基于同义词词林的句子语义相似度方法及其在问答系统中的应用

周艳平 李金鹏 蔡 素
(青岛科技大学信息科学技术学院 山东 青岛 266061)

摘 要 提出一种基于同义词词林的句子语义相似度方法,借助同义词词林来计算句子的词形相似度,使用向量距离法得到句子间的词序相似度。同时,对句子进行语义依存句法分析。通过对词形、词序、语义依存相似度加权结合获得句子之间的最终相似度。将该方法应用于常问问题问答系统(Frequency Asked Questions, FAQ)的问句匹配。实验结果表明,该方法在问句匹配上相比传统方法具有更高的准确率。

关键词 问答系统 句子相似度 同义词词林 语义依存

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.08.012

A SEMANTIC SIMILARITY METHOD OF SENTENCES BASED ON TONGYICI CILIN AND ITS APPLICATION IN QUESTION ANSWERING SYSTEM

Zhou Yanping Li Jinpeng Cai Su
(*College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, Shandong, China*)

Abstract This paper proposed a method of sentence semantic similarity based on TongYiCi CiLin, which used TongYiCi CiLin to calculate the morphological similarity of sentences. The similarity of word order between sentences was obtained by using vector distance method. Meanwhile, the sentences were parsed with semantic dependency. The final similarity between sentences was obtained by weighting the similarity of morphological, word order and semantic dependencies. The proposed method was applied to the question matching of Frequency Asked Questions. Experimental results show that this method has higher accuracy in question matching than traditional methods.

Keywords Question answering system Sentence similarity Tongyici Cilin Semantic dependency

0 引 言

随着智能时代的来临,问答系统引起了国内外研究和教育机构的广泛关注。问答系统按问题集来源可分为三类[1]:基于常问问题的问答系统、基于 Internet 日志的开放域问答系统和基于百科全书知识库的知识问答系统。

在中文问答系统中,问句信息匹配最大的难点就是一词多义问题,如何准确地计算中心词之间同义项的相似程度是当前研究的重要课题[2]。现有的计算句子相似度的方法可以分为四类:文字匹配方法、概率方法、词频-逆向文档频率向量方法和语义依存方法。文字匹配方法基于两个句子中包含的相同词或同义词的数量来计算句子相似度,如基于 overlap 的改进方法[3]、余弦相似度算法。概率方法通过借助语言模型框架,利用概率方法计算两个句子的相似度[4]。词频-逆向文档频率向量方法需要计算句子中心词的词频和权重,然后生成频率向量组,使用余弦相似度计算方法得出句子之间相似程度[5]。上述三种方法只通过句子中心词的表意来比较相似度,无法对整个句子语法结构进行分析判断。语义依存方法借助本体或字典,对词进行语义分析[6-7]以解决句法问题。但传统的语义依存法不能从词义的角度上考虑句子相似度信息,因此会出现句法一致,句子相似度低的情况。

本文提出了一种基于同义词词林的句子语义相似

收稿日期:2018-12-28。国家自然科学基金项目(61402246);山东省高等学校科技计划项目(J14LN31)。**周艳平**,副教授,主研领域:智能优化算法,生产计划与生产调度,自然语言处理。**李金鹏**,硕士生。**蔡素**,硕士生。

度方法,采用同义词词林的编排及其语义特征,完成多义词的信息匹配,通过加权词形、词序和语义的相似性来获得最终句子相似程度。然后将本文提出的方法应用于常问问题的问答系统中,用来提高问句匹配的准确率。

1 一种同义词词林的句子语义相似度方法

1.1 中心词抽取

在语言学中,句子由中心部分(如主语、谓语和宾语等)和修饰语部分(如定语、状语和补语等)组成。中心部分在句子中起主导作用,本文只考虑句子中心部分的相似性。一般来说,句子中的主语和宾语通常是名词或代词,谓词通常是动词或形容词。在计算句子相似度时,重点考虑这些中心部分的词[8]。

本文利用哈尔滨工业大学信息检索研究中心开发的在线语言技术平台(Language Technology Platform, LTP)[9]获得句子之间的依存句法关系。该平台将整个句子转换为结构化语义依存树,依赖弧反映了句子中词与词之间的依赖关系。该平台可以用于分词、词性标注、命名实体、词义消歧、句法分析和语义分析。例如句子"今年我弟弟考上了青岛科技大学"可表示成树状结构,并能得到句子的中心词和词性标注,如图1所示。

