北京大学

本科生毕业论文

题目：基于矩阵分解和矩阵变换的多义词向量研究

On Multi-Sense Word Embeddings via Matrix Factorization and Transformation

姓　　名： 石昊悦

学　　号： 1300012756

院　　系： 信息科学技术学院

本科专业： 智能科学与技术系

指导导师： 胡俊峰 副教授

二〇一八 年 五 月

版权声明

摘要

多义词向量是一种重要且符合人类直觉的词义表示模式，它根据上下文对每个不同的词义学习出一个向量表示。然而，这种模型经常会对上下文信息过度敏感从而产生伪多义现象：即将同一词义学习出多个词向量表示。

本工作着重讨论了已训练出的多义词向量中进行伪多义检测和消除的问题。本文提出了三个伪多义检测算法：(1) 基于外部知识库的伪多义检测算法；(2) 基于邻域相似度的伪多义检测算法；(3) 基于主成分分析的伪多义检测算法。接着，本文通过一个线性变换说明了伪多义检测和消除对于理解词义有一定的帮助。此外，消除伪多义有助于增强词向量在下游任务上的表现。

关键词：词向量，多义词向量，矩阵分解，矩阵变换

On Multi-Sense Word Embeddings via Matrix Factorization and Transformation

Haoyue Shi (Department of Intelligence Science and Technology)
Directed by Prof. Junfeng Hu

Abstract

Multi-sense word embeddings, which learn a vector for each sense, represent the concept of sense with vectors intuitively. However, such embeddings learned without supervision generate many pseudo multi-senses, as such methods are overly sensitive to contextual variations.

In this work, we introduce three different algorithms for pseudo multi-sense detection from different perspectives of view, termed (1) external knowledge based pseudo multi-sense detection, (2) neighborhood similarity based pseudo multi-sense detection, and (3) principal component analysis based pseudo multi-sense detection. We test our algorithms with transformation matrix based elimination algorithms and downstream tasks. We show that eliminating the effect of pseudo multi-sense could lead to better performance to represent senses, as well as better performance of the embeddings on a lot of downstream tasks.

Key Words : word embeddings, multi-sense word embeddings, matrix factorization, matrix transformation

目录

第一章 绪论

词向量 (Word Embeddings, Word Vectors) 是当代计算机科学对语义表示的最重要贡献之一。这种模型基于分布式语义假设，将语料中的每个词映射为一定长向量。所有词所对应的向量所张成的向量空间可以被用作语义的分布式语义表征 [2, 6, 20, 27–29, 32, 36, 41]。一般情况下，该空间中对应向量的距离或余弦夹角可以反应两个词的相似程度：两个词在该空间中向量距离越小或夹角余弦值越大，则相似程度越大。

然而，每个词仅映射为一个向量的表示方法存在一定缺陷。许多语言都存在或多或少的多义词。尽管在现代汉语中这一点表现得不够明显，但如表 1.1所示，古代汉语以及英语等语言中的一词多义现象非常普遍。每个词对应单一词向量的表示方式并不能很好地表示这一点。因此，分布式多义词表示（即多义词向量）的研究也成为了一个分布式语义学当中的重要问题。

语言	词	不同词义
古代汉语	之	(1) 去, 往 (2) 代词 (3) 的
古代汉语	然	(1) 正确，是 (2) 然而 (3) 语气词 (4) 样子
英语	bank	(1) 银行 (2) 河岸
英语	net	(1) 网 (2) 净（含量）
英语	foot	(1) 脚 (2) 英尺 (3) 总计

表 1.1 不同语言中的常见多义词举例。

已有的关于多义词向量的工作 [5, 12, 13, 18, 22, 26, 33] 大多通过一个词的上下文进行聚类和分类来确定词义。这是一个非常自然的方案，考虑如下两句话：

- A **bank** account is considered indispensable by most businesses and individuals.
- In geography, the word **bank** generally refers to the land alongside a body of water.

英语水平达到一定程度的人能够轻易地分出两句话中 bank 分别表示“银行”和“(河)岸”，而人类分辨两个词不同具有词义的依据正是它们所在的上下文。理想情况下，一个多义词向量表示框架应该能够自动挖掘词义信息，使得每个词义都能不重不漏地被学习出一个表示。但是现实中，这一点却很难实现，原因主要有如下两点：

- 无监督的多义词向量学习难以不漏掉一些稀有词义：在根据上下文聚类的过程中，由于稀有词义的频率过低，其上下文很难单独被聚为一类。

- 无监督的多义词向量学习难以不重复地对每个词义学出一个单独的词表示：对于一个词，同一个义项的上下文可能有多种情况而且相差很大，如表1.2所示，在工作 [33] 中，bear(熊) 一词被学习出了三个不同的向量，但具有一定英语水平的人能够根据常识判断这前两个向量表示的词义是相同的，即动物“熊”，而第三个向量则表示了动词“忍受”。在本文中，我们定义如“熊”的前两个义项重复现象为**伪多义现象**。需要特别说明的是，伪多义的概念是针对一对学习出的义项向量而非针对某一个义项向量，即，“bear(1) 和 bear(2) 是一对伪多义”的表述正确，而“bear(1) 是伪多义”的表述错误。根据本文的研究，几乎所有无监督、直接从语料中学习词义的模型都具有严重的伪多义现象。

编号	最近邻
1	emerald, bears, three-toed, snake, periwinkle, ruffed, hoopoe, distinctive
2	bird, wolf, arrow, pelican, emerald, canyon, diamond, buck
3	pride, lady, hide, king, gift, crane, afflict, promise, reap

表 1.2 同一个义项被学习出了多种词义表示：以英文中的“bear”（熊）为例。

本文重点关注第二种现象。将同一个义项学习为多个不同、且在向量空间相似度较低的向量表示（即伪多义现象），是一种不符合语言学和认知科学先验的现象。首先，本文尝试从多种不同的角度出发对词义从数学意义上进行抽象定义，并对伪多义进行检测和消除，从实践上证明了：伪多义是一种降低词向量表示能力、增加整体模型复杂性的现象，消除伪多义能够提升词向量的表达能力。

本文中涉及的伪多义检测算法主要有三种，分别是：

- 基于外部知识库的伪多义检测。对于常见语言，根据已有知识库（如 [31, 50]）可以非常方便地检索出一个词的上下位词。这里，我们认为一个词在具有不同上位词时才是真正的多义，而其余情况如果被学习出不同的词向量则属于伪多义。对于词向量，我们根据每个词近邻的上位词来估计该词的上位。
- 基于邻域相似度的伪多义检测。这种算法可以看成是第一种算法的无监督版本，基于的是由观察得到的假设：伪多义是一种系统现象而非偶然出现。这种算法以一个义项对应词向量周围的近邻词集合来表示一个义项的含义，再通过该集合中词的两两相似度评估两个义项词向量对应的相似度，设定阈值决定伪多义的选取。然而，阈值的选取是一件困难的事，使得这种算法较为主观，在理论上没有足够的支持。

