一种基于双通道 LDA 模型的汉语词义表示与归纳方法

王少楠[1] 宗成庆[1),2)]

[1)](中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

[2)](中国科学院脑科学与智能技术卓越创新中心 北京 100190)

摘 要 语义记忆是人类理解自然语言的基础.人类理解语言的过程可以看作是对词义进行编码、对语义记忆进行检索,进而对词义进行解码的过程.因此,对词义进行合理地表示是计算机理解语言的关键步骤.该文总结分析了已有的词义表示方法与人脑词义表征的关系,针对汉语词汇的歧义现象,重点阐述了如何从歧义词所处的上下文中最大限度地自动获取关于歧义词的词义信息,并将这些信息整合,通过一系列的特征集合表示歧义词的词义.具体地说,该文将出现在歧义词上下文语境中有明确含义的实词作为模型的输入,同时在上下文中获取可以表示歧义词词义的其他特征,最终将这两种信息通过贝叶斯概率模型整合在一起,共同实现歧义词的词义表示和归纳.实验表明,该文提出的方法可以得到更好的词义表示和归纳效果.

关键词 词义表示;词义归纳;词义消歧;主题模型;双通道主题模型

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2016.01652

A Dual-LDA Method on Chinese Word Sense Representation and Induction

WANG Shao-Nan[1] ZONG Cheng-Qing[1),2)]

[1)](*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Semantic memory is the foundation of human language understanding. Human brain needs to encode, retrieve and decode word meanings for language understanding. The semantic representation is the key step to develop natural language processing systems. Some studies have shown that the formation of concepts is affected by the interaction of human brain and the real world, and the concepts in human brain contain rich forms of information including vision, perception and language. Based on the distributional hypothesis which states that "similar words occur in similar contexts", the concepts are represented as vectors by calculating the co-occurrence frequency of each word and its statistical features. In this way, word representation in computer can be seen as the semantic representation in human brain. This article mainly focuses on how to represent word senses and do word senses induction in natural language text. We first investigate the relation between computational models of word representation and semantic representation in human brain. Based on word similarity experiments, we have verified that word representations by statistical methods can capture the relationship of similarity between words in human brain. In the view of Chinese word sense disambiguation, this paper studies the methods to find the semantic features of ambiguities from context automatically. Bayesian probability model can learn word

收稿日期:2015-06-26;在线出版日期:2016-01-24.本课题得到国家自然科学基金(61333018)和中国科学院战略性先导科技专项(B 类)课题(XDB02070007)资助.**王少楠**,女,1990 年生,博士研究生,主要研究方向为计算语言学和神经语言学.E-mail:wangshaonan2013@ia.ac.cn.**宗成庆**,男,1963 年生,博士,研究员,中国计算机学会(CCF)杰出会员,主要研究领域为机器翻译、文本分类和自然语言处理等.

representations and do word sense induction together. Specifically, in order to do word sense induction, Bayesian probability model clusters words with the same topic. The words within the same topic can be seen as the representation of the topic. In the task of word sense induction, the topics are mapped to word senses in evaluation. Therefore, we use latent Dirichlet allocation model to learn word sense representation from large scale of corpus without annotation. On the basis of word sense representation, we do word sense induction on the testing data. In order to better capture the meaning of ambiguous words, this article builds a Dual Latent Dirichlet Allocation (Dual-LDA) model with two input channels. Specifically, we propose an approach to extract content words that have clear meaning in the context and the words that can distinguish the ambiguous words in the context. Then we combine them as two inputs of Bayesian probability model to represent the word senses and induce the word sense. In the experiment, we use the SogouLab data (sogouCS) as our training corpus and extract 120 thousand sentences which contain the target ambiguous words. The ambiguous words are from word sense induction task in CLP2010, which contain 50 sentences for each ambiguous word and each sentence is annotated with the sense of the ambiguous word. For evaluation, we choose *K*-Means clustering model and latent Dirichlet allocation model as baseline and use the accuracy as evaluation metric. The experimental results show that the proposed Dual-LDA model achieves the best results among other models. This indicates that Dual-LDA model can get better word representations by integrating two different information extracted from context. What is more, the better word representations can improve the performance of word sense induction.

Keywords　semantic representation; word sense induction; word sense disambiguation; latent Dirichlet allocation; dual Dirichlet analysis

1　引　言

随着互联网技术的快速发展，爆炸式的信息增长模式凸显了计算机自动理解和处理自然语言文本的重要性. 目前自然语言处理的计算模型大多采用有监督的机器学习方法，以分类为手段，以获得较高的准确率为目标，无法处理训练样本中没有出现过的词汇和字符. 少数计算模型，如主题模型(topic model)、词向量(Word2Vec)模型等，试图利用大规模无标注数据来对文本进行向量化表示，以解决传统计算模型泛化能力差和特征表示困难的问题，但即使这些模型对文本进行了编码，与真正实现文本语义理解的目标还是有较大的距离. 反观我们人类自身，人脑可以毫不费力地理解一句话的含义，即使在阅读含有歧义词的句子时，人脑甚至都感觉不到歧义的存在. 人脑处理语言的过程十分迅速，但由于其认知理解的生物过程尚不清楚，要想完全模拟人脑处理语言的认知过程和建立计算模型是十分困难的. 幸运的是，认知心理学家已经对人脑处理语言的过程和表现进行了深入的研究，并得出了一些重要的结论，我们可以借鉴这些结论来建立更好的自然语言理解模型.

根据认知心理学家研究的结果，人类理解语言是建立在概念存储基础之上的，只有在记忆中存储了相关的世界概念，才可能理解一句话的含义. 认知心理学家将语言理解看作是一个增量式的动态系统，他们认为，读者增量式地理解句子的含义，即阅读到一个词汇时会自动将其概念含义融入到已经出现过的上文的含义之中[1]. 这样，人脑理解一个句子的过程就可以看作是概念信息的检索和语义整合的过程. 大脑首先从记忆中检索单词对应的语义概念及相关信息，然后通过句子中的语法和上下文信息来整合句子的含义. 所以，要想让计算机理解一个自然语言的句子，首先需要对词义进行合理地表示.

相关研究表明，人类与外部世界的互动影响了概念的形成，因此概念中包含了丰富的信息. 语言文字可以看作是人类表达、传递和交流大脑思维的一套符号系统，人类遣词造句从一定程度上也反映了人类对于外部世界的表达. 概念的形成受到了视觉、

知觉和语言上的影响,本文仅关注如何从语言文字中对概念进行表示和归纳.根据语义的分布式假设,即相似的词出现在相似的上下文中,概念可以表示为统计特征的集合.在不利用知识库的前提下,我们可以从文本中利用词汇共现信息对词义进行某种程度上的表示,并将其看作词汇概念的一个原型.这样,计算机对词义进行表示相当于人脑对单词进行编码.对于新的刺激,人需要检索记忆和提取相关编码才能理解其含义.类似地,计算机需要将新的刺激与已有词义表示进行比较才能对其进行归类.

词义消歧和词义归纳是计算语言学领域研究的两个基本任务.一般将词义消歧看作是有监督的分类问题,将歧义词上下文窗口范围内的多个单词及词性等当作可以区分词义的特征,在训练集中训练特征的权重并利用分类器对测试语料中的歧义词进行区分[2].而词义归纳被看作是无监督的聚类问题,通常利用歧义词上下文的相似程度对包含歧义词的句子进行聚类,从而实现对歧义词的词义归纳.不同于上述两种任务和已有的方法,本文根据上下文对歧义词的不同含义建立不同的向量表示,并利用CLP2010词义归纳评测数据集对词义表示的效果进行检验.

与以往工作相比,本文的最大贡献在于提出了一种基于双通道隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)模型的汉语词义表示与归纳方法,该方法将上下文中获取到的可以表示歧义词词义的信息与歧义词所在上下文中的实词进行整合,同时作为双通道的贝叶斯概率模型的输入,从而获得了更好的词义表示和归纳效果.

2 相关工作与对比

2.1 相关研究

人类如何对概念进行表征是认知科学的基础问题.具身认知(embodied cognition)理论认为,语言中包含的语义是人类与外部世界的交互和体验,语义由大脑的构造、神经的结构、感官和运动系统的活动方式共同决定.具身认知理论的观点极大地推动了认知心理学和认知神经科学研究的发展.与此同时,语义的分布式模型也取得了很多进步,语义的分布式假设认为从语音或者书面文字的统计分布中可以建立语义的计算模型.最近的研究趋向于认为,文本的统计数据和人对外部世界的感知都会影响人脑对于语义的表征[3].另外,一些认知心理学研究通过行为学实验验证了语义的分布式假设的正确性.比如,McDonald 等人[4]设计了两个实验,通过分别操控被试对于"略微熟悉的词汇①"和"非词②"的语意信息③"和"非词④"的语境信息⑤,发现了语境中包含的信息对词汇相似度测试的结果造成了影响.该实验说明了词汇的分布式信息的确会影响人脑对于语义的表征.最近,一些研究通过融合词汇的分布式信息和人的感知信息,来建立词义表示的计算模型.实验证明,这种方法比单独利用上述信息可以在词汇相似度等测量指标上得到更好的效果[5-8].

