基于关系向量模型的句子相似度计算

殷耀明,张东站
YIN Yaoming, ZHANG Dongzhan

厦门大学 信息科学与技术学院,福建 厦门 361005
School of Information Science and Engineering, Xiamen University, Xiamen, Fujian 361005, China

YIN Yaoming, ZHANG Dongzhan. Sentence similarity computing based on relation vector model. Computer Engineering and Applications, 2014, 50(2):198-203.

Abstract: Sentence similarity computation is very important in all fields of natural language process. Some of the traditional algorithms only compare sentences based on their surface form such as same words, sentence length, word order and do not consider the sentence deep-level semantic information, some methods considered the sentence semantics get an unsatisfactory performance on the algorithm practicality. Therefore, a relation vector model which taking into account the relationship of sentence structure and semantic information based on space vector model is presented, this model is composed of a mix between the key words of the sentence and the key words synonymous information, which reflects local structural component of the sentence as well as the correlation between the local structure and therefore better reflects the structure and semantics of the sentence. An algorithm of sentence similarity based on relation vector model is put forward. The algorithm is applied to the network news summary generation algorithm in order to avoid redundancy. The experimental results show that, compared with the algorithm which considers the word order and semantic, relation vector model algorithm not only improves the accuracy of sentence similarity calculation, the time complexity of calculation is also reduced.
Key words: sentence similarity; relation vector model; sentence syntax; sentence semantics

摘 要:句子相似度的计算在自然语言处理的各个领域占有很重要的地位,一些传统的计算方法只考虑句子的词形、句长、词序等表面信息,并没有考虑句子更深层次的语义信息,另一些考虑句子语义的方法在实用性上的表现不太理想。在空间向量模型的基础上提出了一种同时考虑句子结构和语义信息的关系向量模型,这种模型考虑了组成句子的关键词之间的搭配关系和关键词的同义信息,这些信息反应了句子的局部结构成分以及各局部之间的关联关系,因此更能体现句子的结构和语义信息。以关系向量模型为核心,提出了基于关系向量模型的句子相似度计算方法。同时将该算法应用到网络热点新闻自动摘要生成算法中,排除文摘中意思相近的句子从而避免文摘的冗余。实验结果表明,在考虑网络新闻中的句子相似度时,与考虑词序与语义的算法相比,关系向量模型算法不但提高了句子相似度计算的准确率,计算的时间复杂度也得到了降低。
关键词:句子相似度;关系向量模型;句子语法;句子语义
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1304-0017

1 引言

句子相似度计算是自然语言处理领域中比较重要的研究课题,应用广泛。在信息检索领域,相似度方法用来对检索结果进行排序。在问答系统中,需要使用相似度方法对用户所提问题和系统知识库中的问题进行比较,找到问题的最佳匹配从而返回最佳答案。在自动文摘的生成过程中,需要使用句子相似度方法排除意思相近的句子,从而避免文摘的冗余。然而,自然语言表达和语法的多样性使得辨别两个语义相近的句子十分困难。

目前,现有的计算句子相似度的方法,按照对语句的分析程度来看,主要存在两种,一种是基于向量空间模型的方法,这种方法把句子看成为词的线性序列,不对语句进行语法结构分析,相应的语句相似度衡量机制

作者简介:殷耀明(1990—),男,硕士研究生,主研方向:数据挖掘和自然语言处理;张东站,副教授,硕士生导师。
收稿日期:2013-04-02 **修回日期:**2013-07-23 **文章编号:**1002-8331(2014)02-0198-06
CNKI**网络优先出版:**2013-08-22, http://www.cnki.net/kcms/detail/11.2127.TP.20130822.1408.007.html

只能利用句子的表层信息,即组成句子的词的词频、词性、长度、词序等信息。由于不加任何结构分析,这种方法在计算语句之间的相似度时不能考虑句子整体结构的相似性;另一种方法是对语句进行完全的句法和语义分析,这是一种深层结构分析法,对被比较的两个句子进行深层的句法分析,找出依存关系,并在依存分析结果的基础上进行相似度计算。本文提出的一种关系向量模型既考虑了句子的表层信息,同时融入了语义分析。实验表明,本文提出的算法在准确率和算法时间效率上都有较大的提高。