- 基于词内义项对差向量矩阵的矩阵降维算法。这种算法同样是一种无监督算法，同样基于伪多义是一种系统现象的假设，和在大部分已有的词向量中，伪多义远多于真多义数目。该算法将每个词内部所有义项两两组成对，并将正反两个方向的差向量同时加入差向量矩阵，从而得到一个行列数分别为 (对数 ×2, 词向量维度) 的矩阵。对这个矩阵进行降维所得到的“主成分”即可张成一个“伪多义空间”。此时，在伪多义空间投影比例较大的词义对即被认为是伪多义。事实上，采用这种方式对伪多义进行伪多义消除时不需要设定阈值，只需要指定一个伪多义空间的维度。这种方法基于严格的矩阵分解理论，同时也避免了选取超参数的主观性，在一定程度上规避了阈值设定的困难。

其次，本文提出了一个给定任意基底进行矩阵变换方式来消除伪多义，并从理论上证明了这种方法能够实现最低损失的伪多义消除，同时保留其余词之间的相对位置关系不变。这种矩阵变换的方法同样可以应用于词向量的偏差消除工作。

本文结构组织如下：第二章介绍研究背景，包括词向量、基于矩阵分解的降维算法以及中英文词义知识库。第三章首先提出了三种多义词检测算法并做了对比分析，接着提出了一种基于矩阵变换的伪多义消除算法。第四章以实验分析了伪多义消除的效果。第五章为总结和展望。

第二章 研究背景

2.1 词向量

最早通过后向传播 (backward propagation) 算法进行词向量表示学习的工作可以追溯到 1986 年 [38]。这种将词表示为一个向量的方法随后被证明在语言模型、语音识别和情感分析等诸多任务上极其有效 [2, 6, 10, 39, 40, 42, 44, 45]，现已成为一个通用的自然语言处理的基础工具。

按照获取方法，词向量大致可以分为如下两类：

1. 基于统计的词向量表示。这种词向量表示方法先通过一系列统计量（如互信息）将每个词的特征进行表示，之后对特征进行基于矩阵分解的降维以取得便于运算的词向量。
2. 基于神经网络和后向传播算法的词向量学习。这种方法通过神经网络结构，根据上下文进行词向量的学习。

接下来的两小节将分别回顾这两种不同的词向量表示方式。

2.1.1 基于统计的词向量表示

基于统计的词向量表示离不开信息论中互信息的概念。我们在这里简单介绍互信息的相关背景知识：对于离散随机变量 X 和 Y，互信息 (Mutual Information)[8] 可以定义为

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{2.1}$$

其中，x, y 是 X, Y 的所有可能取值，$p(x), p(y)$ 分别表示 x 和 y 被取到的概率，$p(x,y)$ 表示 x 和 y 的共现概率。互信息反映的是两个随机变量的共现程度，在自然语言处理的应用中，我们经常定义两个词“共现”为两个词按顺序连续出现。

对于 X 的每个取值 x 和 Y 的每个取值 y，式 (2.1) 求和号之后的后半部分即是 x 与

y 之间的点互信息 (Pointwise Mutual Information, PMI)，记作

$$i(x;y) = \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{2.2}$$

对于给定语料和词典 V，我们可以计算一个大小为 $|V| \times |V|$ 的点互信息矩阵 M^I。在矩阵 M^I 中，第 x 行第 y 列元素 $M^I_{x,y} = i(v_x, v_y)$ 为行列对应词之间的点互信息。根据概率知识，我们能够根据语料中词 v_x 的频数 $c(v_x)$ 和二元组 (x,y) 的频数 $c(v_x,v_y)$ 来估计 $p(x)$ 和 $p(v_x,v_y)$. 它们具有如下关系：

$$\begin{aligned} p(v_x) &= \frac{c(v_x)}{N} \\ p(v_x,v_y) &= \frac{c(v_x,v_y)}{N} \end{aligned} \tag{2.3}$$

其中，N 为语料中的总词数。由此，点互信息公式可改写为

$$i(x;y) = \log\left(\frac{c(x,y)\cdot N}{c(x)\,c(y)}\right) \tag{2.4}$$

在实际应用中，由于该矩阵过大，我们需要对其进行降维处理。然而，这个矩阵为稠密矩阵，难以存储在有限内存中进行存储。所以，在降维前的一个必要步骤就是进行稀疏化。一个有效的稀疏化手段是将点互信息矩阵改为正点互信息 (Positive Pointwise Mutual Information, PPMI) 矩阵，即将点互信息矩阵中所有负元素都视为 0，同时保留所有非负元素。根据常识，一个词只会和少量的词连续出现多次，故正点互信息矩阵为稀疏矩阵，可以利用稀疏矩阵上的降维算法对其进行降维。工作 [44] 提出，降维后得到的较低维数向量即可作为词义的分布式表示，且这种表示能够较好地对词义相似度进行建模以完成其他下游任务 [43]。

工作 [21] 通过推导，在理论上证明了基于 skip-gram 和负采样 (Negative Sampling, NS) 结合 (SGNS) 的词向量训练 [29] 在本质上等价于对点互信息矩阵进行 SVD 降维，同时该工作也证明了对偏移后的正点互信息矩阵降维在词相似度 [4, 9] 和类比 [28, 30] 实验上取得了和 SGNS 模型相似的结果。此外，[23] 从表示学习的角度上证明，SGNS 模型在实质上等价于词的共现矩阵 (co-occurence matrix) 的直接分解。

此外，Pennington 等人的工作 [36] 通过一个热动力场模型在非常大规模的语料上

获取的词向量模型效果得到许多研究者的认可，被很多深度学习工作采用为了初始向量；基于 word2vec [29] 的改进工作 [15] 提出的使用若干诸如合并互信息较高单字/单词再进行训练的 FastText 模型在许多下游任务上取得了非常好的效果。

2.1.2 基于神经网络的连续词向量表示学习

基于神经网络的词向量表示学习大体可以分为如下两类。如图 2.1所示，方便起见，我们以 Mikolov 等人的两篇工作 [28, 29] 为例进行说明。其他模型也大致可以根据该分类法归入其中一类。

1. 连续词袋子 (Continuous Bag-of-Words, CBoW) 模型。这种模型通过给定一个词一定窗口范围内的上下文来预测该词，实际上是通过训练上下文向量的词向量来获取语义。
2. skip-gram 与负采样相结合的模型。这种模型通过最大化词与其上下文词的内积、最小化词与按均匀分布所采出的负样本的内积来完成语义表示的学习。本文的章节2.1.1提到，这种模型其实在本质上等价于对于点互信息矩阵进行降维。

如表 2.1所示，基于神经网络的词向量与基于统计的词向量一样，具有“邻域相似”性质。

此外，基于神经网络的词向量可以通过语言学先验知识使得词向量更具语言学特性：在工作 [19] 中，Omer Levy 和 Yoav Goldberg 提出，除了上下文信息可以在训练 skip-gram 目标的词向量中被考虑之外，在句子中的依存信息也被考虑进来，具有相同句法角色的词被拉近，不同句法角色的词被拉远。[19] 中提出，如表 2.1所示，这种考虑依存信息的词向量更容易把语义、句法同时较为相近的词嵌入到一个较小的邻域内。