为了从大规模文本中抽取语义信息,计算语言学家将词义看作是单词在文本中统计分布的结果,建立了很多计算模型.向量空间模型(Vector Space Model,VSM)是表示文本词义重要的模型之一,它将词义看作高维的空间向量,利用词汇的共现信息对词汇含义进行表示.其中,最受关注的一种向量空间模型是潜藏语义分析(Latent Semantic Analysis,LSA)方法,该方法通过奇异值分解(Singular Value Decomposition,SVD)对词项文档矩阵进行分解,使用降维后的矩阵构建潜在语义空间,试图找出词在文档中的潜在语义.但是,LSA 无法解决一词多义的问题,因而其表达词义的能力有限.为了解决 LSA 模型对于词义表示的限制,Hofmann 在 LSA 的基础上提出了一种概率隐含语义模型(probabilistic Latent Semantic Analysis,pLSA),用概率的方式解释了文档的生成过程.之后,Blei 等人在 pLSA 的基础上又提出了基于隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)的主题模型,为 pLSA 模型的概率分布加上了先验知识,利用贝叶斯推断方法求解参数,解决了 pLSA 容易过拟合的问题.

另一种可以对文本词汇进行向量化表示的方法是近年来非常流行的基于神经网络的词向量(Word2Vec)模型.这种方法通过对词汇上下文进行预测来学习词汇的向量化表示.Bengio 等人[9](2003)提出的方法可以在神经网络语言模型(Neural Network Language Model,NNLM)训练的过程中得到词汇的向量化表示.由于利用整个神经网络结构来学习词汇的向量化表示效率很低,Mikolov 等人[10](2013)提出了词向量(Word2Vec)模型,可以

① "略微熟悉的词汇"指可能曾经遇到过,但是不能很快说出它的含义.
② 非词也称假词,指人工构造的没有含义的"词汇".
③ 这里的语境信息即是词汇的上下文信息.
④ 非词也称假词,指人工构造的没有含义的"词汇".
⑤ 这里的语境信息即是词汇的上下文信息.

利用单层的神经网络结构直接学习词向量，他们在神经网络语言模型的基础上进行了大量优化. 由于 Word2Vev 模型可以快速地处理大规模无标注语料，且在词汇相似度比较(word similarity task)、词汇类比(word analogy task)等任务中表现良好. 但是，Word2Vec 模型生成的词向量维度无法解释其物理意义，且在词义归纳任务中不如贝叶斯模型的可扩展性好.

另外，与本文相关的还有以下利用贝叶斯概率模型来完成词义归纳任务的工作.

Brody 和 Lapata 等人[11]利用主题模型对词义进行归纳，在模型学习过程中可以得到歧义词的词义表示. 具体地，他们提出的方法将传统的 LDA 模型的输入变为包含歧义词的句子，这样学习到的文档(句子)主题即为歧义词在句子中的含义. 他们还提出了一种多层的 LDA 模型来解决不同的特征分布问题，分别将不同的特征输入到不同的 LDA 层中，并在文档主题分布处加以融合. 但是，模型仅仅考虑了词袋信息和常用的词义消歧特征(n 元语法、共现信息、词性和依存句法信息)，且多层次的 LDA 模型效果相比单层的 LDA 模型效果没有明显的提升. 另外，为了简化实验过程，他们对不同种类的信息赋予了相同的权重，没有体现对不同特征分布的输入信息分别进行建模的优越性.

上述工作的基本出发点是对歧义词的词义进行聚类，而词义表示只是模型学习过程中的副产物，没有专门探讨词义的表示问题. 我们认为，只有合理地表示词义才是解决词汇理解问题的关键，因此文本主要关注如何从文本中获取更加丰富的词义表示. 我们的实验证明，用统计方法得到的词义表示可以作为人脑中的词义表征的近似表示.

2.2　对比分析

基于统计方法的词义表示研究是根据词义的分布式假设，通过词汇间的共现关系对词义进行表示. 而大量认知心理学的相关研究表明，相似的词在大脑中有相似的词义表征. 那么，基于统计方法的词义表示是否反映了人脑判断的这种相似性呢？下面我们分别计算从认知心理学角度得到的词汇相似度和从统计模型得到的词汇相似度，并分析其相关性，验证统计方法得到的词义表示与大脑的词义表征是否显著相关.

由于认知心理学研究中公开的实验语料大多为英文，因此本节使用英文语料. 实验中采用的 McRae① 数据集[12]是目前心理学研究中使用最广泛的关于(英语)语义特征的数据集. 我们首先利用语义特征产生实验范式得到的词汇属性值计算不同词汇间的余弦相似度，得到 McRae 数据集中词汇间的相似度信息，将其看作人脑对词汇相似度的编码.

为了直观展示 McRae 数据集中的词汇相似度信息，我们选取了 38 个基本单词，根据余弦相似度计算得到的结果绘制了图 1. 这 38 个单词分别属于不同的类别，包括餐具、水果、厨房用品、交通工具、动物、衣物和家具，见图 1 横坐标或纵坐标轴上的词汇. 图中每一个矩形颜色小方块代表对应横坐标上的词汇与纵坐标上的词汇之间的相似度，颜色越深表示其相似度越高. 其中，图中对角线为词汇与其自身的相似度，因此相似度值最高，为 1. 由图 1 可以看出，McRae 数据集中的信息可以很好地捕捉词汇间的相似度信息.

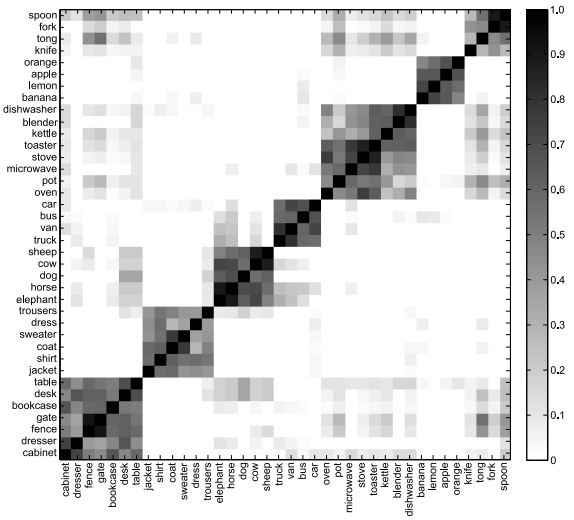


图 1　McRae 数据集词汇相似度图

之后，我们再从统计模型的角度，利用目前广泛使用的 3 种词义表示模型：基于矩阵分解的模型、神经网络模型和贝叶斯概率模型，分别计算这 38 个词的词义表示效果，并对这 3 种模型得到的词义表示效果与人脑的词义表征进行对比分析. 本文选择了 LSA 模型、Word2Vec 模型和 LDA 模型分别作为上述 3 种模型的代表，在 BNC(British National Corpus)②数据集上对词汇进行向量化表示. 下面分别对 LSA 模型和 Word2Vec(Skip-gram)模型的原理进行简要叙述. 关于 LDA 模型的原理介绍见 3.1 节.

① 该数据集由 725 个本科生被试通过行为学实验产生，共有 541 个有生命的(如：狗)和无生命的(如：凳子)基本概念. 附录 3 是数据集中单词“airplane”的样例.

② http://www.natcorp.ox.ac.uk/

在 LSA 模型中，假设有 N 个文档，词典大小为 M，将 M 个单词表示为特征矩阵 $\boldsymbol{X}=[X_1,\cdots,X_M]\in R^{M\times M}$，其中，$X_{ij}$ 为词典中第 i 个单词在文档 j 中出现的频次. 假设有 D 个不相关的隐变量 $U_1,\cdots,U_D$，其中 $U_d\in R^N$ 为单位长度. 由于隐变量之间互不相关，可以得到 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$，其中 $\boldsymbol{I}$ 为单位矩阵. 假设矩阵 $\boldsymbol{X}_j$ 可以表示为隐变量 $U_1,\cdots,U_D$ 的线性组合，即

$$\boldsymbol{X}_j=\sum_{d=1}^{D}a_{dj}U_d+\varepsilon \tag{1}$$

其中，$\boldsymbol{A}=[a_{dj}]\in R^{D\times M}$ 为特征空间到隐空间的映射矩阵，ε 是均值为零的噪声. LSA 模型的目标是求解映射矩阵 $\boldsymbol{A}$. 通过将问题转化为下述优化问题[13]，通过最小化秩为 D 的构造矩阵 $\boldsymbol{UA}$ 与特征矩阵 $\boldsymbol{X}$ 的误差，即可求解映射矩阵 $\boldsymbol{A}$. 如式(2)所示：

$$\min_{U,\boldsymbol{A}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{UA}\| \tag{2}$$

在 Word2Vec 模型中，目前有两种目标函数：一种是利用目标词汇的上下文来预测目标词汇，称为连续的词袋模型(Continuous Bag-Of-Word，CBOW)；另一种是利用目标词汇来预测其上下文的词汇，称为 Skip-gram 模型. 一般认为 Skip-gram 方法在小数据集上的效果优于 CBOW 模型①，因此本文实验采用 Skip-gram 模型. Skip-gram 模型的目标函数是最大化所有单词预测其上下文固定窗口 c 内单词的对数概率，即

$$J(\theta)=\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c,j\ne 0}\log p(w_{t+j}\mid w_t) \tag{3}$$

$$p(w_{t+j}\mid w_t)=\frac{\exp(\boldsymbol{v}'^{\mathrm{T}}_{w_{t+j}}\boldsymbol{v}_{w_t})}{\sum_{w=1}^{W}\exp(\boldsymbol{v}'^{\mathrm{T}}_{w_{t+j}}\boldsymbol{v}_{w_t})} \tag{4}$$

其中：$\boldsymbol{v}'$ 和 $\boldsymbol{v}$ 分别是单词 w 的输入和输出向量；θ 为模型参数.

