向量模型和多源词汇分类体系相结合的词语相似性计算*

梁泳诗，黄沛杰，岑洪杰，唐杰聪，王俊东

（华南农业大学数学与信息学院，广东 广州 510642）

摘要：现有的词语语义相似性计算主要包括基于向量模型以及基于词汇分类体系两类方法，但这两类方法都存在自身的缺点。向量模型所依赖的文本共现中的上下文信息不等同于真正意义上的语义，而词汇分类体系方法则存在构建代价大，并且在一定程度上还不够完善的问题。本文提出一种向量模型与多源词汇分类体系相结合的词语相似性计算方法，采用多源词汇分类体系的近义词关系以及向量模型得到的词向量，计算得到词语的向量表达，并探索了不同类型词汇分类体系提供的知识的选用和融合问题，弥补了单一词向量和单一词汇分类体系在词语相似性计算中的缺点。本文采用了 NLPCC-ICCPOL 2016 词语相似度评测比赛中的 PKU 500 数据集进行评测。在该数据集上，本文的方法取得了 0.618 的斯皮尔曼等级相关系数，比 NLPCC-ICCPOL 2016 词语相似度评测比赛第一名的方法的结果提高了 19.3%。

关键词：词语相似性；向量模型；词汇分类体系；组合方法；多源融合

中图分类号：TP391　　**文献标识码：**A

Word Similarity Computing by Retrofitting Vector-based Models with Multiple Lexical Taxonomies

LIANG Yongshi, HUANG Peijie, CEN Hongjie, TANG Jiecong, WANG Jundong

(College of Mathematic and Informatics, South China Agricultural University, Guangzhou 510642, China)

Abstract: There are two different methods used to compute semantic similarity. One uses vector-based models, and the other is based on lexical taxonomies. But each of these two methods has its own shortcomings. The vector –based models are based on the context's information from the co-occurrence of a word. But this information doesn't portray the authentic semantic relevance accurately. The lexical taxonomies which are not perfect enough to a certain extent suffer from huge construction cost. This paper proposes a way to calculate semantic similarity by linking vector model to multi-source lexical taxonomies. In this method, vector representation of a word is calculated through distributed representation from vectors-based models, and synonym relations are derived from multi-source lexical resource. Furthermore, this paper explores the way to select and fusion the knowledge from multiple lexical taxonomies. These combinational methods can make up the shortcomings of the two traditional semantic similarity computing methods mentioned above. We adopt PKU 500 for evaluation, which was used as the dataset of the NLPCC-ICCPOL 2016 shared task on Chinese word similarity measurement. Our method achieves a Spearman's ρ of 0.618, which gains 19.3% improvement comparing to the best result in the shared task.

Key words: word similarity; vector –based model; lexical taxonomy; combinational method; multi-source fusion

1 引言

词语的相似性是衡量两个词语之间语义相似的程度，是自然语言处理（natural language processing, NLP）的一个重要的任务，也是信息检索、机器翻译、自动文摘、问答系统、情感分析等众多NLP下游应用的基础[1]，所以如何正确地计算词语的相似性显得尤为重要。词语间的相似性主要有两种，一种是关系相似，另一种是属性相似[2]。在属性上有很强相似性的两个词语也被称为同义词。而本文所研究的词语相似性计算就是在属性相似上开展的。

*收稿日期：　　定稿日期：

基金项目：国家自然科学基金(71472068)

目前有两种计算词语相似性的方法，一种是基于训练文本上下文的向量模型，目前最主流的是基于词向量[3-4]。另一种是基于手工构建的词汇分类体系[5-7]。通过向量模型得到词向量，计算词向量间的余弦相似度代表词语间的语义相似性，这种基于向量模型的方法可以在文本语料中提取词语间的关系与词语的特征表达，但是上下文不等同于真正意义上的语义，向量模型的可解释性是受到限制的[8]。词汇分类体系是由人工构建的知识世界体系。根据词汇分类体系的结构特点，可以对词语的语义相似性进行计算，但是人工构建的词汇分类体系词汇量少，词汇分类粒度粗糙，难以对众多词语的语义差别进行细致的评价。

组合方法可以弥补单一词向量和单一词汇分类体系在词语相似性计算中的不足。Guo等人[9]在NLPCC-ICCPOL 2016评测比赛中，运用多种语料库得到的向量表达以及多种词汇分类体系对词语进行相似性计算，然后通过加权组合得到最终的词语相似性，取得了比赛的第一名。但他们的组合方法过于简单，也没有考虑不同类型知识来源的差别。Faruqui等人[10]利用词汇分类体系，在已经训练好的词向量上增强它的语义关系，弥补了词汇分类体系中词汇量不足的缺点，同时改善了词向量在语义表达。然而，他们忽视了不同的词汇分类体系对词语向量表达的修正带来的潜在的差异，本文在词语的向量表达构建中综合了不同类型的词汇分类体系知识，并初步探索了这些差异性知识的选用和融合效果。相比于已有的研究，本文的主要贡献包括：

