软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(4):776−785 [doi: 10.3724/SP.J.1001.2012.04116]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于词语距离的网络图词义消歧*

杨陟卓[1,2+], 黄河燕[1,2]

[1](北京市海量语言信息处理与云计算应用工程技术研究中心(北京理工大学),北京 100081)

[2](北京理工大学 计算机学院,北京 100081)

Graph Based Word Sense Disambiguation Method Using Distance Between Words

YANG Zhi-Zhuo[1,2+], HUANG He-Yan[1,2]

[1](Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing (Beijing Institute of Technology), Beijing 100081, China)

[2](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

+ Corresponding author: E-mail: 10907029@bit.edu.cn, http://cs.bit.edu.cn

Yang ZZ, Huang HY. Graph based word sense disambiguation method using distance between words. *Journal of Software*, 2012,23(4):776−785. http://www.jos.org.cn/1000-9825/4116.htm

Abstract: Almost all existing knowledge-based word sense disambiguation (WSD) methods used exploit context information contain, in certain window size around ambiguous word, are ineffective because all words in the window size have the same impact on determining the sense of ambiguous word. In order to solve the problem, this paper proposes a novel WSD model based on distance between words, which is built on the basics of traditional graph WSD model and can make full use of distance information. Through model reconstruction, optimization, parameter estimation and evaluation of comparison, the study demonstrates the feature of the new model: The words nearby ambiguous word will have more impact to the final sense of ambiguous word while the words far away from it will have less. Experimental results show that the proposed model can improve Chinese WSD performance, compared with the best evaluation results of SemEval-2007: task #5, this model gets MacroAve (macro-average accuracy) increase 3.1%.

Key words: word distance; Markov chain; graph based model; PageRank; parameter estimation

摘 要: 传统的基于知识库的词义消歧方法,以一定窗口大小下的词语作为背景,对歧义词词义进行推断.该窗口大小下的所有词语无论距离远近,都对歧义词的词义具有相同的影响,使词义消歧效果不佳.针对此问题,提出了一种基于词语距离的网络图词义消歧模型.该模型在传统的网络图词义消歧模型的基础上,充分考虑了词语距离对消歧效果的影响.通过模型重构、优化改进、参数估计以及评测比较,论证了该模型的特点:距离歧义词较近的词语,会对其词义有较强的推荐作用;而距离较远的词,会对其词义有较弱的推荐作用.实验结果表明,该模型可以有效提高中文词义消歧性能,与 SemEval-2007:task #5 最好的成绩相比,该方法在 MacroAve(macro-average accuracy)上提高了 3.1%.

关键词: 词语距离;马尔可夫链;网络图模型;PageRank;参数估计

* 基金项目: 国家自然科学基金(61132009); 国防基础基金; 北京理工大学科技创新计划重大项目培育专项计划
收稿时间: 2011-03-18; 定稿时间: 2011-09-02

中图法分类号: TP391 文献标识码: A

词义消歧是指确定多义词在自然语言特定的上下文中的意义,是自然语言处理中的一个核心问题[1],在机器翻译、信息检索、文本分析、自动文摘和知识挖掘研究中均具有十分重要的作用.

词义消歧方法可分为有监督、无监督和基于知识库的方法[2].有监督词义消歧法利用已标注词义的语料库,提取特定词义的特征属性,并且通过机器学习方法生成分类器或分类规则,对新实例进行词义判定;无监督消歧是从原始的数据文集中获取词义的相关特征,对新实例进行词义判定;严格的无监督消歧是不使用任何外部资源的.相关研究表明[3],有监督方法的消歧效果明显优于无监督方法.但是,有监督方法需要大量的人工标注语料,存在数据稀疏问题.无监督方法不需要人工标注语料,可以有效解决知识获取瓶颈问题,但其消歧效果不尽如人意.

随着知识库系统应用范围的扩大,基于知识库的词义消歧方法逐渐流行起来.该方法通过外部知识源和词汇的特定上下文推导出多义词的意义,可以有效解决数据稀疏问题.目前,基于知识库的词义消歧法主要包括利用词语重叠进行词义消歧的方法[4]、基于选择限制的词义消歧方法[5]、基于互联网知识的词义消歧方法[6,7]、基于结构化知识的词义消歧方法[8]和基于网络图的词义消歧方法等[9,10].其中,基于网络图的词义消歧方法近年来在国际词义消歧评测任务中表现出良好的消歧性能.该方法既是基于知识库的方法,也是一种无监督的方法,这类方法的代表性方法主要有以下两种:

一种是 2005 年 Navigli 提出的基于词汇链的方法,即 SSI(structural semantic interconnections)方法[11].该方法首先通过语义词典构造待消歧句子的词语链,词汇链中的词汇均为在字典中语义距离最小的词汇.然后,选择“基于网络图连接度最大的词义”作为最终歧义词的词义.该方法在国际评测 Senseval-3 无监督全词消歧中取得了最好成绩[12].