利用 LSA、Word2Vec 和 LDA 这 3 个模型均可以得到 McRae 数据集所有词汇的向量化表示，然后我们通过余弦相似度计算，可以分别得到 3 种情况下 McRae 数据集中的词汇间相似度值. 然后将这些结果与认知心理学家得到的人脑对这些词汇的相似度进行相关度计算，从而验证统计方法得到的词义表示与大脑的词义表征是否显著相关.

在具体的实验中我们选择的 BNC 数据集中句子的长度范围为 200～300(单词数)，且每个句子中包括至少一个 McRae 中的单词(参照文献[6]). 去除停用词和低频词等预处理后共得到 17 526 个句子，词汇量为 90 459.

表 1 给出的是上述 3 种模型分别得到的词义相似度与认知心理学家得到的词义相似度之间的相关性结果.

表 1 统计模型与人脑对于词汇相似度的相关性

模型	皮尔逊(P 值)	斯皮尔曼(P 值)
LSA	0.360(0.0)	0.162(0.0)
Word2Vec	0.193(0.0)	0.127(0.0)
LDA	0.326(0.0)	0.131(0.0)

表中，“皮尔逊”和“斯皮尔曼”是度量两个变量间相关程度的方法，皮尔逊值和斯皮尔曼值越大，说明两个变量越相关. “P 值”是结果出现的可能性大小. P 值越小，说明效果越显著.

从表 1 可以看出，LSA 模型、Word2Vec 模型和 LDA 模型下得到的 P 值均近似为 0，因此我们可以认为 3 个模型计算出的词义相似度与认知心理学研究得出的词汇相似度均有较大的相似性. 换句话说，统计方法得到的文本表示所包含的词汇相似度信息可以较好地反映认知心理学研究的(人脑)类似的词汇相似度信息. 从表 1 中的结果还可以看出，以 BNC 数据为训练集，LSA 方法和 LDA 方法与人脑对于词汇相似度的相关性高于 Word2Vec，这说明在中等规模训练数据的情况下，LSA 方法和 LDA 方法得到的词义表示能够更好地描述词汇相似度信息.

上述 3 种模型都是利用分布式假设对词义表示进行的建模，那么，这 3 种模型得到的词汇相似度信息是否类似呢？以下对这 3 个模型在上述 38 个基本单词上得到的词汇相似度与人脑反映出的词汇相似度(图 1)分别给出更加直观的比较.

图 2 是利用 LSA 词汇表示模型②和余弦相似度计算方法得到的词汇间相似度信息. 原始矩阵为词汇-文档共现矩阵，即统计词典中每个单词在文档中出现的频次.

图 3 为利用 Word2Vec 词汇表示模型③和余弦相似度计算方法得到的结果.

图 4 为利用 LDA④ 词汇表示模型和余弦相似度计算方法得到的结果.

由图 2、图 3 和图 4 可以看出，相比 LSA 和 Word2Vec 模型，LDA 模型得到的词汇相似度边界

① 参考网页地址：http://licstar.net/archives/tag/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98.
② https://pypi.python.org/pypi/sparsesvd/
③ https://radimrehurek.com/gensim/models/word2vec.html
④ 根据文献中常见的参数设置，令 LDA 模型中参数：$\alpha=50/$主题个数，$\beta=0.01$.

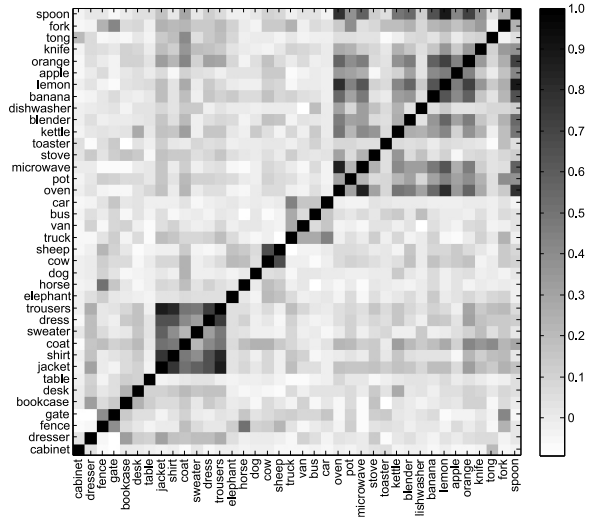


图 2　LSA 模型处理 BNC 数据得到的词汇相似度图

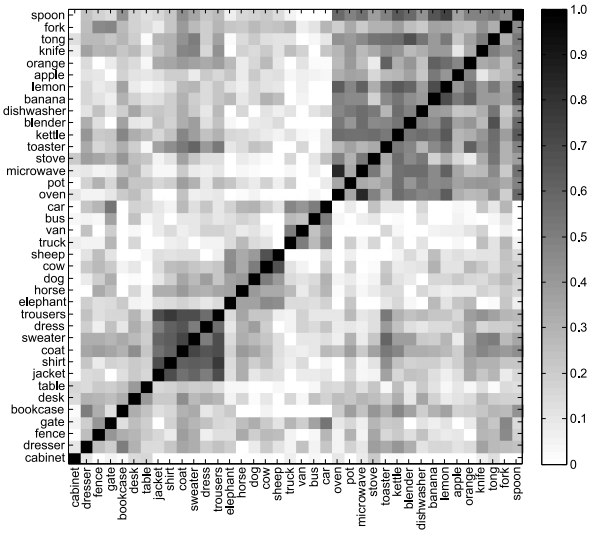


图 3　Word2Vec 模型处理 BNC 数据得到的词汇相似度图

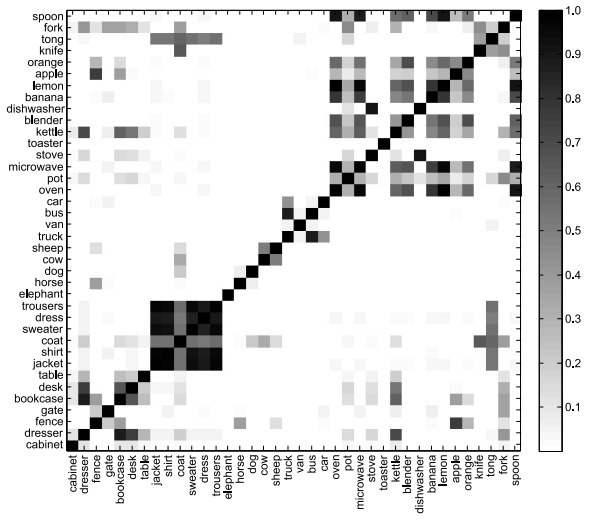


图 4　LDA 模型处理 BNC 数据得到的词汇相似度图

比较清晰. 相比 LSA 和 LDA 模型，Wod2Vec 模型捕捉了较好的属于动物和交通工具类别的词汇间的相似性，但是对于不属于同一类别的词汇赋予了较高的相似性，导致了在 McRae 所有词汇的相似度比较上效果较差. 总体来说，LSA、Word2Vec 和 LDA 这 3 种表示模型从 BNC 语料中计算出的词汇之间的相似度比较类似，3 个模型都捕获了部分与人脑类似的词汇相似度信息（如属于交通工具、动物和衣物类别的词汇）. 尽管与人的认知有一定的差距，但是通过表 1 可以看出，LSA 模型和 LDA 模型得到的词义表示能更好地描述词汇相似度信息，并且 LDA 模型是一种生成式的贝叶斯概率模型，其可扩展性和可解释性要优于 LSA 模型和 Word2Vec. LDA 模型结构比较灵活，利用概率图模型来表示词义可以融合其它语义和句法结构信息[14]，所以概率图模型受到了计算语言学家[15-18]和认知语言学家[19-22]的共同关注. 针对词义表示和归纳任务，由于贝叶斯概率模型具备从大规模无标注语料中学习词义表示的能力，类似人学习概念的过程. 因此，我们认为完全可以利用贝叶斯概率模型来提取文本中有用的信息以建立词义表示，并将其看作词汇概念的一个原型，利用相似度原则对标准数据集进行词义归纳.

