一种自发性口语评测文本语义相似度评分特征提取方法

宋 阳 王 岚

（ 中国科学院深圳先进技术研究院 深圳 518055 ）

摘 要 自发性口语评测中如何提取文本语义相似度评分特征是一个非常困难的问题。针对这个问题本文采用基于词网络（WordNet）的Lesk算法计算词与词之间的语义相似度，在词义相似度的基础上提出了词与文本之间的语义相似度算法，提出了一种完整的基于词网络的文本语义相似度评分特征提取方法。实验利用该方法在考生答案与标准答案之间中提取文本语义相似度评分特征，并利用该特征与老师评分进行相关度分析，实验结果表明该算法可以有效的表征考生答案和标准答案之间的文本语义相似度。

关键词 自发性口语评测；文本语义相似度；词网络；Lesk算法

Method of Text-to-text Semantic Similarity Feature Extraction for Spontaneous Speech Evaluation

SONG Yang WANG Lan

(*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract Due to the difficult of text-to-text semantic similarity feature extraction in spontaneous speech evaluation, this paper presents WordNet based Lesk algorithm to calculate the semantic similarity between words, defines the semantic similarity algorithm between word and text based on the semantic similarity between words, and proposes a complete set of wordnet based text-to-text semantic similarity feature extraction methods. Experiment extracts text-to-text semantic similarity feature between student's answers and the standard answers with this algorithm and analyzes the correlation between the feature and the teacher rating. Experimental results show that the algorithm can effectively characterize the text-to-text semantic similarity between the students' answers and the standard answer.

Keywords spontaneous speech evaluation; text-to-text semantic similarity; wordnet; lesk algorithm

1 引 言

21世纪，随着信息化进程的不断加速人工智能技术正以日新月异速度发展着，同时也迅速的渗透到人们的日常生活当中。语音作为人们最容易接受的交互方式，即将成为下一个主流的人机交互方式。语音评测是教育领域中语言教学的重要组成部分，它能评价并反馈发音人的发音质量，帮助其进行口语学习。传统的口语考试，由于考生众多，教师很难做到一一对学生进行评分和发音纠正。近几年来，随着语音识别的准确度不断提高，计算机辅助语言学习技术（Computer Assisted Language Learning，CALL）受到了越来越多的关注，利用计算机进行口语自动评测的技术也取得了很大的进步。但是目前的口语自动评测技术[1-5]主要停留在朗读题评分上，主要针对考生的发音准确度、流利度、语速等声学层面上特征对考生的口语水平进行评价。如何针对自发性口语表述

基金项目：国家自然科学基金资助项目（NSFC 90920002）。

作者简介：宋阳，研究方向为英语发音自动评测技术；王岚，博士，研究员，国家自然科学基金重点项目负责人，研究方向以构建先进智能信息系统为核心，包括基于语音识别的英语发音纠错，三维说话人头像发音和表情的动态模拟，复杂声学环境下的语音识别。

进行评测一直是口语自动评测的一个难点，主要原因就是没有一个有效的方法来评价考生答案和标准答案之间的文本语义相似度。

语义相似度研究的是用什么样的方法来计算和比较两个词语的相似性，自然语言的词语之间有着非常复杂的关系。在实际应用中，有时需要把这种复杂的关系用一种简单的方式来度量，而语义相似度就是其中一种。计算词与词之间的语义相似度的算法有很多种，比如Michael E.Lesk[6]在1986年提出的基于WordNet的Lesk算法，它最初是应用在词义消歧领域当中，后经Satanjeev Banerjee和Ted Pedersen[7]等人的改进，这种算法得出的语义相似度和人工判断结果的相关度又有了很大的提高。但是在进行口语评测时需要计算的是两个文本之间的语义相似度。传统的基于向量空间模型（Vector Space Model，VSM）[8]的文本语义相似度的算法只适用于比较大的文档之间文本语义相似度的计算，因为它存在由自然语言的歧义性与多样性给特征向量带来的噪声问题。Michael Pucher[9]在2007年曾经提出过一种将两个文本之间所有单词的语义相似度得分取平均值作为两个文本间语义相似度的方法，但由于方法存在考生只说关键词就能得高分情况，因此这种方法不能直接应用在自发性口语评测当中。