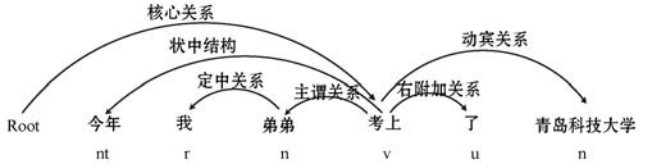

图1 使用LTP抽取句子的中心词和词性标注

1.2 词形相似度计算方法

词形相似度需要考虑到同义词识别判断的情况,本节借助同义词词林的编排及语义特点完成一词多义的信息匹配。

同义词词林利用树状结构表示词语义项间的关系[10],共有5层分支编码,第1、4层为大写英文字母、第2层为小写英文字母、第3、5层为两位十进制数。借助同义词词林计算词语相似度步骤如下:

(1) 判断两个词语的同义词词林编号在第几层不同。两个词语编号不同的层需要乘以该层的系数 d。如:Ae05A02与Ae05A03是第5层不同,则乘以第5层系数 d;Ae05A02与Ae05B03是第4层不同,则乘以第4层系数 d。为了将结果控制在[0,1]范围内,需要乘以调节参数 $\cos(l\pi/180)$,l 为该层分支元素个数。

(2) 词语所在的分支上的元素会影响到词语义项的相似度,为了将分支元素与义项相似度对应起来,需要乘以控制参数 $(l-k+1)/n$,l 为分支元素个数,k 为两个分支的距离。相似度计算如下:

$$SIM(W,W') = d \times \cos\left(l \times \frac{\pi}{180}\right) \times \left(\frac{l-k+1}{l}\right) \quad (1)$$

通过实验分析,本文将不同层数对应的参数 d 设置为:第1层 $d=0.1$,第2层 $d=0.65$,第3层 $d=0.8$,第4层 $d=0.9$,第5层 $d=0.96$。

实验发现,中心词 W、W'之间相似度 $SIM(W,W') \geqslant 0.8$ 时,才能将 W、W'作为同义词或相同词使用。但该方法只是针对同义词词林中存在的词语进行对比,如词林中并无该词语,就会默认不是相同词或同义词。为避免这种情况发生,本文根据词林中有无某个词语进行判断,步骤如下:

(1) 判断同义词词林中有无 W、W',若有,则用上述方法直接计算词语相似度;若没有,则需要对 W 扩展近义词,并将扩展的近义词按顺序加入数组中。

(2) 判断 W 近义词组中是否包含 W'。在近义词组中,序数越小的近义词与原词 W 的相似度越高,因此 W、W'的词语相似度 $SIM(W,W')$ 计算如下式所示:

$$SIM(W,W') = \begin{cases} 1 - \dfrac{orderW(W')}{countW} & arrayW \text{ 包含 } W' \\ 0 & arrayW \text{ 不包含 } W' \end{cases} \quad (2)$$

式中:$arrayW$ 是 W 的近义词组,$orderW(W')$ 是 W'在 W 近义词组中的序数,$countW$ 是 W 近义词组的元素个数。

在语言学中,一个词在不增加任何词缀的情况下转化为另一词性的现象在构词法中称作词性转化[11]。例如句子 A = "怎么做好协调工作"中的"协调"是动词,而句子 B = "和老板的关系一直不是那么协调"中的"协调"是形容词。研究证明,同义词或相同词在不同词性下含义不同。针对词性转化问题,本文提出词性匹配加权方法,对同义词或相同词的不同词性所得出来的相似度进行加权处理,从而使计算得出的中心词相似度更加准确。本文只对词语相似度 $SIM(W,W') \geqslant 0.8$ 并且词性不同的中心词进行加权处理,权值 $\gamma = 0.85$,处理之后的中心词 W 与 W'词形相似度 $SIM_{new}(W,W')$ 按下式计算:

$$SIM_{new}(W,W') = \gamma \times SIM(W,W') \quad (3)$$

设 S、S'为2个句子,S 中有 m 个词($W_i, i=1,2,\cdots,m$),S'中有 n 个词($W'_j, j=1,2,\cdots,n$),则 S、S'的词形相似度 $S_{word}(S,S')$ 为:

$$S_{word}(S,S') = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} SIM_{new}(W_i,W'_j)}{m \times n} \quad (4)$$

