中文主观题自动批改中相似句子检索算法

张小艳

(西安科技大学 计算机系, 陕西 西安 710054)

[摘要] 学生答案与标准答案语义匹配程度的计算是基于中文文字类主观题自动批改中的关键问题. 提出了学生答案与标准答案匹配程度的计算分两步进行: 候选相似语句的检索和基于语义依存的句子相似度计算. 利用动态规划法实现候选语句检索, 确定数量不多但有可能与标准答案相似的候选句子, 然后对标准答案中的句子与少量的候选句子进行深层的句法分析, 找出依存关系, 并在依存分析结果的基础上进行语义相似度计算, 得出最终的结果. 该方法可以提高主观题自动批改的效率及准确性, 具有一定的实用价值.

[关键词] 自动批改, 动态规划法, 语义相似度, 依存树

[中图分类号] TP391.1 [文献标识码] B [文章编号] 1672-1292(2007)02-0062-05

Similar Sentence Search Algorithm in Automated Assessment System of Subjective Test Based on Chinese

Zhang Xiaoyan

(Department of Computer, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract The key problem in the automated assessment of the subjective test is the computation of the semantic marching degree between the student answer and the standard answer. The computation method of the matching degree between the student answer and the standard answer is proposed which includes two steps of searching for the similar candidate sentences and computing the similarity degrees between sentences based on semantic existence dependence. The dynamic programming is used to search for the candidate sentences, and then the candidate sentences are determined. The number of the determined sentences is not much, but they may have the similar meanings with the standard answer. Furthermore, the deep syntax analyses on the standard answer and the few candidate sentences are implemented to find the dependent relationship. The final result is acquired by the computation of the semantic similarity degree based on the result of the dependent analysis. This method has the practical value to a certain degree because it can improve the efficiency and the accuracy of automatic check about the subjective questions.

Key words automatic check system, dynamic programming, semantic similarity degree, dependent tree

0 引言

基于中文的主观题自动批改是实现远程教学系统中在线考试功能的一个关键技术. 由于其涉及到人工智能、模式识别以及自然语言理解等方面的理论和知识, 实现起来相当复杂. 这在一定程度上使得在线考试不能实用化, 从而成为制约网络教学发展的一个重要因素.

目前, 国内基于中文的自动批改系统尚未出现, 已有的研究都是针对简单类主观题如名词解释、简答题自动批改, 且没有实用的系统出台. 这些研究主要有: 李辉阳等[1]研究了有限领域中简述文字的自动判读问题, 提出利用基于关系的带权匹配技术实现 CAI中简单论述正误的判定; 王邯等[2]提出了网络教学中 C程序设计填空题机器批改的实现方法; 高思丹等[3]对语句相似度的计算进行了研究, 提出基于动态

收稿日期: 2006-12-27.
基金项目: 陕西省教育厅专项课题基金(06JK248)资助项目.
作者简介: 张小艳(1967-), 女, 副教授, 主要从事网络集成与数据库技术、知识工程与智能系统、计算机教育技术等方面的教学与研究. E-mail: zhangxy@xust.edu.cn

规划法的语句相似度计算方法, 实现对文字类主观题的自动批改.

这三种方法在一定程度上模拟了老师阅卷过程, 对主观题计算机自动批改有一定的借鉴意义. 但是, 这些自动批改技术普遍存在以下不足: (1) 对学生在答题时的主观性考虑不足. 汉语句子的表示方法有多样性, 在回答问题时对同一个语义可能采用不同的词语表达. (2) 现有的自动批改方法多是以关键词的匹配为基础, 只注重学生答案与标准答案的关系, 而忽略了考试试题本身的要求. (3) 相应的语句相似度的衡量只利用句子的表层信息, 即组成句子的词的词频、词性等信息, 不加任何结构分析, 即不进行句法与语义分析.