（1）提出了向量模型和多源词汇分类体系相结合的词语相似性计算方法。采用HowNet、《同义词词林扩展版》等词汇分类体系的近义词关系以及中文信息学会社会媒体专委会提供的SMP 2015微博数据集训练得到的词向量，计算得到的词语向量表达，取得优于单一词向量、单一词汇分类体系以及单一词汇分类体系修正词向量等方案的词语相似性计算效果。

（2）研究了不同类型词汇分类体系提供的知识的选用和融合，进一步提高词语相似性的计算效果。在中文词语相似性评测的公开数据集 PKU 500 上进行实验，取得了 0.618 的斯皮尔曼等级相关系数，比 NLPCC-ICCPOL 2016 词语相似度评测比赛第一名的方法的结果提高了 19.3%。

本文后续部分安排如下：下一节介绍相关工作。第 3 节介绍本文提出的方法。第 4 节给出测试结果及分析。最后，第 5 节总结了本文的工作并作了简要的展望。

2 相关工作

在现有的词语相似性计算的两类方法中，基于向量模型建立在一个假设上：有相似语义的词语会倾向于在相似的上下文中出现。因此一个词语的语义可以通过对它所在的上下文建模计算出来[11]。尽管所有的向量空间模型都是基于相同的假设，他们又有各自的特色。他们之间最主要的区别在于如何定义上下文[11]。早期的模型是基于文档模型（document-based models）进行潜在语义分析（latent semantic analysis, LSA）[12]。这些模型是以所有的文档或者段落作为上下文，因此在文档中经常共同出现的词语会被视作语义相似。还有一种模型是近年来最受欢迎的分布式向量表示，它就是词向量，也称为词嵌入（Word Embeddings）[3-4]。它的核心思想是通过词的上下文（周围的词）训练出词汇表征[13]。在这种模型里面词语被投射进连续的空间，拥有相似上下文的词语在这个多维空间里面会很相近。

在词汇分类体系方面，过去有很多研究者花了巨大的人力构建词汇分类体系，意在为自然语言处理提供词汇知识库，如在中文上就有 HowNet[6]和《同义词词林扩展版》[7]，在英文上有 WordNet[5]、DBnary[14]等。

WordNet 和《同义词词林扩展版》都是以层次结构的方式呈现的，而词语的相似性就是根据词语在语义分类树上的距离所定义。WordNet 是一个非常闻名的英文词汇资源，它是由普林斯顿大学构建的[5]。WordNet 把名词、动词、形容词和副词连接成一套同义词集（synsets），每套同义词集都代表一个概念，同义词集之间会根据语义、概念和词汇关系相连接。一词多

意的词语会与多个同义词集对应，它们的意思会根据出现频率进行排序。而 HowNet 则与 WordNet 和《同义词词林扩展版》不一样，HowNet 是用复杂的、多个维度的知识描述语言对词语进行定义的。HowNet 选用义原（最小单位）作为标记集去描述词语的语义。通过这些标记集，可以对词语的语义相似性进行计算以及生成词类。

但是如上文提到，这两类传统的词语相似性计算方法在词语表达的语义性、构建代价以及词汇覆盖等方面都存在各自的缺点。本文提出一种向量模型与多源词汇分类体系相结合的词语相似性计算方法，采用多源词汇分类体系的近义词关系以及向量模型得到的词向量，计算得到词语的向量表达，并探索了不同类型词汇分类体系提供的知识的选用和融合问题，弥补了单一词向量和单一词汇分类体系在词语相似性计算中的缺点。

