结合依存关系与同义词词林的相似度计算

付鹏斌,陈帅帅,杨惠荣,李建君
(北京工业大学 信息学部,北京 100124)

摘 要:设计了一种基于依存关系与同义词词林相结合的语义相似度计算方法。该方法通过依存关系分别提取两个文本的关系路径,同时基于同义词词林计算两个文本之间关系路径的语义相似度。在计算两个文本之间的语义相似度时,使用语言技术平台(language technology platform,LTP)对文本进行中文分词以及获取文本的依存关系图,从中提取关系路径,从而可以结合关系路径和同义词词林计算两个文本之间的语义相似度。通过实验,获得的平均偏差率为 13.83%。实验结果表明,结合依存关系与同义词词林的语义相似度方法在准确率上相比较基于同义词词林的语义相似度和基于依存关系的语义相似度有了一定的提高。

关键词:依存关系;同义词词林;语义相似度;关系路径;平均偏差率
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2020)01-0013-06
doi:10.3969/j.issn.1673-629X.2020.01.003

Similarity Calculation between Dependency Relation and Tongyici Cilin

FU Peng-bin,CHEN Shuai-shuai,YANG Hui-rong,LI Jian-jun
(Faculty of Information Technology,Beijing University of Technology,Beijing 100124,China)

Abstract:We present a method of calculating semantic similarity based on the combination of dependency relation and Tongyici Cilin. This method extracts the relationship paths of two texts by the dependency relation, and calculates the semantic similarity of the relationship paths between two texts based on Tongyici Cilin. When calculating the semantic similarity between two texts, we use language technology platform (LTP) to segment the Chinese text and obtain the dependency graph of the text,and extract the relationship path from it,so that we can calculate the semantic similarity between the two texts by combining the relationship path and Tongyici Cilin. The average deviation rate is 13.83% in the experiment which shows that the accuracy of the semantic similarity method based on the dependency relation and Tongyici Cilin is better than that based on Tongyici Cilin and based on the dependency relation.

Key words:dependency relation;Tongyici Cilin;semantic similarity;relationship path;average deviation rate

0 引 言

语义相似度是给定一组文本,评价这一组文本之间内容表达相似程度的量度[1]。语义相似度出自计算语言学领域,目前广泛应用于自然语言处理中的 Web 信息可信分析、搜索引擎、Web 服务发现、文本聚类研究和标识释义等领域[2]。

在语义相似度的研究方法中,主要分为基于词向量的语义距离相似度计算方法和基于语法结构的语义相似度计算方法。其中,基于词向量的语义距离相似度计算,是将文本中的词频转化为词向量的形式,然后在词向量的基础上计算空间距离的长度,以此来表示文本的语义距离相似度。目前,主流的词向量转化方法是 TF-IDF(term frequency - inverse document frequency)方法,TF-IDF 方法是计算出文本中词的词频集合[3]。而使用 TF-IDF 方法将中文文本转化成词向量,比较不同词向量在线性空间中的相似度有余弦距离、欧氏距离、概率分布距离(K-L 距离)等方法。文献[4]使用向量空间模型计算文本的语义相似度,使用 TF-IDF 算法将文本转化为词向量,然后将这些词向量映射到文本向量空间,这样就将一组文本之间的匹配问题转化为求向量空间的距离问题。但是,基于向量空间模型的语义相似度只是单纯地计算词向量之间的空间距离,没有考虑句子中词语的词序和句子的结构信息对句子语义的影响。文献[5]使用基于词

收稿日期:2019-02-21 **修回日期**:2019-06-21 **网络出版时间**:2019-09-25
基金项目:北京市自然科学基金资助项目(4153058)
作者简介:付鹏斌(1967-),男,副教授,硕士,研究方向为智能系统、自然语言处理;陈帅帅(1990-),男,硕士研究生,研究方向为智能信息系统。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1523.042.html

性信息的改进 TF-IDF 算法去计算每个词向量之间的权重系数,然后将这些权重系数应用到向量空间和马尔可夫模型中,分别计算它们的语义相似度,最终获得整体的语义相似度。但是,文献[5]没有精确地反映出每个词向量之间的语义关联,如果在词向量中同时考虑到语义结构,那么该方法在文本语义相似度中有更好的表现。