另一种是逐步确定网络图中词汇词义的方法.2005 年,Mihalcea 提出使用 PageRank 算法进行消歧的方法[13].该方法首先将消歧句子中的同义词作为网络图的节点,网络图中边的权重采用同义词在字典中的语义距离,然后在网络图中运行迭代算法,根据各个词义节点的重要度大小确定歧义词的词义.随后,Mihalcea,Agirre 和 Navigli 又使用不同的语义距离计算方法进行词义消歧;同时,他们又考察了不同网络图算法对消歧效果的影响[14−17].实验结果表明,PageRank 算法相比其他网络图算法在英文词义消歧任务中具有较好的表现.2009 年,Agirre 提出使用 Personalizing PageRank 模型进行词义消歧[18].这种方法和 Mihalcea 提出的方法原理相同,只是对重要度计算公式作了改进,修改了某些词语在算法迭代时传递的重要度.

基于知识库的词义消歧方法,在进行词义消歧时往往忽略了词语之间的距离信息.同样,基于网络图的消歧方法利用词语的语义距离描绘它们之间关系的强弱,使得语义关系较强的词语在算法迭代时相似度相互传递并有所加强.这些方法都假设,一定窗口大小下的每个词对歧义词的词义具有相同的影响.本文认为,这个假设过于牵强.实际上,还应当考虑词语在消歧句中的实际距离.本文提出一种基于词语距离的网络图词义消歧方法.这种方法在传递词语相似度时不仅考虑词语间语义关系的强弱,而且还考虑它们在歧义句中的实际距离,即:距离歧义词较近的词语会对与其词义有较强的影响,而距离较远的词对其词义有较弱的影响.那么,如何将词语在消歧句中的实际距离关系引入(关联)到网络图是一个难点,也是本文解决词义消歧问题的关键.

本文首先介绍基于词语距离的词义消歧方法,重点介绍消歧技术路线以及本文构建的两个词义消歧模型.随后通过实验,对模型参数估计以及测评进行比较.最后给出研究结论并提出下一步工作计划.

1 基于词语距离的网络图词义消歧方法

1.1 基于网络图的词义消歧原则

PageRank 算法[19]是最早被搜索引擎用来计算 Web 网页重要度的方法,目前已成功应用在许多任务中,如对象检索[20]、文本抽取[21]、自动文摘[22]等.该算法提出了一个假设:网络图中节点的重要性和质量,可通过其他与

它相连节点的质量和数量来衡量.也就是说,若一个节点被越多或者质量越好的节点所指向,那么这个节点在整个网络中的重要度就越高.

Google 搜索引擎使用 PageRank 算法确定 Web 网页的重要度.本文提出的模型也是利用 PageRank 算法,确定各个词义节点在句子中的重要度.如果一个词义节点 A 与其他词义节点相连,那么可以认为节点 A 被其他节点推荐.在网络图中,推荐 A 节点的节点越多,或者推荐 A 节点的节点重要度越高,则该节点在网络图中的重要度就越高,可以认为 A 节点与上下文节点联系越紧密.在词义消歧问题中,将句子中歧义词的每个词义分别看作不同的节点,词义消歧的目的是区分各个词义节点的重要度.在算法迭代过程中,各个词义节点的重要度将逐渐区分开来,最终将重要度最大的词义节点作为歧义词的词义.由于重要度最大的词义节点实际上就是与上下文节点相似度最大的节点,因此在以下的讨论中,本文使用节点相似度指代节点重要度.

1.2 基于词语距离的网络图消歧路线

构建网络图时,本文利用句中的实词构建网络图.组成网络图的节点包括待消歧句子中的各个词语和歧义词的各个词义.连接网络图中词语之间的关系可以是语义关系,也可以是共现关系.语句“程序设计表彰大会顺利召开”,该语句中的“程序”和“设计”都是歧义词.“程序”的词义有两个,分别是“软件”和“次序”;“设计”的词义有两个,分别是名词“规划”和动词“计划”;该语句网络图的所有词语节点和节点之间的关系如图 1 所示,其中,椭圆表示词义节点,方框表示上下文词语节点,六边形表示歧义词节点,实线表示词语间的语义或共现关系,虚线表示歧义词与词义节点的关系.由于消歧的目的是区分同一歧义词的各个词义相似度大小,因此,同一歧义词的各个词义间的联系并不加入到网络图中,保证同一歧义词词义间的相似度不会相互传递,防止相似度相互加强.