2.1.3 多义词向量

考虑到一个词对应一个向量的模型在语义表达上的局限性，研究者们提出了多义词向量表示 [13, 37]。后续工作大多数基于神经网络和 skip-gram 模型：工作 [33] 提出，对于每个词，同时学习三种词向量表示

1. 全局词向量。每个词只有一个，代表一个词的全局语义。
2. 义项向量。每个词可以有若干个（视模型而定）。
3. 上下文中心向量。每个义项向量伴随一个上下文中心向量。在给出一个词 w 时，模型首先根据其上下文窗口内的词计算其上下文向量，并根据上下文向量，在 w

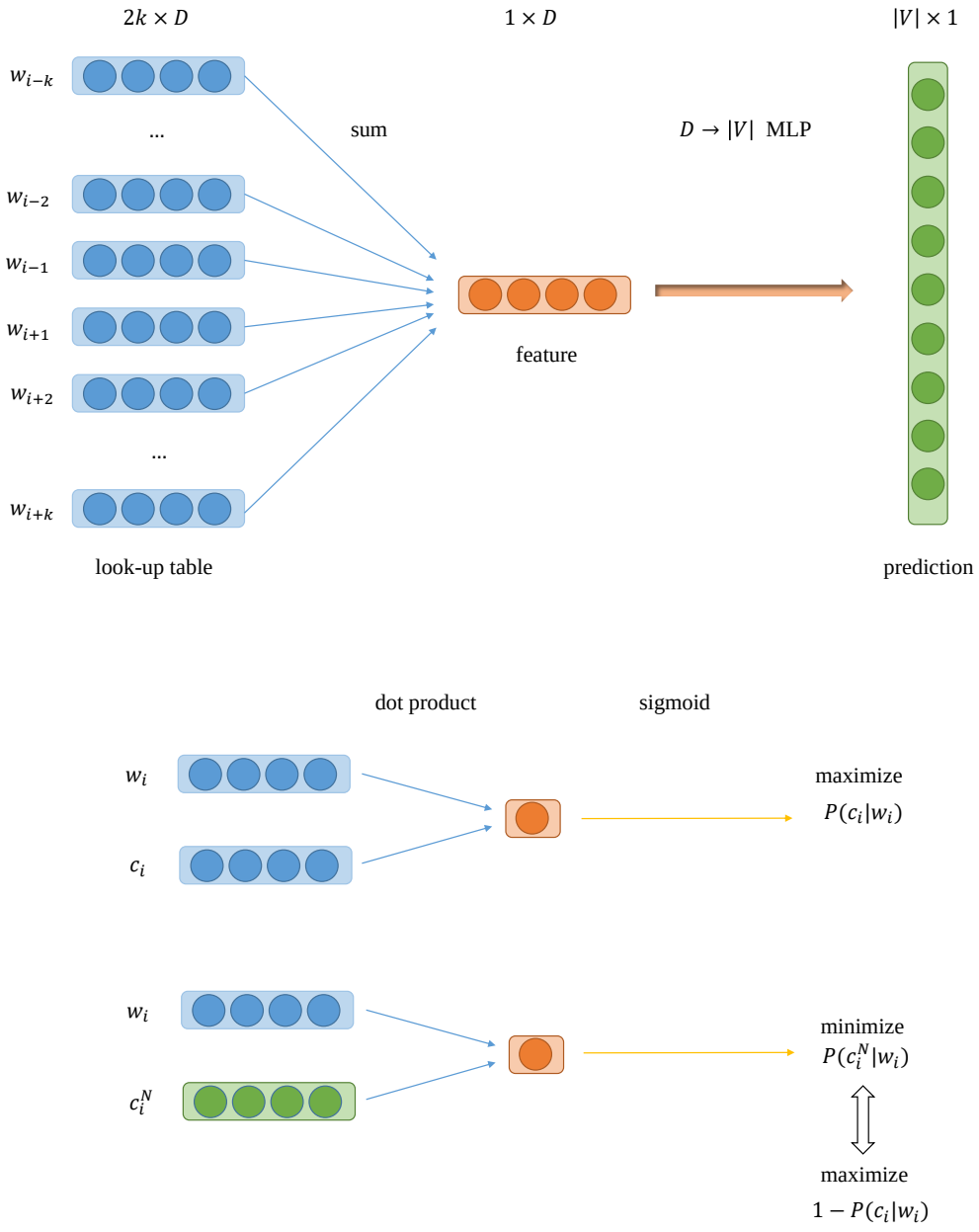

图 2.1 基于窗口的连续词袋子模型（上）和基于负采样的 skip-gram 模型（下）图示。

Target Word	词袋子模型 2-gram [29]	词袋子模型 5-gram [29]	依存模型 [19]
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl	event-driven domain-specific rule-based data-driven human-centered
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning	singing rapping breakdancing miming busking

表 2.1 不同目标函数的基于神经网络训练出词向量的 k 近邻展示 (k=5)。引用自 [19]。容易看出，基于依存关系的词向量更容易学习出“相似”关系，而完全基于上下文的词袋子模型则更容易学习出（话题上的）“相关”关系。

所有义项中选择对应上下文中心向量距离计算出的上下文向量最近的一个义项作为选取的义项，然后再按照 skip-gram 模型，根据每个词的义项向量和其上下文词的全局向量优化模型。此处，上下文中心向量随模型的学习动态更新。容易证明，当模型迭代次数足够多时，上下文中心向量趋于稳定。

