第 43 卷 第 9 期 计 算 机 工 程 2017 年 9 月
Vol. 43 No. 9 Computer Engineering September 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)09-0240-05 文献标志码：A 中图分类号：TP391

基于句法结构与修饰词的句子相似度计算

邓 涵[1a]，朱新华[1a,2]，李 奇[1a]，彭 琦[1b]

（1. 广西师范大学 a. 计算机科学与信息工程学院；b. 网络中心，广西 桂林 541004；
2. 广西区域多源信息集成与智能处理协同创新中心，广西 桂林 541004）

摘 要： 根据汉语句子结构复杂、词语一词多义的特点，提出一种句子相似度计算方法。对句子进行句法分析和依存关系的预处理，提取句子结构中的主、谓、宾、介词等主要成分的词语集合，从而准确地表达出句子的浅层语义，并利用《知网》计算不同句子相同成分之间的语义相似度。考虑依存句法关系中的定中关系和状中关系起到的语义修饰作用，在句法结构基础上进一步融入修饰词，综合计算句子的语义相似度，区分句子主题内容的一致性和句子间的反义关系。以微软研究院释义语料库中抽取的 30 对句子作为测试集，实验结果表明，提出方法的皮尔森相关系数达到 0.89，F 值达到 85.7%，具有较好的准确性与实用性。

关键词： 句子相似度；知网；依存树；句法结构；修饰词

中文引用格式： 邓 涵，朱新华，李 奇，等. 基于句法结构与修饰词的句子相似度计算[J]. 计算机工程，2017，43(9)：240-244，249.

英文引用格式： DENG Han，ZHU Xinhua，LI Qi，et al. Sentence Similarity Calculation Based on Syntactic Structure and Modifier[J]. Computer Engineering，2017，43(9)：240-244，249.

Sentence Similarity Calculation Based on Syntactic Structure and Modifier

DENG Han[1a]，ZHU Xinhua[1a,2]，LI Qi[1a]，PENG Qi[1b]

（1a. College of Computer Science and Information Engingeering；1b. Network Center，Guangxi Normal University，Guilin，Guangxi 541004，China；2. Collaborative Innovation Center of Guangxi Regional Multi-source Information Integration and Intelligent Processing，Guilin，Guangxi 541004，China）

【Abstract】 According to the complex structure and polysemy characteristics of Chinese sentences，this paper proposes a sentence similarity calculation method. It pretreats the sentence through syntactic analysis and dependency relationship，and extracts word set of main components such as subject，predicate，object，preposition and so on，thus the shallow semantics of sentences can be expressed accurately. HowNet is used to calculate the semantic similarity between the same components of different sentences. Considering semantic modification effect of attribute relationship and adverbial relationship in dependency syntactic relations，based on syntactic structure，further integrating into the modifiers，the sentence semantic similarity is comprehensive by calculated to distinguish consistency of sentence topic content and the antonym relationship between sentences. The extracted 30 pairs of sentences are used as test sets，which are taken from paraphrase corpus of Microsoft Research Institute Corpus. Experimental results indicate that the Pearson correlation coefficient of the proposed method reaches 0.89 and the F-measure reaches 85.7%，which has better accuracy and practicability.

【Key words】 sentence similarity；HowNet；dependency tree；syntactic structure；modifier

DOI： 10.3969/j.issn.1000-3428.2017.09.042

0 概述

人工智能的飞速发展使得人们不再满足于一些简单的智能操作，而是进一步要求计算机能够像人类一样进行思考与对话。人类语言的文本常常以句子的形式存储在计算机上，因此首先要求计算机能够理解单个的句子。在英文领域，已经有了简单匹配、莱文斯坦编辑距离、连续子序列匹配、最长公共子序列（(Longest Common Length，LCS）[1]等多种较为成熟的方法，但由于汉语和英语在表达方式、主被动、时态等语法、语义表达等各方面存在比较明显的差异，这些对英文句子的相似度计算方法应用到