本文将基于WordNet语义相似度算法应用到了自发性表述口语评测当中，提出了一套完整的文本语义相似度评分特征提取流程。内容安排如下：第二节介绍了机器自动评分的流程；第三节介绍了基于WordNet的Lesk词义相似度算法；第四节介绍了传统的文本语义相似度算法及存在的问题；第五节提出了改进后的文本语义相似度算法与特征提取流程；第六节对本文工作进行了总结与展望。

2 机器自动评分流程

图1就是目前主流的口语自动评分流程，它可以概括为以下三个步骤：

(1)对音频进行语音识别得到识别结果；

(2)在识别结果中通过评分特征提取算法提取评分特征；

(3)将评分特征送到评分模型当中通过机器学习算法得到机器评分。

这就是口语自动评分的流程，本文将沿着这个流程着重探讨自发性口语评测文本语义相似度评分特征提取问题。

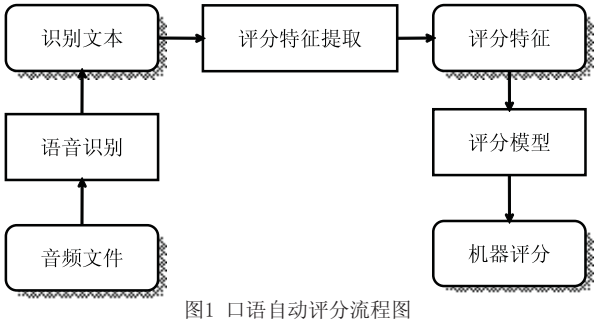

图1 口语自动评分流程图

3 基于WordNet的Lesk词义相似度算法

基于WordNet的词义相似度算法有很多，本文采用的是基于WordNet的Lesk算法，下面将对WordNet和Lesk算法进行一下简单的介绍。

3.1 WordNet简介[10]

WordNet最初在1985年由普林斯顿大学认知科学实验室实现，它是一部在线词典数据库系统。传统词典一般都是按字母顺序组织词条信息的，这样的词典在解决用词和选义问题上是有价值的。然而，它们有一个共同的缺陷，就是忽略了词典中同义信息的组织问题。20世纪以来，语言学家和心理学家们开始从一个崭新的角度来探索现代语言学知识结构以及特定的词典结构，终于由Princeton大学研制成功了一个联机英语词汇检索系统—WordNet，它作为语言学本体库，同时又是一部语义词典，在自然语言处理研究方面应用非常广泛。WordNet与其他标准词典最显著的不同在于：它将词汇分成五个大类：名词、动词、形容词、副词和虚词。实际上，WordNet仅包含名词、动词、形容词和副词。虚词通常是作为语言句法成分的一部分，WordNet忽略了英语中较小的虚词集。

3.2 Lesk算法简介[6]

Lesk算法最初是用来在句子的语境下消歧句子中的单词。主要方法是利用WordNet计算两个单词的词义中共享的词数，重叠的单词越多语义就越相关。一般情况下英文单词大多会具有多重词性，每重词性下又包含多重语义。为了进行语义消歧，Lesk算法对单词的每一个语义与在短语中出现的其它单词的语义来做比较，共享词数最多的那一个语义即为该单词在该语境下的语义。

Satanjeev Banerjee和Ted Pedersen[7]在2003年提出了一种改进的Lesk算法，它的基本思想是在原来Lesk计算词义重叠词的基础上加入各种重叠打分，访

问词典中的词义层次结构来加入多种重叠打分以引入更多的相关信息。这种改进不仅考虑了WordNet中的各语义（gloss）中的单词共享，也考虑了其他如同义词（synset）、上位词（hypernym）与下位词（hyponym）等集合的单词共享，并且制定了新的打分规则：

