基于向量空间模型中义项词语的无导词义消歧*

鲁 松, 白 硕, 黄 雄

(中国科学院 计算技术研究所,北京 100080)
E-mail: songl@public3.bta.net.cn; bai@ict.ac.cn
http://www.ict.ac.cn

摘要: 有导词义消歧机器学习方法的引入虽然使词义消歧取得了长足的进步,但由于需要大量人力进行词义标注,使其难以适用于大规模词义消歧任务.针对这一问题,提出了一种避免人工词义标注巨大工作量的无导学习方法.在仅需义项词语知识库的支持下,将待消歧多义词与义项词语映射到向量空间中,基于 k-NN(k=1)方法,计算二者相似度来实现词义消歧任务.在对 10 个典型多义词进行词义消歧的测试实验中,采用该方法取得了平均正确率为 83.13%的消歧结果.

关 键 词: 词义消歧;无导方法;义项词语;上下文位置权重计算;向量空间模型
中图法分类号: TP391 文献标识码: A

多义词的词义消歧是为了解决自然语言中同形异义词语在不同上下文环境中的义项标注问题.该问题普遍存在于各种自然语言之中.在汉语词典中,多义词约占汉语词语集合的 14.8%,但在汉语语料中,多义词出现频率约占语料总词次的 42%[1].可见,多义词在自然语言中尽管数量不多,但出现频率却极高.

同时,多义词分布的普遍性决定了多义词词义消歧任务必然成为多种应用问题的关注焦点之一,诸如机器翻译、信息检索、自然语言内容语义分析、语法分析、语音识别和文语转换[2].据统计,在信息检索(information retrieval)中引入部分多义词消歧技术以后,可使其整个系统的正确率由 29%提高到 34.2%,取得较明显的改善[3].可见,只要涉及自然语言的计算机应用,多义词的词义消歧工作就是不可回避的基础问题.

从方法论角度来讲,许多计算语言学问题都可以被形式化为一个分类问题(classification).同样,词义消歧问题也是一个典型的分类问题,即一个多义词在一定的上下文环境中的义项被有限个义项类别进行归属.在早期手工规则方式效果不佳的情况下,各种机器学习的分类方法被应用于词义消歧任务中,如决策树[4]、决策表[5]、Naive-Bayes[6]、神经网络[7]、Exemplar-Based Learning[8]、最大熵方法[1]等.与手工提取规则比较,尽管这些有导的机器学习方法在多义词消歧问题中取得了较好的效果,但同手工规则一样,有导分类方法难以实现大规模多义词词义知识的学习和消歧任务.其原因只有一个,即为了获得较好的学习效果和避免数据稀疏问题,必须对训练语料中的多义词进行大量代价高昂的人工义项标注工作.因此,很难实现大规模的多义词消歧工作.

由此,无导的多义词词义消歧方法开始引起关注.其中典型的方法有双语对齐方法[9]、机器可读词典方法[10]和向量空间中的词义识别方法[11].在这些方法中,双语对齐语料的获取本身就是一个有待解决的问题;机器可读词典方法面临大量难以克服的噪音问题;而词空间(word space)中词义聚类方法的学习过程复杂,时间开销过大,且其侧重点在多义词的词义识别上.因此,上述方法针对大规模的多义词词义消歧都有其局限性.

本文也提出了一种基于向量空间模型的无导词义消歧学习方法.向量空间模型(vector space model)来源于

* 收稿日期: 2000-08-01; 修改日期: 2001-03-26
基金项目: 国家自然科学基金资助项目(69773008);国家 863 高科技发展计划资助项目(863-306-2D02-01-3);国家重点基础研究发展规划 973 资助项目(G1998030510)

作者简介: 鲁松(1972－),男,北京人,博士,主要研究领域为计算语言学,信息检索,机器学习;白硕(1956－),男,辽宁沈阳人,博士,研究员,博士生导师,主要研究领域为计算语言学,人工智能,网络信息处理;黄雄(1969－),男,江苏苏州人,博士,主要研究领域为信息检索,计算复杂性理论.

