句子相似度计算及其应用①

景 红, 岳群琴

(西南交通大学 信息科学与技术学院, 成都 610031)

摘 要: 随着互联网技术的发展, 智能答疑系统也受到了更多的关注, 应用它能够及时给学生提供学生疑惑的问题答案. 智能答疑系统通常包括问句理解、信息检索、答案抽取和选择三个主要部分, 其中句子相似度计算是问句理解的一部分, 它的性能将直接影响到最后答案的准确性. 本文通过对词型和普通的编辑距离算法为基础, 加入了词性的语义信息, 提出了一种新的句子相似度算法, 并将其应用到计算机基础课程答疑系统中, 使得系统的正确率有了较大的提高.

关键词: 句子相似度; 词型; 编辑距离; 智能答疑

Sentence Similarity Computation and Application

JING Hong, YUE Qun-Qin,

(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: With the development of Internet technology, the Intelligent Answering also has drown more attention, because it can solve user unsure questions timely. Intelligent Answering system typically includes comprehension questions, information retrieval and answer extraction and selection. Sentence Similarity Computation is a part of comprehension questions, which directly affects the accuracy of the final answers. This is based on the morphological pattern and improved edit-distance method, adding semantic information of words and taking weight into account to the new sentence similarity method in. And Intelligent Answering is implemented, to make the correct rate of the system greatly improved.

Key words: sentence similarity; morphological pattern; edit-distance; intelligent answering

随着互联网技术发展, 智能答疑系统也越来越受到关注. 对于用户而言, 在通过计算机基础课程进行学习的过程中, 会遇到一些疑惑问题, 往往只能通过留言或者邮件等待老师的答疑. 而利用智能答疑却能够很好地解决用户的上述情况, 及时对用户遇到的一些难点问题进行答疑. 本文关于计算机基础课程的智能答疑, 让用户能够快速查找问题对应的答案.

就答疑系统而言, 在答疑过程中, 当用户输入问题时, 需要通过相关算法查找到用户所需要的答案, 这就需要涉及到句子相似度计算. 现在很多领域都涉及到这种方法, 例如, 答疑系统中进行问题答案匹配; 在基于实例机器翻译中得到需要的译文等等. 因此, 该方法的研究一直是智能答疑系统领域内的热点问题, 尤其是关于中文的句子相似度计算.

对于文本进行相似度计算时, 通常因含有的信息量比较大, 导致计算工作量比较大, 以及准确率也会比较低. 而对于问句而言, 通常实际含有的信息相对会比较少, 一般在十几个字内, 由此进行问题相似度计算时, 其准确率也会比较高. 目前, 研究句子相似度的方法有很多种, 比如基于语义依存, 基于编辑距离, 基于词型等等. 本文主要是对后两种方法进行研究. 其中, 基于词型的方法没有考虑到句子的语义特征, 而编辑距离只考虑了字词之间的变换代价, 具体的操作还是比较单一.

① 基金项目:中央高校基本科研基金(A092050205130425)
收稿时间:2014-03-11; 收到修改稿时间:2014-04-14

一个句子不仅包含语义信息，而且包括词汇信息. 本文主要以词型和编辑距离为基础，添加词性权重来改进计算方法，使得句子相似度的正确率有了比较大的提高.

1 句子相似度计算方法研究

句子相似度[1]是指两个句子的匹配符合程度，通常把计算的结果规范到实数[0,1]之间，0 代表完全不相同，1 代表完全符合要求. 相似度越接近 1，说明句子越相似度，也就越符合要求. 而句子相似度计算有三种计算方法，分别是基于语法、基于语义和基于语用这三种计算方法[2]. 由于语用相似度具有一定的难度，目前的研究还处于初级阶段，所以现阶段主要研究基于语法和语义这两种方法. 本文主要研究基于词型和基于编辑距离的两种方法.

1.1 基于词型相似度计算方法

基于词型的相似度计算方法[3]主要考虑两个句子当中所含有相同词的多少，通过计算两个句子之间所含有相同词汇量来表示相似度. 词型相似度计算公式如下:

Sim1(A,B)=Same(A,B)/(Len(A)+Len(B)-Same(A,B))

其中 A,B 表示两个句子，Same(A,B)表示含有相同单词个数，Len(A),Len(B)表示句子的长度，Sim1(A,B)表示句子相似度.

在进行句子匹配时，通过使用以上公式进行句子的相似度匹配，然后选取计算结果最大的值作为相似句子返回. 但是这种计算方法只考虑了句中的词型因素，并没有考虑句中的语义因素，而问句中的语义因素在问句当中占有很大的比例，如果没有考虑语义因素，通常计算结果会因句中一些变化而会产生一定的误差.