3 向量模型和多源词汇分类体系相结合的词语相似性计算

3.1 总体技术架构

图 1 是本文提出的方法的总体技术架构。

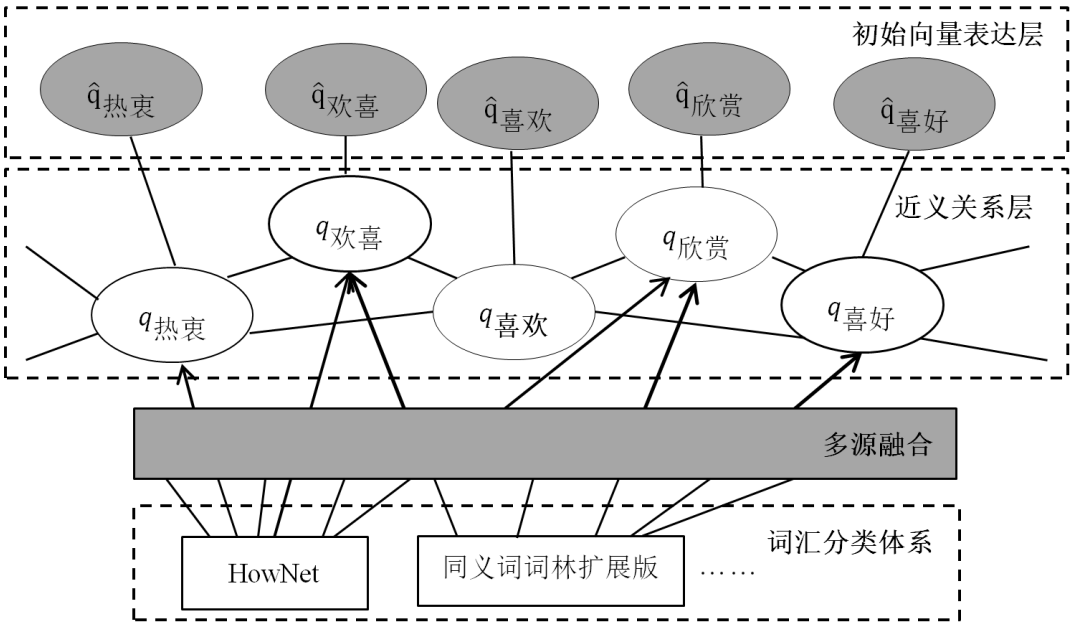

图 1　向量模型与多源词汇分类体系相结合的技术架构

在图 1 所示的技术框架中，主要分为四个部分：第一个部分是利用大型的语料库，通过向量模型训练得到词向量，构成初始向量表达层；第二个部分是词汇分类体系，本文选用 HowNet 和《同义词词林扩展版》两种中文词汇分类体系；第三部分是近义关系层，有别于 Faruqui 等人[10]的采用的单一词汇分类体系的近义词关系修正词语向量表达，本文采用多源词汇分类体系的近义词关系结合向量模型得到的词向量，计算得到词语的向量表达；第四个部分是多源融合层，与 Guo 等人[9]采用的简单组合方法不同，本文提出对不同类型的词汇分类体系提供的差异性知识进行选用和融合，通过不同的关联强度对词向量进行修正，最后得到更能体现词语语义的向量表达。

3.2 向量模型

目前训练词向量的主流方法是在训练语言模型的同时得到词向量。基于统计的语言模型能够表示成一个已出现的词和当前词的条件概率的极大似然估计：

$$\hat{P}(w_1^T)=\prod_{t=1}^{T}\hat{P}(w_t \mid w_1^{t-1}) \quad (1)$$

其中，w_t 表示句子中第 t 个词，$w_i^j=(w_i,w_{i+1},...,w_j)$ 表示句子中下标 i 到 j 的子串。

针对不同的上下文构造方法，在训练词向量时主要有 CBOW（continuous bag-of-words）和 Skip-gram 两种语言模型[4]。Skip-gram 模型允许某些词被跳过，在训练数据少的情况用 Skip-gram 可以创造更多的训练例子，而连续的 CBOW 则可以有较快的训练速度[4]。由于本文选用的词向量训练数据不论是新闻语料还是微博数据都是数量较大，因此本文使用 CBOW 语言模型对词语的语义层面建模。CBOW 语言模型不仅限于已出现的词为 w_t 的上下文，而是考虑了句子中距离当前词为 n 以内的词都看作是当前词的上下文环境。

用一个函数 f 表示当前词 w_t 的上下文的向量到当前词 w_t 条件概率的映射[3]，并结合 CBOW 的机制，则当前词的上下文和当前词的条件概率可以表示为：

$$\hat{P}(w_t \mid context(w_t))=f(w_t,C(w_{t-n}),...,C(w_{t-1}),C(w_{t+1}),...,C(w_{t+n}))=f(w_t,\sum_{0<|i-t|<n}C(w_i)) \quad (2)$$

其中，$C(w_i)$ 是词语 w_i 的分布式特征向量。

在训练语言模型及词向量时，对于 w_t 都要扫一遍词库大小|V|，计算复杂度过高。可以采用负采样(negative sampling)[15]和分层的 softmax(hierarchical softmax)[16]的方法来降低计算复杂度。

3.3 词汇分类体系

基于词汇分类体系计算词语相似度的方法是在某种世界知识库上展开的，这些世界知识库一般都采用一颗或者几颗树状的层次结构对词语的概念进行描述，在这些层次结构图中，一个概念代表一个点，任何两个节点之间有且只有一条路径，这条路径的长度就可以反映这两个概念的语义距离。本文主要研究的是两个中文方面的词汇分类体系，分别是HowNet[6]以及《同义词词林扩展版》[7]，并根据词汇分类体系各自的结构特点，制作近义词词典。

