基于词向量的无监督词义消歧方法

吕晓伟,章露露

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

摘 要:词义消歧在多个领域有重要应用。基于Lesk及其改进算法是无监督词义消歧研究的典型代表,但现有算法多基于上下文与义项词覆盖,通常未考虑上下文中词与歧义词的距离影响。为此提出一种基于词向量的词义消歧方法,利用向量表示上下文以及义项,并考虑融合上下文与义项的语义相似度及义项分布频率进行词义消歧。在Senseval-3数据集上测试,结果表明,该方法能有效实现词义消歧。

关键词:词义消歧;词向量;自然语言处理;机器翻译;Word2vec

DOI:10.11907/rjdk.181100

中图分类号:TP391 **文献标识码**:A **文章编号**:1672-7800(2018)009-0193-03

Unsupervised Word Disambiguation Method Based on Word Embeddings

LV Xiao-wei,ZHANG Lu-lu

(*Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming* 650500,*China*)

Abstract:Word sense disambiguation have important applications in many fields. Lesk algorithm and its improved algorithm are typical representatives of unsupervised word-sense disambiguation. However,most of the existing algorithms are mostly based on word coverage of context and gloss. In addition,the effect of distance between ambiguous words and word in context is not considered. This paper proposes a method of word-sense disambiguation based on word vectors,which uses vectors to represent contexts and gloss and also considers combined semantic similarity between context and gloss with the distribution frequency of gloss. The test results on the Senseval-3 dataset show that this method can effectively achieve word-sense disambiguation.

Key Words:word sense disambiguation; word embedding; natural language processing;machine translation;Word2vec

0 引言

在自然语言中存在大量一词多义现象,这些词被称为歧义词。词义消歧指在给定的语境中识别歧义词的正确含义[1]。词义消歧是自然语言处理领域的基础研究,也是核心研究,在机器翻译、语音识别、文本分类、信息检索等方面有着十分重要的作用。

目前,解决词义消歧任务主要有4种方法[2]:①基于知识的方法,②基于语料库的无监督方法,③基于语料库的有监督方法,④组合以上方法的方法。基于知识的方法主要使用丰富且系统的语义知识库进行消歧,例如《知网》[3]、WordNet[4]等;基于语料库的有监督方法使用经过标注的语料库进行消歧。有监督的方法消歧效果较好,但这种方法需要人工标注语料库,现实中大量人工标注的语料库往往难以获取,故多数特定场合难以采用此方式。

基于语料库的无监督方法使用未经标注的语料库进行消歧,典型代表为Lesk算法[5]。该算法利用机读词典,将歧义词的每个义项与上下文中词的每个义项进行匹配,单词覆盖的最多义项为该歧义词上下文中的正确含义。

Lesk算法虽能进行词义消歧任务,但存在两个问题[7]:①计算单词覆盖度的次数与概念中的单词数量有关,单词数量越多,计算次数越多;②词汇覆盖只是基于义项中词汇的共现。针对第一个问题,有研究者提出简化版的Lesk算法[8]:将歧义词的各个义项分别与歧义词所在的上下文计算单词的覆盖度,覆盖度最大的义项为最佳含义。针对第二个问题,有研究者[9]提出根据语义相关,使用WordNet作为语义网络,扩充歧义词各个义项,以增加覆盖度。王永生[10]以WordNet为基础,使用CBC算法扩充目标词的相似词集合进行词义消歧。Pierpaolo Basile

收稿日期:2018-02-08

作者简介:吕晓伟(1989-),女,昆明理工大学信息工程与自动化学院硕士研究生,研究方向为数据挖掘、词义消歧;章露露(1992-),女,昆明理工大学信息工程与自动化学院硕士研究生,研究方向为数据库、信息检索。本文通讯作者:吕晓伟。

等[11]考虑扩展后的义项频率等信息，并在分布语义空间中计算相似度以消歧。基于改进的 Lesk 算法，通过不同方式扩展同义词、义项，再与上下文计算相似度进行消歧。

随着 Word2vec、Glove 的提出以及普及，大量研究者使用词向量[6]完成自然语言处理中的许多任务，词义消歧任务也不例外。

词向量是使用一个向量表示一个词。目前，有两种词向量表达方式：① one-hot representation 方式；②Distributed representation[12]。one-hot representation 方式表示的词向量非常简单，向量的长度为辞典大小，向量中的每一维由 0 或 1 表示，词在辞典中对应的维为 1，其它为 0。这种方式虽然可简单表示一个词，但不能有效表达词之间的词义信息，而且存在数稀疏问题。Distributed representation 这种方式能很好地克服 one-hot representation 方式的两个缺点。该方法将一个词映射到一个实数向量空间中，一般为 100～300 维，这种方法使得词义之间的相似性可以用空间距离表示，两个词向量的空间距离越近，表示两个词的相似性越高。

