文章编号:1671-9352(2017)09-0007-06 DOI:10.6040/j.issn.1671-9352.1.2016.PC7

基于COAE2016数据集的中文实体关系抽取算法研究

孙建东,顾秀森,李彦,徐蔚然*
(北京邮电大学模式识别与智能系统实验室,北京 100876)

摘要: 实体关系抽取是知识图谱技术的重要环节之一。英文实体关系抽取的研究已经比较成熟,相比之下,中文实体关系抽取的发展却并不理想。由于相关语料的匮乏,中文实体关系抽取的发展受到了一定的限制。针对这一问题,COAE2016在任务三中提出了中文实体关系抽取任务。通过分别使用了基于模板、基于SVM与基于CNN的实体关系抽取算法解决了这一问题,并根据其在COAE2016任务三的评测数据集上的效果,对比分析了三种实体关系抽取算法的优缺点。实验证明,基于SVM的算法和基于CNN的算法均在评测数据集上表现出了良好的效果。

关键词: 关系抽取;模板匹配;SVM;CNN
中图分类号: TP391 **文献标志码:** A
引用格式: 孙建东,顾秀森,李彦,等.基于COAE2016数据集的中文实体关系抽取算法研究[J].山东大学学报(理学版),2017,52(9):7-12,18.

Chinese entity relation extraction algorithms based on COAE2016 datasets

SUN Jian-dong, GU Xiu-sen, LI Yan, XU Wei-ran*
(Beijing University of Posts and Telecommunications, Lab of Pattern Recognition and Intelligent System, Beijing 100876, China)

Abstract: Entity relation extraction is one of the important procedures of knowledge graph technology. Research on entity relation extraction in English is comparatively developed. By contrast, the development of Chinese entity relation extraction is not ideal, and it is mainly because the lack of corpus. In order to solve this problem, COAE2016 proposes a Chinese entity relation extraction task in task 3. In this paper, we use three algorithms to solve the problem: a pattern based algorithm, a SVM based algorithm and a CNN based algorithm respectively. Then, we analyze the advantages and the disadvantages of the three algorithms according to the effects of the dataset in COAE2016 Experiments show that the SVM based algorithm and the CNN based algorithm are useful to extract entity relation.

Key words: feature extraction; pattern match; SVM; CNN

0 引言

近年来,结构化知识图谱(structured knowledge graph)技术的发展已引起了研究人员的广泛关注。相比于英文中已有的大型知识库,中文知识图谱的发展相对较慢。作为构建知识图谱的基本环节之一,实体关系抽取的重要性不言而喻。然而,由于中文实体关系抽取语料的匮乏,中文实体关系抽取的发展受到了一定的限制。针对该问题,COAE2016提出了中文实体关系抽取任务。传统的实体关系抽取算法主要分为两大类:分别是基于模板的算法与基于特征的算法。而随着深度学习在自然语言处理领域的应用,出现了基于深度学习的实体关系抽取算法。基于COAE2016任务三的数据集,本文分别使用了基于模板、基于支持向量机(support vector machine, SVM)以及基于卷积神经网络(Convolution Neural Network, CNN)的算法解决了中文实体

收稿日期:2016-11-25; 网络出版时间:2017-06-14 09:04
网络出版地址:http://kns.cnki.net/kcms/detail/37.1389.N.20170614.0904.020.html
基金项目:111计划资助项目(B08004);国家自然科学基金资助项目(61300080,61273217,61671078);国家教育部博士点基金资助项目(20130005110004)
作者简介:孙建东(1994—),男,硕士研究生,研究方向为自然语言处理. E-mail:sunjd@bupt.edu.cn
* 通讯作者:徐蔚然(1975—),男,副教授,硕士生导师,研究方向为自然语言处理. E-mail:xuweiran@bupt.edu.cn

关系抽取问题。进而,根据其在 COAE2016 任务三的评测数据集上的表现,对比分析了 3 种算法的特点。实验证明,基于 SVM 的算法与基于 CNN 的算法均在 COAE2016 任务三的评测数据集上表现出了良好的效果。