在HowNet中，“义原”是描述概念的最基本单位，不同义原的集合表述不同的概念。HowNet中的词语有一个或者多个概念[17-18]。如在HowNet中词语“男人”的表述如图2所示。

```
NO.=061553
W_C=男人
G_C=N
E_C=
W_E=husband
G_E=N
E_E=
DEF=human|人,family|家,male|男
```

图2　HowNet结构示例

从图2可以看到，在HowNet中，词语“男人”的概念是DEF=human|人,family|家,male|男，人、家、男就是组成概念的义原。

HowNet中的义原有1600多个[18]，HowNet中的中文词语就由这些义原的组合进行描述。义原又以树状结构的层次体系进行组织，通过义原在层次体系中的深度求出义原的相似度，进而逐步求出词语概念的相似度以及词语的相似性。本文利用HowNet的词语相似性的计算方法，计算出HowNet中所有词语两两之间的相似性，并把一个词语以及与之相似度最高的词语视为该词语的近义词词集，所有近义词词集组合成HowNet的近义词词典。

而《同义词词林》则是由梅家驹等人[19]在1983年整理编写的，后由哈尔滨工业大学信息检索实验室进行更新而成的一部具有汉语大词表的“哈工大信息检索研究室同义词词林扩展版”[7]。本文实验中采用扩展版，下文简称《同义词词林扩展版》。《同义词词林扩展版》包含约7万条词语，按照词语的意思进行编码，是一部同义类词典，如图3所示。

```
Ae16E01= 创意者 创意人
Ae17A01= 演员 艺人 优伶 戏子 优 伶 艺员 饰演者 表演者 扮演者
Ae17A02= 明星 影星 星 超巨星 超新星 大腕
Ae17A03= 女演员 坤角儿 女星
Ae17A04= 歌手 歌姬 歌舞伎 伎 歌者 歌星 演唱者 唱工 唱头
```

图 3 《同义词词林扩展版》示例

《同义词词林扩展版》在秉承《同义词词林》编撰风格的基础上，对《同义词词林》进行修正与扩充。与《同义词词林》编码规则类似，《同义词词林扩展版》按照树状层次结构把词条进行组织，把词语分为大、中、小、词群和原子词群五类，大类有12组，中类有95组，小类有1425组，词群有4223组，原子词群有17807组。每一个原子词群中都有若干个词语，同一原子词群的词语不是语义相同或十分接近就是语义有很强的相关性[7]。每一行都有自身所属的编码，在《同义词词林扩展版》中词语的相似性就是根据每一行的编码计算的。编码的最后一位标记符用于说明同一个原子词群中的词语关系，共有3种标记符，分别为“=”、“#”、“@”，“=”代表相等、同义，“#”代表同行词语属于相关词语，是同类，不能视为相等，“@”代表独立，表示在词典中该词既没有同义词也没有相关词。《同义词词林扩展版》自身就是一部同义词类的词典，每一行词语视为语义上具有强相关性，可以直接利用在词向量的修正上。并且，《同义词词林扩展版》中近义词的不同标记符也成为了本文对其提供的知识进行选用的依据。

3.4 向量模型和多源词汇分类体系相结合

向量模型和词汇分类体系相结合的方法可以弥补单一词向量和单一词汇分类体系在词语相似性计算中的不足。Guo等人[9]在NLPCC-ICCPOL 2016评测比赛中也运用了多种语料库得到的向量表达以及多种词汇分类体系对词语进行相似性计算，但他们的组合方法过于简单，仅仅通过加权组合得到最终的词语相似性。Faruqui等人[10]利用词汇分类体系，在已经训练好的词向量上增强其语义关系，在英文语料上取得了较好的应用效果。本文在其基础上，进一步考虑不同的词汇分类体系对词语向量表达的修正带来的潜在的差异，在向量表达的构建中综合了不同类型的词汇分类体系知识，并研究了这些差异性知识的选用和融合效果。具体上，如图1所示，由近义关系层、初始向量表达层以及多源融合层共同完成本文方案中词语向量表达的构建。

近义关系层提供了特定词语在词汇分类体系中的近义词关系信息。通过不同类型的词汇分类体系，可以得到多组语义上具有强相关性的词集，如上文提到的HowNet中的相似度最大近义词词集以及《同义词词林扩展版》中的原子词群。

