任务描述与特征选择

微博实体链接是将微博中给定的实体指称链接到知识库中无歧义实体的过程. 本文选取两个特征进行实体消歧, 词向量语义分类特征 (Semantic categorization by word embeddings, SCWE) 和实体流行度特征 (Entity frequency, EF). 实体链接过程表述如下:

$$\boldsymbol{E}^* = \arg \max_{\forall e_{ij} \epsilon \boldsymbol{e}_i} [\lambda \cos(SCWE_{m_i}, e_{ij}) + (1 - \lambda) f(e_{ij})] \tag{10}$$

其中, $\boldsymbol{E}^*$ 表示候选实体的最终得分, e_{ij} 表示指称 m_i 对应的候选实体, $f(e_{ij})$ 表示候选实体流行度得分, λ 表示词向量语义分类特征的权重.

2.2 实体链接过程

图 3 所示是整个实体链接的过程. 整个过程可以分为三个部分: 实体指称标准化、候选实体扩充和实体消歧. 微博中许多实体有若干不同的名称、提法, 有的是别名 (如小飞侠)、昵称 (如大姚), 有的是全名的一部分或是缩写 (如北京理工、北理工、北理等). 因此, 首先需要对微博中出现的指称映射到一种标准的表达形式. 具体地, 构建一个同义词词表[17] (见表 2) 来解决这个问题. 其中, Key 值表示实体的不规则指称, Value 表示标准实体.

表 2 同义词表举例

Table 2 Examples of synonym lexicon

文中实体表示 (Key)	标准实体表示 (Value)
迈克尔乔丹	迈克尔 · 乔丹
飞人	
篮球之神	
迈克尔 · 杰弗里 · 乔丹	
Michael Jordan	
Michael Jeffrey Jordan	
乔丹	

将实体指称标准化之后, 需要为待消歧的命名实体构建一个候选实体列表. 本文构建了歧义词表 (见表 3), 表 3 中存储的是实体的标准形式 (Key) 及其对应的无歧义实体列表 (List).

表 3 歧义词表举例

Table 3 Examples of ambiguity lexicon

标准实体表示 (Key)	无歧义真实实体 (List)
苹果	苹果 (果树)
	苹果 (果实)
	苹果 (公司)
	苹果 (人物)
	苹果 (动漫角色)
	苹果 (歌曲)

在链接阶段, 需要对扩充后的候选实体列表进行消歧, 本文通过词向量语义分类特征和实体流行度特征进行消歧. 词向量语义分类特征由第二部分构建的模型获得, 实体流行度特征[4] 则由 Wikipedia 页面中实体在所有描述页面中出现的次数来度量 (见表 4), 并根据经验对实体流行度的权值进行设置 (见表 5).

