

新时代背景下语言学特征在中文实体语义关系抽取中的应用

风 罡

(河套学院 内蒙古通辽 015000)

【摘要】 新时代互联网科技高速发展,电子文本的巨大信息量带给人们便利的同时也提高了信息提取的难度。中文实体语义关系抽取是从纷繁复杂的文本信息中找出语言的语义关系并对其进行分类转化储存,这样人们再进行查询时就可以做进一步分析研究,这将对语言学探究和信息抽取研究提供重要依据。而我们想要从中文文本中获取信息,就要关注实体抽取和关系抽取的应用。因此,识别文本实体和根据上下文确定、抽取实体间的语义关系变得越来越常见。探索新时代背景下语言学特征,寻找合适的研究工具,在中文实体语义关系抽取上将大大提高其准确率和效率。基于此,我们试分析传统信息抽取研究现状,并探讨在新背景下中文实体语义关系抽取适合采用的方法及运用方向。

【关键词】 信息抽取;知识库;语义关系

DOI: 10.18686/jyyxx.v2i10.36673

中文语言学主要存在的词汇特点,如量词和句末语气词;还有些句法,如句子之间的处理式等等。基于中文语言学特征下的知识抽取也有适合的研究方法。知识抽取通常需要涉及的是清楚的、符合事实的信息,而这些信息抽取的难度取决于他们的来源和结构,因此对于不同数据源进行的信息抽取方法各不相同,传统知识库的方法效率较低并且容易产生误差,而机器学习方法可以比较科学地根据人工干预的多少进行指导,在探索语言学特征的同时也能提高中文实体间语义关系抽取的性能。还有国外最新的一些新型学习方法也在试运行中,结合语义抽取的研究经验来看,我们去探究新背景下的新方法也是很有必要的。

1 传统语义关系抽取的方法及现状

语义关系抽取近年来成为自然语言理解研究的重要研究方向。从语言学层面来理解语义关系,如美国前任总统特朗普。实体“美国”和实体“特朗普”之间存在雇佣关系。人们最开始都是采用知识库的方法解决中文实体语义关系抽取的问题,但是知识库的构建需要大量的专家团队日以继夜进行大规模的知识系统联系,而且在此基础上,具有语言学专业技能的专家也都要准备好投入大量的人力、物力及时间成本,来进行知识库的日常维系运作。因此,知识库这种方法的缺点是比较明显的。

后来人们开始使用机器学习的方法,这类方法首先解决了专家建造知识库的难题,因为只需要先建立两个实体间的关系系统即可,将任意两个实体之间的关系做出与需要关系之间的对比,然后以此作为训练数据构造分类器。机器学习的方法在人工干预过程中也要分有指导和无指导的。这两种也各有利弊,通常机器学习算法需要构造特征向量形式的训练数据。然后使用各种机器学习算法,也叫做基于特征向量的学习算法。但是同时

也需要人工去进行标注,对人力成本要求也比较高,而无指导方法不需要事先定义关系类别和人工标注数据,不需要人工指导或干预,但是正确率较低,无法确定实例所属类别。

2 基于特征向量的有指导学习方法

在这个信息极度膨胀的社会中,如何迅速有效地检索获取信息已经成为人们迫切想解决的问题。机器学习中基于语言学特征的指导学习方法是用以处理用户提出的自然语言问题,抽取有效信息,最后以自然语言给出答案的一个很好的工具。在已知一些关系信息的基础上,通过从文本中解析、分类中文实体语义关系,对分词、代词等信息抽取处理本文给出解决方案,并以内容提取为主,语法分析为辅;对结果进行细致地标注;中间过程允许用户进行修改,建立问题集合,方便机器学习;回答的语言人性化,内容准确而形式多变。因此,本文介绍了目前语言学领域关于中文实体语义关系研究上所做的一些探索性工作,希望对后续的研究和工作者提供一些参考和帮助,也为将来研究和开发全开放的言语学与计算机技术的合作打好基础。

基于特征向量的有指导学习方法的标准流程包括从训练文本提取到进行断句和分词,再到词法分析以及命名实体识别,再到短语块分析以及句法分析及特征抽取、模型文件以及测试文档。

准备工作是先从语料库中提取出文本文件,实体标注信息和实体间关系。然后进行原文切分句子、断词处理、分词处理、词性标记、语法标记等。为了提高效率,通常会训练两个分类器,判断命名实体间是否有关系,如果有关系,再送到第二个分类器,给实体分配关系类别,这样做的好处是通过排除大多数的实体来加快分类器的训练过程。另外,对每个任务可以使用 task-specific feature-set。或者从另一个角度考虑,可以分为轻量级实

