一种基于成分的句子相似度计算

郑 诚,夏青松,孙昌年
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:当前信息数据量庞大、冗余度高,如何在自动问答系统中快速查询所需要的信息成为一个关键课题。句子相似度计算作为该领域的一个基础并且是核心的部分,一直受到人们的关注。当前的方法各有其不足之处,文中提出了一种基于成分的句子相似度计算方法。通过将句子划分为主语、谓语、宾语、定语等成分,根据知网计算各个成分间的相似度,最后将所有成分的相似度加权求和得到句子相似度。这种方法不仅能够明显提高句子相似度计算的准确率,同时也极大地降低了计算时的时空消耗,可以有效地提高自动问答系统的准确性。

关键词:句子相似度;句子成分;自然语言处理
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)12-0101-04

Sentence Similarity Calculation Based on Composition

ZHENG Cheng, XIA Qing-song, SUN Chang-nian
(Dept. of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: The current information data has large high redundancy, how to find fast the information needed in automatic question answering system has become a key issue. Sentence similarity calculation as the field of the foundation and the core part, has got the attention of people. In this paper, propose a new method which is based on the composition of sentence. Divide the sentence into subject, predicate, object, attribute and other parts. Calculate the similarity weight between corresponding parts according to HowNet, and the sentence similarity is the summation of all the weight above by some proportion. It not only significantly improves the accuracy of sentence similarity calculation, but also greatly reduces the calculation time and space consumption, and it can effectively improve the accuracy of the automatic question answering system.

Key words: sentence similarity calculation; composition of sentence; natural language processing

0 引 言

在信息检索、机器翻译、自动问答和自动文摘等自然语言处理的领域中,句子相似度计算是一个基本同时也是非常重要的研究课题。从目前国内外的研究成果来看,主要是基于统计学方面的知识。相关句子相似度计算方法主要包括词形相似度、句长相似度、词序相似度、依存关系相似度、基于编辑距离相似度和基于知网的相似度等。文中从句子的语法结构出发,提出一种基于句子成分的相似度计算方法,有效提高了计算的准确率。

1 句子相似度概念及常用的计算方法

句子相似度表示两个句子含义匹配的程度,一般用一个介于0和1之间的数字来表示。1表示两个句子完全匹配,0表示两个句子完全不匹配,句子的相似程度与相似度成正比。

1.1 词形相似度

词形相似度是一种基本的句子相似度计算方法,用两个句子共有的词所占的比例作为句子的相似度[1]。公式如下:

$$word\_sim(X,Y) = 2 | X \cap Y | /(| X | + | Y |) \quad (1)$$

其中,X和Y分别表示要计算相似度的两个句子,X与Y的交集表示两者共有的词语,模表示句子中所有的词语或者字的个数。例如,X代表句子"今年的四六级考试报名的人很多",Y代表句子"四六级今年好像有很多人报名了",X去分词后得到的词语序列是"今年,的,四六级,考试,报名,的,人,很多",同样Y的词语序列是"四六级,今年,好像,有,很多,人,报名,了"。因此有$| X | = 8$, $| Y | = 8$, $| X \cap Y | = 5$, $word\_sim(X,Y) = 2 * 5/(8 + 8) = 0.625$。

1.2 词序相似度

词序相似度根据两句话中字(词语)的顺序来判

收稿日期:2012-03-31;**修回日期**:2012-07-02
基金项目:安徽省自然科学基金资助项目(11040606M133)
作者简介:郑 诚(1964-),男,副教授,硕导,研究方向为语义信息检索、数据挖掘。

定句子相似的程度。通常用一句话中的字(词语)在另外一句话中的逆序数来判定[2]。计算公式如下:

$$\text{order\_sim}(X,Y)=\begin{cases}1-\dfrac{\text{rev\_ord}(X,Y)}{|X+Y|-1} & \text{if } |X+Y|>1\\ 1 & \text{if } |X+Y|=1\\ 0 & \text{if } |X+Y|<1\end{cases} \tag{2}$$

$|X+Y|$ 表示两个句子 X 和 Y 中不同字(词语)的数目,并将所用字(词语)排序,然后计算字(词语)在 X 和 Y 中的排列情况,进而求出句子 X 中的字(词语)在句子 Y 中各相邻分量的逆序数,也就是 rev_ord(X, Y)。最后根据公式(2)可以得出两个句子的词序相似度。例如对句子"今天的股市价格又跌停了"和"今天的股市价格依然没跌停"进行词序相似度计算。以字为基本划分单位的话,$|X+Y|=14$,rev_ord(X,Y) = 2,order_sim(X,Y) = 0.846。这种方法的优点在于考虑了句子中字(词)之间的相对位置关系,当一个句子中的短语或者分句整体发生长距离移动后,句子相似度也不会改变。