信息检索领域[12];Schutze[13,14]将其用于解决词义消歧问题,提出一种基于聚类的无导方法;文献[1]借鉴Schutze[13,14]的方法,也提出了结合机器可读词典《同义词词林》的无导方法.

本文提出的方法是基于这样一种假设:在上下文环境分布上,多义词某一义项与指示该义项的义项词语(sense words)所具有的相似性,比指示其他义项的义项词语具有更强的相似性.以此为基础来完成多义词词义的消歧工作.

与已有工作相比,本文提出的方法其主要特点体现在以下几个方面:

(1) 方法的无监督性:消歧工作无须昂贵的人工标注工作,但需要构造为数不多的义项词语知识库;

(2) 方法的简洁性和高效性:由于人为指定多义词义项词语的数量有限,避免了解决高维向量空间中停用词删除、特征提取、特征选择和聚类带来的诸多问题,因此具有很高的学习效率和运行效率.

(3) 对词的形式化表示方法的改进:通过词矩阵的概念和计算上下文中词语在刻画该词语时的重要性,即计算词语权重,实现了词语在向量空间中的精确定位.与仅依靠词语共现频率的形式化方法相比,更客观,也更具说明性.这在本文的实验结果中已经得到了证明.

(4) 采用信息增益方法量化确定上下文有效范围,为词语向量化提供了依据.

(5) 采用 k-NN(k=1)方法,取相似性最近的义项词语来完成标注,避免了采用语义类别向量计算所带来的误差.

本文第 1 节介绍我们提出的无导多义词词义消歧方法,其中第 1.1 节简单介绍经典向量空间模型,第 1.2 节对本文提出的方法进行详细的描述.第 2 节详细描述实验过程、训练和测试数据情况及实验结果.第 3 节对本文提出的方法的优、缺点进行全面的总结和讨论.

1 基于向量空间模型的词义消歧无导学习方法

与词空间[11]的词语知识向量化表示方法一样,本文提出的方法也是以向量空间为基础的.但不同的是,本文将多义词的义项词语映射到实数域向量空间中,通过计算多义词特定上下文与义项词语向量的相似度来对其进行标注.此外,本文将多义词的词义消歧任务形式化为一个与信息检索中自然语言查询(query)和答案文档(document)之间相似度计算方法完全相同的过程.在这一过程中,本文将多义词特定上下文视为信息检索中的查询;而多义词义项词语的词矩阵被视为信息检索中的答案文档.

1.1 向量空间模型

传统的信息检索是指用户给出一个查询(query)以后,在知识库中搜索能回答这个查询的答案文档(document),并通过一定测评机制对答案文档(document)与查询(query)进行相关性计算.其中在诸多模型中,由于向量空间模型(vector space model)[12]具有较强的可计算性和可操作性,已经被广泛地应用于文本检索、自动文摘、关键词自动提取、文本分类和搜索引擎等信息检索领域的各项应用中,并且取得了较好的效果.

本文的词义消歧思想和方法是在借鉴这一模型框架的基础上实现和完成的.

1.1.1 文档的形式化表示方法

在向量空间模型中,文档被形式化为 n 维空间中的向量,空间的一维是倒排表(inverted index)中的一个词语,形式如下:

$$D=\langle w_{\text{term\_1}},w_{\text{term\_2}},w_{\text{term\_3}},\ldots,w_{\text{term\_}n}\rangle.$$

该向量中每一分量表示该词语在此文档中的权重,用以刻画该词语在描述此文档内容时所起作用的相对重要程度.

