第46卷第9期 中国科学技术大学学报 Vol. 46, No. 9
2016年9月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Sep. 2016

文章编号:0253-2778(2016)9-0719-08

基于低维语义向量模型的语义相似度度量

蔡圆媛,卢 苇

(北京交通大学软件学院,北京 100044)

摘要:语义相似性度量能够提高信息检索的准确性和效率,已成为文本处理中的一个核心任务.为解决一词多义等词汇歧义问题,提出一种基于低维向量组合的语义向量模型.该模型引入了知识库与语料库的多语义特征的融合,主要的语义融合对象包括连续的分布式词向量和从WordNet结构中的语义特征信息.首先利用深度学习技术中的神经网络语言模型,预先从文本语料中学习得到连续的低维词向量;然后从知识库WordNet中抽取多种语义信息和关系信息;再将多语义信息融入词向量进行知识扩展和强化,生成语义向量,从而实现基于向量空间的语义相似性度量方法.在基准测试集上的实验结果表明,该方法优于基于单一信息源(知识库WordNet或文本语料)的语义相似性度量方法,其皮尔森相关系数比基于原始词嵌套向量的方法提高了7.5%,说明在向量特征层面上的多语义信息的融合有助于度量词汇间的语义相似性.

关键词:语义向量;特征融合;分布式词嵌套;语义相似度

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.09.002

引用格式:蔡圆媛,卢苇. 基于低维语义向量模型的语义相似度度量[J]. 中国科学技术大学学报,2016,46(9):719-726.

CAI Yuanyuan, LU Wei. Semantic similarity measurement based on low-dimensional sense vector model[J]. Journal of University of Science and Technology of China, 2016,46(9):719-726.

Semantic similarity measurement based on low-dimensional sense vector model

CAI Yuanyuan, LU Wei

(*School of Software Engineering, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: Semantic similarity measurement enables the improvement of information retrieval in terms of accuracy and efficiency, so it has become one of the core components in text processing. To solve the problem of lexical ambiguity like polysemy, a sense vector model based on vector composition was proposed, which integrates knowledge base with corpus by fusing multiple semantic features derived from both of them. This model focuses on the continuous distributed word vectors and the inherent semantic properties in WordNet. Firstly, the continuous word vectors were trained from a textual corpus in advance by the neural network language model in deep learning. Then multiple semantic information and relationship information were extracted from WordNet to augment original vectors and generate sense vectors for words. Hence, the semantic similarity between concepts can be measured by the similarity of

收稿日期:2016-03-11;**修回日期**:2016-09-17

基金项目:国家自然科学基金(61272353),国土资源部地质信息技术重点实验室开放基金(201606)资助.

作者简介:蔡圆媛,女,1985年生,博士生. 研究方向:语义计算、信息检索. E-mail:yycai@bjtu.edu.cn

通讯作者:卢苇,博士/教授. E-mail:luwei@bjtu.edu.cn

sense vectors. The experimental results on benchmark indicate that this measure outperforms state-of-the-art measures based on either WordNet or corpora. Compared with the measures based on original distributed word vectors, the proposed measure has an improvement of Pearson correlation coefficient (7.5%). The outstanding results also show the contribution of multiple feature fusion to measuring the conceptual semantic similarity.

Key words: sense vector; feature fusion; distributed word embedding; semantic similarity

0 引言

语义相似性度量(semantic similarity measurement)指从特定的知识表意中计算出对象的相似程度,它是人工智能和自然语言处理中的重点研究内容之一.语义相似性度量能够有效提高数据匹配的准确率和效率,因此被广泛应用于自然语言处理和信息检索等领域[1],尤其是单词、句子等短文本的语义相似性计算,已经在微博等社交媒体相关的推荐系统[2]、舆情分析[3]等方面得到广泛应用.

依据语义信息来源,语义相似性计算方法主要分为基于语料库和基于知识库.语料库是大规模的文本集合,如英国国家语料库(British National Corpus),维基百科(Wikipedia)等.基于语料库的计算方法往往依赖于提取的文本特征,用于构造特征向量或统计模型.前者利用单词和文档的向量表示,将语义相似程度表示为欧式空间中的向量距离;后者基于概率统计学原理在概率向量空间中进行语义相似性的评估.此外,知识库是领域专家依据先验知识所构造的结构化数据,如语义词典、领域本体等.其中,本体 WordNet[4] 作为一类典型的知识库,与其相关的概念语义计算已成为核心研究内容.基于 WordNet 的语义相似性计算方法依据计算对象又被归纳为四类:路径距离、信息量、特征[5-6]和混合式[7-8].

