统计与语义相融合的词语相似度计算

郭丽 蔡东风 季铎 白宇

（沈阳航空工业学院知识工程中心，沈阳，110034）

E-mail: gg4179@163.com

摘要： 词语相似度计算是自然语言处理领域的基础和研究难点。本文利用知网的相关概念场，将词语的相关性引入到基于平均互信息的词语相似度计算中，并提出了统计和语义相结合的词语相似度计算方法，采用了国家公务员考试“词语替换”题型进行测试，实验结果显示，准确率为 0.65。

关键词： 词语相似度，知网，统计，语义

Word Similarity Calculation based on Statistic and Semantic

Guo Li Cai Dong feng Ji Duo Bai Yu

（Knowledge Engineering center of Shenyang Institute of Aeronautical Engineering，ShenYang，110034）

E-mail: gg4179@163.com

Abstract: Word Similarity calculation is one of the fundamental and key problems in the research of Natural language processing .This paper imports the word correlation from related concept field in Hownet for calculating word similarity which is based average mutual information, a method based on the fusion of statistic and semantic is proposed to calculating word similarity, besides tests from word replacement problems of national civil service examinations, the result of the experiment shows that the precision is 0.65.

Keywords: word similarity calculation; Hownet; statistic; semantic

1 引言

词语相似度计算是自然语言处理领域研究的基础，在机器翻译、聚类、信息检索等方面有着重要的应用价值。而在不同的应用中词语相似度都有不同的含义，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，而在信息检索中，词语相似度的计算能够提高信息检索的准确率和召回率，如问答系统中问句和答案的符合程度可以通过两者含有词语的词语相似度来衡量。另外，在构造统计语言模型的过程中，由于数据稀疏导致未登录词的统计信息无法计算，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，从而提高模型的表现力。

现有的词语相似度计算方法主要有两种，基于统计的和基于语义知识的词语相似度计算方法。基于统计的词语相似度计算方法假设，如果两个词语的语义在某种程度上相似，那么它们上下文的概率分布也相似，因此可以将词语上下文的概率分布作为计算词语语义相似度的依据。如Brown[3]的基于平均互信息的方法计算词语相似度。基于统计的相似度计算方法能够对词语间的语义相似度进行有效的度量，但是，计算量大，方法复杂，并且训练采用的语料库的规模和质量决定了词语相似度计算结果的优劣。基于语义知识的计算方法，根据所使用语义资源的不同可以分为两类，一类是基于层次结构的树状语义词典，如《同义词词林》和 WordNet 等等；另一类是基于网状结构的语义网络，如知网。基于知网的概念相似度计算是当前研究的热点，如文献[2]提出的实词概念相似度计算的方法。基于语义知识的词语相似度计算方法，是从语义资源中直接获取信息，能够较准确地反映词语之间的复杂关系，并且计算简单。但是，因为对词语概念的描述详细程度并没有一个明确的规定，所以，获得的语义信息，有时并不能真实准确的反映词语间

的语义差别。

针对这个问题，本文提出了一种统计和语义相结合的词语相似度计算方法。在第二部分中提出采用平均互信息计算词语相似度，分析导致部分词语相似度结果不正确的原因，利用知网的语义资源，针对近义词间的相似度，提出了统计和语义相结合的词语相似度计算方法，使词语相似度的计算结果更加合理，并详细介绍了计算方法；第三部分采用国家公务员考试的词语替换题作为词语相似度计算的评测方法，实验结果显示正确率为 65%；第四部分对于实验结果进行了详细分析；第五部分为结论与展望。

2 统计和语义相结合的词语相似度计算

2.1 基于平均互信息的词语相似度计算

词语的相关性反映的是词语间互相关联的程度，而词语相关性在一定程度上体现了词语间的语义距离。如提起“足球”，人们就会联想到“球星”，“比赛”，“世界杯”等与“足球”存在联系的词语。这种联系既反映了这些词语都与一个主题相关，又从侧面体现出它们的语义相似。本文将词语间的相关性引入到词语相似度计算中，使得词语间尤其是近义词间细微的语义差别得以体现。例如，“苹果”、“梨”、“栗子”，这些词语的概念都是一种水果，但是“苹果”和“梨”是仁果的一种，而“栗子”是坚果的一种，“苹果”和“梨”的相互联系的程度要比“苹果”和“栗子”，“梨”和“栗子”更强，而这一点也正体现了词语间的语义距离和语义差别。因此，引入词语相关度计算可以改善词语相似度计算结果。

