句子相似度计算新方法及在问答系统中的应用

周法国,杨炳儒
ZHOU Fa-guo,YANG Bing-ru

北京科技大学 信息工程学院,北京 100083
School of Information Engineering,University of Science and Technology Beijing,Beijing 100083,China

ZHOU Fa-guo,YANG Bing-ru.New method for sentence similarity computing and its application in question answering system.Computer Engineering and Applications,2008,44(1):165-167.

Abstract: Sentence similarity computing plays an important role in machine question-answering systems,machine-translation systems,text categorization systems,etc.Aiming at a sentence similarity model based on key words,an improved method is put forward,including the extraction of keywords,and the induction of synonyms in sentence similarity definition.And on this basis,a question answer system based on FAQ(Frequently Asked Question) is implemented.This system involves automatically searching for candidate question set,computing sentence similarity and returning the answer to the user.This system can also automatically update and maintain FAQ.Experiments' result shows that the new method has more accuracy than the others in matching questions of question answering system.

Key words: natural language processing;sentence similarity;Frequently Asked Question;question answer

摘 要:计算句子的相似度在机器问答、机器翻译、文本分类等系统中有着非常重要的作用。该文对基于相同关键词的句子相似模型作了进一步的改进,包括关键词抽取,以及在句子相似度的定义中引入同义词以及近义词的情形。并以此为基础,实现了一个基于常问问题集的中文自动问答系统,对用户以自然语言输入的问题,该系统能够自动地在FAQ(Frequently-Asked Question)库中寻找候选问题集,通过计算句子相似度,将匹配的答案返回给用户。该系统还能够自动地更新和维护FAQ库。实验结果表明,这种新方法在问答系统中匹配问句时比其他方法具有较高的准确率。

关键词:自然语言处理;句子相似度;常问问题集;问答系统
文章编号:1002-8331(2008)01-0165-03 **文献标识码**:A **中图分类号**:TP391

1 引言

在自然语言处理领域,尤其是在中文信息处理中,句子相似度计算是一项基础而核心的研究课题,长期以来一直是人们研究的一个热点和难点。句子相似度计算在现实中有着广泛的应用,它的研究状况直接决定着其他一些相关领域的研究进展,句子相似度的计算在自然语言处理的各个领域都有着非常重要的作用,如在基于实例的机器翻译系统[1]中、在文档自动文摘系统[2]中、在基于常见问题集(FAQ)的机器问答系统[3]中以及信息检索、信息过滤等方面,句子相似度的计算都是其中关键的技术之一。本文给出了一种计算句子相似度的新方法,并给出了该方法在问答系统中的应用,设计并实现了一种简单的基于常问问题集的中文问答系统。

2 句子相似度计算的新方法

2.1 常见句子相似度计算方法

在相似度计算中,按照对语句的分析深度来看,主要存在两种方法:(1)基于向量空间模型的方法,即基于词的方法。该方法把句子看成词的线性序列,不对语句进行语法结构分析,相应的语句相似度衡量机制只能利用句子的表层信息,即组成句子的词的词频、词性等信息[4]。由于不加任何结构分析,该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。(2)基于语义的方法,对语句进行完全的句法与语义分析,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,找出语义依存关系,并在依存分析结果的基础上进行相似度计算[5]。本文是在基于词的方法的基础上充分考虑了同义词与近义词。

2.2 关键词抽取

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。进行句子相似度计算时,只要考虑句中的关键成分。基于词的方法不考虑句法结构分析,因此,不能确定句子的内部成分,包括关键成分和修饰成分。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,可以将一个句子中的所有名词、代词、动词和形容词作为关键词,并在计算句子相似度时只考虑这些关键词。例如,句子“我 当然 愿意 了解 她们 的

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675030)。

作者简介:周法国(1976-),男,博士研究生,主要研究方向为自然语言处理,知识发现与智能系统;杨炳儒(1943-),男,教授,博士生导师,主要研究方向为知识发现与智能系统,柔性建模与集成技术。

要求。”的关键词序列为“我 愿意 了解 她们 要求。”。对于特定句中的某个名词、代词、动词或形容词,不一定就是该句中的主语、宾语或谓语成分,但相对于句中所有的词构成的词序列而言,关键词序列却具有一定的句法结构信息表达能力,至少可以了解句子中的哪些词在组成句子框架结构方面是比较重要的。在此基础上进行相似度计算,比一般基于词的方法更准确一些。