本文计算句子之间的词形相似度的步骤:

(1) 计算句子 S、S'中的任意两个词 W_i、W'_j 的相似度。如果同义词词林中有 W_i、W'_j 两个词,则借助同义词词林求相似度 $SIM(W_i,W'_j)$,否则用式(2)计算相似度 $SIM(W_i,W'_j)$。

(2) 如果 $SIM(W_i,W'_j)\geqslant 0.8$ 并且 W_i、W'_j 词性不同,则按式(3)计算 $SIM_{new}(W_i,W'_j)$,否则 $SIM_{new}(W_i,W'_j)=SIM(W_i,W'_j)$。

(3) 按式(4)计算句子 S、S'的词形相似度 $S_{word}(S,S')$。

1.3　词序相似度计算方法

词序相似性反映的是相同词或同义词在两个句子中的位置相似性,在某些情况下词序会直接影响到句子所要表达的意思[12]。例如句子 S = "我今天买了从青岛到北京的车票",句子 S' = "我今天买了从北京到青岛的车票"。经过词形匹配发现这两句话相似度是100%,但实际意义有差别。通常用逆序数法计算词序相似度,但其时间复杂度要高。为了降低时间复杂度,本文采用基于向量的词序相似度算法进一步提高词序相似度的计算效率。

下面举例说明计算句子 S 与句子 S'之间的词序相似度的方法。

句子 S 的中心词 = {"我","今天","买","青岛","北京","车票"};

句子 S'的中心词 = {"我","今天","买","北京","青岛","车票"};

句子 S、S'的公共词语:$C_{SS'}$ = {"我","今天","买","青岛","北京","车票"}。

对应的标准排列向量 $\boldsymbol{u}=(1,2,3,4,5,6)$。

将句子 S'中的公共词语映射为位置向量 $\boldsymbol{u}'=(1,2,3,5,4,6)$。则 $\boldsymbol{u}'$到 $\boldsymbol{u}$ 的向量距离可用下式表示:

$$distance(\boldsymbol{u},\boldsymbol{u}')=\sum_{i=1}^{n}|u'_i-u_i| \tag{5}$$

从而得出 $distance(\boldsymbol{u},\boldsymbol{u}')=2$,基于向量的词序相似度算法定义为:

$$S_{ord}(S,S')=\begin{cases}1-\dfrac{distance(\boldsymbol{u},\boldsymbol{u}')}{maxDistance} & c>1\\ 1 & c=1\\ 0 & c=0\end{cases} \tag{6}$$

式中:$maxDistance$ 为 $distance(\boldsymbol{u},\boldsymbol{u}')$的最大值:$maxDistance=n^2/2$,$c$ 是两个句子中相同中心词的个数。最终得出句子 S 与 S'的词序相似度 $S_{ord}(S,S')=0.89$。

1.4　语义相似度计算方法

词形相似度算法通过句子中心词的表意来比较相似度,不能分析和判断整个句子的句法结构。本文引入了语义依存树来弥补词形相似度算法的不足。因句子中心部分占主导作用,所以在使用依存句法进行相似度计算时,只需计算两个句子中心部分的词语组成的关键配置对的相似程度。关键配置对是指句子中的核心词以及由中心词组成的配置对[9]。这里中心词定义为名词、代词、动词和形容词,它是通过在线语言技术平台分词后的词性标注决定的。相似度 $S_t(S,S')$的计算公式为:

$$S_t(S,S')=\frac{\sum_{i=1}^{n}Q_i}{\max\{TC_1,TC_2\}} \tag{7}$$

式中:Q_i 为句子 S 和句子 S'关键配置对的权重,TC_1 为句子 S 的关键配置对数,TC_2 为句子 S'的关键配置对数。对于任意两个配置对:(1) $W1-W2$;(2) $W1'-W2'$。若 $W1=W1'$并且 $W2=W2'$,则配置对(1)和配置对(2)的权重为 1;若 $W1\neq W1'$但 $W2=W2'$,或 $W1=W1'$但 $W2\neq W2'$,则配置对(1)和配置对(2)的权重为 0.5;否则为 0。需要指出的是,这里的 $W=W'$表示这两个词是相同词或同义词,可通过 1.2 节的方法判断两个词是否为相同词或同义词。