基于统计的方法是在大规模语料中通过对词语相关性的计算来体现词语的相似性，并且多使用互信息或平均互信息来衡量词语相关度。在信息论的噪声信道模型中，点互信息 I（x,y）可以取正值也可取负值，如果点互信息取负值，说明输出端由于受到所接收消息的影响反而使对输入端是否输入 x 的熵增大，即不确定性增加了，这显然违背了本文计算词语间语义相似度的初衷。因此，本文采用平均互信息来度量词语间的语义相似度，平均互信息是对点互信息求数学期望的结果，因此更能反应点互信息的趋势。平均互信息值非负，这符合人们对词语相似度的常规理解，即使两个词语在语义上相差甚远，它们的语义相似度也应该为零而不是负值。

平均互信息公式：

$$I(X,Y)=\sum_{xy}P(XY)\log\frac{P(xy)}{P(x)P(y)} \qquad (1)$$

本文使用平均互信息作为词语间相关度的度量，对于两个词语 X，Y，它们的相似度表示如下：

$$Rel(X,Y)=\sum_{i=1}^{2}\sum_{j=1}^{2}p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \qquad (2)$$

其中X含有两种情况，x_1表示单词X出现，x_2表示单词X不出现，Y的含义同理，$p(x_1,y_2)$表示单词X和单词Y同时出现的概率，$p(x_1,y_2)$表示单词X出现，但是单词Y不出现的概率；同理，$p(x_2,y_1)$表示单词Y出现，但是单词X不出现的概率；$p(x_2,y_2)$表示单词X和单词Y都不出现的概率。

本文将利用平均互信息计算出的词语相似度称作统计相似度，由于词语的多义性，使得统计数据可靠性降低，所以引入语义相似度计算。

2.2 基于语义的词语相似度计算

基于语义的词语相似度计算多借助于大型的语义词典，而词典的完备性在很大程度上影响了词语相似度计算的准确性。知网（英文名称为HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。知网对概念的描述是要着力体现概念与概念和概念的属性与属性之间的相互关系，例如，“打”有{DEF=buy|买}，{DEF=beat|打}等概念。

知网是一个知识系统，并不是一部语义词典[6]，它利用一种知识描述语言将词语构成一个大型的语义网络，为语义计算提供了大量的宝贵信息。本文利用知网的宝贵资源进行语义相似度计算，并采取统计相似度和语义相似度相结合的方法计算词语相似度，计算两个词语X，Y的相似度方法如下：

$$Sim_{corr}(def_1, def_2) = \alpha Rel(X,Y) + (1-\alpha) \times Semantic(def_1, def_2) \qquad (3)$$

其中def_1和def_2分别表示词语X，Y要计算相似度的知网中的概念，α是语义相关度的权重。$Semantic(def_1, def_2)$是基于知网的语义相似度计算，计算公式如下：

$$Semantic(def_1, def_2) = \sum_{i=1}^{4} \beta_i sim_i(def_1, def_2) \qquad (4)$$

下面介绍公式（4）中$sim_i(def_1, def_2)$，i=1，2，3，4的具体计算方法。

本文引用文献[5]中的词语相似度计算方法中的前三步计算公式，具体如下：

1) 第一独立义原描述式相似度$sim_1(def_1, def_2)$；
2) 其它独立义原描述式：语义表达式中除第一独立义原以外的所有其它独立义原相似度$sim_2(def_1, def_2)$；
3) 关系义原描述式：语义表达式中所有的关系义原描述式$sim_3(def_1, def_2)$；

知网中的相关是概念间的相关，相关概念场是相关概念的集合。知网中描述这些概念的词语（知网称为义原）与被描述的概念在语义上是一致的，由此，本文将概念间的相关转换为词语（义原）间的相关，即把词语相关概念场用于词语相似度计算。

计算两个词语X,Y的概念(def_1, def_2)之间的相关度，分别提取概念(def_1, def_2)的相关概念场，得到两个概念的相关词语的集合Sem_X和Sem_Y，利用Dice系数计算两个词语的相关度为：

$$sim_4(def_1, def_2) = \frac{2|Sem_X \cap Sem_Y|}{|Sem_X| + |Sem_Y|} = \frac{2S}{S_X + S_Y} \qquad (5)$$

其中，$S = |Sem_X \cap Sem_Y|$是同时出现在集合Sem_X和Sem_Y的相关词语的个数，$S_X = |Sem_X|$和$S_Y = |Sem_Y|$分别表示集合Sem_X和Sem_Y中的相关词汇个数。例如，“男人”，取其DEF的表达式为“{human|人：modifier={male|男}}”，利用知网得到它的相关概念场，即词汇的集合为{男礼服，男装，二老，双亲，兄长，兄弟，老兄……}，集合中共含有512个相关词汇；“女人”取其DEF表达式为“{human|人:modifier={female|女}}”，得到它的相关概念场为{女孩子气，二老，双亲，妇德，女衬衫，女装，大姐，女足……}，集合中共含有915个相关词汇，同时出现在两个集合中词语为117个，则“男人”和“女人”的语义相似度为0.16398。

