doi: 10.3969/j.issn.1007-855x.2010.02.017

基于自然语言处理的主观题评分算法研究

梁晓诚[1],岳晓光[1],麦范金[1],赵子强[2],路 英[3],王 挺[4,5]
(1.桂林理工大学 信息科学与工程学院,广西 桂林 541004;2.太原科技大学 机械电子工程学院,太原 030024;
3.莫纳什大学 管理学院,澳大利亚 维多利亚州 墨尔本 3800;4.利物浦大学 计算机科学系,英国 利物浦 L697ZF;
5.西交利物浦大学 计算机科学与软件工程系,江苏 苏州 215123)

摘要:从自然语言处理的角度来看,现有的主观题评分算法都以相似度计算为核心的,没有考虑语义对立度问题.提出了一种基于中文分词技术、相似度计算和对立度计算的新的主观题评分算法.对主观题评分算法的以下三个方面进行了重点研究:怎样改进中文分词算法中的歧义切分的方法;怎样引入参数限制计算中分数过高或过低;怎样设计一个融合相似度计算和对立度计算的主观题分数的计算公式.

关键词:主观题;相似度;对立度
中图分类号:TP391 **文献标识码**:A **文章编号**:1007-855X(2010)02-0081-04

Algorithm Research of Subjective Question Assessment Based on Natural Language Processing

LIANG Xiao-cheng[1], YUE Xiao-guang[1], MAI Fan-jin[1], ZHAO Zi-qiang[2],
LU Ying[3], WANG Ting[4,5]
(1. School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi 541004, China;
2. Mechanical and Electronic Engineering College, Taiyuan University of Science and Technology, Taiyuan 030024, China;
3. Department of Management, Monash University, Melbourne 3800, Australia; 4. Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, UK; 5. Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China)

Abstract: From the perspective of natural language processing, all existing subjective questions scoring algorithms use similarity calculation as a key method while the contrary degree is ignored. A new subjective questions scoring algorithm based on Chinese words segmentation technology, similarity calculation and contrary degree calculation is presented in this paper. The three areas about subjective questions scoring algorithm are mainly researched: how to improve the method of disambiguation in Chinese words segmentation; how to introduce parameters to restrict the score being too high or too low in calculation; and how to design a formula to combine similarity calculation and contrary degree calculation for subjective questions scoring.

Key words: subjective question; similarity; contrary degree

0 引 言

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向[1].自然语言处理可以认为是一种多学科交叉的语言文字的处理技术.运用自然语言处理技术可以发现语言文字所表达的"义"[2].通过分析这个"义",分析语句得到相关数据,用来给主观题评分.从另外一个角度来说,自然语言处理的理解方式与人工阅卷也是相似的,人工阅卷是通过比较考生答案和参考答案,并受到阅卷老师情绪等因素影响下给出的分

收稿日期:2009-06-26.**基金项目**:广西自然科学基金资助项目(项目编号:桂科自 0991254).
第一作者简介:梁晓诚(1958-),男,高级工程师.主要研究方向:计算机应用技术. **E-mail:** lxc@glite.edu.cn

数,而基于自然语言处理的主观题阅卷在同样的情况下,就可以很好地避免一些不公平的因素的影响.

目前的主观题评分主要采用的方法都是侧重于相似度计算[3-5]方面,没有考虑对立度的问题.本文引入对立度的概念,避免了相似度计算中因为词、句子出现重复率高而导致的评分偏高现象.对立度在阅卷中的意义不仅仅限于限制分数,而且表明了学生观点和参考答案的对立程度.主观题评分算法的第一个需要考虑的问题就是中文分词,其需要解决的主要问题就是歧义切分;第二个需要考虑的问题就是如何将中文分词、相似度计算、对立度计算结合起来并用公式表达出来.

1 主观题评分算法相关理论

1.1 最大匹配算法

分词方法通常分为基于统计的分词方法、基于词典的分词方法和二者相结合的分词方法.基于词典的分词方法最常用的是最大匹配算法.考虑到分词算法实现的难度,这里使用最大匹配分词算法,因为最大匹配算法通过一个词表就可以实现分词的目的[6].正(逆)向最大匹配算法是从待处理的字符串的开头(末端)开始匹配,查找分词的词表,如果词表中有这个词,则匹配成功,否则去掉最后(最前)的一个字符,直到匹配成功,如此循环进行,匹配整个待处理的字符串[7].

1.2 词形相似度

对于语句 A 和语句 B,它们的相似度通常是由词形相似度决定的,词形相似度的计算公式[8]如公式(1)所示:

$$WordsSim(A,B) = \frac{SameWords(A,B)}{\max(Length(A),Length(B))} \tag{1}$$

其中 $SameWords(A,B)$ 代表相同的单字长度,$\max(Length(A),Length(B))$ 代表语句 A 和语句 B 中最大词长.

1.3 语义对立度

对于词语 a 和词语 b,记 $Similarity(a,b)$ 为词语 a 和词语 b 的相似度,且 $Similarity(a,b) \in [-1,1]$.