在基于语法结构的语义相似度计算方法中,应用最广泛的语法结构是依存句法结构。依存句法是法国语言学家 L. Tesnier[6] 提出的,这种句法结构将句子的内部成分之间的依赖关系更加清楚地呈现到开发者的面前。语义依存关系能准确反映句子成分之间的搭配关系,李彬[7] 利用句子的关键依存关系来匹配相似度,但是只使用依存关系中的词来计算依赖关系的相似性,不能准确地反映句子的内部语义关系。文献[8]对中文依存句法树进行研究和分析,提出一种细粒度依存关系的相似度计算方法,该方法基于依存句法树中的各节点的词语、词性以及它们之间的依赖关系及其重要性权重等多个特征值,给出了两个依存句法结构的相似度计算方法。但是,文献[8]计算复杂度特别大,当文本的句长特别大时,消耗的时间较多,影响文本的语义相似度计算的效率。

而目前针对基于依存关系的语义相似度计算方法中,只考虑文本中词语的词序信息和句子的结构信息,而忽略文本中单个词语之间的词义信息。因此,文中在基于依存关系的语义相似度计算方法的基础上,增加了基于同义词词林的词语语义相似度计算方法,较好地解决上述问题,弥补以上不足。

1 相关技术

1.1 依存关系图

定义 1:依存关系图 $R_s = (V_s, E_s)$,V_s 为图中所有顶点的集合,E_s 为图中所有相邻边的集合。且满足条件:$\forall e \in E_s$,$\exists u, v \in V_s (u \neq v)$,使得 $e = (u, v)$。

依存关系图是根据标注关系连接分词的,图中的每个顶点表示的是一个分词,子节点表示文本的依存词,父节点表示文本的中心词,子节点是依赖于父节点,它们直接使用连接弧来反映它们之间的依存关系。其中依存关系的标志类型有 15 种[9],如表 1 所示。

表 1 依存句法分析标注关系

类型	描述	Example
SBV	主谓关系	俄罗斯举办世界杯(俄罗斯←举办)
VOB	动宾关系	俄罗斯举办世界杯(举办→世界杯)
IOB	间宾关系	荷兰队给荷兰球迷一个失望的结局(给→荷兰球迷)
FOB	前置宾语	荷兰球迷的眼泪都流完了(眼泪←流)
DBL	兼语	俄罗斯的朋友邀请我看世界杯(邀请→我)
ATT	定中关系	橙色球衣(橙色←球衣)
ADV	状中结构	非常完美(非常←完美)
CMP	动补结构	丢失了球门(丢→失)
COO	并列关系	荷兰队和意大利队(荷兰队→意大利队)
POB	介宾关系	在球门范围内(在→内)
LAD	左附加关系	荷兰队和意大利队(和←意大利队)
RAD	右附加关系	球员们(球员→们)
IS	独立结构	两个单句在结构上彼此独立

1.2 同义词词林

《同义词词林》[10] 是以树状的形式将所有的词语编织在一起,将所有的词语可以分为大类、中类和小类这三类形式。为了更能细化各个词语之间的语义关系,《同义词词林》将小类又细分为词群和原子词群。词群是将小类中的词语根据词语之间的词义相关性和词义相似性进行划分,而原子词群又在词群的基础上进行划分,每个原子词群之间的词语相关性特别的接近而且词义相似性几乎相同。根据上述分析,可以将《同义词词林》分为 5 层树状结构,它们以编码的形式进行体现。第一层的编码形式使用英文大写字母表示;第二层的编码形式使用英文小写字母表示;第三层的编码形式使用两位阿拉伯数字表示;第四层使用英文大写字母表示;第五层使用两位阿拉伯数字表示。同时为了体现第五层的词义相关性和词义相似性,单独增加一位编码进行标记,标记有 3 种,分别是“=”、“#”、“@”,其中“=”代表“相等”、“同义”;“#”代表“不等”、“同类”,属于相关词语;“@”代表“自我封闭”、“独立”,它在词典中既没有同义词,也没有相关词。具体的编码描述如下:

<词义编码>=<大类><中类><小类><词群><原子词群><标记>

例如:编码“Ba01A02 =”的词语类别为“物质 质素”,它的编码描述见图 1。

编码位	1	2	3	4	5	6	7	8
符合	B	a	0	1	A	0	2	=
符号类别	大类	中类	小类		词群	原子词群		标记
层数	第一层	第二层	第三层		第四层	第五层		

图 1　测试用例的编码描述

2　依存关系与同义词词林相结合的语义相似度计算方法

文献[11]提出一种基于句法依存分析的路径相似度计算方法,该方法首先对文本进行句法依存分析,获得依存树,然后在依存树中提取关系路径,最后进行路径间相似度的计算。

文献[12]提出并实现了一种基于同义词词林的词语相似度计算方法,该方法从词语的语义出发,根据词语的义项在同义词词林的位置和编码,计算出词语的相似度。

文献[11]只考虑文本中词语的词序信息和句子的结构信息,而忽略文本中单个词语之间的词义信息。而文献[12]只是从词语的语义出发,没有考虑文本的句子结构。因此文中提出了使用依存关系与同义词词林相结合的语义相似度计算方法,建立了一种结合依存关系与同义词词林的语义相似度模型,如图 2 所示。该模型以哈工大自然语言处理平台为基础,将文本 A 和文本 B 进行中文分词、词性标注、语法分析和语义分析等包装,最终获得依存关系图;在依存关系图的基础上提取关系路径;使用基于《同义词词林》的词汇语义相似度和基于搭配对的关系路径计算文本之间的语义相似度。

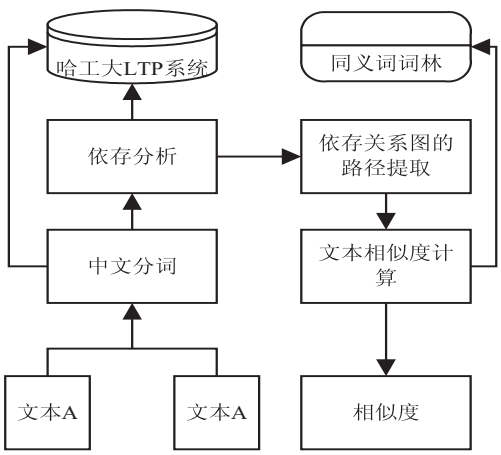

图 2　依存关系与同义词词林相结合的语义相似度模型

2.1　依存关系图中的关系路径提取

定义 2:关系路径 p 可以表示为从依存关系图的节点 v_0 开始,到节点 v_n 结束中间所经过的一系列边 $e_s \subseteq E_s$ 和顶点 $v_s \subseteq V_s$ 所构成的集合。且满足以下两个条件:

连接性: $\forall i:(v_{i-1},v_i) \in E_s \vee (v_i,v_{i-1}) \in E_s$;

无环性: $\forall i \forall j: i \neq j \rightarrow v_i \neq v_j$ 。

传统的计算文本之间语义相似度的方法是通过对词语之间的语义相似度进行加权求和,而文中在计算语义相似度时加入了依存句法结构,所以可以将计算文本之间的语义相似度转化为求关系路径间词语的加权之和。关系路径即通过遍历依存关系图,获得图中任意两个顶点之间的通路,并根据通路得到两个顶点之间的依存关系,它可以表示文本中词语之间直接或间接的关系。每一个依存关系表示一个直接的语义关系,而一条关系路径表示两个词语之间非直接的语义关系。因为关系路径是整个句子的一部分,所以可以通过不同文本间对应的关系路径的相似度来计算出文本间的相似度[13]。

通过下面的算法流程在依存关系图中提取关系路径[14]。

(1)算法输入:依存关系图 $R_s = (V_s,E_s)$, V_s 为图中所有顶点的集合, E_s 为图中所有相邻边的集合。且满足条件: $\forall e \in E_s$, $\exists u,v \in V_s$ ($u \neq v$),使得 $e = (u,v)$;

(2)初始化顶点集合 S 为空集,初始化关系路径集合 C 为空集;