2 句子相似度计算方法

2.1 相关工作

Palakorn A,Hu XiaoHua等[1]总结了三类计算句子相似度的算法,对当前主要的算法进行了分类,同时分析了各种方法的原理并将它们进行了对比。Li Yuhua,David M等[2]提出了一种计算两个词相似度的方法,并将其用在了句子相似度计算的同义词分析上。Donald M,Susan D等[3]提出了一种计算短语之间相似度的算法。李彬,刘挺等[4]提出了基于语义依存的汉语句子相似度计算方法,该方法基于《知网》的知识资源,首先采用哈尔滨工业大学计算机科学与技术学院信息检索研究室所作的依存句法分析器建立句子依存树,然后利用依存结构计算有效搭配对之间的相似程度,这种方法测试结果的准确率严重依赖于所生成的句法依存树,在分析句子较长,动词较多的网络文章时,正确率往往比较低。穗志方,俞士汶[5]设计了一种基于骨架依存树的语句相似度计算模型SBCM。该模型在计算语句相似度时,可以同时考虑语句之间的整体结构信息以及词汇语义信息。郭艳华,周昌乐[6]以依存语法作为语言模型的基础,提出了语句依存关系网协动生成的分析策略。裴婧,包宏[7]通过对传统的汉语句子相似度模型进行改进,提出了一种基于关键词加权的汉语句子相似度计算方法,并实现了一个基于常问问题库的中文问答系统。董自涛,包佃清等[8]以及郑实福,刘挺等[9]也研究了句子相似度在问答系统中的应用。邸书灵,刘晓飞等[10]对基于分词的语句相似度计算进行了改进,针对基于分词的语句相似度计算过于依赖实际的分词效果的问题,在原相似度计算模型中增加了两个句子不分词时的词形相似度计算,以缓解因为句子分词不准确而导致相似度计算结果偏低的情况。杨思春[11]对基于相同词的句子相似度模型作了改进。刘群,李素建[12]研究了《知网》中知识描述语言的语法,采用了一种更为结构化的方式改写了《知网》中词的定义,同时研究了义原的相似度计算方法、集合和特征结构的相似度计算方法,并在此基础上提出了利用《知网》进行词语相似度计算的算法,给句子相似度计算中同义词的判别提供了参考。张奇,吴立德等[13]通过回归方法将几种相似度结果综合起来,提出了一种新的句子相似度度量方法并研究了其在文本自动摘要中的应用。李玉红,柴林燕等[14]针对网络考试系统中主观题自动评分面临的困难和问题,提出了一种基于中文分词技术结合语句相似度的主观题自动判分算法。Li Yuhua,David M等[15]提出了一种同时考虑语义和词序的句子相似度计算方法,并将其用在了问答系统中。该方法计算两个句子的语义相似度和词序相似度,然后进行加权得到两个句子的相似度。由于需要和本文提出的算法进行比较,下面简单介绍下该算法的计算过程。

2.2 基于语义和词序的句子相似度计算算法

为便于说明,此处在对原文进行了深入研究的情况下对其中涉及的相关概念作如下描述:

定义1 给定一个句子 T_i,经过汉语分词系统分词后,得到的所有词 w_i 构成的向量称为句子 T_i 的向量表示,表示为 $T_i=\{w_1, w_2, \cdots, w_n\}$。

例1 对于语句 T_1:我是中国人。 T_2:我爱中国。分词后为 T_1:我/r 是/v 中国人/n。 T_2:我/r 爱/v 中国/ns。则 T_1、T_2 的向量表示如下:

$T_1=\{$我,是,中国人$\}$, $T_2=\{$我,爱,中国$\}$

定义2 给定一个句子 T_i 的向量表示, T_i 中词的个数称为 T_i 的向量长度,表示为 $Len(T_i)$。

例2 对于定义1中的两个句子 T_1、T_2, $Len(T_1)=3$, $Len(T_2)=3$。

定义3 给定两个句子 T_i、T_j 的向量表示,将 T_i、T_j 中的所有词 w_i 进行合并且对于重复出现的词只保留一个,由此得到两个向量的集合 T 称为 T_i、T_j 的并集,表示为 $T=T_i \cup T_j=\{w_1, w_2, \cdots, w_n\}$。

例3 对于定义1中的两个句子 T_1、T_2,它们的并集表示如下:

$T=T_1 \cup T_2=\{$我,是,中国人,爱,中国$\}$

很显然,并集的长度 $Len(T) \leqslant Len(T_1)+Len(T_2)$,因为两个句子中可能出现相同的词。

定义4 给定一个句子 T_i 的向量表示 $T_i=\{w_1, w_2, \cdots, w_n\}$ 和一个词 w_i,依次计算 w_i 和 T_i 中每个词的相似度(值为0到1之间),所得所有结果中的最大值称为 w_i 在 T_i 中的语义分数,表示为 C_i。

定义5 给定两个句子 T_i、T_j 的向量表示, T_i 和 T_j 的集合表示为 $T=\{w_1, w_2, \cdots, w_n\}$,对 T 中的每个词 w_i,计算 w_i 在 T_i 中的语义分数 C_i, T 中每个词的语义分数组成的一个向量称为 T_i 基于 T 的语义向量,表示为 $S_i=\{C_1, C_2, \cdots, C_n\}$。

在该算法中,基于 T 分别计算 T_i 和 T_j 的语义向量 S_i、S_j,以计算 S_i 作为说明,过程如下:

(1)对于 T 中的每个词 w_i,如果 w_i 在 T_i 中出现,则在语义向量 S_i 中将 w_i 的语义分数 C_i 设为1。

(2)如果 T_i 中不包含 w_i,则计算 w_i 在 T_i 中的语义分数 C_i,如果 C_i 大于预先设定的阈值 δ,则 C_i 保持不变,否则 $C_i=0$,本文中 δ 取为0.2。

根据语义向量计算语义相似度方法如式(1)所示:

$$S_s=\frac{S_1\times S_2}{\|S_1\|\times\|S_2\|} \tag{1}$$

词序相似度计算法方法如式(2)所示:

$$S_r=1-\frac{\|r_1-r_2\|}{\|r_1+r_2\|} \tag{2}$$

其中 r_1、r_2 分别为 T_1、T_2 的词序向量,以 T_1 为例,其计算方法如下:

(1)对于 T 中的每个词 w_i,如果 T_1 中包含该词,则 r_1 中该词的取值为该词在 T_1 中出现的次序。否则在 T_1 中找出与 w_i 最相似的词 w_i^*。

(2)如果 w_i 和 w_i^* 的相似度大于一个给定的阈值(实验过程中取为0.4),w_i 在 r_1 中的取值设为 w_i^* 在 T_1 中出现的次序。

(3)如果以上两种情况均未发生,则 w_i 在 r_1 中的取值设为null。

该方法既考虑了句子的语义信息,也考虑了词序信息,实验中得到了不错的准确率,但是这种方法考虑了句子中的所有词,且需要多次计算两个词之间的相似度,时间复杂度很高,如果测试数据比较多,花费的时间相当长。

3 基于关系向量模型的句子相似度计算算法

3.1 关系向量模型

Li Yuhua,David M等[15]提出的算法考虑了句子中的所有词,这在很大程度上增加了算法的运行时间。由语言学知识可知,任何句子都是由关键成分(主、谓、宾等)和修饰成分(定、状、补等)构成的。关键成分对句子起主要作用,修饰成分对句子起次要作用,因此在进行句子相似度计算时只考虑关键成分。在通常情况下,一个句子中作主语和宾语的多为名词或代词,作谓语的多为动词或形容词。因此,将一个句子中的所有名词、代词、动词、形容词和副词作为关键词,并在计算句子相似度方面只考虑这些关键词,这样既可以保证正确率,又可以提高算法的效率。在计算词的相似度时,使用了《知网》的知识资源。