1.2 基于编辑距离的相似度计算方法

编辑距离[4]是指从一个以字为单位的句子经过变换后需要最小的编辑操作个数. 编辑操作共有“插入”，“删除”和“替换”三种. 在计算编辑距离时，考虑两个句子通过三种操作后使句子相同的最小操作代价. 但是，这种以字为操作的编辑距离，没有考虑词汇相关信息，使得操作结果具有很大的出入. 一般在使用编辑距离进行相似度计算时，通常考虑以词为单位进行编辑[5]. 在进行编辑距离计算时，通常使用 HowNet 和《同义词林》这两种资源. 同时还分为几种不同的语义体系，在这个语义体系中，每个词汇都有相应的语义编码，通过编码来计算词间的距离[6]. 同时，对于每个词汇来说，在同义词林中可能含有多个相近的词汇，这时需要考虑多个词汇中的最小编辑距离，使得取得的编辑距离的代价最小. A，B 两个句子计算距离的公式如下:

$$Dist\ (A,B) = \min_{a\in A, b\in B}(dist\ (a,b))$$

a, b 之间的距离为:

dist(a,b)=2*(4-n)

其中 a, b 表示两个词，n 为两个词的语义代码从第几层开始不同. 这样计算代价如表 1 所示.

表 1 编辑距离相关操作代价

编辑操作	操作代价
A->A	0
A->A’	0.4
A->A”	Dist(A,A”)/10+0.5
其他	1

其中，->表示操作，A’表示由 HowNet 中的语义信息进行转换的同义词，A”表示由同义词林中的语义集合进行转换的同义词.

通常基于编辑距离的方法采用动态规划的方式来计算两个句子之间的最小编辑距离，其计算公式如下，其中 dist(t)表示替换代价.

$$dist(i,j)=\begin{cases}0\\ Min\begin{cases}dist(i,j-1)+1\\ dist(i-1,j)+1\\ dist(i-1,j-1)+dist(t)\end{cases}\end{cases}$$

计算所得 dist(n,m)为两个句子间的最小编辑距离.

1.3 改进的句子相似度计算方法

虽然一个问句的通常含有比较少的文本信息，但是它一定含有能体现出问句主体信息的语义和词型信息. 所以，问句相似度通常包含词型和句法信息，它既关注于句子中相同词型的多少，又关注于句子中语法信息. 本文基于词型和编辑距离的两种方法，将这两部分信息通常加入到句子相似度计算过程中.

对于上文中的编辑距离算法，由于计算的结果在中文变换时有很大的出入，比如，问题“什么是主存”和“主存是什么”，在对着两个问题句计算句子相似度时，其相似度会比较低. 为了解决该问题，本文采用支持非相邻块交换的编辑距离算法，在计算句子相似

度过程中，考虑了非相邻块词之间的变换[7]，将句型变换因素考虑进去，这样，当问句句型变换时，计算出的句子相似度也比较准确，不会因为语型变换因素而造成比较大的误差，使得结果不准确.

一般一个问题的主要的成分是名词和动词. 在计算编辑距离过程中，对各个词性的编辑距离采用的权重相同，都为 1，不能体现出名词和动词在问句中的重要程度. 本文在计算编辑距离的过程中，对于名词的编辑权重为 5，动词的编辑权重为 3，其他词性的编辑权重为 1.

通过上述改进的编辑距离算法，与原始的编辑距离算法相比，在计算两个句子相似度时，返回给用户最大的相似度，相对后者而言正确率有了很大的提高. 由于编辑距离通操作不灵活，使得其正确率也有一定的差距.

通过上述编辑距离的改进，然后融合基于词型的方法进行相似度计算，其相似度计算方法得公式如下:

$$Sim(A,B)= \alpha Sim1(A,B)+(1-\alpha)Sim2(A,B)$$

其中 α 表示各部分在计算过程中的权重，Sim2(A,B)表示编辑距离相似度计算结果，Sim1(A,B)表示词型相似度计算结果. 通过计算可知，当 α=0.6 时，该算法获得最高的正确率，三种算法计算相似度的典型代表句如表 2 所示.

表 2　各种方法句子相似度计算结果

句子对	词型相似度算法	编辑距离算法	本文算法
什么是主存 内存是什么	1.0	0.33	0.91
显卡和显示器有什么区别 显示器与显卡有什么异同	0.67	0.76	0.73
人会感染电脑病毒吗 什么是计算机病毒	0.33	0.55	0.45
什么是分组 分组有什么特点	0.75	0.25	0.59

综上所述，本文算法同时将基于词型和编辑距离两种算法的优点融合和克服两者间的不足，使得相似度计算的结果具有较高的准确率，取得了比较满意的计算结果.

2　句子相似度在计算机基础课程答疑中的应用

2.1　计算机基础课程答疑流程

计算机基础课程答疑流程图，如图 1 所示，其具体流程为: 当用户输入问句时，系统首先通过分词模块进行问题分词和词性标注，将分词的结果传给问句理解模块; 问句理解模块根据所接收到的分词结果进行问句分类和关键词提取，然后将处理的结果传给答案生成模块; 答案生成模块根据上一模块的处理结果进行候选集生成，以及进行问句相似度比较，然后将答案返回给用户，同时对问题库进行更新. 通过上述的流程，便能将与用户匹配的答案返回给用户，进而及时解决用户在学习过程中遇到的问题，提高用户的学习效率.

