

基于多特征的汉语句子相似度计算模型的研究

李春梅, 徐庆生

(云南楚雄师范学院 计算机科学系, 云南 楚雄 675000)

摘要: 句子相似度的计算在自然语言处理的各个领域中都占有很重要的地位。文中深入分析了现有的一些句子相似度计算的方法, 这些方法各自从词特征、词义特征或句法特征等某一侧面描述了句子相似的情况, 未能全面地描述一个句子的完整信息。文中提出了一种新的基于多特征的汉语句子相似度的计算模型。该方法在基于词的基础上, 从句子中词的表层到词的逻辑联系, 从句子的局部结构到整体结构, 用句子的区分度、相同词的相似度、长度相似度、词性相似度及词序相似度五个方面来综合考虑两个句子相似度的计算。实验结果表明, 该方法合理、简便、可行。

关键词: 自然语言处理; 区分度; 词性; 词序; 句子相似度

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2014)06-0136-04

doi: 10.3969/j.issn.1673-629X.2014.06.034

Research on Chinese Sentence Similarity Calculation Model Based on Multi-features

LI Chun-mei, XU Qing-sheng

(Department of Computer Science, Chuxiong Normal University, Chuxiong 675000, China)

Abstract: Sentence similarity calculation plays an important role in various areas of natural language processing. Analyze the existing some sentence similarity calculation method. These methods describe the sentence similarity from the word characteristics, semantic features or syntactic features, all the information of a sentence can't be described fully. A new model of Chinese sentence similarity based on the multi-feature is proposed. This method is based on the word, from the surface to the logical connection of the word, from local structure to the overall structure of a sentence, five aspects of sentence similarity such as degree of differentiation, the same word similarity, length similarity, the part of speech similarity and word order similarity have been studied in depth. Experimental results show that the method is reasonable, simple and feasible.

Key words: natural language processing; degree of differentiation; discrimination part of speech; word order; sentence similarity

0 引言

句子相似度的计算在自然语言处理的各个领域中都占有很重要的地位, 它的研究状况直接决定着其他一些相关领域的研究进展。在基于实例的机器翻译中, 相似度主要用于衡量文本中词语的可替换程度, 句子相似度衡量的准确性, 直接影响到最后翻译结果的正确性; 在多文档文摘系统中, 相似度可以反映出局部主题信息的拟合程度, 将与用户相关的文档形成文摘结果提交给用户, 帮助用户在较少的时间获得较多的信息, 大大提高了获取信息的效率; 在自动问答中, 相似度反映的是问题与答案的匹配程度; 而在信息检索中, 句子的相似度更多的是反映文本与用户查询在意

义上的符合程度, 它是从相关文档中检索出与查询相关的句子。在这些领域中, 句子相似度计算是一个非常关键的问题, 如果相似度能得到进一步提高, 这些领域都将受益。

1 句子相似度计算方法分析

伴随着自然语言处理相关领域的发展, 到目前为止, 关于汉语句子相似度的计算, 有很多的方法被提出来, 比较典型的相似度计算模型是布尔模型和向量空间模型(VSM)。布尔模型因为只有0和1两种相关度, 采用精确匹配, 并不能提供更细微的排名, 因此不能全面反映用户的需求。向量空间模型(VSM)采用

夹角余弦来衡量相似程度,可以根据计算得到的相似度按从大到小的顺序对句子信息排序,是一种“松弛匹配”。它的缺点是假设句子中所有的词项都是独立的,而事实上句子的词项之间常常是具有某种内部关联的。向量空间模型也不能解决同义词的问题,例如,如果一个句子中有“计算机”一词,而另一个句子中有“电脑”一词时,这两个将被看成两个不同的词。另外相似度的计算量大,当有新文档加入时,则必须重新计算词的权值与相似度。哥伦比亚大学的 Goldsdein 等人通过最大边缘相关的方法(Maximal Marginal Relevance)进行相似度计算^[1],学者 Chris H. Q. Ding 等人采用了隐含语义索引(Latent Semantic Indexing)的方法^[2],穗志方等利用骨架依存的方法^[3]计算汉语句子间相似度,李彬等使用语义依存的方法^[4],其他研究者提出了基于统计的方法^[5]和基于词性的方法^[6]等,另外,也有一些改进的方法^[7-15]。