1 问题的分析

教师评阅文字类主观题, 一般是预先制定好一套评分标准, 然后将每道试题的总分划分成若干部分, 将分数分配到试题的求解过程中的一些关键的步骤或关键的语句上, 通常称之为得分点. 教师在人工评阅主观题时, 一般首先检查学生答案中有几个得分点, 得分点多则分数高; 然后再看学生答案和标准答案意义相近度, 意义相近度高则分数高; 最后再考虑学生答案语言是否通顺, 条理性是否强等因素, 适当对分数进行调整.

通过分析可以发现, 主观题自动批改系统中的关键问题在于学生答案与标准答案意义匹配程度的计算. 本文提出学生答案与标准答案匹配程度的计算分候选相似语句的检索和语义相似度匹配两步进行: 利用动态规划法实现候选语句检索, 确定数量不多但有可能与标准答案相似的候选语句; 对标准答案中的语句与少量的候选语句进行深层的句法分析, 找出依存关系, 并在依存分析结果的基础上进行语义相似度计算, 得出最终的结果.

2 候选句子粗检索算法

对标准答案中的某一关键语句 (得分点), 在学生答案中寻找与之相似的语句, 当相似度达到某个阈值时, 则可视为相似语句, 即候选语句. 选择候选句的依据是: 如果一个学生答案中的句子与标准答案中的句子相同或同义的词越多, 就越有可能是正确答案.

首先, 将标准答案、学生答案断句, 然后对各句进行分词, 将分词后的语句视为词的向量. 对于标准答案、学生答案的词向量中的词求其相似度, 形成相似矩阵, 然后利用动态规划算法求出最佳匹配路径及相似度.

2.1 相关定义

令 A 表示标准答案中的某一关键语句, B 表示学生答案中的语句. 通过分词和词性、词义标注, 将 A 和 B 分别表示如下: $A = \{A_1, A_2, \cdots, A_m\}$, $B = \{B_1, B_2, \cdots, B_n\}$, 其中 A_i 表示 A 语句中的一个孤立词, B_j 表示语句 B 中的一个孤立词, 且 $A_i = A_{mi} \cup A_{ti}$, $B_j = B_{mj} \cup B_{tj}$. 其中, A_{mi} 表示语句 A 中第 i个词的词义集合, A_{ti} 表示语句 A 中第 i个词的词性集合; B_{mj} 表示语句 B 中第 j个词的词义集合, B_{tj} 表示语句 B 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** 集合相似度.

令 $|M|$ 表示集合 M 的元素个数, $|N|$ 表示集合 N 的元素个数, 定义集合 M 和 N 的相似度为:

$$SM(M, N) = \frac{2 \times |M \cap N|}{|M| + |N|}. \tag{1}$$

显然, 如果 $M = N$, 则 $SM(M, N) = 1$; 如果 M, N 交集为空, 则 $SM(M, N) = 0$

定义 **2** 词义、词性集合的相似度.

词义、词性集合的相似度可分别用以下方式表示:

$$SM_{ij} = SM(A_{mi}, B_{mj}), \tag{2}$$

$$ST_{ij} = SM(A_{ti}, B_{tj}). \tag{3}$$

定义 **3** 关键词相似度.

定义词 A_i 与词 B_j 的相似度是词义相似度和词性相似度加权和:

$$W_{ij} = a \times SM_{ij} + b \times ST_{ij}. \tag{4}$$

其中, a, b分别为词义相似度、词性相似度的权值.

定义 **4**　词向量的相似矩阵.

使用式(4)计算出语句 A 与语句 B 的所有对应关键词的相似度 W_{ij}($i=1,2,\cdots,m$; $j=1,2,\cdots,n$), 形成一个 $m\times n$ 矩阵 ***SMAT***, 称该矩阵为语句词向量相似矩阵.

$$\boldsymbol{SMAT}=\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ W_{m1} & W_{m2} & \cdots & W_{mn} \end{bmatrix}.$$

定义 **5**　语句长度相似度

$$\mathrm{Sim\ len}=\frac{|A|}{||A|-|B||+|A|}. \tag{5}$$

其中, $||A|-|B||$表示 $|A|-|B|$的绝对值.