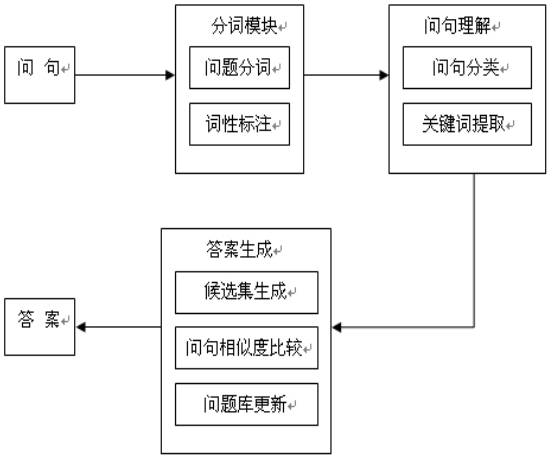

图 1　答疑流程图

2.2　句子理解模块

问句理解模块主要根据分词模块处理的结果进行下一步操作，主要包含问句分类和关键词提取. 其中，问句分类子模块主要根据关键词进行问句分类，以便在候选处理过程中能够让问句在同一类别中进行相似度匹配，而减小了问句比较的规模，提高了运行效率; 关键词提取子模块主要是对问句中的分词结果对于噪声词进行处理，比如助词等，处理后达到减小关键词的数目，提高在比较过程中的运行速率.

2.3　答案生成模块

答案生成模块将根据问句理解模块的处理进行下一步操作，主要包含候选集生成,问句相似度比较和问题库更新. 其中: 候选集生成子模块主要根据问句分

类结果, 将与用户输入问题的类别相同的问题作为相似度比较的候选集; 问题相似度比较模块在该候选集中进行问句的相似度匹配, 并返回满足阀值而且相似度最大的问句和答案; 问题库更新子模块负责对问题库中的相关信息进行更新.

3 实验结果

本文以大学计算机基础课程为背景, 进行了句子相似度计算在答疑系统中的应用研究和测试. 首先, 作者对该门课程的专业术语进行了收集, 以便在进行问句分词过程中, 能够快速准确的进行分词. 实验根据用户输入的问句采用基于词型相似度算法, 编辑距离算法, 和本文改进算法这三种方法进行相似度匹配, 并返回答案给用户. 同时, 采用正确率对这三种算法进行实验评价, 正确率的定义如下:

正确率=正确回答问题数/测试问题数*100%

本文测试的问题数为 50, 其中 10 条在问题库中没有答案, 10 条在问题库中有完成的问题和答案, 其他 30 条问题则是经过问句的变形提问并在问题库中都有答案. 问题库都是有关计算机基础课程的相关问题, 总共有 600 条. 三种算法的测试结果如表 3 所示. 本文对于问题相似度的阀值, 设置为 0.65, 即当句子相似度计算结果大于 0.65 时, 认为该句子与问句相似.

表 3 实验结果

方法	测试句子	结果正确句子数	正确率
词型相似度算法	50	37	74%
编辑距离算法	50	32	64%
本文算法	50	40	80%

3 结语

本文采用了基于词型和编辑距离结合的算法对句子相似对进行计算, 不仅融合了词汇信息和语义信息, 而且在计算编辑距离过程中, 对于不同的词性赋予不同的编辑权重, 这主要考虑问句中, 一般主要结构式名称和动词, 需要对这类词添加在句中的权重, 实验结果也获得了比较好的效果. 但是, 本文尚存在不足, 一是在进行分词和词性标注过程中, 由于有些词汇在不同的问句中词性不同, 而分词工具不能明确的识别出其在问句中的词性, 这也就影响到了编辑距离的大小, 进而影响到句子相似度的大小, 造成了一定的误差; 二是在使用《同义词林》进行问句语义分析时, 由于该词典收藏的词汇有限, 而且对于一些专业术语的同义词理解有误, 对某些问句进行句子相似度计算时会有较大的误差. 为了进一步提高实验结果的正确率, 在以后的工作中, 需要对这些不足的方法进行改进, 使得句子相似度的正确率得到进一步的提升.

参考文献

1 杨思春,陈家骏.中文自动问答中句子相似度计算研究.情报学报,2008(1):35–41.
2 周永梅,陶红,陈姣姣,等.自动问答系统中句子相似度算法的研究.计算机技术与发展,2012,(5):75–78.
3 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.计算机工程与应用,2008,(1):165–167.
4 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.高通技术通讯,2004,7:15–19.
5 赵作鹏,尹志民,王潜平,等.一种改进的编辑距离算法及其在数据处理中的应用.计算机应用,2009,(2):424–426.
6 裴倩,包宏.汉语句子相似度计算在 FAQ 中的应用.计算机工程,2009,(17):46–48.
7 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.计算机应用与软件,2008,(7):33–37.