为便于说明,首先给出本文算法的相关概念和定义如下:

知网[16](英文名称HowNet)是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常用知识库,它是一个网状的有机的知识系统。

知识词典是知网系统的基础文件。在这个文件中每一个词语的概念及其描述形成一个记录。每一种语言的每一个记录都主要包含4项内容。其中每一项都由两部分组成,中间以"="分隔。每一个"="的左侧是数据的域名,右侧是数据的值。它们排列如下:

W_X=词语

E_X=词语例子

G_X=词语词性

DEF=概念定义

其中的W_X、G_X、E_X构成每种语言的记录,X用以描述记录所代表语种,X为C则为汉语,为E则为英语。每个词语有DEF来描述其概念定义,DEF的值由若干个义原及它们与主干词之间的语义关系描述组成,义原是知网中最基本、不易于再分割的意义的最小单位。本文中采用刘群、李素建[12]提出的基于《知网》的词汇语义相似度计算方法计算两个词的语义相似度。

定义6 给定一个句子 T_i,经过汉语分词系统分词后,所得到的关键词 m_i 构成的向量称为句子 T_i 的关键词向量表示,表示为 $T_i=\{m_1,m_2,\cdots,m_n\}$。

定义7 给定一个句子 T_i 的关键词向量表示 $T_i=\{m_1,m_2,\cdots,m_n\}$,在向量中关键词 m_i 的前一个关键词 m_{i-1} 称为 m_i 的前关键词,m_i 的后一个关键词 m_{i+1} 称为 m_i 的后关键词。

定义8 给定一个句子 T_i 的关键词向量表示 $T_i=\{m_1,m_2,\cdots,m_n\}$,$T_i$ 的向量长度 $Len(T_i)=n$,给每一个关键词 m_i 赋一个初始权重值 $basevalue=\frac{1}{n}$,所有关键词的权重值构成一个向量称为 T_i 的初始权重值向量,表示为 $TB_i=\{basevalue,basevalue,\cdots,basevalue\}$。

定义9 给定两个句子 T_i、T_j 的关键词向量表示,对于 T_i 中的任一关键词 m_i,如果 m_i 也在 T_j 中出现,则称 m_i 在 T_j 中存在,T_i 中所有在 T_j 中存在的关键词构成的向量称为 T_i 基于 T_j 的存在向量,表示为 $E_{i,j}=\{e_1,e_2,\cdots,e_p\}$。存在向量中相应关键词的权重值构成的向量称为 T_i 基于 T_j 的存在值向量,表示为 $TE_{i,j}=\{v_1,v_2,\cdots,v_p\}$。

定义10 给定两个句子 T_i、T_j 的关键词向量表示,在计算 T_i 和 T_j 的相似度时,在考虑关键词词序和语义的基础上,同时考虑关键词的前关键词和后关键词与该关键词形成的局部结构关系的计算模型,称为关系向量模型。

在上述定义的基础上,接下来详细介绍关系向量模

型的算法实现过程。

3.2 基于关系向量模型的算法实现

关系向量模型不但考虑一个句中的关键词是否在另一个句中出现,还考虑了与这个关键词最紧密的两个词(前关键词和后关键词)的影响,这样,句中所有关键词之间的结构关系得到了体现,因而增加了分析的全面性和准确性。本文提出的基于关系向量模型计算两个句子 T_i 和 T_j 相似度的算法过程如下:

首先利用中科院开源的SharpICTCLAS分词系统[17]分别对 T_i 和 T_j 进行分词,从分词结果中提取 T_i 和 T_j 相应的关键词构成 T_i 和 T_j 的关键词向量,然后考虑向量长度更短的关键词向量,这里假设 $Len(T_i) \leqslant Len(T_j)$,计算 T_i 的初始权重值向量 $TB_i=\{b_1, b_2, \cdots, b_n\}$。对于 T_i 中的每一个关键词 m_i,依次做如下处理:如果 m_i 在 T_j 中存在,考虑 m_i 在 T_i 和 T_j 中的前关键词,如果这两个前关键词为相同的词或者为同义词,则将 TB_i 中 m_i 相应的权重增大 σ 倍,对于 m_i 的后关键词做相同的处理。如果 m_i 在 T_j 中不存在,则 TB_i 中 m_i 的相应权重不变。然后计算 T_i 基于 T_j 的存在向量 $E_{i,j}$,对于 $E_{i,j}$ 中的每个关键词,从 TB_i 中取出相应关键词的权重值构成存在值向量 $TE_{i,j}=\{v_1, v_2, \cdots, v_p\}$。最后根据下列公式(3)计算出两个句子的相似度 $Sim(T_i, T_j)$。

$$Sim(T_i, T_j)=\frac{\sum_{i=1}^{p} v_i}{\sum_{i=1}^{n} b_i} \times \frac{Len(T_i)}{Len(T_j)} \tag{3}$$

其中,b_i 为 TB_i 向量中第 i 项的值,v_i 为 $TE_{i,j}$ 向量中第 i 项的值。

算法的伪代码描述如下所示:

算法 1 基于关系向量模型计算两个句子 T_1 和 T_2 的相似度

输入:两个完整的句子 T_1 和 T_2

输出:T_1 和 T_2 的相似度 $Sim(T_1, T_2)$ $(0 \leqslant Sim(T_1, T_2) \leqslant 1)$

方法:

(1)input two chinese sentences T_1、T_2

(2)foreach sentence T_i (i = 1, 2)

(3) SharpICTCLAS (T_i) //利用中科院SharpICTCLAS分词系统对句子 T_i 进行分词

(4) form a key word vector T_i with key words

(5)set T_1 to be a shorter vector

(6)calculate the initial vector TB_1 value of T_1

(7)foreach keyword m_i in T_1

(8) if m_i exist in T_2

(9) if front keyword of m_i in T_1 and T_2 are the same

(10) $TB_1(m_i)=\sigma \times TB_1(m_i)$

(11) endif

(12) if back keyword of m_i in T_1 and T_2 are the same

(13) $TB_1(m_i)=\sigma \times TB_1(m_i)$

(14) endif

(15) endif

(16)calculate T_1's exist value vector $TE_{1,2}$ according to T_2

(17)calculate similarity of T_1 and T_2

举例说明如下:

给定两个分词后的句子 T_1 和 T_2 如下:

T_1:无论如何/l 我们/r 永远/d 不/d 会/v 使用/v 化学武器/n 对抗/v 自己/r 的/uj 人民/n 。/w

T_2:无论/c 在/p 何/r 种/q 情况/n 下/f ,/w 永远/d 不/d 会/v 对/p 本国/r 人民/n 使用/v 化学武器/n 。/w

针对 T_1、T_2 分词后的结果,可以得到 T_1、T_2 的关键词向量表示如下:

$T_1=\{$我们,永远,不,会,使用,化学武器,
对抗,自己,人民$\}$

$T_2=\{$情况,永远,不,会,本国,人民,
使用,化学武器$\}$

然后,考虑长度更小的向量,如上例的 T_2,首先给出 T_2 的初始权重向量 $TB_2=\{\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\}$。接下来针对 T_2 中的每一个关键词 m_i,按算法1中的步骤(7)进行处理。如对于 T_2 中的关键词“情况”,其不在 T_1 中出现,因此该关键词的权重不变,关键词“永远”在 T_1 中出现,但其在 T_1、T_2 中的前关键词分别为“我们”、“情况”,这两个关键词不同,也不是同义词,因此也不用增大“永远”在 TB_2 中相应的权重。再考虑 T_2 中的关键词“永远”,其在 T_1、T_2 中的后关键词分别为“不”,“不”。这两个词相同,因此在 TB_2 中将关键词“永远”相应的权值增大 σ 倍。因此关键词“永远”最后的权重为 $\sigma \times \frac{1}{8}$。依次类推,$T_2$ 中所有词处理结束后,计算 T_2 基于 T_1 的存在值向量 $TE_{2,1}$。