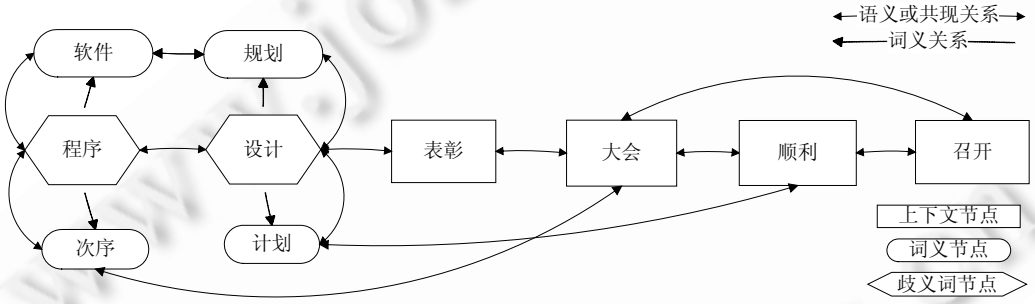


Fig.1 Graph for ambiguous sentence

图 1 歧义句子所构建的网络图

消歧模型采用 PageRank 算法,迭代计算待消歧句中各个节点的相似度,节点的相似度通过语义或共现关系相互传递,最终,网络图各个节点的相似度收敛于一个定值,不再发生变化.从图 1 所示的网络图可以看出,上下文节点“大会”与 3 个上下文节点“表彰”、“顺利”、“召开”和一个词义节点“次序”相连接.根据 PageRank 算法原理,经过算法迭代后,最终,“大会”节点在句子中的相似度会较大.因此,与其相连的“次序”节点就会收到较多的相似度分数.同理,“顺利”节点也会将较多的相似度分数传递给“计划”节点.由于传统的网络图并不考虑词语之间的距离关系,无论在一定窗口大小内的两个词语相距多远,都会传递相同的相似度,那么在传统的网络图词义消歧模型中,“次序”和“计划”很有可能成为歧义词“程序”和“设计”的最终词义.

但是,如果将词语距离的信息引入到网络图模型中,增强短距离词语之间的推荐作用,弱化长距离词语的推荐作用,就可以使节点的相似度分数更合理地在网络图中传递.在本文提出的改进网络图模型中,由于“大会”和“顺利”节点距离“次序”和“计划”节点较远,因此只会将较少的相似度分数传递给连接的两个节点;而“程序”与“软件”、“设计”与“规划”、还有两个词义节点“软件”和“规划”的距离较近,算法迭代时会将较多的分数传递给对方,使得两个词义节点“软件”和“规划”的相似度在网络图中逐渐增大,最终在词义节点中获得最大的分数,成

为歧义词的词义.从上面的例子中还可以看出,基于网络图的模型进行全词消歧时是有优势的,因为不仅上下文节点的相似度可以传递给词义节点,而且词义节点的相似度也可以相互传递.例如,“软件”和“规划”的相似度也可以相互加强.

1.3 基于词语距离的网络图消歧模型

基于词语距离的网络图消歧模型修改了原始 PageRank 算法的计算公式,本文使用马尔可夫链和随机行走过程[23,24]解释基于词语距离的网络图消歧模型工作原理.本文使用 $G=(V,E)$表示消歧网络图,V 表示上下文节点和歧义词词义节点,E 表示词语节点之间的语义或者共现关系.在马尔可夫模型中,将网络图 G 看作马尔可夫链,网络图中的每个节点或者每一个词义表示一个状态,词义节点之间的关系表示从一个节点到另一个节点的状态转移,随机行走者可以根据节点之间的关系跳转到各个节点.假设与当前词义节点 v_j 相连的词义节点个数是 $Out(v_j)$,如果随机行走者以相同的概率行走到与当前词义节点相连的其他词义节点,那么在马尔可夫链中,从该词义节点转移到其他词义节点的概率就是 $1/Out(v_j)$.综合考虑网络图中所有的节点,可以将马尔可夫链的转移概率矩阵用 $\boldsymbol{A}$ 表示,矩阵中的每个元素为

$$A_{ij}=\begin{cases}1/Out(v_j), & \text{当}(i,j)\in E\text{时}\\ 0, & \text{其他}\end{cases} \tag{1}$$

并且满足条件:

$$\sum_{j=1}^{n}A_{ij}=1 \tag{2}$$

也就是说,从某个词义节点出发,转移到其他节点的概率之和为 1.

根据马尔可夫链的各态经历理论[24,25],如果矩阵 $\boldsymbol{A}$ 既是不可约的又是非周期的,那么随机转移矩阵 $\boldsymbol{A}$ 定义的有限马尔可夫链具有唯一的静态概率分布.转移矩阵具有唯一的静态概率分布意味着在经过若干步的迭代后,无论网络图中每个词义节点的初始概率是多少,随机浏览者最后到达每个词义节点的概率是收敛的.使用 $\boldsymbol{P}$ 代表静态概率分布,有如下公式:

$$\boldsymbol{P}=\boldsymbol{A}^T\times\boldsymbol{P} \tag{3}$$

也可以写成

$$p(v_i)=\sum_{j=1}^{n}A_{ji}\times p(v_j) \tag{4}$$

为了将随机转移矩阵转变为不可约和非周期的,可以向网络图中人为地添加从每一词义节点到其他 n 个词义节点的有向边,并且使用参数 d 控制这些节点之间的跳转概率.由此,公式(4)可以写成:

$$p(v_i)=\sum_{j=1}^{n}[(1-d)/n+d\times A_{ji}]\times p(v_j) \tag{5}$$

公式(5)是原始 PageRank 模型计算公式.这里,本文将词义节点的距离信息引入到公式(5)中,借鉴 Time PageRank[26]的思想,本文提出模型 1:

$$p(v_i)=\sum_{j=1}^{n}\{[1-rate(v_j,v_i)]/n+rate(v_j,v_i)\times A_{ji}\}\times p(v_j) \tag{6}$$