3　模型描述

3.1　词义归纳模型

词义归纳任务的目的是自动地从无标注语料中获取单词的含义. 传统的归纳方法假设上下文信息可以对歧义词的含义进行区分，因此一般将词义归纳看作是无监督的聚类问题，利用上下文信息将歧义词划分为不同的类别. Brody 等人[11]解决词义归纳问题采用生成式模型，并利用贝叶斯方法进行求解，利用包含歧义词的句子作为 LDA 主题模型的输入，将模型求解得到的文档主题看作句子中歧义词的词义. 本文提出的模型建立在该模型的基础之上，因此以下首先对该基准模型做简要介绍.

LDA 主题模型如图 5 所示. 在词义归纳任务中 LDA 模型的输入为包含歧义词的句子.

假设语料中共有 M 个包含给定歧义词的句子，每个句子中包含 N_m 个单词，w_{mn} 表示第 m 个句子中的第 n 个词，s_{mn} 表示 w_{mn} 的词义. 假设歧义词共有 K 个词义，则上下文中词 w_{mn} 的分布为

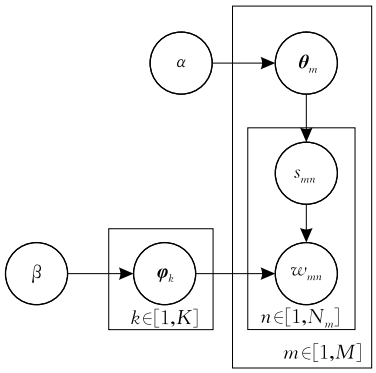


图 5 LDA 的图模型表示

$$p(w_{mn})=\sum_{k=1}^{K}p(w_{mn}\mid s_{mn}=k)\times p(s_{mn}=k) \quad (5)$$

令 $\boldsymbol{\varphi}_k=p(w_{mn}\mid s_{mn}=k)$ 为 LDA 模型的主题词项分布,即某个主题下单词出现的概率矩阵,矩阵每一行是一个词典大小为 V 维的向量,由狄利克雷分布产生,即 $\boldsymbol{\varphi}_k\sim Dir(\beta)$,图中 α 和 β 为狄利克雷分布的参数. 令 $\boldsymbol{\theta}_m=p(s_{mn}=k\mid d=m)$ 为句子主题分布,即某个句子中主题出现的概率矩阵,矩阵每一行是一个主题个数 K 维的向量,由狄利克雷分布产生,即 $\boldsymbol{\theta}_m\sim Dir(\alpha)$. 由 LDA 主题模型描述的句子中词的产生由如下过程组成:

对于主题 $k\in(1,\cdots,K)$:
采样主题词项向量: $\boldsymbol{\varphi}_k\sim Dir(\beta)$
对于包含歧义词的句子 $m\in(1,\cdots,M)$:
采样文档主题向量: $\boldsymbol{\theta}_m\sim Dir(\alpha)$
对于包含歧义词的句子 m 中的每个词 $n\in(1,\cdots,N_m)$:
采样词项 n 的主题: $s_{mn}\sim Mult(\boldsymbol{\theta}_m)$
采样词项 n: $w_{mn}\sim Mult(\boldsymbol{\varphi}_{s_{mn}})$

利用吉布斯采样(Gibbs sampling)方法对模型进行推断[23],每一次循环需要从条件概率分布函数式(6)中采样当前词的词义 s_i. $\boldsymbol{s}_{\neg i}$ 表示除了当前词的主题以外其他所有词项的主题; $\boldsymbol{w}_{\neg i}$ 表示除了当前词以外其他所有的词项; $n^k_{m,\neg i}$ 表示除了当前词以外句子 m 中属于主题 k 的词项个数; $n^t_{k,\neg i}$ 表示除了当前词以外主题 k 中属于词项 t 的个数.

$$\begin{aligned}p(s_i=k\mid \boldsymbol{s}_{\neg i},\boldsymbol{w})&\propto p(s_i=k,w_i=t\mid \boldsymbol{s}_{\neg i},\boldsymbol{w}_{\neg i})\\&=E(\theta_{mk})\times E(\varphi_{kt})\\&=\frac{n^k_{m\neg i}+\alpha_k}{\sum_{k=1}^{K}(n^k_{m\neg i}+\alpha_k)}\times\frac{n^t_{k\neg i}+\beta_t}{\sum_{t=1}^{V}(n^t_{k\neg i}+\beta_t)} \quad (6)\end{aligned}$$

LDA 主题模型中参数 β 控制主题词项分布的平滑程度, β 越大,某个主题下的词项概率分布越平滑. 参数 α 控制句子主题分布的平滑程度, α 越大,某个句子中的主题概率分布越平滑.

3.2 我们的模型

3.2.1 基本思路

贝叶斯概率模型将词义表示与词义归纳同步进行,即在实现词义归纳的同时,生成了关于词义的主题分布,使其具有了词义表示的功能. 借鉴 Brody 等人[11]的方法,我们首先利用 LDA 主题模型在大规模无标注数据集上学习词义的表示,在词义表示的基础上对测试集数据进行词义归纳. 模型流程见图 6. 考虑到不同词类的词需要不同的特征来表达[24],本文仅以名词歧义词的词义表达为实验目标.

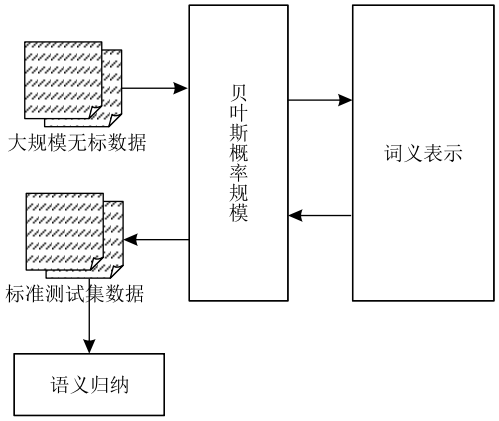


图 6 模型流程图

为了更好地捕捉句子中关于歧义词的词义信息,本文建立了双通道的狄利克雷分布(Dual Latent Dirichlet Allocation, Dual-LDA)主题模型,在基准 LDA 模型的基础上加入可以表达歧义词词义的信息作为另一个通道的输入. 下面对该双通道模型进行介绍.

3.2.2 Dual-LDA 主题模型

假设不同种类的数据分布不同,因此需要利用不同的输入通道来处理不同分布的数据. 图 7 描述了执行词义归纳任务的 Dual-LDA 模型,图中符号含义同图 5,下标 1 和 2 表示两个通道. 对于不同种

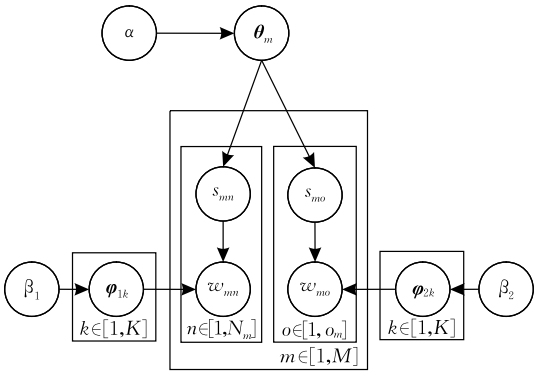


图 7 Dual-LDA 的图模型表示

类的数据赋予一个主题词项分布，在文档主题分布中进行融合. 利用双通道模型融合不同种类的数据，得到的有关句子的主题词项分布可以更好地对歧义词的词义进行表示.

考虑到人一般利用实词对概念进行描述，因此本文只提取文本中有明确语义的实词作为模型的输入. 根据宾州中文树库中的词性标注，我们认为具有如下词性标签的词具有明确的词义信息：其他动词(VV)、形容词(VA)、专有名词(NR)、普通名词(NN)、除名词之外修饰名词的词(JJ)、外来词(FW)和量词(M). 在我们的 Dual-LDA 模型中，从句子中提取的这类实词作为模型一个通道的输入，记作“通道一”. 如歧义词“黄牛”出现在如下句子中：

“刘翔博尔特，480 的票 100 了!”卖力的吆喝，无法阻止**黄牛**迎来赔钱的夜晚，上海八万人体育场，上座人数只怕还不到八千. 群众不再对“飞人”抱有信心，他们很明智，刘翔只得了第三，甚至没跑赢史冬鹏.

词性标注①后抽取出了该句子中所有上文所述词性标签的词：刘翔，博尔特，票，卖力，吆喝，阻止，黄牛，迎来，赔，钱，夜晚，上海，人，体育场，上，座，人数，怕，到，群众，飞人，抱有，信心，明智，刘翔，得，跑，赢，史，冬鹏.