体的特征,包括实体前后的词、实体类型、实体之间的距离等,中等量级考虑短语,重量级考虑实体间的依存关系,实体间树结构的距离及其他特定的结构信息。

3 语言学特征在中文实体语义关系抽取中的应用

与世界其他国家语言相比中文语言学有个明显的特征是缺少时态的表达,在关系小句和名词的词序,介词词组和动词的词序等方面还是比较特殊的。并且汉语中很多词存在一个字和两个字的形式,比如,“煤/煤炭”“蒜/大蒜”等。这些词的两种形式在词义上基本没有区别,但是由于汉语韵律和语法之间的某些特殊关系,这两种形式在句子中很多时候是不能随便互换的。在信息抽取这项技术背景下,中文语言学特征可以聚焦在词汇特征,也是目前我们能够提取的最简单有效的特征,主要是对词性的标记;而句法分析也是比较常见的中文语言学特征之一,就是在词法分词对基础上生成一个句子的完全句法脉络,以提取结构化信息以及为基本短语分析提供基础。

下面先看基于句法分析的开放式中文实体关系提取的几个简单句子,宾语可以不当作关系表述的一部分:

主谓宾关系:方小块生于重庆。这个三元组很明显(方小块、生于、重庆)。动补结构,如“方小块洗干净了衣服”,如果套用主谓宾关系就是:方小块洗衣服,但是这里描述的是一个状态,是方小块把衣服洗干净了,“干净”是动词“洗”的补语,所以还应该提取出一个三元组(方小块、洗干净了、衣服);再比如状动结构:妈妈非常喜欢画画,这句和上面有点像,主谓宾关系是妈妈喜欢画画,“非常”用于修饰“喜欢”或者是“父亲非常喜欢跑步”。下一个是介宾关系“方小块就读于清华大学”如果直接把这个三元组抽取为(方小块、就读、清华大学),很别扭,“于”和“清华大学”是介宾关系,它们的关系应该是就读于。(方小块、就读于、清华大学)。再者是宾语前置关系:海洋由水组成,“海洋”是“组成”的前置宾语,“由”是“组成”的状语,“水”和“由”是介宾关系,所以上面的句子没有明确的主谓关系,需要我们判断。抽出的三元组应该为(水、组成、海洋)。

依存句法分析就是将句子分析成一棵依存句法树,描述各个词语之间的依存关系,即指出词语之间在句法

上的搭配关系。有了上面所说的依存句法树,其实我们只需要进行各种判断就可以了。先做出下面的一点说明,就拿前文第一个例子来说。

原文:方小块生于重庆,这是分词结果;[方小块、生于、重庆],这是句法分析结果。抽取对应的语义关系则是,方小块—(主谓关系)→生于;生于—(核心关系)→核心;重庆—(动宾关系)→生于。为每一个词语都建一个依存句法字典,对于分词后的每个词语都进行这样操作。“方小块”和“重庆”,关系字典都为空。而对于“生于”,关系列表里面既有主谓也有动宾,而自己本身就是动词,主谓宾就出来了。直接从主谓关系中拿出来词语作为 entity1,再拿上自己作为关系,最后拿出动宾关系中的词语作为 entity2。很明确的三元组(方小块、生于、重庆)就出来了。

还有一种关于 Stanza 的研究方法,这种研究方法目前国内并不盛行,原因还是因为中文语言学有着相对明显的特征,而词汇和句法分析的特点用外国的软件产生的准确率还有待考证,所以目前国内对于中文语言实体间语义关系提取还是比较信赖机器学习的方法。新时代背景下,关于语言学实体语义关系探索的意义还是十分深远的,目前人们工作学习社交社会活动都充斥着语言运用的魅力,最主要的是中文语言在不同语境、语气、情绪、人物关系等因素影响下会有十分详尽的语义关系特征,这也是中文与英文最大的不同,中文实体语义关系抽取的过程对了解语言学,更成熟的运用中文语言具有深远的意义。再者关于这方面的方法研究还存在很多问题,外国语言学研究工具如何与中文语言特征兼容也是需要研究的方向,进而帮助中文语言学的发展。

4 结语

互联网科技和计算机技术让人们重视信息提取,如何利用信息提取工具来获得有效信息的研究也越来越受到关注。我们基于中文语言学特征进行的实体语义关系提取是信息提取环节十分重要的一部分,以后也将引起更广泛的重视。下一步我们应该寻找更多有效的语言特征进行实例验证,并探索更多新时代背景下的运用方法。

作者简介: 凤罡(1979.3—),男,内蒙古通辽人,副教授,研究方向:语言学。

【参考文献】

- [1] 黄鑫. 基于特征向量的中文实体间语义关系抽取研究 [D]. 苏州大学, 2009.
- [2] 季元叶. 语言学特征在中文命名实体间语义关系抽取中的应用研究 [D]. 苏州大学, 2010.
- [3] 奚斌, 钱龙华, 周国栋, 等. 语言学组合特征在语义关系抽取中的应用 [J]. 中文信息学报, 2008 (3): 44-49+63.