定义 **6**　拓展词向量相似矩阵.

对矩阵 ***SMAT***进行如下拓展, 形成矩阵 ***M***:

令 $M_{00}=0, M_{i0}, M_{0j}=0$($i=1,2,\cdots,m$; $j=1,2,\cdots,n$), 则

$$M_{ij}=\max\{M_{i-1,j-1}+W_{ij}, M_{i,j-1}+\gamma, M_{i-1,j}+\gamma\}. \tag{6}$$

其中, γ表示词位置不对应时的惩罚系数.

$M_{m,n}$即为语句中所有词的相似度之和. 可利用上述初值和递推公式, 递推得到输入语句同答案的匹配最优路径, 最后形成相似度.

2.2　语句相似度求解算法

利用动态规划法[4]求解***M*** 矩阵.

2.2.1　***M*** 矩阵的初始化

创建一个$(m+1, n+1)$矩阵, 矩阵的行对应标准答案中语句A的每个词, 矩阵的列对应学生答案中语句 B的每个词. 利用定义(6)进行初始化, 将***M*** 矩阵的M_{i0}、M_{0j}填充为 0, 其中 $i=0,1,2,\cdots,m$; $j=0,1,2,\cdots,n$.

2.2.2　***M*** 矩阵元的求解

利用式(1)、(2)、(3)、(4)及递推公式(6)从(1, 1)依次求解***M*** 矩阵中的元M_{ij}($i=1,2,\cdots,m$; $j=1,2,\cdots,n$), 最终产生 ***M*** 矩阵.

2.2.3　求解最大相似度路径

从点(m, n)开始, 到(1, 1)结束. 在点(i, j)上, 优选$M_{i-1,j-1}+W_{ij}$、$M_{i,j-1}+\gamma$、$M_{i-1,j}+\gamma$三者中最大者所对应的$M_{x,y}$作为路径的前一个节点(x, y). 三者中有两部分相同且最大时, 如果斜路径$M_{i-1,j-1}$在候选, 则优选斜路径$(i-1, j-1)$作为路径的前一个节点. 三者中有两部分相同且最大时, 如果斜路径不在候选, 则优选水平方向路径$(i-1, j)$作为路径的前一个节点. 如果 $i=1$, 则路径的前一个节点是$(i, j-1)$. 从点(1, 1)到(m, n)路径上的点表示了语句中词的对应关系, 路径最后一个点的值$M_{m,n}$表示了语句中所有词的相似度之和.

设 m 是标准答案语句的词数, 则语句相似度为: $\mathrm{Sim}=\dfrac{M_{m,n}}{m}$.

假如还要考虑语句长度差异对语句相似度的影响, 可通过下面的公式计算句子相似度:

$$W\mathrm{Sim}=\alpha\times\mathrm{Sim}+\beta\times\mathrm{Sim\ len}$$

其中, α, β分别表示 Sim的相似度和 Sim len的相似度的权值.

3　语义相似度计算

经过以上分析, 可从学生答案中筛选出少量候选语句. 本文用基于语义依存树的句法分析对标准答案与候选语句进行句法与语义分析.

句子的意义是由句子中词之间的句法语义关系体现出来的, 而依存语法描述的正是具有直接句法语义联系的词之间的关系, 因此本文在描述句子中词之间的语义关系时, 采用了相对成熟的依存语法理论. 依存语法认为, 词之间的关系是有方向的, 通常是一个词支配另一个词, 这种支配与被支配的关系就称作依存关系.

在语义依存语法中, 支配词又称为被支配词的中心词. 中心词通常可以表现它所在短语的主要语法、语义特征, 例如动词、名词短语中的中心词是动词、名词, 方位词短语的中心词是地点名词[5].

