文章编号: 1003-0077(2010)03-0024-05

一种基于语境的词语相似度计算方法

蔡东风,白 宇,于 水,叶 娜,任晓娜

(沈阳航空工业学院 知识工程中心, 辽宁 沈阳 110034)

摘 要: 词语相似度计算是机器翻译、信息检索等自然语言处理领域的关键问题之一。传统的词语相似度计算方法,未能很好地考虑上下文信息对词语语义的约束,从而不能对语境变换带来的词语间相似度的差异进行有效的区分。该文引入模糊数学中隶属函数的概念计算词语上下文信息的模糊重要度,并结合基于《知网》的语义相似度计算方法,提出一种基于语境的词语相似度计算方法。实验表明,该算法可以根据语境有效地区分语义相近的词语。

关键词: 计算机应用;中文信息处理;语境;模糊重要度;词语相似度;隶属函数

中图分类号: TP391 **文献标识码**: A

A Context Based Word Similarity Computing Method

CAI Dongfeng, BAI Yu, YU Shui, YE Na, REN Xiaona

(Knowledge Engineering Center, Shenyang Institute of Aeronautical Engineering, Shenyang, Liaoning 110034, China)

Abstract: Word similarity computation is one of the key issues in natural language processing fields, such as machine translation, information retrieval etc. As traditional methods ignore the context information of the word, they can not effectively distinguish the differences among the word similarities when the context information changes. This paper presents an approach for word similarity computation based on the context information, which employs the fuzzy membership functions to compute the fuzzy significance of the words and combines a method of word similarity calculation using HowNet. The experimental results indicate that our approach distinguish the semantic similar words effectively by the context information.

Key words: computer application; Chinese information processing; context; fuzzy degree of significance; word similarity computation; membership function

1 引言

词语相似度计算是自然语言处理领域中的关键问题之一,在机器翻译、信息检索等方面有着重要的应用价值。在不同的应用中,词语相似度有不同的用途,比如,在基于实例的机器翻译中,词语相似度能够体现文本中两个词语的可替换程度;在信息检索中,利用词语相似度能够提高信息检索的准确率和召回率;在问答系统中,答案和问句的符合程度可以通过计算两者含有词语之间的相似度来衡量。另外,在构造统计语言模型的过程中,由数据稀疏导致未登录词的统计信息无法计算的问题,可以通过词语相似度计算对词语进行聚类,以词类作为统计信息,改善统计语言模型的数据稀疏问题,从而提高语言模型的表现力。

词语相似度计算不应该忽略词语所处的语境。在实际应用中,某个词语的具体语义根据所处的语境不同而有一定的差异,因此计算词语的相似度不应该忽略词语的上下文信息。本文利用模糊数学中隶属函数计算词语上下文信息的模糊重要度,并结合基于《知网》的词语语义相似度的计算方法,提出

收稿日期: 2009-08-12 **定稿日期**: 2010-03-04

基金项目:国家自然科学基金资助项目(60842005);辽宁省教育厅科技研究资助项目(2007T140)

作者简介:蔡东风(1958—),男,教授,主要研究方向为人工智能、自然语言处理;白宇(1982—),男,助教,主要研究方向为自然语言处理;于水(1984—),男,硕士生,主要研究方向为自然语言处理。

一种基于语境的词语相似度计算方法,通过对测试语料集中的词语进行测试,该方法准确率达到 70 %。

2　相关研究工作

国内外对词语语义相似度的计算方法大体可分为两类:基于统计的词语语义相似度计算方法和基于语义知识的词语相似度计算方法。

基于统计的词语语义相似度计算方法是一种经验主义方法,它把词语相似度的研究建立在可观察的语言事实上,而不仅仅依赖于语言学家的直觉。它是建立在两个词语语义相似当且仅当它们处于相似的上下文环境中这一假设的基础上,它利用大规模语料库,将词语的上下文信息作为语义相似度计算的参照依据[1]。基于统计的定量分析方法能够对词语间的语义相似性进行比较精确和有效的度量,但该方法依赖于训练所使用的语料库,计算量大且方法较复杂。另外,受数据稀疏和数据噪声的干扰较大,有时会出现明显的错误。