基金项目： 国家自然科学基金(61363036，61462010)；广西师范大学自然科学青年基金“词汇语义相似度计算研究”。

作者简介： 邓 涵(1991—)，女，硕士研究生，主研方向为自然语言处理；朱新华(通信作者)，教授；李 奇、彭 琦，硕士研究生。

收稿日期： 2016-08-16 **修回日期：** 2016-10-18 **E-mail：** zxh429@263.net

汉语计算时不能考虑汉语的自身特点。因此,国内学者针对汉语的特点做了大量实验研究并提出了多种适用于汉语语法语义的句子相似度计算方法,比如由文献[2]提出的基于骨架依存的汉语句子相似度计算,首次提出了骨架依存的概念;文献[3]提出的基于语义计算的语句相关度研究,将词语的相似度和相关度一起考虑作为句子的相似度。这些句子相似度计算方法在相对应的不同语境环境和领域内取得了较好的实验效果,但仍然不能达到最终的实际应用需求,如许多方法[1-3]句子的浅层语义分析不够精确,没有考虑句子程度词的影响等特点。本文综合考虑句子语义、结构等多方面的特点,从句子结构和语义修饰词两方面对句子语义相似度进行研究,更为精确地表达句子的浅层语义,正确区分句子内容的一致性。

1　相关知识

1.1　基于《知网》的词语相似度计算

本文通过对文献[4-5]算法思想的分析和研究,把整体相似度还原为部分相似度的加权平均策略[6],得出两个义项间的语义相似度,具体方法为:

计算义原或具体词与空值、义原与具体词之间的相似度,分别将其设置为一个较小的常数 δ 和 γ。具体词指的是《知网》中尚未给出定义的词条,一般用括号括起来。本文把常数 δ 和 γ 均设置为 0.2。将上述 3 个部分相似度根据加权平均策略组合起来就可以得到义项的相似度,即根据词语的意义相似性得到语义相似度。义项间的语义相似度计算公式如下:

$$sim(C_1,C_2)=\sum_{i=1}^{3}\beta_i\prod_{j=1}^{i}sim(C_1,C_2) \tag{1}$$

其中,$\beta_i(1\leqslant i\leqslant 3)$是调节参数,且有 $\beta_1+\beta_2+\beta_3=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3$。本文的实验中 β_i取 0.7,β_2取 0.17,β_3取 0.13。公式中采用了多个义原相似度连续相乘的形式,其主要目的是为了充分发挥主要部分相似度的权重,同时又降低了后面次要部分的相似度在计算相似度时所起的作用,有效地避免了计算的主次颠倒引起的整体相似度过高的不合理现象。

由于大多数词语都存在多个义项,因此本文将 2 个词语的所有义项相似度最大的值作为 2 个词语的最终相似度值,计算公式如下:

$$sim(W_1,W_2)=\max_{1\leqslant i\leqslant m,1\leqslant j<n}sim(C_{1i},C_{2j}) \tag{2}$$

其中,W_1和 W_2分别是 2 个不同或相同的词语。

1.2　句子相似度的研究现状

目前,针对汉语的一词多义、结构复杂等语法特点,国内学者提出了很多句子相似度的计算方法,如文献[7]通过计算词语语义之间的相似度得到句子的相似度,计算公式如下:

$$sim(A,B)=\frac{1}{2}\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{j=1}^{n}b_j}{n}\right) \tag{3}$$

其中,A 句中有 m 个词语;B 句中有 n 个词语;$A_i(1\leqslant i\leqslant m)$为句子 A 中的第 i 个词语;$B_j(1\leqslant j\leqslant n)$为句子 B 中的第 j 个词语;$sim(A_i,B_j)$表示词语 A_i和 B_j的相似度;a_i是 A_i与 B 句中所有词语相似度的最大值;b_j是 B_j与 A 句中所有词语相似度的最大值。

句法结构方法首先通过对 2 个句子进行分词来构建句法结构树,抽取结构树中的主要信息来计算 2 个句子之间的相似度,其中比较有代表性的是基于依存关系的计算方法。文献[2]在分析句法结构的基础上,提出了一种基于骨架依存树的句子相似度计算模型。该骨架依存树首先提取句子的核心谓词作为树的根节点,而与核心谓词相依存的句子则作为叶子节点。基于该思想文献[8]提出了一种基于语义依存的方法来计算句子的相似度,该方法通过依存分析得到句子中各成分之间的依存关系,并将词语的语义与依存关系构成句子语义依存树。其句子相似度的计算公式如下:

$$sim(Sen1,Sen2)=\frac{\sum_{i=1}^{n}W_i}{\max\{PairCount1,PairCount2\}} \tag{4}$$