规则1：如果两个语义重复共享一个连词，那么他们之间的语义相似度得分score就加上N^2分。

规则2：计算两个单词的语义相似度得分时，取两个单词之间各组语义中得分最高的一组作为其各自的语义，相应的得分作为两个单词之间的语义相似度得分。

我们用$S(\omega)$表示单词ω的语义集合，c_i表示ω的第i个语义，并且$c_i \in S(\omega)$，则规则2可用以下公式表示：

$$\mathrm{score}(\omega,\omega') = \max_{c_i \in S(\omega) c_j \in S(\omega')} \mathrm{score}(c_i, c_j) \qquad (1)$$

制定这样规则的原因就是两个单词的语义中出现N连词共享的概率很低，N越大相应的语义也就越接近，因此应当给一个更高的得分。我们以英语单词中的car和motorcycle为例，它们各自有这样一层语义：

gloss(car): *a motor vehicle with four wheels*

gloss(motorcycle): *a motor vehicle with two wheels*

我们看到两个单词的语义中一共共享了四个单词，一个三连词a motor vehicle和一个一元词wheels。如果按着传统的Lesk算法两个单词的语义相似度得分为4分，但是按着改进的Lesk算法两个单词的语义相似度得分为一个三连词积9分再加上一个一元词积1分，总共的语义相似度得分将为10分。

假设我们只考虑WordNet中语义、同义词信息，结合上这种新的打分机制，在计算任意两个单词ω和ω'的语义相似度score(ω,ω')时，这种新的Lesk语义相似度算法可以表示为：

$$\begin{aligned}\mathrm{score}(\omega,\omega') = {} & \mathrm{score}(\mathrm{gloss}(\omega),\mathrm{gloss}(\omega')) \\ & + \mathrm{score}(\mathrm{gloss}(\omega),\mathrm{synset}(\omega')) \\ & + \mathrm{score}(\mathrm{synset}(\omega),\mathrm{synset}(\omega'))\end{aligned} \qquad (2)$$

公式（1）中不但考虑到了两个单词中语义之间的单词共享，还考虑到了语义与同义词之间、同义词之间的单词共享。

4　传统的文本语义相似度算法及存在的问题

得到了词与词之间的语义相似度得分，接下来我们就需要计算文本之间的语义相似度得分了。在介绍新的算法之前，我们首先分析一下传统文本语义相似度算法应用在自发性口语评测时存在的问题。

4.1　基于向量空间模型文本语义相似度算法及存在的问题

传统的文本语义相似度都是通过文档向量空间模型来计算的，向量空间模型[8]是20世纪60年代末期由提出的，最早是应用在SMART信息检索系统中，目前已成为自然语言处理的常用模型。

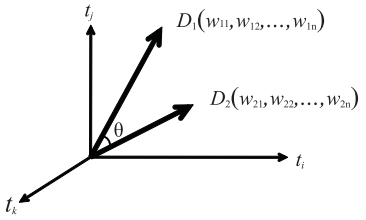

图2　文档的向量空间模型示意图

如图2所示，文档向量空间模型由三部分组成：

（1）文档(document)：通常是文章中具有一定规模的片段，如句子、句群、段落。

（2）项/特征项（term/feature term）:特征项是VSM中不可分割的语言单元，可以是字、词、词组或短语等，这样文档可表示为：Decoment=D(t_1,t_2,…,t_n)，其中t_k是特征项，并且1≤k≤n。

（3）项的权重(term weight)：对于含有个特征项的文档D(t_1,t_2,…,t_n)，每一个特征项t_k都依据一定的原则被赋予一个特征权重λ_k，表示它们在文档中的重要程度。

设文档D_1和D_1表示VSM中的两个向量：

$$D_1=D_1(\lambda_{11},\lambda_{21},\cdots,\lambda_{1n})$$
$$D_2=D_2(\lambda_{21},\lambda_{22},\cdots,\lambda_{2n})$$