L. Lillian 利用相关熵, P. Brown 等人采用平均互信息来计算词语之间的相似度[2-3]。Dagon 等人使用了更为复杂的概率模型来计算词语的距离[4]。胡俊峰等人利用上下文的词语向量空间模型来近似地描述词语的语义,再在此基础上定义词语的相似关系[5]。由于该文概念相似的计算只停留在词汇层面,使检索结果中很多词与检索概念相关,但整首诗的意境未必与检索概念相关。章志凌等人基于统计的方法提出一种优化的 Corpus 库[7],目的是把在大规模语料库中统计得来的丰富信息进行筛选并存储,作为词和词之间相似度量化的信息基础。Corpus 库可以把浩瀚的语料库中所蕴涵的词和词之间的关系通过统计的方法提取出来并进行存储,为上层的词语关系量化的计算提供支持。

基于语义词典的词语相似度计算方法是一种基于语言学和人工智能的理性主义方法,它利用语义词典,依据概念之间的上下位关系和同义关系[8],通过计算两个概念在树状概念层次体系中的距离来得到词语间的相似度。基于概念词典的方法建立在两个词汇具有一定的语义相关性,并且它们在概念间的结构层次网络图中存在一条通路这一假设的基础上。这种方法直观、简单有效且易于理解,但是它依赖于比较完备的按照概念间结构层次关系组织的大型语义词典,受人的主观影响比较大,有时不能反映客观现实。

R. Rada 和 J. H. Lee 等人就是通过计算 WordNet 中词节点之间上下位关系构成的最短路径来计算英文词语之间的相似度的[9-10]。有些研究者考虑的情况更复杂,例如 Resnik 根据两个词的公共祖先节点的最大信息量来衡量两个英文词语的语义相似度[11]。在汉语词语相似度的计算研究方面,王斌采用树形图中节点之间路径的方法[12],利用《同义词词林》来计算汉语词语之间的相似度。刘群等人提出一种基于《知网》的词汇语义相似度计算方法[6],该方法在计算两个概念的语义表达式之间的相似度时,采用了"整体的相似度等于部分相似度加权平均"的做法,对于两个义原的相似度,采用根据上下位关系得到语义距离并进行转换的方法。

3　基于语境的词语相似度计算

3.1　存在问题及主要工作

基于《知网》的语义相似度算法简单易行,但是某些语义相近的词无法单纯利用《知网》对其进行区分。例如:对于词语"尊重"、"崇拜"、"敬仰"、"佩服"和"尊敬"等近义词语,利用文献[6]提出的基于《知网》的词汇语义相似度计算方法,得到上述任意两个词语之间的相似度都为 1,因而无法对它们进行语义区分。

针对上述词语相似度计算方法存在的不足,本文利用统计学方法与之融合,提出一种语义与语境相融合的词语相似度算法。在实际应用中,需要进行相似度计算的词语往往出现在具体的语言环境中。如:"尊重"一词与"尊敬"、"崇拜"、"敬仰"、"佩服"等词语的语义相近,若给定上下文信息"领导的心理是想让人尊重他,特别是在一些大众场合,领导者都很注重自己的形象",那么这里的"尊重"应与"尊敬"一词的语义更相近。本文将上下文信息称为词语的语境,将其引入到词语相似度的计算过程中。

在对词语上下文信息进行统计时,其结果可能存在噪声(例如某个数值过大),对相似度计算的值会产生较大的影响。对数据的有效处理是改进统计方法计算词语相似度效果的途径之一。因此,本文引入模糊数学的相关理论,有效地解决了由上述情况引起的问题。

3.2　相关定义

将模糊数学的相关概念引入到相似度的计算

中,设词语 k 的上下文向量为 key,k 的候选相似词集合为 $W=\{w_1, w_2, \cdots, w_N\}$。对平均共现次数($AVG$)、中间共现次数($MID$)、第二中间数($SecondMID$)和重要度标尺分别给出如下定义:

定义1:平均共现次数(AVG):表示词语 w_i 与 key 的共现次数的平均数,如式(1):

$$AVG=\frac{\sum_{i}^{N} counter(w_i, key)}{N} \tag{1}$$

$counter(w_i, key)$ 表示词语 w_i 与 key 的共现次数。

定义2:中间共现次数(MID):表示共现次数的中间数,如式(2):

$$MID=\frac{\max_{w_i\in W}(counter(w_i, key))+\min_{w_i\in W}(counter(w_i, key))}{2} \tag{2}$$