其中,$W_i(1\leqslant i\leqslant n)$为 2 个句子的有效搭配对匹配数的权重;$PairCount_i$为句子 i 的有效搭配对数;i 为2 个所比较的句子序号。以上 2 种方法都是句子中关键词提取后再重新组合的过程,这些方法在计算以语义为核心的句子相似度时效果欠佳,无法很好地区分句子主题内容的一致性与句子间的反义关系。

2　句子语义相似度计算

2.1　计算公式

文献[7]通过计算词语之间的相似度得到句子的相似度,只考虑了单纯的词语的语义,但汉语有着语法结构很复杂、一词多义等特点。因此,通过单一的考虑句子中词语的语义很难准确地提炼句子信息而得到较准确的相似度。于是在考虑词语语义的基础上加入句子的句法结构,提出初始的基于句法结构的句子语义相似度计算公式如下:

$$CSim(x_i,y_j)=\frac{1}{2}\left(\frac{\sum_{a=1}^{num(x_i)}s(x_{ia})}{num(x_i)}+\frac{\sum_{b=1}^{num(y_j)}s(y_{jb})}{num(y_j)}\right) \tag{5}$$

其中,$s(x_{ia})$表示一个完整“主谓宾介”结构 x_i中第 a 个词分别与 y_j结构中相同成分词语相似度的最大值,且 x_{ia}中的 a 满足:$1\leqslant a\leqslant num(x_i)$,$num(x_i)$表示 x_i结构中词语的个数;$s(y_{jb})$表示一个完整“主谓宾介”结构 y_j中第 b 个词分别与 x_i中相同成分词语相似度的最大值,且 y_{jb}中的 b 满足:$1\leqslant b\leqslant num(y_j)$,$num(y_j)$表示 y_j结构中词语的个数。

由于对句法结构的分析只考虑了主、谓、宾、介词等主要成分,而通过对句子的依存关系分析,得出的在句法结构中定中和状中关系中的修饰词也对句

子的语义相似度有重要影响。因此在句法结构分析的基础上再融入修饰词的语义影响因素,得到下面的句子相似度计算公式:

$$CDegSim(x_i,y_j)=CSim(x_i,y_j)\times DegSim(c_{x_i},c_{y_j}) \tag{6}$$

其中,$CSim(x_i,y_j)$为式(5)中基于主谓宾介结构的句子相似度计算;$DegSim(c_{xi},c_{yj})$为程度副词对句子相似度的影响。由于只有当 2 个句子表达的主题内容相似时融入情感词才能更好地进一步精确句子的相似度,因此给 $CDegSim(x_i,y_j)$设定一个阈值,只有当大于这个阈值时才考虑融合正面情感词语与负面情感词语,本文设定为 0.8。将式(6)进行改进得到以下公式:

$$SDegSim(x_i,y_j)=\begin{cases}1-CDegSim(x_i,y_j), neg(c_1,c_2)=0\\ CDegSim(x_i,y_j), neg(c_1,c_2)=1\end{cases} \tag{7}$$

其中,$neg(c_1,c_2)=1$ 表示句子中都存在程度副词且同为正面或负面;$neg(c_1,c_2)=0$ 表示句子中都存在程度副词且不同为正面或负面。若 $neg(c_1,c_2)=0$,本文设定为用 1 减去这 2 个句子的相似度,得到在句法结构的基础上融入修饰词的句子相似度。由于每个完整的句子不可能都只是一个简单句,有的句子里面可能包含几个小的完整主谓宾介结构,最后得到 2 个句子的语义相似度计算公式:

$$\begin{aligned}SGDegSim(A,B)&=\left(\frac{\sum_{x=1}^{n}A_i}{n}+\frac{\sum_{y=1}^{m}B_j}{m}\right)/2\\ A_i&=\max_{1\leqslant b\leqslant n}(SDegSim(x_i,y_b))\\ B_j&=\max_{1\leqslant a\leqslant m}(SDegSim(x_a,y_j))\end{aligned} \tag{8}$$

其中,假设 A 中有 n 个完整的主谓宾介结构的小句子,B 中有 m 个完整的主谓宾介结构的小句子。且 $A_i(1\leqslant i\leqslant n)$和 $B_j(1\leqslant j\leqslant m)$分别表示一个句子中包含所有完整“主谓宾介”结构单元和另一个句子中包含的“主谓宾介”结构单元之间相似度的最大值。