初始向量表达是通过语料库训练出来的词向量 $\hat{q}_i \in R^d$，其中 d 为分布式向量的维度，则有矩阵 $\hat{Q}=(\hat{q}_1,\hat{q}_2,\mathrm{L}\ ,\hat{q}_n)$ 为初始词向量集合。本文词语向量表达构建的目标就是利用 $\hat{Q}$ 结合近义关系层的近义关系，进行词语向量表达的迭代改变构建出词语矩阵 $Q=(q_1,q_2,\mathrm{L}\ ,q_n)$，这样的 Q 不但携带上了词汇分类体系中准确的语义信息，而且保留着分布式向量原有的信息。如图 1 所示，每一个词语的词向量被投射到平面中，其中灰色的椭圆代表利用语料库训练出来的词语初始向量表达，白色的椭圆之间的连线 $(w_i,w_j)\in E$ 则表示一个来自于某一词汇分类体系的词语近义关系。图 1 显示了词语“喜欢”的部分近义关系，如近义词“热衷”来自于 HowNet，近义词“喜好”来自于《同义词词林扩展版》，而近义词“欢喜”和“欣赏”则同时来自于 HowNet 和《同义词词林扩展版》。Faruqui 等人[10]利用词汇分类体系提供的近义词关系对初始词向量进行修正，实现词向量语义关系上的加强。

本文在Faruqui等人[10]提供的方法基础上，增加了多源融合层。考虑到不同词汇分类体系，以及同一词汇分类体系内部的不同近义情况对于词语语义向量表达价值的差异，本文增

加多源融合层对多源的词汇分类体系提供的知识进行选用和修正权重的赋予。目前本文仅在一定数量案例分析的基础上尝试了一些较为基础的选用考虑因素，更为系统地选用机制还有待进一步研究。一方面，对来自于《同义词词林扩展版》的强关联词，本文选取了编码的最后一位的标记符为“=”的原子词群，而弃用了标记符为“#”的原子词群，因为“#”代表的词语间是相关的，是同类，但在很多情况下和同义有一定差距。另一方面对于来自于HowNet相似度最大的近义词词集里，本文只保留最大相似度为α以上的近义词词集（在后面的实验中，我们采用了α=0.75的设置，更优化的参数可以通过验证得到），因为相似度过低的词语，对词向量的修正可能会造成负面影响。本文的实验表明，上述词汇分类体系的知识选用有助于近义词词集质量的提高。

给定某一词语的近义关系，以及该词语与近义词集的初始化向量表达，词语的向量表达构建可以通过以下方式进行。向量与向量之间的距离采用了常用的欧几里得距离计算，最终构建的词语向量表达 q_i 与该词语的初始词向量 $\hat{q}_j$ 及其任意的近义词 q_j 差距应该尽可能的小，并考虑不同来源的词汇分类体系知识的价值差异，得到以下迭代公式（3）：

$$\Psi(Q)=\sum_{i=1}^{n}\left[\alpha_i\left\|q_i-\hat{q}_i\right\|^2+\sum_{(i,j)\in E_1}\beta_{ij}^1\left\|q_i-q_j\right\|^2+\sum_{(i,j)\in E_2}\beta_{ij}^2\left\|q_i-q_j\right\|^2+\mathrm{L}+\sum_{(i,j)\in E_k}\beta_{ij}^k\left\|q_i-q_j\right\|^2\right] \tag{3}$$

其中，α 和 β 是控制关联相对强度的系数，β^k 代表不同来源的词汇分类体系知识权重，i 代表需要构建的词语，j 代表词语 i 的近义词。

根据要求，对公式 $\Psi(Q)$ 求最小值，对 $\Psi(Q)$ 进行一阶求导，通过迭代，求解线性方程组得到求解公式：

$$q_i=\frac{\sum_{j:(i,j)\in E_1}\beta_{ij}^1 q_j+\sum_{j:(i,j)\in E_2}\beta_{ij}^2 q_j+\mathrm{L}+\sum_{j:(i,j)\in E_k}\beta_{ij}^k q_j+\alpha_i\hat{q}_i}{\sum_{j:(i,j)\in E_1}\beta_{ij}^1+\sum_{j:(i,j)\in E_2}\beta_{ij}^2+\mathrm{L}+\sum_{j:(i,j)\in E_k}\beta_{ij}^k+\alpha_i} \tag{4}$$

其中，$\sum_{j:(i,j)\in E_1}\beta_{ij}^1+\sum_{j:(i,j)\in E_2}\beta_{ij}^2+\mathrm{L}+\sum_{j:(i,j)\in E_k}\beta_{ij}^k+\alpha_i=1$。任意 q_j 在训练前会被初始化成 $\hat{q}_j$。

4 实验

4.1 实验数据集

本文采用两个来源的训练语料库训练词向量，分别是搜狗实验室提供的搜狗新闻数据集以及中国中文信息学会社会媒体专委会提供的SMP2015微博数据集(SMP 2015 Weibo DataSet)，其中在SMP 2015 Weibo DataSet 中取了4G和10G的微博作为训练语料库，得到三个语料库用于对比试验。

在中文的词汇分类体系选用方面，本文采用了HowNet 2000版（HowNet的开源版本，实验结果中标记为“HowNet”）以及哈工大信息检索研究室的《同义词词林扩展版》。