表 4 实体流行度表举例

Table 4 Examples of entity frequency

无歧义真实实体 (Entity)	实体出现次数 (Frequency)
苹果 (果树)	26
苹果 (果实)	39
苹果 (公司)	158
苹果 (人物)	2

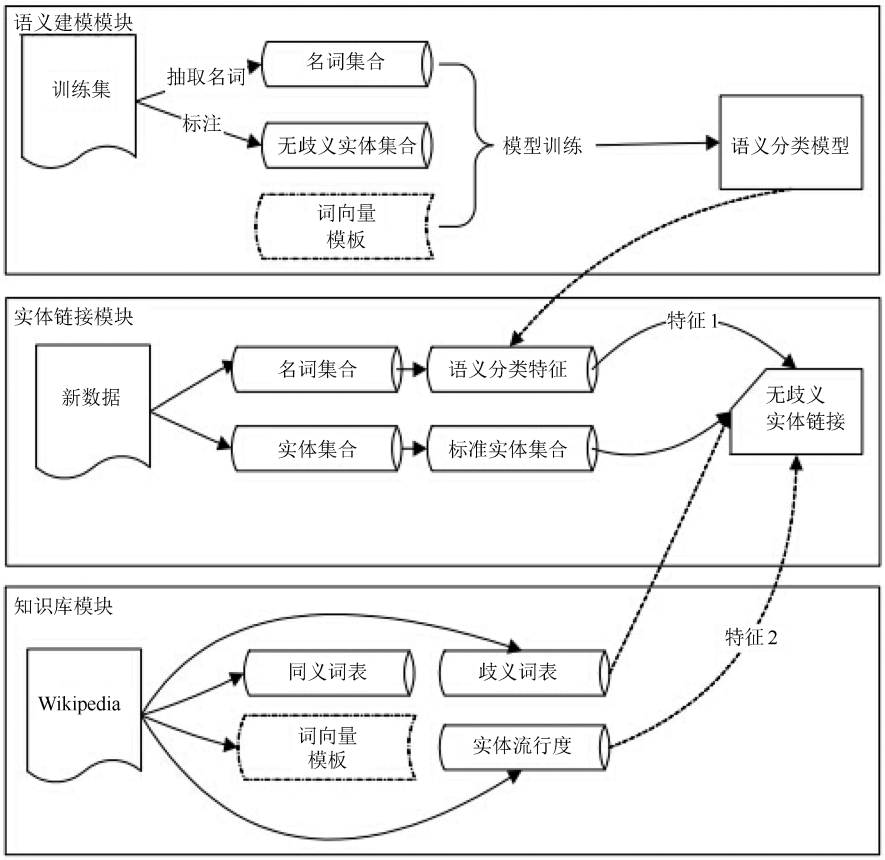


图 3　实体链接过程

Fig. 3　Process of entity linking

表 5　实体流行度权值

Table 5　Weights of entity frequency

流行度排行	一	二	三	四	五
权值	1.0	0.8	0.7	0.6	0.5

综上所述, 本文提出的基于词向量语义分类的微博实体链接过程的算法如下.

算法 1. 基于词向量语义分类的微博实体链接方法

输入. 微博及其对应的待链接实体

输出. 链接到知识库的无歧义实体

步骤 1. 根据知识库中的同义词表, 对 $\boldsymbol{M}$ 中的指称进行描述标准化, 得到标准化后的指称集合.

步骤 2. $\boldsymbol{E}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)$, 并对 $\boldsymbol{e}_i$ 中的候选实体按流行度排序. 当 $|\boldsymbol{e}_i|=0$ 时, 说明知识库中没有相应的实体, 即指称 m_i 是不可链接的, 返回标签 NIL; 当 $|\boldsymbol{e}_i|=1$ 时, 直接返回唯一候选实体作为最终链接实体; 当 $|\boldsymbol{e}_i|>1$ 时, 执行步骤 3 和步骤 4.

步骤 3. 再通过计算各候选实体到该标签的余弦距离得到各候选实体的语义分类特征.

步骤 4. 按式 (10) 计算每个候选实体两个特征的加权和, 并输出结果最高的候选实体作为最终链接实体. 如果加权和小于阈值 α, 则返回标记 NIL.

3　实验

3.1　数据集描述

本文建立同义词表、歧义词表、实体流行度表以及训练的词向量模板所用数据均为 Wikipedia[18], 使用的数据版本是 2015 年 7 月 19 日的中文百科. 通过规则对知识库抽取信息并进行统计, 获得数据规模如表 6 所示.

表 6　实验数据规模

Table 6　Scale of experiment data

数据类型	数据规模
同义词表 Key 总数	4 293 406
同义词表 Value 总数	1 948 277
歧义词表 Key 总数	213 764
歧义词表 Value 总数	2 354 687
实体总数	4 369 348

实验中选取 NLPCC 2014[9] 中文实体链接评测任务训练集中包含的 177 条中文微博数据和人工标注的 400 条新浪微博数据 (从抓取的 10 000 条数据中随机抽取标注) 作为训练数据, 从剩余的 9 600 条微博中随机抽取 100 条标注作为验证集, 测试集使用 NLPCC 2014 官方提供测试集, 共包含 1 152 个实体指称.

3.2 实验设计

1) 选取链接效果好的算法进行对比. 通过对比验证本文算法是否有效. 该部分选取 NLPCC 2014 评测中的最优方法 (NLPCC)[19]、基于上下文概率模型的实体链接方法 (EF*)[20] 以及基于维基百科和搜索引擎 (CMEL)[21] 方法进行对比.

NLPCC 采用百度百科分类属性和实体流行度相结合的方法进行消歧, EF* 在概率模型基础上添加平滑方法, CMEL 采用结合维基百科实体描述页面和搜索引擎结果相结合的方法进行实体消歧.