2.2 句法结构分析

汉语句子本身所表达的含义是通过各个相对独立的词语的语义来体现的,词语是最小的能够独立活动且代表一定含义的语言成分,故词语相似度是本文的基础。在计算词语相似度时,通过对知网的研究,利用式(2)对词语之间的语义相似度进行计算。考虑到句子的结构,对句子进行预处理,其中包括句子的分词、指代消解、词语集合的筛选、去停用词以及构建句法分析和对依存关系的判断。本文采用的是 ICTCLAS 分词系统[9]进行分词。句法分析使用的是斯坦福大学自然语言实验室开发的 Stanford 中文句法分析器得到一个基于句法标注的句法分析树[10],依存句法分析器是由哈尔滨工业大学自然语言处理实验室开发的,可以得到句子主要成分之间的依存关系[11],得到句子的主语、谓语、宾语以及介词短语。

例如以“在备忘录中,巴尔默对微软重申了开源的危险。”为例说明构建过程。首先,将句子“在备忘录中,巴尔默对微软重申了开源的危险。”进行分词,得到分词后的句子“在/p 备忘录/n 中/f ,/wd 巴尔默/nrf 对/p 微/ag 软/ng 重申/v 了/ule 开/v 源/ng 的/ude1 危险/an 。”再通过 Stanford 中文句法分析器得到句法分析树,如图 1 所示。利用依存句法分析器得到句子各成分之间的依存关系,如图 2 所示。

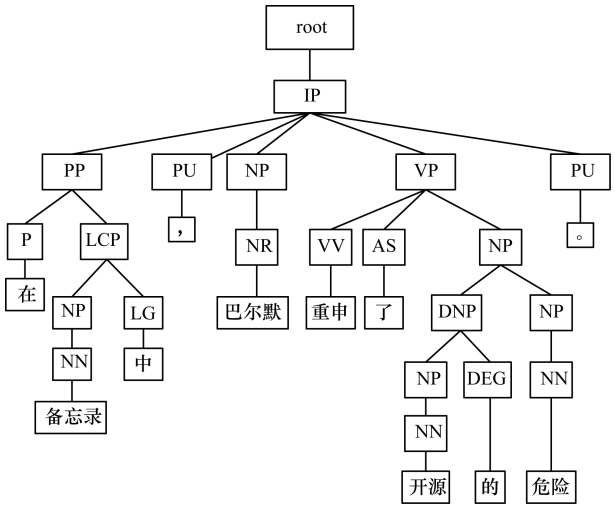

图 1 句子 A 的句法分析树

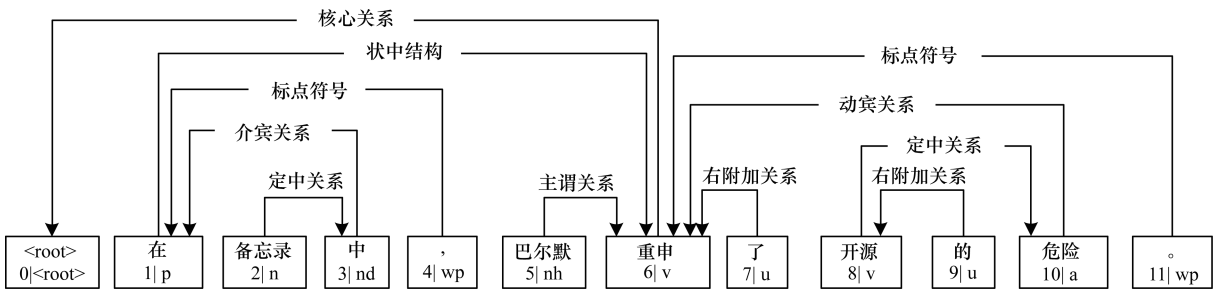

图 2 句子 A 的句法依存树

图 1 根据句法结构分析,将以 root 为标注的 IP 的下一层为划分依据,将句子分为 PP,NP 和 VP,这里 PP 表示介词短语,NP 表示名词短语,可作主语或宾语,VP 表示动词短语。其中,VP 里面的 VV 是句子中的核心谓词。

在图 2 的依存树中,使用如下的依存关系规则

决定相关核心成分:1)与依存树根结点构成核心依存关系的 VP 成员为句子的谓语;2)与所确定的谓词构成主谓依存关系的名词是句子的主语;3)与所确定的谓语有着动宾关系的名词是句子的宾语;4)与所确定的谓语存在状中关系的 PP 短语为所要提取的介词性短语。