1.5　问句相似度算法计算方法

问句相似度反映了两个问句之间的相似程度,通常用[0,1]之间的数值表示。数值越大,两个问句的相似度越高。若 S 为问句,S'为问题集中的任意一个问句,则问句相似度为:

$$SIM(S,S')=k_1\times S_{word}(S,S')+k_2\times S_{ord}(S,S')+k_3\times S_t(S,S') \tag{8}$$

式中:$S_{word}(S,S')$为词形相似度;$S_{ord}(S,S')$为词序相似度;$S_t(S,S')$为语义相似度;k_1、k_2、k_3 为相似度系数,且满足 $k_1+k_2+k_3=1$。

根据本文上述方法对问句 S 和问题集 M 中的任意问句 S'进行 FAQ 问句相似度计算,如图 2 所示,具体步骤如下:

(1) 将问句分词并抽取中心词。

(2) 使用本文提出的词形相似度计算方法计算问句的词形相似度 $S_{word}(S,S')$。

(3) 使用基于向量的词序相似度算法计算关键词之间的词序相似度 $S_{ord}(S,S')$。

(4) 借助语义依存方法计算句子之间的语义相似度 $S_t(S,S')$。

(5) 使用式(8)对上述相似度加权求和,最终得到 S、S'问句的相似度 $SIM(S,S')$。

对给定相似度阈值 σ,选择 $SIM(S,S')$中相似度

的最大值 *MAX*，若该最大值大于 σ 则返回相应的答案，若该最大值小于 σ 则默认问题集没有该问题的答案。

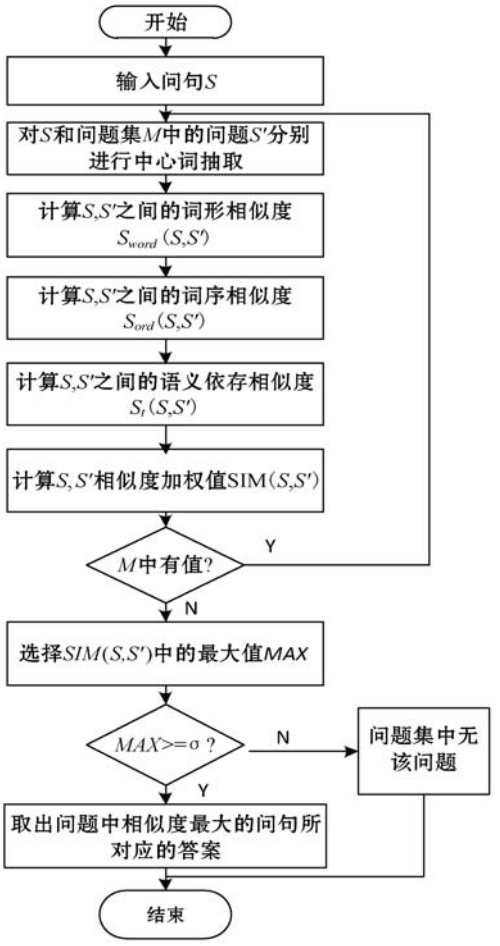

图 2 FAQ 问句相似度计算流程图

2 应用及结果分析

将本文所提出的基于同义词林的句子语义相似度算法应用于 FAQ 机器问答系统中。本文使用同义词词林扩展版，并通过在线词典来对同义词词林没有的词进行扩展，本文只将扩展后的前 5 个近义词作为计算元素加入近义词组中。本文开发环境为 Window7 X64，开发工具为 PyCharm4.5.4，开发语言为 python3.6.1。