(3) $\forall x \in V_s$,将 x 添加到 S 中;

(4)若($\exists u \in S$) $\wedge$ ($\exists v \in V_s - S$)满足(u, v) $\in E_s \vee (v, u) \in E_s$,则将 v 添加到集合 S 中;

(5)寻找 S 中所有节点之间存在的路径 $P = <v_i, \cdots, v_j, \cdots, v_n>$, v_i, v_j, $v_n \in S$。令 $P' = <v_n, \cdots, v_j, \cdots, v_i>$, v_n, v_j, $v_i \in S$。若($P \notin C$) $\wedge$ ($P' \notin C$),则将 P 添加到关系路径集合 C 中;

(6)若 $S \neq V_s$,转到 3。否则,算法结束,返回关系路径集合 C。

2.2 基于同义词词林的词语语义相似度计算

根据《同义词词林》的分析可得,若是两个词语的编码形式在第一层上有所区别,则说明两个词语不在同一个大类中,它们之间的词义几乎没有相关性,如果在第一层的编码相同,说明它们之间的词义具有相似性,具体的相似性大小可以根据下方的算法流程进行计算。文中采用的算法定义是:在树形结构中,两个词语的语义相似性与它们所处的层级成反比,对于标记位进行特殊处理[15]。具体的词语语义相似度计算方法如下方的算法流程所示:

(1)算法输入:两个词语 S_1 和 S_2;

(2)查询同义词词林,分别获得词语 S_1 和 S_2 的编码形式 $code_1$ 和 $code_2$;

(3)遍历 $code_1$ 和 $code_2$,如果 $code_1$ 和 $code_2$ 的编码形式都相同,则 SenseSim(S_1, S_2) = 1,同时返回到第 5 步,反之,到第 3 步;

(4)如果 $code_1$ 和 $code_2$ 的编码形式除标记位相同,若标记位等于“=”符号或“@”符号,则 SenseSim(S_1, S_2) = 1,否则,SenseSim(S_1, S_2) = 0.5,同时返回到第 5 步,反之,到第 4 步;

(5)如果 $code_1$ 和 $code_2$ 的编码形式的前 $i-1$ 位编码都相同,而第 i 位编码不同,确定 i 在同义词词林树状结构中的层数 j(其中层数 j 的获得在 1.2 小节中有介绍),则 SenseSim(S_1, S_2) = 1/(12-(2 * j));

(6)返回词语语义相似度 SenseSim(S_1, S_2)。

2.3 关系路径间语义相似度计算

通常情况下,计算文本的语义相似度是通过计算词语之间的语义相似度的加权求和。类似地,计算路径间语义相似度,可以将 2.1 小节中提取的关系路径转化为计算词语语义相似度计算的方法,但是这种计算方法不能完全体现词语之间的无歧义性的依存关系,以及词语之间的直接或间接依存关系[16]。因此,可以使用式 1 表示关系路径 p_i:

$$p_i = << w_1, r_1>, \cdots, <w_i, r_i>, \cdots, <w_n, r_n>>, i = 1, 2, \cdots, n \quad (1)$$

其中,w_i 为关系路径 p_i 上的一个顶点;r_i 为指向顶点 w_i 的有向边的依存关系。

则两条关系路径 P_i 和 P_j 的语义相似度可以用式 2 计算得到。

$$S(P_i, P_j) = \begin{cases} \dfrac{\sum_{k=1}^{n} W(r_k^{P_i}) * \text{SenseSim}(w_k^{P_i}, w_k^{P_j})}{\sum_{k=1}^{n} W(r_k^{P_i})}, & m = n \\ 0, & m! = n \end{cases} \quad (2)$$

其中,$\text{SenseSim}(w_k^{P_i}, w_k^{P_j})$ 为关系路径 P_i 和关系路径 P_j 中第 k 个位置上对应的词组之间的语义相似度(具体的计算过程在 2.2 小节中有介绍);$W(r_k^{P_i})$ 表示依存关系 $r_k^{P_i}$ 的权重。使用文献[17]中的研究结果对依存关系 $r_k^{P_i}$ 进行赋值,每一种依存关系的权重如表 2 所示。