那么，可以借助于n维空间中两个向量之间的某种距离来表示文档间的相似度，常用的方法是使用两个向量夹角的余弦值来表示相似度，如公式（3）所示：

$$\mathrm{Sim}(D_1,D_2) = \cos\theta = \frac{\sum_{k=1}^{n} \lambda_{1k} \times \lambda_{2k}}{\sqrt{(\sum_{k=1}^{n} \lambda_{1k}^2)(\sum_{k=1}^{n} \lambda_{2k}^2)}} \qquad (3)$$

但是这种方法存在一个缺陷，它适用于比较大的文档之间的语义相似度计算。如果测试文档比较小数据非常稀疏，用这种方式计算语义相似度将不再适用，因为它没有考虑到同义词之间或者同一单词不同形态之间的语义关系。通常自发性口语评测中无论是考生答案还是标准答案所涵盖到的词汇范围都很少，并且经常出现同义词的使用（如mother与mom）、时

态的变换（如is与was）、单复数的变换（如student与students）等情况，特征向量变得更加稀疏，让机器学习算法很难从中提取到用于计算特征权重的统计特性。因此，需要一种算法来解决自发性口语评测时单词的歧义性和多样性给特征向量带来的噪声问题。

4.2 基于WordNet文本语义相似度算法存在的问题

Michael Pucher[9]在2007年针对自动语音识别的任务曾经提出一种文本语义相似度计算方案，他首先通过基于WordNet的Lesk算法计算两个文本所有词之间的语义相似度得分，这样就很好的解决了小文本之间语义相似度计算时数据稀疏的问题，因为基于WordNet语义相似度算法可以很好的表征出同义词、近义词、同一单词不同形态等多种语义联系。

有了词与词之间的语义相似度的分，下一步就需要制定一个算法计算词与文本之间的语义相似度得分。我们定义任意一个单词ω和一段文本T之间的语义相似度得分score_W，那么score_W可以取为ω_i和文本T中所有单词词义相似度得分的平均值。因为考生答案中如果出现了ω，那么我们可以认为ω和文本T中所有的单词都存在一定的语义联系，如公式（4）所示：

$$\text{score}_W(\omega,T)=\frac{1}{N}\sum_{i=1}^{N}score(\omega,\omega_i) \qquad (4)$$

其中$\omega_i \in \text{T}$，N表示文本T中的单词个数。这样就得到了单词与文本之间的语义相似度得分。下一步需要计算文本之间的语义相似度得分$score_T$。我们可以想到最直观的办法是：首先分别计算出考生答案文本A中所有的单词ω_j与标准答案文本T之间的语义相似度得分score_W，然后再将这些得分累加到一起就可以作为两个文本之间的文本语义相似度得分，用这种求和方式得到的文本语义相似度得分我们用$score_{\text{sum }T}$表示，如公式（5）所示：

$$score_{\text{sum }T}(A,T)=\sum_{j=1}^{M}score_W(\omega_j,\text{T}) \qquad (5)$$

其中$\omega_j \in A$，M表示文本A中的单词个数。但是这样做有一个很明显的缺点：只要答案A中包含的单词越多，相应的A与T之间的文本语义相似度得分也就越高，也就是说考生只要尽可能多说一些内容得分也就越高，这样评出来的分数明显不合理。为了解决这个问题，Michael Pucher又提出了一种新的解决方案：在公式（5）基础上求平均值，用这种求平均值方式得到的文本语义相似度得分我们用$score_{\text{aver }T}$表示，如公式（6）所示：

$$score_{aver\ T}(A,T)=\frac{1}{M}\sum_{j=1}^{M}score_W(\omega_j,\text{T}) \qquad (6)$$

这样就很好的解决了考生只要说得越多得分也就越高的问题。但是这种算法应用在自发性口语评测时又会出现另外一个问题，考生在进行作答时是事先知道话题内容的，按照这个评分规则，考生只需围绕话题最相关的一个或几个单词说一下他就绝对能够得高分，因为这种方法无法考察考生作答内容的丰富程度。