1 相关工作

1.1 实体关系抽取概述

实体关系抽取是指从文本中获取指定实体之间的关系,即预先给定文本与指定的实体,根据文本对实体之间的关系进行抽取。实体关系使用 3 元组(triple)表示〈subject, relation, object〉,其中,subject 表示关系的主语,object 表示关系的宾语,relation 表示对应的关系。例如,在句子"李忠诚又译为〈e1〉李忠成〈/e1〉(1985 年 12 月 19 日-),原名大山忠成出生于〈e2〉日本东京都〈/e2〉"中,"李忠成"表示关系的主语,"日本东京都"表示关系的宾语,而实体关系抽取的目的就是抽取出这两个实体间的关系"出生地",对应的 3 元组是〈李忠成,出生地,日本东京都〉。

在实体关系抽取任务的基础上,COAE2016 提出了具有特定关系的实体对的抽取任务。例如,从句子"李忠诚又译为李忠成(1985 年 12 月 19 日-),原名大山忠成出生于日本东京都"中,我们可以抽取出具有"出生日期"关系的实体对〈subject = "李忠成",object = "1985 年 12 月 29 日"〉与具有"出生地"关系的实体对〈subject = "李忠成",object = "日本东京都"〉。对具有特定关系的实体对进行抽取是中文知识图谱搭建的重要步骤之一,对中文知识图谱技术的发展具有重要的现实意义。

1.2 中文实体关系抽取研究现状

中文实体关系抽取是中文自然语言处理领域的一个重要课题。近年来,随着知识图谱技术的发展,中文实体关系抽取已经引起了广泛的关注。传统的中文实体关系抽取算法[1]主要分为两类:一是基于模板匹配的中文实体关系抽取算法,二是基于特征的中文实体关系抽取算法。

使用模板进行实体关系抽取是一种最为基础的方法[1]。该方法通过运用语言学知识,在抽取之前就构造出若干基于词语、基于词性或基于语义的模板存储起来,在进行实体关系抽取时,将经过处理的包含实体对的语句与该模板进行匹配,一旦匹配成功,就认为该实体对具有对应模板的关系属性。

另外一类解决实体关系抽取问题的传统算法主要是基于特征。该类算法将实体关系抽取问题看作是一个分类问题,继而搭建关系分类器进行实体间的关系抽取。如文献[2]中使用基于特征选择的方法进行实体关系抽取,文献[3]中使用了 Winnow 和 SVM 来进行实体关系抽取,文献[4]中使用对序列核进行 k 均值聚类的方法进行实体关系抽取,文献[5]中采用将极大熵算法和 Boot-strapping 算法相结合的方法进行实体关系抽取。这些方法都在一定程度上取得了较好的效果。一般来说,基于特征的算法其性能依赖于人工选择出的特征的适用程度。当选择出的特征能够为关系抽取提供足够的信息时,这些算法会表现出良好的效果;反之,当选择出的特征不足以为实体关系抽取提供足够的信息时,这些算法的性能会变得很差。此外,使用人工选择的特征在引入与实体关系有关的信息的同时,也会引入与实体关系无关的信息。在进行实体关系抽取时,这些无关信息就变成了噪声。随着样本数据集的增大,这些噪声会严重影响实体关系抽取性能的提升。因而,基于特征的实体关系抽取算法虽然在样本数据集较小时可以取得很好的效果,但随着样本数据集的增大,其性能提升却并不明显。

1.3 基于 CNN 的实体关系抽取

近年来,随着深度学习技术[6]的发展,人们希望可以通过对特征进行表示学习,以取代传统的人工选择特征的方式,从而减少特征工程所带来的复杂性与不确定性。作为一种极为有效的深度网络,CNN 已被广泛地应用在了图像处理领域[7]与语音处理领域[8]。

CNN 是一种比较高效的深度神经网络,其与最基本的多层神经网络的不同之处在于,它是以部分连接的方式来提取局部特征的。此外,CNN 还通过权值共享解决了传统多层神经网络中参数过多、容易造成过拟合的问题。CNN 每一层的输出可以看作是对前一层的更为抽象的表示。通过学习,CNN 可以提取出对目标任务有用的信息,并过滤与目标任务无关的信息。卷积神经网络的这一特性可以有效地避免人工选择特征所带来的问题,而且,随着训练数据集的增加,CNN 的性能往往会有很大的提升。