实验过程中, 对所有方法均采用相同的资源模板和预处理. 首先用训练集数据训练得到多分类回归模型, 再用验证集进行模型调参, 得到最优权重 $\alpha = 1.4$, $\lambda = 0.6$, 最后在测试集进行方法验证. 选取准确率 (Precision)、召回率 (Recall) 以及 F1 值作为评价指标. 其中, in-KB 部分表示知识库中已收录实体的准确率, NIL 表示未收录到知识库中的实体链接准确率. 对比实验结果如表 7 和表 8 所示.

表 7 in-KB 实验结果

Table 7 Results of in-KB

系统	准确率	召回率	F1 值
NLPCC	0.7927	0.8488	0.8198
SCWE + EF	**0.8137**	**0.8593**	**0.8358**
EF*	0.7641	0.8142	0.7884
CMEL	0.7951	0.8345	0.8143

表 8 NIL 实验结果

Table 8 Results of NIL

系统	准确率	召回率	F1 值
NLPCC	0.9024	0.8653	0.8835
SCWE + EF	**0.9144**	**0.8763**	**0.8949**
EF*	0.8871	0.8648	0.8758
CMEL	0.8543	0.8694	0.8461

实验结果中, 粗体部分表示本文方法. 实验表明, 本文方法在准确率、召回率方面均明显优于其他三种算法, 特别是准确率有显著提升. 而本文方法与 NLPCC、EF* 和 CMEL 方法的主要区别在于本文方法加入词向量语义分类特征, 实验结果的提升表明词向量语义分类特征是有效的. 例如, 评测数据样例 "好怀念当时的那支队伍啊! 弗朗西斯、麦迪、巴蒂尔、大姚、斯科拉、穆托姆博、诺瓦克", NLPCC 和 EF* 将 "大姚" 链接至 "姚晨", 而本文方法将该指称链接至 "姚明".

2) 不同比重下的词向量语义分类特征 (SCWE) 对链接结果的影响. 本文采用 SCWE 和实体流行度特征 (EF) 两个特征, 选取不同的 λ 进行实验结果比对. 结果如表 9 和表 10 所示.

表 9 in-KB 实验结果

Table 9 Results of in-KB

λ 值	准确率	召回率	F1 值
0	0.7532	0.8016	0.7766
0.2	0.7621	0.8158	0.7880
0.4	0.7943	0.8375	0.8153
0.6	**0.8137**	**0.8593**	**0.8358**
0.8	0.8032	0.8432	0.8227
1.0	0.7983	0.8488	0.8228

表 10 NIL 实验结果

Table 10 Results of NIL

λ 值	准确率	召回率	F1 值
0	0.8432	0.8532	0.8482
0.2	0.8643	0.8713	0.8678
0.4	0.8917	0.8732	0.8824
0.6	**0.9148**	**0.8762**	**0.8951**
0.8	0.9032	0.8754	0.8891
1.0	0.9013	0.8743	0.8876

从实验结果看, 当 $\lambda = 0$ 时, 链接方法中只选取了实体流行度作为特征, 此时 F1 值最低. 随着 λ 的增长, F1 值也随之增长. 在 $\lambda = 0.6$ 时, F1 值最高. $\lambda > 0.6$ 之后, F1 值开始降低. 表明本文构建的实体链接方法效果的提升依赖于词向量语义分类特征. 为了更清晰地表示 F1 值与参数 λ 之间的关系, 构建图 4.

3) 词向量语义分类特征与聚类特征数目 k 的关系. 只采用 SCWE 模型进行实体链接, 选取不同的 k 值来观测模型与 k 之间的关系. 如图 5 所示, 当 $k = 10$ 时, 模型取得最高的 F1 值, 在 $5 \sim 15$ 之间, k 值的变化对模型的预测效率影响不大. 但 $k = 20$ 时, SCWE 的 F1 值大幅度下降. 通过对评测数据中每条微博含有的名词数目进行统计, 发现每条微博中

平均有 7.91 个名词. 分析认为当 $k = 20$ 时 F1 值下降是由于特征选取过多, 训练数据稀疏所致.