公式(6)中使用两个词义节点的距离函数 $rate(v_j,v_i)$代替固定的跳转因子 d,使得与词义节点在歧义句中实际距离较远的两个词义跳转概率较小,而距离较近的两个词语跳转概率较大.当跳转概率较小时,从源节点向目标节点传递的相似度分数就较小;而概率较大时,传递的相似度分数就越大.但是应当注意到,当 v_i 与 v_j 两个节点距离较小时,虽然 $rate(v_j,v_i)$的值在相对增加,而$[1-rate(v_j,v_i)]$的值却在相对减少.这就意味着减少了所有其他 n 个节点向 v_i 节点传递的相似度,最终导致 v_i 节点的相似度不一定增大.因此,模型 1 存在一定的缺陷.针对模型 1 的缺点,本文对 PageRank 公式作了新的修改,提出模型 2:

$$p(v_i)=\sum_{j=1}^{n}[(1-d)/n+d\times A_{ji}\times rate(v_j,v_i)]\times p(v_j) \tag{7}$$

可以看出,公式(7)中距离函数仅仅影响与节点 v_i 实际相连词所传递的相似度.模型 2 相比模型 1 更加合理.在加入距离函数 $rate(v_j,v_i)$之后,当两个词义节点的距离较大时,$rate(v_j,v_i)$取较小的值,这样,从 v_j 节点传递到 v_i 节点的相似度就较小;当两个节点的距离较小时,$rate(v_j,v_i)$取较大的值,此时传递的相似度就较大.改进后的相似度计算公式在源词义节点向目标词义节点传递相似度时,不仅考虑了源节点的重要度 $p(v_j)$、与源节点相连的权重 A_{ji},而且还有源节点与目标节点的距离函数 $rate(v_j,v_i)$.其中,$rate(v_j,v_i)$函数值的估计应符合词义消歧任务的需要.由于距离函数经常使用指数函数估计,因此,本文使用指数函数作为两个节点之间的距离函数,即

$$rate(v_j,v_i)=0.5^{distance(v_j,v_i)/\lambda} \tag{8}$$

其中,$distance(v_j,v_i)$表示两个词语之间的距离,它的值即为词语 v_j 和 v_i 之间的词汇数加 1.λ的最佳取值估计在下面的实验中加以说明.

1.4 参数估计

在基于网络图的词义消歧方法中,网络图中词义之间的关系可以使用语义关系,也可以使用共现关系,同时也可以是两种关系的优化组合.本文分别用 *simi* 和 *cooc* 表示语义和共现关系.在两种关系条件下,分别测试了基于词语距离的网络图模型的消歧性能.下面介绍网络图中共现关系 $p(v_j|v_i,cooc)$和语义关系 $p(v_j|v_i,simi)$的计算方法,也就是公式(5)和公式(6)中的随机转移矩阵 $\boldsymbol{A}$ 中的元素.

(1) $p(v_j|v_i,cooc)$的估计.本文使用搜狗实验室提供的中文词语搭配库[27]估计 $p(v_j|v_i,cooc)$.词语搭配关系库来自 SOGOU 搜索引擎索引到的中文互联网语料的统计分析,统计涉及到的互联网语料规模在 1 亿个 Web 页面以上.涉及到的搭配样例超过 2 000 万,涉及到的高频词超过 15 万.如果两个中文词语在词语搭配库中的共现频率较大,那么就为这两个词语分配较高的共现概率.令 $CoocNum(v_j,v_i)$表示两个词语的共现次数,则词语之间共现概率计算公式为

$$p(v_j\mid v_i,cooc)=CoocNum(v_j,v_i)\Big/\sum_{v_k\in Out\_Cooc(v_i)}CoocNum(v_k,v_i),$$

其中,$Out_Cooc(v_i)$表示在歧义句中与 v_i 词语共现的所有节点的集合.

(2) $p(v_j|v_i,simi)$的估计.本文使用 Hownet[28]提供的 API 函数:HowNet_Get_Concept_Similarity 估计语义关系 $p(v_j|v_i,simi)$.该函数可以给出两个概念(义项)之间的相似度,计算方法综合考虑了概念类的相似度、框架的相似度和定义的相似度等.一个词语在 Hownet 中往往具有多个义项,令 $Similarity(x_i,y_j)$表示词语间的词义相似度,并定义为:设 X 词语具有义项$(x_0,x_1,\ldots,x_j)$,Y 词语具有义项$(y_0,y_1,\ldots,y_j)$,$i>0$,$j>0$,则 X 与 Y 的语义相似度为

$$Similarity(X,Y)=\max_{x_i,y_j}[HowNet\_Get\_concept\_similarity(x_i,y_j)].$$

词语之间相似概率计算公式为

$$p(v_j\mid v_i,simi)=Similarity(v_j,v_i)\Big/\sum_{v_k\in Out\_Simi(v_i)}Similarity(v_k,v_i),$$

其中,$Out_Simi(v_i)$表示在歧义句中与 v_i 词语以语义关系类型相联系的所有词语的集合.

(3) 中英文映射.利用国际语义评测的中英文词汇任务,对本文所提出的方法进行评测.在该任务中,中文歧义词词义是用英文标注的,因此在估计不同语言间的共现和语义关系时,需要使用一种映射方法将英文标注映射为中文.然后,用映射好的中文代替英文标注,计算不同语言之间的共现和语义关系.映射过程中选取的中文词义,应当是与句子上下文共现和语义关系较强的词语.