现有的概念语义相似性度量方法采用单一的语义信息源,即语料库或知识库.例如,文献[8]利用 WordNet 的结构属性和概念间的语义关系衡量相似性,而文献[9]利用语料库提供的上下文和句法信息衡量概念间的语义相似性.虽然单个语义来源可以满足语义计算的任务要求,但是语料库和知识库均具有一定的局限性.大规模的语料不能直接反映出词汇的语义信息,导致相似性度量的准确性低下;而本体,尤其是领域本体的知识覆盖范围有限,在实际应用中适用性不强.针对上述问题,本文在概念的语义相似性的计算模型中引入知识库与语料库的语义知识整合,利用知识库提供的词汇语义和关系定义,增强词汇向量中隐含的语义知识.此外,在基于语料的语义相似度度量算法中,构建具有潜在语义特征的、高质量的词向量是提高性能的关键因素之一.由于一词多义和同义词带来词汇歧义问题,即单个词汇包含多种语义(概念),影响了语义相似度的准确度量,因此我们提出一个基于语义向量的语义相似度计算方法(semantic similarity measure based on sense vector, SSM-SV),通过向量组合对低维的连续词向量进行重构,整合 WordNet 和语料的语义特征,将词汇表示为对应的多个语义向量(本文也称其为词义向量或概念向量),实现语义消歧.

1 语义向量的构建

1.1 基于语料的低维词向量的构建

基于语料库的相似性计算方法通常将词表示为离散的分布式向量,基于特征选择和统计构造向量空间,通过一个词语的上下文建立 one-hot 分布式词向量(distributional word vector).离散的向量空间模型利用词语之间的同现关系来描述某个特定词,构建词与文档的共现矩阵.这种表示方法也成为词袋模型,即将文本表示为词的集合,基于统计模型来表示词的重要程度,如 TF/IDF、点互信息、Dice 系数等,但这类模型仅仅考虑了词形,无法反映词性以及语义层面上的特征,导致"词汇鸿沟".此外,one-hot 词向量数据稀疏问题严重,常用的平滑技术、降维技术(如矩阵分解)均从统计学角度解决该问题,未考虑语法、语义等语言学作用.

近年来,随着深度学习技术在自然语言处理领域的发展,词嵌套向量(distributed word vector, word embedding)因其连续、低维的特征表示以及挖掘文本潜层语义知识的能力,逐渐在自然语言处理中受到重视.词嵌套向量属于神经网络语言模型的训练产物,语言模型将文本的潜层语义信息、语法信息以及词态信息,通过神经网络模型的训练嵌入词向量的维度特征中.作为一个有效的知识表示的

手段，词嵌套向量在很大程度上提高了语义相似性计算、语义消歧及组合[10]和语义关系推理的性能.

1.2　基于语义消歧的语义向量的构建

基于特征选择和统计所构建的向量空间模型只能反映出语料词汇的词形和语法，基于此类传统模型的语义相似度度量方法往往依赖于语义消歧区分多义词的具体词义. 同样，用于特征学习的神经网络语言模型大部分[11-12]未考虑一词多义问题，无法获得多义词的多个语义向量.

为了解决词汇歧义问题，提高基于语料的语义相似度度量的性能，往往需要进行语义消歧从而获得语义概念的分布式向量表示，因此语义消歧的策略和性能成为基于有关度量方法的瓶颈. 传统的文本消歧基于上下文和人工标注，过程困难且耗时；而且用于语义标注的大规模语料难以搜集，导致人工标注的准确性较低，因此一些自动语义消歧的方法被提出，主要分为：利用额外的语义知识库和利用无监督的聚类算法.