另外，我们认为按上述方法抽取出的句子中的实词只能捕捉歧义词的主题或者部分上下文语境信息，并不能完全表示歧义词的词义. 通过对大量歧义词文本的观察，本文认为包含歧义词的基本单元中，与歧义词具有并列关系的词、与歧义词紧邻的形容词、名词及动词②，也可以表达歧义词的含义. 对于并列关系，本文利用“顿号”、“和”、“或”、“与”、“并”、“及”、“甚至”等关键字作为并列规则进行抽取. 对于基本单元的切分边界，本文利用标点符号“，”、“. ”、“?”、“!”、“：”等作为边界信息，从句子中抽取包含歧义词的基本单元. 对于上述包含歧义词‘黄牛’的句子，得到基本单元为：无法阻止黄牛迎来赔钱的夜晚. 对于歧义词的局部紧邻实词，为了避免语法分析工具带来的错误，本文抽取基本单元中歧义词前的一个形容词、动词和名词，以及歧义词后的一个动词和名词，来近似表示句子中歧义词的局部紧邻实词. 对于上述包含歧义词‘黄牛’的句子，抽取出的词为：阻止，迎来，钱. 为了缓解数据稀疏的问题，对于抽取的非歧义单词，本文在词义表示中加入了同义词词林中的词类信息. 这些抽取出来的词和词类信息一起作为 Dual-LDA 主题模型的另一个通道的输入，记作“通道二”.

同 LDA 模型，Dual-LDA 模型同样利用吉布斯采样方法对模型进行推断，每一次循环需要从条件概率分布函数公式(7)中采样当前词的词义 s_i. 与 LDA 模型不同的是，更新 n_m^k(句子 m 中属于主题 k 的词项个数)时需要利用两个通道的信息，因为两个通道的数据共同影响文本的主题词项分布. 其中 λ_1 和 λ_2 为两个通道的权重.

$$\begin{aligned} p(s_i=k\mid \boldsymbol{s}_{\neg i},\boldsymbol{w}) &\propto p(s_i=k,w_i=t\mid \boldsymbol{s}_{\neg i},\boldsymbol{w}_{\neg i}) \\ &=E(\theta_{mk})\times E(\varphi_{kt}) \\ &=\frac{n_{m\neg i}^k+\alpha_k}{\sum_{k=1}^{K}(n_{m\neg i}^k+\alpha_k)}\times\frac{n_{k\neg i}^t+\beta_t}{\sum_{t=1}^{V}(n_{k\neg i}^t+\beta_t)} \end{aligned} \tag{7}$$

其中，$n_m^k=\lambda_1\times n_m^k\{1\}+\lambda_2\times n_m^k\{2\}$，$\{1\}$表示通道一，$\{2\}$表示通道二，其他符号含义同式(6).

我们提出的 Dual-LDA 模型与传统的 LDA 模型有以下几处不同：Dual-LDA 模型的输入为从包含歧义的句子中抽取的不同特征，训练得到主题的词汇分布即为歧义词的词义表示，测试得到的文档主题即为歧义词的词义. 传统的 LDA 模型的输入为整个文档，训练得到的主题词汇的分布为整个文档的主题词汇分布，测试得到的主题为整个文档的主题. 图 8 示意性地描述了执行词义归纳任务的 Dual-LDA 模型的工作原理.

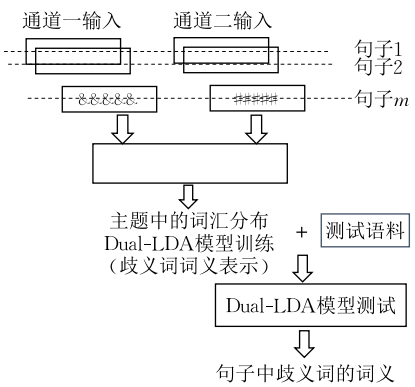


图 8　执行词义归纳任务的 Dual-LDA 的图模型表示

图 8 中通道一和通道二的输入为从句子中抽取的特征，通过 Dual-LDA 模型训练可以得到 φ_1 和 φ_2，分别为通道一和通道二的主题词项分布，即某个主题下单词出现的概率矩阵，可以看作是歧义词不同词义的表示. 这里需要强调的是，Dual-LDA 模型的两个通道共享 θ，即句子的主题分布，这样会影响

① 词性标注工具包：http://nlp.stanford.edu/software/tagger.shtml.

② 下文中将与歧义词紧邻的形容词、名词及动词统称为局部紧邻实词.

φ_1 和 φ_2 中主题下单词的分布，即影响对于歧义词词义的编码. 利用训练阶段得到的 φ_1 和 φ_2 之后，通过 Dual-LDA 模型就可以得到测试语料中句子内歧义词的词义编号.

4 实验与分析

实验中我们利用 *K*-Means 聚类模型[①]和单通道的基本 LDA 模型作为基线模型，一方面验证基于词义表示的方法[②]有利于词义归纳任务；另一方面验证本文提出的双通道的贝叶斯概率模型可以更好地获取词义表示和词义归纳的效果.

4.1 实验数据

我们选择的歧义词为 CLP2010 词义归纳评测任务[③]中的部分名词. 从大规模无标注训练语料中分别抽取包含歧义词的句子，筛选出 29 个出现句子数大于 1000 但小于 10000 的歧义名词.

利用搜狗新闻数据作为词义表示实验的训练数据，大小为 5.6 GB. 为了满足实验需求，首先从训练语料中抽取包含歧义词的句子约 13 万条. 然后对句子进行清洗(去重和乱码过滤). 最后对句子进行分词和词性标注，删除句子中词数少于 10 的句子，并去除停用词. 最终训练集包含句子约 12 万个. 测试语料中每个歧义词包含 50 个句子，每个歧义词的词义分布均匀.

4.2 评价方法

由于词义归纳可以被看作是对歧义词实例的聚类问题，因此一般利用标准聚类方法的评价指标来对词义归纳效果进行评价. 类别中实例的正确率是正确分类的样例数与总体样例数目的百分比，实例的召回率为返回的样例数目与相关样例(标准答案)数目的比值.

参考文献[11]对歧义词词义的评价方法，我们利用有监督模型的评价方法对实验结果进行评价. 由于本文提出的模型得出的主题个数与测试集中的歧义词的词义个数不同，我们需要将模型得到的主题词集合映射到相应的词义上才能对结果进行评价. 对于主题 $1\sim K$ 和词义 $1\sim S$，我们用下面的式(8)计算 KS(topic-sense)值，将主题空间映射到词义空间中：

$$p(s\mid k)=\frac{\#(k,s)}{\#k} \tag{8}$$

这样，给定某个文档的主题分布，相比于将概率最大的主题作为歧义词的词义，式(9)可以更好地预测歧义词的词义：

$$\arg\max_{s=1}^{S}\sum_{k=1}^{K}\theta_{jk}p(s\mid k) \tag{9}$$

其中，θ_{jk} 为测试集文档的主题分布，即包含歧义词句子中歧义词的词义分布.

有监督模型的评价方法包括准确率(*P*，*precision*)、召回率(*R*，*recall*)和 *F* 值. 由于词义归纳时会对每个句子的歧义词赋予词义，因此本文用准确率作为实验最终的评价指标. 令标准词义类别 S_r 的大小为 n_r，令经过 KS 计算映射到词义空间的类别 h_j 的大小为 n_j，$n_{r,j}$ 为上述两个类别相同实例的个数，则准确率可由式(10)计算得出

$$p(s_r,h_j)=\frac{n_{r,j}}{n_j} \tag{10}$$

4.3 模型选择

在模型选择时我们首先利用高频词义(Most-Frequent-Sense，MFS)模型对歧义词赋予常见的词义，作为一种基准模型，用来评定本文所选测试数据集的复杂度. 然后利用 *K*-Means 聚类模型对测试数据集中的歧义词进行自动聚类，聚类中心个数与 LDA 模型相同，也作为基线模型之一，用来比较基于词义表示的模型(LDA 模型和 Dual-LDA 模型)和聚类模型在词义归纳任务上的效果. 最后，本文选择 LDA 模型作为基准模型，用来与本文提出的 Dual-LDA 模型比较词义归纳效果，验证合理的词义表示是否可以提升模型的性能. 为了调试贝叶斯概率模型的参数，我们选择 3 个歧义词作为开发集分别对 LDA 模型和 Dual-LDA 模型进行调参.

在 LDA 模型中，需要确定 Dirichlet 分布的先验 α，β 和主题个数 *K*. 同文献[11]，本文固定参数 $\beta=0.1$，调节 α 和主题数目 *K*，α 调节范围为 0.005～1，*K* 调节范围为 2～9. 其中在开发集中，主题数目 *K* 对模型效果的影响如图 9 所示. 超参数 α 对模型效果的影响如图 10 所示.

通过开发集进行调参，最终确定参数为 $\alpha=0.32$，$K=3$. Dual-LDA 模型的狄利克雷分布的先验参数 α，β 及主题个数 *K* 选择和 LDA 模型参数相同. 为了简化调参过程，我们没有分别调节 β_1 和 β_2，令 $\beta_1=\beta_2=\beta$. 我们在开发集上对两个通道的权重 λ_1 和 λ_2 进行了调节实验，λ_1 和 λ_2 的调节值分别为[0.8，1.0，1.2].