图 2.2展示了多义词向量的 skip-gram 模型。值得注意的是，这种模型中的全局向

量和义项向量被统一在了一个空间内。大多数后续模型也都能将模型参数归纳为这三种向量，本文也着重讨论能够将模型参数归纳为这三类向量的无监督多义词向量学习模型。

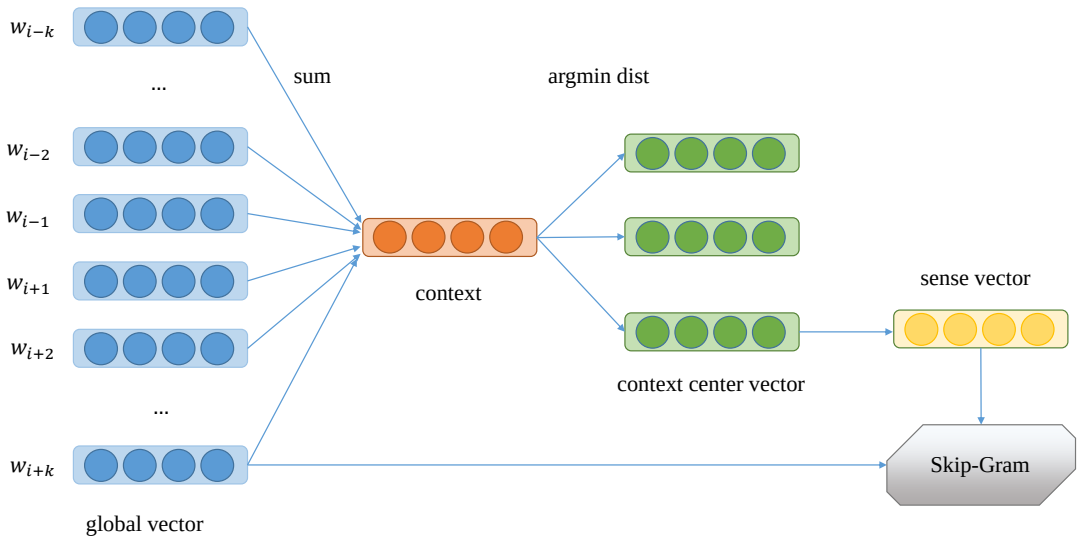

图 2.2 工作 [33] 中提出的多义词 skip-gram 模型。

在本文中，我们主要讨论已开源代码或易于复现的工作 [12, 18, 22, 33] 中的多义词向量模型。

2.2 基于矩阵分解的降维算法

这一节首先介绍两种基础且常见的的降维算法，之后将针对这两种算法对于强噪声不稳定的劣势介绍它们的改进版。

2.2.1 奇异值分解

奇异值分解 (Singular Value Decomposition, SVD) [11] 是线性代数里重要的一种分解形式，其分解后所提取的矩阵一般可以认为是在最大限度保留了原矩阵信息的基础上所进行的降维。奇异值分解是特征值分解 (Eigen Value Decomposition, EVD) 对奇异

矩阵和非方阵的扩展，其目标是将 $n \times m$ 的任意矩阵 M 分解为三个矩阵之积:

$$M = U\Sigma V^* \tag{2.5}$$

其中，U 为 n 阶正交酉矩阵；Σ 为 $n \times m$ 对角矩阵（只有前 $\min(n,m)$ 行的对角线上有值）；V^* 为 $m \times m$ 阶正交酉矩阵 V 的共轭转置，也为正交酉矩阵。将 Σ 中的元素从大到小排列为 $\{\sigma_1, \sigma_2, ..., \sigma\min(n,m)\}$，其中任意 $\sigma_i (1 \le i \le \min(n,m))$ 都对应了矩阵 M^TM 一个特征值的平方根，且 $\sigma_i \le 0$。σ_i 被称作是矩阵 M 的**奇异值**。

我们保留矩阵的前 k 大特征值，此时可以得到一个 $n \times k$ 的矩阵 Σ_k，$U\Sigma_k$ 矩阵即为 M 对行降维的结果。同时，保留 V^* 中对应 Σ 中前 k 大奇异值的列向量（按照 Σ 中排序奇异值的方式的初等列变换方式进行变换），得到的 V_k^* 即为以 $U\Sigma_k$ 中数据点（行）为坐标的一组正交基底。

一般情况下，对一个实矩阵 M 进行 SVD 分解，需要求矩阵 M^TM 的特征值。而矩阵 M^TM 是半正定的（定理 2.1），其所有特征值 λ_i 均满足 $\lambda_i \ge 0$，故 M 的特征值 $\sigma_i = \sqrt{\lambda_i} \in \mathbb{R}$，即实矩阵的 SVD 分解可以在实数域内完成。

定理 2.1 对于任意实矩阵 A, A^TA 是正定或半正定矩阵。

证明 不妨设 A 为 n 行 m 列矩阵，则 A^T 为 m 行 n 列矩阵。对于任意 $\mathbf{x} \in \mathbb{R}^m$

$$\mathbf{x}^TA^TA\mathbf{x} = (\mathbf{x}^TA^T)(A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = \sum_{i=1}^{m}(A\mathbf{x})_i^2 \ge 0 \tag{2.6}$$

故 A^TA 为正定或半正定矩阵。

证毕。

2.2.2 主成分分析

类似于奇异值分解，主成分分析 (Principal Component Analysis, PCA)[14, 46] 同样是一种常见的矩阵降维算法。但与由特征值分解演化出的奇异值分解不同，主成分分析的核心思想是提取数据点中“方差较大的部分”。基于这个思想，数据的每一维度在

进行分解前应该减去该维度的均值，即对数据进行常见的“0 均值化”操作。

记 Q 为对 M 进行 0 均值化操作后的矩阵。此时，$\frac{Q^T Q}{n-1}$ 是一个中 M 中每个数据点（行）的所有特征对之间协方差的无偏估计，称为 M 的协方差矩阵，记作 M_{cov}。同样地，由定理 2.1 可知，M_{cov} 是一个正定或半正定矩阵，其所有特征值均不小于 0。我们对于 M_{cov} 进行特征值分解，得到的每个特征值即正比于 M 矩阵中的所有数据点在对应的特征向量投影的方差，这些特征向量就是分离出的“主成分”。由于 M_{cov} 为实对称矩阵，这些分离出来的不同主成分必定两两正交。将 M 中的所有数据点（行）变换为其在主成分上的投影坐标后，即完成了对 M 的降维。

特别地，在数据矩阵中，如果数据每一维度均值都为 0，则主成分分析得到的结果完全等价于奇异值分解。

另一种常见的对于主成分分析的表达是，将 M 表达为一个低秩序矩阵 L 和一个噪声矩阵 E 之和，一般地，在实际应用中，E 可被视为高斯噪声矩阵。这里，提取出的主成分为 L 行空间的一组正交基，而 L 的每一行都是对应的 M 中的一行在主成分空间上的投影。

2.2.3 健壮主成分分析和健壮奇异值分解

主成分分析和奇异值分解都基于所有样本点没有被强噪声污染过的假设，即，假设每次采样都是服从数据的原始分布外加一个方差可忽略的高斯误差所得到的。然而，在实际应用中，很多数据采样都会受到较强噪声的干扰。主成分分析和奇异值分解对这种强噪声异常地敏感：几个甚至只有一个强噪声数据点会极大地影响均值的估计，进而影响主成分分析的结果。所以，工作 [47] 提出，应当将可能被强噪声污染的数据矩阵 M 表达为三个矩阵 L、E 和 S 之和，其中，L 和 E 的定义与主成分分析相同，而 S 则是一个稀疏的强噪声矩阵。这类问题被称为健壮主成分分析 (Robust Principal Component Analysis, RPCA)。

在提出这个问题的同时，[47] 同时提出了一个基于凸优化的主成分分析解决方案：首先，健壮主成分分析的优化目标可以表达为

$$\begin{aligned} \min \quad & \text{rank}(L) + \lambda_1 ||E||_F + \lambda_2 ||S||_0 \\ \textit{subject to} \quad & L + E + S = M \end{aligned} \tag{2.7}$$