1.3 句长相似度

句子的相似度在一定程度上与句子的长度相关。在其他条件相同时,两个句子的长度越接近,句子也越相似。其计算公式为:

$$\text{len\_sim}(X,Y)=1-\left|\frac{|X|-|Y|}{|X|+|Y|}\right| \tag{3}$$

1.4 基于 VSM(向量空间模型)的句子相似度

这是一种最常见的 TF-IDF 方法,其对语料库中的语料进行关键词词频统计,并计算关键词出现的逆文档频率以及当前句子中该关键字的词频,进而得出关键词的权重。然后将句子表示成向量,句子的相似度用权重向量的相似度表示[3]。具体的算法如下:

(1)统计语料库中出现的所有词语 word_1, word_2, word_3, $\cdots$, word_n;

(2)将句子中所有的词项权重表示成一个 n 维的向量:$<w_1,w_2,w_3,\cdots,w_n>$;

(3)其中,$\text{word}_i(1\leqslant i\leqslant n)$ 表示词项 word_i 的权重。$W_i=tf_i*\log(N/d_i)$,tf_i 表示词项 word_i 在一句话中出现的频率,N 表示语料库中句子的总数目,d_i 表示该词项出现的句子个数。$\log(N/d_i)$ 就是常说的逆文档率,简称 IDF。从公式可以看出:词语的权重和在该句子中出现的频率成正比,和出现该词的句子数成反比。

设任意的两个句子表示成的 n 维向量 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$,两者之间的相似度可以用权重向量差的模表示,即 $\text{Sin}(\boldsymbol{V}_1,\boldsymbol{V}_2)=|\boldsymbol{V}_1-\boldsymbol{V}_2|$。现实中更偏向于使用两个向量之间的余弦相似度表示,公式如下:

$$\text{Sim}(v_1,v_2)=\frac{\sum_{i=1}^{n}w_i w_i''}{\sqrt{\sum_{i=1}^{n}(w_i')^2\times\sum_{i=1}^{n}(w_i'')^2}} \tag{4}$$

1.5 基于编辑距离的句子相似度

所谓编辑距离,简单的说就是一个串转换成目标串所需要的最少的编辑操作的数目[4]。编辑距离有三种基本的操作:"插入"、"删除"和"替换"。传统的编辑操作以串中的字为基本单位,比如"爱写程序"与"喜欢写代码"之间的编辑距离是 4。这是因为"爱"替换为"喜",插入"欢","程"转换为"代","序"转换为"码"。编辑距离越大,句子越不匹配。这种简单的基于字为单位的方法,有悖于人们理解句子以词为单位的思维习惯,同时也没有深究词语之间的深层次的联系。例如,"爱"和"喜欢"在一定条件下可以替换,同样"程序"和"代码"也有着更深层次的联系。其次,在句子中添加一个词有时并不会改变句子的意思,或者改变不是很大。因此,插入操作和替换操作具有相同的编辑长度在一定条件下也不是合理的。

1.6 基于依存文法的句子相似度

所谓依存文法就是通过分析句子成分间的依存关系从而揭示其句法结构,认为动词是支配句子中其他部分的核心,动词本身不受任何其他成分的支配,被支配成分从属于支配者[5]。国外专家提出了依存关系的四条基本公理,在研究汉语语言处理时,国内专家提出了第五条公理[6]:

(1)每个句子的成分中只有一个独立;

(2)其它成分直接依存于某一成分;

(3)任何一个成分最多依存于一个成分;

(4)如果 X 成分直接依存于 Y 成分,X 和 Y 中间的成分支配于 Y 或者 X 与 Y 之间的成分;

(5)核心成分两边的成分间没有任何关系。

在使用依存关系进行相似度计算时,只考虑有效搭配对之间的相似程度。有效搭配对指的是句子中的核心词和直接依存于核心词的有效词所构成的搭配对,有效词包括动词、形容词和名词。

相似度计算公式如下:

$$MS(\text{Sen1},\text{Sen2})=\frac{\sum_{i=1}^{n}S_i}{\text{Max}\{\text{PairC1},\text{PairC2}\}} \tag{5}$$

其中分子是两个句子有效搭配对的总权重,Pair $C1$ 是句子 1 的有效搭配对数,Pair $C2$ 是句子 2 的有效搭配对数。

虽然基于依存文法的句子相似度计算有其合理性,但是此方法需要对依存关系进行分析,这不仅需要大量的数据进行训练,另外分析结果也有待提高,在实际中应用比较少。

1.7　基于知网的句子相似度计算

知网是董振东和董强创建的系统的语义知识资源。它是一个以汉语和英语所代表的概念为描述对象,以揭示概念之间以及概念具有的属性之间的关系为基本内容的常用知识库[7]。这是一个网状的知识系统,作为面向汉语计算需求的知识库,为自然语言处理提供了丰富的研究资源[8]。知网中概念是对词语语义的一种描述,每一个词语可能有不同的语义,即一个词可以表达为几个概念[9]。义原是用于描述一个概念的最小意义单位。知网是同对每个概念用一系列义原来表示。如对于"野牛"和"狮子",两者的共性是"动物",而"野牛"的个性是"食草","狮子"的个性是"食肉"、"猫科"等。另外,知网还反映了概念之间和概念的属性之间的各种关系。主要关系有:上下位关系、同意关系、翻译关系、对义关系、部件-整体关系、属性-宿主关系等等。