① *K*-Means 聚类工具包：http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

② 本文中基于词义表示的方法指 LDA 模型和 Dual-LDA 模型.

③ CLP2010 词义归纳评测任务的数据来自新华日报和人民日报，并参照 HowNet 标准对其中的歧义词进行了词义标注.

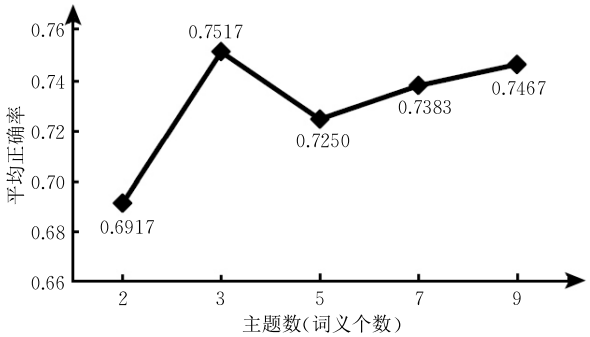


图 9　主题数对模型正确率的影响示意图

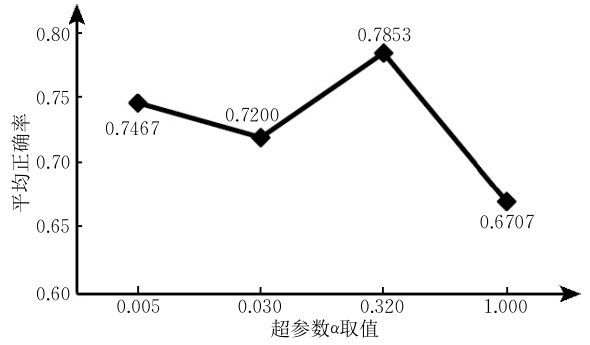


图 10　超参数 α 对模型正确率的影响示意图

经过收敛性验证，上述 LDA 模型和 Dual-LDA 模型设置吉布斯采样的迭代次数均为 1000.

由于模型中涉及随机采样过程，我们对每个模型都进行了 3 次实验，最后对结果取平均作为最终的实验结果.

4.4　实验结果

表 2 展示了 3 个基准模型、Dual-LDA 模型和调节了权重的 LDA 模型在词义归纳标准测试数据集上的效果. 根据 CLP2010 标准测试集中歧义词词义分布均匀的原则，可以得到 MFS 模型的平均正确率. 其中，每个歧义词的词义个数见附录 1.

由表 2 可以看到，LDA 模型和 Dual-LDA 模型的正确率高于 *K*-Means 模型，这说明基于词义表示的方法可以提升词义归纳任务的效果，但是效果提升并不十分明显，我们会在结果分析中解释原因. 由表 2 还可以看出，再加入歧义词的词义信息并对通道的权重进行调节后，Dual-LDA 模型在这个测试集上取得了最好的效果. 由于词义表示的好坏直接影响了模型的正确率，从而验证了本文的假设，即在句子中抽取出的可以表示歧义词词义的信息与句子的部分实词信息作为不同的输入通道，并通过贝叶斯模型进行整合，可以更好地表示歧义词的词义.

另外，我们对 Dual-LDA 模型的不同通道赋予了不同的权值后，观察到改变主题和词义的权重对词义表示效果的影响. 由表 2 可以看出，Dual-LDA (1.2{1}+0.8{2}[①])平均正确率最高，这说明在本文的实验数据集上，通道一的特征比通道二的特征对词义归纳任务的效果要好. 据此推断，如果对影响词义的重要信息赋予较大的权值，应该可以学习到更好的词义表示. 因此，我们在 LDA 模型的基础上，利用词频×逆文档频率(Term Frequency×Inverse Document Frequency，TF×IDF)和正值点互信息(Positive Pointwise Mutual Information，PPMI)指标对输入信息赋予不同的权值进行了进一步实验，但结果与预期相差较大，TF×IDF 和 PPMI 指标的引入相反地降低了模型的平均正确率. 对此我们也会在 4.5.3 节中解释其中的原因.

表 2　模型正确率

模型	平均正确率/%
MFS	49.06
K-Means	75.79
LDA	78.97
Dual-LDA	81.59
Dual-LDA(0.8{1}+1{2})	78.23
Dual-LDA(1{1}+0.8{2})	81.38
Dual-LDA(1.2{1}+0.8{2})	**82.92**
LDA-TF×IDF	63.00
LDA-PPMI	60.92

详细的实验结果见附录 1，词语正确率分布图见附录 2.

4.5　实例与结果分析

下面以“光环、结晶、黄牛、春秋”4 个词作为实例，对模型的效果进行进一步分析和说明. 如表 3 所示，因为“光环”和“结晶”在 *K*-Means 模型和基于词义表示的模型上的正确率相差较大，所以选择这两个歧义词来比较 *K*-Means 模型与基于词义表示的模型之间的差异. 而“黄牛”和“春秋”在 LDA 模型和 Dual-LDA 模型上的正确率相差较大，所以选择“黄牛”和“春秋”来比较 LDA 模型和 Dual-LDA 模型词义表示的差异. 表 4～表 7 分别展示了我们所选示例及贝叶斯模型得到的歧义词的词义表示结果. 完整的歧义词测试正确率结果见附录 2.

表 4 和表 5 中 S1～S3 表示包含歧义词 3 个含义的测试集实例，由于 *K*-Means 模型根据实例间词汇的重合程度对句子进行聚类，因此，表 4 和表 5 中用测试集实例来反应 *K*-Means 算法的聚类效果.

① 1.2{1}+0.8{2}表示通道一的特征权重为 1.2，通道二的特征权重为 0.8.

表 3 部分实例正确率

歧义词	*K*-Means	LDA	Dual-LDA
光环	76.00	60.67	67.33
结晶	64.67	94.67	95.33
黄牛	72.00	75.33	94.67
春秋	73.33	73.33	68.00

4.5.1 实例(光环)

“光环”一词在汉语词典[①]中有 3 个含义:(1)某些行星周围明亮的环状物,由冰和铁等构成,如土星、天王星等都有数量不等的光环;(2)发光的环子,如象征奥运会的五彩光环;(3)特指神像或者圣象头部周围的环形光辉. 词义归纳任务测试集中只给出了含义(1)和(3).

对于“光环”这个词,简单的 *K*-Means 聚类模型比 LDA 模型正确率高. 表 4 比较了语料中的“光环”和 LDA 模型对“光环”的词义表示. 由于测试集中包含“光环”的句子含义比较明确,用词比较一致,而利用大规模无标注语料学习到的词义表示得到的结果和词义归纳标准测试集的领域差异较大,因而导致了基于词义表示模型正确率的减少.

表 4 模型词义表示实例(“光环”)

例句
S1 天文学家认为,他们将很快会直接观测到黑洞,并且能够观测到这些〈head〉光环〈/head〉.
S1 太阳缓缓落山时,云层中的小冰珠将太阳光折射成一圈圈不同的〈head〉光环〈/head〉,继而又连缀成一根直达天顶的太阳光柱. 月亮光柱则更为罕见.
S2 佛像头后表现出的圆形〈head〉光环〈/head〉.
S2 当时多钦哲在他处见到多智钦的光蕴身在虚空中,〈head〉光环〈/head〉围绕,赫赫放光.
LDA
T1 光环 冠军 比赛 场 赛季 球员 世界 球队 中国 顶
T2 光环 中国 市场 公司 品牌 企业 基金 顶 元 国际
T3 光环 明星 生活 想 头 身上 笼罩 奥运 女 父亲

4.5.2 实例(结晶)

汉语词典给“结晶”一词定义了 3 个含义:(1)物质从液态(溶液或熔化状态)或气态形成晶体;(2)原子、离子或分子按一定空间次序排列而成的固体,具有规则的外形. 如食盐、石英、云母、明矾等;(3)比喻珍贵的成果,如劳动的结晶. 词义归纳任务测试集中将(1)和(2)归为一个含义,加上含义(3),共两个含义.

表 5 中比较了 *K*-Means 模型和 LDA 模型对于“结晶”的效果. 由表 5 可以看出 LDA 模型学习到“结晶”的词义表示与测试集标准词义一致. 对比实例“光环”并结合表 3 可以看出,LDA 模型和 Dual-LDA 模型在这种词义表示与测试集一致的实例上能得到更好的词义归纳效果.