其中,$||\cdot||_F$ 表示矩阵的 Frobenius 范数，$||\cdot||_0$ 表示矩阵的 0 范数，λ_1 和 λ_2 为选定的超参数。然而，由于目标函数非凸，这个问题是非凸的，无法通过简单的优化方式进行优化。[47] 给出的方案是对原问题进行凸松弛，即将原问题松弛为

$$\begin{aligned} \min \quad & ||L||_* + \lambda_1||E||_F + \lambda_2||S||_1 \\ \textit{subject to} \quad & L + E + S = M \end{aligned} \tag{2.8}$$

其中，$||\cdot||_*$ 表示矩阵的核范数（矩阵奇异值之和），也是矩阵秩的一个凸近似，$||\cdot||_1$ 是矩阵的 1 范数。这样，目标函数就成为了一个凸函数。对问题 (2.8)，可以采用已有的凸优化算法 [1, 24, 34] 进行求解。同时， [47] 也展示了一些该算法求解健壮主成分分析问题在图像处理方面的结果。实验表明，健壮主成分分析对于消除图像中的强噪声、并对图像进行尽可能的低秩表示具有很大的帮助。

健壮奇异值分解 (Robust Singular Value Decomposition, RSVD) [25, 48] 与健壮主成分分析所基于的假设相同，即应当忽略那些强噪声对奇异值计算的影响。鉴于在本文讨论的问题中，主成分分析和奇异值分解完全等价，且健壮主成分分析的提法较为普遍，所以本文以健壮主成分分析来定义解决的问题。

2.3 词义知识库

WordNet [31] 是由普林斯顿大学的研究者发起的词关系项目，现今已有各种语言上的扩展版本，已经成为了当下自然语言处理领域最为通用的知识库之一。

WordNet 的元单位为 Synset(synonym set)，每个 Synset 对应一个义项。每个义项可以与多个词 (word) 相关。WordNet 中存储了每个 Synset 的词性和详细解释和大量相关关系。最常用的相关关系为上下位关系和整体-部分关系。所有 Synset 根据某一种关系形成一个拓扑图（每个 Synset 可能与一个或多个其他 Synset 有同种关系）。常见的 WordNet 的扩展版本还包含了话题域信息等。在本研究中，我们将着重采用上下位关系信息。

对应 WordNet，同义词词林（及其扩展板）是在中文上的词义知识库。同义词词林同样收录了中文中上下位词和同义词的关系，并对每个词义进行了有效的编码。同 WordNet 一样，同义词词林也是自然语言处理，尤其中文自然语言处理中的宝贵而有

价值的资源。

第三章 多义词向量中的伪多义检测和消除

3.1 多义词向量中的伪多义检测

本节将提出三个伪多义检测的算法。其中，前两个算法将显示地检测出伪多义对，而第三个算法则提取的是隐式的“伪多义空间”。对比之下，第三个算法更具系统性，效果也更为显著。

3.1.1 算法一：基于外部知识库的伪多义检测

什么是多义词？这让我们开始从头思索多义词的定义。本文认为，多义词，一定不是系统产生的——假如一个词有两个“义项”，而与它相似的词有类似对应关系的两个“义项”，那么这两个“义项”有很大可能是隐喻 (metaphor)[16] 关系，而非多义。比如，如果进行比较细致地分类，“攻击”有实体的“攻击”义项：“攻击一座城市”，也有抽象的义项：“网络攻击”；对应地，“防御”一词也有同样关系的两个词义。一部分语言学家认为，这两个义项并非“多义”(multi-sense)，而是抽象的义项是具体义项的隐喻表示，称抽象义项为隐喻表达 (metaphorical expression)，具体义项为字面表达 (literal expression)[17, 43]。本文采纳这种观点，不考虑这部分有争议的“多义词”的影响，并将它们默认为“伪多义”。实验表明，这种伪多义在所有伪多义的例子中占极少的一部分，更多检测出的伪多义则是毫无争议的同义词被映射到了不同的词向量，这种映射实际上非常有损词向量的语义表达。

对于多义词的表达，本章里，我们借助外部知识库的帮助：对于英文，我们以 WordNet[31] 中的上位词 (hypernym) 来表达多义；同样地，对于中文，我们可以用同义词词林 [50] 中的上位词来表达。正如在相关工作一章中所提到得到，上位词一般描述了一个词或 Synset 的上一层的特征，如“动物”、“哺乳动物”和“宠物”都是“猫”的上位词。在 WordNet 中，上位关系（有向边）和 Synset（节点）共同构成了一个拓扑图。对于多义词，我们定义：如果一个词对应的不同 Synset 具有不同的上位词集合，我们认为这两个 Synset 具有不同的含义，即该词为**多义词**。

对于给定的词 w 的两个义项 s，记其所对应的词向量为 $\mathbf{v}_{s_1}^{w}$，给定这个词向量和

WordNet 或其他外部知识库，我们可以找到该词向量所“最可能对应的”词义。具体步骤如下：

1. 在 WordNet 或其他外部知识库中找到词 w 所对应的所有 Synset，并且将它们记作 $Syn_1^w, Syn_2^w, \cdots, Syn_m^w$.
2. 对于这些可能的候选项，每个 Synset 都可能对应了一些上位 Synset，将这些上位 Synset 记为 $SynH_1^w, SynH_2^w, \cdots, SynH_n^w$。这里，我们忽略单个词的不同 Synset 对应相同上位词的极个别情况，从而使词 w 的 Synset 与 w 上位 Synset 的集合成为一个满射，且是一个逆向的单射。
3. 根据 $\mathbf{v}_{s_1}^w$ 的近邻的可能上位 Synset 对各个可能的上位 Synset 进行打分，打分方案为

$$score(\mathbf{v}_{s_1}^w, SynH_i^w) = \sum_{\mathbf{v}_{s_k}^{w'} \in NN(\mathbf{v}_{s_1}^w)} \cos(\mathbf{v}_{s_1}^w, \mathbf{v}_{s_k}^{w'}) isPossibleHypernym(SynH_i^w, \mathbf{v}_{s_1}^w) \quad (3.1)$$

该方案提出如何确定一个词的某个特定词向量在 WordNet 中最可能对应的 Synset，即对每个可能的 Synset，考虑其上位 Synset，如果该上位 Synset 也成为了其近邻的可能的上位 Synset，则这个近邻为该上位 Synset 所对应的该词的 Synset 贡献值等于余弦相似度的分数。

我们取每个词向量所对应的最高分的 Synset 作为其表达的含义，如果一个词的不同词向量对应到了相同的 Synset，我们称识别出了一对“伪多义”。①

3.1.2　算法二：基于邻域相似度的伪多义检测

上一小节提出的方案需要外部知识库进行半监督，本小节试图降低对这部分对外部知识库的依赖，设计一个无监督的伪多义检测算法。

考虑上一小节的方案，每个词向量对于 Synset 的归属完全由其近邻向量所对应的词义决定。由此，我们可以衡量词 w 的一对词义向量 $\mathbf{v}_{s_0}^w$ 和 $\mathbf{v}_{s_1}^w$ 的邻域，根据其对应的全局词向量的相似程度来进行打分。两个词义向量 $\mathbf{v}_{s_0}^w$ 和 $\mathbf{v}_{s_0}^w$ 的“伪多义”程度被定义为

$$P_{pseudo}(\mathbf{v}_{s_0}^w, \mathbf{v}_{s_1}^w) \propto \sum_{\mathbf{v'}_{s_0} \in NN(\mathbf{v}_{s_0}^w)} \sum_{\mathbf{v'}_{s_1} \in NN(\mathbf{v}_{s_1}^w)} \cos(\mathbf{v'}_{s_0}, \mathbf{v'}_{s_1}) \quad (3.2)$$