经过深入的分析,以上这些不同的句子相似度计算方法归结起来主要可概括为三大类:基于词特征的句子相似度计算、基于词义特征的句子相似度计算,以及基于句法分析特征的句子相似度计算,这三类方法也各自反映出了句子的三个重要特征:词特征、词义特征以及句法特征。但是,这三类方法都存在着各自的缺点,给计算带来了一定的误差和不便:基于关键词特征的方法体现了句子表面的信息,要求词形上的精确匹配,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。基于语义特征的方法体现了组成句子的每个词的深层语义信息,但需要一定的语义知识资源作为基础,资源通常是由手工构建的,需耗费大量时间和人力。基于句法特征的方法借助于树形图体现了句子中词与词之间的相互依存关系,能从某种程度上达到意义理解的层次。但是,用从属树来进行自动生成时,根据从属树的互斥条件,从属树中节点之间的支配关系和前于关系是互相排斥的,必须把表示句子层次结构的从属树按照一定的规则转变成按线性排列的句子,无论是生成树还是转换规则,其实现过程也是相当复杂。

针对以上所述问题,在文中提出了一种新的汉语句子相似度度量方法,该方法是从多角度,在基于词的基础上,从句子中词的表层到词的逻辑联系,从句子的局部到整体,对句子各个方面的相似作了深入的研究,在理论上更具有合理性。实验表明,该方法简便,可行。

2 句子描述模型

在处理过程中,把句子定义成一个二元组。

定义1:一个句子可描述为一个二元组 $S = \{W_i,$

$PW_i\}$ 的模型结构。

其中, S 表示句子; W_i 表示句子中已经过切词处理后得到的若干词; PW_i 表示第*i*个词相对应的词性,如动词、名词、形容词等。

3 句子相似度计算模型

通过对以上句子模型的深入分析与研究,在文中以词为基础,用五个特征来度量句子的相似度,下面分别从五个方面来进行详细的描述。

3.1 句子区分度计算

句子由一组不同含义的单词组成,不同词降低了两个句子的相似度,但是能区别出不同的句子,区别越大,相似度越低。

定义2:句子区分度。

句子区分度反映了两个句子之间的不相似程度。定义其值为一个0~1之间的数值,0表示两个句子完全相似,1表示两个句子完全不相似,数值越小表示两句的区分度越小,句子也就越相似。设 Q_s 是用户输入的查询句子, R_s 是可能的检索结果句子,经过分词后, Q_s 的词序序列为 $(qw_1, qw_2, \dots, qw_m)$ 、 R_s 的词序序列为 $(rw_1, rw_2, \dots, rw_n)$,则定义句子 Q_s 与句子 R_s 的区分度为:

$$\text{difsim}(Q_s, R_s) = \frac{\text{nnotword}}{|(\text{llen}(Q_s.w) \cup \text{llen}(R_s.w))|} \quad (1)$$

其中, $\text{llen}(Q_s.w)$ 是指用户输入查询句子的所有词; $\text{llen}(R_s.w)$ 是结果句子的所有词; nnotword 指两个句子之间不同词的总个数; $|(\text{llen}(Q_s.w) \cup \text{llen}(R_s.w))|$ 表示两个句子中只出现一次的所有词的集合。两个句子中的不同词越多,区分度越大,说明两个句子的相似度越低。

例如下面两个句子:

我/r% 去/v% 看/V% 电影 n%

他/r% 要/v% 买/V% 衣服 n%

两个句子的不同词总个数 $\text{nnotword} = 8$, $|(\text{llen}(Q_s.w) \cup \text{llen}(R_s.w))| = 8$, 计算结果这两个句子的区分度=1,表明这两个句子完全不相似,符合生活中的实际情况。

