基于可靠词汇语义约束的词语向量表达修正研究*

梁泳诗，黄沛杰，黄培松，杜泽峰

（华南农业大学数学与信息学院，广东 广州 510642）

摘要：词语向量表达（word vector representation）是众多自然语言处理（natural language processing，NLP）下游应用的基础。已有研究采用各种词汇分类体系提供的词汇语义约束，对海量语料训练得到的词向量进行修正，改善了词向量的语义表达能力。然而，人工编制或者自动构建的词汇分类体系普遍存在语义约束可靠性不稳定的问题。本文基于词汇分类体系与词向量之间、以及异构词汇分类体系之间的交互确认，研究适用于词语向量表达修正的可靠词汇语义约束提炼方法。具体上，对于词汇分类体系提供的同义词语类，基于词向量计算和评估类内词语的可靠性。在其基础上，通过剔除不可靠语义约束机制避免词语类划分潜在不够准确的词语的错误修正；通过不同词汇分类体系的交互确认恢复了部分误剔除的语义约束；并通过核心词约束传递机制避免原始词向量不够可靠的词语在词向量修正中的不良影响。本文采用 NLPCC-ICCPOL 2016 词语相似度测评比赛中的 PKU 500 数据集进行测评。在该数据集上，将本文提出的方法提炼的可靠词汇语义约束应用到两个轻量级后修正的研究进展方法，修正后的词向量都获得更好的词语相似度计算性能，取得了 0.6497 的 spearman 等级相关系数，比 NLPCC-ICCPOL 2016 词语相似度测评比赛第一名的方法的结果提高 25.4%。

关键词： 词语向量表达修正；可靠词汇语义约束；核心词约束传递

中图分类号： **文献标识码：**

Refining Word Vector Representation with Reliable Lexical Semantic Constraints

LIANG Yongshi, HUANG Peijie, HUANG Peisong, DU Zefeng

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

Abstract: Word vector representation is the basic of multiply downstream applications in natural language processing (NLP). Studies have shown that word vectors trained from large corpora gain improving representation by refining with semantic constraints in various lexical taxonomies. However, the issue about the stability of reliability is general in manual or auto-constructed lexical taxonomies. Based on lexicon-vectors interaction and the heterogeneous taxonomies' interaction, we present the method about extracting reliable lexical semantic constraints which could be applied to refine word vectors representation. In this method, the word class knowledge from lexical taxonomies will be assessed for reliability based on word vectors' calculation. In detail, the deletion of unreliable constraints can avoid the negative effect that the misclassification of word-class brought while the lexical taxonomies interaction can recover the constraints deleted by mistake. Moreover, the conductive mechanism of core words will avoid the negative effect the unreliable word vectors brought in word vectors amendment. We adopt PKU 500 for evaluation, which was used as the dataset of the NLPCC-ICCPOL 2016 shared task on Chinese word similarity measurement. The word vectors refined in the way, incorporating reliable semantic constraints extracted by our way and the two state of the art methods, outperform in the word similarity calculation. The refined word vectors achieve a Spearman score 0.6497, which gains 25.4% improvement comparing to the best result in the shared task.

Key words: word vector representation refinement, reliable lexical semantic constraints, conductive mechanism of core words

*收稿日期： 定稿日期：

基金项目：国家自然科学基金(71472068)

1 引言

词语向量表达（word vector representation）是机器翻译（Zou et al., 2013）、文本分类（Le et al., 2014）、情感分析（Socher et al., 2013）等自然语言处理（natural language processing，NLP）下游应用中的重要基础。作为词语的向量化形式，词语向量表达通过计算后能够捕捉语言的特性，因此其被用于解决各种 NLP 的任务。近年来，以分布假说——词的语义由其上下文决定（Harris,1954；Firth,1957）作为理论基础，由神经网络模型训练语言模型时生成的词分布表示（Bengio et al., 2003; Mikolov et al., 2013），又叫词嵌入（word embedding）或词向量，在许多 NLP 任务上，取得超越传统的词袋（BOW）特征表达方法的效果，这一提升归功于神经网络语言模型可以使用组合方式，以线性复杂度对复杂的 n 元上下文进行建模，解决了传统 BOW 特征表达方法高维稀疏的问题。然而，上下文不等同于真正的语义，词分布表示也存在局限性。

近年来，研究者们采用各种词汇分类体系提供的词汇语义约束（如词语间的同义关系或者反义关系），对海量语料训练得到的词向量进行修正，改善了词向量的语义表达能力。词向量的修正主要分为两种方法：一种方法是直接在训练词向量的过程中加入词汇语义约束（Bian et al., 2014; Yu and Dredze, 2014; Xu et al, 2014 ; Bollegala et al., 2016; Niu et al., 2017）。另一类方法则是对训练好的词向量根据词汇语义约束进行后处理（Faruqui et al., 2015; Mrkšić et al., 2016; Mrkšić et al., 2017; Vulić et al., 2017）。相比于前者，后者适用于任何模型训练得到的词向量的修正，并且效率较高，被称为轻量级后修正方法。然而，人工编制或者自动构建的词汇分类体系，普遍存在一些词语类划分或者词语语义构成不完善的地方，一定程度上影响了从中提取的语义约束的正确性。而训练词向量时词频稀疏的词语容易产生不可靠的原始词向量，也影响了词汇语义约束在词汇向量表达修正中的效果。针对词汇语义约束的可靠性问题，本文提出一种提炼可靠性词汇语义约束的方法。本文的方法适用于各种不同的词向量修正方法，是对词语向量表达修正研究领域的有益补充。本文的主要贡献包括：