在实验效果评价方面，采用了中文词语相似度评测数据集 PKU 500 数据集[1]。PKU 500 共有 500 对词语，每对词语都有人工标注的相似度(范围为 0-10)。PKU 500 被采用到第五届国际自然语言处理与中文计算会议暨第 24 届国际东方语言计算机处理会议（NLPCC-ICCPOL 2016）的评测比赛中。

4.2 实验设置

本文采用斯皮尔曼等级相关系数（Spearman rank correlation coefficient，实验结果中标记为“Spearman’s ρ”）去衡量词向量计算词语相似性的效果。通过计算PKU 500中每对词语人工标注的相似性和词向量计算出的词语相似性之间的斯皮尔曼等级相关系数，借以判断各

实验方案对词语相似性的计算效果。

实验方案为：

（1）基于向量模型的词语相似性计算：对比不同的训练语料训练得到的词向量对词语的相似性计算的效果。

（2）基于词汇分类体系的词语相似性计算：对比不同词汇分类体系，本文中为HowNet和《同义词词林扩展版》应用于词语相似性计算的效果。

（3）基于向量模型与词汇分类体系相结合的词语相似性计算：分别利用HowNet与《同义词词林扩展版》所提供的知识参加词语向量表达的构建，考察其结合方法的效果。

（4）基于向量模型与多源词汇分类体系相结合的词语相似性计算：对比了本文提出的方法在不同类型的词汇分类体系的知识选用及其在词语向量表达构建中的不同权重的效果。

（5）研究进展方法在中文词语相似性计算上的性能对比：对比了本文提出的方法与研究进展方法在词语相似性计算上的性能。

本文的方法，向量模型和多源词汇分类体系相结合的词语相似性计算方法，对比的两个研究进展的方法如下：

（1）NLPCC-ICCPOL 2016评测比赛第一名的方法：Guo等人[9]在NLPCC-ICCPOL 2016评测比赛中也运用组合多种语料库得到的向量表达以及多种词汇分类体系对词语进行相似性计算的方法，对比实验中直接引用了其在比赛中得到的结果。

（2）向量模型与单一词汇分类体系相结合的方法：Faruqui 等人[10]利用词汇分类体系，在已经训练好的词向量上增强它的语义关系，在英文语料上取得了较好的应用效果的方法。本文将其应用于中文词语相似性计算，在实验中，以实验效果最好的单一词汇分类体系（本文实验中为《同义词词林扩展版》）修正词向量的结果代表该方法的结果。

4.3 实验结果分析

4.3.1 基于向量模型的词语相似性计算

本实验运用word2vec的CBOW模型在三个语料上进行词向量的训练，“搜狗新闻”代表搜狗新闻语料库训练词向量。“4G微博”代表用4G的微博数据来进行词向量训练的方案。“10G微博”代表用10G的微博数据作为语料库来训练词向量。三个语料库对PKU 500的数据集词语的覆盖率，如表1所示。

表 1 不同的词向量训练语料库对 PKU 500 词语的覆盖率

词语覆盖性	缺失词语数	词语覆盖率
搜狗新闻	12	98.8%
4G 微博	8	99.2%
10G 微博	0	100%

表 1 可以看到，搜狗新闻语料库对 PKU 500 的 1000 个词语中有 12 个词语不存在，4G 微博语料库缺失了 8 个，而 10G 微博的语料库覆盖了全部词语。进一步地，我们对比不同语料库训练的词向量的词语相似性计算效果，如表 2 所示。

表 2 不同语料库训练的词向量的词语相似性计算效果

训练语料库	Spearman's ρ
搜狗新闻	0.412
4G 微博	0.413
10G 微博	0.418

表 2 显示，通过这三种语料库计算出来的斯皮尔曼等级相关系数为分别为 0.412、0.413 和 0.418。证明语料库越大，词语覆盖率就会越高，计算出的词语相似度就会越准确，斯皮尔曼等级相关系数也会越高。本文后续实验都采用 10G 微博训练的词向量。

4.3.2 基于词汇分类体系的词语相似性计算

本实验采用HowNet以及《同义词词林扩展版》作为词汇分类体系，词语的相似性计算分别采用了李峰等[18]以及田久乐等[20]的方法。

利用 HowNet 和《同义词词林扩展版》计算词语相似度的参数设置分别如表 3 和 4 所示。

表 3　HowNet 相似性计算参数设置

参数	数值	参数解释
β	0.7	不带符号义原集合在整体相似度中
γ	0.02	义原（词语）和空元素之间的相似度
η	0.01	词语和义原之间的相似度
h	5	词语的“层次深度”
α	1.6	计算义原相似度的调节参数（公式 19）
δ	0.01	不在同一颗树上两个义原之间的相似度