词语权重计算惟一的准则就是要最大限度地区分不同文档.其中最为典型并被广泛使用的文档词语权重计算方法为 tf.idf[15],如公式(1)所示:

$$w_{ik}=\frac{tf_{ik}\log\left(N/n_k+0.01\right)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2\left[\log(N/n_k+0.01)\right]^2}}. \tag{1}$$

公式(1)中的 w_{ik} 为〈文档 $_i$〉中〈词语 $_k$〉的权重;tf_{ik} 是〈词语 $_k$〉在〈文档 $_i$〉中出现的频率;$\log(N/n_k+0.01)$是〈词语 $_k$〉

在多义词所有义项词语中分布情况的量化,其中 N 为文档集合中的文档数目,n_k 为出现过〈词语 $_k$〉的文档数目;公式(1)的分母是对各分量进行标准化.

公式(1)的提出是基于这样一种假设:对区别文档最有意义的词语应该是那些在文档中出现频率足够高,但在整个文档集合的其他文档中出现频率足够少的词语.可以看出,向量空间模型的量化基础是词语的出现频率和出现文档频率.

1.1.2 距离计算的方法

查询(query)和文档(document)之间的相似度计算是通过 cosine 距离计算来完成的,见公式(2):

$$sim(Q,D)=\frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t}\left(w_{qj}\right)^2 * \sum_{j=1}^{t}\left(w_{d_{ij}}\right)^2}}. \tag{2}$$

向量空间模型的最大优点在于它在知识表示方法上的巨大优势.在该模型中,文档的内容被形式化为多维空间中的一个点,以向量的形式给出.也正是因为把文档以向量的形式定义到实数域中,才使得模式识别和其他领域中各种成熟的算法和计算方法得以采用,极大地提高了自然语言文档的可计算性和可操作性.同时,在多项实际应用中也已证明了向量空间模型的 tf.idf 文档表示方法的有效性.

本文对这两个完全不同的应用进行一个有趣的类比和映射,使向量空间模型 tf.idf 文档表示方法成为本文提出的词义消歧无导方法的一个有利工具.

1.2 基于向量空间模型的词义消歧无导学习方法

1.2.1 词义消歧无导学习的基本框架

本文提出的词义消歧无导学习方法是基于这样一种假设:具有相同或近似上下文分布的词语同样具有相同或近似的语义属性和语义类别.在这一假设的基础上,我们提出义项词语的概念,即与多义词的一个义项具有相同、近似或相关语义范畴的词语.本文用义项词语来定义和描述多义词在某一义项上的语义概念和属性.由此,可将多义词词义消歧的无导学习问题转化为一个对多义词上下文进行分类的问题,即通过计算多义词上下文与多义词不同义项的义项词语之间的相似度来实现多义词不同词义的标注.

引入义项词语的意义在于将一个复杂的、较难控制的无导学习问题转化为处理过程较为简单的分类问题.具有相同或近似的语义属性或者语义类别的义项词语与多义词语义义项之间具有相同或近似的上下文分布,这是问题得以简化和转化的关键.

下面将在第 1.2.2 节~第 1.2.4 节详细介绍义项词语的形式化定义方法,在第 1.2.5 节介绍多义词的形式化定义和消歧方法,在第 1.2.6 节阐明未使用机器可读词典《同义词词林》的原因和人为指定义项词语的优点.

1.2.2 词的上下文表现形式

我们认为,一个词的上下文环境作为一种可用资源,已经为该词语的定义提供了较为充分的语言信息.基于这一认识,本文用词语左、右一定范围内上下文环境的集合来定义这个词,更明确地讲,这个集合就是词的定义.为了形象地描述,我们将它称为词矩阵(word matrix).在词矩阵中,每行是词语的一个上下文环境,每列是词语上下文环境中相对位置上的上下文词语.词矩阵的形式为词语向量表示的进一步形式化提供了一种直接的表示方式.

1.2.3 上下文有效范围大小的确定

词语周围一定范围的上下文可以为词语的定义提供较为充分的语言信息,如何确定上下文范围的大小是本节讨论的主要问题.