（1）提出可靠词汇语义约束的提炼方法。本文目前针对的是同义语义约束的可靠性提炼。在给定词汇语义源（如本文采用《同义词词林扩展版》（Li et al., 2005））提供的词汇语义约束的基础上，采用词汇分类体系与词向量之间、以及异构词汇分类体系之间交互确认的方法，确定了核心约束，剔除了不可靠约束，有效地降低了错误词汇语义约束以及训练不充分的词向量在基于语义约束的词向量修正中的不良影响。

（2）在中文词语相似性评测的公开数据集 PKU 500 上进行实验。将本文提出的方法提炼的可靠词汇语义约束应用到两个轻量级后修正的研究进展方法，修正后的词向量都获得更好的词语相似度计算性能，取得了 0.6497 的 Spearman 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结果提高了 25.4%。

（3）本文的方法提炼的可靠词汇语义约束，不仅有助于词语向量表达修正方法获得更符合词汇语义约束的词向量，其核心词约束传递机制也阻隔了原始词向量不够可靠的词语在词向量修正中的信息传递，提高了修正后词汇向量表达的质量。

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做出了简要的展望。

2 相关工作

词语向量表达修正中语义信息的来源，一般是人工或者半人工的方法构建的词汇分类体系，如在英文上的 WordNet (Miller, 1995)、PPDB（Paraphrase Database） (Ganitkevitch et al., 2013)，中文上的 HowNet（Dong and Dong, 2006）、《同义词词林扩展版》（Li et al., 2005）等。

基于词汇语义约束的词向量修正主要分为两种方法：一种方法是直接在训练词向量的过程中加入词汇语义约束，使词向量在学习上下文信息的过程中保持语义约束信息。这类方法

是通过修改词向量训练的神经网络模型的目标函数实现的，如 Xu 等人（2014）提出的 RC-NET 模型，通过改造 skip-gram 模型(Mikolov et al., 2013)的目标函数把相关语义知识与分类知识注入词向量训练过程；Bollegala 等人（2016）在 Glove 模型（Pennington et al.,2014）的基础上使用同义、反义的语义约束对词向量进行训练。

另一类方法则是对训练好的词向量根据词汇语义约束进行后处理（post-processing），属于轻量级后修正的方法（Faruqui et al., 2015; Mrkšić et al., 2016; Mrkšić et al., 2017; Vulić et al., 2017）。这类方法不需要对语言模型进行修改，只需要对训练好的原始词向量进行修正，因此适用于任何语言模型的词向量，且训练效率高。Faruqui 等人（2015）提出 Retrofitting 后修正方法，利用 WordNet 和其它词汇分类体系中提取出来的同义词约束对词向量进行后修正，使语义上相似的词语相互靠近，优化词向量的质量。Mrkšić 等人（2016）提出了 Counter-fitting 的后修正方法，在 PPDB（Ganitkevitch et al., 2013）和 WordNet（Miller, 1995）中提取同义约束和反义约束，修正过程中使具有同义关系的词向量拉近、具有反义关系的词向量相互进行推远，通过推拉机制获得表达能力更强的词向量。在其基础上，Mrkšić 等人（2017）和 Vulić 等人（2017）进一步研究了推拉机制中新的词汇语义约束的使用方法以及研究跨语言词向量的生成，获取高质量的跨语言词向量。

上述的基于词汇语义约束的词向量修正方法，都在一定程度上改善了词向量的语义表达能力。然而，由于词汇语义约束一般是来自于人工编制或者自动构建的词汇分类体系，其可靠性不稳定。此外，对于相对正确的词汇语义约束，如果约束包含的词语在训练原始词向量时词频稀疏，依然容易产生不可靠的原始词向量，也会影响词汇语义约束在词汇向量表达修正中的效果。针对词汇语义约束的可靠性问题，本文提出一种提炼可靠性词汇语义约束的方法。值得注意的是，虽然本文只将提出的方法提炼到的可靠词汇语义约束应用在两个轻量级后修正的研究进展方法上，但本文的方法适用于各种不同的词向量修正方法。