$\max(counter(w_i, key))$ 和 $\min(counter(w_i, key))$ 分别表示集合 W 中的词语 w_i 与需要处理词语 k 的上下文向量 key 的共现次数中的最大值和最小值。

定义3:第二中间数($SecondMID$):如式(3):

$$SecondMID=\frac{secondmax_{w_i\in W}(counter(w_i, key))}{2}+\frac{secondmin_{w_i\in W}(counter(w_i, key))}{2} \tag{3}$$

$secondmax(counter(w_i, key))$ 和 $secondmin(counter(w_i, key))$ 分别表示集合 W 中的词语与需要处理词语 k 的上下文向量 key 的共现次数中次大值和次小值。

定义4:重要度标尺:

对每组测试数据得到的 AVG, MID, $SecondMID$, $\max(counter(w_i, key))$, $\min(counter(w_i, key))$ 按照数值从小到大排序,分别定义为 $sort_i(i=1,2,\cdots,5)$,并且按照这个次序定义一个标尺,本文将其称为重要度标尺。本文将刻度标记为 0.5, 0.625, 0.75, 0.875, 1,如图1所示。

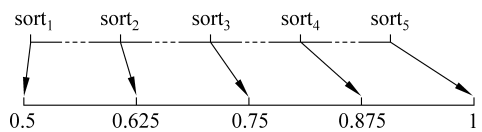

图1 重要度标尺

3.3 模糊重要度计算

在模糊数学[13]中,若对论域 U 中的任一元素 x,都有一个数 $A(x)\in[0,1]$ 与之对应,则称 A 为 U 上的模糊集,$A(x)$ 称为 x 对 A 的隶属度。本文借鉴了隶属度的相关定义,构造词语在上下文语境中的模糊重要度。对于同一组测试集合中的词语,利用3.2节中定义的重要度标尺,依照公式(4)得到词语 w 的模糊重要度 $degree(w)$。

模糊重要度定义为:

$$degree(w)=lowsign(w)+0.125\times\frac{counter(w)-lowtab(w)}{uptab(w)-lowtab(w)} \tag{4}$$

这里 $lowsign(w)$ 表示为词 w 在重要度标尺中相对应重要度区间的下界($lowsign(w)\in\{0.5, 0.625, 0.75, 0.875\}$),$uptab(w)$ 和 $lowtab(w)$ 分别对应词 w 在重要度标尺中相应区间的上界和下界($uptab(w), lowtab(w)\in\{sort_i \mid i=1,2,\cdots,5\}$)。

考虑到两个词之间的同现关系对计算词语的相似度计算有着重要的作用,将两个词之间的同现信息引入到相似度计算公式中,利用公式(5):

$$I(w_1; w_2)=\log\frac{p(w_1 w_2)}{p(w_1)\,p(w_2)} \tag{5}$$

上式得到两个词语之间的点互信息 $I(w_1; w_2)$,并将其利用上述的模糊重要度计算方法进行计算得到模糊互信息重要度 $Ifuzzy(w_1, w_2)$。

综合考虑语义网络,词语的语境以及词语间的统计互信息,将各个部分的相似度信息进行组合,得到如下相似度计算公式:

$$similtide(w_1, w_2)=(1-\alpha-\beta)\times Semantic(w_1, w_2)+\alpha\times degree(w_2)+\beta\times Ifuzzy(w_1; w_2) \tag{6}$$

其中,$Semantic(w_1, w_2)$ 为利用《知网》相似度计算工具计算得到的词 w_1 与词 w_2 的相似度。在采用《知网》相似度计算工具进行词语间相似度计算时,若词语包含多个义项,则选择义项间相似度的最大值作为词语间的相似度。

4 实验及结果分析

由于词语相似度是一个主观性很强的概念,现在还没有一个普遍适用的词语相似度计算测试集。本文在选用国家公务员考试题型中的词语替换题作为测试语料,该题目具有以下特点:

a) 待计算相似度的词语具有一个具体上下文。

b) 有一组语义相似的词作为替换词。

c) 每一组测试题答案都有一个确定的答案可

以进行评价。

采用从 2002～2008 年国家公务员考试真题中选取的 50 个词语替换题作为测试问题集,题型实例如下:

这个故事听起来很真实,但它是[杜撰]的。

A 草拟　　B * 虚拟

C 撰写　　D 写真

在对外关系上,我们一贯[奉行]独立自主的和平外交政策。

A 遵守　　B 遵循

C * 实行　　D 实施

每道题都有 4 个备选答案,解答者从中选择一个与题干括号中词语最接近的词。

采用准确率(*precision*)和平均排序倒数(Mean Reciprocal Rand, *MRR*)两个指标对算法进行评价,当正确选项与括号中词语的相似度大于其余选项与该词语的相似度时,认为系统返回该题的正确答案,否则认为答案错误。评测公式如下:

$$precision = \frac{\text{系统返回正确答案个数}}{\text{测试语料总数}} \times 100\ \%$$

排序倒数(Reciprocal Ranking, RR)是算法返回结果中正确结果出现位置的倒数,平均排序倒数是多次计算的 RR 的结果的平均值,利用以下公式:

$$MRR = \frac{1}{N} \times \sum_{i}^{N} \frac{1}{n_i}$$

其中, N 表示题目总数, n_i 表示对于第 i 个题目算法返回的第 n_i 个答案为正确答案。

4.1　实验步骤

本文以网络上的文本作为统计语料,利用网络搜索引擎①返回词语共现次数。具体的算法描述如下:

针对问题 T_i (i = 1 to n),得到待替换的词语 k_i, k_i 的上下文向量 key_i 以及相似词语候选集合 W_i。

for 每个 T_i 的候选词集 W_i do

　　for W_i 中的每个项 w_{ij} do

　　计算 w_{ij} 与 key_i 的共现次数, w_{ij} 与 k_i 的出现概率及共现概率,利用《知网》计算 w_{ij} 与 k_i 的相似度 $Semantic_{ij}$;

　　for W_i 中的每个项 w_{ij} do

　　计算 w_{ij} 的模糊上下文重要度 $degree_{ij}$ 和互信息重要度 $Ifuzzy_{ij}$,计算得出 w_{ij} 与 k_i 的相似度 $similarity_{ij}$;

　　选取 $similarity_{ij}$ 最大的 w_{ij} 作为 k_i 的替换词。

4.2　参数选定

实验在测试题库中随机抽取了《知网》的几个不同类别的词语进行相似度计算,通过对 α, β 取不同的值,得到的准确率如表 1 所示,从表 1 观察可得:当 α = 0.2, β = 0.2 时实验结果最优。

表 1　参数选择

α \ β	0	0.1	0.2	0.3	0.4	0.5
0	0.26	0.40	0.40	0.40	0.44	0.44
0.1	0.40	0.40	0.50	0.48	0.48	0.46
0.2	0.54	0.60	**0.68**	0.64	0.60	0.40
0.3	0.50	0.62	0.64	0.60	0.52	0.42
0.4	0.50	0.44	0.54	0.50	0.50	0.50
0.5	0.52	0.46	0.50	0.56	0.60	0.52

4.3　词语替换题测试结果

本文采用如下的词语相似度计算方法在相同的测试集上进行测试:

A) 传统的《知网》相似度计算方法;

B) 文献[14]中采用的改进的《知网》词语相似度计算方法;

C) 统计与语义结合计算方法[15];

D) 基于语境的词语相似度计算方法(本文方法)。

对整个测试集,所得到的正确率与 *MRR* 值如表 2 所示。

表 2　正确率及 MRR 值

方法	正确率	MRR
A	0.40	0.58
B	0.46	0.70
C	0.65	0.77
D	0.70	0.84

通过表 2 可以看到,本文所采用的基于语境的相似度计算方法可以有效地区分出语义相近的相关词。通过 *MRR* 值可以看出,采用该方法得到的词语相似度计算结果,可以对相似词语与原词语的相似程度给出一个较合理的排序,即可以给检索系统提供一个准确的待扩展词的相似度排序。