随着深度神经网络在自然语言处理领域的应用[9],CNN 已经被广泛地应用于英文的实体关系抽取[10-11]之中。然而,CNN 在中文关系抽取中的应用还处于初始阶段。本文基于 COAE2016 任务三所提供的评测数据集,使用基于 CNN 的算法进行了中文实体关系抽取。实验证明,基于 CNN 的中文

实体关系抽取算法在 COAE2016 任务三的评测数据集上取得了良好的效果。

总的来说，本文的创新主要有以下两点:

1) 基于 COAE2016 任务三的评测数据集，分别使用了基于模板匹配、基于 SVM 与基于 CNN 的算法完成了中文实体关系抽取任务与具有特定实体对的关系抽取任务，并取得了良好的效果;

2) 根据 3 种中文实体关系抽取算法在 COAE2016 任务三的评测数据集上的实验结果，对比分析了 3 种算法的特点。

2　方法

2.1　基于模板的实体关系抽取算法

基于模板的实体关系抽取算法主要通过模板匹配的方式抽取出对应的实体关系。本文中，首先对句子进行句法解析①，然后找出实体对之间的最短依存路径，通过最短依存路径上是否可以匹配到关系对应的模板来判断是否存在既定关系。

具体的过程如图 1:

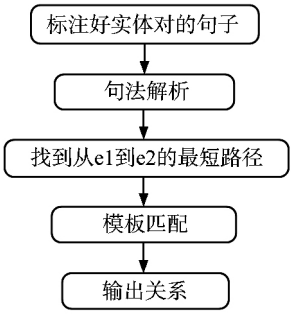

图 1　基于模板的实体关系抽取

Fig.1　Pattern-based entity relation extraction

举例来说，首先对“李忠诚又译为〈e1〉李忠成〈/e1〉(1985 年 12 月 19 日-)，原名大山忠成出生于〈e2〉日本东京都〈/e2〉”进行句法解析，然后从该句对应的句法依存树中获取实体间的最短依存路径:“李 忠 成-VOB-为-CMP-译-COO-出 生-CMP-于-POB-日本东京都”，最后，根据该路径上是否可以匹配到模板词来判断实体间是否存在对应的关系。在该最短路径上出现了对应于关系“出生地”的模板词“出生”，那么该实体对之间的关系即可判定为模板词“出生”所对应的“出生地”关系。

2.2　基于特征的实体关系抽取算法

基于特征的算法是目前最为普遍的实体关系抽取算法。本文主要使用基于支持向量机(Support Vector Machine，SVM)的算法进行实体关系的抽取。具体流程如图 2:

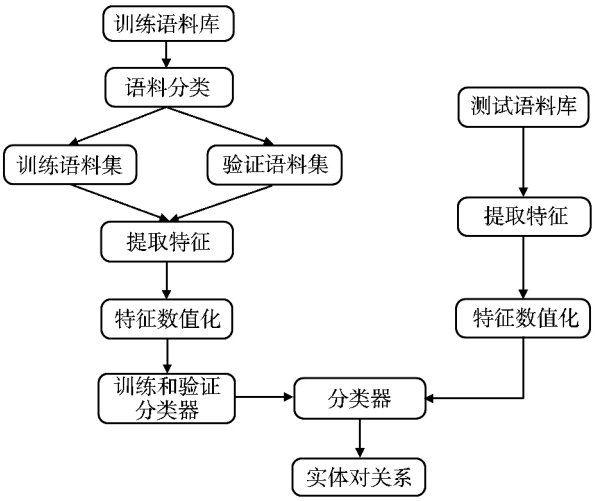

图 2　基于特征的实体关系抽取

Fig.2　Feature-based entity relation extraction

具体来说，本文选取了实体顺序、实体间距离、实体对上下文等特征，这些特征的含义为:

1) 实体顺序表示实体 1 与实体 2 在句中的位置关系，若实体 1 在实体 2 之前，则表示为 2，否则为 1;

