基于 HowNet 的信息量计算语义相似度算法①

游 彬[1], 严岳松[1], 孙英阁[2], 刘 靖[3]

[1](海军指挥学院 信息战研究系, 南京 211800)
[2](中国人民解放军海军 92665 部队, 常德 415300)
[3](海军指挥自动化站, 北京 100841)

摘 要: 语义相似度计算的应用范围广泛, 从心理学、语言学、认知科学到人工智能都有其应用. 提出了仅依赖于知网(HowNet)的信息量计算来估计两个词汇间的语义相似度. 经实验证明, 相比于传统的基于词网(WordNet)和大型语料库的计算信息量来估计语义相似度的算法, 本文的算法更容易计算, 并更接近于人工的语义相似度判断.

关键词: 语义相似度; HowNet; 信息量; 语义距离; 相似度计算

Method of Information Content Evaluating Semantic Similarity on HowNet

YOU Bin[1], YAN Yue-Song[1], SUN Ying-Ge[2], LIU Jing[3]

[1](Depart. of Information Warfare Research, Naval Command College, Nanjing 211800, China)
[2](People's Liberation Army Navy Force 92665, Changde 415300, China)
[3](Navy Command Automation Workstation, Beijing 100841, China)

Abstract: Evaluating Semantic similarity is widely used in areas range from Psychology, Linguistics, Cognitive Science to Artificial Intelligence. This paper means to the merely use of HowNet to evaluate Information Content as the semantic similarity of two terms or word senses. While the conventional ways of measuring the IC of word senses must depend on both an ontology like WordNet and a large corpus, the experiment proves that the semantic similarity measured in this method is easier to calculate and more closely with human judgments, as HowNet has an elaborate way to represent descriptive object.

Key words: semantic similarity; HowNet; information content; semantic distance; similarity measuring

语义相似度计算在自然语言处理和信息抽取等领域都发挥着重要的作用, 如提高信息抽取任务的精确度, 进行词义排歧, 进行本体实体间的映射, 计算短文本间的相似度[1], 和舆情分析处理的情感倾向性判断[2]等.

通常, 计算两个词汇间的语义相似度会用到两种经典的方法: 基于边的数量(节点距离)的算法[3]——由于共享知识分类(IS-A 关系)的概念被表示为每个概念对应一个节点的分层的树形结构, 因此可以将节点间的距离作为两概念间的语义相似度, 因为一个节点到另一节点的路径越短, 它们就越相似. 另一种是基于信息理论的算法[4,5]——一个概念越抽象, 它的信息量越低, 并且如果一个分类系统有一个唯一的顶层节点, 那么该顶层节点的信息量为 0.

基于边数量的方法虽然计算简单, 但是由于没有语料库的词汇统计信息支持, 计算的结果将会过分依赖于公式的设计和参数的选取. 而基于信息理论的算法则依赖语义词典(如 WordNet)和语料库的词汇统计信息, 其计算出的语义相似度比其他方法更接近人工判断, 因此被广泛采用.

由于语义词典 WordNet 本身就是大型知识库, 其中蕴涵了各类词汇间的统计关系, N. Seco 在文献[6]中

① 收稿时间:2012-06-23;收到修改稿时间:2012-09-04

提出了仅依赖 WordNet 而不使用语料库的方法. 但是 WordNet 是一部英语语义词典, 无法胜任中文词语的相似度计算, 于是国内学者通常基于 HowNet 计算中文词语的语义相似度. 对于国内中英双语语义词典 HowNet, 目前计算词语相似度的研究均是基于语义距离的方法, 而且也没有利用 HowNet 计算英语词汇相似度的研究.

本文据此提出了一种仅依赖 HowNet 的使用信息量来计算词语间语义相似度的算法, 分别研究了 HowNet 用于中文和英文词汇语义相似度的计算. 为了提高计算相似度的精确性, 本文在 HowNet 的分类系统中使用信息量取代义原(Primitive)距离作为义原的相似度.

1　信息理论算法与WordNet信息量算法

1.1 经典信息理论算法

Philip Resnik[4]首次提出使用信息量来计算语义相似度. 对于概念 c, 其信息量计算公式为:

$$ic_{res}(c) = -\log p(c) \tag{1}$$

在公式(1)中, WordNet 中的概念 c 的信息量(Information Content, IC)表示为概念 c 在某给定语料库中出现的概率 p(c)的负对数函数. 那么, 两概念间的语义相似度的计算公式为:

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1, c_2)} ic_{res}(c) \tag{2}$$

根据 Resnik 的思想, 两概念间的语义相似度等于它们间共有的信息含量, 即最详尽的共同抽象(Most Specific Common Abstraction , MSCA)父概念的信息量. 公式(2)中, $S(c_1; c_2)$表示包含子节点 c_1 和 c_2 的概念集.