表 5 模型词义表示实例(“结晶”)

例句
S1 现对外承接各类中草药的水提. 醇提层析、分离、萃取、〈head〉结晶〈/head〉加工服务及技术支持.
S1 日本研究人员称,他们制造出氧化钛的一种新的〈head〉结晶〈/head〉形式,可以用于制造“超级”蓝光光盘,这种光盘不仅价格更加低廉,而且其数据存储能力是 DVD 的几千倍.
S2 一片绿叶不仅是大自然的恩赐,更是人类辛勤劳作的〈head〉结晶〈/head〉.
S2 所以汉字是汉民族文化的〈head〉结晶〈/head〉,是民族文化历经数千年凝练而成的精华,值得善待珍视,而不应随便对汉字动“手术”.
LDA
T1 结晶 过程 研究 工艺 结构 作用 影响 水 技术 温度
T2 结晶 爱情 孩子 爱 岁 儿子 生活 女儿 想 妻子
T3 结晶 中国 智慧 文化 发展 技术 思想 国 精神 公司

4.5.3 实例(黄牛)

表 6 中 T1～T4 表示模型对主题通道的表示,ST1～ST3 表示模型对词义通道的表示. 在双通道模型中,由于主题相同的约束,T1 和 ST1 表示相同的词义,其他 T 和 ST 的含义与上述描述一致. 词义表示实例中的字母,如“Di02”,是为了改善词义通道数据的稀疏性引入的同义词信息,由同义词词林中得到.

表 6 模型词义表示实例(“黄牛”)

LDA
T1 黄牛 元 票 门票 买 钱 名 球迷 球票 价格
T2 黄牛 车 二手 元 市场 车牌 价格 头 牛 工作
T3 老 黄牛 头 牛 公牛 工作 人民 村民 岁 长
Dual-LDA
T1 黄牛 元 票 价格 门票 车 市场 二手 买 钱
T2 黄牛 头 牛 发展 产业 改良 县 元 企业 经济
T3 老 黄牛 工作 足球 成都 精神 影片 小罗 村长 党员
ST1 黄牛 买 票 手中 倒卖 告诉 钱 门票 车牌 黑车
ST2 黄牛 改良 养殖 县 产业 基地 发展 Di02 品种 Di18
ST3 老 拿出 黄牛 体育 恳恳 求票 奔走 遍地 勤勤 精神
LDA-TFIDF
T1 黄牛 倒卖 拍牌 假票 兜售 赠票 车票 挂号 号源 售票处
T2 黄牛 黄牛票 锵 肉用 倒号 鲁力 冻配 返现 驱使 空子
T3 黄牛 黄牛党 倒票 返券 贺根 车牌 黄牛们 出站 水牛 卖完
LDA-PPMI
T1 黄牛 倒票 黄牛党 炒到 假票 售票处 卖票 球票 返利 水牛
T2 黄牛 倒卖 车牌 拍牌 黄牛党 返券 二手 挂号 贺根 号源
T3 黄牛 兜售 贩子 赠票 卖完 票贩 万荣 二郎山 育肥 黄牛票

“黄牛”一词在汉语词典中有两个意思:(1)牛的一种,角短,皮毛黄褐色,或黑色,也有杂色的,毛短,用来耕地或拉车,肉供食用,皮可以制革;(2)指恃力气或利用不正当手法抢购物资以及车票、门票后高价出售而从中取利的人.

由表 6 可以看出,LDA 模型的词义表示结果捕

① http://xh.5156edu.com/

捉到了歧义词的常用含义. 表中 T1 表示含义(2)，T2 和 T3 表示含义(1). 在 Dual-LDA 中，T1 和 ST1 表示含义(2)，T2、T3、ST2 和 ST3 表示含义(1). 可以看出，ST1～ST3 对“黄牛”的词义做出了正确的表示，可以帮助提高模型的词义表示效果. 另外，从 LDA 模型和 Dual-LDA 模型的 T3 中可以看出，模型还捕捉到了“黄牛”的比喻义. 标准测试集中没有给出这个含义，这也会影响基于词义表示模型的效果. LDA-TF×IDF 和 LDA-PPMI 模型对输入信息按 TF×IDF 和 PPMI 指标赋予了不同的权重. Wilson 和 Chew[25] 利用这个方法在多语言的检索任务上得到了很好的效果. 但是，在我们的实验中(见表 6)该模型只捕捉到了“黄牛”的含义(2). 我们认为，这是因为训练语料是搜狗网络新闻文本，大部分包含“黄牛”的句子都是表示“黄牛”的含义(2)，而且模型的输入单词已经做过筛选，再对单词赋予权值可能会影响单词权重的真实表达. 因而 TF×IDF 和 PPMI 指标将“黄牛”含义(2)的相关词赋予了较高的权重，导致词义表示部分含义的缺失.

4.5.4　实例(春秋)

“春秋”一词在汉语词典中有 5 个含义：(1) 年岁；光阴：苦度春秋|他在讲台上耕耘了 40 个春秋；(2) 泛指历史：甘洒热血写春秋；(3) 时代名，因鲁国编年史《春秋》得名，一般指前 770 年～前 476 年这个时期；(4) 儒家经典之一，编年体春秋史；(5) 古代史书的通称. 词义归纳任务测试集中只给出了 3 个含义，分别是“春夏秋冬、春秋两个季度”和含义(3)、(1)，如表 7 所示.

表 7　模型词义表示实例(“春秋”)

LDA
T1 春秋 中国 时期 文化 历史 国 战国 淹城 荆州 时代
T2 航空 春秋 公司 旅客 元 服务 航班 上海 黑名单 延误
T3 春秋 墓 传 文物 发现 时代 岁 考古 世界 长
Dual-LDA
ST1 航空 春秋 公司 元 旅客 服务 航班 上海 黑名单 延误
ST2 春秋 中国 发展 社会 活动 世界 孔子 文化 传 生活
ST3 春秋 时期 国 文化 历史 战国 淹城 荆州 中国 时代

表 7 比较了 LDA 模型和 Dual-LDA 模型在“春秋”一词上的词义表示效果. 可以看到，LDA 模型和 Dual-LDA 模型捕捉到的词义与测试集给出的标准词义不符，这是由于语料中大量出现“春秋航空”和“春秋儒家经典”的原因，导致了贝叶斯模型得到“春秋”一词其他的含义. 由表 7 可以看到，Dual-LDA 模型的正确率并没有单通道 LDA 模型的正确率高，一方面由于贝叶斯模型捕捉到的词义与测试集不同，导致了在语料中捕捉到的表示词义的信息并不能帮助提高模型的正确率. 另一方面，“春秋”是抽象词，本文抽取的可以表示词义的特征并不适用. 比如包含歧义词“春秋”的句子：他在职业拳坛上奋战了 25 个春秋，一共打了 202 场职业赛，抽取的特征为：Dd05，Di19，奋战，拳坛，职业. 显然，这些特征对于词义归纳并没有帮助.

针对这个问题，由于本文针对词义归纳任务中的每个歧义词都训练一个模型，因此，在实际应用中选择模型融合的方式可以得到更好的效果，即对于每个歧义词，选择 LDA 模型和 Dual-LDA 模型在这个歧义词上测试正确率高的那个模型.

4.5.5　领域差异性分析

大规模无标注语料与标准测试集的领域差异性影响了基于词义表示的模型在词义归纳任务上的正确率. 实际上，训练集与测试集的差异性影响模型性能是所有统计方法存在的普遍问题. 针对这一问题，我们手工将测试样例中与基于词义表示模型得到的词义不一致的样例排除，最后利用剩余的测试样例重新评估模型的效果. 结果显示，排除了部分大规模无标注语料与测试集数据主题不一致的样例后，除了 K-Means 以外(K-Means 模型不需要训练)，LDA 模型和 Dual-LDA 模型的平均正确率均有了一定幅度的提高(见表 8). 具体实验结果见附录 1. 由表 8 可以看出，与 K-Means 模型相比，LDA 模型和 Dual-LDA 模型利用在大规模无标注语料中获取的词义表示信息对提升词义归纳任务的效果有明显的帮助，并且本文提出的 Dual-LDA 模型比 LDA 模型得到更好的词义表示和词义归纳效果.

表 8　部分测试集上模型正确率

模型	平均正确率/%
K-Means	74.89
LDA	83.44
Dual-LDA	86.93

5　总结与展望

本文尝试换一个角度来研究词义表示和归纳任务. 首先利用两种数据的相似度矩阵说明文本的词义表示可以捕捉大脑词义表征的部分信息，然后总结分析了已有的词义表示方法与人脑词义表征的关系，最后提出了一种新的汉语词义表示与归纳的方法，利用双通道的贝叶斯概率模型在大规模无标注数据集上学习歧义词的共现信息，建立词义的向量

化表示,并在此基础上,对测试集中的歧义词自动地进行词义归纳.实验表明,本文提出的方法可以提升词义归纳模型的性能.

为了更好地表示单词的含义,在下一步的工作中,我们将从以下几点进一步开展相关研究:

(1)人在进行词义归纳时,经常会从多个角度进行考虑,因此词义的表达应在多个维度上(主题、句法、语用、属性等)进行,词义类别可以由选择的维度确定.

(2)不同的上下文会影响词义的表达,因此应该建立更加灵活的模型以捕捉上下文对于词义的影响.另外,仅从无结构的文本中挖掘词义是十分局限的,大规模知识库的出现为词义表示提供了良好的条件.如果词义表示可以将知识库中的概念作为一般的世界知识,上下文可以对一般世界知识进行调整,那么应该可以得到更好的词义表示效果.