映射的方法为:假设一个英文译文在 Hownet 中有若干个中文义项$(x_0,x_1,\ldots,x_j)$,$i>0$,则选取其中的一个义项 x_i,能够满足:

$$\arg\max_{x_i}\left\{\sum_{y_j\in context\_word}[\gamma\times p(y_j\mid x_i,simi)+(1-\gamma)\times p(y_j\mid x_i,cooc)]\right\},$$

其中,$p(y_j|x_i,simi)$表示义项 x_i 与歧义句上下文词汇 y_j 的相似关联概率,$p(y_j|x_i,cooc)$表示义项 x_i 与歧义句上下文词汇 y_j 的共现关联概率,γ用于平衡两种类型的关联权重.如果侧重语义关系,则γ取较大的值,反之亦然.实验中,γ的

值取 0.5.

2 实验与结果讨论

2.1 测试语料评价标准与基线方法

利用 ACL2007 的一个组成部分 SemEval-2007[29],国际语义评测的中英文词汇任务(task#5 multilingual Chinese English lexical sample task)对本文方法进行评测.该任务共含有 40 个歧义词(所有词在后面的表 3 中详细列出),语料由训练语料(本文方法是完全无指导的方法,没有利用任何训练语料,而是对其测试语料直接进行测试)以及测试语料两部分组成,见表 1.同时,采用其提供的标准评测工具及相应的评价指标 p_{mar}(macro average accuracy),如下面公式所示:

$$p_i = m_i / n_i, p_{mar} = \sum_{i=1}^{N} p_i \Big/ N,$$

其中,N 为所有的目标词数,m_i 是对每一个特定的词标注正确的例句数,n_i 是对该特定词所有的测试例句数.

Table 1 Basics of gold standard dataset

表 1 标注评估语料情况

	歧义词平均词义个数	训练实例个数	测试实例个数
19个名词	2.45	1 019	364
21个动词	3.57	1 667	571

实验采用 5 个 baseline,分别为:

(1) TorMD.该方法为多伦多大学参加 SemEval-2007 评测的无指导方法[30],获得了 SemEval-2007Task#5 评测第 1 名(p_{mar}=43.1%);

(2) PMI.该方法是北京大学的无指导词义消歧方法[7],采用的是用双语词汇 Web 间接关联的完全无指导消歧方法;

(3) BL_MFS.该方法选取测试集答案内的测试实例最常用词义(most frequent sense)的结果,由标准测试集直接给出;

(4) Original_Cooc.该模型采用传统的 PageRank 模型消歧,词语在网络图中的相似度采用公式(5)来计算,并且利用词语之间的共现关系构建网络图;

(5) Original_Simi.该模型同样采用传统的 PageRank 模型消歧,与 Original_Cooc 方法的不同之处在于,利用词语之间的语义关系构建网络图.

实验采用的 4 个改进的网络图模型分别为:

(1) Cooc_model_1.该方法的相似度计算公式采用模型 1,并且利用词语之间的共现关系构建网络图;

(2) Cooc_model_2.该方法的相似度计算公式采用模型 2,并且利用词语之间的共现关系构建网络图;

(3) Simi_model_1.该方法的相似度计算公式采用模型 1,并且利用词语之间的语义关系构建网络图;

(4) Simi_model_2.该方法的相似度计算公式采用模型 2,并且利用词语之间的语义关系构建网络图.

2.2 实验结果

(1) 4 种方法实验结果.在实验中,利用句中所有的实词构建异构关系网络图,因此网络图中的词语一般不会超过 20 个.这样,一般算法迭代次数在 100 次以内,节点相似度分数就不再发生变化.4 种方法的实验结果见表 2.从表中可以看出,基于词语距离的网络图消歧方法取得了不错的效果,歧义词的平均消歧准确率超过 TorMD 系统 3.1%、PMI 系统 1.8%.但是,所有的方法均没有超过 BL_MFS 方法,说明无监督词义消歧方法的性能还有很大的提升空间.

为了更加客观地考察各种方法对各个歧义词的消歧准确率,将 MFS,TorMD,PMI 和基于词语距离网络图模型 2(Cooc_Model_2)这 4 种方法对 40 个词的消歧精度结果整理在表 3 中.在表 3 列举的各种方法的消歧结果中,

左侧为 19 个名词的结果,右侧为 21 个动词的结果,最后一行是 Cooc_Model_2 方法对 TorMD 方法以及 PMI 方法名词和动词性能分别提升的百分比.表中粗体表示该词消歧性能的最好结果.