依次处理完所有词后,TB_2 和 $TE_{2,1}$ 向量取值如下:

$$TB_2=\{\frac{1}{8}, \frac{1}{8}\times\sigma, \frac{1}{8}\times\sigma\times\sigma, \frac{1}{8}\times\sigma, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\times\sigma, \frac{1}{8}\times\sigma\}$$

$$TE_{2,1}=\{\frac{1}{8}\times\sigma, \frac{1}{8}\times\sigma\times\sigma, \frac{1}{8}\times\sigma, \frac{1}{8}, \frac{1}{8}\times\sigma, \frac{1}{8}\times\sigma\}$$

最后,根据式(3)计算 T_1 和 T_2 的相似度:

$$Sim(T_1, T_2)=\frac{\sum_{i=1}^{6} v_i}{\sum_{i=1}^{8} b_i} \times \frac{Len(T_2)}{Len(T_1)}$$

由此,上例 T_1 , T_2 的相似度为:

$$Sim(T_1, T_2)=\frac{\frac{1}{8}\times\sigma+\frac{1}{8}\times\sigma\times\sigma+\frac{1}{8}\times\sigma+\frac{1}{8}+\frac{1}{8}\times\sigma+\frac{1}{8}\times\sigma}{\frac{1}{8}+\frac{1}{8}\times\sigma+\frac{1}{8}\times\sigma\times\sigma+\frac{1}{8}\times\sigma+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\times\sigma+\frac{1}{8}\times\sigma}\times\frac{8}{9}$$

实验中取 $\sigma=1.3$,可得 $Sim(T_1, T_2)=0.71$ 。

3.3 基于关系向量模型的算法与基于语义和词序的算法比较

对于需要计算相似度的两个句子,如果一个关键词同时出现在两个句子中,且该词在两个句子中的前关键词或者后关键词是同一个词或者是同义词,则这两个句子相似的概率更大,因为这种关系反应了句子的局部结构成分,比单纯的词序更能体现句子的语义,下面对此给出理论上的分析。

给定两个句子 T_1、T_2 的关键词向量表示如下:

$T_1=\{a_0, a_1, \cdots, a_i, \cdots, a_m\}$

$T_2=\{b_0, b_1, \cdots, b_i, \cdots, b_n\}$

如果 T_1、T_2 是完全相同的句子,即 $a_i=b_i, m=n$,无论是基于词序还是基于关系向量模型的计算方法,得出的句子相似度都很高,这种情况下两种方法都表现良好。但对于常见的倒装句,可以证明关系向量模型的计算方法比词序法更优越。现假设 T_2 是 T_1 的倒装句, T_2 表示如下:

$T_2=\{b_j, \cdots, b_n, b_0, \cdots, b_{j-1}\}$

这相当于将一个关键词组成的循环链表中的每个关键词向后移 j 位,此时,对于 T_1 中的任意关键词 a_i ,它在 T_1 中的序号为 $(i+1)$, 在 T_2 中的序号为 $(i+j)\%(n+1)$,其中%表示取模。 a_i 在 T_1、T_2 中的序号相差 j, 因此,根据公式(2), T_1、T_2 的词序相似度表示如下:

$$S_r=1-\frac{(n+1)\times j}{\sum_{i=0}^{n}(i+1+(i+j)\%(n+1))}$$

取 $n=9, j=5$ 代入上式可得 $S_r=\frac{4}{9}$,可见倒装句的词序相似度较小,对于关系向量模型的方法而言,由于只是边缘部分个别关键词的前关键词或后关键词有所改变,因此对式(3)的影响较小,由此说明了关系向量模型的方法在处理倒装句的相似度问题上比词序法更优。

