基于语句结构及语义相似度计算主观题评分算法的研究

贾电如，李阳明
（燕山大学 信息科学与工程学院，河北 秦皇岛 066004）

摘 要：文字类主观题的自动评分是实现远程教育中在线考试系统的一个关键技术，由于其自动评判具有相当难度，使自动评分系统中在对语句结构、关键字匹配、词性、词义以及语义方面的判断还存在很多问题。通过对已有的算法分析，提出了一种方法，采用浅层次句法结构分析和深层次语义分析相结合的算法计算相似度，该方法可以提高主观题自动评分的效率和准确度，具有一定的实用价值。

关键词：自动评分；动态规划；语句相似度；语义相似度
中图分类号：TP301.6 **文献标识码**：A

Research on assessment algorithm of subjective issue based on sentence structure and semantic similarity computation

JIA Dian Ru, LI Yang Ming
(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: Automated assessment technology of the subject issue based on free texts is one key techniques of online test in the distance teaching system. There is so much difficulty to realize the automated scoring system of subjective issue. There are many problems about sentence structure, keyword matching, gender, acceptation and semantic analysis and so on. This paper presents a new algorithm, which adopted low level sentence and deep semantic similarity computation. This method has the practical value to a certain degree because it can improve the efficiency and accuracy of automatic check about the subject issue.

Key words: automated assessment; dynamic programming; metric of sentence similarity; semantic similarity degree

目前，在线考试系统正在逐渐代替传统的考试系统，能否实现主观题自动评分是在线考试系统中一个重要环节。对于主观题的考查，由于它的答题涉及到人工智能、模式识别以及自然语言理解等方面的理论知识，评阅时就需要解决很多技术上的问题，因而成为阻碍在线考试系统发展的一个技术难点。

当前的主观题自动评分算法中，多数使用的是对学生答案和标准答案中关键字匹配来计算语句相似度，如基于向量空间模型 TF-IDF 方法、词性词序相结合的方法以及基于语义依存树等[1-4]。已有的这些方法要么从句子的表层结构信息进行匹配而忽略了语句语义分析，要么就是从语义分析而影响了整体语句的相似性，这些都会影响到自动评分计算的精确度。由于汉语语言的结构和语义的复杂性，一种意思可以用多种形式和多种关键字表达，单从一方面很难对语句的意思作出准确的判断，因此提出了一种新的主观题自动评分算法策略，主要思想是采用浅层次句法结构分析和深层次语义分析相结合的算法计算相似度，将这两种思想结合起来使用可以互补不足，提高了主观题自动评分的准确度。

1 语句相似度计算算法

在主观题自动批改系统中，语句相似度是用来评价学生答案和标准答案的接近程度。针对汉语的特殊性和机器翻译领域内一些对语句相似度的研究，采用动态规划法来计算语句相似度，主要思想是对语句进行层次句法分析。首先用正向最大匹配（MM）和基于词频统计的方法对句子分词，将分词后得到的语句视为词的向量，分别对各个关键词进行匹配。然后在此基

础上利用动态规划算法求出最优路径及语句相似度[5]。

1.1 相关定义

令 P 表示标准答案中的某一语句，Q 表示学生答案中的某一语句。P 和 Q 分别表示如下：$P=\{P_1,P_2,\cdots,P_m\}$,$Q=\{Q_1,Q_2,\cdots,Q_n\}$,其中 P_i 表示 P 语句中的一个关键词，Q_j 表示语句 Q 语句中的一个关键词，且 $P_i=P_{mi}\cup P_{gi}$, $Q_j=Q_{mj}\cup Q_{gj}$，其中 P_{mi} 表示语句 P 中第 i 个词的词义集合，P_{gi} 表示语句 P 中第 i 个词的词性集合；同理 Q_{mj} 表示语句 Q 中第 j 个词的词义集合，Q_{gj} 表示语句 Q 中第 j 个词的词性集合。为了便于进一步讨论给出以下几个定义：

定义 1：词义、词性相似度。词义、词性相似度可分别表示为：$SM_{ij}=SM(P_{mi},Q_{mj})$，$SG_{ij}=SM(P_{gi},Q_{gj})$。

定义 2：关键词相似度。关键词相似度 $W_{ij}=a\times SM_{ij}+\beta\times SG_{ij}$ 其中 a、β 分别为词义、词性相似度的权值。