3 基于可靠词汇语义约束的词语向量表达修正

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。加粗部分是本文的主要工作。

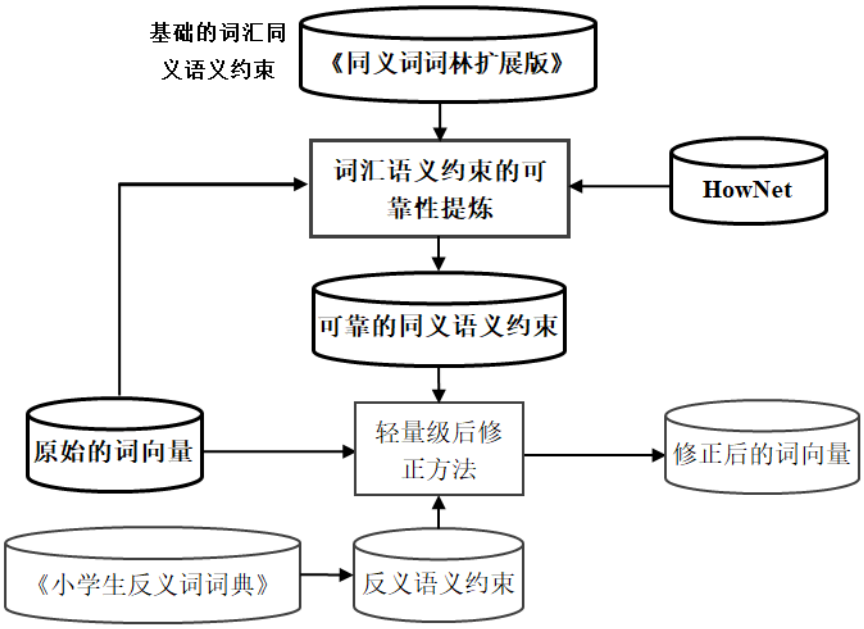

图 1　总体技术架构

包括两个主要部分：

（1）可靠同义语义约束的提炼。首先利用词汇同义语义源（如本文采用的《同义词词林扩展版》）获得大量的词汇语义约束（也即是同义词语类，具体上是《同义词词林扩展版》中的原子词群）；然后利用词向量与词汇语义约束进行交互确认，计算和评估类内词语的可靠性，将词语类中的词语按一定比例划分成核心词（提供可靠的词向量修正来源的词语）、非核心可靠词（潜在的原始词向量不够可靠的词语）、不可靠词（潜在的词语类划分不够准确的词语）；最后，针对划分到不可靠词的词语，通过不同词汇分类体系（本文中是《同义词词林扩展版》与 HowNet）的交互确认，恢复了一部分词语到非核心可靠词中，也即是恢复了部分误剔除的语义约束。由最终确定的核心词和非核心可靠词共同构成了可靠的词汇语义约束。

（2）词语向量表达修正阶段。根据可靠的同义语义约束、反义语义约束（本文的反义语义源采用了《小学生反义词词典》），以及原始词向量，采用轻量级后修正方法得到修正后的词语向量表达。

3.2 词语向量表达的轻量级后修正方法

在研究进展中，词向量的轻量级后修正方法的代表包括 Retrofitting（Faruqui et al., 2015）和 Counter-fitting（Mrkšić et al., 2016）。两个方法的区别主要在于：在计算两个词向量的距离时 Retrofitting 采用的是欧式距离， Counter-fitting 采用的是余弦距离。Counter-fitting 比 Retrofitting 多采用了反义语义约束，尽管效果并不明显。

在本文实验中发现 Counter-fitting 的效果优于 Retrofitting，下面简要介绍一下 Counter-fitting 的轻量级后修正方法，Retrofitting 方法可见文献（Faruqui et al., 2015）。

通过输入初始词向量集 $V=\{v_1,v_2,...,v_N\}$，模型会自动把词汇语义注入到初始词向量的空间中，形成一组新的词向量集 $V'=\{v'_1,v'_2,...,v'_N\}$，而被注入的同义和反义关系以同义词词对和反义词词对的形式组成同义词约束集 S 以及反义词约束集 A。模型的目标函数包括三个部分。

（1）同义词拉近：将同义词约束 $(i,j)\in S$ 指向的词对词向量 v_i、v_j 尽可能地拉近：

$$SA(V')=\sum_{(u,w)\in S}\tau(d(v'_u,v'_w)-\gamma) \quad (1)$$

其中，$d(v_u,v_w)$ 来自于余弦相似度，$d(v_i,v_j)=1-\cos(v_i,v_j)$；$\tau(x)\triangleq\max(0,x)$ 是损失函数，γ 定义了同义词词向量间理想的距离，文献（Mrkšić et al., 2016）中设置 γ=0，使得单因素优化的目标是 v_i、v_j 的词向量相同。

（2）反义词推远：将反义词约束 $(i,j)\in A$ 指向的词对词向量 v_i、v_j 尽可能地推开：

$$AR(V')=\sum_{(u,w)\in A}\tau(\delta-d(v'_u,v'_w)) \quad (2)$$

其中，$d(v_u,v_w)$ 来自于余弦相似度，$d(v_i,v_j)=1-\cos(v_i,v_j)$；$\tau(x)\triangleq\max(0,x)$ 是损失函数，δ 定义了反义词词向量间理想的距离，文献（Mrkšić et al., 2016）中设置 δ=1.0，使得单因素优化的目标是 v_i、v_j 的词向量正交。

（3）保持自身信息：使词向量一定程度上也保持初始词向量的值，一定程度上可避免不完全同义或者反义的语义约束的过度拉和推：

$$SR(V,V')=\sum_{i=1}^{N}\tau(d(v_i,v'_i)-\lambda) \quad (3)$$

其中，$\tau(x)\triangleq\max(0,x)$ 是损失函数，λ代表新词向量与原词向量之间的理想距离。我

们设置 λ=0，确保在加入语义约束的同时保持词向量本身所具有的信息与价值。

综上，轻量级后修正模型的总体目标函数为：

$$C(V,V') = k_1 SA(V') + k_2 AR(V') + k_3 SR(V,V') \quad (4)$$

其中，k_1、k_2 和 k_3 用于控制不同部分损失之间的权重。$k_1,k_2,k_3 > 0$ 且 $k_1 + k_2 + k_3 = 1$。本文在 k_1、k_2 和 k_3 上使用 grid search，并用随机梯度下降法（stochastic gradient descent）对目标函数的最小值进行求解，迭代 20 次，选取 Spearman 相关系数在 PKU 500 中的最高分数的模型超参数。

3.3 可靠词汇语义约束的提炼

人工编制或者自动构建的词汇分类体系普遍存在语义约束可靠性不稳定的问题，通过分析本文语义约束基础来源的《同义词词林扩展版》的同义词语类，以及相应词语的原始词向量，发现问题主要来自两个方面：一是由于现实生活中语义是渐变和存在多义，人很难精准的判定词语之间的语义界限，所以对于人工构造的同义词语类很容易会出现划分潜在不够准确的词语（如词语“分解”划分到了“解释”、“说明”等词语所在的词语类）。二是由于训练词向量的语料中，不同词语的词频存在差异，词频稀疏的词语容易产生不可靠的原始词向量，这些词语就算是词语类划分正确，在采用词语类词汇语义约束进行类内词语的词向量修正时也会带来不良影响。