随机从哈尔滨工业大学信息检索研究室（HIT-IRLab）提供的问答集合中选取 500 条问句作为初始数据集 *S*。随机选出 100 条问句作为初始标准集 *M*，剩余 400 条作为初始噪声集 *N*，依次选取 *S* 中的问句作为百度知道的查询条件，利用 BeautifulSoup 解析库[13]对查询返回的网页 *H* 进行标签处理，提取出 *H* 中的前 3 个标题。为了提高标题与问句的相似度，需要人工进一步筛选。*M* 中每个问句都会有 1～3 个相似句子，文本最后得到扩充标准集 *MT* 的元素个数为 253。将 *M* 与 *MT* 混合起来作为标准测试集 *MMT* 的元素个数为 353。同样将 *N* 进行扩充得到噪声测试集 *NNT* 的元素个数为 1 200。最后我们把 *MMT* 与 *NNT* 混杂起来作为测试集。

实验流程：按顺序从标准测试集 *MMT* 的 353 个句子中抽出 1 个问句 *X*，然后计算这个问句与测试集中的问句之间的相似度，并按照相似度值对测试集中问句进行倒序排序，输出前三个问句，如果所得到的这三个问句包含了扩展标准集 *MT* 中的问句 *X* 对应的所有问句（1～3 个），则说明这个问句的相似度计算是成功的。

句子成分是由一个个词组成的，即使在句子不通顺的情况下，也可以根据词来判断整个句子要表达的意思。因此句子相似度应以词形相似度为主，以语义结构相似度为辅，同时考虑词序相似度，并进行综合计算得出句子相似度。本文相似度系数的取值原则：当 k_1 过小时，词形相似度比例过低，会导致词义不同但句法相同的两个句子相似度变高。当 k_1 过大时，词形相似度比例过高，会导致词义相同但句法和顺序不同的两个句子相似度变高。因此相似度系数应符合 $k_1 \geqslant k_2 + k_3$ 且 $k_3 > k_2$（k_1, k_2, k_3 不为 0）。根据人工测试经验，当 $k_1 = 0.5$、$k_2 = 0.2$、$k_3 = 0.3$ 时较为合适。为了防止遗漏正确答案并且剔除冗余数据，本文取相似度阈值 $\sigma = 0.7$。

采用信息检索技术中的准确率（*P*）计算：

$$P = \frac{CorrectCount}{AllCount} \tag{9}$$

式中：*CorrectCount* 表示实验中正确的句子总数，*AllCount* 表示实验的句子总数。

分别用词频-逆向文档频率向量方法和文献[7]方法与本文提出的方法做测试试验，实验结果如表 1 所示。从表 1 可以看出，使用本文所提出的方法，问句相似度准确率可以达到 92.63%，明显优于使用词频-逆向文档频率向量方法和文献[7]方法。

表 1 实验结果

实验	CorrectCount	AllCount	P
词频-逆向文档频率向量方法	247	353	71.10%
文献[7]方法	307	353	86.97%
本文方法	327	353	92.63%

3 结 语

本文提出了一种基于同义词词林的句子语义相似度方法，并将该方法应用与问答系统的问句匹配。相

（下转第 81 页）

按照此方案,延误航班进港用时总和为 123.9 min,按照本文的优化方案,延误航班的进港用时总和为 104.55 min,减少了旅客的进港用时。

4 结　语

本文研究了在航班发生大面积性延误时,机场对登机桥的再调度问题。考虑了顾客满意度、登机桥类型、航班类型及航班的进离港时间等因素,以减少使用远机位登机梯的航班数、减少对原调度方案的改动性及减少旅客进港用时为目标。根据机场实际标准建立约束条件,建立了航班延误时登机桥再调度的多目标模型,以国内某机场航班延误情况为实验样本,验证了所提方法的可行性。本文所研究的登机桥再调度问题属于再调度的 NP-hard 问题,目前还没有找到可以高效解决该问题的方法,本文根据改进的 ABC 算法仿真结果来看算法的收敛性较好。结合模型、算法和实验数据进行仿真研究,证明了实验得出的登机桥再调度方案在三个目标上都得到较优的效果。后续工作还要去深入分析机场登机桥在多跑道条件下的多目标综合调度问题,并设计出更合理有效的调度求解方法,以此来获得更优的调度方案是未来工作的关键和难点。

参　考　文　献

[1] 董念清. 中国航班延误的现状、原因及治理路径[J]. 北京航空航天大学学报(社会科学版), 2013, 26(6): 25-32.