我们在这里采用信息论中信息增益的计算方法来实现上下文有效范围的量化确定.其基本思路是:将词矩阵形式化为一个信号系统,从两年的《人民日报》中统计的 5 000 高频词(为获得基于语料库的上下文位置权重统计性平均结果,文中未对 5 000 高频词作任何人为加工和词性上的限制)作为核心词构成的词矩阵基础上,每个词作为一类,计算整个系统的统计不确定性(entropy),即为公式(3)中的 $H(D)$;其次,计算在已知每个相对位置

的前提下整个系统的不确定性,即:条件熵,公式(3)中的$\sum_{v\in Vp}P(v)\times H(D|v)$;二者求差后(公式(3))即为该相对位置对整个系统提供的信息量.这个信息量作为信息增益就是该位置在整个系统中的权重.由此我们可以凭借量化的权重来确定上下文范围选取的大小.计算公式如下:

$$IG_p=H(D)-\sum_{v\in Vp}P(v)\times H(D|v). \tag{3}$$

公式(3)中的 $H(D)$为

$$H(D)=-\sum_{d\in D}P(d)\times\log_2 P(d), \tag{4}$$

其中

$$P(d)=\frac{|\ fre(d)\ |}{\sum_i|\ fre(d_i)\ |}. \tag{5}$$

这里,$\sum_i|fre(d_i)$为 5 000 高频词的在语料中出现的总频率;$|\ fre(d)|$为〈词语 $_d$〉在语料中出现的频率.

设核心词语(focus-word)的上下文形式为

$$\langle wd_{-8},wd_{-7},wd_{-6},wd_{-5},wd_{-4},wd_{-3},wd_{-2},wd_{-1},\ \boldsymbol{focus\text{-}word},wd_{+1},wd_{+2},wd_{+3},wd_{+4},wd_{+5},wd_{+6},wd_{+7},wd_{+8}\rangle,$$

其上下文各个位置权重的计算结果见表 1.

Table 1 Information gain of every position of context

表 1 上下文位置信息增益

Left context①		Right context②	
Position	Information gain	Position	Information gain
wd_{-1}	3.979 875 226 272 497	wd_{+1}	4.005 737 046 263 377
wd_{-2}	2.800 943 865 861 443	wd_{+2}	2.931 834 057 123 61
wd_{-3}	2.183 287 198 587 985	wd_{+3}	2.287 020 518 109 051
wd_{-4}	1.709 504 251 213 968	wd_{+4}	1.810 530 959 811 459
wd_{-5}	1.361 637 860 661 606	wd_{+5}	1.437 952 775 004 01
wd_{-6}	1.074 606 203 376 334	wd_{+6}	1.137 979 746 437 988
wd_{-7}	0.304 606 203 376 334	wd_{+7}	0.821 330 432 383 498 7
wd_{-8}	0.298 992 304 039 87	wd_{+8}	0.419 472 048 769 573 4

①左上下文,②右上下文,③位置,④信息增益.

计算结果反映了这样一个情况:上下文对核心词语的描述能力随着相对位置由近及远而逐渐递减,即通过已知上下文推断空缺词语时,近距离的上下文在推断中所起的作用比远距离的上下文更有价值.

这一计算结果与人们的认知过程基本一致.虽然是近似结果,但在一定程度上具有统计意义.故本文中上下文范围确定为前后 6 个位置的范围大小.

1.2.4 义项词语的形式化表现形式

词矩阵概念的定义使得在词义消歧问题和信息检索两个问题之间可以做个有趣的类比.在这个类比中,将词义消歧里多义词的一个上下文与信息检索中的一个自然语言查询(query)进行对应,而将词义消歧中指示多义词义项的义项词语与信息检索中的答案文档(document)进行对应.其中,义项词语与文档的对应实质上是在义项词语的词矩阵与文档之间进行的.

如此一来,同样可以为义项词语构造一个向量空间模型,向量空间中的每一维是义项词语矩阵中的一个上下文词语,其值是该上下文词语在表示此义项词语时所具有的重要程度,即权重.其中,义项词语矩阵中有多少个上下文词语,向量空间中就有多少维.最终,通过文档内词语权重计算 tf.idf 方法获得义项词语矩阵中每个上下文的权重,进而实现义项词语到向量空间的映射,完成其形式化定义.