5 改进后的文本语义相似度算法与特征提取流程

前面我们已经讨论了传统的文本语义相似度特征提取算法应用到自发性口语评测时所出现的问题，针对这些问题本文提出了一种改进型的特征提取算法。

5.1 改进后的文本语义相似度算法

公式（6）应用到自发性口语评测时无法解决考生只说关键词就能的高分的情况，也就是说这种算法只考虑到了考生作答内容是否与话题相关，无法考察考生作答内容的丰富程度。针对这一问题本文在公式（6）的基础上做了进一步的改进，我们用$N_{\text{uniq}}(A)$表示考生答案中去重后的单词数，用它乘以公式（6）所有单词的语义相似度平均值，用这种平均值乘以去重单词数方式得到的文本语义相似度得分我们用$score_{\text{uniq }T}$表示，则有：

$$score_{uniq\ T}(A,T)=\frac{1}{M}\sum_{j=1}^{M}score_W(\omega_j,\text{T})\times N_{\text{uniq}}(A) \qquad (7)$$

这样做的目的就是将考生作答内容丰富程度引入到评分特征当中，不会再出现考生只围绕关键词作答就能的高分的情况。

这种算法应主要是针对自发性口语评测文本语义相似度特征提取而设计的，它不但考察了考生作答内容是否与话题相关还能够考察考生作答内容是否丰富，从话题相关性和内容丰富性两个角度考察了考生答案的文本语义相似度，在第5节中本文将通过实验来证明这种算法的合理性和有效性。

5.2 语义相似度评分特征提取流程

图3是一个完整的文本语义相似度评分特征提取流程，主要包括以下步骤：

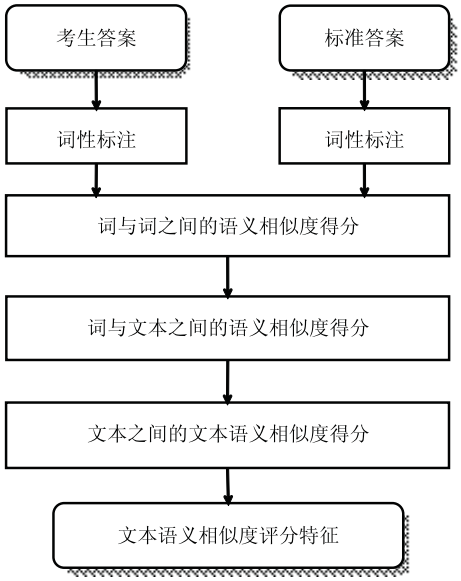

图3　文本语义相似度评分特征提取流程

（1）词性标注：在进行语义相似度计算之前，必须对文本中的单词进行词性标注，本文采用的是基于决策树的词性标注的方法，详见参考文献[11]。

（2）词与词之间的语义相似度得分计算：标注完词性之后，利用基于WordNet的Lesk算法计算考生答案和标准答案的所有词之间的语义相似度得分。

（3）词与文本之间的语义相似度得分计算：利用公式（4）计算词与文本之间的语义相似度得分。

（4）文本之间的文本语义相似度得分计算：利用公式（7）计算文本之间的文本语义相似度得分得到最终的评分特征。

6　相关度实验

这一节我将通过实验来证明这种算法的有效性，试验将通过利用该方法在考生答案与标准答案之间中提取文本语义相似度评分特征，并利用该特征与老师评分进行相关度分析。

6.1　相关系数

相关系数r_{XY}又称线性相关系数，它是衡量两个变量X和Y之间线性相关程度的指标。如公式（8）所示：

$$r_{XY}=\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2\sum_{i=1}^{N}(Y_i-\bar{Y})^2}} \tag{8}$$