[2] 顾兆军,安一然,潘杰,等. 不正常航班旅客流恢复方法研究[J]. 计算机应用与软件, 2016,33(6):79-83.

[3] 王力,刘长有,涂奉生. 民用机场停机位优化配置[J]. 南京航空航天大学学报,2006(4):433-437.

[4] Jiang Y, Zeng L, Luo Y. Multiobjective Gate Assignment Based on Passenger Walking Distance and Fairness[J]. Mathematical Problems in Engineering,2013, 2013: 361031.

[5] 陈骁睿. 基于改进粒子群算法的机位分配问题研究[J]. 软件,2015,36(1):72-76.

[6] 李亚玲,李毅. 基于可变禁忌长度的优化停机位分配[J]. 计算机应用,2016,36(10):2940-2944.

[7] 薛清文,姜雨,刘照明,等. 基于航空公司运行成本和公平性的停机位指派[J]. 航空计算技术,2016,46(1):64-69.

[8] 陈杰,沈艳霞,陆欣. 基于信息反馈和改进适应度评价的人工蜂群算法[J]. 智能系统学报, 2016, 11(2): 172-179.

[9] Van Gorp P, Stenten H, Mens T, et al. Towards Automating Source-Consistent UML Refactorings[J]. Lecture Notes in Computer Science,2003, 2863:144-158.

[10] 周长喜, 毛力, 吴滨. 基于细菌趋药性和当前最优解策略的人工蜂群算法[J]. 计算机应用与软件,2016,33(1):268-272,285.

[11] 火久元,张政,孟凡明. 一种劣解突变策略引导的混合人工蜂群算法[J]. 计算机应用与软件, 2018,35(2):267-272,293.

[12] 魏锋涛,岳明娟,郑建明. 基于改进邻域搜索策略的人工蜂群算法[J]. 控制与决策,2019(5):72-79.

[13] 彭巍,赖怀南. 浅析基于顾客需要的民航运输服务[J]. 空运商务,2018(7):14-17.

(上接第 68 页)

比传统句子相似度方法,该方法能有效提高问句相似度准确率。本文所提出的方法可以适用于所有句子语法情况,后续研究中,将进一步简化该方法的计算复杂度和提高 FAQ 回答效率。

参　考　文　献

[1] 朱新华, 郭小华, 邓涵,等. 基于抽象概念的知网词语相似度计算[J]. 计算机工程与设计,2017,38(3):664-670.

[2] 赵臻, 吴宁, 宋盼盼. 基于多特征融合的句子语义相似度计算[J]. 计算机工程, 2012, 38(1):171-173.

[3] Metzler D, Bernstein Y, Croft B, et al. Similarity measures for tracking information flow[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 517-524.

[4] Dagan I, Lee L, Pereira F C N. Similarity-Based Models of Word Cooccurrence Probabilities[J]. Machine Learning, 1999, 34(1/3):43-69.

[5] 张俊飞. 改进 TF-IDF 结合余弦定理计算中文语句相似度[J]. 现代计算机, 2017(32):20-23.

[6] Pang X L, Jia K L. Chinese question similarity calculation based on word sense disambiguation[C]//International Conference on Machine Learning and Cybernetics. IEEE,2009: 2217-2220.

[7] 黄洪, 陈德锐. 基于语义依存的汉语句子相似度改进算法[J]. 浙江工业大学学报, 2017, 45(1):6-9.

[8] 李玲, 何聚厚. 基于语义依存分析的句子相似性度量算法及应用研究[J]. 计算机应用与软件, 2017,34(7):244-248,313.

[9] 王晶, 黄广君. 信息检索中的句子相似度计算[J]. 计算机工程, 2011, 37(12):38-40.

[10] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版),2010,28(6):602-608.

[11] 高航. 认知语法与汉语转类问题[M]. 上海交通大学出版社, 2009.

[12] 董刊生, 方金云. 基于向量距离的词序相似度算法[J]. 中文信息学报, 2009, 23(3): 45-51.

[13] Liao W H, Nie X. Spatial Association Analysis for Urban Service Based on Big Data[J]. Scientia Geographica Sinica, 2017,37(9):1310-1317.