通过句法结构和依存关系,可以共同确定句子中的核心成分,最终得到例句的"主语 + 谓语 + 宾语 + 介词短语"结构单元为(主语"巴尔默",谓语"重申",宾语"危险",介词短语"在备忘录中")。但需要声明的是简单的句子可能也只含有主谓宾的结构。本文在划分结构单元时都采用主谓宾介结构,将不存在的结构成分设置为空。

本文假设待计算的 2 个句子为:句子 A(包含 n 个分句)和句子 B(包含 m 个分句),每个分句被划分为以"主语 + 谓语 + 宾语 + 介词短语"的结构集合,记为 $x_i(1\leqslant i\leqslant n)$ 和 $y_j(1\leqslant j\leqslant m)$。同时,将第 x_i 个短句的"主谓宾介"集合中所包含的第 a 个词语记为 x_{ia},将第 y_j 个短句的"主谓宾介"集合中所包含的第 b 个词语记为 y_{jb}。

为了让主语和主语、谓语和谓语、宾语和宾语以及介词短语这样相同成分的词语进行词语语义相似度计算,为每一个词语 x_{ia} 和 y_{jb} 分别分配一个成分标志:v_{ia} 和 v_{jb}。以"主 + 谓 + 宾 + 介词"为结构划分成分信息标志,将主语成分标志设为 1,谓语成分标志设为 2,宾语成分标志设为 3,介词短语成分标志设为 4。2 个词语之间的成分标志相似度计算公式为:

$$vsim(x_{ia},y_{jb})=\begin{cases}1, & v_{ia}=v_{jb}\\ 0, & v_{ia}\neq v_{jb}\end{cases} \tag{9}$$

公式表明若相比较的 2 个词语分别在句子中的成分相同,则它们的成分相似度计算为 1,否则记为为 0。最终,可以得到 2 个完整的"主谓宾介"单元结构相似度的最大值为:

$$\begin{aligned}s(x_{ia})&=\max_{1\leqslant q\leqslant num(y_j)}(sim(x_{ia},y_{jq})\times vsim(x_{ia},y_{jq}))\\ s(y_{jb})&=\max_{1\leqslant p\leqslant num(x_i)}(sim(x_{ip},y_{jb})\times vsim(x_{ip},y_{jb}))\end{aligned} \tag{10}$$

其中,$s(x_{ia})$ 表示句子 A 中第 x_i 个"主谓宾介"结构中的第 a 个词分别与 y_j 结构中的每个词语相似度的最大值,且 y_{jq} 中的 q 满足:$1\leqslant q\leqslant num(y_j)$;$s(y_{jb})$ 表示一个完整的"主谓宾介"结构 y_j 中的第 b 个词分别与 x_i 结构中的每个词语相似度的最大值,且 x_{ip} 中的 p 满足:$1\leqslant p\leqslant\leqslant num(x_i)$, $num(y_j)$。然后得到完整的"主谓宾介"的结构之间的相似度计算式。因为每个词语需要计算 2 遍,所以结果除以 2。

由于每个句子有可能由几个完整结构单元的短句组成,A_i 和 B_j 分别表示一个完整句子中的"主谓宾介"结构单元和另一个句子中的"主谓宾介"结构单元之间相似度的最大值。其中,$y_q(1\leqslant q\leqslant m)$;$x_p(1\leqslant p\leqslant n)$。

$$\begin{aligned}A_i&=\max_{1\leqslant q\leqslant m}(CSim(x_i,y_q))\\ B_j&=\max_{1\leqslant p\leqslant n}(CSim(x_p,y_j))\end{aligned} \tag{11}$$

将式(11)中的 A_i 和 B_j 代入式(5),得到仅基于词语相似度的基础上考虑句法结构的句子语义相似度计算公式:

$$CGSim(X,Y)=\frac{1}{2}\left(\frac{\sum_{i=1}^{n}A_i}{n}+\frac{\sum_{j=1}^{m}B_j}{m}\right) \tag{12}$$

本节在句法结构分析的基础上提出基于"主语 + 谓语 + 宾语 + 介词短语"单元结构的相似度计算方法。这种方法是将一个复杂的句子划分为一个或多个"主语 + 谓语 + 宾语 + 介词短语"结构单元。同时,本文还考虑了这些结构中主要成分之间的不可置换性。引入成分信息,使主语和主语、谓语和谓语、宾语和宾语以及介词短语之间进行相似度计算,消除了主语、谓语、宾语和介词短语之间混杂计算存在的弊端。