1.2.1　利用额外的语义知识库

为了实现语义消歧，一些基于特征学习和预测的研究侧重于利用额外的知识库对语料中的词汇进行语义标记，生成多个相对应的“词-词义”对. 再通过神经网络对语义标记过的语料进行训练，生成与词义相对应的连续低维向量. 例如，Iacobacci 等[13]实现了将语料中的词汇与 BabelNet 语义网络中的语义概念的映射，并标记词汇所属的概念. 利用 CBOW(continuous bag of word)模型训练被标记的语料，生成词义对应的向量. BabelNet 整合了 WordNet 以及一些语义资源，因此词汇覆盖面广，适用性较强. 其中丰富的概念之间的语义关系也被 Iacobacci 等用于对词对的向量相似度结果进行缩放. 例如，若两个概念之间有直连边，则其相似度被放大，反之亦然. 类似地，Chen 等[14]的消歧过程基于词义注释的平均向量与上下文向量的相似度. 基于神经网络语言模型学习得到词嵌入向量，他们从 WordNet 中抽取目标词的多个注释，分别计算每一个注释中有效关键词的平均词向量，作为目标词对应的多个词义的初始向量. 再将目标词上下文的向量平均求和得到上下文初始向量. 找到与上下文初始向量具有最大余弦相似度的语义向量，通过对应的注释确定目标词所属的语义概念. 经过语义消歧和语义标注后的语料被用于训练一个改进的 Skip-gram 模型，更新并得到目标词最终的词义向量和上下文向量. Goikoetxea 等[15]不考虑 WordNet 中概念节点的定义和结构信息，仅仅利用 WordNet 连通图模型中的边信息，借助随机游走(random walk)算法抽取 WordNet 中的词汇组成语料，通过神经网络语言模型的训练学习得到词向量. 由于 WordNet 本身包含了对词汇的语义标注，因此得到的连续词向量就是词汇的语义向量.

1.2.2　利用无监督的聚类算法

当不存在额外的知识库可用，聚类算法也常被用于区分多义词的多个语义. 基于聚类的相关研究[16-20]针对一词多义的问题，利用聚类算法将词向量聚类，实现语义归类，再通过向量的线性操作得到语义向量或者通过神经网络将词义向量化. 这类研究通常包含两个步骤：向量聚类和对语料进行语义标注. 向量聚类是指对多义词的上下文向量进行聚类，再将每个聚类的质心向量当作一个语义的向量表示. 语料语义标注的目的是利用原始词嵌入向量的聚类结果对语料中的词进行消歧和语义标注，再将标注后的语料重新输入神经网络，学习得到低维连续的词义向量. Reisinger 等[16]基于词与上下文的共现关系和向量特征的加权策略(TF-IDF 和 χ^2)，提出一个多原型(multi-prototype)向量空间模型. 他们利用聚类算法将每个单词的上下文聚类为不同类别，分别表示单词所包含的多种不同的词义. Huang 等[17]扩展了上述基于高维向量的方法，他们将文档的全局语义信息引入递归神经网络，学习得到低维连续的多原型语义向量. Guo 等[18]基于双语平行数据，利用预先聚类和语义标注的手段，进行词义归纳(word sense induction). 针对源语言中的目标词，对其在翻译语言中对应的词汇直接聚类，生成包含不同语义的聚类集合，利用这些语义聚类的结果和双语数据对原语料中的目标词进行语义标注，将标注后的语料作为 RNN 神经网络的输入，学习得到具有消歧能力的词义向量. 上述基于聚类的方法仅仅是对神经网络学习得到的词向量进行后加工，没有改变神经网络本身的模型结构. 与上述方法不同，Neelakantan 等[19]改进了 Skip-gram 模型. 为了生成多语义的原型向量，他们基于两种聚类策略，提出了固定每个词汇语义数量的 MSSG(multi-sense Skip-gram)模型和不固定聚类数量的 NP-MSSG(non-parametric multi-sense Skip-gram)模型. MSSG 模型的基本思想是将上下文向量的聚类过程融入 Skip-gram 模型从而生成新的神经网络.

虽然基于聚类的语义向量的模型和算法在很大程度上改进了词汇向量的表义能力和辨别能力，但仍存在四点不足：

(Ⅰ)难以预先设定聚类算法中类的个数，因此确定每个目标词所包含的词义个数成为一个难题. 现有的方法[17]往往随机给所有的词都分配一个相同的、固定的聚类个数，而 Neelakantan 等的对比实验证明，固定词汇的语义数量将导致低质量的语义向量表示.