然而, 这种算法必须依赖一个大型语料库的统计信息(p(c)的值), 使得信息量的计算比较困难.

1.2 基于 WordNet 的信息量计算

WordNet 是由 Princeton 大学的心理学家, 语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典. 在文献[6]中, N. Seco 提出了仅依赖语义词典 WordNet 而不需任何其他语料库的信息量算法, 如公式(3)所示:

$$ic_{wn}(c) = \frac{\log(\frac{hypo(c)+1}{\max_{wn}})}{\log(\frac{1}{\max_{wn}})} = 1 - \frac{\log(hypo(c)+1)}{\log(\max_{wn})} \tag{3}$$

这种算法的思想源于 WordNet 本身就是一部语义详尽的规则词典, 或者说是共享的知识分类系统(通用本体), 拥有很多下位关系节点的概念比叶子节点概念所含的信息量少. 函数 hypo 返回给定概念所含下位关系节点的个数, $\max_{wn}$ 是 WordNet 分类系统中所有概念节点的数量.

为了计算两个概念的语义相似度, 在公式(2)中, 使用 icwn(c)代替 $ic_{res}(c)$计算出的 $sim_{wn}(c_1,c_2)$即为概念 c_1 和 c_2 的相似度.

由于 WordNet 无法计算中文词语间的语义相似度, 国内学者集中研究了基于 HowNet 的语义相似度计算, 但是几乎均是基于 HowNet 义原距离的计算, 这将需要大量的实验来设计公式和选取参数. 对于义原描述的本体 HowNet, 下节将介绍本文如何基于 HowNet 义原的信息量来计算语义相似度.

2　基于HowNet的信息量计算

2.1 HowNet 简介

HowNet[9]是一个揭示概念间关系和概念的属性间的关系的在线知识库. 与 WordNet 不同, HowNet 分层系统不是简单地使用一个概念表示一个节点, 对于每一个“义项(概念)”, 使用一系列的“义原”来描述. “义原”是描述“义项”的基本单位.

HowNet 2000 版包含 55501 个中文义项, 58582 个英文义项和 1621 个义原. 图 1 是“Entity|实体”义原分类系统的片段, 这种结构和 WordNet 的概念分类体系类似.

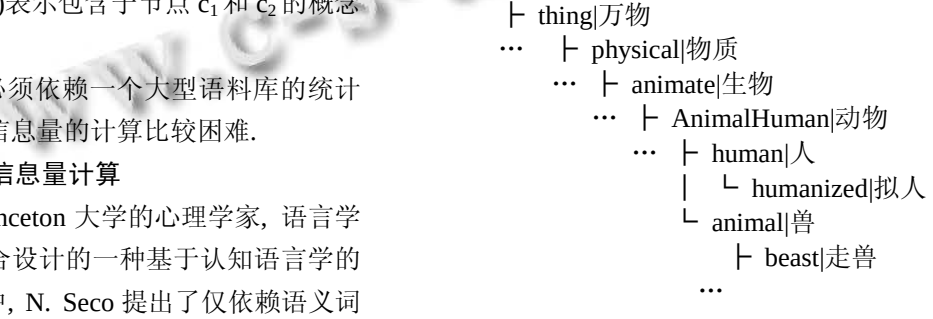

图 1　HowNet 义原分类层次图

对于义原的语义相似度计算, 文献[10]提出基于语义距离的计算公式:

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \tag{4}$$

其中 p_1 和 p_2 表示两个义原(primitive), d 是 p_1 和 p_2 在义原层次体系中的路径长度, 是一个正整数. α 是一个可调节的参数, 需要通过大量实验来确定.

考虑到一些词经常有多个概念, 而且 HowNet 对于一个词语也经常有多个记录, 即描述了词语的多义现象, 本文将待计算相似度的术语或是词汇均当作词语而不是概念来计算.