2.3　修饰词

在上一节计算句子语义相似度时,只考虑了"主谓宾介",而没有考虑依存关系中的定中和状中关系起到的修饰作用。特别是在句子间的反义关系的比较上,比如否定副词"不"可以起到语义反转的作用,"只有一点"可以起到削弱程度的作用。于是在这一基础上,本文继续融入了修饰词来考虑句子之间的相似度。

从人文心理学的相关知识[12]和《知网》对词语的分类为基础,将修饰词[13]分为 3 种。其基本的划分标准为:

1)正面情感词语:对所描述的内容持有肯定和赞同态度,带有明显的称赞、颂扬、拥护、积极向上等意味,比如喜欢、喜爱、爱戴、其乐融融等词语。

2)负面情感词语:对所描述的内容持有否定、消极、排斥态度,带有明显的自卑、失意、绝望、质疑、讽刺、鄙视、批评等意味,比如讨厌、烦躁、深恶痛绝等词语。

3)程度级别词语:对所描述的内容起程度修饰作用的副词。根据修饰方向和程度大小的不同,在《知网》中将程度副词分为 6 类,其中"极其"类词有 69 个词,"很"类词有 42 个,"较"类词有 37 个,"稍"类词有 29 个,"欠"类词有 12 个,"超"类词有 30 个,共计 219 个程度类别词。

本文再根据语义程度将这 6 类程度副词分成语气加强型程度副词和语气减弱型程度副词 2 类。其中,语气加强型程度副词为极其、超、很、较,且加强程度为极其 > 超 > 很 > 较。语气减弱型程度副词为稍、欠,且减弱程度为稍 < 欠。其中,语气加强型程度副词赋值区间为 1 ~ 2,分别将它们从 1 开始以 0.1 为单位逐渐递增,带入式(8)中对语料库中的句子进行相似度计算,并计算其与人工判定值的 Pearson 相关系数,直到 Pearson 相关系数下降为止;最终确定

“极其”类词赋值 1.4,“超”类词赋值 1.3,“很”类词赋值 1.2,“较”类词赋值 1.1。语气减弱型程度副词赋值区间为 0 ~1,分别将它们从 1 开始以 0.1 为单位逐渐递减,带入式(8)中对语料库中的句子进行相似度计算,并计算其与人工判定值的 Pearson 相关系数,直到 Pearson 相关系数下降为止;最终确定“稍”类词赋值 0.8,“欠”类词赋值 0.4。

程度副词对句子相似度的影响计算公式为:

$$Deg(c_1,c_2)=1-\frac{\text{abs}(Ad(c_1)-Ad(c_2))}{\max(Ad(c_1),Ad(c_1))} \quad (13)$$

其中,$Deg(c_1,c_2)$为程度副词 c_1 和 c_2 对句子相似度的影响程度;$Ad(c_1)$ 和 $Ad(c_2)$ 为程度副词 c_1 和 c_2 对应的权值;abs($Ad(c_1)-Ad(c_2)$)为 $Ad(c_1)$ 和 $Ad(c_2)$ 差值的绝对值;max($Ad(c_1)$,$Ad(c_2)$)为 $Ad(c_1)$ 和 $Ad(c_2)$ 中的较大值。

在考虑程度副词对句子的影响时,要将程度副词融入到每一个“主语 + 谓语 + 宾语 + 介词短语”结构单元的相似度计算中去。因此将式(13)融入式(5),得到在句法结构的基础上融入修饰词的计算公式。

然后考虑正面、负面情感词语。本文设定 c_1 和 c_2 都是正面情感词语或者负面情感词语,则 $neg(c_1,c_2)=1$;否则,设定为 0。但有的句子表达的主题内容不同,且相似度本就不大的 2 个句子就没有必要再融入修饰词去计算句子的相似度,所以需要强调的是给 $CDegSim(x_i,y_j)$ 设定一个阈值,当大于这个阈值时才考虑融合正面情感词语与负面情感词语,当小于或等于这个阈值时将不考虑正面或负面情感类词语对句子相似度的影响。本文将这一阈值设为 0.8。

在句法结构上融入修饰词,不仅考察了句子结构中主要成分之间的相似度,还将词语间的同义、近义、反义等关系进行了考虑,特别是正面、负面情感词语以及程度副词,这样使相似度计算结果与人工判定值更接近。