基于上述分析，本文提出的词汇语义约束的可靠性提炼机制包括以下几个主要方面：

（1）基于语义约束源《同义词词林扩展版》提取同义语义约束（词语类），并基于词向量计算和评估类内词语的可靠性；

（2）通过剔除不可靠语义约束机制避免词语类划分潜在不够准确的词语的错误修正；

（3）通过不同词汇分类体系（本文中为 HowNet）的交互确认恢复了部分误剔除的语义约束。

（4）通过核心词约束传递机制避免原始词向量不够可靠的词语在词向量修正中的不良影响；

3.3.1 基础的词汇同义语义约束的提取

《同义词词林扩展版》是比较著名的中文词汇分类体系，它是哈尔滨工业大学信息检索实验室在《同义词词林》（梅家驹等，1983）基础上修正与扩充而成（Li et al., 2005）。《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，是一部同义词语类词典，例子如图 2 所示。图中每一行为一个原子词群，提供具有同义语义的词语类。

```
Aa01C01= 众人 人人 人们
Aa01C02= 人丛 人群 人海 人流 人潮
Aa01C03= 大家 大伙儿 大家伙儿 大伙 一班人 众家 各户
Aa01C04= 们 辈 曹 等
Aa01C05@ 众学生
Aa01C06# 妇孺 父老兄弟 男女老少 男女老幼
```

图 2　《同义词词林扩展版》示例图

《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。每一行都有自身所属的编码，在《同义词词林扩展版》中

词语的相似性是根据每一行的编码计算得到的，编码的最后一位有三种符号：“=”、“#”、“@”，用于说明同一个原子词群中的词语关系，分别代表语义相等、语义相关（同行词语是同类，但不能视为相等）、独立（表示在词典中该词语既没有同义词也没有相关词）。本文以最后一位标记符号为“=”的语义相等原子词群提供基础的同义语义约束。

3.3.2 同义词语类内词语的可靠性评估

本文基于词向量与词汇语义约束的交互确认，对《同义词词林扩展版》中获得的每个同义原子词群的词语类中的词语进行类内评估，算法流程如图 3 所示。

算法：Synonym-evaluation. 给定一个原子词群，评估类内词语的可靠性与核心性。
输入：一个 “=”的原子词群 W，即$(W_1,W_2,\cdots,W_n)$及其对应的词向量 V，即$(V_1,V_2,\cdots,V_n)$。
输出：一个“=”的原子词群中的核心同义词集 C，及可靠同义词集 R
方法：

(1) $count$=0;sum=0;
(2) for 原子词群中的每个词向量 V_i
(3) $sum \leftarrow sum+V_i$
(4) $count \leftarrow count+1$
(5) end for
(6) $middle\_Vec \leftarrow sum/count$
(7) 创建数组 *all_measure*
(8) for 原子词群中的每个词 W_i 词向量 V_i
(9) 计算词向量 V_i 与中心词向量 *middle_Vec* 的欧式距离 d_i
(10) 将(W_i, d_i)插入数组 *all_measure*
(11) end for
(12) 对数组 *all_measure* 中的元素(W_i, d_i)根据 d_i 进行从小到大排序
(13) 将排序后的数组 *all_measure* 前 α%的元素插入 C
(14) 将排序后的数组 *all_measure* 前 β%的元素插入 R（α<β）

图 3　同义词语类的类内评估算法

在图 3 的算法中，（1）-（6）步采用原子词群中全体词语的词向量计算出代表该原子词群中心的中心词向量。（8）-（12）步分别计算该原子词群中每个词语的词向量与中心词向量的距离，并由小到大排序。离中心词向量越近的同义词的向量表达越能反映该原子词群的类语义，而远离中心词向量的词语则可能是人工错误划分的词语（与该词语类中其它词语不够同义），或者也可能是词向量训练时词频稀疏导致的向量表达不好。算法第（13）和（14）步分别根据一定的阈值从话语类中划分出核心词和可靠词。考虑到《同义词词林扩展版》具有较好的质量，尽管存在一定的噪声，本文实验中 α%和 β%分别设为 60%和 85%。

根据算法的输出，可以进一步将每个原子词群的词语类中的词语按一定比例划分成：

- 核心词：距离词语类中心最近的 α%的词语，一般有着正确词语类归属，以及训练良好的词向量，因此代表着提供可靠的词向量修正来源的词语。
- 非核心可靠词：距离词语类中心 α%之外，但在 β%之内的词语，仍然较大可能是正确词语类归属，但潜在存在原始词向量不够可靠的问题。因此，在根据语义约束修正词向量时，我们通过核心词约束传递机制限制了非核心可靠词的词向量在修正中的影响。
- 不可靠词：距离词语类中心 β%之外的词语，较大概率属于潜在的词语类划分不够准确的词语。这类词语被判定为不可靠词，从同义语义约束中剔除。

下面，我们以“Hg12A01=”原子词群为例介绍词语类的划分，如图 4 所示。

Hg12A01=1 解释 2 讲 3 说明 4 释疑 5 说 6 诠 7 训诂 8 解说 9 诠释 10 分解 11 释

图 4 “Hg12A01=”原子词群词语排序

图 4 中给出了该原子词群中每个词语的词向量与中心词向量的距离排序情况（缺失词向量的词语没有展示）。从图 4 可以看到，整体上，原子词群中的词语随着与中心词向量的距离增大，其词语与词语类语义的关联强度也在逐步弱化。排序的前 5 名，分别是“解释”、“讲”、“说明”、“释疑”、“说”，都表达着说明、阐明的核心概念，其语义紧密相关，但是离中心向量较远的词语如“分解”，阐明、说等的语义概念不在其常用语义中，“分解”属于潜在的误划分词语。实际上，在 PKU 500 数据集中，词语对（“解释”，“分解”）的人工标注的相似度也只有 1.7 （范围为 0-10，10 代表最相似）。当然，也确实存在一些词语属于词语类归属正确，但是因为训练词向量时的词频过于稀疏导致词向量不准确而远离中心词向量，如上面例子中的“释”，这属于误剔除词语，我们将通过恢复误剔除词语机制“恢复”它们进语义约束。