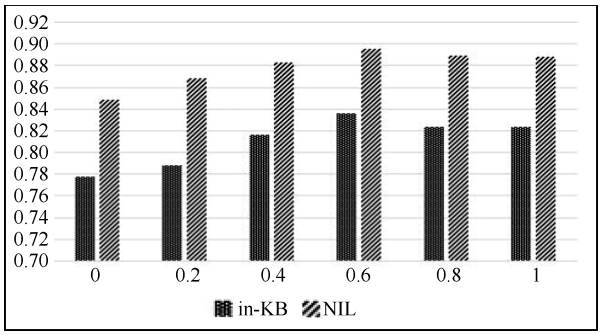


图 4 本文方法在不同参数 λ 下的 F1 值

Fig. 4 F1 scores of the combined measure with the λ parameter

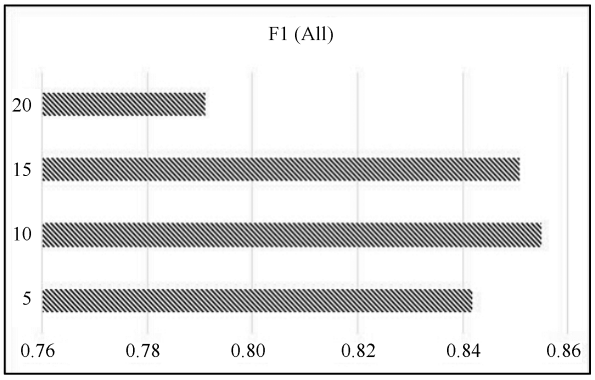


图 5 SCWE 在不同参数 k 下的 F1 平均值

Fig. 5 F1 scores of SCWE with the k features

4 结论与展望

基于微博中名词位于相近的语义空间的假设, 本文提出了利用词向量语义分类对微博实体进行语义消歧的思路, 设计了完整的实体链接方法, 并在 NLPCC 2014 发布的评测数据上进行验证. 实验结果表明使用本文提出的基于词向量语义分类的实体链接方法, 链接效果优于 NLPCC 已公开的最好结果, 链接准确率有显著提升. 后续工作主要集中在两点, 一是结合词向量和图模型进行实体链接, 二是探索不同的多分类模型在实体链接中的应用.

References

1 Chinese Microblog Service. Sina Weibo User Development Report in 2014 [Online], available: http://www.199it.com/archives/324955.html. November 24, 2015
(中国微博服务. 2014 年新浪微博用户发展报告 [Online], available: http://www.199it.com/archives/324955.html. November 24, 2015)

2 Guo Y H, Qin B, Liu T, Li S. Microblog entity linking by leveraging extra posts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA: Association for Computational Linguistic, 2013. 863−868

3 Yang Jin-Feng, Yu Qiu-Bin, Guan Yi, Jiang Zhi-Peng. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, 2014, **40**(8): 1537−1562
(杨锦锋, 于秋滨, 关毅, 蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, **40**(8): 1537−1562)

4 Shen W, Wang J Y, Han J W. Entity linking with a knowledge base: issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 2015, **27**(2): 443−460

5 Jiang L, Yu M, Zhou M, Liu X H, Zhao T J. Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: 2011. 151−160

6 Shen W, Wang J Y, Luo P, Wang M. Linking named entities in tweets with knowledge base via user interest modeling. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2013. 68−76

7 Liu X H, Li Y T, Wu H C, Zhou M, Wei F R, Lu Y. Entity linking for tweets. In: Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, 2013. 1304−1311

8 Odbal, Wang Zeng-Fu. Emotion analysis model using compositional semantics. *Acta Automatica Sinica*, 2015, **41**(12): 2125−2137
(乌达巴拉, 汪增福. 一种基于组合语义的文本情绪分析模型. 自动化学报, 2015, **41**(12): 2125−2137)

9 NLPCC [Online], available: http://tcci.ccf.org.cn/conference/2014/pages/page04_sam.html. October 31, 2015

10 Hachey B, Radford W, Nothman J, Honnibal M, Curran J R. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 2013, **194**: 130−150

11 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.