3 实验与结果分析

本文的训练句子集含有 720 个句子,所有句子均来自于微软研究院释义语料库[14],在这 720 个句子中找出有代表性的 30 对句子,按照最接近最通顺的原则翻译成对应的中文语句对,并让实验室 20 个成员对数据集进行语义相似度判断,最后取他们的平均值作为本文的测试集的人工判定值。主要采用皮尔森相关系数以及和 F 值(F-measure)来衡量本文提出方法的优劣。其计算公式为:

Pearson 相关系数:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (14)$$

其中,r 值表示两组值的相关程度,越大则越相关;反之,越小则越不相关。

计算 F 值的公式:

$$p=\frac{\text{正确识别的文本总数}}{\text{识别出的文本总数}}\times 100\%$$

$$R=\frac{\text{正确识别的文本总数}}{\text{测试集中存在的文本总数}}\times 100\%,F=\frac{2PR}{P+R}\times 100\% \quad (15)$$

本文提出的基于句法结构和修饰词的句子相似度计算方法得到:基于词语语义的 Pearson 相关系数为 0.51,式(8)、式(12)的 Pearson 相关系数为0.89,0.83,人工判定值为 1.0。

实验数据表明,本文提出的基于句子结构与修饰词的句子相似度计算方法与人工判定值之间具有较高的皮尔森相关系数,说明该方法计算的句子相似度比较准确。特别是对句子主题内容一致,含有反义关系的一些句子,计算效果较好。

本文方法与同类方法在正确率、召回率和 F 值方面的对比结果如表 1 所示。

表 1 句子相似度计算方法的性能对比 %

方法	正确率	召回率	F 值
基于词语语义[7]方法	62.9	43.8	51.7
融合多特征[15]方法	65.1	66.1	65.2
基于单核树[16]方法	86.8	72.6	79.1
基于本体[17]方法	91.0	69.0	80.0
本文方法	92.3	80.0	85.7

通过表 2 得知,本文提出的基于句法结构和修饰词的句子相似度计算方法具有较高的正确率、召回率和 F 值,证明了本文计算方法具有较好的准确性和实用性,也符合人们对汉语语言的认知。

4 结束语

本文所提出的基于句法结构和修饰词的句子语义相似度计算方法,在词语相似度研究的基础上通过对句法结构和依存关系的分析,进而提取每个小短句中的“主谓宾介”结构,然后对 2 个句子中相同成分的词语进行词语的相似度计算。当 2 个句子结构比较相似时,为了进一步计算相似度的准确度,融入了修饰词进行研究。将得到的实验结果和人工判定值进行对比,并计算皮尔森相关系数和 F 值分别为 0.89 和 85.7% ,达到了比较好的实验结果。但是,本文在计算句子相似度时,还有其他的句法成分没有考虑,遗漏了很多句子信息。下一步将综合考虑句子的句法结构,尽可能多地包含句法成分,使得句子相似度的计算更为准确。

参考文献

[1] LEUSEH G, UEFFING N, NEY H, et al. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation[J]. Journal of Magnetic Resonance,2003,8(6):28-36.

[2] 穗志方,俞士汶. 基于骨架依存树的语句相似度计算模型[C]//中文信息处理国际会议论文集. 北京:清华大学出版社,1998:458-465.

(下转第 249 页)

度控制相结合,既提高了模型的数据分析能力,又防止了可能的过拟合,使得算法在提高稀疏度的同时预测精度也有所提高。同时,训练后模型中相关向量对应的小波尺度因子为自适应选择的最优值,各尺度因子可选取不同的值,使得模型能够更好地描述样本数据的结构特点。

参考文献

[1] TIPPING M E. Sparse Bayesian Learning and the Relevance Vector Machine[J]. Journal of Machine Learning Research,2001,1(3):211-244.

[2] 赵春晖,齐　滨,张　燚. 基于改进型相关向量机的高光谱图像分类[J]. 光学学报,2012,32(8):256-261.

[3] WIDODO A,KIM E Y,SON J D,et al. Fault Diagnosis of Low Speed Bearing Based on Relevance Vector Machine and Support Vector Machine[J]. Expert Systems with Applications,2009,36(3):7252-7261.

[4] CAESARENDRA W,WIDODO A,HONG T P,et al. Machine Degradation Prognostic Based on RVM and ARMA/GARCH Model for Bearing Fault Simulated Data[C]//Proceedings of Prognostics and Health Management Conference. Washington D. C.,USA:IEEE Press,2010:1-6.