(Ⅱ)聚类是预先计算，无法适应新出现的词汇，且新词出现后需要重新计算，因此计算量较大. 如 Huang 等的方法需要花费约一周的时间才完成了基于一个包含十亿词汇的语料对 6 000 个词汇的向量训练[19]；

(Ⅲ)生成的不同聚类与词义的对应关系难以直接获得，往往需要借助其他的知识库才能真正实现语义消歧；

(Ⅳ)聚类算法的性能对于初始聚类质心和词义向量的初始值较为敏感，而上述基于聚类的方法仅仅采用简单的随机初始化或者利用 WordNet 的概念注释中词的平均向量.

针对上述问题，一些研究提出了相应的改进方法，如在聚类算法中加入额外知识的监督. Chen 等[20]的 VMSSG(variant of MSSG)模型基于两点对 MSSG 模型进行了改进，一是利用 WordNet 所给定的词汇的概念个数来决定语义向量的聚类个数；二是将 MSSG 模型对于语义向量的随机初始化取代为由 WordNet 的同义词注释经卷积组合模型 CNN 训练所得到的向量. 同样，为了学习得到多义词的多个语义向量，Wang 等[21]也将上下文的聚类与在线词典中的词义个数结合，标记相应的词-词义对并将其向量化.

上述基于语义消歧的两类方法均侧重于对语料进行预处理. 而本文提出的语义向量模型侧重于利用额外的语义信息源，对已有的词向量进行向量组合的后加工操作，从而增强词向量中的语义知识. 该方法充分利用了不同信息源中互补的语义信息，采用简单直观的向量组合，实现在特征层面的语义融合.

2 基于低维语义向量的语义相似度算法描述

组合操作最初源于基于逻辑的语义框架研究[22]，其理论依据是：一个句子所表达的主题和内容可以由句中包含的子句、短语或是单个的词的含义来解释. 换句话说，语言学上的每一种语法都应该对应着一个语义操作. 此外，Mikolov 等提出的词汇类比(Analogy)理论——基于嵌入词向量，向量代数操作(如加减对应着词汇语义属性的增加和删减)，说明向量操作可以保留、反映词汇之间隐含的语义关系特征[12]. 例如，单词“国王”的词向量约等于“男人”的词向量与“统治者”的词向量的和，而“王后”的词向量约等于“女人”的词向量与“统治者”的词向量的求和. 受上述研究的启发，我们认为，语义的融合和增强同样可以通过向量操作实现. 词向量不仅包含并且能够拆分出词汇的多个语义特征，而且同样包含词汇所对应的多个语义义项.

与其他能够生成多原型向量的研究不同，本文的 MSF 模型只生成一个语义向量，其理论依据在于：我们认为虽然多义词具有多种词义，存在于不同的同义词集合中，但是在某一语料中词汇可能突出表现为一个最主要的词义.

2.1 生成词嵌套向量

本文基于文献[12]提出的 log 线性神经网络 CBOW 进行词嵌套向量的学习. CBOW 模型移除了其他神经网络模型中的非线性隐藏层，简化了模型训练的复杂度，提高了计算速度，因此适用于大数据语料；但 CBOW 模型本身没有考虑语义消歧，学习得到的词汇只是单个向量表示，无法处理多义词可能存在的多原型向量表示，即没有考虑词汇的词义向量.

定义 2.1 对于一个在词表 W 中出现的单词 w_t，它的上下文为 K 窗口大小的单词集合 $C_t=\{w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k}\}$，即目标词的左边和右边各取 K 个近邻单词.

CBOW 模型首先将 C_t 中的单词词向量随机初始化作为输入层，在投影层对其求和平均，在输出层使目标词词向量出现的可能性最大. 其输出层的目标函数如下：

$$\mathrm{Obj}=\max\frac{1}{T}\sum_{t=1}^{T}\sum_{(-K\leqslant j\leqslant K,j\neq 0)}\log P(wt\mid wt+j) \tag{1}$$

式中，w_t是给定的目标词，w_{t+j}是其上下文 C_t集合.

$$P(wt\mid wt+j)=\frac{e(\mathrm{vec}'(wt)^{\Gamma}\mathrm{vec}(wt+j))}{\sum_{w=1}^{W}e(\mathrm{vec}'(w)^{\Gamma}\mathrm{vec}(wt+j))} \tag{2}$$

2.2　抽取 WordNet 中概念实体的定义与语义关系

本体 WordNet 是一个由美国普林斯顿大学开发的英语词典，它将词汇按照词性以同义词集(synset)的形式归类为名词、动词、形容词和副词. 一个多义词可能存在于不同的同义词集中，而一个同义词集合中的词互相存在同义关系. 不同的同义词集之间存在多种语义关系，包括上位/下位关系(IS-A 继承关系)、整体/部分关系、反义关系等. 图 1 展示了名词“coast”在 WordNet 3.0 中的 IS-A 层级结构.

