以大规模语料库为基础的语义空间词向量类比关系研究

文献综述

* 1. **前言**

文献及相关资料主要来源于Spriger、中国知网和NAACL HLT会议数据库。主要领域为机器学习、CNN卷积神经网络、机器翻译、自然语言处理等。

随着计算机应用领域的不断扩大，自然语言处理受到了人们的高度重视。机器翻译、语音识别以及信息检索等应用需求对计算机的自然语言处理能力提出了越来越高的要求。

为了使计算机能够处理自然语言，首先需要对自然语言进行建模。自然语言建模方法经历了从基于规则的方法到基于统计方法的转变。从基于统计的建模方法得到的自然语言模型称为统计语言模型。有许多统计语言建模技术，包括n-gram、神经网络以及log\_linear模型等。在对自然语言进行建模的过程中，会出现维数灾难、词语相似性、模型泛化能力以及模型性能等问题。寻找上述问题的解决方案是推动统计语言模型不断发展的内在动力。

依靠计算机技术的发展和机器计算能力的跃迁，打破了性能障碍，曾经难以实现的工作也有了实践的基础，伴随着机器学习热潮，国内外对自然语言处理技术提出了越来越多的解决方案，解决了许多实际问题。

* 1. **主要工作介绍**

1. **Word2vec技术的提出**

Word2vec技术是由Google的T.Miklov等提出的新一代自然语言处理技术，具有训练快、准确率高等优点，已经被广泛应用于自然语言处理的机器翻译、语音识别等领域。Miklov等在模型中提出了两种大规模语料库的向量空间表示框架CBOW和Skip-gram[1]，本质上来说，这两种框架都是神经网络概率语言模型的具体表现形式。CBOW模型主要用于根据上下文通过出现概率来预测当前词，Skip-gram模型主要用于根据当前词来预测上下文[2]。这两种框架的应用极大地提高了模型的适用性和功能性，具有操作简便、预测速度快的特点[3]。

1. **以大规模语料库为基础的研究前沿**

大规模语料库是近年兴起的新概念，传统意义上的语料库是以文本为基础，通过标签和引用进行评价的文本语料库，具有构建成本高、准确率高、人工筛选成本高等特点。传统语料库的主要内容是对语料进行词性标注，功能单一缺乏大规模推广的价值，并且由于其依赖于人工筛选纠错，使得构建语料库成本高昂。在传统语料库研究中，研究者提出了许多模型原型和理论方法，如N-gram和log\_linear方法，但这些方法预测的准确性存在问题，其结果通常需要人工标注来进行修正，即便是精心设计的半自动标注模型也存在着标注成本的问题[4]，这制约了研究的进一步深入。对于语料库研究者来说，在Word2vec技术公开之前，他们不得不过多地借用已有的规模语料库成果，否则就必须承受准确性较差的自建语料库。

随着Word2vec技术的推出，大量的实证研究表明，Word2vec构建的语料库模型具有训练快、成本低、准确度高的特点，直接推动了大规模语料库的建设和相关研究的发展[5]。

1. **从语义空间到向量空间**

根据Google的T. Mikolov等提出的两种大规模语料库的向量空间表示框架CBOW和Skip-gram，国内外许多研究机构和研究者根据提出的框架实现原理进行了许多实证研究，并改进提出了不同领域的Word2vec算法拓展，特别的，国内研究者逐步攻克了中文语料库建模的难题，对中文语料的切片、分类提出了许多新方法，取得了一定的效果[6][7][8]。但目前应用效果最好的，还是英文语料库，其中以Google训练并开源的大规模语料库最为完整、规范。目前对word2vec技术的学习和研究多以这一语料库为主要研究对象。

综合国内外研究发现，Word2vec技术的应用主要是对语料的快速建模和语义分析，寻找相似关系以及衡量语义相似度，也被用于建立新的分类模型。在这些研究中，注意到类比关系及类比计算是Word2vec技术研究的基础。类比计算是通过训练好的语料库，在词向量之间进行向量运算，以相似度函数值的形式来衡量不同词向量之间的相似度和语义相关性[9]。目前的研究主要是对词向量两两之间的关系进行探索，尝试找到更加有效的关系评价指标体系。而对词向量对之间的关系仍然缺乏系统的研究，如“man-woman”和“king-queen”两个之间的存在的关系可以通过向量做差得到的新向量捕捉到，即新向量之间存在极大的相似性（模、方向）[10]。因此，不同词向量之间形成的差空间（逐项作差得到的向量空间）存在着丰富的语义关联研究方向，但现在仍缺乏系统的差空间研究。因此适合作为本文的研究方向。

1. **工作可行性及意义**

研究大规模语料库语义空间是当前的热门研究方向，类比关系是语义中一种主要的语义关系。对这一方面的研究，将丰富国内对Word2vec技术和大规模语料库的研究内容，为Word2vec技术的推广提供可复现的研究成果。

研究内容涉及面广，技术细节丰富，代码量较大。通过本研究的实践，可以充分锻炼建模能力、计算机编程能力和信息处理能力，体现了交叉学科的特点，锤炼了基本的学科技能，培养了处理复杂科研问题和理论转化为实践成果的思维能力和动手能力。

* 1. **参考文献**

1. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
2. Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.
3. 熊富林,邓怡豪,唐晓晟. Word2vec的核心架构及其应用[J]. 南京师范大学学报(工程技术版),2015,15(01):43-48.
4. Ker S, Huang C R, Hong J F, et al. Design and Prototype of a Large-Scale and Fully Sense-Tagged Corpus[M]// Large-Scale Knowledge Resources. Construction and Application. Springer Berlin Heidelberg, 2008:186-193.
5. 张为泰. 基于词向量模型特征空间优化的同义词扩展研究与应用[d].北京邮电大学,2015.
6. 陈培,景丽萍. 融合语义信息的矩阵分解词向量学习模型[J/OL]. 智能系统学报,:1-7(2017-08-31).
7. 曾琦,周刚,兰明敬,王濛. 一种多义词词向量计算方法[J]. 小型微型计算机系统,2016,37(07):1417-1421.
8. 李峰,侯加英,曾荣仁,凌晨. 融合词向量的多特征句子相似度计算方法研究[J]. 计算机科学与探索,2017,11(04):608-618.
9. 刘敏. 基于词向量的句子相似度计算及其在基于实例的机器翻译中的应用[D].北京理工大学,2015.
10. Enríquez F, Troyano J A, López-Solaz T. An approach to the use of word embeddings in an opinion classification task[J]. Expert Systems with Applications, 2016, 66:1-6.