算法时间复杂度方面,本文算法取一个句子(如 T_1)中的每个关键词依次与另一个句子中的关键词进行词的相似度计算,因此本文算法的时间复杂度为 $O(m\times n)$,其中 m、n 为 T_1、T_2 的关键词向量长度。基于语义和词序的算法首先生成句子 T_1、T_2 的并集 T , T 的长度记为 k, 显然, $m, n\leqslant k\leqslant m+n$ 。然后针对 T 中的每一个词,依次计算其与 T_1、T_2 中的每个词的相似度,因此,其时间复杂度为 $O((m+n)\times k)$ 。由此可见,本文算法在算法效率方面有一定的提高。

4 实验及结果分析

本文实现的句子相似度的计算方法主要用于在生成网络热点新闻的自动摘要时排除意思相近的句子,从而提高自动文摘的准确率,避免文摘句的冗余。当前人工判断句子相似度很困难且因人而异、因上下文而异。但在网络新闻的自动文摘中,两个描述同一事件的意思相近的表述不能同时出现在自动文摘中,在这种情况下可以将描述同一事件的这种意思相近的表述看成是同义句,从而在自动文摘生成的过程中避免冗余。因此,从当前的热点新闻中抽取了军事、体育、教育等7个分类的相关文章,每个分类4~6篇文章,对于每篇文章,抽取文章标题和文章中最能表达标题意思的一个句子或者正文中两个最能表达文章观点的意思相近的语句作为相似句。从7个分类中总共抽取了35组相似句构成标准集。同时将相应文章中的其他句子以及一些相关新闻中的共1 037个句子作为噪音集。从标准集的35组相似句中每组取出一个句子共35个句子和噪音集合并得到1 072个句子作为测试集。

首先采用文献[4]中提出的实验方法,对标准集中剩下的35个句子,按顺序从中抽出一个句子,然后计算这个句子与测试集中每个句子之间的相似度,并按照所得相似度的大小对测试集中的句子进行排序并输出相似度最大的前3个,最后人为地观测输出结果,如果该句的相似句作为第一个相似句输出,则说明这个句子的相似度计算是成功的。

分别用Li Yuhua,David M等[15]提出的考虑语义和词序的句子相似度计算方法和本文提出的方法做了实验,实验结果如表1所示。其中正确率的计算公式如式(4)所示:

$$\text{正确率}=\frac{\sum\text{测试结果正确的句子总数}}{\sum\text{测试句子总数}}\times 100\% \quad (4)$$

表1 实验结果对比1

方法	测试句子总数	结果正确的句子	正确率/(%)
基于语义和词序的方法	35	25	71.43
关系向量模型算法	35	30	85.71

第二个实验采用邸书灵,刘晓飞等[10]在论文中提出的实验方法,在7个分类35组相似句的每组中再人工加入两个与原来句子较为相似的句子。如此每组就有4个较为相似的句子,其中两个是相似句,另两个是较为相似句。类似实验1,还是从每组的相似句中抽取一个句子组成一个待测试集合,剩下的句子作为测试集。一次从待测试集中选择一个句子和测试集中的句子进行相似度计算,并按顺序排序,输出前三个最相似的句子。实验中假定相似度大于0.7的两个句子为相似句。如果每组中的相似句作为最相似句输出,且相似度大于