于是当 $Similarity(a,b) = -1$、$Similarity(a,b) \in (-1,0)$、$Similarity(a,b) = 0$、$Similarity(a,b) \in (0,1)$、$Similarity(a,b) = 1$ 分别表示词语 a 和词语 b 为反义词、不完全相对的反义词、无相似性、近义词、同义词.当 $Similarity(a,b) \in [-1,0)$ 时,记为 $Contrary(a,b)$,称之为语义对立度[9].

2 主观题评分算法设计

算法设计是基于自然语言处理的思想,自然语言处理既然属于人工智能的一个范畴之内,所以仍然是用计算机来模拟人的思维.

考虑到人的思维的联想性,文章设计的算法的思路如下:首先是比较参考答案 $keysAnswer$ 和学生答案 $studentsAnswer$,通过相似观点(或者说得分点)给出应得分数;然后比较学生答案和参考答案中的不合适的观点,即扣除应扣分数,最后得出学生答案的最终得分.所以,可以转换为一个分词过程加上一个相似度的计算问题,然后比较学生答案和参考答案的语义对立度,给出得分.

2.1 分词算法设计

分词算法的设计有两个问题要解决,那就是分词规范、歧义切分.分词规范主要是用在词表设计上,分词规范决定着词的切分形式.歧义切分的解决将在算法描述里面讨论.

2.1.1 词表设计

词表设计是分词的基础,可以采用 GB/T 13715-92信息处理用现代汉语分词规范[10]这套标准,这样保证切分形式的统一性和规整性.分词规范确定之后,词表设计可以采用数据库的里面的表来实现.表的属性列要有词的频次记录,为精确分词做准备.表要有一个反义词选项用来计算语义对立度.

2.1.2 分词算法描述

分词算法可以采用双向匹配算法,即同时进行正向和逆向匹配,如果切分结果一致,则认为切分成功,否则存在歧义.以逆向最大匹配算法为例,描述如下:

设 maxLength为最大切分长度,则有下面的步骤.

①从待处理 NewString字符串末端选择长度为 maxLength的字符串 String, CurrentLength = maxLength;

②把 String与词表 Dictionary进行比较;

③如果 Dictionary中存在这个字符串,则 String为词, NewString = NewString - String (即除去 String,并得到新的 NewString),返回①;

④如果 Dictionary中不存在这个字符串,并且 CurrentLength > 1,则得到新的 String,此时 CurrentLength = CurrentLength - 1,返回②;如果 CurrentLength = 1,则得到单字,从 NewString末端选取长度为 CurrentLength的 String,返回②;

⑤如果某个 String一直匹配到 CurrentLength = 1,仍然无法识别,当做歧义识别.

总体而论,逆向最大减字匹配算法是一个基于词表的分词算法,通过长词优先的原则进行切分,达到了获取词的目的,为下一步语义相似度的计算打下基础.

2.1.3 歧义切分问题

切分时,例如形如 ABC的字串,本应该切分 AB/C,但根据长词优先的原则,可能会被切分为 A/BC等形式,这样就出现了歧义切分错误现象. 歧义问题经常使用到的两个概念是:交集型歧义字段和链长. 交集型歧义字段指是形如 ABC可以被切分成 AB或者 BC的字段. 链长是交集型歧义字段中含有交集字段的个数,例如前面的 ABC的链长为 1.

目前歧义切分尚无十分精确的算法来解决,根据实际情况,我们遇到大概 85%的歧义字段是属于交集型字段,遇到链长为 1和 2的字段占歧义字段的 95%以上,所以解决这两个问题就能解决大多数歧义切分的问题. 这里引入两条规则[11].

【规则一】链长为 1的交集型歧义字段 ABC约有 80%切分为 AB/C或 A/BC,有计算公式(FreqLeft表示左边字段出现的概率, FreqRight表示右边字段出现的概率),有公式(2):

$$FreqLeft(AB) + FreqRight(C) > FreqLeft(A) + FreqRight(BC) \tag{2}$$

若式子成立,则切分为 AB/C,否则切分为 A/BC;

【规则二】链长为 2的交集型歧义字段 ABCD约有 98%的概率切分为 AB/CD.

2.1.4 算法改进原则

两个常用规则引入算法中,由规则一, FreqLeft和 FreqRight对应词表中的词频,考虑到可能有 17%的概率出现 ABC仍然切分为 ABC (剩下 3%的概率不再考虑),因此比较时要引进公式(3):

$$|FreqLeft(AB) + FreqRight(C) - FreqLeft(A) + FreqRight(BC)| \geq \alpha \tag{3}$$

若式子成立,则切分为 ABC. 其中 α是一个经验值,是一个极小的值,即当 FreqLeft和 FreqRight概率接近时,切分为 ABC. 另外,由规则二, ABCD直接切分为 AB/CD. 剩下的情况不再考虑.