表 4 《同义词词林扩展版》相似性计算参数设置

参数	数值	参数解释
a	0.65	分支在第二层
b	0.8	分支在第三层
c	0.9	分支在第四层
d	0.96	分支在第五层
e	0.5	仅在符号编码不一样，且符号为“#”
f	0.1	词语不在同一颗树上

首先考察这两个词汇分类体系在 PKU 500 数据集中词语的覆盖情况，表 5 所示。

表 5 不同的词汇分类体系对 PKU 500 词语的覆盖率

词汇分类体系	缺失词语数	词语覆盖率
HowNet	115	88.5%
同义词词林扩展版	42	95.8%

从表 5 可以看到，在 PKU 500 的数据集的 1000 个词中，HowNet 和《同义词词林扩展版》的词语覆盖率分别为 88.5%和 95.8%，可见，词汇分类体系的词语覆盖率还存在不足。因为这些词汇分类体系都是人工打造，要收录所有词语显得十分困难，这是词汇分类体系方法的不足之处之一。而且我们发现，不同于词向量计算得到的结果，在 HowNet 的计算结果中，相似度为 1 词语有 85 对（占 17%），在《同义词词林扩展版》中相似度为 1 的有 134 对（占 26.8%），所以相对于词向量计算词语相似性而言，人工打造的词汇分类体系词语的区分粒度不够细致，很多情况下都不能区分相似度较高的词语。进一步检验词汇分类体系计算词语相似度的效果，如表 6 所示。

表 6 词汇分类体系的词语相似性计算效果

词汇分类体系	Spearman's ρ
HowNet	0.483
同义词词林扩展版	0.481

在表 6 中，来自 HowNet 的词语相似度与人工标注的词语相似度的斯皮尔曼等级相关系数为 0.483，《同义词词林扩展版》的为 0.481。通过与上一个实验方案的斯皮尔曼等级相关系数对比，可以发现基于词汇分类体系求得的斯皮尔曼等级相关系数都比基于向量模型求出的斯皮尔曼等级相关系数高，说明尽管词汇分类体系有自身的缺点，但是利用词汇分类体系求出的词语相似度比利用词向量求出的词语相似度更能反映真实的词语语义相关性情况。

4.3.3 基于向量模型与词汇分类体系相结合的词语相似性计算

在本实验中，“w2v”代表采用 10G 微博训练词向量的计算方法，“w2v+HowNet”代表运词语向量表达构建中采用了 10G 微博训练词向量以及来自于 HowNet 的知识。“w2v +同义词词林扩展版” 代表运词语向量表达构建中采用了 10G 微博训练词向量以及来自于《同义词词林扩展版》的知识。不同方案的词语相似性计算效果如表 7 所示。

表 7 不同方案的词语相似性计算效果

计算方法	Spearman's ρ
w2v	0.418
HowNet	0.483
同义词词林扩展版	0.481
w2v+ HowNet	0.517
w2v +同义词词林扩展版	0.538

从表 7 可以看到，采用词汇分类体系对词向量进行修正的方法，实验效果比传统的两种词语相似性的计算方法都要好，说明这种利用词汇分类体系对词向量进行修正从而计算词语相似性的方法是可行的。这种方法弥补了词汇分类体系中词汇量不足的缺点，同时也补充了词向量语义表达上的欠缺。同时我们也发现《同义词词林扩展版》的修正效果在 PKU 500 数据集上比 HowNet 的修正效果好一些。如 4.2 所述，本文将效果较好的“w2v +同义词词林扩展版”方案代表 Faruqui 等人[10]方法在中文词语相似性计算的应用。

4.3.4 基于向量模型与多源词汇分类体系相结合的词语相似性计算

本实验中，“w2v+ (HowNet,同义词词林扩展版)”代表采用了HowNet和《同义词词林扩展版》两个词语语义关系的知识源。“HowNet(>0.75)”代表HowNet的知识来源只保留相似度大于0.75的近义词加入到词语的近义词集。“同义词词林扩展版(=)”表示《同义词词林扩展版》的知识来源只保留编码最后一位标记符为“=”的原子词群。“差异权重”代表对来自于HowNet、《同义词词林扩展版》以及两者的交集的近义词差异对待，考虑到4.3.3的实验中《同义词词林扩展版》的单源结合效果优于HowNet，在公式（4）的词语向量表达构建中，本文分别给词语本身w2v、仅来自于HowNet的近义词的w2v、仅来自于《同义词词林扩展版》的近义词的w2v、同时来自于HowNet和《同义词词林扩展版》的近义词的w2v设定了0.2、0.1、0.2和0.5的权重。而非差异权重的方案，则给予来自于不同词汇分类体系的近义词的w2v相同的权重。结果如表8所示。