3.3.3 基于异构词汇分类体系交互确认的误剔除词语恢复

一般可以认为，被不同的词汇分类体系同时认为是同义关系较大概率是可靠的同义约束。本文通过异构词汇分类体系（HowNet 与《同义词词林扩展版》）的交互确认恢复了部分误剔除的语义约束。

HowNet（Dong and Dong, 2006）是一部用一个或者多个“义原”去描述词语概念的中文词汇分类体系。“义原”是描述概念的最基本单位，不同的义原集合表述不同的概念，HowNet 中的词语有一个或多个概念（刘群等, 2002; 李峰等, 2007）。如图 5，为词语“家庭”在 HowNet 中的表述。

NO.=041970 W_C=家庭 G_C=N E_C= W_E=family G_E=N E_E= DEF=community|团体,family|家

图 5 HowNet 中词语示例图

词语“家庭”的概念是 DEF=community|团体,family|家，团体、家就是组成概念的义原。

HowNet 中的义原有 1600 多个（李峰等, 2007），HowNet 中的词语概念由这些义原的组合进行描述。义原之间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，进而逐步求出词语间概念的相似度以及词语间的相似度，本文就是利用李峰（2007）关于 HowNet 的词语相似性的计算方法计算出 HowNet 的关于每个词语的近义词词集。

经过第前面的词语类内词语的可靠性评估，一个原子词群上的词语被分为三类：核心词部分（设为$\{W_1,W_2,...,W_i\}$）、非核心可靠词（设为$\{W_{i+1},W_{i+2},...,W_t\}$）、以及排序在β%之后的准备剔除的不可靠词$\{W_{t+1},W_{t+2},...,W_n\}$。对于候选剔除词$\{W_{t+1},W_{t+2},...,W_n\}$，本文利用 HowNet 与《同义词词林扩展版》的交互确认，对其中的部分词进行恢复：

$$(H_1 \cup H_2 \cup ... \cup H_i) \cap \{W_{t+1}, W_{t+2}, ..., W_n\} \quad (5)$$

其中，$H_j = \{w | HowNet$近义词词典中关于W_j 的近义词$\}$。

该原子词群中的候选剔除词中，属于核心词在 HowNet 中的近义词词集的并集的词语将被恢复到非核心可靠词中。如图 4 列举的“Hg12A01=”原子词群为例，候选剔除词{“分解”，“释”}，“释”被恢复成非核心可靠词，而“分解”因为在 HowNet 角度也没有被视为是核心词的同义词而被最终删除。

3.3.4 核心词的约束传递机制

在留下的可靠词中，相比于非核心可靠词，核心词一般既有着正确的词语类归属，又具有训练良好的原始词向量，因此代表着提供可靠的词向量修正来源的词语。图 6 展示了核心词约束传递机制示意图。

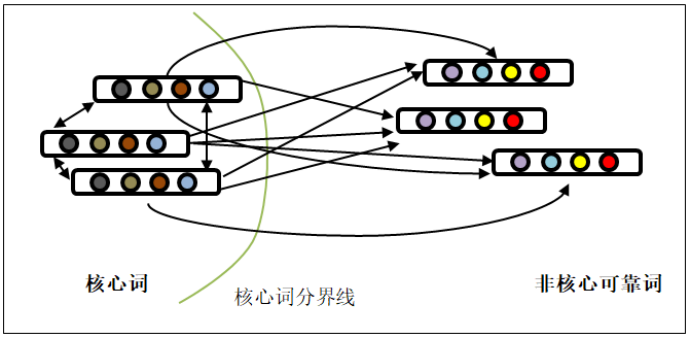

图 6　核心词约束传递机制示意图

在核心词分界线右边的为靠近中心词向量的核心词，在核心词分界线右边的为经过提炼的非核心可靠词，箭头的指向为约束引导下的词向量修正时的约束作用方向。核心词的约束传递机制使得核心词的词向量修正时，只接受核心词的语义约束，而“阻止”了来自非核心可靠词的词向量影响；非核心可靠词修正时，同样只接受核心词的语义约束引导下来自于核心词的词向量影响。这个机制下可预期能使得词语修正后的词向量，朝着核心词原始词向量的方向靠拢，远离非核心可靠词原始词向量，获得质量较好的词语向量表达。

4 实验

4.1 实验数据集

本文采用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)中的 10G 微博作为原始词向量的训练语料库。运用 word2vec 的 CBOW 模型（Mikolov et al., 2013）在语料上进行词向量的训练，获得 400 维的词向量。

可靠同义词约束的提炼运用到以下两个词汇分类体系：

（1）《同义词词林扩展版》：来自于哈工大信息检索研究室的《同义词词林扩展版》最大的特点是它是一部类语词典，同一词语类之间通过相互组合可以获得多对同义词词对约束，并且质量较好。本文以最后一位标记符号为“=”的原子词群作为基础的同义语义约束来源。

（2）HowNet：用义原标记每个词语概念的中文知识库。本文用李峰等（2007）关于 HowNet 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，对于每一个词语，与之相似度最高的词语组织起来，得到每个词语的最高相似度词集，也视为该词语的近义词词集。词语最高相似度达到 0.75 的近义词词集组合成 HowNet 的近义词词典，用于对误剔除词语的恢复。