表 2 依存关系的权重

依存关系	权重值
SBV(主谓关系)	0.433
VOB(动宾关系)	0.347
IOB(间宾关系)	0.347
FOB(前置宾语)	0.347
DBL(兼语)	0.347
其他关系	0.250

2.4 结合依存关系与同义词词林的语义相似度计算

文本 A 由关系路径集合 $\Pi_A = \{p_1^A, p_2^A, \cdots, p_n^A\}$ 组成,文本 B 由关系路径集合 $\Pi_B = \{p_1^B, p_2^B, \cdots, p_m^B\}$ 组成(关系路径集合的获取过程在 2.1 小节有具体介绍)。首先根据关系路径长度对关系路径集合进行分类,然后计算相同关系路径长度之间的语义相似度,最后加权求和计算文本的语义相似度。

设文本 A 的关系路径集合中最大的一条关系路径的长度为 max_path_count_A,文本 B 的关系路径集合中最大的一条关系路径的长度为 max_path_count_B,设 length(p) 为关系路径 $p(p \in \Pi_A \cup p \in \Pi_B)$ 的长度,且 $0 < \text{length}(p) \leqslant \min(\text{max\_path\_count\_}A, \text{max\_path\_count\_}B) \cap \text{length}(p) \in Z$,分别使用 T_1 和 T_2 表示关系路径长度为 $i(0 < i \leqslant \min(\text{max\_path\_count\_}A, \text{max\_path\_count\_}B) \cap i \in Z)$ 的关系路径集合,设 $|T_1| = x$, $|T_2| = y$。

其中关系路径集合 T_1 和 T_2 满足下面的规则:

$$T_1 = \{p_j \mid p_j \in \Pi_A \cap \text{length}(p_j) = i\}$$

$$T_2 = \{p_j \mid p_j \in \Pi_B \cap \text{length}(p_j) = i\}$$

关系路径集合 T_1 和 T_2 中的关系路径一一对应,构建 $x * y$ 维的相似度矩阵 $\boldsymbol{M}_{AB}(i)$,具体的计算方法[18]如下:

$$\boldsymbol{M}_{AB}(i)=\begin{pmatrix} S(p_1^{T_1},p_1^{T_2}) & \cdots & S(p_1^{T_1},p_y^{T_2}) \\ \vdots & S(p_s^{T_1},p_t^{T_2}) & \vdots \\ S(p_x^{T_1},p_1^{T_2}) & \cdots & S(p_x^{T_1},p_y^{T_2}) \end{pmatrix} \tag{3}$$

(1)使用式 2 计算相似度矩阵 $\boldsymbol{M}_{AB}(i)$ 中的每个元素 $S(p_s^{T_1},p_t^{T_2})$ 的值。

(2)使用式 4 计算关系路径集合 T_1 和 T_2 的语义相似度 $X_{\text{wss}}(i)$，$X_{\text{wss}}(i)$ 具体表示关系路径长度 i 在关系路径集合 T_1、T_2 上的语义相似度,具体计算过程如下:

$$X_{\text{wss}}(i)=\frac{\sum_{k=1}^{x}(\max(S(p_k,p_j)),j\in[1,y])}{x} \tag{4}$$

(3) $X_{\text{wss}}(AB)$ 表示文本 A 和文本 B 之间的加权语义相似度,具体的计算过程如下:

$$X_{\text{wss}}(AB)=\sum_{i=1}^{h}\theta_i X_{\text{wss}}(i) \tag{5}$$

其中，设 $h=\min(\text{max\_path\_count\_}A,\text{max\_path\_count\_}B)$，$\theta_i$ 表示不同关系路径长度上的语义相似度权值。长的关系路径更多地表示文本中词语与词语的间接关系,所以 θ_i 和关系路径的长度成反比,文中 $\theta_i=\frac{k}{i}$,且 k 满足 $\sum_{i=1}^{h}\frac{k}{i}=1$。