2.2 基于 HowNet 义原信息量的语义相似度计算

类似于 WordNet 概念的信息量计算, HowNet 本身就是一部语义详尽的规则词典, 是共享的知识分类系统(通用本体), 其拥有很多下位关系节点的义原比叶子节点义原所含的信息量少. 根据公式(3), 给出计算义原 p 的信息量计算公式:

$$ic_{hn}(p)=1-\frac{\log(hypo(p)+1)}{\log(\max_{hn})} \tag{5}$$

函数 hypo(p)返回给定义原的子节点数量, max_{hn} 是义原所存在的分类系统的总数量, 由于 HowNet 2000 版包含 1621 个义原, 本文中取 max_{hn}=1621.

那么, 根据经典信息量相似度算法, 两义原间的语义相似度为:

$$Sim_{hn}(P_1,P_2)=\max_{p\in(p_1,p_2)} ic_{hn}(p) \tag{6}$$

由于 HowNet 描述的词语一般有多个义项(多义词), 每个义项又由多个义原描述, 因此可以假设义项 n_1 有 n 个义原: $N_1=\{p_{11}, p_{12}, \ldots, p_{1n}\}$, 义项 n_2 有 m 个义原: $N_2=\{p_{21}, p_{22}, \ldots, p_{2m}\}$, 那么义项 n_1 和 n_2 间的相似度为:

$$Sim_{hn}(n_1,n_2)=Sim_L(N_1,N_2)\bullet\frac{\min(C_1,C_2)}{\sqrt{C_1\bullet C_2}} \tag{7}$$

式(7)中, $sim_L(N_1, N_2)$是刘群在文献[10]中计算两个义项集合的相似度算法, 即集合的相似度等于其元素对的相似度的算术平均. C_1 和 C_2 分别表示义项 n_1 和义项 n_2 的记录数目, 用于修正 $sim_L(N_1, N_2)$在计算两个极其相似义项的误差.

对于词语的相似度, 在计算两个词语的义项相似度后, 假设词语 w_1 有 k 个义项: $\{n_{11}, n_{12}, \ldots, n_{1k}\}$, 词语 w_2 有 p 个义项: $\{n_{21}, n_{22}, \ldots, n_{2p}\}$, 那么词语 w_1 和词语 w_2 间的语义相似度为:

$$Sim_{hn}(w_1,w_2)=\frac{\sum_{i=1}^{k}\sum_{j=1}^{p}Sim_{hn}(n_{1i},n_{2j})}{k\bullet p} \tag{8}$$

至此, 对于 HowNet 用义原、义项描述的中英文词汇, 通过公式(5)至公式(8)可以计算出两词语间的语义相似度. 公式(8)在计算词语对的相似度时考虑了一词多义现象, 由于取词语各义项的平均值, 所以计算出的结果将更接近 HowNet 对于词语的客观描述, 能区分相似度极大时的情况, 下一节的实验数据对此进行了论证.

3 相似度实验结果及分析

3.1 实验数据集的选取

为了评估词语语义相似度计算算法的质量, 可以将计算结果与大量统计的人工判断结构比较[1]. 如果计算结果越接近人工判断, 那么该算法就越精确.

本文实验所采用的数据集选自文献[11]数据集中的 28 对词语. 表 1 显示了近年来语义相似度实验所采用的数据集:

表 1 近年来的语义相似度实验数据集[11]

实验	年份	词语对数目	参与者数目
R&G	1965	65	51(all native speakers)
M&C	1991	30	38(all native speakers)
Resnik	1995	30	10(all native speakers)
P&S	2008	65	101(76 native, 25 non)

由于 P&S 数据集是 G. Pirró 和 N. Seco 通过互联网从著名的计算机科学组织(如 DBWORLD, CORPORA, LINGUIST 等)尽可能广范围采集的数据, 同改进前的统计数据相比更具权威性, 因此本文将从 P&S 数据集中选取最常用的 28 对词语作为实验数据.

3.2 实验结果及分析

如表 2 所示的部分词语语义相似度, 分别是根据本文提出的算法用公式(8)计算的 sim_{hn}, 刘群在文献[10]的算法计算的 simL 和 P&S 数据集中 native speakers 的人工判断的结果.