2.3 有关定义和计算

汉语句子就是一个字符串,是由一组不同含义的单词组成,它不同于数值型变量,可以用一个特定的数值来确定它的大小或位置,所以用何种方式来描述两个字符串之间的距离,成为了一个值得探讨的问题。

通常情况下,用于分析的数据类型有如下几种:区间标度遍历、二元变量、标称型变量、序数型变量、比例标度型变量、混合类型变量等。

综合这些变量类型,本文认为字符串变量更适合于归类于二元变量,我们可以利用分词技术将字符串分成若干个单词,每个独立的单词作为二元变量的一个属性。把所有单词设定为一个二元变量属性集合 R,字符串 1 和字符串 2 的单词包含于这个集合 R。设 q 是字符串 1 和字符串 2 中都存在的单词的总数,s 是字符串 1 中存在,字符串 2 中不存在的单词总数,r 是字符串 2 中存在,字符串 1 中不存在的单词总数,t 是字符串 1 和字符串 2 中都不存在的单词总数。称 q,r,s,t 为字符串比较中的 4 个状态分量。如图 1 所示。

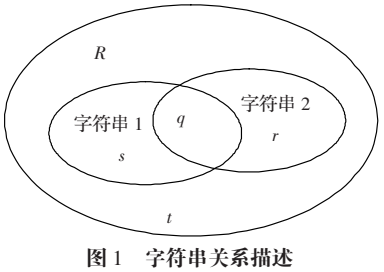

图 1 字符串关系描述

由于两个字符串都不存在的单词对两个字符串的比较没有任何作用,所以忽略 t,于是采用非恒定的相似度评价系数(Jaccard 系数)来描述两个字符串间的相异度表示公式为:相异度=$(r+s)/(q+r+s)$,不难推断,他们的相似度公式为:相似度=$q/(q+r+s)$。

由此,可以得到句子的词形相似度。句子的相似度除了与关键词有关外,还与句子长度、句子中关键词的顺序、关键词之间的距离有关,下面给出具体的定义与计算方法。

定义 1 词形相似度 $WordSim(S_1,S_2)$

从句子形态以及词形上来标注句子的相似性,反映句子形态上的相似性。$WordSim(S_1,S_2)$表示 S_1 与 S_2 中相同关键词的个数。则词形相似度可以根据 Jaccard 系数来计算。其计算方法如下:

$$WordSim(S_1,S_2)=\frac{SameWord(S_1,S_2)}{Word(S_1)+Word(S_2)-SameWord(S_1,S_2)}$$

其中,$SameWord(S_1,S_2)$表示 S_1 与 S_2 相同关键词的个数,如果同一关键词出现多次则只算一次,其中的关键词不包含句子中的疑问词及停用词表中的词,如:为什么、怎么样、如何、的、地、得等。$Word(S_i)$表示 S_i 中的关键词个数,i=1,2。

在实践过程中发现名词和动词在句子中起着非常重要的作用,并且名词比动词承载着更多的信息量。一个句子的中心信息基本上都是围绕着动词和名词来展开的,所以在进行计算的时候也特意加大了名词和动词的重要程度,将句子的重心落在名词和动词上面。这样,在此处计算相同关键词的个数时,若两个词相同并且都是名词,相同个数以 5 计,若两个词相同并且都是动词,相同个数以 3 计,在计算 S_i 中的关键词个数时,名词的个数也按 5 计,动词个数以 3 计,即一个名词实际出现一次计算为 5 次,一个动词实际出现一次计算为 3 次。编程时,对每个句子分词后,然后要进行词性标记从而区分是否为名词和动词。同时为了更进准确的计算句子的相似度,我们引入了同义词词典。如:句子“怎么杀计算机病毒?”和句子“怎么杀电脑病毒?”是基本一样的。其中“计算机”和“电脑”是同义词。

定义 2 句长相似度 $LenSim(S_1,S_2)$

从句子长度上来标注句子的相似性,在一定程度上也反映句子形态上的相似性。其计算方法如下:

$$LenSim(S_1,S_2)=1-\text{绝对值}\left(\frac{Len(S_1)-Len(S_2)}{Len(S_1)+Len(S_2)}\right)$$