反义词约束来源：由于在中文语言中，还没有覆盖比较全面的反义词语义信息的语义词典，本文在 Counter-fitting 相关实验中，选择从《小学生反义词词典》中抽取反义词约束。

利用 HowNet 计算词语相似度的参数设置如表 1。

表 1　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的权重
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
δ	0.01	不在同一颗树上两个义原之间的相似度

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（Wu et al., 2016）。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）的词语相似度计算评测比赛中。

4.2 实验设置

本文采用斯皮尔曼等级相关系数（Spearman rank correlation coefficient）去衡量词向量计算词语相似性的效果。通过计算 PKU 500 中人工标注的分数列和词向量计算词语相似性的列之间的 Spearman 等级相关系数 ρ，借以判断各实验方案对词语相似性的计算效果。本文的实验方案围绕词语向量表达修正过程中运用的可靠词汇语义约束的效果，以及核心词约束传递机制的价值两个方面展开，实验方案如下：

（1）研究进展方法中中文词语相似性上的性能对比：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。

（2）典型词对在可靠词汇语义约束下的探讨：通过典型词对的观察，展示本文的方法可靠词汇语义约束的保持和不可靠词汇语义约束的去除情况。

（3）核心词语义传递对词汇向量表达质量的影响：通过典型词对的观察，展示本文采用核心词约束传递机制对词汇向量表达质量的影响。

4.3 实验结果与分析

4.3.1 研究进展方法中中文词语相似性上的性能对比

本实验中基线方法 Retrofitting 及 Counter-fitting 采用的同义词汇语义约束为来自于《同义词词林扩展版》中标记有“=”的原子词群，另外 Counter-fitting 中采用的反义词词汇语义来自于《小学生反义词词典》。研究进展方法在中文词语相似性上的性能对比如表 2 所示。“Core”代表采用了核心词约束传递机制（本文以靠近每个原子词群中心向量前 60%的词语作为核心词）；“DEL”代表采用了剔除不可靠语义约束机制，即对离中心词向量最远的 15%的词语判为不可靠词，不参加词语向量表达修正；“REC”代表恢复了部分误剔除的语义约束机制。NLPCC-ICCPOL 2016 测评比赛第一名的方法的结果直接采用了其文献中的结果。

从表 2 中可以看到，基线方法 Retrofitting 及 Counter-fitting 在 PKU 500 上分数都超过 60，比原始词向量的 41.8 高出 20，表明了轻量级后修正方法的良好效果；本文的方法取得最好的 PKU 500 分数，比 NLPCC-ICCPOL 2016 测评比赛第一名的方法高出 25.4%。将本文提出的方法中的“Core“、“DEL”、“REC” 逐步应用到 Retrofitting 及 Counter-fitting 上后，

PKU 500 的分数都得到逐步提升，累计分别提高了 3.63 和 0.84。从提升数值上看不是太显著，主要是因为本文的方法目标是有效地减少错误词汇语义约束，以及降低训练不充分的词向量在基于语义约束的词向量修正中的不良影响，这部分词语在语义约束中占的只是有限的比例。在下面的实验中，我们将进一步展示具体的改善效果。

表 2　研究进展方法在中文词语相似性上的性能对比

对比方案		ρ *100
原始词向量		41.8
基线方法	NLPCC-ICCPOL 2016测评比赛第一名的方法（Guo et al., 2016）	51.8
	Retrofitting （Faruqui et al.,2015）	61.20
	Counter-fitting （Mrkšić et al., 2016）	64.13
本文的方法	Retrofitting+Core	64.42
	Retrofitting+Core+DEL	64.76
	Retrofitting+Core+DEL+REC	64.83
	Counter-fitting+Core	64.61
	Counter-fitting+Core+DEL	64.74
	Counter-fitting+Core+DEL+REC	64.97
	Counter-fitting+Core+DEL+REC -Antonyms	64.91

此外，我们观察了反义语义约束的价值，“Counter-fitting+Core+DEL+REC -Antonyms”代表在“Counter-fitting+Core+DEL+REC”的基础上去除了反义约束。从结果可以看到，影响非常微弱，其主要原因可能是词语的反义语义约束覆盖面过小（相比于来自于《同义词词林扩展版》的同义约束）。实际上，目前在英文的词汇向量表达修正的研究中，反义约束的应用效果也不太理想（Bollegala et al., 2016; Mrkšić et al., 2016）。这意味着在语料库建设方面，急需构建足够覆盖面的反义词典，尤其是提供中文反义语义约束的反义词典。

4.3.2 典型词对的词汇语义约束的分析

我们进一步通过 PKU 500 中典型词对的观察，展示本文的方法可靠词汇语义约束的保持和不可靠词汇语义约束的去除情况。表 3 上显示了 PKU 500 中的 17 对高误差词对（原始词向量词对相似度计算排序在 300 名之后，人工评分分数排名在前 100 名）的修正情况。

表 3 PKU 500 中基于原始词向量计算的高误差词对

（修正后词向量词语相似度计算排序进入前 200 名，用“√”表示）

高误差词对	Counter-fitting	Counter-fitting +Core	Counter-fitting+ Core+DEL	Counter-fitting+ Core+DEL+REC
出神，发楞	√	√	√	√
几乎，差点儿	√	√	√	√
闯祸，惹是生非	√	√	√	√
清晰，一清二楚	√	√	√	√
解闷，排遣	√	√	√	√
私语，嘀咕		√	√	√
在行，熟练	√	√	√	√
种，栽				
支柱，骨干	√	√	√	√
犹豫，踌躇不前	√	√	√	√