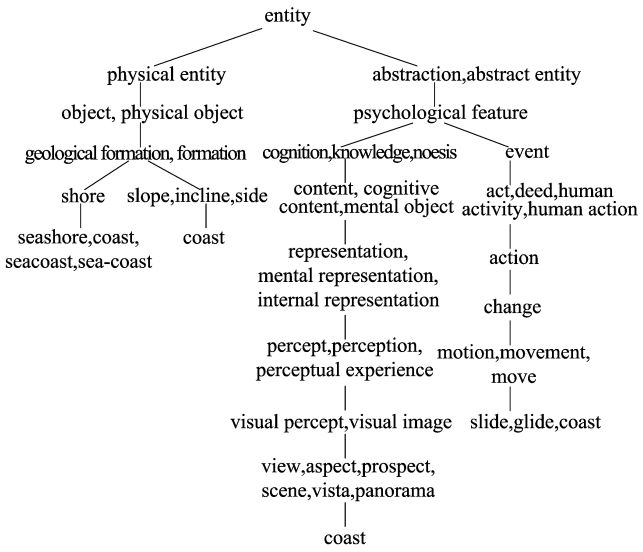

图 1　名词 coast 在 WordNet 3.0 中的 IS-A 关系

Fig. 1　IS-A relationship ofnoun *coast* in WordNet 3.0

定义 2.2　将 WordNet 视为一个无向图(V, E)，其中 V 代表概念节点的集合，E 代表连接节点的边的集合，每个边表示概念之间存在某个语义关系. 三元组$<C_1, r, C_2>$表示概念 C_1 与概念 C_2 存在的语义关系 r，$r\in E$ 且 $C_i\in V$.

对于给定的目标词 w，可能包含多个词义，存在于多个同义词集合中. 我们找出其在 WordNet 中对应的概念集合 $C(w)$，$C(w)=\{C_1(w), C_2(w), \cdots, C_i(w), \cdots, C_n(w)\}$. $C_i(w)$代表 w 在 WordNet 中的第 i 个词义. 本文中，我们定义 $r=\{$Synonymy, Hyponym, Hypernym$\}$，因为在 WordNet 中 80% 的语义关系是继承关系和同义关系. 对于每一个 $C_i(w)$，我们把与其相关的三元组抽取出来. 例如，与名词“coast”的第一个词义组成 Synonymy 关系三元组的集合为{seashore, seacoast, sea-coast}.

2.3　基于向量组合融合多种语义特征

我们将语义向量 SV(w) 定义为 WordNet 语义特征和从语料中得到的语义特征在相同维度空间中的特征组合，形式如下：

$$\mathrm{SV}(w)=F(V(w),V(s),V(a),V(d)) \quad (3)$$

式中，$V(w)$ 代表目标单词的词嵌套向量. s 代表单词 w 的同义词集，$V(s)$ 是该单词集合的平均词嵌套向量. a 代表单词 w 在 WordNet 中的祖先概念所对应的同义词集，$V(a)$ 是 a 所包含的多个同义词集合的词嵌套向量的求和平均. d 代表单词 w 在 WordNet 中的子孙概念所对应的同义词集，$V(d)$ 是 d 所包含的多个同义词集合的词嵌套向量的求和平均. 特征 s 侧重于利用同义词强化目标词的语义，而特征 a 和 d 侧重于利用上位/下位关系增强目标词的语义. F 代表对于上述特征向量的代数操作，包含特征求和、特征串联这两种操作.

通过公式(3)的语义向量模型，能够利用 WordNet 所包含的固有消歧语义信息对目标词的唯一的词向量进行语义强化，生成对应的多个词义向量. 其目的是将概念的语义相似度度量转化为基于词汇的概念向量的相似度计算：将词汇表示为概念向量，通过原始词向量的代数操作生成概念向量. 将词向量转换为概念向量的好处在于将词汇的多种语义都混合在一个向量中，减小生成词汇的多个语义对应的向量所需要的计算量. 此外，由于原始词向量基于神经网络模型训练生成，因此词向量本身就包含了词与词之间潜在的语义关联信息，避免了传统词袋模型中的“词汇鸿沟”问题，也避免了对词向量的特征进行选择时需要的人工干预. 为了寻找到最优的向量操作及特征属性的组合结果，我们基于语义向量的构建模型，提出了多种不同的语义组合策略，而在每一种策略下，一个单词只对应着一个唯一的语义向量. ①对 $V(w)$、$V(s)$、$V(a)$、$V(d)$两两组合，分别进行特征求和及特征串联操作. ②从 $V(w)$、$V(s)$、$V(a)$、$V(d)$中抽取三元进行向量求和.