其中 $Len(S_i)$表示 S_i 中(关键)词的个数,i=1,2。

定义 3 词序相似性 $OrdSim(S_1,S_2)$

从关键词的顺序上来标注句子的相似性,反映两个句子中所含相同词或同义词在位置关系上的相似程度,以两个句子中所含相同词或同义词的相邻顺序逆向的个数来衡量。其计算方法如下:

$$OrdSim(S_1,S_2)=1-\frac{Rev(S_1,S_2)}{MaxRev(S_1,S_2)}$$

其中,$MaxRev(S_1,S_2)$:表示 S_1 与 S_2 相同关键词的个数的自然数序列的最大逆序数,例:若 S_1 与 S_2 相同关键词的个数为 4,则自然数序列为{4,3,2,1},它的逆序数为 6。$Rev(S_1,S_2)$:表示 S_1 中关键词在 S_2 中的位置构成的自然数序列的逆序数。

反映两个句子中所含相同词或同义词在位置关系上的相似程度,以两个句子中所含相同词或同义词的相邻顺序逆向的个数来衡量。设 S_1、S_2 为两个句子,$OnceWord(S_1,S_2)$为 S_1、S_2 中所含相同词或同义词的集合,重复出现的词仅计一次,$P_{first}(S_1,S_2)$为 $OnceWord(S_1,S_2)$中的词在 S_1 中出现关键词的先后顺序所构成的向量(为一自然数顺序序列,重复出现的关键词计第一次出现),$P_{second}(S_1,S_2)$为 $P_{first}(S_1,S_2)$中的分量按对应词在 S_2 中的次序排序生成的向量,$RevOrd(S_1,S_2)$为序列 $P_{second}(S_1,S_2)$的逆序数。

定义 4 距离相似性 $DisSim(S_1,S_2)$

从相同关键词的距离上来标注句子的相似性。其计算方法如下:

$$DisSim(S_1,S_2)=1-\text{绝对值}\left(\frac{SameDis(S_1)-SameDis(S_2)}{Dis(S_1)+Dis(S_2)}\right)$$

其中 $SameDis(S_i)$表示 S_1,S_2 中相同的关键词在 S_i 中的距离,i=1,2。若关键词重复出现多次,以产生最大距离为准。

$Dis(S_i)$:表示 S_i 中非重复关键词中最左及最右关键词之间的距离,i=1,2。若关键词出现多次,以产生最小距离值为准。

定义 5 句子相似度

反映两个句子之间的相似程度。通常为一个 0~1 之间的数值,0 表示不相似,1 表示完全相似,数值越大表示两句越相似。

记两个要比较的句子为 S_1 和 S_2,S_1 与 S_2 的相似度记为 $SenSim(S_1,S_2)$,则:

$SenSim(S_1,S_2)=\lambda_1 WordSim(S_1,S_2)+\lambda_2 LenSim(S_1,S_2)+\lambda_3 OrdSim(S_1,S_2)+\lambda_4 DisSim(S_1,S_2)$

其中:$\lambda_1+\lambda_2+\lambda_3+\lambda_4=1$ 且 $\lambda_1\geq 0.5\geq\lambda_2\geq\lambda_3\geq\lambda_4>0$。

2.4 算法描述

算法 一种改进的计算句子相似度计算算法

输入:要计算相似度的两个句子 S_1 和 S_2

输出:S_1 和 S_2 的相似度

步骤 1 对输入的两个句子 S_1 和 S_2 进行分词,得到字符串 S_1' 和 S_2';

步骤 2 从 S_1' 和 S_2' 中得到两个句子相同或相近的关键词;

步骤 3 计算词形相似度、句长相似度、词序相似度和距离相似度;

步骤 4 求取句子 S_1 和 S_2 的相似度。

与其他算法相比,该算法中的关键词抽取部分涉及分词与词性标注(其他算法大部分仅涉及分词),在计算词形相似度时还需要借助一部同义词词典。该算法具有以下特点:

(1)简单,所利用的信息仍为句子的表层信息。

(2)保留了其他已有算法的优点,可以保证句子中的分句或短语整体移动后仍与原来的句子相似。

(3)比原算法更准确,所抽取的关键词可以近似地表达部分句法结构信息。

3 基于常问问题集的中文问答系统

中文问答系统的研究开始于 20 世纪末,最近 10 年是中文问答系统的高速发展期,众多学者在中文问答系统方面做了大量的研究,取得了大量有益的研究成果,主要有基于本体的中文问答系统[6],基于语义相似度的中文问答系统[7],知识驱动的中文问答系统[8],基于数据挖掘的中文问答系统[9],基于检索的中文问答系统[10]以及聊天机器人、基于检索的问答系统,各种形式的网络答疑系统,客户服务系统等等。其中基于知识库的问答系统是其中最主要的一种,基于知识库的问答系统中基于 FAQ 的问答系统是最常见的一种。