定义 3：词向量的相似矩阵。用定义 2 计算出语句 P 和 Q 的所有关键词的相似度 $W_{ij}(i=1,2,\cdots,m;j=1,2,j=i=1,2,\cdots,n)$,形成一个 $m\times n$ 矩阵 M,称该矩阵为语句向量的相似矩阵。

$$M=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$$

定义 4：拓展词向量相似矩阵,对矩阵 M 进行如下拓展，形成矩阵 M',令 $M'_{0,0}=0$, $M'_{i,0}$, $M'_{0,j}=0(i=1,2,\cdots,m;\ j=1,2,\cdots,n)$,则 $M'_{i,j}=\max\{M'_{i-1,j-1}+W_{ij}, M'_{i,j-1}+\gamma, M'_{i-1,j}+\gamma\}$,其中，$\gamma$ 表示词位置不对应时的惩罚系数。

1.2 语句相似度求解算法

（1）利用动态规划法先求出 M' 矩阵[6]。

（2）M' 矩阵的初始化

创建一个 $(m+1,n+1)$ 矩阵，矩阵的行表示标准答案语句 P 的每个词，矩阵的列表示学生答案语句 Q 的每个词，利用定义 4 进行初始化，将 M' 矩阵的 $M'_{i,0}$, $M'_{0,j}$ 设置为 0. 其中 $i=0,1,2,\cdots,m$; $j=0,1,2,\cdots,n$。

（3）利用定义 1、2、3、4 依次求解 M' 矩阵中的每个元素 $M'_{i,j}$。

（4）求解最优相似矩阵

先从点 (m,n) 开始，到 $(1,1)$ 结束。在点（i,j）上选择 $M'_{i-1,j-1}+W_{ij}$,$M'_{i,j-1}+\gamma$,$M'_{i-1,j}+\gamma$ 最大者为最优点，所对应的 $M_{x,y}$ 作为路径的前一个节点 (x,y)。如果出现三者中两部分值相同且最大时，若该值在斜路径上则选择斜路径上 $(i-1,j-1)$ 作为路径的前一个节点；若不在斜路径上，优选水平方面 $(i-1,j)$ 作为路径的前一个节点；依次递推则选择一条最优路径。这样得到的路径上就是一条最优的路径，路径上最后一个点的值 $M'_{m,n}$ 表示了语句中所以词的相似度之和。

设 L 是标准答案语句的词数，则语句相似度为 $Sim=M'_{m,n}/L$。

2 语义相似度计算算法

Dekang Lin[7]认为任何两个事物的相似度取决于它们的共性(Commonality)和个性(Differentces)，然后从信息理论的角度给出任意两个事物相似度的通用公式：

$$Sim(A,B)=\frac{\log p(common(A,B))}{\log p(description(A,B))}$$

其中分子是描述 A、B 共性所需要的信息量的大小；分母是完整地描述出 A、B 所需要的信息量大小。刘群[8]认为两个词语的相似度是它们在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性大小。在本文中计算语义相似度是利用《知网》中词语相似度的定义[9]，可以把词语语义相似度的计算归结为“概念”相似度的计算;“概念”的相似度由描述它的“义原”的相似度得到。词语存在着一词多义的现象,知网中的一词多义表现为单个词语有多个概念,每个概念由一项定义来描述。比如:“打”在“打架”，“打太极”，“打猎”中的意义各不相同,知网中对应的概念描述分别是:

DEF = fight| 争斗

DEF = exercise| 锻炼,sport| 体育

DEF = catch| 捉住, # animall 兽

词语语义相似度的计算,严格来讲应该是计算概念之间的语义相似度。本文中采用刘群的思路,认为两个词语的语义相似度是其所有概念之间相似度的最大值。

$$Sim(c_1,c_2)=\max Sim(C_{1i},C_{2j})\ (i=1,2,\cdots,m;j=1,2,\cdots,n)$$

其中，C_{1i} 是词 C_1 的 m 项概念,C_{2j} 是 C_2 的 n 项概念。这样就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。本文利用语句相似度中分词方法将词语标注为概念,然后再对概念计算相似度。

2.1 义原相似度的计算

由于所有的概念都最终归结于用义原来表示，词语整体相似度由部分相似度合成的，所以义原的相似度计算是概念相似度计算的基础。所有的义原根据上下位关系构成了一个树状的义原层次体系,这里采用刘群的公式计算语义相似度的方法。

$$Sim(S_1,S_2)=\frac{a}{a+distence(S_1,S_2)}$$

其中，S_1、S_2 表示两个义原，$distance(S_1,S_2)$表示它们的路径长度，a 是一个可调节的参数。在知网的知识描述语言中,在一些义原出现的位置都可能出现一个具体词(概念),并用圆括号( )括起来。所以本文在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的

相似度计算。理想的做法应该是先把具体词还原成知网的语义表达式,然后再计算相似度。这样做将导入函数的递归调用,有可能导致死循环,反而使算法变得很复杂。由于具体词在知网的语义表达式中只占很小的比例,因此可以作如下处理:具体词与义原的相似度定义为一个比较小的常数 γ。具体词和具体词的相似度按两个词相同则为 1 否则为 0。