基于上述特征融合的流程，本文利用两个单词的语义向量的相似性度量其语义相似性，公式如下：

$$\mathrm{Sim}(w_1,w_2)=\mathrm{MaxSim}(\mathrm{SV}(w_1),\mathrm{SV}(w_2)) \quad (4)$$

式中，w_1，w_2为两个单词，SV(w) 是单词 w 的语义向量. 单词在 WordNet 中具有多个对应的概念，因此用 MaxSim 表示将词对的所有概念相似度的最大值定义为其相似度. 向量距离(相似性)的计算选用多种度量方法，包括余弦相似度、欧氏距离和 Jaccard 相似性.

3 实验结果与分析

构建具有潜在语义特征的、高质量的词向量是提高基于语料的语义相似度计算方法的性能的关键因素之一.由于多义词和同义词等可能包含多种语义(概念),因此如何实现词向量到语义向量(sense vector)的转化计算,则成为本文语义相似度计算研究的难点和重点.为了应对大规模语料的预处理和训练,本文基于服务器将实验环境设置为16G内存和16核CPU,操作系统为Ubuntu12.4.

3.1 基准测试数据集与评价标准

实验中将RG-65[23]数据集作为基准.RG-65包含了65对英语词对,并由51名评分者给出得分区间为0-4的平均分数.该数据集合中词对的人工打分被用于与计算出的相似性值进行比较.实验中,皮尔森相关系数[24]被用于评价语义相似性计算方法的效果,即与人工评分值的相关度,公式如下:

$$\mathrm{Pcorr}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E(X-\mu_X,Y-\mu_Y)}{\sigma_X\sigma_Y} \tag{5}$$

式中,分子是协方差,分母是两个变量标准差的乘积.显然要求X和Y的标准差都不能为0.

皮尔森系数是一个区间为[-1,1]的浮点数,它通常用于衡量两个变量值或向量的线性相关性.当两个变量的线性关系增强时,相关系数趋于1或-1.正相关时趋于1,负相关时趋于-1.当两个变量相互独立时相关系数趋近于0,反之不成立.

3.2 训练语料与实验参数设置

本文的实验使用了三个语料,包括Wikipedia中约10 G大小的文本数据、Brown语料①和BNC②语料.对这三个语料进行整合,经过去除重复句、停用词和特殊字符等预处理步骤,得到包含约8亿个词的词表,使用谷歌公司开源软件word2vec③训练.该工具实现了CBOW模型,在训练中,窗口大小K被设置为3到8,生成的向量的维度为设置为100到700,步长为100.word2vec训练算法设置为层级softmax,且欠采样阈值参数设置为10^{-5}.

3.3 结果分析

针对三种向量相似性函数,包括余弦相似性、jaccard相似性和欧式距离,分别进行实验.结果表明基于余弦相似性的算法效果均优于基于后两者.因此,以下展示的结果均为由向量余弦相似性计算出的词对相似性与测试集中人工打分的一致程度.

实验1 单一语义特征向量

我们首先用单一的语义特征的向量化表征语义向量,结果如表1所示.维度和窗口分别对应词向量训练时所规定的向量维度和采用的窗口大小.

表1 单一语义特征向量在RG-65上的结果

Tab.1 Results of individual feature vectors on RG-65

语义特征	相关系数	维度	窗口
子孙(d)	0.651	200	8
祖先(a)	0.737	100	7
同义(s)	0.760	200	6
目标词(w)	0.812	300	5

由表1可知,WordNet包含的多种语义属性的向量表示降低了基于目标词词嵌套向量的相似性(0.812)的计算效果.这说明这些来源于知识库的语义属性包含的语义信息少于从语料库中提取的语义信息.我们以该结果为基准,分别进行单一语义特征向量与目标词向量的串联和叠加操作.