6.2　实验数据介绍

实验数据选用深圳西乡中学527名中学生一次复述题考试作答音频作为实验数据，共计1054分钟的数据量（每位考生2分钟录音时间）。聘请了专业英语教师对音频进行了文本转写和评分，评分分数最低0分最高24分，最小评分粒度为1分。实验选用复述题的话题文本作为标准答案。

6.3　试验结果分析

实验首先将在考生的作答音频的转写文本上进行，分别利用公式（5）（6）（7）计算考生答案和标准答案之间的文本语义相似度得分$score_{\mathrm{sum}\ T}$，$score_{aver\ T}$，$score_{uniq\ T}$，然后再利用三种文本语义相似度得分同老师评分数据进行相关度分析。

表1 相关度实验结果

相似度算法	$score_{\mathrm{sum}\ T}$	$score_{aver\ T}$	$score_{uniq\ T}$
相关系数	0.493088	0.335502	0.775135

由表1我们可以看出采用求和方式得到的相似度得分$score_{\mathrm{sum}\ T}$与老师评分的相关度只有0.493088，而采用了平均值方式得到的相似度得分$score_{aver\ T}$与老师评分的相关度只有0.335502，下降了近0.13，说明本次考试中考生作答内容相对比较丰富，但是作答的话题相关性普遍不高。而采用了新的算法得到的相似度得分$score_{uniq\ T}$与老师评分的相关度达到了0.775135明显优于前两种算法，达到了预期的实验效果，说明这种基于WordNet的文本语义相似度评分特征提取方法应用在自发性口语评测中是有效的。

作为对比，我还在考生作答音频的识别结果上利用公式（7）做了相同的实验。由于本此实验数据是中学生英语口语口试，考生的英语水平普遍都不高，属于非母语说话人自发性表述语音识别。目前我们所取得的识别正确率只有44.14%，在这种识别率下利用新算法所提取到的文本语义相似度评分特征与老师评分的相关度只有0.554424，说明这种算法对语音识别的准确率也是有一定的要求的。

7　结　论

本文将基于WordNet的语义相似度算法应用到自发性口语测评文本语义相似度评分特征提取当中，提出了完整的文本语义相似度评分特征提取流程，取得了良好的实验效果。但该方法仅从考生做答内容的语义方面分析了文本语义相似度，若在此基础上同时对考生作答内容的语法结构进行相似度分析，评分准确度应当会进一步提高。另外，该方法是基于自发性口语表述语音识别结果的，因此需要进一步提高语音识别的准确率以利于语义相似度的分析，这也是我们的

下一步的工作方向。

参 考 文 献

[1] Xie S, Evanini K, Zechner K. Exploring content features for automated speech scoring [J]. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, 2012.

[2] Franco H, Neumeyer L, et al. Automatic pronunciation scoring for language instruction [J]. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, 2: 1471-1474.

[3] Neumeyer L, Franco H, et al. Automatic scoring of pronunciation quality [J]. Speech Communication, 2000, 30(2-3): 83-93.

[4] Neumeyer L, Franco H, Weintraub M, et al. Automatic text-independent pronunciation scoring of foreign language student speech [C] // Fourth International Conference on IEEE, 1996, 3: 1457-1460.

[5] Franco H, Neumeyer L, et al. Combination of machine scores for automatic grading of pronunciation quality [J]. Speech Communication, 2000,30(2-3): 121-130.

[6] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone [C] // Proceedings of the 5th annual international conference on Systems documentation, 1986: 24-26.

[7] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness [C] // International Joint Conference on Artificial Intelligence, 2003, 18: 805-810.

[8] 宗成庆. 统计自然语言处理 [M]. 北京: 清华大学出版社.2008:340-353.

[9] Pucher M. WordNet-based semantic relatedness measures in automatic speech recognition for meetings [C] // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007: 129-132.

[10] Fellbaum C. WordNet [J]. Theory and Applications of Ontology: Computer Applications, 2010: 231-243.

[11] H Schmid. Probabilistic part-of-speech tagging using decision trees [C] // In Proceedings of International Conference on New Methods in Language Processing, 1994, 12: 44-49.