表 2 词语相似度的计算结果

词语对	$P\&S_{nat}$	sim_{hn}	sim_L
gem-jewel	3.296	0.517	1.000
midday-noon	3.270	0.816	1.000
automobile-car	3.544	0.816	1.000
boy-lad	3.103	0.750	1.000
implement-tool	2.654	0.541	1.000
coast-shore	3.016	1.000	1.000
journey-voyage	3.028	0.260	0.040
magician-wizard	2.857	0.507	0.722
furnace-stove	2.386	0.588	1.000
asylum-madhouse	2.699	0.528	0.896

brother-monk	1.997	0.528	0.800
food-fruit	1.794	0.465	0.444
bird-cock	1.714	0.576	1.000
bird-crane	1.722	0.565	1.000
…	…	…	…

P&S 数据集的每对词语的相似度从 0(无相似性)到 4(极其相似)取值，为了有一个直观的比较，本文将 $P\&S_{nat}$ 数据集的统计数据进行归一化(除以 4)处理，使其在[0, 1]间取值. $P\&S_{nat}$ 数据集的人工判断统计值、sim_{hn} 和 sim_L 的对比折线图如图 2 和图 3 所示.

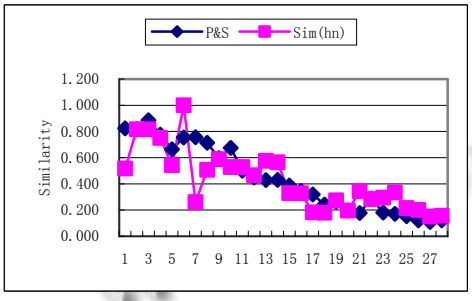

图 2　sim_{hn} 与 $P\&S_{nat}$ 数据集对比折线图

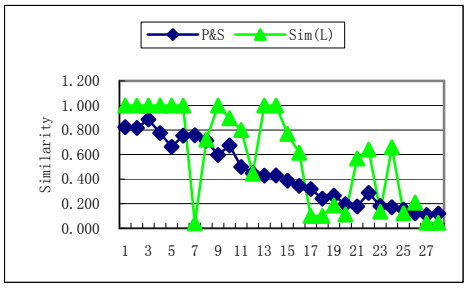

图 3　sim_L 与 $P\&S_{nat}$ 数据集对比折线图

sim_L 是采用刘群的算法计算的结果，由于该算法提出用义项的最大相似度作为词语的相似度，从图 3 可以看出该算法无法区分相似度极大的情况.

图 2 中本文算法计算的相似度 simhn 的折线图更接近于 P&S 的人工判断值. 对于少数词语对，如“gem-jewel”，“journey-voyage”和“coast-shore”，sim_L 和 sim_{hn} 相比于 P&S 均有较大误差，但基于义原距离的算法计算的 sim_L 的误差更大. 这些误差来源于 HowNet 本身的知识描述方式，而且实验说明基于义原距离的误差大于基于信息量计算的误差.

对于中文词语的语义相似度计算，本文选取文献[10]实验一(基于语义距离的算法)的词语对作为实验数据集，计算的结果如表 3 所示.

表 3　中文词语对实验结果

词语对	方法 1	sim_L	sim_{hn}
男人-父亲	1.000	1.000	0.766
男人-女人	1.000	0.861	0.727
男人-母亲	1.000	0.861	0.724
男人-和尚	1.000	0.861	0.662
男人-经理	1.000	0.630	0.394
男人-鲤鱼	0.347	0.209	0.374
男人-苹果	0.285	0.171	0.342
男人-工作	0.186	0.112	0.279
男人-收音机	0.186	0.112	0.222
男人-责任	0.016	0.126	0.177
男人-高兴	0.016	0.048	0.150

在表 3 中，方法 1 参考文献[10]，即仅使用 HowNet 语义表达式中第一基本义原来计算词语相似度； sim_L 为刘群在文献[10]中提出的算法计算的结果，将义原集合相似度最大值作为词语相似度. 本文提出的信息量算法的计算结果 sim_{hn} 与这两种基于语义距离算法的计算结果折线图如图 4 所示.

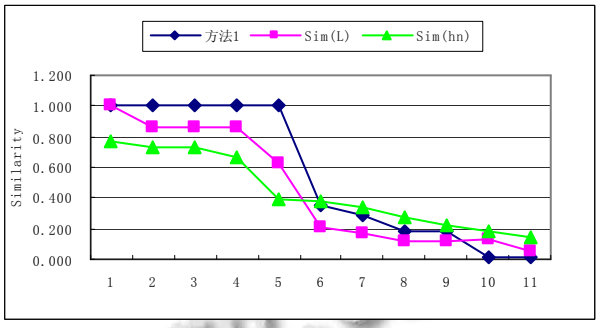

图 4　sim_{hn} 与基于语义距离算法计算结果对比折线图

方法 1 由于将第一基本义原的相似度作为词语相似度，因此对于接近一半的测试词语计算结果均为 1.000，因为前四个词语的第一基本义原均为“human|人”，这种方法明显过于粗略，但计算最为简单，可以应用于相似或不相似两种结果的判断.