上下文词语权重计算公式仍为公式(1),但其中每个变量均被赋予了新的含义:公式(1)中的 w_{ik} 在这里被表示为〈义项词语 $_i$ 的词矩阵〉中〈词语 $_k$〉的权重,tf_{ik} 是〈词语 $_k$〉在〈义项词语 $_i$ 的词矩阵〉中出现的频率;$\log(N/n_k+0.01)$是〈词语 $_k$〉在多义词所有义项词语中分布情况的量化,其中 N 为多义词义项词语数目,n_k 为出现过〈词语 $_k$〉的义项词语数目.

通过义项词语词矩阵中上下文词语的权重计算,义项词语以向量的形式被映射到实数域空间中,为无导的词义消歧提供了一个有效的知识表示方法和计算平台.

1.2.5 多义词的形式化表示及其与义项词语的相似性计算

上下文范围为±6 的待消歧多义词,其形式如下:

$$w_{-6},w_{-5},w_{-4},w_{-3},w_{-2},w_{-1},\ \boldsymbol{polysemous\text{-}word},w_{+1},w_{+2},w_{+3},w_{+4},w_{+5},w_{+6},$$

同样也需映射到向量空间中,其向量表示仍为

$$V_{\text{polysemous-word}}=\langle w_{\text{term\_1}},w_{\text{term\_2}},w_{\text{term\_3}},\ldots,w_{\text{term\_}n}\rangle.$$

其中,向量中每个分量 $w_{\text{term\_}i}$ 为词语 term_i 的权重值.

通过第 1.2.3 节位置权重计算结果,可以将待消歧多义词的上下文映射到向量空间中,其形式见公式(6):

$$W(\text{term\_}i)=\begin{cases}0 & \text{other}\\ weight(pos) & \text{若term\_}i\text{出现在待消歧多义词的上下文中}\end{cases}. \tag{6}$$

进一步可以解释为如果 term_i 在此待消歧多义词上下文中出现,即

$$\text{term\_}i\in\{w_{-6},w_{-5},w_{-4},w_{-3},w_{-2},w_{-1},w_{+1},w_{+2},w_{+3},w_{+4},w_{+5},w_{+6}\},$$

则词语 term_i 的权重值 $w_{\text{term\_}i}=Weight(pos)$;否则,$w_{\text{term\_}i}=0$.其中,$Weight(pos)$是该 term_$i$ 在上下文中所在位置的权重(见表 1),计算过程及其结果见第 1.2.3 节.

向量之间的相似性计算公式仍是典型的 cosine 距离计算方法(公式(2)).

1.2.6 无导词义消歧方法总结

在引入义项词语的概念之后,我们倾向于人工指定少数几个义项词语的方法.多义词的一个义项可以由多个义项词语表明,各义项词语之间关系平等.计算多义词上下文与义项词语相似程度时,无须利用同一义项的不同义项词语向量在义项类中的凝聚点来计算相似度,可以直接采用 k-NN 方法(k=1)计算每个义项词语向量与多义词上下文的相似度,并取距离最近的义项词语来标注该词语在此上下文中的语义义项.

根据多义词义项具有领域差异这一特点,我们对义项词语的选取有两个原则:(1) 语义的相关性,即在相同的描述领域有相同的描述对象;(2) 语义的相似性,即互为同义词或近义词.

文献[1]中采用的机器可读词典《同义词词林》(以下简称为《词林》)来获取语义类向量.我们认为,《词林》由于在 3 方面存在问题会在词义消歧中产生很大误差:

(1) 分类的颗粒度仍然偏大,这使得义类向量的确定不够准确;

(2) 由 6 万多词构成的《词林》面临严重的词量不足问题;

(3) 《词林》是在层次树的语义框架体系上建立起来的,可以体现很好的上下位关系,但不能体现词语间的语义相关性,特别是领域相关性.而针对多义词消歧问题时,多义词和义项词语的上下文语义相关性在其中起到了举足轻重的作用.例如,在本文的实验中对多义词“健康”进行消歧时,添加一个义项词语“医疗”之后,正确率由原来的 67.25%提高到 75.07%,而这种关系是《词林》无法涉及的.