Table 2 Experimental results of 4 methods

表 2 4 种方法实验结果

	TorMD	PMI	Cooc_Model_2	MFS
平均准确率(p_{mar})	0.431	0.444	0.462	0.481
提高百分比(%)	3.1	1.8	0	−1.9

Table 3 Detail nouns|verbs results of 4 methods (p_{mar})

表 3 各种方法名词|动词实验结果(p_{mar})

名词	词义数	TorMD	PMI	MFS	Cooc_Model_2	动词	词义数	TorMD	PMI	MFS	Cooc_Model_2
本	3	**0.720**	0.560	0.400	0.600	补	3	**0.550**	0.400	0.500	0.300
表面	2	0.556	0.444	**0.611**	0.278	成立	3	**0.481**	0.296	0.370	0.185
菜	2	0.474	**0.789**	0.579	0.317	吃	4	0.174	0.217	**0.435**	0.174
长城	3	0.429	0.381	0.476	**0.524**	出	9	**0.169**	0.117	0.130	0.142
单位	2	0.706	**0.588**	**0.588**	0.529	带	8	0.119	0.060	**0.150**	0.149
道	3	**0.500**	0.278	**0.500**	0.222	动	4	0.300	0.150	**0.500**	0.250
队伍	3	0.318	**0.591**	0.455	0.455	动摇	2	0.500	**0.625**	**0.625**	0.500
儿女	2	0.500	0.450	0.500	**0.800**	发	5	0.25	0.278	0.278	**0.305**
机组	2	0.643	**0.786**	0.714	0.571	赶	3	0.389	0.278	**0.500**	0.444
镜头	2	0.467	0.667	0.533	**0.733**	叫	4	0.256	0.256	0.256	**0.333**
面	3	0.348	**0.739**	0.435	0.500	进	5	0.250	0.227	0.227	**0.454**
牌子	2	0.353	0.353	0.353	**0.588**	开通	2	0.500	0.500	0.500	**0.550**
旗帜	3	0.500	0.444	**0.556**	0.333	看	4	0.294	**0.500**	0.294	0.323
气息	2	**0.857**	**0.857**	0.714	0.714	平息	2	0.375	0.625	0.500	**0.750**
气象	2	0.438	0.438	0.625	**0.750**	使	2	0.563	**0.625**	**0.625**	0.500
日子	3	0.28l	0.219	0.313	**0.406**	说明	2	0.444	0.444	0.556	**0.833**
天地	3	**0.560**	0.320	0.400	0.520	挑	2	0.143	0.214	0.429	**0.428**
眼光	2	**0.714**	0.357	**0.714**	0.500	推翻	2	0.300	0.600	0.600	**0.700**
中医	2	0.438	**0.750**	0.625	0.588	望	2	0.462	0.538	**0.769**	0.538
						想	4	0.216	**0.432**	0.270	0.216
						震惊	2	**0.714**	0.357	**0.714**	0.500
平均准确率(p_{mar})		0.516	0.527	**0.528**	0.522			0.335	0.369	**0.440**	0.408
提高百分比(%)		0.6	−0.5	−0.6	0			7.3	3.9	−3.2	0

从表 3 的实验数据可以看出,本文提出的方法虽然在名词消歧效果上表现平平,但在动词消歧效果上却远远超过 TorMD 和 PMI.分别超出 7.3%和 3.9%,并且有 8 个动词的消歧准确率超过了 MFS 方法.由于动词的词义数目比名词的词义数要多,而且动词使用灵活,因此可以看出,基于词语距离的网络图消歧方法对动词消歧效果要优于其他方法,具有更好的消歧稳定性.

为了横向比较本文提出的两个模型在不同相似度计算方法下的消歧效果,我们做了如下一组实验,实验结果见表 4.从表中可以看出,在使用相同词语关系构建网络图时,消歧性能最好的模型是 Model_2,其次是原始的 PageRank 模型,最差的是 Model_1.实验结果表明,本文所提出的 Model_2 确实能够提高网络图词义消歧的性能.但是根据 Time PageRank 思想改进的 Model_1,由于距离函数影响词语范围太大的原因,不适合词义消歧.

Table 4 Experimental results of 6 graph based disambiguation methods

表 4 6 种基于网络图消歧方法实验结果

	Original_Simi	Simi_Model_1	Simi_Model_2	Original_Cooc	Cooc_Model_1	Cooc_Model_2
平均准确率(p_{mar})	0.446	0.427	0.457	0.444	0.427	0.462
提高百分比(%)	0	−1.9	1.1	0	−1.7	1.8

为了考察不同窗口大小对本文提出的各种方法消歧性能的影响,我们做了下面一组实验,其实验结果如图 2 所示.横坐标表示不同窗口大小,纵坐标表示消歧准确率 p_{mar}.应当注意的是,本文所指的窗口大小是去除句中虚词后所取的窗口,而词语之间的距离是在原始的歧义句中词语之间的距离.因此,即使当窗口大小为 1 时,歧义

词左右两个实词到歧义词的距离都不一定相等.这样,基于词语距离的网络图模型在任何大小窗口条件下都能发挥作用.