sim_L 计算值考虑了每个义项的义原集合的相似度，并且将义原集合相似度最大值作为词语相似度，第一个词语对计算结果为 1.000，因为这对词均有义项“human|人,family|家,male|男”，即取“男人”的“丈夫”的义项，而“男人”的另一表示“男性”的义项“human|人,male|男”将被算法忽略. 除此之外，该算法能够较好区分其他词语的相似度.

本文算法计算的结果 sim_{hn} 由于考虑了每个词语的所有义项，因此没有相似度为 1.000 的词语对，除非

一个词语和它本身的相似度为 1.000. 算法的整个折线图趋势较为平缓, 在相似度小于 0.5 时的折线图也位于另两种算法的上方, 说明对于低相似度的词语对也能较好地区分. 这种相似度精确的计算可应用于舆情分析等处理短文本的情形.

4 结语

本文提出了基于 HowNet 的信息量计算语义相似度的方法, 虽然不同于传统的基于 WordNet 和语料库的思想, 在没有大型语料库的词语统计信息的支持下, 仍然能仅依赖于 HowNet 的义原描述方式计算信息量及其词语间的相似度.

P&S 数据集的实验证明了本文的算法计算的结果很接近人工判断的经验值, 比基于义原距离的算法的误差小, 同时也表现出了 HowNet 对于英文词语的丰富语义描述能力. 在中文词语对的实验中, 本文的算法也表现出了对于极其相似词语对的区分能力.

参考文献

1 Pirro G. A Semantic Similarity Metric Combining Features and Intrinsic Information Content, Data & Knowledge Engineering 68, 2009:1289–1308.

2 Xiao B, Xue LM, Zhao Y. Sememe Description Based Estimating for Semantic Orientation of Chinese Vocabulary, in Proceedings 0f The 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 2010: 671–674.

3 Li Y, Bandar A, McLean D. An approach for measuring semantic similarity between words using multiple information sources, IEEE Trans. on Knowledge and Data Engineering, 2003,15(4):871–882.

4 Resnik P. Information content to evaluate semantic similarity in a taxonomy. Proceedings of IJCAI, 1995,448–453.

5 Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 1999,11:95–130.

6 Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet. Proc. of ECAI. 2004. 1089–1090.

7 Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. Proc. of the International Conference on Research in Computational Linguistics. 1998.

8 Lin D, An information-theoretic definition of similarity, in Proc. of the 15th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, 1998. 296–304.

9 HowNet. HowNet's Home Page. http://www.keenage.com.

10 刘群,李素建.基于《知网》的词汇语义相似度计算.第三届汉语词汇语义学研讨会.台北,2002,5.

11 Pirró G, Seco N. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. ODBASE, 2008. 1271 –1288.

(上接第 86 页)

for Flexible Indoor Environment. Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, Harbin, China, July 26–30, 2011, 1050 –1053.

4 俱莹,刘开华,史伟光,等.基于 RFID 的边界虚拟参考标签定位算法.计算机工程,2011,37(6):274–276.

5 Ni LM, Liu YH, Yiu CL, et al. LAND MARC: Indoor Location Sensing Using Active RFID. Proc. of the First IEEE International Conference on Pervasive Computing and Communications. Dallas, Texas, USA, March, 2003. 407– 415.

6 李兴鹤,胡咏梅,宋吉波,等.基于 LANDMARC 系统的室内定位仿真研究.计算机工程与应用,2008,44(27):209–212.

7 Patwari N, O'Dea RJ. Relative location in wireless networks. Vehicular Technology Conference, Rhodes, Greece, May 6-9, 2001. 1149–1153.

8 Xie YG. On RFID Positioning Base on LANDMARC and Improved Algorithm. Proc. of the 29th Chinese Control Conference. Beijing, China, July 29–31,2010,4831 –4836.

9 王静,张会清.基于信号强度的室内定位技术的研究.计算机测量与控制,2009,17(12):2500–2503.

10 Wing WYNG. Efficiency Of Applying Virtual Reference Tag to Neural Network Based RFID Indoor Positioning Method. Proc. of the 2011 International Conference on Machine Learning and Cybernetics. Guilin, China, July 10-13, 2011. 447–453.