2　文中提出的基于成分的句子相似度计算方法

词形相似度计算方法统计两个句子中相同词出现的比例,根据该比例的高低判断句子的相关性。这种方法比较简单直观,但并没有考虑到词语之间的内在联系以及词语之间的顺序对句子意思的影响,因此很难真实判断两个句子是否真正相关。最常见的一种缺陷是两个句子的词语组成完全相同,而词语顺序不同,此时可能意思完全相反的句子判断为相同涵义的句子[10],如"霸权的美国打败了独裁的伊拉克"和"独裁的伊拉克打败了霸权的美国"。

词序相似度方法考虑了句子中词语的顺序的重要性,这相对词形相似度方法来说是一种较大的改进,但句子中的所有词语占有的权重比例相同,这和现实情况不是很符合。

句长相似度方法一般和其他的计算方法结合来计算句子的相似度,作为一种辅助的方法。这种方法在一定程度上可以弥补其他方法的不足,但其也有缺点,有时会显得多余[11]。例如对句子"中国正在走向辉煌的明天"和"中华人民共和国正在走向辉煌的明天",两句意思完全相同,采用基于词序与知网相结合的方法计算的相似度很高,若考虑句长的话相似度反而会降低了。

向量空间模型方法根据句子中的词项分布情况来对词项设定相应的权重,然后采用余弦公式计算句子的相似度[12]。但是这种方法也有其缺陷:

(1)词项权重计算方法过于简单,如没有考虑到不同的词性对词项权重的影响;

(2)没有考虑句子词语之间的前后关系,即使相同的词语构成的句子意思可能完全不一样;

(3)由于句子中的词项数目有限,难以满足向量空间模型方法对特征项数目的要求。在实际应用中,这种方法需要大量的语料进行训练,结果并不理想。

其实,无论是汉语还是其他语言,句子的主要成分都是主语、谓语、宾语、定语、状语和补语,我们也是根据句子的成分来理解句子的。因此可以根据两个句子的成分来进行相似度计算:相同的成分越多,句子也就越相似。具体来说就是将句子划分为以下部分:主语 S、修饰主语的定语 SA;谓语 P、修饰谓语的状语 B;宾语 O,修饰宾语的定语 OA 以及补语 C。然后根据知网和同义词词林计算两个句子相同成分间的词语相似度,也就是两个句子主语之间、谓语之间、宾语之间以及其他的修饰成分之间分别进行词语相似度计算,最后将这些相似度进行加权求和,从而得到整个句子的相似度。通过上述计算,可以得到句子的相似度,如果该值大于阈值就可以认为这两句是相似的。另外在句子进行相似度计算之前,需要将"把"字句和"被"字句转换为一般的句子形式,这不仅不会改变句子的意思,同时也更易于处理。

句子的相似度计算公式如下:

$$\mathrm{Sim}(S1,S2) = M + E \tag{6}$$

$$M = \lambda_1 * W(S) + \lambda_2 * W(P) + \lambda_3 * W(O) \tag{7}$$

$$E = \lambda_4 * W(SA) + \lambda_5 * W(OA) + \lambda_6 * W(B) + + \lambda_7 * W(C) \tag{8}$$

其中,$S1$ 和 $S2$ 表示参与相似度计算的两个句子,M 表示句子的主干部分,E 表示句子中除了主干后的其他部分。W 表示相应成分的权重,根据知网判断相应成分间的权重,如果两者含义完全相同或者相应成分都缺失则权重为 1,毫无关系权重为 0,一句中的相应成分缺少权重为 0.5。$\sum_{i=1}^{7}\lambda_i = 1, \lambda_k > \lambda_j (k = 1,2,3; j = 4,5,6,7)$,$\lambda_i$ 为调和因子,之所以主谓宾的调和因子高于其他成分,是因为句子中它们起主要作用,定状补起修饰作用。

3　实验结果与分析

由于没有用于句子相似度计算的开源语料库,所以人工总结了 1000 对句子进行实验。实验之前每对句子已经标注好是否相似,便于统计比较实验结果。

实验的过程如下:

(1)分词和词性标注。

分词的方法比较多,常用的方法有最大正向匹配法、最大逆向匹配法等,这方面中科院和复旦大学的技术比较成熟。在词性标注方面一般基于隐形马尔科夫模型,近似估计词语的词性。由于只作为实验所用,故