3.2 公共词串相似度计算

在两个句子中,相同词体现了两个句子的共同点,对两个句子相似起到了较大的贡献作用,两个句子中相同词在整个句子中的相似度公式为:

$$\text{comkeysim}(Q_s, R_s) = \frac{\text{ncomword}}{|(\text{llen}(Q_s.w) \cup \text{llen}(R_s.w))|} \quad (2)$$

其中, $\text{llen}(Q_s.w)$ 、 $\text{llen}(R_s.w)$ 、 $|(\text{llen}(Q_s.w)$

$|\cup|len(Rs.w)|$ 与公式(1)中的含义一样; $ncomword$ 表示在两个句子中都出现的公共词串的个数。

3.3 句子长度相似度计算

以上的区分度和公共词串只是从词表面来考虑,如果仅从词法上来度量相似,则这样的相似只是表层的相似,鉴于以上考虑,对于句子相似,可从句子的整体特征上来考虑。从整体来说,句子长度会影响句子的相似度,如果两句子的长度差变大时,则相似度会变低;两个句子的长度越是接近,则两个句子越相似,因此从整体上考虑句子的相似程度。两个句子如果长度一样,分词一样,则句子相似度的值为 1。

$$lensim(Qs,Rs)=1-\frac{abs(|len(Qs.w)|-|len(Rs.w)|)}{\max(|len(Qs.w)|,|len(Rs.w)|)} \tag{3}$$

3.4 词性相似度计算

词性是语言中的词在语法意义上的性别,它表示词所属的类别,是语言的基本结构,词性不同,其词义及作用也不同。

如以下两个句子:中国人发明了造纸术;造纸术的发明有重大意义。这两个句子中的“发明”词性不同,在句子中所起的作用不同,在第一句中是动词,作谓语,在第二句中是名词,作主语。

设 Qs 是用户输入的查询句子, Rs 是可能的结果句子,经过分词后, Qs 的词语序列及词的词性为 $(qw_1/\%ws_1,qw_2/\%ws_2,\cdots,qw_m/\%ws_m)$ 、 Rs 的词语序列为 $(rw_1/\%ws_1,rw_2/\%ws_2,\cdots,rw_n/\%ws_n)$,则定义句子 Qs 与句子 Rs 的词性相似度定义为:

$$pswsim=\frac{psmatchcount}{eword} \tag{4}$$

比较规则:对两个句子从词的最左边起始位置开始,依次进行比较,如果词性相同,就匹配, $psmatchcount$ 表示词性匹配的总数目, $eword$ 表示两个比较的句子中分词较少的句子的词个数,如果其中有一个句子的所有词都比较完了,则整个比较就结束。以下是三种匹配情况:

①输入句子与结果句子一样长。

例 1 Qs : 购买笔记本电脑。切分后: 购买/ $v\%$ /笔记本/ $n\%$ /电脑/ $n\%$

Rs : 联想笔记本电脑。切分后: 联想 $n\%$ /笔记本 $n\%$ /电脑 $n\%$

对应的词性序列如下:

Qs : $v\ n\ n$

Rs : $n\ n\ n$

$$pswsim=2/3=0.667$$

②输入句子比结果句子短。

例 2 Qs : 购买电脑。切分后: 购买/ $v\%$ /电脑/ $n\%$

Rs : 联想笔记本电脑。切分后: 联想 $n\%$ /笔记本 $n\%$ /电脑 $n\%$

Qs : $v\ n$

Rs : $n\ n\ n$

$$pswsim=1/2=0.5$$

③输入句子比结果句子长。

例 3 Qs : 购买联想笔记本电脑。切分后: 购买/ $v\%$ /联想/ $n\%$ /笔记本/ $n\%$ /电脑/ $n\%$

Rs : 笔记本电脑。切分后: 笔记本/ $n\%$ /电脑/ $n\%$

Qs : $v\ n\ n\ n$

Rs : $n\ n$

$$pswsim=1/2=0.5$$

3.5 词序结构相似计算

汉语与英语在语法上有明显的不同。英语的词形变化非常丰富,依靠词形变化表达句子丰富多彩的语言关系和逻辑关系。而汉语没有词的形态学的变化,不靠词形变化表达语法意义,它主要靠词语、词序及暗含的逻辑关系来表达句子的语言意义。句法结构相同,用词相同的句子,如果词序变化了,句子的意义也随之变化。例如以下两对句子:我请朋友去吃饭,朋友请我去吃饭;我打你,你打我。很明显,这两对句子中每组句子的用词是一样的,句法结构也一样,但受事者与施事者不一样,导致了语法关系与修辞意义不同,它们所表达出的句子意义也就不同。因此,词的逻辑顺序即词序在句子中所起的作用与意义是重大的,通过句子中的词序区别语义是最自然、最便捷、最节约的方法,也是经济、高效率的方法。

定义 3:词序相似度。

词序相似度反映了两个句子中所含相同词在位置关系上的相似程度。定义其值为一个 0 ~ 1 之间的数值,数值越大两个句子越相似。

设用户输入句子 Qs , 结果句子 Rs , 设 $comword = (cw_1,cw_2,\cdots,cw_t)$ 表示这两个句子经过分词后的相同词的集合, $|comword|$ 表示这个集合中词的个数, $Qsnum$ 为 $comword$ 集合中的词在 Qs 中的词序向量; $Rsnum$ 为 $comword$ 集合的词在 Rs 中的词序向量; $pairnum$ 为 $Qsnum$ 和 $Rsnum$ 中相同词序向量的个数,则定义 Qs 与 Rs 的词序相似度为:

$$ordersim(Qs,Rs)=\frac{pairnum}{|comword|} \tag{5}$$

例如句子:我请朋友去吃饭,其词序如表 1 所示。

表 1 例句词序 1

我	请	朋友	去	吃饭
0	1	2	3	4

句子:朋友请我去吃饭,其词序如表 2 所示。

表2 例句词序2

朋友	请	我	去	吃饭
0	1	2	3	4

comword=(我 请 朋友 去 吃饭)
则 Qsnum=(0,1,2,3,4)
Rsnum=(2,1,0,3,4)
根据以上相似度公式,上例中这两个句子的词序相似度为0.6。

3.6 句子综合相似度计算模型

综合以上五个方面的考虑,给出如下具有多特征的汉语句子综合相似度计算模型。

Zsim(Qs,Rs)= αdifsim(Qs,Rs) + βcomkeysim(Qs,Rs) + θlensim(Qs,Rs) + λpswsim(Qs,Rs) + γordersim(Qs,Rs) (6)

其中,α,β,θ,λ,γ分别是区分度的权重,公共词相似的权重,句子长度相似的权重,词性相似度的权重及词序相似度的权重,0 ≤ α ≤ 1,0 ≤ β ≤ 1,0 ≤ θ ≤ 1,0 ≤ λ ≤ 1,0 ≤ γ ≤ 1,且α + β + θ + λ + γ = 1。这是可调节的值,考虑区分度和句子长度对整个句子相似的贡献小,公共词和词性及词序对两个句子相似的贡献大,因此取α = 0.1,β = 0.2,θ = 0.1,λ = 0.3,γ = 0.3,突出了公共词、词性与词序在句子中的作用,因此其权重大。

3.7 算法描述

输入:要计算相似度的查询句子 Qs 和一系列的结果句子 Rs_i;

输出:Qs 和 Rs_i的相似度。

步骤1:对输入的两个句子 Qs 和 Rs_i进行分词,得到分词后的字符串新句子 WQs 和 WRs_i;

步骤2:分别在 WQs 和 WRs_i中,从最左端开始依次读取每一个词;

步骤3:分别计算两个句子的区分度、公共词串相似度、句子长度相似度、句子词性相似度和句子词序相似度;