① 这部分工作发表在了计算语言学与语言复杂度 Workshop(COLING Workshop CL4LC 2016) 上。

将上式对所有词对归一化，即可得到两个不同的词义向量对应了相同含义的概率。通过人工观察得到的阈值，即可确定选出的“伪多义对”。

然而，这个步骤存在如下两个问题：

- 人工观察的阈值可能并不准确。
- 由于训练语料的样本不均匀性，使得每个词所对应的伪多义的阈值可能并不一致。因此，可能不存在一个可以供选取的全局阈值来检测伪多义向量对。②

3.1.3　算法三：基于词内部意义对的伪多义检测

“伪多义”是一种词内部意义对之间关系的描述。本小节将基于这一点思考，重新形式化定义伪多义检测问题为一个矩阵分解问题，并针对矩阵分解问题的主成分分析问题给出对应稠密词向量的一个普适算法。这一算法同样适合伪多义检测问题。

首先，我们定义词内部意义对差矩阵 (intra-word sense-wise difference matrix)。对于词向量 V，令 $\mathbf{v}_s^w$ 表示词 w 的第 s 个含义所对应的向量（默认为列向量），记词向量维度为 d，记词 w 在词向量 V 中所具有的的意义对数为 n_w 则词内部意义对矩阵可以写为

$$M = \bigoplus_w [\oplus_{i \neq j} (\mathbf{v}_i^w - \mathbf{v}_j^w)] \tag{3.3}$$

其中，$\oplus$ 表示对于列向量或矩阵的在行上的连接。这里，M 是一个 d 行 $\sum_w n_w \times (n_w - 1)$ 列的矩阵。图 3.1直观地展示了词内部意义对矩阵的形态。

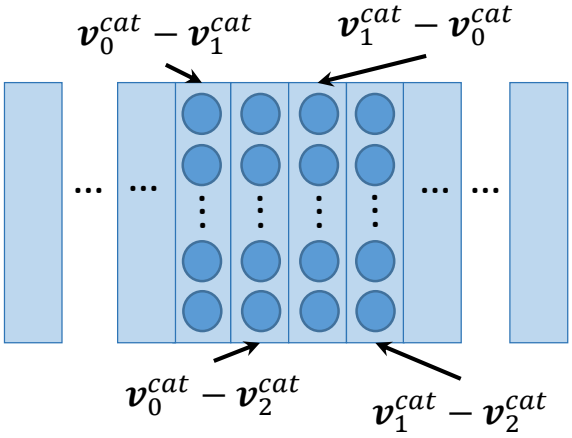

图 3.1　词内部意义对差矩阵。

② 这部分工作发表在了语言资源与评测会议 (LREC2018) 上。

这里，由于每一对词义向量的差向量 $\mathbf{v}_i^w$ 和 $\mathbf{v}_j^w$ 和 $\mathbf{v}_i^w - \mathbf{v}_j^w$ 与 M 中的列形成了一一对应，则 M 是一个在列上的 0 均值矩阵。因此，对于 M 进行主成分分析和奇异值分解在数值上没有差别。

由于健壮主成分分析被应用得比较多，本文中统一将以下对于 M 的分解方案称为主成分分析和健壮主成分分析。

基于之前的观察，伪多义在学习出的多义词向量中占了绝大多数。那么，我们如果将矩阵 M 的每一列看作一个数据点进行主成分分析（即一般意义下地对 M^T 进行主成分分析），则分离出来的一部分主成分应当张成了一个“伪多义空间”——即大多数伪多义对的差向量在这个空间上的投影长度占了向量本身长度的一大部分。

在以上算法的基础上，我们考虑真正多义词之差在 M 中的存在。[33] 提出，真多义的词义向量之间都应该具有显著差距，显然，将它们简单地视作主成分分析中的高斯噪声是不合适的。于是，本文采用健壮主成分分析来解决这个问题。

回顾健壮主成分分析算法：该算法将矩阵 M 写为三个矩阵 L、E、S 之和, 其中 L 是 M 的低秩近似，E 为 0 均值、小方差的高斯噪声，S 为稀疏噪声矩阵。这里，稀疏噪声矩阵的成分是主成分分析算法没有考虑到的。这里引入 S 的目的就是为了处理 M 中真多义词义向量差的情况。

然而，已有的基于凸松弛的健壮主成分分析算法并不适用于解决词向量词内部意义对矩阵的分解问题：回顾式 (2.8)，其分解结果依赖于超参数 λ_1 和 λ_2 的设定，而其所分解出的低秩矩阵的秩更是经过了核范数的松弛。其分解出的低秩矩阵 L 的秩对于超参数格外敏感, 即微调超参数时会带来 $\text{rank}(L)$ 的跳跃——尽管在计算机视觉的应用中，这种敏感没有太大的影响，因为视觉中的图片表征向量一般维数非常高 (10^4 或以上量级维数)。但回到其对于词向量的应用上来，对于 10^2 量级维数的词向量，“伪多义空间”的维数几乎只有个位数，这样对于超参数敏感的 $\text{rank}(L)$ 显然不适用于提取“伪多义空间”。同时，尽管真多义的存在对于伪多义来说是一个较强噪声，但词内部意义对矩阵的性质决定了真多义词对差不会对均值产生任何影响；其数量较少，也不会对方差产生过大影响。考虑到以上几点，本文提出了一种基于主成分分析的迭代健壮主成分分析算法：

考虑主成分分析算法 $M = L + E$，如果在该算法中有少量较强噪声，则其在 E 中则有大概率仍然表现为较强噪声。本文算法的就是基于逐步迭代地消去这些较强噪声

考虑的。同时，主成分分析算法对矩阵进行低秩近似需要指定一个秩，本文的健壮主成分分析算法同样需要这个指定的秩。

在健壮主成分分析算法的第 t 步，我们首先将 $M^{(t)}$ 分解为低秩矩阵 $L^{(t)}$ 和高斯噪声矩阵 $E^{(t)}$ 之和。之后，我们从 $E^{(t)}$ 里提取稀疏噪声 $S^{(t)}$，提取的方法是根据 $3-\sigma$ 准则进行一个过滤。记 $A^{(t)}$ 为该步骤过滤 $E^{(t)}$ 的掩码，其计算方式为

$$A_{i,j}^{(t)} = \begin{cases} 0, & -3\sigma_{E^{(t)}} \le E_{i,j}^{(t)} \le 3\sigma_{E^{(t)}} \\ 1, & otherwise \end{cases} \tag{3.4}$$