实验2 双语义特征向量的串联操作

基于连续的词嵌套向量,本文对不同的语义特征向量进行两两串联操作,实现目标词的知识扩展.对向量串联后实现了概念语义向量的维度扩展,并且能够突出不同属性的语义信息量.表2展示了向量串联操作的结果.

表2 双语义特征向量的串联在RG-65上的结果

Tab.2 Results of a couple of concatenate feature vectors on RG-65

语义特征	相关系数	维度	窗口
祖先(a),子孙(d)	0.705	500	3
同义(s),子孙(d)	0.761	500	4
同义(s),祖先(a)	0.770	200	4
目标词(w),同义(s)	0.808	200	3
目标词(w),子孙(d)	0.819	500	3
目标词(w),祖先(a)	0.831	200	4

由表2可知,不同语义属性的向量串联结果基本低于单纯采用目标词向量的结果,说明大部分特征的串联没有增加原词向量中的语义.上位词集与目标词在特征向量上的串联的结果有小幅提高,这

① http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml
② http://www.ota.ox.ac.uk/
③ http://word2vec.googlecode.com/svn/trunk

是因为在计算概念对之间的相似性时引入语义上更加抽象的上位词，扩充了概念的上层语义知识，有利于找出概念对之间共享的语义信息.

实验 3　双语义特征向量的求和操作

表 3 展示了目标词向量与不同语义特征向量的求和操作以及除目标词以外不同的语义特征向量间的两两相加操作. 由表 3 可知，双语义特征的求和操作得到的皮尔森系数普遍高于双语义特征的串联操作. 其中，上位词或下位词的语义特征叠加到目标词向量上能够明显提高系数，且上位词的贡献相对于上位词更大. 上位词的平均向量与目标词词向量求和操作使得皮尔森系数在 RG-65 数据集上达到了 0.875，优于其他特征之间的组合结果. 这说明，为了更精确地度量两个词汇之间的语义相似度，适当地在语义向量中加入更加抽象的语义特征能够增强语义向量发现共有语义信息的能力. 这一结论与实验 2 所得出的结论一致.

表 3　双语义特征向量的求和操作在 RG-65 上的结果

Tab. 3　Results of summation of two feature vectors on RG-65

语义特征	相关系数	维度	窗口
目标词(w)，同义(s)	0.776	200	6
祖先(a)，子孙(d)	0.781	200	4
同义(s)，祖先(a)	0.800	300	7
同义(s)，子孙(d)	0.815	200	6
目标词(w)，子孙(d)	0.857	500	3
目标词(w)，祖先(a)	0.875	500	4

实验 4　多语义特征向量的求和操作

基于 $V(w)$、$V(s)$、$V(a)$、$V(d)$，我们尝试了多个 3 元组合策略和 4 元组合策略，当训练模型的参数设置为维度 500 和窗口 3 时，学习得到的词向量使得 $V(a)$、$V(d)$ 与 $V(w)$ 之间的求和操作获得了 0.893 的最好结果，优于其他文献中提出的度量方法，如表 4 所示. 文献[5-6]为基于 WordNet 特征向量的相似度度量算法，文献[7-8]为基于 WordNet 的混合式度量算法. 并且，相对于基于原始词嵌套向量的基准结果(维度 500、窗口 3)0.818，我们提出的算法在 RG-65 数据集上使得皮尔森系数提高了 7.5%.

在 RG-65 数据集上的实验结果表明，本文所提出的基于语义向量的语义相似度计算方法优于基于 WordNet 的方法以及基于原始词向量的方法. 因此，基于简单直接的向量组合操作，在向量特征层面上的多语义信息的融合，有助于提高基于单一信息源的语义相似性度量方法的准确性. 值得一提的是，向量的操作方式对于高质量语义向量的构建起着决定性作用.

表 4　多语义特征向量的相加在 RG-65 上的结果

Tab. 4　Results of summation of multiple feature vectors on RG-65

度量方法	相关系数
Patwardhan[6]	0.797
Liu[5]	0.810
Pirro[7]	0.829
Gao[8]	0.863
目标词(w)，子孙(d)，祖先(a)	0.893

4　结论

本文提出了一种基于低维语义向量的语义相似性度量算法 SSM-SV. 主要工作包括：①基于低维词向量，通过向量组合将词向量转换为概念的词义向量，解决了一词多义问题；②整合英文语料和本体 WordNet 的语义概念及语义关系信息，采用多个语义属性实现了语义知识扩展及特征层面上的语义融合.