将$sim_i(def_1, def_2)$，i=1，2，3，4这四部分的相似度计算公式组合，得到基于知网的语义相

似度计算公式（4），其中β_i代表各部分相似度的权重，取值为

$$\beta_1 = 0.1, \beta_2 = 0.1, \beta_3 = 0.6, \beta_4 = 0.2$$

3 实验结果

本文采取对实验结果进行人工评价的方法确定式（3）中的参数α，评价方法主要是采用搜狗实验室互联网语料库 2.0 作为训练语料，随机抽取知网中文词表中 200 个词语根据公式（3）进行词语相似度计算，对计算得到的语义相似度的序列和人工排列结果进行比较，结果表明当α=0.2 时，系统的表现最好。本次实验采用了词语替换测试题和对比实验两种测试方式，词语替换测试题选取了国家公务员考试中的词语替换题，对比实验的测试数据为知网中随机抽取的 10 组同类词。

3.1 词语替换题测试结果

鉴于人工评测的主观性，本文选用了公务员考试中的“词语替换题型”的 40 道题（包括真题、训练题，带有标准答案），共计 200 个词语作为测试数据，列举其中的部分试题如表 1 所示。

试题题目中标有括号的词语是需要进行替换的词语，答案选项中的词语是可供替换选择的候选答案词语，需要从这些候选答案中选择最准确的词语进行替换，测试方式如下：

1) 将题目中括号内的词语和选项中每一个候选答案进行词语相似度计算；
2) 选择与需要替换的词语相似度最大的候选答案作为正确答案返回。

表 1 测试试题

题目	替换词	选项 A	选项 B	选项 C	选项 D	正确答案
公司的重组使公司(摆脱)了严重亏损的局面，避免了终止上市。	摆脱	脱离	离开	走出	甩掉	脱离
这些学子，来自(五湖四海)，相聚在知识的殿堂	五湖四海	四面八方	四通八达	五花八门	五颜六色	四面八方

对测试语料中所有试题采用上述方法进行测试，正确率为 65%，本文采用知网相似度计算方法[5]，基于语境相似度计算方法[7]，和本文相似度计算方法进行对比测试，实验结果如图 1，图 2 所示。

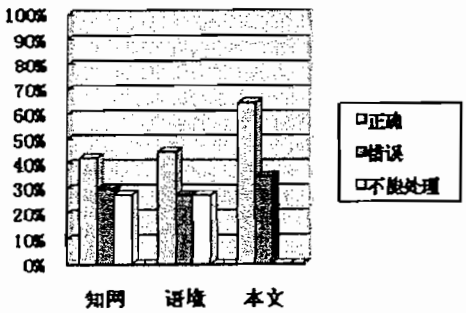

图 1 测试结果

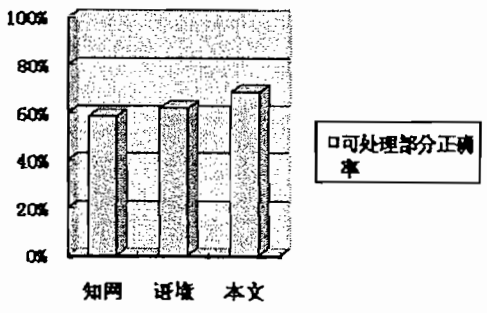

图 2 正确率

3.2 对比实验结果

表 2 对比实验结果

单词 X	单词 Y	知网相似度	基于语境相似度	本文相似度
水果	橘子	0.95	0.95	0.954
	菠萝	0.95	0.95	0.954
	瓜果	0.95	0.95	0.961
	花生豆儿	0.95	0.95	0.951
花生豆儿	瓜果	0.95	0.95	0.977
	水蜜桃	0.95	0.95	0.951
	栗子	0.95	0.95	0.965
沉湎	沉迷	1	1	0.932111
	沉浸	0.95	0.95	0.932741
	沉溺	0.95	0.95	0.935938
	沉醉	1	1	0.932176
趋势	局面	0.028571	0.028571	0.0303245
	前景	0.95	0.95	0.831833
	趋向	1	1	0.832081
	苗头	0.000624	0.000624	0.00343768

在这里随机抽取了知网的 10 个不同类别的词语进行相似度计算，分别采用知网词语相似度计算方法[5]，语境相似度计算方法[7]，和本文相似度计算方法三种方法进行对比试验，部分结果如表 2 所示。