(4)计算所得的 $X_{\text{wss}}(AB)$,就是文中所求的文本语义相似度。

3　实验结果与分析

3.1　实验数据

采用陕西省某重点中学 2019 届高二年级共 1 038 名学生的第二学期历史期末考试试题作为实验数据集,一共采集了 2 076 条文本数据,每条文本数据包括学生答案、教师给分、该试题总分,及该题的标准答案。从实验数据集中选取典型的 236 条数据作为实验数据,使用文中基于改进的依存关系的文本相似度算法,以及胡宝顺[11]和田久乐[12]提出的相似度方法分别在实验数据上计算相似度,最后比较它们在评价指标上的效果。

3.2　实验结果与分析

文中引入了偏差率和平均偏差率分析它们之间的显著效果。偏差率是表示各种方法计算的相似度和标准相似度(文中使用专家标记的相似度)之间的偏差,局部反映了文本之间语义相似度的稳定程度和正确性。而平均偏差率是从整体上去反映文本之间语义相似度的稳定程度和正确性。偏差率和平均偏差率的具体计算过程分别如式 6 和式 7 所示:

$$\text{偏差率}=\frac{|\text{标准相似度}-\text{各种方法的相似度}|}{\text{相似度的总量度}} \tag{6}$$

$$\text{平均偏差率}=\frac{1}{236}\sum_{i=1}^{236}\frac{|\text{标准相似度}-\text{各种方法的相似度}|}{\text{相似度的总量度}} \tag{7}$$

其中,由于计算的相似度范围在[0,1]之间,所以相似度的总量度恒等于 1。

对文中的相似度方法和胡宝顺的相似度计算方法绘制折线图,具体如图 3 所示。

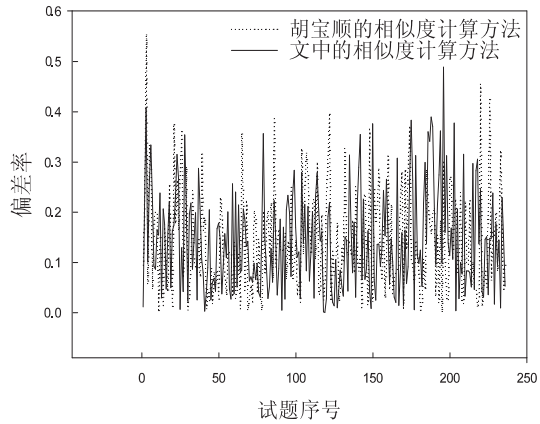

图 3　方法对比(1)

对文中的相似度方法和田久乐的相似度计算方法绘制折线图,具体如图 4 所示。

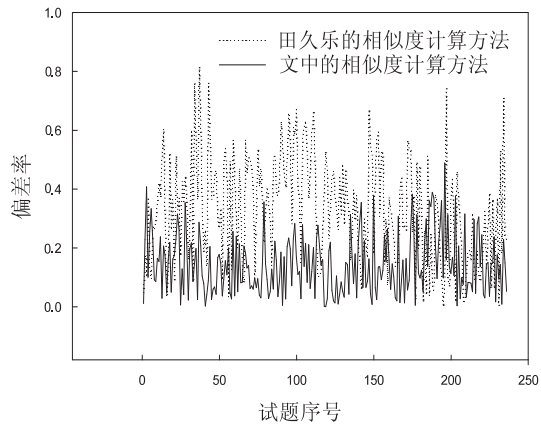

图 4　方法对比(2)

通过分析图 3 和图 4 可得:在图 3 中,基于依存关系与同义词词林相结合的语义相似度计算方法相比胡宝顺的相似度方法在偏差率上有了小幅度的降低;在图 4 中,文中方法相比田久乐的相似度方法在偏差率上有了大幅度的降低,同时文中方法相比田久乐的相似度方法在折线图上的上下幅度波动明显较小,说明文中方法的稳定性相有了明显的提高。通过分析可得,文中的相似度方法和胡宝顺的相似度方法都使用了依存关系计算文本的相似度,在计算相似度的过程中增加了语序结构,计算所得的文本相似度更能反映出语义层面的含义,所以两种相似度方法在偏差率和稳定性上的差别不是很大。但是文中方法在胡宝顺的