从实验结果可以看出:

(1) 在所有窗口大小的条件下,无论是在共现关系,还是在相似关系网络图中,Model_2 的消歧效果总是比原始的 PageRank 模型的消歧效果要好,说明 Model_2 比原始的 PageRank 模型更适合词义消歧;

(2) 大多数基于网络图的方法在窗口为 1 时具有最佳的消歧性能,表明在基于网络图的词义消歧方法中,歧义词附近的两个实词对歧义词词义影响是最大的.随着窗口的扩大,不仅会带来对消歧有利的信息,同时也会带来一定的噪声.继续增加窗口大小,各种模型的消歧性能略有下降并最终趋于稳定;

(3) 当使用相同的消歧模型时,在绝大多数情况下,利用词语间共现关系的模型比利用语义关系的模型的消歧准确率要高.这是由于词语的共现关系从真实文本出发,描述并刻画词语之间的相关度,它相比语义字典中的关系具有更强的领域适应性和针对性,因此具有更好的消歧能力.

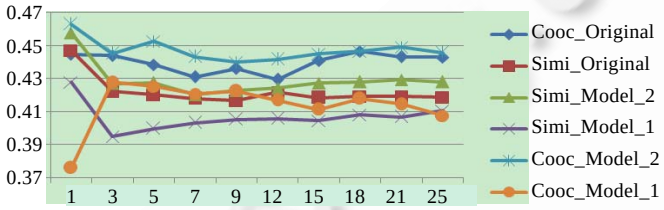


Fig.2 Comparison of several models with different context window size

图 2 不同窗口大小对各种模型消歧性能的影响

最后考察了距离函数(8)中,λ的取值对本文所采用的 4 种模型的影响.实验中,窗口大小选择图 2 中消歧性能最好时的窗口大小.实验结果如图 3 所示,其中,横坐标表示λ的值,纵坐标表示消歧准确率 p_{mar}.在 Model_2 中,当λ取值为 2 时获得了最佳性能.在 Model_1 中,当窗口大小取值为 10 时获得了最佳性能.表明在两个模型取最佳λ值时,模型 1 随着词语距离的增加,由源节点向目标节点传递的相似度减少得较快;而在模型 2 中,随着词语距离的增加,传递的相似度减少得较慢.

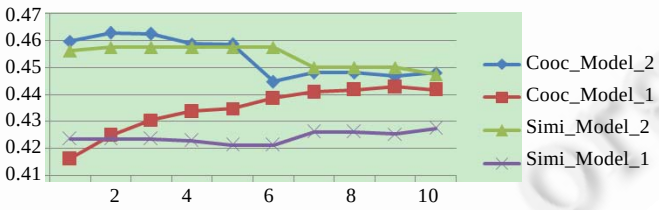


Fig.3 Comparison of several models for different value of λ

图 3 λ取值对消歧效果的影响

3 结束语

基于网络图词义消歧方法,将词义消歧问题转换成在网络图上节点的排序问题.词义消歧不再是通过一条简单的规则实现,而是在算法迭代过程中将歧义词的词义逐渐区分开来.本文在传统的网络模型中引入了词语距离信息,提出了基于词语距离的网络图词义消歧模型.在新的词义消歧模型中,距离歧义词较远的词对歧义词的词义有较大的影响,而距离较近的词对它的词义有较弱的影响.在 SemEval-2007 上的测试结果表明,该模型不但可以提高传统的基于网络图的词义消歧性能,而且消歧效果优于参加该项评测的最好系统.

本文下一步的改进工作可以从两个方面进行:第一,深入分析大规模中文词义消歧数据集的特征,挖掘更多可用的中文词义消歧知识,进一步提高模型的消歧性能;第二,在同一中文数据集中比较其他网络图算法,如 Degree,HIT 等的消歧性能,分析各种网络图消歧模型的适用范围.

References:

[1] Chan YS, Ng HT. Scaling up word sense disambiguation via parallel texts. In: Howe A, ed. Proc. of the 20th National Conf. on Artificial Intelligence, Pittsburgh: Association for the Advancement of Artificial Intelligence, 2005. 1037−1042.

[2] Navigli R. Word sense disambiguation: A survey. ACM Computing Surveys, 2009,41(2):1−69. [doi: 10.1145/1459352.1459355]

[3] McCarthy D, Koeling R, Weeds J, Carroll J. Finding predominant word senses in untagged text. In: Scott D, ed. Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 279−286. [doi: 10.3115/1218955.1218991]

[4] Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pinecone from an ice cream cone. In: Burger S, ed. Proc. of the 5th Special Interest Group on Document. Morristown: Association for Computational Linguistics, 1998. 24−26. [doi: 10.1145/318723.318728]

[5] Agirre E, Martinez D. Learning class-to-class selectional preferences. In: Traum DR, ed. Proc. of the 5th Conf. on Computational Natural Language Learning. Morristown: Association for Computational Linguistics, 2001. 15−22. [doi: 10.3115/1117822.1117825]

[6] Liu PY, Zhao TJ. Unsupervised translation disambiguation by using semantic dictionary and mining language model from Web. Journal of Software, 2009,20(5):1292−1300 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3367.htm [doi: 10.3724/SP.J. 1001.2009.03367]

[7] Liu PY, Zhao TJ. Unsupervised translation disambiguation based on web indirect association of bilingual word. Journal of Software, 2010,21(4):575−585 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3574.htm [doi: 10.3724/SP.J.1001.2010.03574]

[8] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Williams J, ed. Proc. of the 14th Int'l Joint Conf. on Artificial Intelligence. Montreal: Association for Artificial Intelligence, 1995. 448−453.