4 实验结果分析

因为在计算相似度时，不同的词语间会计算出相同的相似度值，本文称这样的试题为不能处理部分，测试语料中含有 11 道知网不能处理的试题。将可以处理的试题抽取出来，对三种相似度进行对比，知网计算的正确率为 58.6%，基于语境的相似度计算正确率为 62.1%，本文方法计算的正确率为 68.9%。对不能处理部分的试题，用本文的计算方法进行计算，单纯使用统计手段得到的正确率为 50%，正确率为 55%。对结果进行分析，发现相关度对计算近义词之间的相似度时起到了很大的作用，而知网在计算词语相似度的时候，因为缺少词语共现的相关数据，对于近义词（同类词）之间的相似度不能给出区分，这点也证明了词语相关度对词语相似度起到了一定了影响。例如下题：

有相当一部分人没有受到科学技术观的熏陶，对于科学技术茫然无知，他们还(沉湎)于封建迷信的毒害中，任凭落后愚昧的愚弄。（正确答案是 C）

A 沉迷　B 沉浸　C 沉溺　D 沉醉

知网词语相似度、基于语境相似度以及本文相似度表 3 所示：

表 3 试题测试结果

单词 x	单词 y	知网相似度	基于语境相似度	本文相似度
沉湎	沉迷	1	1	0.932111
	沉浸	0.95	0.95	0.932741
	沉溺	0.95	0.95	0.935938
	沉醉	1	1	0.932176

由表 3 可以看出无论是知网还是改进的基于语境的词语相似度计算结果，都因知网对词语概念的描述详细程度不同，“沉湎”与选项 A、D 的相似度都是 0.95，而不能给出确切的答案，而本文的计算方法在这里得到了很好的效果，说明词语相关度对于词语相似度具有一定的影响，把词语相关度引入到词语相似度计算中是合理的，同时也看到词语的共现信息在词语的相似度计算中具有重要的意义。

另外，导致答案出错的主要原因是词语相似度计算是在确定的语境中进行的，这里提到的语境是指，词语当前所处句子的上下文，然而本文的相似度计算方法并没有考虑到真实语境对词语相似度的影响，以及词语在语用上的相似度，从而导致了错误的出现。

5 结论与展望

语义相似度计算的研究是很多重点研究领域的基础，尽管已经有很多学者进行了大量的研究，但是由于汉语词汇的复杂性和多义性，再加上词汇语义概念较强的主观性、具体应用领域的专业性等因素，使得准确的计算词语相似度一直是自然语言处理领域的难题，本文采用的基于统计和语义相结合的方法不失为一种实践上行之有效的语义相似度量化方法。

本文简单介绍了现有基于知网的词语相似度计算方法，提出了基于统计和语义相融合的词语相似度计算方法，证明了词语相关度在词语相似度的计算中具有重要的意义。对实验结果进行综合分析后，提出了词语当前所处的句子的上下文，即实际语境对词语相似度计算有较大影响，因在后续的工作中，将要把词语相似度计算结果用于自动问答系统的答案抽取部分，所以，会将词语实际语境加入到词语相似度计算中，并考察词语相似度计算在实际应用中的贡献。

参考文献：

[1].Peter D. Turney. Similarity of Semantic Relations. Computational Linguistics Journal. Volume 32, Issue 3. September 2006. p 379-416.

[2].刘群，李素建 基于知网的词汇相似度计算.http://www.keenage.com，2002

[3].P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer.(1991) Word sense disambiguation using statistical methods. In Proceedings of the 29th Meeting of the Association for Computational Linguistics(ACL-91),pages 264-270,Berkley,C.A.,19 91.

[4].余超，蔡东风，张桂平 词汇语义相似度计算中相关技术的分析. 2006 年第三届学生计算语言学研讨会论文集 p127-133

[5].Dong ZhenDong.HowNet and Computation of Meaning[M]. Singapore: World Scientific press,2006.187-195

[6].董振东，董强. 知网[EB/OL]. http://www.keenage.com, 2002

[7]余超.基于知网的词语相似度计算研究及应用.2006 年沈阳航空工业学院硕士毕业论文

[8].赵军，金千里，徐波 面向文本检索的语义计算.2005 年计算机学报第 12 期 28 卷

[9]关毅，王晓龙.基于统计的汉语词汇间语义相似度计算. 2003 年全国第七届计算语言学联合学术会议

[10].颜伟，荀恩东 基于 WordNet 的英语词与相似度计算[A] 2004 年全国计算语言学学生会议论文集[C] p282-288

[11].搜狗实验室互联网语料库 2.0 http://www.sogou.com/labs/