Google 公司 2013 年开放了 Word2vec[16]这一可以训练词向量的工具。Word2vec 以大量文本训练语料作为输入，通过训练模型快速有效地将一个词语表达成向量形式。该工具包含 CBOW 和 Skip_gram 两种训练模型。CBOW 模型通过上下文预测当前词，Skip_gram 模型通过当前词预测其上下文。Word2vec 开放后，有研究者使用 Word2vec 训练所得的词向量进行词义消歧实验。杨安等[13]考虑义项与上下文相似度分数、领域相关性分数、WordNet 相似度分数以及义项频度分数 4 种因素进行消歧。Kaveh Taghipour[14]等结合 IMS 系统，加入词向量进行消歧。

上述方法考虑了扩展注释、相似词集、参考领域信息、利用语义网络等因素，但是未考虑上下文中词与歧义词的距离对消歧的影响。本文使用文档向量表示歧义词所在的上下文，使用义项向量表示歧义词的各个义项，进行词义消歧。同时考虑到义项频度对消歧的准确度影响，最终实现无监督词义消歧方法。通过在 Senseval-3 数据集上测试，表明本文方法能有效实现词义消歧。

1 基于词向量的词义消歧方法

1.1 方法描述

本文提出的词义消歧方法主要思想是，使用向量分别表示歧义词的各个义项及歧义词所在的上下文，分别计算向量表示的上下文与歧义词的各个义项之间的相似度，再计算歧义词各个义项的分布频率，结合相似度以及义项频度，选择出歧义词的最佳含义。消歧步骤如下：①数据预处理；②上下文以及义项的向量表示；③上下文－义项相似度计算；④义项分布；⑤最终词义选择。如图 1 所示。

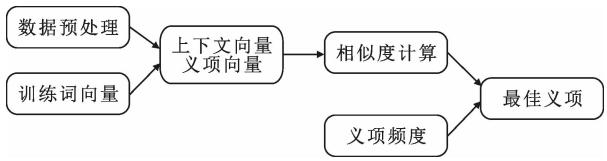

图 1 消歧方法步骤

在数据预处理步骤中，本文只进行去标点、分词、大写转换小写操作，得到歧义词的上下文，后续分别描述上下文及义项的向量表示、上下文－义项相似度计算、义项分布以及最终词义选择。

1.2 上下文及义项向量表示

直观来看，若上下文中的词距离歧义词越近，对歧义词正确含义的判定影响就越大。为体现上下文中词与歧义词距离的影响，本文采用局部加权法计算上下文中词相对歧义词的位置权重。如公式(1)所示。

$$w_i = e^{\left(-\frac{|i-t|}{\tau^2}\right)}, x_i \in |C| \tag{1}$$

式(1)中，wi 表示上下文中第 i 个词位置权重，xi 表示上下文中第 i 个词位置，t 表示歧义词位置，|C| 表示上下文大小，τ^2 是调节参数，表示上下文中的词相对歧义词位置的权重关系，距离关键词距离越近，权重越大。

在定义位置权重基础上，上下文向量计算公式如下：

$$c_i = \sum_{i}^{n} e_{1i} \cdot w_i, w_i \in [0, |C|] \tag{2}$$

式(2)中，c_i 表示第 i 篇上下文的向量表示，w_i 表示上下文中第 i 个词的位置权重，e1 表示上下文向量，e_{1i} 表示上下文词集中第 i 个词的词向量，|C| 表示歧义词所在上下文的大小。