(3)针对训练语料与测试语料的领域不一致性和新义项问题,研究领域自适应方法,提高词义归纳模型的效果.

(4)研究人类学习概念、表征概念和整合不同种类信息的认知机制,建立计算模型以组合不同类型的信息,从而获得更好的词义表示结果.

致 谢 审稿专家对本文提出了深刻而到位的修改意见和建议,在此表示衷心地感谢!通过对这些审稿意见的学习和对本文的多次修改,我们感到受益匪浅.

参 考 文 献

[1] Mitchell J, Lapata M. Composition in distributional models of semantics. Cognitive Science, 2010, 34(8): 1388-1429

[2] Zong Cheng-Qing. Statistical Natural Language Processing. Beijing: Tsinghua University Press, 2013 (in Chinese)
(宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2013)

[3] Andrews M, Frank S, Vigliocco G. Reconciling embodied and distributional accounts of meaning in language. Topics in Cognitive Science, 2014, 6(3): 359-370

[4] McDonald S, Ramscar M. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity//Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Edinburgh, Scotland, 2001: 611-616

[5] Fyshe A, Talukdar P P, Murphy B, et al. Interpretable semantic vectors from a joint model of brain-and text-based meaning//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 489-499

[6] Lazaridou A, Pham N T, Baroni M. Combining language and vision with a multimodal skip-gram model//Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. Denver, USA, 2015: 153-163

[7] Andrews M, Vigliocco G, Vinson D. Integrating experiential and distributional data to learn semantic representations. Psychological Review, 2009, 116(3): 463

[8] Roller S, Im Walde S S. A multimodal LDA model integrating textual, cognitive and visual modalities//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, USA, 2013: 1146-1157

[9] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. The Journal of Machine Learning Research, 2003, 3(2): 1137-1155

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the International Conference of Learning Representations. Scottsdale, USA, 2013

[11] Brody S, Lapata M. Bayesian word sense induction//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece, 2009: 103-111

[12] McRae K, Cree G S, Seidenberg M S, et al. Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, 2005, 37(4): 547-559

[13] Chen X, Qi Y, Bai B, et al. Sparse latent semantic analysis//Proceedings of the 2011 SIAM International Conference on Data Mining. Phoenix, USA, 2011: 474-485

[14] Heinrich G. Parameter estimation for text analysis. Technical Report, 2005

[15] Griffiths T L, Steyvers M, Tenenbaum J B. Topics in semantic representation. Psychological Review, 2007, 114(2): 211

[16] Séaghdha D O, Korhonen A. Probabilistic distributional semantics with latent variable models. Computational Linguistics, 2014, 40(3): 587-631

[17] Erk K. What is word meaning, really?: (and how can distributional models help us describe it?)//Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics. Uppsala, Sweden, 2010: 17-26

[18] Boyd-Graber J L, Blei D M, Zhu X. A topic model for word sense disambiguation//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Pague, Czech Republic, 2007: 1024-1033

[19] Silberer C, Lapata M. Grounded models of semantic representation//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea, 2012: 1423-1433

[20] Baldewein U, Keller F. Modeling attachment decisions with a probabilistic parser: The case of head final structures//Proceedings of the 26th Annual Conference of the Cognitive Science Society. Chicago, USA, 2004: 73-78

[21] Jurafsky D. A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science, 1996, 20(2): 137-194

[22] Andrews M, Vigliocco G. The hidden Markov topic model: A probabilistic model of semantic representation. Topics in Cognitive Science, 2010, 2(1): 101-113

[23] Frermann L, Lapata M. Incremental Bayesian learning of semantic categories//Proceddings of the 14th International Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 2014: 249-258

[24] Xue N, Chen J, Palmer M. Aligning features with sense distinction dimensions//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydeny, Australia, 2006: 921-928

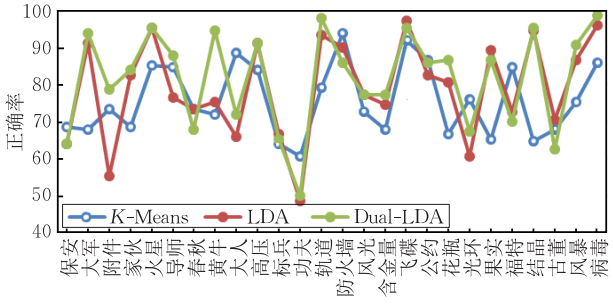
[25] Wilson A T, Chew P A. Term weighting schemes for latent dirichlet allocation//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 465-473

附录 1.

歧义词列表

歧义词	HOWNET 词义个数	歧义词	HOWNET 词义个数
保安	2	防火墙	2
大军	2	风光	2
附件	2	含金量	2
家伙	2	飞碟	2
火星	2	公约	2
导师	2	花瓶	2
春秋	3	光环	2
黄牛	2	果实	2
大人	2	福特	2
高压	2	结晶	2
标兵	2	古董	2
功夫	3	风暴	2
轨道	2	病毒	2

附录 2.



词语正确率分布图

模型正确率　　(%)

歧义词	*K*-Means	LDA	Dual-LDA
保安	68.67	64.00	64.00
大军	68.00	91.33	94.00
附件	73.33	55.33	78.67
家伙	68.67	82.67	84.00
火星	85.33	95.33	95.33
导师	84.67	76.67	88.00
春秋	73.33	73.33	68.00
黄牛	72.00	75.33	94.67
大人	88.67	66.00	72.00

(续　表)

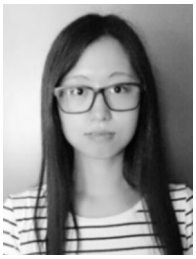
歧义词	*K*-Means	LDA	Dual-LDA
高压	84.00	91.33	91.33
标兵	64.00	66.67	65.33
功夫	60.67	48.67	50.00
轨道	79.33	93.33	98.00
防火墙	94.00	90.00	86.00
风光	72.67	77.33	77.33
含金量	68.00	74.67	77.33
飞碟	92.00	97.33	95.33
公约	86.67	82.67	86.00
花瓶	66.67	80.67	86.67
光环	76.00	60.67	67.33
果实	65.33	89.33	86.67
福特	84.67	72.67	70.00
结晶	64.67	94.67	95.33
古董	68.00	70.67	62.67
风暴	75.33	86.67	90.67
病毒	86.00	96.00	98.67

注：表中灰色区域为标准测试集中与大规模无标注语料的领域不一致的词汇. 如：歧义词‘导师’，大规模无标注语料中有大量‘中国好声音’相关的实例，这个主题没有出现在标准测试集中.

附录 3.

语义特征产生实验数据(McRae, 2005)

概念	特征	频次/人
airplane	beh_-_flies	25
airplane	has_wings	20
airplane	used_for_passengers	15
airplane	is_fast	11
airplane	requires_pilots	11
airplane	used_for_transportation	10
airplane	found_in_airports	8
airplane	is_large	8
airplane	made_of_metal	8
airplane	inbeh_-_crashes	7
airplane	used_for_travel	7
airplane	has_a_propeller	5
airplane	has_engines	5



WANG Shao-Nan, born in 1990, Ph. D. candidate. Her research interests include computational linguistics, and neurolinguistics.

ZONG Cheng-Qing, born in 1963, Ph. D., professor. His research interests include machine translation, text classification and natural language processing.

Background

Word sense representation is a foundational problem both in cognitive science and in natural language processing. There have been two increasingly popular approaches to the study of word representation in the two disciplines. One, based on theories of embodied cognition, treats meaning as a simulation of perceptual and motor states of human beings. An alternative approach treats word meanings as a consequence of the statistical distribution of words in spoken and written language corpus. Recent studies tend to argue that both statistical data and perception data have influence on the expression of word representation in the brain. A lot of researches found that brain areas related to perception and action will be activated when people are reading, which proved that word meaning is based on perceptual and motor states. Other psychology studies have proved the effect of distributional hypothesis on word representation in the brain.

Computational linguists treat word meanings as the result of statistical distribution. In order to build natural language processing systems, a lot of computational models have been proposed. Vector space model (VSM) is one of them which addresses the problem of language representation. The idea of the VSM is to represent each language unit in a collection as a point in a space (a vector in a vector space). Points that are close together in this space are semantically similar and points that are far apart are semantically distant. Latent semantic analysis (LSA) is one important method of VSMs. LSA model builds a latent semantic space by constructing a matrix containing word counts per paragraph and using singular value decomposition (SVD) to reduce the number of rows while preserving the similarity structure among columns. On the basis of LSA, Blei et al. proposed the LDA model, which allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Another approach to represent word meaning as a vector is neural network language model (NNLM). In order to improve the efficiency of NNLM models to generate word representations, Mikolov et al. proposed Word2Vec model, which is a single layer neural network.

To sum up, we wish to build the next generation of semantic models by studying how human brain handles the complexities of language.