麻利，灵活	√	√	√	√
轻蔑，藐视	√	√	√	√
服气，心服口服	√	√	√	√
凌乱，乱七八糟	√	√	√	√
工作，干活儿	√	√		
拍马屁，谄谀	√	√	√	√
赢，胜	√	√	√	√

在“Counter-fitting”、“Counter-fitting+Core”、“Counter-fitting+Core+DEL”和“Counter-fitting+Core+DEL+REC”方案中，被有效修正（本文以从后200名前进到前200名为标准）的高误差词对都是15对。由此可以看出，“Counter-fitting+Core+DEL”及“Counter-fitting+Core+DEL+REC”方案在采用了不可靠词剔除级之后，依然能保证有效约束的实现。

表4显示了PKU 500中的11对相对低误差词对（原始词向量词对相似度计算排序在300名之后，人工评分分数排名也在300名之后）是否能不被不可靠约束影响的情况。

表4 PKU 500中基于原始词向量的相对低误差词对
（修正后词向量词语相似度计算排序进入前200名，用“×”表示）

低误差词对	Counter-fitting	Counter-fitting +Core	Counter-fitting+ Core+DEL	Counter-fitting +Core+DEL+REC
亏，幸亏		×	×	×
琢磨，镂刻				
崛起，凸起	×	×		
榜样，样子				
僻静，冷静	×	×		×
消极，四大皆空	×	×	×	×
爱怜，同病相怜	×	×		
功夫，素养	×	×		×
公司，代销店	×	×		
解释，分解	×	×		
轻，善				

在“Counter-fitting”方案及在“Counter-fitting+Core”方案中，分别有7对、8对词语对被错误修正，而在“Counter-fitting+Core+DEL”和“Counter-fitting+Core+DEL+REC”方案中，仅有2对和4对词语被错误修正。由此可以看出，不可靠词的剔除机制可以有效地从同义约束中剔除部分不可靠词语地影响。

4.3.2 核心词约束传递机制对词汇向量表达质量的影响

最后我们继续通过典型词对的观察，展示具体词对在采用核心词约束传递机制对词汇向量表达质量的影响。我们选取在表3中均被“Counter-fitting”、“Counter-fitting +Core”两个方案有效修正的词对进行观察。

每个词对的词语均来自于同一个“=”号原子词群。为了观察词汇向量表达的修正质量，我们给每个词对所在原子词群都定义一个核心词代表和非核心词代表。例如图7所示的词对（“出神”、“发愣”）所在的原子词群“Ic04B01=”中，设定第一位为核心词代表（本例中为“愣住”），最后一位为非核心词代表（本例中为“泥塑木雕”）。核心代表词能反映该词所在原子词群的语义核心，而非核心代表词明显地与原子词群的语义核心出现了偏离，所以从词语向量表达修正的角度，词语更应该往该原子词群中的核心词靠近，而远离非核心词（甚至是边缘词）。

Ic04B01= 1 愣住 2 愣 3 目瞪口呆 4 傻眼 5 呆 6 愣神 7 发愣 8 张口结舌 9 瞠目结舌 10 干瞪眼 11 发呆 12 发傻 13 出神 14 直勾勾 15 眼睁睁 16 木然 17 呆若木鸡 18 发楞 19 愣神儿 20 泥塑木雕

图 7 “Ic04B01=”原子词群词语排序

我们以“与核心代表词的相似度”表示与核心词代表的原始词向量的相似度，“与非核心代表词的相似度” 表示与非核心词代表的原始词向量的相似度。通过计算可以得到，采用核心词约束传递机制（Counter-fitting+Core）修正的词向量和没有采用核心词约束传递机制（Counter-fitting）修正的词向量，与核心词代表词和非核心词代表词的原始词向量的距离关系，结果如表 5 所示。

表5 修正后词对词向量与核心词代表、非核心词代表原词向量的距离情况

词对	词对所属词语类的代表词		Counter-fitting			Counter-fitting +Core		
	核心代表词	非核心代表词	词对相似度	与核心代表词的相似度	与非核心代表词的相似度	词对相似度	与核心代表词的相似度	与非核心代表词的相似度
出神，发楞	愣住	泥塑木雕	0.572	0.397,0.331	0.200,0.1371	0.539	0.451,0.383	0.114,0.046
几乎，差点儿	差一点	殆	0.552	0.388,0.541	0.234,0.146	0.581	0.438,0.548	0.066,0.109
闯祸，惹是生非	出事	肇祸	0.655	0.593,0.432	0.341,0.212	0.572	0.621,0.451	0.215,0.184
清晰，一清二楚	明晰	冥	0.675	0.704,0.428	0.131,0.172	0.723	0.706,0.455	0.035,0.067
解闷，排遣	消遣	消	0.594	0.589,,0.372	0.316,0.411	0.592	0.626,0.375	0.207,0.388
在行，熟练	纯熟	稳练	0.707	0.592,0.739	0.039,0.056	0.739	0.611,0.750	0.004,0.022
支柱，骨干	中坚	基于	0.591	0.455,0.545	0.131,0.168	0.599	0.459,0.533	0.087,0.127
犹豫，踌躇不前	踌躇	犹豫不前	0.752	0.656,0.684	0.260,0.291	0.776	0.670,0.695	0.216,0.244
麻利，灵活	灵活	眼疾	0.560	0.456,0.942	0.085,0.083	0.574	0.482,0.941	0.052,0.062
轻蔑，藐视	唾弃	轻	0.697	0.549,0.542	0.115,0.152	0.715	0.594,0.585	0.035,0.073
服气，心服口服	信服	伏	0.795	0.548,0.607	0.121,0.155	0.844	0.595,0.644	-0.002,0.028
凌乱，乱七八糟	杂乱	纷	0.682	0.607,0.619	0.045,0.054	0.702	0.631,0.635	0.031,0.043
工作，干活儿	做事	视事	0.597	0.552,0.516	0.054,-0.078	0.618	0.559,0.571	0.066,-0.064
拍马屁，谄谀	溜须拍马	买好	0.683	0.689,0.599	0.010,-0.001	0.719	0.719,0.647	-0.059,-0.079
赢，胜	旗开得胜	捷	0.644	0.518,0.501	0.108,0.131	0.664	0.527,0.499	0.062,0.100