通过掩码 $A^{(t)}$，超过 $E^{(t)}$ 标准差 $\sigma_{E^{(t)}}$ 正负三倍范围的强噪声被过滤出来，记作 $S^{(t)}$。

$$S^{(t)} = A^{(t)} \circ E^{(t)} \tag{3.5}$$

此处，$\circ$ 表示两个相同大小矩阵各个元素之间相乘所得到的矩阵。得到 $S^{(t)}$ 之后，我们将 $S^{(t)}$ 从 $M^{(t)}$ 中减去得到 $M^{(t+1)}$，进行下一步迭代。

算法 1总结了以上描述的基于主成分分析的健壮主成分分析算法。

算法 1 基于主成分分析的迭代健壮主成分分析算法。

Require: 词内部意义对矩阵 M, 指定主成分数量 d
Ensure: 低秩矩阵 L, 稀疏噪声矩阵 S, 高斯噪声矩阵 E

令 $M^{(0)} = M, t = 0, S = 0$ 矩阵
while 未收敛 **do**
  利用 PCA 计算 $L^{(t)} + E^{(t)} = M^{(t)}$
  利用式 (3.4)计算 $A^{(t)}$
  利用式 (3.5)计算 $S^{(t)}$
  令 $S = S + S^{(t)}$
  令 $M^{(t+1)} = M^{(t)} - S^{(t)}$
  令 $t = t + 1$
end while
return $L^{(t)}, E^{(t)}, S$

定理 3.1 基于主成分分析的迭代健壮主成分分析算法对任意词内意义对差矩阵 M 都是收敛的。

证明　令 $\rho_L(M)$ 表示低秩矩阵 L 所表出的 M 中数据点的的方差比率。对于任意 $t \in \mathbb{N}$, 我们有

$$\rho_{L^{(t)}}(M^{(t)}) \leq \rho_{L^{(t)}}(M^{(t+1)}) \leq \rho_{L^{(t+1)}}(M^{(t+1)}) \tag{3.6}$$

这表明，$\{\rho_{L^{(t)}}(M^{(t)})\}_t$ 是一个单调不下降序列，且存在上界 1（$L = M$ 的情况）。根据单调有界收敛定理 [49]，$\lim_{t\to\infty} \rho_{L^{(t)}}(M^{(t)})$ 存在，即对于任意词内意义对差矩阵 M，原算法均收敛。

证毕。

更为一般地，对于任意 0 均值化后的矩阵 M，算法 1均收敛。

值得注意的是，与算法 1 和算法 2 不同，算法 3 中的伪多义检测算法在不指定阈值的情况下，不会选出哪些对是伪多义，但能够给定一些提取出的主成分作为“伪多义方向”。③

3.2　基于矩阵变换的伪多义消除

3.2.1　给定样本对的伪多义消除

给定3.1.1或3.1.2中提取出的伪多义词向量对 $\{(\mathbf{v}^{w_0}_{s_{0,0}}, \mathbf{v}^{w_0}_{s_{0,1}}), (\mathbf{v}^{w_1}_{s_{1,0}}, \mathbf{v}^{w_1}_{s_{1,1}}), \cdots (\mathbf{v}^{w_m}_{s_{m,0}}, \mathbf{v}^{w_m}_{s_{m,1}})\}$。考虑通过特定变换使这些词向量对在空间中的距离尽可能地近，一种直观的方案是：训练一个实矩阵 T，使得任意一个对中的两个向量能够尽可能地接近中点向量，即

$$\min\ \mathcal{L}(T) = \sum_{i=1}^{m} \frac{1}{2}(||T\mathbf{v}^{w_i}_{s_{i,0}} - \frac{(\mathbf{v}^{w_i}_{s_{i,0}} + \mathbf{v}^{w_i}_{s_{i,1}})}{2}||_2 + ||T\mathbf{v}^{w_i}_{s_{i,1}} - \frac{(\mathbf{v}^{w_i}_{s_{i,0}} + \mathbf{v}^{w_i}_{s_{i,1}})}{2}||_2) \tag{3.7}$$

由于 T 中任意一个元素都为实数，且目标函数为凸函数，故该问题为一个凸优化问题，在给定目标函数的情况下存在唯一解。

但这种方法存在一个很明显的问题：这种将伪多义对变换到中点的方案只考虑到了涉及到伪多义的向量，而没有考虑任何其他的向量。只对涉及到伪多义的向量进行矩阵变换，可能损害它们对于其它词义的相对意义；而对全局所有向量都进行矩阵变换，则对其它向量来讲的物理意义不明。于是，我们考虑以下对于给定子空间的伪多义消除算法。

③ 这部分工作投稿到了自然语言处理中的经验性方法会议 (EMNLP2018)。

3.2.2 给定子空间的伪多义消除

根据3.1.3中所描述的算法，如果指定一个“伪多义空间”的维度，我们可以比较容易地提取出一个“伪多义空间”。所有的伪多义方向在提取出的伪多义空间上的投影应该占据比较高的一个比例。基于这个伪多义空间，我们可以训练一个变换矩阵 T，使得其满足如下条件:

- 伪多义空间中的每个向量都被 T 投影到 0 向量;
- 原空间中每个与伪多义空间垂直的向量都被 T 投影到自身。

通过定理 3.2 容易得到：对于任意一个伪多义空间，这样的矩阵 T 是唯一的。其中，$\alpha_1, \alpha_2, \cdots, \alpha_k$ 对应了伪多义空间的一组正交基。

得到变换矩阵 T 后，我们可以通过简单的方式将整个词向量空间 V 变换到新空间 $\tilde{V}$:

$$\tilde{V} = \{\tilde{\mathbf{v}}_s^w = T\mathbf{v}_s^w | \mathbf{v}_s^w \in V\} \tag{3.8}$$

定理 3.2 给定一组线性空间 $\mathbb{R}^n$ 的正交基 $\alpha_1, \alpha_2, \cdots, \alpha_n$ 和整数 $k(1 \le k \le n)$, 存在唯一的一个矩阵 T，使得：$T\alpha_\mathbf{i} = 0, \forall 1 \le i \le k$ 且 $T\alpha_\mathbf{i} = \alpha_i, k \le i \le n$.