句子成份间相互支配与被支配、依存与被依存的现象普遍存在于汉语的词汇(合成词)、短语、单句、复合直到句群的各级能够独立运用的语言单位之中, 这一特点为依存关系的普遍性. 依存句法分析可以反映出句子中各成分之间的语义修饰关系, 它可以获得长距离的搭配, 并与句子成分的物理位置无关[6].

利用依存结构计算句子间的相似度, 关键的一步是如何获得句子各成分间的依存关系信息. 本文利用哈尔滨工业大学信息检索研究室的"中文依存句法分析"获得句子各成分间的依存关系.

例如: 标准答案中的句子 *A*: 数据元素被存入连续的存储单元中. 学生答案中的句子 *B*: 所有数据被放入连续的存储空间中. 句子各成分之间的依存关系可以表示为图 1和图 2.

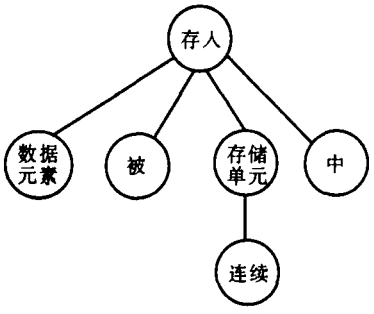

图 1 句子 *A* 的依存树

Fig.1 Dependent tree for sentence *A*

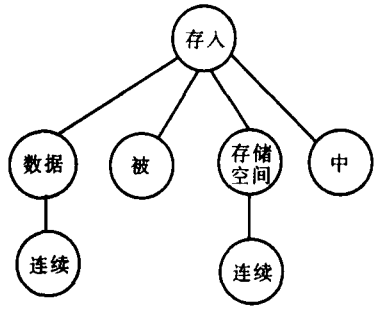

图 2 句子 *B* 的依存树

Fig.2 Dependent tree for sentence *B*

依存树是一个复杂的非线性关系, 如果对整个依存树进行完全匹配的话, 所花费的代价是巨大的. 另外, 一个完整的汉语句子是由句子的关键成分和修饰成分所构成, 而人们往往从关键成分就可以了解一个句子的大概意思. 但由于汉语表达形式的多样性, 相同的关键成分可用不同的修饰成分来修饰, 如果强调修饰成分, 这无疑会给句子间相似度的计算增加噪音. 基于以上两点, 在利用依存结构进行相似度计算时, 只考虑那些有效搭配对之间的相似程度. 所谓有效搭配对是指全句核心词和直接依存于其有效词组成的搭配对, 针对简答题的特点, 这里有效词定义为动词、名词以及形容词, 它是由分词后的词性标注决定的. 这样一来算法的复杂度就大大降低, 而准确率也会得到一定程度的提高.

句子 *A* 的有效搭配对为: 数据元素—存入、存入—存储单元、连续—存储单元. 句子 *B* 的有效搭配对为: 数据—放入、放入—存储空间、连续—存储空间. 只要比较它们之间的语义相似程度即可. 语义相似度计算公式如下[7]:

$$Sim(A,\ B) = \frac{\sum_{i=1}^{n} W_i}{\max\{Paircount1,\ Paircount2\}} \tag{7}$$

Paircount1、Paircount2分别返回的是句子 *A*、*B* 中有效语义搭配对的总数, $\sum_{i=1}^{n} W_i$ 是指句子 *A* 和 *B* 两种有效语义搭配对匹配的总权重.

将搭配对匹配的权重定义如下:

设有两个搭配对: Word*A*1 ~ Word*A*2 和 Word*B*1 ~ Word*B*2, 若 Word*A*1 与 Word*B*1、Word*A*2 与 Word*B*2 的语义相同, 则权重为 1; 若仅有一组语义相同, 则权重为 0.5; 否则为 0.

如上面的两个句子中: 存入与放入语义相同; 存储空间、存储单元语义相同; 数据元素与数据的语义不同, 则: $SM(A,\ B) = \frac{1+1+0.5}{3} = 0.83$.