因此,人工指定少数几个可替换的相关或相近的义项词语,可以避免上述问题的出现.本文的实验证明,指定义项词语在实验结果上具有更好的效果,且在多数情况下改进幅度较大.详细评测数据见第 2 节.

2 实验及其结果

2.1 实验过程及数据

实验的整个过程如下:

(1) 从语料库中提取左、右上下文范围为 6 个词的多义词义项词语矩阵;

(2) 分别计算多义词所有义项词语的向量表现形式;

(3) 逐一计算一个多义词的上下文与所有义项词语向量的相似度;

(4) 根据相似度距离最近的义项词语所表示的义项标注多义词在该上下文中的词义.

采用本文提出的思想和方法对 10 个多义词进行无导词义消歧的详细数据和结果见表 2.

实验中的几点说明:

(1) 实验所需的训练数据来自 1996 年和 1997 年两年的《人民日报》语料;

(2) 义项词语上下文提取以句为单位,上下文词语不足 5 个的将被剔除掉;

(3) 多义词义项数量和解释来自《现代汉语词典》1996 年版;

(4) 由于是无导学习方法,所以不存在开放测试和封闭测试的区别.

Table 2 Experimental data and results

表 2 实验数据及结果

Sense①		Examples②	Sense-Words and number of examples③	Number of sense examples④	Accuracy⑤ (%)	
						Average⑥ (%)
材料	s1	建筑~	物资(710)/设备(1658)	438	72.83	83.13
	s2	搜集~	素材(38)/题材(748)	24		
	s3	人事~	文件(1762)/档案(839)	220		
	s4	唱歌的~	人材(2)/人才(2197)	6		
改	s1	~国号	改变(1805)/更改(27)	240	78.71	
	s2	~文章	修改(771)/改善(1840)	55		
	s3	毛病要~	改正(32)/修正(45)/更正(4)	166		
表现	s1	~出疲倦	体现(2033)/显露(86)	488	90.39	
	s2	工作~好	成绩(2233)/言行(146)	143		
	s3	爱~自己	卖弄(4)/招摇(1)/出风头(5)	21		
发表	s1	~谈话	表达(931)/表明(1000)/表示(7624)	1 873	89.11	
	s2	~论文	登载(18)/刊登(485)	305		
健康	s1	恢复~	健壮(17)/强壮(12)/康复(133)/医疗(1532)	934	75.07	
	s2	市场~发展	正常(1365)/正确(2929)	1 663		
造就	s1	~人才	培养(2564)/栽培(266)	389	92.13	
	s2	很有~	造诣(44)/成绩(2233)	3		
保守	s1	~秘密	严守(16)	9	78.97	
	s2	思想~	守旧(6)/迂腐(4)/落后(1007)	67		
放手	s1	~发动群众	大胆(387)/敢于(182)	30	87.5	
	s2	一~,本掉了.	松手(3)/松开(1)/放开(131)	2		
漏洞	s1	财政~	周密(86)/毛病(24)/疏漏(10)	36	86.84	
	s2	房屋~	缝隙(17)/窟窿(12)	2		
保管	s1	粮食~	保藏(2)/管理(11137)/维护(2746)	86	79.78	
	s2	有 2 个~	员工(642)/工人(1448)	2		
	s3	这方法~最好.	肯定(825)/必定(72)	1		

①义项,②举例,③义项词语及其样本数量,④词义样本数量,⑤正确率,⑥平均.