3.1 基于 FAQ 的中文问答系统的流程

在目标问句进入基于 FAQ 的问答系统之前,需要将中文句子分成词语的集合。分词部分包括对库中问题的分词,也包括对目标问句的分词。然后通过建立知识库的全文检索,选择与目标问句比较相似的一小部分集合,在这个小集合中进行相似度计算,即计算各个句子与目标问句的相似度。选择相似度的最大值,与设定的阈值进行比较。如果大于设定的闭值,则返回该答案,如果小于设定的阈值,则不返回答案,通过信息检索、答案抽取等技术来更新问题库。大致流程如图 2 所示。

3.2 候选问题集的建立

这一步骤的目的是要从常问问题库(FAQ)中找出若干个候选的问题组成候选问题集,以缩小查找的范围,使后续的相似度计算等较复杂的处理过程都在候选问题集这个相对较小的范围内进行。在系统中,问题集存储在 Sql Server 2000 数据库中,在建立候选问题集时,我们采用了 Sql Server 2000 数据库管理系统自带的全文检索系统。首先,对用户输入的目标问句进行分词、关键词抽取,过滤掉停用词后,对关键词在问题域字段上进行全文检索,把和目标问句相关的记录中的问句作为候选问题集。

3.3 FAQ 库的更新

利用 2.3 中介绍的方法计算出用户所输入的目标问句和候选问题集中每个问句的相似度,如果所有这些计算出来的相似度的最大值大于或等于一定的阀值 m(m=0.65),那么就认为最大的相似度所对应的问句和用户的目标问句问的是同一个问题。可以直接将这个问句对应的答案输出给用户。

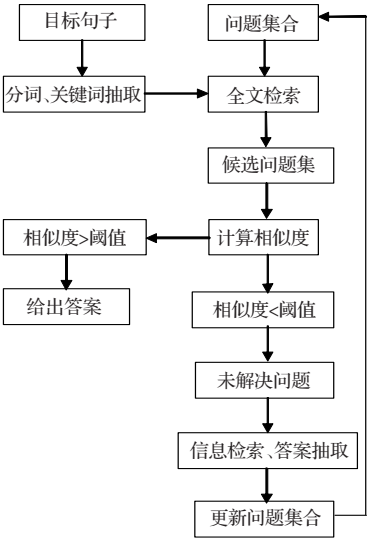

图 2 问答系统一般流程

如果最大相似度的值小于阀值 m(m=0.65),就可以认为 FAQ 库中没有用户所问的问题,那么必须利用其他的方法(如信息检索,答案抽取等)来找出答案。如果能够找到答案,就可以将用户所问的这个问题和对应的答案加入 FAQ 库。

4 实验结果

算法在基于 FAQ 的机器问答系统中应用,在有 1 千多个问题的问题集中进行测试,取 λ_1=0.6,λ_2=0.2,λ_3=0.1,λ_4=0.1,匹配问句时选择相似度(阈值)大于等于 0.65 的问题中相似度最大的问句,将其答案返回。对相似度小于 0.65 的问句,则认为问题集中没有该问题的答案。测试平均准确率在 85%以上,比文献[4]中基于词形和词序的计算方法匹配问句要高出 10 个以上左右的百分点。

在基于 FAQ 的中文问答系统中,选择了 3 个人进行独立测试,每个人随机地选择 100 个问题进行测试,测试结果如表 1 所示。

表 1 实验测试结果

测试人	测试问题数	问题平均长度	准确率/%
1	100	12.3	81
2	100	9.8	89
3	100	10.9	85

5 结束语

在计算句子相似度时,通过关键词抽取、以及扩充同义词词典和加大名词和动词在句子中的重要性可以明显地提高计算的准确性,自动分词和词性标注的质量也直接影响本方法的准确率。本文在一定程度上提高了计算句子相似度的正确率,但并没有对句子的语法、句法、语义等方面进行详细的分析,如

(下转 178 页)