4 实验及分析

以《数据结构与算法》课程为实验素材,选取了计算机系本科 05级学生的 120份试卷中的 4道简答题,共 480道试题,每道试题 5分,进行了自动批改. 与人工批改对比结果如表 1所示.

表 1 实验结果

Table 1 Results of experiment

	误差 = 0分	误差 = 1分	误差 = 2分	误差 > 2分
试题 1	108	5	6	1
试题 2	110	7	3	0
试题 3	106	8	5	1
试题 4	100	12	6	2
合计	424	32	20	4

其中误差为自动批改与人工批改所得分数之差值. 可看出误差为 1分的有 32例,误差为 2分的有 20例,误差大于 2分的有 4例. 经过仔细分析,导致误差的原因有如下几种:

(1) 系统的分词词典中缺少某些词汇的相关信息;

(2) 句子较长,而且句中有较多动词,导致句子中心词找不准确,最终导致得分偏差;

(3) 学生答案中的句子繁琐,逻辑关系不清晰;

(4) 教师主观给分或扣分.

总之,本文所提出的方法,去除人为因素,自动批改的结果与人工批改结果趋向一致. 在下一步工作中,将针对长句子(多动词)的分析进行深入研究.

5 结束语

语句相似度的计算是主观题自动批改技术的核心内容,只有成功获得学生答案与标准答案各得分点的相似度,才能给出相应成绩. 本文提出的分两步计算语句相似度的方法综合考虑了语句的词法、句法、语义三个层次特征的相似度,不仅考虑了词的局部相似度,还从语句整体出发,考察了语句在整体上的相似情况,提高了相似度的计算性能. 同时,通过第一个层面的计算,确定了数量不多但有可能与标准答案相似的候选句子,缩小了进行深层的句法分析、语义相似度的计算量;在计算依存树之间的相似度时,仅计算那些有效搭配对之间的语义相似程度,使计算的时间复杂度大大降低,提高了主观题自动批改的效率及准确性,具有一定的实用价值.

[参考文献] (References)

[1] 李辉阳,韩忠愿. 有限领域简述文字的自动判读及其在 CAI中的应用[J]. 计算机工程与应用, 2002, 38(8): 76-78.
LiHuiyang, Han Zhongyuan. Auto-judging to the simple description in the specific field and its application in CAI[J]. Computer Engineering and Applications, 2002, 38(8): 76-78. (in Chinese)

[2] 王邯,肖俊,冯刚. 网络教学中 C程序设计填空题机器批改的实现[J]. 计算机与数字工程, 2003, 31(1): 37-39.
Wang Han, Xiao Jun, Feng Gang. Implementation of a testing system on web-based course of C programming language[J]. Computer and Digital Engineering, 2003, 31(1): 37-39. (in Chinese)

[3] 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.
Gao Sidan, Yuan Chunfeng. The application of sentence similarity measurement in automated assessment technology of subjective tests[J]. Computer Engineering and Applications, 2004, 40(14): 132-135. (in Chinese)

[4] Anany Levitin. The Design and Analysis of Algorithm [M]. Beijing: Electronics Industry Press, 2003.

[5] 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.
LiMingqin, Li Juanzi, Wang Zuoying, et al. Semantic analysis and structured language models[J]. Journal of Software, 2005, 16(9): 1523-1533. (in Chinese)

[6] 李明琴,李涓子,王作英,等. 中文语义依存关系分析的统计模型[J]. 计算机学报, 2004, 27(12): 1679-1687.
LiMingqin, Li Juanzi, Wang Zuoying, et al. A statistical model for parsing semantic dependency relations in a chinese sentence[J]. Chinese Journal of Computers, 2004, 27(12): 1679-1687. (in Chinese)

[7] 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.
LiBin, Liu Ting, Qin Bing, et al. Chinese sentence similarity computing based on semantic dependency relationship analysis[J]. Application Research of Computers, 2003, 20(12): 15-17. (in Chinese)

[责任编辑: 严海琳]