采用了复旦大学的开源分词软件 FudanNLP。这款软件可以高效地进行分词并对词性进行标注。实验时，在分词的基础上，根据专有名词词库进行二次分词，从而使得机器的分词结果更接近于人工分词。

(2)句子成分提取。

在词性标注的基础之上，结合句子的结构可以相对容易地划分出句子的成分。首先找出句子中的动词，如果有那么其为谓语。在没有动词的情况下，第一个名词或者代词后的形容词为谓语。谓语之前的名词或者代词为句子的主语，主语之前的形容词为修饰主语的定语。在主语和谓语之间的部分为状语，谓语后面的名词、代词为宾语，宾语之前的形容词为修饰宾语的定语。在谓语后面的其他成分为补语。“的”字前面是定语，“地”字前面是状语，“得”字后面是补语。根据这些基本规则对句子进行成分提取。提取之后，人工纠正错误的部分。

(3)成分相似度计算。

句子中的每一个成分其实就是一个词语或者短语，相应成分间的相似度可以根据词语的语义相似度来判定。知网全面阐述了词语的语义信息，其根据义原的上下位关系组成了多个不同的语义树。在同一语义树下的两个义原所在的层次越近，含义越相似；同一节点下的叶子越多，两个义原也越相似；在不同的语义树上的两个义原的相似度为 0。对于同一语义树上的两个义原 A 和 B 的相似度计算公式如下：

$$\mathrm{Sim}(A,B) = H/(P_1 + P_2)\eta_1\eta_2 \tag{9}$$

其中，

$$H = |\eta_1 \cup \eta_2| - |\eta_1 \cap \eta_2| \tag{10}$$

H 表示语义距离；η_1 表示树的深度系数；η_2 表示树的密度系数；$|\eta_1 \cup \eta_2|$ 表示从根节点到两个节点的中间节点数；$|\eta_1 \cap \eta_2|$ 表示从根节点到两个节点的公共节点数。对于测试的两句中有一方缺少相应成分，则该成分间的相似度设为 0.5；双方都没有相应成分，则该成分间的相似度为 1；其他情况根据公式(9)和(10)得到成分的相似度。

(4)句子相似度计算。

得出成分相似度以后，根据公式(6)计算出句子间的相似度。实验中，调和因子 $\lambda_1 = \lambda_2 = \lambda_3 = 0.2$，$\lambda_4$ 到 λ_7 设为 0.1，句子相似度阈值设为 0.7。

实验结果如表 1 所示：

表 1 实验结果对比

方法	测试句子	结果正确的句子	正确率
词形与句长结合	1000 对	832	83.2%
依存文法	1000 对	894	89.4%
句子成分	1000 对	941	94.1%

实验中采用正确率进行评估，其计算公式如下：

$$\text{正确率} = \frac{\text{测试正确句子数}}{\text{总句子数}} \tag{11}$$

从实验结果中，可以看出基于成分的句子相似度方法能够明显地提高结果的正确率。

4 结束语

文中采用了一种基于成分的句子相似度计算方法，此方法综合考虑了词语间的含义联系以及句子的成分分析，隐含了词语顺序、词语间的依存关系和句子长度，因此更加符合人们理解句子的方式。这种方法提高了句子相似度计算的准确率，同时也大大提高了时空效率。由于需要依赖于句子的成分分析，因此下一步的研究重点是句子成分的自动分析。

参考文献：

[1] 吕学强，任飞亮，黄志丹，等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报：自然科学版，2003，24(6)：531-534.

[2] 周法国，杨炳儒. 句子相似度计算新方法及在问答系统中的应用[J]. 计算机工程与应用，2008，44(1)：165-178.

[3] 胡明涵. 面向领域的文本分类与挖掘关键技术研究[D]. 沈阳：东北大学，2009.

[4] 黄 品，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.

[5] Li Yuhua，McLean D，Bandar Z A，et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics[J]. IEEE transactions on knowledge and data engineering，2006，18(8)：1138-1150.

[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，2000，20(4)：24-32.

[7] 董振东，董 强. 知网[EB/OL]. 2003. http://www.keenage.com.

[8] 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(1)：46-48.

[9] 吴全娥，熊海灵. 一种综合多特征的句子相似度计算方法[J]. 计算机系统应用，2010(11)：110-114.

[10] 李 彬，刘 挺. 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(12)：15-17.

[11] Moschitti A，Pighin D，Basili R. Engineering of Syntactic Features for Shallow Semantic Parsing[C]//Proceedings of the ACL05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing. Ann Arbor (MI)，USA：[s. n.]，2005：48-56.

[12] Collins M，Duffy N. Convolution Kernels for Natural Language [C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Table of Contents. Barcelona，Spain：[s. n.]，2004.