各个义项的向量计算如下：

$$g_i = \sum_{i \in a}^{n} e_{2i} \tag{3}$$

式(3)中，g_i 表示歧义词第 i 个义项的向量表示，a 表示义项中的词，e2 表示义项向量，e_{2i} 表示义项中的第 i 个词的词向量。

1.3 上下文－义项相似度计算

本文使用余弦相似度判断上下文与各个义项的相似度。公式(4)为余弦相似度计算公式。

$$\cos(c, g_i) = \frac{c \cdot g_i}{\| c \| \cdot \| g_i \|} \tag{4}$$

式(4)，cos(c，g_i)表示上下文向量与歧义词第 i 个义项的余弦相似度，c 表示上下文向量，gi 表示第 i 个义项的义项向量。

1.4 义项分布频率

义项分布频率指歧义词的各个义项在包含该歧义词的文档中出现的概率。根据式(5)计算各个义项分布概率：

$$P_i = \frac{n_i}{N} \tag{5}$$

式(5)中，N 表示包含该歧义词的上下文数量，n_i 表示在上下文中歧义词的含义是第 i 个义项的上下文数目。

1.5 最终词义选择

最佳义项选择采用评分方式，对上下文和义项的相似

度以及义项频度综合考虑。根据公示(6)计算每一个义项得分，最高得分的义项为歧义词在该上下文的最佳含义。

$$score_i = a \cdot \cos(c, g_i) + b \cdot P_i \quad (6)$$

式(6)中，a、b 是参数，本文方法中 a=b=0.5。

2　实验

本文使用维基百科数据集，采用 Word2vec 训练词向量，使用 CBOW 模型，窗口大小为 5，词向量维度为 300。

本文使用 Senseval-3 数据集，该数据集包含 57 个歧义词，其中动词 32 个，名词 20 个，形容词 5 个。训练集包含 7 860 篇文档，测试集包含 3 944 篇文档，每个词平均义项为 6.473 个，义项分布频率在 Senseval-3 数据集中得到。使用本文方法在 Senseval-3 测试集上测试，并与基于改进的 Lesk 算法[15](L_1)及文献[10]中的方法(L_2)进行对比，本文方法使用 L_3 表示，结果见表 1。

表 1　实验结果对比

算法	动词	名词	形容词	全部词
L_1	——	——	——	0.525
L_2	——	0.452	——	——
L_3	0.578	0.551	0.482	0.558

使用本文方法全部词的平均准确率达到 0.558，高于文献[15]中改进的 Lesk 算法准确度，也高于文献[10]中没有使用义项频度只计算名词消歧的准确度，表明本文考虑上下文中词与歧义词的距离及融合义项频度方法有效。

3　结语

语义消歧在机器翻译、语音识别、文本分类、信息检索等方面有着十分重要的作用。考虑歧义词周围词语对歧义词正确含义判定的影响，以及歧义词各个义项在数据集中出现的概率，使用词向量进行消歧，消歧效果优于改进的 Lesk 算法。歧义词的有些义项在数据集中并不存在，消歧准确率还有进一步提升空间。下一步拟研究更准确的歧义词义项概率及用更准确的方法表示上下文以及歧义词义项方法。

参考文献：

[1] NAVIGLI R. Word sense disambiguation：asurvey[J]. ACM Computing Surveys，2009，42(2)：1-69.

[2] AGIRRE E，EDMONNDS P. Word sense disambiguation[J]. Algorithm and Application，2007(10)：1-28.

[3] 董振东，董强. 知网和汉语研究[J]. 当代语言学，2001，3(1)：33-44.

[4] FELLBAUM C. WordNet：An electronic lexical database[M]. Cambridge：MIT press，1998.

[5] LESK M. Automatic sense disambiguation using machine readable dictionaries：how to tell a pine cone from an ice cream cone[C]. Proceedings of the 5th Annual International Conference on Systems Documentation，1986：24-26.

[6] 蒋振超，李丽双，黄德根，等. 基于词语关系的词向量模型[J]. 中文信息学报，2017，31(3)：25-31.

[7] BASILE P，CAPUTO A，SEMERARO G. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model[C]. International Conference on Roceedings of Coling，2014.

[8] KILGARRIFF A，ROSENZWEING J. Framework and Results for English SENSEVAL[J]. Computers and the Humanities，2000，34(1-2)：15-48.

[9] BANERJEE S，PEDERSEN T. An adapted Lesk algorithm for word sense disambiguation using WordNet[J]. Computational Linguistics and Intelligent Text Processing，2002(2276) 136-145.

[10] 王永生. 基于改进的 Lesk 算法的词义排歧算法[J]. 微型机与应用，2013 (24)：69-71.

[11] BASILE P，CAPUTO A，SEMERARO G. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model[C]. International Conference on Roceedings of Coling，2014.

[12] HINTON G E. Learning distributed representation of concepts.[C]. Proceedings of CogSci，1986：1-12.

[13] 杨安，李素建，李芸. 基于领域知识和词向量的词义消歧方法[J]. 北京大学学报：自然科学版，2017，53 (2)：204-210.

[14] TAGHIPOUR K，NG H T. Semi-supervised word sense disambiguation using word embeddings in general and specific domains[J]. The 2015 Annual Conference of the North American Chapter of the ACL，2015(5)：314-323.

[15] BASILE P，CAPUTO A，SEMERARO G. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model[C]. International Conference on Roceedings of Coling，2014.

[16] 周练. Word2vec 的工作原理及应用探究[J]. 图书情报导刊，2015 (2)：145-148.

（责任编辑：杜能钢）