[5] ZHANG Lei. A Multivariate Relevance Vector Machine Based Algorithm for On-line Fault Prognostic Application with Multiple Fault Features[C]//Proceedings of International Conference on Intelligent Computation Technology & Automation. Washington D. C.,USA:IEEE Press,2012:26-32.

[6] 胡昌华,王兆强,周志杰,等. 一种 RVM 模糊模型辨识方法及在故障预报中的应用[J]. 自动化学报,2011,37(4):503-512.

[7] 丁二锐,曾　平,丁　阳,等. 一种新的回归型约简多分辨率相关向量机[J]. 控制与决策,2008,23(1):65-69.

[8] 汪洪桥,孙富春,蔡艳宁,等. 多核学习方法[J]. 自动化学报,2010,36(8):1037-1050.

[9] FAUL A,AVENUSE J J T. Fast Marginal Likelihood Maximisation for Sparse Bayesian Models[C]// Proceedings of the 9th International Workshop on Artificial Intelligence & Statistics. Berlin,Germany:Springer,2003:3-6.

[10] TOLAMBIYA A,KALRA P K. Relevance Vector Machine with Adaptive Wavelet Kernels for Efficient Image Coding[J]. Neurocomputing,2010,73(7-9):1417-1424.

[11] SCHMOLCK A,EVERSON R. Smooth Relevance Vector Machine:A Smoothness Prior Extension of the RVM[J]. Machine Learning,2007,68(2):107-135.

[12] TZIKAS D G,LIKAS A C,GALATSANOS N P. Sparse Bayesian Modeling with Adaptive Kernel Learning[J]. IEEE Transactions on Neural Networks,2009,20(6):926-937.

[13] TZIKAS D,LIKAS A,GALATSANOS N. Incremental Relevance Vector Machine with Kernel Learning[C]// Proceedings of Conference on Artificial Intelligence. Berlin,Germany:Springer,2008:301-312.

[14] CHEN Yuehui,YANG Bo,DONG Jiwen. Time-series Prediction Using a Local Linear Wavelet Neural Network[J]. Neurocomputing,2006,69(4):449-465.

[15] ZHANG Li,ZHOU Weida,JIAO Licheng. Wavelet Support Vector Machine[J]. IEEE Transactions on Systems Man & Cybernetics Part B,2004,34(1):34-39.

[16] WANG Xinying,HAN Min. Online Sequential Extreme Learning Machine with Kernels for Nonstationary Time Series Prediction[J]. Neurocomputing,2014,145:90-97.

[17] TAKENS F. Detecting Strange Attractors in Turbulence[M]// RAND D,YOUNG Lai-sang. Dynamical Systems and Turbulence. Berlin,Germany:Springer,1981:366-381.

编辑　顾逸斐

(上接第 244 页)

[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] 刘　群,李素建. 基于《知网》的词汇语义相似度计算[D]. 北京:中国科学院计算技术研究所,2002.

[5] 江　敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报,2008,22(5):84-89.

[6] 朱征宇,孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用,2013,33(8):2276-2279,2288.

[7] 李家南. IT 领域问答系统的研究与实现[D]. 广州:华南理工大学,2016.

[8] 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.

[9] 张华平. NLPIR 简介[EB/OL]. (2014-12-12). http://ictclas.nlpir.org/docs.

[10] 王利局. 基于语义分析树核的句子相似度计算[D]. 大连:大连理工大学,2008.

[11] 蓝雁玲,陈建超. 基于词性及词性依存的句子结构相似度计算[J]. 计算机工程,2011,37(10):47-50.

[12] 腾少冬,王志良,王　莉,等. 基于马尔可夫链的情感计算建模方法[J]. 计算机工程,2005,31(5):17-19.

[13] 王志良,解　仑,董　平. 情感计算数学模型的研究初探[J]. 计算机工程,2004,30(21):33-34.

[14] 微软研究院释库[EB/OL]. (2015-07-18). http://www.datatang.com/data/14263.

[15] 李佳媛. 汉语句子相似度计算技术及其应用[D]. 北京:北京信息科技大学,2013.

[16] 庄成龙,钱龙华,周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报,2009,23(1):3-8.

[17] 刘宏哲. 一种基于本体的句子相似度计算方法[J]. 计算机科学,2013,40(1):251-256.

编辑　顾逸斐