12 Hartigan J A, Wong M A. Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society — Series C (Applied Statistics)*, 1979, **28**(1): 100−108

13 Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 2014, **15**: 3133−3181

14 Mao Yi, Chen Wen-Lin, Guo Bao-Long, Chen Yi-Xin. A novel logistic regression model based on density estimation. *Acta Automatica Sinica*, 2014, **40**(1): 62−72
(毛毅, 陈稳霖, 郭宝龙, 陈一昕. 基于密度估计的逻辑回归模型. 自动化学报, 2014, **40**(1): 62−72)

15 Zhou Xiao-Jian. Enhancing ε-support vector regression with gradient information. *Acta Automatica Sinica*, 2014, **40**(12): 2908−2915
(周晓剑. 考虑梯度信息的 ε-支持向量回归机. 自动化学报, 2014, **40**(12): 2908−2915)

16 King G, Zeng L C. Logistic regression in rare events data. *Political Analysis*, 2001, **9**(2): 137−163

17 Guo Y H, Qin B, Li Y Q, Liu T, Lin S. Improving candidate generation for entity linking. In: Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems. Salford, UK: Springer, 2013. 225−236

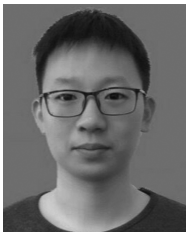
18 Wikipedia [Online], available: http://download.wikipedia.comzhwikilate-stzhwiki-latest-pages-articles.xml.bz2. October 31, 2015

19 Zhu Min, Jia Zhen, Zuo Ling, Wu An-Jun, Chen Fang-Zheng, Bai Yu. Research on entity linking of Chinese microblog. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2014, **50**(1): 73−78
(朱敏, 贾真, 左玲, 吴安峻, 陈方正, 柏玉. 中文微博实体链接研究. 北京大学学报 (自然科学版), 2014, **50**(1): 73−78)

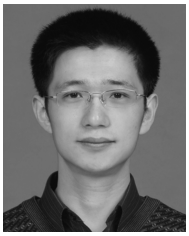
20 Guo Yu-Hang. Research on Context-based Entity Linking Technique [Ph. D. dissertation], Harbin Institute of Technology, China, 2014.
(郭宇航. 基于上下文的实体链指技术研究 [博士学位论文], 哈尔滨工业大学, 中国, 2014.)

21 Meng Z Y, Yu D, Xun E D. Chinese microblog entity linking system combining Wikipedia and search engine retrieval results. In: Proceedings of the 3rd CCF Conference on Natural Language Processing and Chinese Computing. Berlin Heidelberg: Springer, 2014. 449−456

冯 冲 北京理工大学计算机学院副研究员. 2005 年获中国科学技术大学计算机科学系博士学位. 主要研究方向为自然语言处理, 信息抽取, 机器翻译. 本文通信作者.
E-mail: fengchong@bit.edu.cn
(**FENG Chong** Associate professor at the College of Computer Science and Technology, Beijing Institute of Technology. He received his Ph. D. degree from the Department of Computer Science, University of Science and Technology of China in 2005. His research interest covers natural language processing, information extraction, and machine translation. Corresponding author of this paper.)



石 戈 北京理工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, 实体链接, 问答系统.
E-mail: shige713@126.com
(**SHI Ge** Ph. D. candidate at the College of Computer Science and Technology, Beijing Institute of Technology. His research interest covers natural language processing, entity linking, and question answering system.)



郭宇航 北京理工大学计算机学院讲师. 2014 年获哈尔滨工业大学计算机科学与技术学院博士学位. 主要研究方向为自然语言处理, 信息抽取, 机器翻译.
E-mail: guoyuhang@bit.edu.cn
(**GUO Yu-Hang** Lecturer at the College of Computer Science and Technology, Beijing Institute of Technology. He received his Ph. D. degree from Harbin Institute of Technology in 2014. His research interest covers natural language processing, information extraction, and machine translation.)

龚 静 北京理工大学计算机学院硕士研究生. 主要研究方向为自然语言处理, 机器翻译, 问答系统.
E-mail: gongjing@bit.edu.cn
(**GONG Jing** Master student at the College of Computer Science and Technology, Beijing Institute of Technology. Her research interest covers natural language processing, machine translation, and question answering system.)

黄河燕 北京理工大学计算机学院教授. 1989 年获中国科学院计算技术研究所计算机科学与技术博士学位. 主要研究方向为自然语言处理和机器翻译社交网络与信息检索, 智能处理系统.
E-mail: hhy63@bit.edu.cn
(**HUANG He-Yan** Professor at the College of Computer Science and Technology, Beijing Institute of Technology. She received her Ph. D. degree from the Institute of Computing Technology, Chinese Academy of Sciences. Her research interest covers natural language processing, machine translation, social network, information retrieval, and intelligent processing system.)