2.2 实验结论

由于现在没有统一的汉语词义消歧测试集,所以只能选用文献[1]的数据结果作为参照.文献[1]中 5 个多义词的无导消歧实验结果数据见表 3(摘自文献[1]第 76 页).文献[1]的多义词义项来源于《词林》,故在“材料”的义项上与《现代汉语词典》有所区别.

Table 3 Results of the experiment in reference [1]

表 3 文献[1]实验结果

Polysemous words①	Sense ID②	Accuracy③ (%)
材料	Dk17/ba06/al03	81.7
改	ih02/hg18/hj66	70.6
表现	Jd06/di20/hj59	68.9
发表	Hc11/hi14/jd03	73.4
健康	ed43/eb37	70.1

①多义词,②歧义类型,③正确率.

通过比较可以看出,本文提出的方法总体上具有更好的正确率且改进幅度较大.分析其原因如下:

(1) 采用信息检索中的 tf.idf 文档词语权重计算方法,与简单的共现频率方式比较,提供了更为精确的词语表示方法,为知识的准确表示和后继计算处理提供了基础.

(2) 义项词语的人工指定在一定程度上避免了机器可读词典组织非面向词义消歧和知识颗粒度较大带来的噪声;同时,义项词语数量不多的特点可以带来学习和应用效率的提高.

(3) 上下文位置权重的计算为上下文向量表示提供了一种更好的量化的方法.

3 结论与讨论

认知语言学家认为,人在进行词语的语义类划分的过程中,上下文的相似性起到了至关重要的作用.由此提出了一个假设:词语间上下文的相似性决定了它们语义的相似性[16].这一假设又被扩展为:词语间语义的相似性反过来也同样决定其上下文的相似性[11].利用这一扩展的假设,Schutze[11]开展了多义词词义自动识别(word sense discrimination)的工作,取得了开创性的结果.但实验对象是受限的,而且算法复杂、效率低等问题限制了它的推广.

本文也同样基于这一假设的扩展,即本文中所定义的义项词语与多义词某一义项之间的语义相似性决定义项词语上下文与多义词在同一义项上的上下文的相似性.

在这一假设的框架下,利用信息检索中的文档词语计算技术将义项词语映射到向量空间中,通过 k-NN 方法(k=1)的相似性计算完成词义消歧的标注任务.

实验证明,基于这样一种假设,采用本文提出的思想和方法是可行的,并且是有正确性保证的.部分多义词的无导消歧结果甚至可以与有导学习的高正确率结果相比较.但一些问题仍然存在,需要进一步解决:

(1) 义项词语的敏感性问题.由于方法是基于义项词语与多义词在某一语义义项上的相似性进行的,故义项词语如何确定是一个关键问题.例如,在对多义词“健康”进行消歧时,添加一个义项词语“医疗”后,正确率提高了近 8%.正确率对义项词语的敏感性由此可见.

(2) 义项词语的自动获取.尽管人为地指定义项词语与多义词标注相比节省了巨大人力,但可以进一步考虑采用针对《现代汉语词典》电子版的多义词义项词语自动获取方法.

(3) “词袋”的缺陷.借用信息检索向量空间模型中 tf.idf 文档权重表示方法可以较为合理地形式化表示词语向量,但同时失去了许多上下文词语语序所提供的语言信息.尽管本文提出的上下文位置权重可以部分平滑这一问题,但仍有不足之处.

(4) 扩大测试对象的范围.多义词消歧方法的系统评价是困难的,这作为一个研究问题已经引起了关注.本文提出的方法其通用性和应用价值需要进一步在大规模测试集和其他语言中进行检验.

致谢 感谢清华大学的李娟子博士对本文的研究工作给予的帮助和支持.

References:

[1] Li, Juan-zi. The research on Chinese word sense disambiguation [Ph.D. Thesis]. Beijing: Tsinghua University, 1999 (in Chinese).

[2] Ide, N., Veronis, J. Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics, 1998,24(1):1~40.