2) 实体间距离是指对句子分词②后实体 1 与实体 2 之间的词语数量;

3) 实体对上下文特征指的是实体对上下文的位置特征，例如“李忠诚又译为〈e1〉李忠成〈/e1〉(1985 年 12 月 19 日-)，原名大山忠成出生于〈e2〉日本東京都田無市〈/e2〉”，这句话对应的上下文特征是“v_bm1l 译_bm1l p_bm1f 为_bm1f v_wbs2 出_wbs2 v_wbsl 生于_wbs1”。其中“v_bm1l 译_bm1l”指实体 e1 前第二个词是“译”，标记为 bml1，所对应的词性为 v，“p_bm1f 为_bm1f”指实体 e1 前第一个词是“为”，标记为 bmlf，其词性为 p，“v_wbs2 出_wbs2”指实体 e2 前第二个词语是“出”，位置标记为 wbs2，其词性是 v，“v_wbs1 生于_wbs1”指实体 e2 前第一个词是“生于”，其位置标记为 wbs1，词性是 v。由于实体 e2 已经处于句尾，所以实体对的上下文特征不再提取实体 e2 后的特征。对于实体处于句首的情况，同样不提取第一个实体之前的位置特征。

在提取了以上特征之后，使用线性 SVM 分类器对实体关系进行分类。该分类器由 K 个二类分类器构成，每个二类分类器对应于一种实体对关系。其中，K 表示二类分类器的个数，即实体对关系的个数。在分类时，首先将输入的实体对表示为相应的特征向量，然后将该特征向量输入到 K 个二类分类

① 使用语言技术平台(LTP)提供的句法解析工具

② 使用 python 自然语言处理工具包结巴分词

器中,并由每个二类分类器给出一个对应的分值 s_k,并取 $s=\max\limits_{1\leqslant k\leqslant K} s_k$ 作为该数据点的最终得分。当 s 高于阈值时,判定实体间关系为 $k=\arg\max\limits_{1\leqslant k\leqslant K} s_k$;否则,判定该实体对之间的关系为"其他关系"。

2.3 基于 CNN 的实体关系抽取算法

随着深度神经网络的发展及其在自然语言处理领域的应用,人们趋于使用深度网络学习特征的方法替代传统的人工选择特征方式。本文采用了基于 CNN 的方法来完成这一任务并进行实体关系抽取。首先,使用工具将句子分词为$\{w_1,w_2,\cdots,w_n\}$,然后将每个词表示为向量[①]$s(w_i)\in\mathbf{R}^{d_1}$,此时,每个句子对应于一个 $s\in\mathbf{R}^{n\times d_1}$ 的词向量矩阵,同样,将每个词到实体 e1 与 e2 的距离分别表示为 p_1 维与 p_2 维的向量。最后将句子对应的词向量矩阵与位置向量矩阵连接起来,作为该句子的特征向量表示 $V\in\mathbf{R}^{n\times d}$,其中 $d=d_1+p_1+p_2$。输出关系所对应的概率使用 softmax 函数表示:

$$p(y)=\frac{e^y}{\sum_{y'}e^{y'}} \tag{1}$$

$$y=\boldsymbol{h}^{\mathrm{T}}\boldsymbol{w}_y \tag{2}$$

其中,$\boldsymbol{h}$ 表示最后一个隐层的输出,$\boldsymbol{w}_y$ 表示类别 y 对应的权值向量。具体结构如图 3 所示:

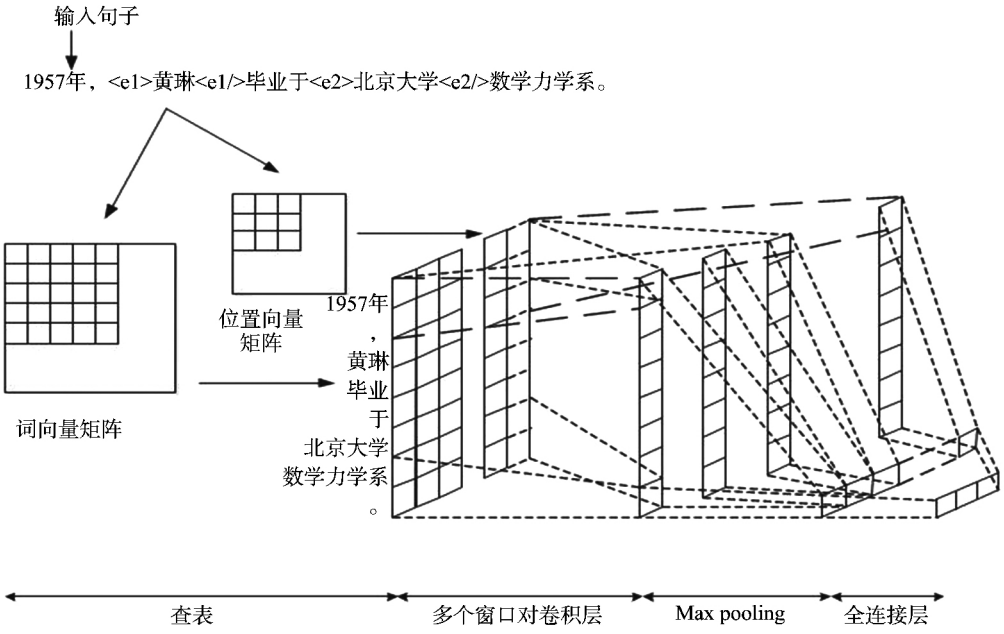

图3 基于 CNN 的实体关系抽取算法

Fig.3 CNN-based entity relation extraction

在基于 CNN 的算法中,输出表示为一个 K 维向量 $\boldsymbol{y}\in(0,1)^K$,其中 $\boldsymbol{y}$ 的第 k 维的值 $1\leqslant k\leqslant K$ 的值 $\boldsymbol{y}_k$ 表示实体关系为关系 k 的概率,且满足

$$\sum_k \boldsymbol{y}_k=1 \tag{3}$$

在 CNN 的训练中,使用了 dropout 来防止过拟合。dropout 是指在模型训练时以一定的概率让网络中某些隐藏层节点暂时不工作,不工作的那些节点可以暂时认为不是网络结构的一部分,但是保留其权值,当下一次输入后重新分配这些不工作的结点。实验证明,使用了 dropout 的 CNN 算法在实体关系抽取和具有特定实体关系的实体对抽取上都取得了较好的效果,在语料充足的情况下,基于 CNN 的算法的性能仍有很大的提升空间。

2.3 具有特定关系的实体对抽取

对于中文知识图谱的搭建,对具有特定关系的实体对进行抽取也是一个重要的环节。针对这一问题,COAE2016 任务三提出了具有特定关系的实体对的抽取任务。

本文采取在中文实体关系抽取之前添加规则的方法进行实体对的抽取。首先,使用命名实体识别工具识别出句子中的所有实体并保留其实体类别,这些实体包括人名、地名、组织名、日期与数量词;然后,根据具体关系选择出所有候选实体对,例如对于关系"人物的出生日期",那么该候选实体对限定 e1 为句中出现的人名,e2 为句中出现的日期;最后,使用上述 3 种关系抽取算法进行关系抽取,并将符合特定关系的实体对抽取出来。

3 实验

3.1 数据集

本次实验使用的数据集是由 COAE2016 任务

① 训练工具 word2vec(Mikolov 等,2013) CBOW 模型。

三提供的评测数据集。其中,共有 10 类候选实体关系,具体见表 1:

表 1　COAE2016 任务三的候选实体关系

Table 1　The entity relations in COAE2016 Task 3

标号	关系名称
cr2	人物的出生日期
cr4	人物的出生地
cr16	人物的毕业院校
cr20	人物的配偶
cr21	组织机构的子女
cr28	组织机构的高管
cr29	组织机构的员工数
cr34	组织机构的创始人
cr35	组织机构的成立时间
cr37	组织机构的总部地点

COAE2016 任务三提供了 988 条训练数据,937 条测试数据。在测试数据中,有 483 条为已人工标注好指定实体对的文本,用于实体关系抽取任务; 454 条为未进行标注的文本,用于具有特定关系的实体对抽取任务。