2.2 相似度计算

分词结果出现之后,基于公式(1),利用分词所得的词的 keysAnswer和 studentsAnswer集合,计算学生答案和参考答案的相似度 WordsSim(keysAnswer, studentsAnswer). 对于第 i道题,相似度高于 h_i的学生给出最高分 $highestScore_i$,语句相似度低于 l_i ($l_i < h_i$)的学生最低分 $lowestScore_i$,相似度在 (l_i, h_i)之间的学生按照公式(4)给出学生的分数. 于是对于第 i道题,学生的应得分数 $addScore_i$公式是:

$$addScore_i = \frac{WordsSim(keysAnswer, studentsAnswer)_i}{h} \times highestScore_i \tag{4}$$

其中的 l_i和 h_i都是可调的经验值, $WordsSim(keysAnswer, studentsAnswer)_i$的取值范围是 [$l_i$, h_i],通过设置 $highestScore_i$和 $lowestScore_i$既可以避免评分过高或过低.

2.3 语义对立度

仅仅评判学生的语句相似度还是不够的,通过比较学生答案中的语义对立度,语义对立度高的同学要扣除一定的分数,从而保证学生观点前后严密、论断一致.

对于第 i道题都有 p_i 个关键词,在评分前,人工标注出关键词汇 $keyWords$. 然后学生答案再和 $keyWords$ 通过与词表匹配,主要是对反义词选项 F进行比较,得出学生答案的语义对立度. 算法大致描述如下:

①把 $studentsAnswer$中的每个词 $studentWords$与 $keyWords$中的每个词 $word$进行比较,对于第 i道题总共要比较 p_i 次;

②对于第 i道题,可以计算第 $j(j \leq p)$ 对 $studentsWord$和 $word$之间的 $currentContrary_j$,即当前语义对立度,当 $studentsWord$ 和 $word$ 的 F 属性匹配时, $currentContrary(studentsWord, word)_j = -1$,此时 $studentsWord$和 $word$是反义词,为了提高精度,应当尽可能多的收录词的近义词和反义词;

③对于第 i道题,累计每个 $currentContrary_j$,对于 $q(j \leq q \leq p)$ 对匹配的学生答案的总体语义对立度 $totalContrary_i$ 可以用公式(5)表示:

$$totalContrary_i = \frac{\sum_{j=1}^{q} currentContrary_j}{p_i} \tag{5}$$

这样可以得到一个不大于 1的负值.

④对于第 i道题,学生的应扣分数可以用总体语义度用公式(6)计算:

$$subtractScore_i = totalContrary_i \cdot X_i \tag{6}$$

其中 X_i 是每道题设置的一个可调的参数,应该是一个较小的正值.

2.4 学生分数评定

对于有 m 道主观题的试卷,学生的总分 $totalScore$计算评定可由公式(7)求得:

$$totalScore = \sum_{i=1}^{m} (addScore_i - subtractScore_i) \tag{7}$$

学生总分是对于 m 道主观题的分数的求和,其中,第 i道主观题的分数是第 i道主观题的应得分数减去应扣分数.

3 结 论

本文首次将相似度计算和对立度计算结合,设计了一个多概念融合的主观题评分算法. 该主观题评分算法在整体上模拟了阅卷教师评定主观题时的思维,并在已有的自然语言处理的相关理论基础上,改进了部分算法,为主观题评分提供了一系列的计算公式,为下一步工作中的算法的模拟实现提供了理论支持. 当然,不同科目考试的主观题内容不一样、评估所使用的参数值的设定问题、算法在实践中的修正等方面还有待于进一步的深入研究.

参考文献:

[1] 梁娜,耿国华,周明全. 自然语言处理中的语义关系与句法模式互发现[J]. 计算机应用研究, 2008, 25(8): 2295 - 2299.
[2] 吴晨,张全,缪建明,等. 语义理解下的自然语言处理及信息检索模型[J]. 计算机科学, 2008, 35(5): 113 - 117.
[3] 南铉国. 基于语句相似度计算的主观题自动评分技术研究[D]. 延吉:延边大学, 2007.
[4] 贾电如,李明阳. 基于语句结构及语义相似度计算主观题评分算法的研究[J]. 信息化纵横, 2009, (5): 5 - 7.
[5] 麦范金,赵子强,岳晓光. 基于语义相似度的主观题阅卷系统模型设计[J]. 微计算机信息, 2009, 25(6 - 3): 255 - 256.
[6] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2008.
[7] 陈飞,王秀峰,饶一梅. 一种混合的中文分词算法[J]. 南开大学学报(自然科学版), 2007, 40(5): 27 - 31.
[8] 王常亮,腾至阳. 语句相似度计算在 FAQ中的应用[J]. 计算机时代, 2006, (2): 24 - 26.
[9] MA I Fan - jin, WANG Ting, SONG Rui. A Model of the Contrary Degree among Different Semantic meanings[C]. Chinese Computing Technologies and Related Linguistic Issues: Proceeings of the 7th International Conferen - ce on Chinese Computing. Beijing, China: PHEI, 2007: 204 - 209.
[10] 国家技术监督局. GB/T 13715 - 92信息处理用现代汉语分词规范[S]. 北京:中国标准出版社, 1993.
[11] 许嘉璐,傅永和. 中文信息处理现代汉语词汇研究[M]. 广州:广东教育出版社, 2006.