步骤4:计算句子 Qs 和 WRs_i的综合相似度并输出。

4 实验结果及分析

4.1 实验用例

为了验证这个句子相似度方法的有效性,做了一个模拟实验,实现语言用Java,数据库用Sqlserver。该实验是针对有关计算机购买行为的,通过在Google上输入句子,在返回的20个句子中选取了15个句子作为待选的结果描述句子,表3只列出其中四个待选的结果描述句子的相似度结果。经过分词系统分词后,

应用文中的计算公式(1)至(6),得到如表3的数据。

表3 实验用例							
输入句子 Qs	结果描述句子 Rs _i	difsim	comkey-sim	lensim	pswsim	ordersim	Zsim
	购买/v%/台式/n%/电脑/n%	0.5	0.5	1	1	1	0.85
购买/v%/笔记本/n%/电脑/n%	购买/v%/计算机/n%	0.75	0.25	0.67	1	1	0.79
	联想 n%/笔记本 n%/电 脑 n%	0.5	0.5	1	0.67	1	0.75
	笔记本 n%/电 脑 n%/购 买 v%/指南 n%	0.25	0.75	0.75	0.33	0	0.35

4.2 实验分析

从计算所得的结果可判断出,对于用户方的描述句子来说,用户目的是要购买电脑,而且是笔记本式的,句子1和句子2都突出了购买这一动作,句子1把电脑式样“具体化”为台式,句子1比句子2的相似度更高,更能深刻反映用户的意图。如在该系统中加入同义词(电脑=计算机),则句子2的相似度将进一步提高到0.833,在句子1和句子3中,当它们的区分度、长度、共同词及词序相同时,句子的词性相似度起到了区别的效果,句子4与用户的意图相差较远,该句子要表现的是买笔记本电脑时的指导思想,而不是要买笔记本电脑。

5 结束语

文中提出了一种新的基于多特征的汉语句子相似度衡量方法,该方法避免了语义本体的手工建立和从属树及其转换规则的复杂性操作,从词的角度出发,从句子的区分度、相同词的相似度、句子长度相似度、词性相似度及词序相似度五个方面来综合考虑两个句子相似度,句子的区分度与相同词的相似度体现了句子的局部信息,句子长度相似度体现了句子的整体信息,词性相似度及词序相似度体现了句子深层蕴含的语法、逻辑联系与修辞关系。该方法合理、简便、可行,具有一定的实用价值。

参考文献:

[1] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries [C]//Proceedings of ACM-SIGIR'98. New York, NY, USA: ACM,1998;335-336.
[2] Ding C H Q. A similarity-based probability model for latent semantic indexing[C]//Proc of 22nd international ACM SIGIR conference on research and development in information

总体而言,文中构造的联想词表可以提高检索系统的准确率,具有一定的实用性。


4 结束语

文中提出了一种基于 PMI-IR 的联想词表自动构造方法。该方法首先利用 Prefix Span 算法进行频繁项集的挖掘,生成候选联想词表;然后利用 PMI-IR 算法计算词间相似度,构造联想词表;最后利用检索词与文档的相关性算法比较了利用 PMI-IR 算法得到的联想词表与利用简单聚类得到的联想词表的扩展查询效果。实验表明,利用 PMI-IR 算法得到的联想词表进行扩展查询具有更好的效果。

参考文献:

- [1] 陆 勇,章成志,侯汉清. 基于百科资源的多策略中文同义词自动抽取研究[J]. 中国图书馆学报,2010,36(1):56-62.
- [2] Salton G. The smart retrieval system - experiments in automatic document processing[M]. Upper Saddle River, NJ: Prentice Hall, 1971.
- [3] Gauch S, Wang Jianying. Corpus analysis for TREC 5 query expansion[C]//Proceedings of the text retrieval conference. [s. l.]: [s. n.], 1996.
- [4] Schutze H, Pedersen J O. A co-occurrence-based thesaurus and two applications to information retrieval[J]. Information Processing and Management, 1997, 33(3): 307-318.
- [5] Crouch C J. A cluster-based approach to thesaurus construction[C]//Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM, 1988: 309-320.
- [6] Crouch C J. An approach to the automatic construction of global thesauri[J]. Information Processing and Management, 1990, 26(5): 629-640.
- [7] 陈建超,郑启伦,李庆阳,等. 基于特征词关联性的同义词集挖掘算法[J]. 计算机应用研究, 2009, 26(7): 2517-2519.
- [8] 吴云芳,石 静,金 澎. 基于图的同义词集自动获取方法[J]. 计算机研究与发展, 2011, 48(4): 610-616.
- [9] Turney P D. Mining the web for synonyms: PMI-IR versus LSA on TOEFL[C]//Proceedings of the twelfth European conference on machine learning. London, UK: Springer-Verlag, 2001: 491-502.
- [10] Pei Jian, Han Jiawei, Mortazavi-Asl B, et al. Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth[C]//Proc of 17th international conference data engineering. [s. l.]: [s. n.], 2001: 215-224.
- [11] 余慧佳,刘奕群,张 敏,等. 基于大规模日志分析的网络搜索引擎用户行为研究[J]. 中文信息学报, 2007, 21(1): 109-114.
- [12] 岑荣伟,刘奕群,张 敏,等. 基于日志挖掘的搜索引擎用户行为分析[J]. 中文信息学报, 2010, 24(3): 49-54.
- [13] 陈红涛,杨放春,陈 磊. 基于大规模中文搜索引擎的搜索日志挖掘[J]. 计算机应用研究, 2008, 25(6): 1663-1665.
- [14] 窦志成,袁晓洁,何松柏. 大规模中文搜索日志中查询重复性分析[J]. 计算机工程, 2008, 34(21): 40-41.
- [15] 曹 雷,郭嘉丰,白 露,等. 基于半监督话题模型的用户查询日志命名实体挖掘[J]. 中文信息学报, 2012, 26(5): 26-32.
- [16] 王荣波,池哲儒. 基于词类串的汉语句子结构相似度计算方法[J]. 中文信息学报, 2005, 19(1): 21-29.
- [17] 车万翔,刘 挺,秦 兵,等. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯, 2004, 14(7): 15-19.
- [18] 王 宇,战学刚,蔡建山. 基于网络的中文问答系统的研究[J]. 计算机工程与应用, 2006(7): 162-165.
- [19] 闰宏飞,陈 钟. 词汇与中心词的距离信息对问句相似度匹配的影响[J]. 清华大学学报(自然科学版), 2005, 45(S1): 1873-1877.
- [20] 李文杰,赵 岩. 基于本体结构的概念间语义相似度算法[J]. 计算机工程, 2010, 36(23): 4-6.
- [21] 刘宝艳,林鸿飞,赵 晶. 基于改进编辑距离和依存文法的汉语句子相似度计算[J]. 计算机应用与软件, 2008, 25(7): 33-34.
- [22] 周永梅,陶 红,陈姣姣,等. 自动问答系统中的句子相似度算法的研究[J]. 计算机技术与发展, 2012, 22(5): 75-78.
- [23] 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(ICCIIP'98论文集). 北京:清华大学出版社, 1998: 458-465.
- [24] 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.
- [25] Chatterjee N. A statistical approach for similarity measurement between sentences for EBMT[C]//Proceedings of symposium on translation support systems. [s. l.]: [s. n.], 2001: 45-48.
- [26] 蓝雁玲,陈建超. 基于词性及词性依存的句子结构相似度计算[J]. 计算机工程, 2011, 37(10): 47-49.
- [27] Li Sujian, Zhang Jian, Huang Xiong, et al. Semantic computation in a Chinese question-answering system[J]. Journal of Computer Science and Technology, 2002, 17(6): 933-939.
- [28] 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, 2005, 19(2): 93-99.

基于多特征的汉语句子的相似度计算模型的研究

作者: 李春梅, 徐庆生, [LI Chun-mei](#), [XU Qing-sheng](#)
作者单位: [云南楚雄师范学院 计算机科学系, 云南 楚雄, 675000](#)
刊名: [计算机技术与发展](#) 
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2014(6)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjfz201406034.aspx