[3] Schutze, H., Pedersen, J. Information retrieval based on word senses. In: Andew, H., Mooery, K., eds. Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas: University of Nevada at Las Vegas, 1995. 161~175.

[4] Black, E. An experiment in computational discrimination of English word senses. IBM Journal of Research and Development, 1988, 32(2):185~194.

[5] Yarowsky, D. Decision lists for Lexical ambiguity resolution: application to accent restoration in Spanish and French. In: Mooney, R., ed. Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics. Las Cruces, NJ: Association for Computational Linguistics, 1994. 88~95. http://www.cs.jhu.edu/~yarowsky/pubs.html.

[6] Mooney, R.J. Comparative experiments on disambiguating word senses: an illustration of the role of bias in machine learning. In: Brill, E., Church, K., eds. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Somerset, NJ: Association for Computational Linguistics, 1996. 82~91.

[7] Kawamoto, A.H. Distributed representations of ambiguous words and their resolution in a connectionist network. In: Small, S., ed. Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence. San Mateo, CA: Morgan Kaufman, 1998. 195~228.

[8] Ng, H.T. Exemplar-Based word sense disambiguation: some recent improvements. In: Johnson, M., Allegrini, P., eds. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing. Providence, Rhode Island, 1997. 208~213. http://www.comp.nus.edu.sg/~nght/publicat.htm.

[9] Dagan, I., Itai, A., Markovitch, S. Two languages are more informative than one. In: Brown, P., Kameyama, M., eds. Proceedings of the 29th Annual Meeting of Association for Computational Linguistics. Berkeley, CA: Association for Computational Linguistics, 1991. 130~137.

[10] Yarowsky, D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Zampolli, A., ed. Computation Linguistic'92. Nantas: Association for Computational Linguistics, 1992. 454~460. http://www.cs.jhu.edu/~yarowsky/pubs.html.

[11] Schutze, H. Automatic word sense discrimination. Computational Linguistics, 1998,24(1):97~124.

[12] Salton, G. Automatic Information Organization and Retrieval. New York: McGraw-Hill Press, 1968.

[13] Schutze, H. Dimensions of meaning. In: Whitelock, P., ed. Proceedings of the Supercomputing'92. Los Alamitos, CA, 1992. 787~796. ftp://parcftp.parc.xerox.com/pub/qca/papers/.

[14] Schutze, H. Word space. In: Stephen, J.H., Cowan, J., Giles, C.L., eds. Advances in Neural Information Processing Systems 5. San Mateo, CA: Morgan Kaufmann, 1993. 895~902.

[15] Salton, G., Buckley, B. Term-Weighting approaches in automatic text retrieval. Information Processing and Management, 1988,24(5):513~523.

[16] Miller, G.A., Charles, W. Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, 1991,6(1):1~28.

附中文参考文献:

[1] 李娟子.汉语词义消歧方法研究[博士学位论文].北京:清华大学,1999.

An Unsuptervised Approach to Word Sense Disambiguation Based on Sense-Words in Vector Space Model*

LU Song, BAI Shuo, HUANG Xiong

(*Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: songl@public3.bta.net.cn; bai@ict.ac.cn

http://www.ict.ac.cn

Abstract: WSD (word sense disambiguation) based on supervised machine learning made a great progress, but it is hard to deal with large-scale WSD because of its 'big' labor cost. An unsupervised WSD method is provided in this paper to solve this problem. Only under the knowledge database of sense-words, this method formulates the sense-words and polysemous words in vector space, and based on k-NN (k=1) it calculates the similarity between them to disambiguate polysemous words. The average accuracy is 83.13% for 10 polysemous words in open test by this method.

Key words: word sense disambiguation; unsupervised approach; sense-word; weight of context position; vector space model

* Received August 1, 2000; accepted March 26, 2001

Supported by the National Natural Science Foundation of China under Grant No.69773008; the National High Technology Development 863 Program of China under Grant No.863-306-2D02-01-3; the National Grand Fundamental Research 973 Program of China under Grant No.G1998030510