表8 不同知识选用及权重方案的词语相似性计算效果

计算方法	Spearman's ρ
w2v+ (HowNet,同义词词林扩展版)	0.526
w2v+差异权重(HowNet,同义词词林扩展版)	0.579
w2v+ (HowNet(>0.75)+同义词词林扩展版(=))	0.579
w2v+差异权重(HowNet(>0.75),同义词词林扩展版(=))	0.618

可以看到，不做任何选取地选用 HowNet 和《同义词词林扩展版》的多源方案结果并没有优于单独采用《同义词词林扩展版》的方案（见表 7），可见探索不同类型词汇分类体系提供的知识的选用和融合问题是有价值的。而本文采用的最优方案，在 PKU 500 数据集上取得了高达 0.618 的斯皮尔曼等级相关系数。

4.3.5 研究进展方法在中文词语相似性计算上的性能对比

本文的方法与研究进展方法的对比如表 9 所示。

表9 本文的方法与研究进展方法在中文词语相似性计算上的性能对比

计算方法	Spearman's ρ
NLPCC-ICCPOL 2016 评测比赛第一名的方法[9]	0.518
Faruqui 等人[10]的方法	0.538
本文的方法	0.618

从表 9 可以看到，在 PKU 500 数据集上，以斯皮尔曼等级相关系数标准，本文的方法比 Faruqui 等人[10]的方法在中文词语相似性计算的效果提高 14.9%，比 NLPCC-ICCPOL 2016 评测比赛中第一名的方法[9]高出 19.3%。

5 结束语

本文提出一种向量模型与多源词汇分类体系相结合的词语相似性计算方法，采用多源词汇分类体系的近义词关系以及向量模型得到的词向量，计算得到词语的向量表达，并探索了不同类型词汇分类体系提供的知识的选用和融合问题，弥补了单一词向量和单一词汇分类体系在词语相似性计算中的缺点。在公开数据集 PKU 500 数据集的评测取得了 0.618 的斯皮尔曼等级相关系数，比 NLPCC-ICCPOL 2016 词语相似度评测比赛第一名的方法的结果提高了 19.3%。进一步的工作主要集中在研究更为系统的不同类型词汇分类体系提供的知识的选用和融合方案。

参考文献

[1] Wu Y.F., Li W.. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement[J]. Lecture Notes in Artificial Intelligence, 2016, 10102:828–839.

[2] Turney P. D.. Similarity of semantic relations[J]. Computational Linguistics, 2006, 32(3):379-416

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

[4] Mikolov, T., Chen, K., Corrado, G. et al. Efficient estimation of word representations in vector space[C]// Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013.

[5] Miller, G. A. WordNet: a lexical database for English[J]. Communications of the ACM, 38(11): 235-244, 1995.

[6] Dong Z.D., Dong Q.. Hownet and the computation of meaning[M]. World Scientific Publishing Company, Singapore, 2006.

[7] Li W., Liu T., Zhang Y., et al. Automated generalization of phrasal paraphrases from the web[C]// Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005: 49-56.

[8] Panchenko, A. Best of both worlds: Making word sense embeddings interpretable[C]// Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016), 2016: 2649-2655.

[9] Guo S.R.,Guan Y.,Li R. and Zhang Q..Chinese word similarity computing based on combination strategy[J]. Lecture Notes in Artificial Intelligence, 2016,10102: 744–752.

[10] Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E.,and Smith, N. A. Retrofitting word vectors to semantic lexicons[C]// Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL (NAACL 2015), 2015:1606–1615.

[11] Heylen K., Peirsmany Y., Geeraerts D., et al. Modelling word similarity: An evaluation of automatic synonymy extraction algorithms[C]// Proceedings of the Sixth International Language Resources and Evaluation, 2008, 3243-49.

[12] Landauer, T. K. and Dumais, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge[J]. Psychological Review, 1997,104(2):211–240.

[13] Baroni, M. and Zamparelli, R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). 2010:1183–1193.

[14] Sérasset G.. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf[J]. Semantic Web Journal-Special issue on Multilingual Linked Open Data, 2015, 6(4):355–361.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013), 2013b: 3111-3119.

[16] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[C]// Proceedings of the International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), 2005: 246-252.

[17] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59–76 .

[18] 李峰, 李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报, 2007, 21(3): 99-105.

[19] 梅家驹, 竺一鸣, 高蕴琦等. 同义词词林 [M] .上海:上海辞书出版社, 1993:106-108.

[20] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010,28(6): 602-608.

作者简介：

梁泳诗（1994—），硕士研究生，主要研究领域为自然语言处理。
Email: ysliang@stu.scau.edu.cn

黄沛杰（1980—），通讯作者，博士，副教授，主要研究领域为人工智能、自然语言处理、口语对话系统。
Email: pjhuang@scau.edu.cn

岑洪杰（1996—），本科，主要研究领域为自然语言处理。
Email: cenhongjie@stu.scau.edu.cn