3.2　评测结果与分析

本次评测使用的是由 COAE2016 任务三提供的 3 个技术指标,分别是准确率 P,召回率 R 和 $F1$ 值。对于每个指标,分别考虑了其宏平均的性能与微平均的性能。

3.2.1　中文实体关系抽取任务结果分析

实体关系抽取任务的评测结果如表 2 所示:

表 2　实体关系抽取任务结果(%)

Table 2　Results of entity relation extraction (%)

方法	Micro	Macro		
	$P(R,F1)$	P	R	$F1$
模板匹配	24.64	55.24	16.96	23.97
SVM	77.23	76.78	64.70	66.29
CNN	64.80	60.83	55.60	56.69
平均水平	66.88	69.26	61.69	62.74

注: 微平均(Micro) 是以整个数据集为一个评价单元,计算整体性能的评价指标; 宏平均(Macro) 是以每个领域为一个评价指标,计算参评系统在该领域中的评价指标,最后计算所有领域上各指标的平均值

实验结果说明,使用模板匹配方法的实体关系抽取算法虽然实现起来比较简单,但其性能也是最差的,其原因主要有以下 3 点:

一是模板覆盖度不足,很多关系对应的模板词数量很少,这会直接影响到关系的匹配结果。如对于关系“组织机构的总部地点”对应的模板为〈‘总部’〉,仅含有一个模板词,那么当实体间的最短依存路径上无法匹配到该词时,就判定实体对不具有该关系,这显然会对该类关系上的抽取效果产生很大的影响。当然,对于一些模板词较多的关系,如关系“人物的配偶”,对应的模板为〈‘夫人’, ‘妻子’, ‘老婆’, ‘太太’, ‘其妻’, ‘原配’, ‘前妻’, ‘结婚’, ‘离婚’〉,其抽取效果要明显好于模板覆盖度小的关系,但是,由于实体关系表达的多样性,该模板的覆盖度仍有缺陷,诸如“丈夫”, “其夫”, “媳妇”等可以表达“人物的配偶”关系的模板词,仍没有被该模板所包含;

二是基于模板匹配的关系抽取算法在很大程度上依赖于句法解析的结果,当模板词落在实体对之间的最短依存路径之外时,同样不能匹配到对应的关系;

三是基于模板的算法受人工的影响很大。因为在该算法中使用的模板往往是根据相关的知识进行构建的。人工构建模板的缺点在于人们往往只能选择一些明显的特征来构建模板,如“夫人”一词可以直接代表实体间具有“人物的配偶”关系。对于一些隐藏的特征,如“实体间距离”,人们很难直接发现其与实体关系的关联。

使用了多种特征的 SVM 算法取得了最好的效果,在人工选择的特征能够为实体关系抽取提供足够的信息时,基于特征的算法可以取得很好的效果。实验结果说明,实体对的位置关系特征、实体对距离特征与实体上下文特征为实体关系抽取提供了足够的信息,这也充分说明了基于 SVM 的关系分类算法可以被有效地用于中文关系抽取问题。

使用了词语语义信息和词到实体的距离信息作为初始输入的 CNN 算法的实验结果略低于平均水平。虽然,基于 CNN 的算法可以有效地表示出与目标任务有关的特征,且不会受到人工选择特征的影响,但其性能却要受到训练集大小的限制。由于 COAE2016 任务三提供的训练数据集相对较小,基于 CNN 的算法的优势并不能体现出来。当训练数据集增大时,基于 CNN 的算法性能仍有很大的上升空间。

3.2.2　具有特定关系的实体对抽取任务

具有特定关系的实体对抽取任务的评测结果如表 3 所示:

表 3　具有特定关系的实体对抽取任务结果(%)

Table 3　Results of entity pair extraction of special relation (%)

方法	Micro	Macro		
	$P(R,F1)$	P	R	$F1$
模板匹配	31.50	56.51	24.21	29.54
SVM	77.53	49.40	40.02	41.22
CNN	66.30	61.20	48.16	52.76
平均水平	65.55	62.46	45.79	48.94