0.7,同时较为相似的句子的相似度低于0.7时,认为结果正确。实验结果的正确率也采用式(4)计算。结果如表2所示。

表2 实验结果对比2

方法	测试句子总数	结果正确的句子	正确率/(%)
基于语义和词序的方法	35	28	80.00
关系向量模型算法	35	31	88.57

从上述两个实验的结果可以看出,本文提出的算法所得的正确率高于基于词义和词序的方法。通过分析实验1和实验2中发生错误的语句特征,发现本文算法在处理两个长度相差较大的句子时出现错误的可能性更大,如对于测试集中的两个句子"'平明二号'探测船隶属于越南国家油气集团","中国渔船从后面撞向'平明二号',并将该船的勘探电缆扯断,此举遭到越南国家油气集团的抗议",由于长句的关键词向量几乎包含了短句的关键词向量。因此使用本文算法计算时得出的相似度较大,而事实上这两个句子在文章中并不是相似句。此外,句中存在的缩写词、简称等也会降低本文算法的正确率。如将"印度媒体"简称"印媒",将"中国人民解放军"缩写为"PLA"。通过跟踪算法的执行过程发现,分词系统不能识别一些简称和缩写词,从而将它们拆开,导致关键词向量中不存在简称,从而降低了准确率。一些没有被分词系统拆开的简称在与全称进行同义词判断时出现错误,如全称"俄罗斯军方"被分词系统拆分为"俄罗斯"、"军方",而简称"俄方"与"俄罗斯"、"军方"都不是同义词,这也会导致算法的准确率下降。

5 结束语

本文提出的一种新的关系向量模型不但体现了句子的局部结构成分,同时综合利用了《知网》的知识资源,在对句子较长、动词较多的网络文章的句子进行相似度计算时不但提高了句子相似度计算的准确率,也提高了算法的运行效率。虽然实验取得了较好的正确率,但是由于关键词元组关系并不能完全反应一个句子所包含的所有语义信息且分析的结果也受分词的准确性以及同义词判断的准确性影响,因此为了达到更好的效果,需要进一步研究句子的语法和语义表示方法。

参考文献:

[1] Palakorn A, Hu Xiaohua, Shen Xiajiong.The evaluation of sentence similarity measures[C]//Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery,2008:305-316.

[2] Li Yuhua, Zuhair A B, David M.An approach for measuring semantic similarity between words using multiple information sources[J].IEEE Transactions on Knowledge and Data Engineering,2003:871-882.

[3] Donald M, Susan D, Christopher M.Similarity measures for short segments of text[C]//Proceedings of ECIR, 2007:16-27.

[4] 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.

[5] 穗志方,俞士汶.基于骨架依存树的语句相似度计算模型[C]//中文信息处理国际会议(ICCIP'98),1998.

[6] 郭艳华,周昌乐.一种汉语语句依存关系网协动生成方法研究[J].杭州电子工业学院学报,2000,20(4):24-32.

[7] 裴婧,包宏.汉语句子相似度计算在FAQ中的应用[J].计算机工程,2009,35(17):46-48.

[8] 董自涛,包佃清,马小虎.智能问题系统中问句相似度计算方法[J].武汉理工大学学报:信息与管理工程版,2010,32(1):31-34.

[9] 郑实福,刘挺,秦兵.中文自动问答系统综述[J].中文信息学报,2006,6(16):46-52.

[10] 邸书灵,刘晓飞,李欢.基于分词的语句相似度计算的改进[J].石家庄铁道大学学报:自然科学版,2011,24(4):94-97.

[11] 杨思春.一种改进的句子相似度计算模型[J].电子科技大学学报,2006,35(6):956-959.

[12] 刘群,李素建.基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会,2002.

[13] 张奇,黄萱菁,吴立德.一种新的句子相似度度量及其在文本自动摘要中的应用[J].中文信息学报,2004,19(2):93-99.

[14] 李玉红,柴林燕,张琪.结合分词技术与语句相似度的主观题自动判分算法[J].计算机工程与设计,2010,31(11):2663-2666.

[15] Li Yuhua, David M, Zuhair A B, et al.Sentence similarity based on semantic nets and corpus statistics[J].IEEE Transactions on Knowledge and Data Engineering, 2006:1138-1150.

[16] 董振东,董强.知网(HotNet)[EB/OL].[2012-10-11].http://www.keenage.com.

[17] 中科院.ICTCLAS汉语分词系统[EB/OL].[2012-10-11].http://www.ictclas.org.