[9] Mihalcea R, Tarau P, Figa E. Pagerank on semantic networks, with application to word sense disambiguation. In: Geffet M, ed. Proc. of the 20th Int'l Conf. on Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 1126−1132. [doi: 10.3115/1220355.1220517]

[10] Véronis J. Hyperlex: Lexical cartography for information retrieval. Computer Speech & Language, 2004,18(3):223−252. [doi: 10.1016/j.csl.2004.05.002]

[11] Navigli R, Velardi P. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(7):1075−1086. [doi: 10.1109/TPAMI.2005.149]

[12] Mihalcea R, Chklovski T, Kilgarriff A. The Senseval-3 English lexical sample task. In: Scott D, ed. Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 25−28.

[13] Mihalcea R. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In: Mooney R, ed. Proc. of the Joint Conf. on Human Language Technology/Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2005. 411−418. [doi: 10.3115/1220575.1220627]

[14] Sinha R, Mihalcea R. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: Brew C, ed. Proc. of the IEEE Int'l Conf. on Semantic Computing. Morristown: Association for Computational Linguistics, 2007. 215−222. [doi: 10.1109/ICSC.2007.87]

[15] Agirre E, Soroa A. Using the multilingual central repository for graph-based word sense disambiguation. In: Korbayova K, ed. Proc. of the Int'l Conf. on Language Resources and Evaluation. Marrakesh: Language Resources Association, 2008. 27−34.

[16] Navigli R, Lapata M. Graph connectivity measures for unsupervised word sense disambiguation. In: Veloso M, ed. Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence. Hyderabad: Association for Artificial Intelligence, 2007. 46−53.

[17] Navigli R, Lapata M. An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(4):678−692. [doi: 10.1109/TPAMI.2009.36]

[18] Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation. In: Lascarides A, ed. Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2009. 33−41. [doi: 10.3115/1609067.1609070]

[19] Brin S, Page L. The anatomy of a large-scale hyper textual Web search engine. In: Nivre J, ed. Proc. of the 7th Conf. on World Wide Web. Sydney: Association for World Wide Web, 1998. 107−117. [doi: 10.1016/S0169-7552(98)00110-X]

[20] Liu XM, Bollen J, Nelson ML, van de Sompel H. Co-Authorship networks in the digital library research community. Information Processing and Management, 2005,41(6):681−682. [doi: 10.1016/j.ipm.2005.03.012]

[21] Mihalcea R, Tarau P. Textrank: Bringing order into texts. In: Lin D, ed. Proc. of the Conf. on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2004. 404−411.

[22] Wan XJ, Yang JW. Improved affinity graph based multi-document summarization. In: Gunopulos D, ed. Proc. of the Human Language Technology Conf. on North America Chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2006. 181−184.

[23] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Vol.66. Technologies Project, San Francisco: Stanford InfoLab., 1998. 281−287.

[24] Liu B. Web Data Mining. Heidelberg: Springer-Verlag, 2007. 245−254.

[25] Aiello W, Chung F, Lu LY. A random graph model for massive graphs. In: Giannotti F, ed. Proc. of the ACM Symp. on Theory of Computing. Pittsburgh: Association for the Advancement of Artificial Intelligence, 2000. 171−180. http://bigcheese.math.sc.edu/~lu/papers/random.pdf [doi: 10.1145/335305.335326]

[26] Li X, Liu B, Yu P. Time sensitive ranking with application to publication search. In: Giannotti F, ed. Proc. of the 8th IEEE Int'l Conf. on Data Mining. Pisa: IEEE Computer Society Technical Committee on Intelligent Informatics, 2008. 893−898. [doi: 10.1109/ICDM.2008.155]

[27] http://www.sogou.com/labs/dl/r.html

[28] Dong ZD, Dong Q. Hownet. 2000. http://keenage.com

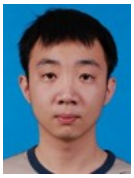
[29] Jin P, Wu YF, Yu SW. SemEval-2007 task 5: Multilingual Chinese-English lexical sample task. In: Agirre E, ed. Proc. of the 4th Int'l Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Prague: Association for Computational Linguistics, 2007. 19−23.

[30] Mohammad S, Hirst G, Resnik P. TOR.TORMD: Distributional profiles of concepts for unsupervised word sense disambiguation. In: Agirre E, ed. Proc. of the 4th Int'l Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Prague: Association for Computational Linguistics, 2007. 326−333.

附中文参考文献:

[6] 刘鹏远,赵铁军.利用语义词典 Web 挖掘语言模型的无指导译文消歧.软件学报,2009,20(5):1292−1300. http://www.jos.org.cn/1000-9825/3367.htm [doi: 10.3724/SP.J.1001.2009.03367]

[7] 刘鹏远,赵铁军.基于双语词汇 Web 间接关联的无指导译文消歧.软件学报,2010,21(4):575−585. http://www.jos.org.cn/1000-9825/3574.htm [doi: 10.3724/SP.J.1001.2010.03574]



杨陟卓(1983−),男,山西临汾人,博士生,主要研究领域为自然语言处理,词义消歧.

黄河燕(1963−),女,博士,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.