可以看出,在具有特定关系的实体对抽取任务中,基于模板的方法与实体关系抽取任务的结果基本一致。而基于SVM的算法的微平均结果与实体关系抽取的结果基本一致,宏平均结果则出现了较大的差异,这是因为基于SVM的算法在COAE2016任务三测试数据集中的常见类别上的表现较佳,而在一些稀有类别上的表现较差。基于CNN的算法在小类别和大类别上的表现比较均衡,其宏平均结果与微平均结果均超过了COAE2016任务三的平均水平。随着训练数据集的增大,基于CNN的算法的性能仍有很大的上升空间。

4 结论

基于COAE2016任务三所提供的评测数据集,本文实现了基于模板、基于SVM以及基于CNN的3种中文实体关系抽取算法,并解决了具有特定关系的实体对抽取问题。实验结果说明,基于SVM的算法与基于CNN的算法在COAE2016任务三的数据集上表现出了良好的性能。这说明基于SVM的算法与基于CNN的算法可以被有效地用于中文实体关系的抽取,而基于模板的方法性能相对较差,在具体应用中,可以将该算法与其余两种算法结合使用。

根据3种算法在COAE2016评测数据集上的表现,可以得出以下结论:

1) 在数据集的大小有限时,基于SVM的算法的性能要优于其他两种算法。因为SVM算法可以充分利用人工选取的特征,因而可以获得良好的效果。相比之下,基于模板的方法只能利用一些与分类直接相关的特征,对于一些隐含的特征,人工很难发现其与实体关系的联系;而基于CNN的算法在训练数据集较小时并不能对初始输入进行最为有效的特征表示,因而在数据集较小时,其效果要略差于基于SVM的算法;

2) 在训练数据集足够大时,我们需要考虑使用基于CNN的算法。此时,由于人工选择的特征中含有与关系抽取无关的信息,SVM的算法性能受到了一定的限制。而基于CNN的算法可以根据目标任务对初始输入进行更为有效的特征表示,故而随着数据集的增大,基于CNN的算法的性能会有显著提升。

本文通过对3种中文实体关系抽取算法的对比与分析,论证了基于SVM的算法与基于CNN的算法在中文实体关系抽取上的有效性。随着深度学习在自然语言处理领域的进一步应用,我们将尝试使用其他的深度网络进行中文实体关系的抽取。同时,我们仍将对基于CNN的算法做进一步的优化,以提高该算法的有效性。

参考文献:

[1] 徐健,张智雄,吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术,2008(8):18-23.
XU Jian, ZHANG Zhixiong, WU Zhenxin. Review on techniques of entity relation extraction [J]. New Technology of Library and Information Service, 2008(8):18-23.

[2] 毛小丽,何中市,邢欣来,等. 基于特征选择的实体关系抽取[J]. 计算机应用研究,2012,29(2):530-532.
MAO Xiaoli, HE Zhongshi, XING Xinlai, et al. Entity relation extraction based on feature selection [J]. Application Research of Computers, 2012, 29(2):530-532.

[3] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报,2004,19(2):1-6.
CHE Wanxiang, LIU Ting, LI Sheng. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2004, 19(2):1-6.

[4] 刘建舟,邵雄凯. 一种改进的中文实体关系抽取方法[J]. 软件导刊,2011,10(4):27-29.
LIU Jianzhou, SHAO Xiongkai. An improved method of chinese entity relation extraction [J]. Software Guide, 2011, 10(4):27-29.

[5] 张素香,文娟,秦颖,等. 实体关系的自动抽取研究[J]. 哈尔滨工程大学学报,2006,27(S1):370-373.
ZHANG Suxiang, WEN Juan, QIN Ying, et al. Study about automatic entity relation extraction [J]. Journal of Harbin Engineering University, 2006, 27(S1):370-373.

[6] LECUN Yann, BENGIO Yoshua, HINTON Geoffrey. Deep learning [J]. Nature.2015, 521(7553): 436-444.

[7] KRIZHEVSKY Alex SUTSKEVER Ilya, HINTON Geoffrey. ImageNet classification with deep convolutionalneural networks [J]. International Conference on Neural Information Processing Systems, 2012, 25(2):1097-1105.