4.4　错误结果分析

通过对算法返回错误结果集的分析发现,计算

① www.baidu.com。

错误的部分包含大量的《知网》不可区分的题目(题目中含有计算所得相似度相同或者《知网》中不包含的词语),这一部分占整个测试集的 56 %。把《知网》可区分的题目单独进行了相似度计算实验,实验结果表明在可区分部分测试集上,本文方法的准确率达到 0.80。

对实验语料分析发现在一些长度较短的测试题目中,其查询关键词的上下文不能完全表现关键词的真实语境,例如:

领导的心理是想让人(尊重)他

该句中待替换的词语"尊重"的上下文是"领导"和"心理"两个词语,而从主观上我们可以看出单纯利用这两个词不能明确的表现"尊重"的真实语境,那么在计算替换词的模糊重要度时,得出的结果就不能正确地表征待替换词与其他替换词的相似度远近关系。

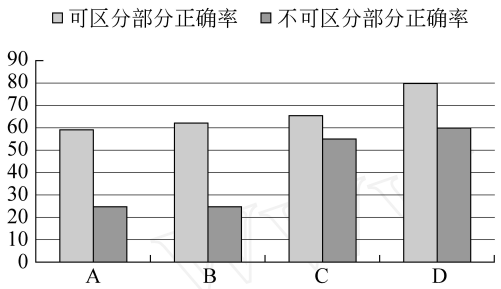

图 2 对比实验结果

针对这种情况,可以从改进语境的方面进行下一步工作:当算法抽取的上下文的长度小于某个阈值时,利用句子相似度算法从语料库中选取扩充问句,并从中提取特征语境,以达到提高基于语境的词语相似度计算方法性能的目的,这也是本文下一步要进行的工作。

5 结束语

研究词语相似度不能离开词语具体的语境,本文将模糊隶属度的概念引入到语境相似度的计算过程中。提出模糊重要度的概念并将其与《知网》计算出的相似度结合,得到一种《知网》语义资源与上下文语境相融合的词语相似度计算方法。通过对公务员考试题库中选取的词语替换题型进行测试,算法准确率达到 0.70,表明该算法可以根据语境较有效地区分语义相近的词语。

参考文献

[1] Miller GA, Fellbaum C. Semantic network of English [M]. Levin B, pinker S. lexical & conceptual semantics Amsterdam, Netherlands: Elsevier Science Publishers, 1991.

[2] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. (1991) Word sense disambiguation using statistical methods[C]// Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91), Berkley, C. A., 1991:264-270.

[3] Lillian Lee. Similarity-Based Approaches to Natural Language Processing[D]. Ph. D. thesis. Harvard University Technical Report, TR-11-97.

[4] Dagan I, Lee L. Similarity-based models of word cooccurrence probabilities [J]. Machine Learning. Special Issue on Machine Learning and Natural Language, 1999.

[5] 于江生, 俞士汶. 中文概念词典的结构[J]. 中文信息学报, 2002, 16 (4): 13-21.

[6] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. Computational Linguistics and Chinese Language Processing, 2002, 7(2):59-76.

[7] 章志凌, 等. 基于 Corpus 库的词语相似度计算方法[J]. 计算机应用, 2006, 26 (3): 638-640.

[8] 秦春秀, 赵捧未, 刘怀亮. 词语相似度计算研究[J]. 信息系统, 2007, 30(1):105-108.

[9] Rada R. Development and application of a metric on semanticnets [C]// IEEE Transactions on System. Man and Cybernetics, 1989.

[10] Lee J H. Information retrieval based on conceptual distance in ISA hierarchies [J]. Journal of Documentation, 1993.

[11] Philip R. Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language [J]. Journal of Artificial Intelligence Research, 1999, (11): 95-130.

[12] 王斌. 汉英双语语料库自动对齐研究[D]. 北京: 中国科学院计算技术研究所, 1999.

[13] 谢季坚, 刘承平. 模糊数学方法及其应用[M]. 华中科技大学出版社 2006. 15-37.

[14] 余超. 基于知网的词语相似度计算研究及应用[D]. 沈阳: 沈阳航空工业学院, 2006.

[15] 郭丽. 基于上下文的词语相似度计算及其应用[D]. 沈阳: 沈阳航空工业学院, 2009.