实验结果表明，这一方法在 RG-65 数据评测集上优于现有的基于单语义资源的相似性度量方法，说明基于向量的多语义特征的融合有助于提高于相似性度量的准确率. 下一步可以基于语义组合给一个给定的多义词生成对应多个义项的同义词向量，与本文所生成的词汇语义向量进行比较，并在句子相似度度量任务中验证模型的有效性和可扩展性. 此外，可以构造一个即包含本体语义关系、又包含文本上下文的矩阵，通过矩阵的分解来构造受本体语义关系约束的神经网络模型.

参考文献(References)

[1] PALIWAL A V, SHAFIQ B, VAIDYA J, et al. Semantics-based automated service discovery [J]. IEEE Transactions on Services Computing, 2012, 5(2): 260-275.

[2] WANG X, ZHAO Y L, NIE L, et al. Semantic-based location recommendation with multimodal venue semantics [J]. IEEE Transactions on Multimedia, 2015, 17(3): 409-419.

[3] QUAN C, REN F. Unsupervised product feature extraction for feature-oriented opinion determination [J]. Information Sciences, 2014, 272(8): 16-28.

[4] MILLER G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38 (11): 39-41.

[5] 刘宏哲. 文本语义相似度计算方法研究[D]. 北京交通大学, 2012.

[6] PATWARDHAN S, PEDERSEN T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts[C]//Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy: EACL Press, 2006, 1501: 1-8.

[7] PIRRÓ G. A semantic similarity metric combining features and intrinsic information content[J]. Data & Knowledge Engineering, 2009, 68(11): 1289-1308.

[8] GAO J B, ZHANG B W, CHEN X H. A WordNet-based semantic similarity measurement combining edge-counting and information content theory [J]. Engineering Applications of Artificial Intelligence, 2015, 39: 80-88.

[9] PENNINGTON J, SOCHER R., MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2014: 1532-1543.

[10] XU R, CHEN T, XIA Y, et al. Word embedding composition for data imbalances in sentiment and emotion classification [J]. Cognitive Computation, 2015, 7(2): 226-240.

[11] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013, arXiv: 1301.3781v3.

[13] IACOBACCI I, PILEHVAR M T, NAVIGLI R. SensEmbed: Learning sense embeddings for word and relational similarity [C]// Proceedings of the 53rd Association for Computational Linguistics and 7th International Conference on Natural Language Processing. 2015: 95-105.

[14] CHEN X, LIU Z, SUN M. A unified model for word sense representation and disambiguation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014: 1025-1035.

[15] GOIKOETXEA J, SOROA A, AGIRRE E, et al. Random walks and neural network language models on knowledge bases[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1434-1439.

[16] REISINGER J, MOONEY R J. Multi-prototype vector-space models of word meaning [C]// Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles: ACL Press, 2010: 109-117.

[17] HUANG E H, SOCHER R, MANNING C D, et al. Improving word representations via global context and multiple word prototypes [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2012: 873-882.

[18] GUO J, CHE W, WANG H, et al. Learning sense-specific word embeddings by exploiting bilingual resources [C]// Proceedings of COLING. Dublin Ireland: ACM Press, 2014: 497-507.

[19] NEELAKANTAN A, SHANKAR J, PASSOS A, et al. Efficient non-parametric estimation of multiple embeddings per word in vector space [J]. arXiv e-print, 2015, arXiv:1504.06654.

[20] CHEN T, XU R F, HE Y L, et al. A gloss composition and context clustering based distributed word sense representation model[J]. Entropy, 2015, 17(9): 6007-6024.

[21] WANG H, GAO B, BIAN J, et al. Solving verbal comprehension questions in IQ test by knowledge-powered word embedding [J]. arXiv e-print, 2015, arXiv:1505.07909v4.

[22] MONTAGUE R. English as a formal language [J]. Linguaggi Nella Società E Nella Tecnica Edizioni Di Comunita, 1970: 188-221.

[23] RUBENSTEIN H, GOODENOUGH J B. Contextual correlates of synonymy [J]. Communications of the ACM, 1965, 8(10): 627-633.

[24] SIMONOFF J S. Smoothing methods in statistics[J]. Journal of the American Statistical Association, 1997, 92(2): 379-384.