[8] ZHANG Shiliang, LIU Cong, JIANG Hui, et al. Feedforward sequential memory networks: a new structure to learn long-term dependency [J]. Computer Science, 2015, arXiv: 1510.02693.

[9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2014(3):1137-1155.

[10] HASHIMOTO K, STENETORP P, MIWA M, et al. Task-oriented learning of word embeddings for semantic relation classification [J]. Computer Science,2015, arXiv: 1503.00095.

(下转第18页)

系统有所提高。

未来的工作将重点聚焦于未能成功识别的意图边界,尝试提出新的特征及模型,比如可以考虑深度学习的方法,利用人工神经网络或者引入知识库来解决这个问题。

参考文献:

[1] SILVERSTEIN C, MARAIS H, HENZINGER M, et al. Analysis of a very large web search engine query log [J]. SIGIR Forum, 1999, 33(1): 6-12.

[2] LI Yanan, ZHANG Sen, WANG Bin, et al. Characteristics of chinese web searching: A large-scale analysis of chinese query logs [J]. Journal of Computational Information Systems, 2008, 4(3): 1127-1136.

[3] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析 [J]. 中文信息学报, 2007, 21(1): 109-114.
YU Huijia, LIU Yiqun, ZHANG Min, et al. Research in search engine user behavior based on log analysis [J]. Journal of Chinese Information Processing, 2007, 21(1): 109-114.

[4] BRODER A. A taxonomy of web search [J]. SIGIR Forum, 2002, 36(2): 3-10.

[5] 江雪, 孙乐. 用户查询意图切分的研究 [J]. 计算机学报, 2013, 36(3): 664-670.
JIANG Xue, SUN Le. Study on segmentation of user's query intents [J]. Chinese Journal of Computers, 2013, 36(3): 664-670.

[6] HE Daqing, GÖKER A, HARPER D J. Combining evidence for automatic web session identification [J]. Information Processing & Management, 2002, 38(5): 727-742.

[7] JANSESN B J, SPINK A, BLAKELY C, et al. Defining a session on web search engines [J]. Journal of the American Society for Information Science and Technology, 2007, 58(6): 862-871.

[8] DOWNEY D, DUMAIS S, HORVITZ E. Models of searching and browsing: languages, studies, and applications [C]//Proceedings of the International Joint Conference on Artificial Intelligence. Hyderabad: ACM, 2007: 1465-1472.

[9] NIKOLAI B, BERNARD J B J. Limits of the web log analysis artifacts [C]//Proceedings of Workshop on Logging Traces of Web Activity. Edinburgh: World Wide Web Conference, 2006: 152-156.

[10] MURRAY G C, LIN J, CHOWDHURY A. Identification of user session with hierarchical agglomerative clustering [J]. Journal of American Society for Information Science, 2006, 43(1): 1-9.

[11] OZMUTLU H C, CAVDUR F. Application of automatic topic identification on excite web search engine data logs [J]. Information Processing and Management, 2005, 41(5): 1243-1262.

[12] OZMUTLU S, CAVDUR F. Neural network applications for automatic new topic identification [J]. Online Information Review, 2005, 29(1): 34-53.

[13] OZMUTLU S, OZMUTLU H C, SPINK A. Automatic new topic identification in search engine transaction logs using multiple linear regression [J]. Hawaii International Conference on System Sciences, 2008, 16(3): 140.

[14] OZMUTLU S, OZMUTLU H C, BUYUK B. Using Monte-Carlo simulation for automatic new topic identification of search engine transaction logs [J]. Winter Simulation Conference, 2007, 16(5): 2306-2314.

[15] LI Xiao, WANG Yeyi, ALEX A. Learning query intent from regularized click graphs [C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2008: 339-346.

(编辑:于善清)

(上接第12页)

[11] HENDRICKX I, KIM S N, KOZAREVA Z, et al. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominal [C]//Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions Boulder: Association for Computational Linguistics Stroudsburg, PA, USA, 2009: 94-99.

[12] MIKOLOV Tomas, SUTSKEVER Ilya, CHEN Kai, et al. Distributed representations of words and pharses and their coposi-tionality [J]. Computer Science, 2013, arXiv: 1310.4546.

(编辑:于善清)