2.2 概念相似度的计算

由义原相似度可以计算概念相似度,词语整体相似要建立在部分相似的基础上。把一个复杂的整体分解成部分,通过计算部分之间的相似度得到整体的相似度。假设两个整体 A 和 B 都可以分解成以下部分:A 分解成 $A_1, A_2,\cdots, A_n$, B 分解成 $B_1, B_2,\cdots, Bm$,那么这些部分之间的对应关系就有 $m \times n$ 种。但并不是这些部分之间的相似度都对整体的相似度发生影响,所以应该选择那些发生影响的部分之间的相似度,选择出来后再进一步得到整体的相似度。在比较两个整体的相似性时,首先要做的工作是对这两个整体的各个部分之间建立起一一对应的关系,然后在这些对应的部分之间进行比较。如果某一部分的对应物为空则将任何义原(或具体词)与空值的相似度定义为一个比较小的常数 δ;其他整体的相似度通过部分的相似度加权平均得到[10]。对于实词概念的语义表达式,可以将其分成四个部分:

第一独立义原描述式:将两个概念的这一部分的相似度记为 $Sim1(S1,S2)$;

其他独立义原描述式:语义表达式中除第一独立义原以外的所有其他独立义原(或具体词),将两个概念的这一部分的相似度记为 $Sim2(S1,S2)$;

关系义原描述式:语义表达式中所有的关系义原描述式,将两个概念的这一部分的相似度记为 $Sim3(S1,S2)$;

符号义原描述式:语义表达式中所有的符号义原描述式,将两个概念的这一部分的相似度记为 $Sim4(S1,S2)$。

于是两个概念语义表达式的整体相似度记为:

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2)$$

其中,$\beta_i(1 \leqslant i \leqslant 4)$是可调节的参数,且有 $\beta_1+\beta_2+\beta_3+\beta_4=1$, $\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$。后者反映了 $Sim1$ 到 $Sim4$ 对于总体相似度所起到的作用依次递减。由于第一独立义原描述式反映了一个概念最主要的特征,所以应该将其权值定义得相对比较大,一般应在 0.5 以上。

3 实验测试与分析

以《操作系统》课程为实验素材,选取 2006 级计算机专业学生的 90 份考卷中 4 道简答题为例。每道试题的分数是 10 分,分别通过计算机自动评分和人工阅卷,所得到的评分结果进行分析,得到如下表所示:

表 1 实验结果

	误差=0	误差=1	误差=2	误差≥3
试题 1	85	2	2	1
试题 2	80	6	3	1
试题 3	83	2	4	1
试题 4	76	10	2	2
合计	324	20	11	5

其中误差为自动评分与人工评分所得分数之差。由于系统中分词词典中缺少某些专用词汇或由于语句繁琐较长,可能导致得分的偏差。但是对于主观题来说,在人工评阅时,也受到教师情绪等诸多因素的影响,因此认为只要误差小于 1 分的就认为得到了正确的评分。

$$正确率=\frac{324+20}{90\times4}\times100\%\approx96\%$$

本文综合运用了语句层次结构、句法、词性、语义等特征来计算相似度,不仅考虑词语的局部相似,还从语句的整体出发,考查了语句语义整体相似性,大大提高了相似度计算性能,降低了计算的时间复杂度,同时也提高了主观题自动评分的准确性,具有一定的实用价值。

参考文献

[1] 孟爱国,胜贤.一种网络考试系统中主观题自动评分的算法设计与实现[J]. 计算机与数字工程, 2005, 33(7):147-150
[2] 李彬, 刘挺, 秦兵.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.
[3] 李素建.基于语义计算的语句相关度研究[J]. 计算机工程与应用, 2002,38(7):75-76.
[4] 李明琴, 李涓子, 王作英, 等. 语义分析和结构化语言模型[J]. 软件学报, 2005,16(9): 1523-1533.
[5] 高思丹,袁春风. 语句相似度计算在主观题自动批改技术中的初步应用[J].计算机工程与应用, 2004,40 (14): 132 -135.
[6] LEVITIN A . The Design and Analysis of Algorithm [M]. Beijing: Electronics Industry Press, 2003.
[7] LIN De Kang. An Information2Theoretic Definition of Similarity Semantic distance in Word Net [A] .In: Proceedings of the Fifteenth International Conference on Machine Learning [C].1998.
[8] 刘群, 李素建.基于《知网》的词汇语义相似度计[C]//第三届汉语词汇语义学研讨会, 台北,2002, 5.
[9] 董振东,董强.“知网”网站[DB/OL].http://www.keenage.com.
[10]朱征宇,苑昆峰,陈杏环.一种基于最大权匹配计算的信息检索方法[J]. 计算机工程与应用, 2007,43(33):176-179.

(收稿日期:2008-11-28)