方法的基础上增加了基于同义词词林的词语语义相似度,在计算文本的相似度过程中不仅考虑了语义结构,而且还考虑了词形之间的词义信息,所以相比较胡宝顺的方法在相似度的偏差率上有了小幅度的降低。但是田久乐的相似度计算方法是基于同义词词林计算文本的相似度,只考虑了词形的词义信息,忽略了语义结构对文本相似度的影响,所以田久乐的相似度计算方法不仅在偏差率上还是在稳定性上都不如文中的相似度计算方法。

使用式 7 分别计算文中相似度方法、胡宝顺的相似度方法和田久乐的相似度方法的平均偏差率,文中相似度方法的平均偏差率为 13.83%,略低于胡宝顺相似度方法的平均偏差率 14.36%,明显低于田久乐相似度方法的平均偏差率 32.92%。因此,提出的结合依存关系与同义词词林的语义相似度计算方法,不但可以缩小与标准相似度之间的偏差率,同时可以提高该方法的稳定性。

4 结束语

笔者针对语义相似度计算方法的研究,设计了一种基于依存关系与同义词词林相结合的语义相似度计算方法,并在某高中历史科目中进行实验验证。通过实验分析可得,该方法的准确率相比较基于同义词词林的语义相似度和基于依存关系的语义相似度有了一定的提高。但是,笔者发现该方法虽然对于所有学科都能使用,但是由于各学科中的差异性,所以造成计算的精确性不是很高。在今后的研究中,可以根据不同的学科选择不同的相似度方法进行相似度计算,这样可以大大地提高相似度的精度。

参考文献:

[1] FURLAN B,BATANOVIC V,NIKOLIC B. Semantic similarity of short texts in languages with a deficient natural language processing support[J]. Decision Support Systems,2013,55(3):710-719.

[2] FERREIRA R, CAVALCANTI G D C, FREITAS F, et al. Combining sentence similarities measures to identify paraphrases[J]. Computer Speech & Language,2018,47(1):59-73.

[3] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(S1):167-170.

[4] 蔡 玮,黄陈蓉,林 忠,等. 一种基于向量空间模型的主观题批改算法[J]. 计算机与现代化,2008(12):88-90.

[5] 周丽杰,于伟海,郭 成. 基于改进的 TF-IDF 方法的文本相似度算法研究[J]. 泰山学院学报,2015,37(3):18-22.

[6] 刘海涛. 依存语法和机器翻译[J]. 语言文字应用,1997(3):91-95.

[7] 李 彬,刘 挺,秦 兵,等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究,2003,20(12):15-17.

[8] 熊 晶,王继鹏,魏墨济. 基于细粒度依存关系的中文长句相似度计算[J]. 科学技术与工程,2017,17(11):277-281.

[9] LIU Ting,MA Jinshan,LI Sheng. Building a dependency treebank for improving Chinese parser[J]. Journal of Chinese Language and Computing,2006,16(4):207-224.

[10] 梅家驹,竺一鸣,高蕴琦,等. 同义词词林[M]. 上海:上海辞书出版社,1993:106-108.

[11] 胡宝顺,王大玲,于 戈,等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报,2008,31(4):662-676.

[12] 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.

[13] 马 婷. 事实型中文问答系统中片段检索方法的研究[D]. 沈阳:东北大学,2008.

[14] LIN Dekang,PANTEL P. Discovery of inference rules for question answering[J]. Natural Language Engineering,2001,7(4):343-360.

[15] 关 毅,王晓龙. 基于统计的汉语词汇间语义相似度计算[C]//语言计算与基于内容的文本处理:全国第七届计算语言学联合学术会议论文集. 北京:清华大学出版社,2003:221-227.

[16] 李 琳,李 辉. 一种基于概念向量空间的文本相似度计算方法[J]. 数据分析与知识发现,2018,2(5):48-58.

[17] OLIVA J,SERRANO J,CASTILLO M D,et al. SyMSS:a syntax-based measure for short-text semantic similarity[J]. Data & Knowledge Engineering,2011,70(4):390-405.

[18] 刘亚军,徐 易. 一种基于加权语义相似度模型的自动问答系统[J]. 东南大学学报:自然科学版,2004,34(5):609-612.