从表 5 中可以看到，Counter-fitting+Core 和 Counter-fitting 都可以取得相近的词语语义约束修正效果。两种方法修正后的词对的词向量相似度均值分别达到了 0.650 和 0.664。

然而，核心词约束传递机制的优势在于使得修正的词向量往核心代表词靠拢。从表 5 可以看到，Counter-fitting+Core 修正后的词对词向量，普遍都要比 Counter-fitting 修正后的词对词向量更靠近核心代表词的原始词向量（与核心词代表的原始词向量的相似度高），以及更远离非核心词代表词的原始词向量（与非核心词代表的原始词向量的相似度低）。

例如第一对词，“出神”在“Counter-fitting +Core”中与核心代表词相似度为 0.451，相似度比“出神”在“Counter-fitting”中与核心代表词的相似度（0.397）大；“发楞”在“Counter-fitting +Core”中与核心代表词相似度为 0.383，相似度也比“发楞”在“Counter-fitting”中与核心代表词的相似度（0.331）大。而“出神”在“Counter-fitting +Core”中与非核心代表词相似

度为 0.114，相似度比“出神”在“Counter-fitting”中与非核心代表词的相似度（0.200）小；“发楞”在“Counter-fitting +Core”中与非核心代表词相似度为 0.046，相似度也比“发楞”在“Counter-fitting”中与非核心代表词的相似度（0.137）小。

由此我们可以得到这样的观察结论：采用核心词约束传递机制的词语向量表达修正，可以在有效保持同义语义约束效果的条件下，获得质量更好的词语向量表达。

5 总结

本文研究了适用于词语向量表达修正的可靠词汇语义约束提炼方法。通过基于词汇分类体系与词向量之间、以及异构词汇分类体系之间的交互确认，对《同义词词林扩展版》提供的基础同义语义约束进行可靠性评估和可靠词选取。提炼得到的可靠词汇语义约束应用到两个轻量级后修正的研究进展方法，修正后的词向量都获得更好的词语相似度计算性能，在公开数据集 PKU 500 数据集的评测上，比 NLPCC-ICCPOL 2016 词语相似度测评比赛第一名的方法的结果提高 25.4%。此外，核心词约束传递机制有效地帮助词语向量表达修正方法在有效实现同义语义约束效果的条件下，获得质量更好的词语向量表达。

未来进一步的工作主要集中在进一步完善同义语义约束的可靠性提炼方法（比如通过评估同义词语类的质量，有差异性地对待不同词语类的可靠词划分标准），以及尝试人工或半人工的方式构建较好覆盖面的反义词典，并研究反义语义约束的可靠性提炼方法。

参考文献

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 .

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.

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, 1983:106-108.

Bian J, Gao B, and Liu T. Knowledge-powered deep learning for word embedding[C]// Proceedings of the 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2014), 2014, 8724 :132-148.

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

Bollegala D, Mohammed A, Maehara T, et al. Joint word representation learning using a corpus and a semantic lexicon[C]// Proceedings of the 13th AAAI Conference on Artificial Intelligence(AAAI 2016), 2016, 2690-2696.

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

Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons[C]// Proceedings of the 11th Annual Conference of the North American Chapter of the ACL (NAACL 2015), 2015:1606–1615.

Firth J R. A synopsis of linguistic theory[J]. Studies in Linguistic Analysis, 1957:1930-1955.

Ganitkevitch J and Burch C C. The multilingual paraphrase database[C]// Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), 2014:4276–4283.

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

Harris Z S. Distributional structure[J]. Word, 1954:146-62.

Li W, Liu T, Zhang Y, et al. Automated generalization of phrasal paraphrases from the web[C]// Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005), 2005: 49-56.

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

Mrkšić N, Séaghdha D Ó, Thomson B, et al. Counter-fitting word vectors to linguistic constraints[C]// Proceedings of the 12th Annual Conference of the North American Chapter of the ACL(NAACL 2016), 2016:142–148.

Mrkšić N, Vulić I, Séaghdha D Ó, et al. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 2017:309-324.

Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of 1st the International Conference on Learning Representations (ICLR 2013), 2013.

Niu Y, Xie R, Liu Z, et al. Improved word representation learning with sememes[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 2017:2049-2058.

Pennington J, Socher R and Manning C D. Glove: global vectors for word representation[C]// Proceedings of the 20th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014:1532 – 1543.

Le Q and Mikolov T. Distributed representations of sentences and documents[C]// Proceedings of the 31st International Conference on Machine Learning (ICML 2014), 2014.

Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 2013:1631–1642.

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.

Xu C, Bai Y, Bian J, et al. RC-NET: A general framework for incorporating knowledge into word representations[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM 2014), 2014:1219-1228.

Vulić I, Mrkšić N, Reichart R, et al. Morph-fitting: fine-tuning word vector spaces with simple language-specific rules[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 2017:56-68.

Yu M and Dredze M. Improving lexical embeddings with semantic knowledge[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014),2014, 1(1): 545-550.

Zou W Y, Socher R, Cer D, et al. Bilingual word embeddings for phrase-based machine translation[C]// Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 2013:1393–1398.

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

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