证明 令矩阵

$$A = \begin{pmatrix} \alpha_1^T \\ \alpha_2^T \\ \cdots \\ \alpha_n^T \end{pmatrix} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,n} \\ & & \cdots & \\ \alpha_{n,1} & \alpha_{n,2} & \cdots & \alpha_{n,n} \end{pmatrix} \tag{3.9}$$

$$\begin{cases} T\alpha_1 = 0 \\ T\alpha_2 = 0 \\ \cdots \\ T\alpha_k = 0 \\ T\alpha_{k+1} = \alpha_{k+1} \\ \cdots \\ T\alpha_n = \alpha_n \end{cases} \Leftrightarrow \begin{cases} \alpha_1^T T_i^T = 0 \\ \alpha_2^T T_i^T = 0 \\ \cdots \\ \alpha_k^T T_i^T = 0 \\ \alpha_{k+1}^T T_i^T = \alpha_{k+1,i} \\ \cdots \\ \alpha_n T_i^T = \alpha_{n,i} \end{cases} \quad \forall i, 1 \le i \le n \tag{3.10}$$

这里, T_i^T 表示矩阵 T^T 的第 i 列。式(3.10)的右半部分可以整理为

$$AT_i = (\underbrace{0, 0, \cdots, 0}_{k\times 0}, \overbrace{\alpha_{k+1,i}, \cdots, \alpha_{n,i}}^{\mathbf{c}_i})^T, \forall i, 1 \le i \le n \tag{3.11}$$

由于矩阵 A 的行向量的构成线性空间 $\mathbb{R}^n$ 的一组正交基，故 $\text{rank}(A) = n, A^{-1}$ 存在。因此，式 (3.11)中，T_i 有唯一解 $T_i = A^{-1}\mathbf{c}_i$。

综上所述，T 在给定 $\alpha_1, \alpha_2, \cdots, \alpha_n$ 和 k 的时候有唯一解。

证毕。

更为一般地，T 在给定 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 时即可唯一确定。

值得一提的是，这种给定子空间的伪多义消除算法与工作 [3] 中提出的 “De-Biased Word Embeddings’’ 比较相似。容易看出，如果将 [3] 中的 “bias’’ 方向当作待消除的伪多义方向，且 “de-bias’’ 的范围是所有词向量全集的时候，两个算法是等价的：两项工作分别从不同的角度出发得到了类似的方法和相同的效果。

3.3　基于点互信息矩阵分解的多义词向量再思考

在工作 [21] 中，作者通过推导证明了基于神经网络学习的词向量在理论上等价于对点互信息矩阵进行分解。本节将讨论一些该文章中所提出的 “等价性” 在多义词向量上的扩展。

简单起见，我们考虑一个语料的正点互信息 (PPMI) 矩阵 P。显然，矩阵 P 是稀疏的。我们记 P 的第 i 行为 P_i，令其表示词 w_i 对于所有其他词的点互信息向量。考虑一个多义词的两个不同义项（非伪多义），如英文中的 bank（银行）和 bank（河岸）。基于分布式假设的语义表示会假设 bank（银行）与 money（钱）、check（支票）等词的表示距离较近，而 bank（河岸）则与 river（河），lake（湖）等词表示距离较近。

我们对于真多义词的一个基本假设是，其出现是非系统的、单一的（这一点在实验中会进一步证明）。由此，我们不妨假设真多义词的每个义项相互独立，即不受其它义项干扰。这时，我们考虑词 w 的两个义项 s_1 和 s_1 所分别对应的点互信息向量，取 t_1 和 t_2 两个词作为上下文中词的代表。

$$\begin{aligned}\mathbf{v}_{PMI}^{s_1} &= [\mathrm{PMI}(s_1,t_1);\mathrm{PMI}(s_1,t_2)] \\ &= \left[\frac{c(s_1,t_1)\cdot N}{c(s_1)\,c(t_1)};\frac{c(s_1,t_2)\cdot N}{c(s_1)\,c(t_2)}\right]\end{aligned} \tag{3.12}$$

$$\begin{aligned}\mathbf{v}_{PMI}^{s_2} &= [\mathrm{PMI}(s_2,t_1);\mathrm{PMI}(s_2,t_2)] \\ &= \left[\frac{c(s_2,t_1)\cdot N}{c(s_2)\,c(t_1)};\frac{c(s_2,t_2)\cdot N}{c(s_2)\,c(t_2)}\right]\end{aligned} \tag{3.13}$$

如果两个义项加以合并，我们得到的点互信息向量则是：

$$\begin{aligned}\mathbf{v}_{PMI}^{w} &= [\mathrm{PMI}(w,t_1);\mathrm{PMI}(w,t_2)] \\ &= \left[\frac{c(w,t_1)\cdot N}{c(w)\,c(t_1)};\frac{c(w,t_2)\cdot N}{c(w)\,c(t_2)}\right] \\ &= \left[\frac{(c(s_1,t_1)+c(s_2,t_1))\cdot N}{(c(s_1)+c(s_2))\,c(t_1)};\frac{(c(s_1,t_2)+c(s_2,t_2))\cdot N}{(c(s_1)+c(s_2))\,c(t_2)}\right]\end{aligned} \tag{3.14}$$

显然，只有 $c(t_1)=c(t_2)$ 的时候，$\mathbf{v}_{PMI}^{s_1}$ 和 $\mathbf{v}_{PMI}^{s_2}$ 才能线性表出 $\mathbf{v}_{PMI}^{w}$，而二者在语料中相等只有极小的可能性。

另有，$\mathbf{v}_{PMI}^{s_1}$ 和 $\mathbf{v}_{PMI}^{s_2}$ 分别和各自含义相关的词向量距离较近，即，可以认为如果对词向量进行聚类，理想状态下，每个类别所对应的点互信息向量分别在一个较小的噪声容忍范围内秩为 1。

将 s_1 和 s_2 融合为 w 后，实际上以较大的概率提升了 w 所对应的点互信息向量所“不能被 s_1 和 s_2 所在的代表向量线性表出”的程度（扩大了对线性表出的噪声容忍范围的要求）。而如果 s_1 和 s_2 为伪多义，且各个语料中的词被均匀地分配给了两个词义，

则可以认为 $c(s_1)$ 约等于 $c(s_2)$，此时融合 s_1 和 s_2 没有上述扩大噪声容忍程度的要求。

然而，鉴于以上方法难以自动生成新的词义，本文暂时不讨论它的应用，而只将其作为一个推论和可能的词向量评测方式加以说明。

第四章 实验结果

由于 [33] 提出的动态聚类过程方法和 [22] 提出的狄利克雷过程（中餐馆过程）方法非常相像，且前者更具确定性，以下的实验主要在工作 [33] 所公开的词向量上进行。如未加特定说明，则所有结果都出自于其公布的 300 维 NP-MSSG，对前 6000 高频词汇自动学习出的多义词向量。

对于矩阵训练中的优化问题，本文中的工作使用 PyTorch 工具 [35] 进行优化。

4.1 经典算法中的伪多义现象

模型	词义	词义间相似度	k 近邻	k 近邻平均相似度
[13]	cat_{s0} cat_{s4}	0.265	$cats_{s1}, dog_{s2}, mouse_{s3}, dogs_{s1}, chicken_{s0}$ $dog_{s4}, cats_{s0}, cats_{s2}, bat_{s3}, mouse_{s1}$	0.764 0.758
[33]	cat_{s0} cat_{s3}	0.294	$dog_{s2}, cat_{s3}, mr_{s9}, girl_{s1}, mouse_{s0}$ $dog_{s4}, bat_{s8}, cat_{s0}, air_{s4}, pan_{s8}$	0.703 0.588
[22]	cat_{s0} cat_{s1}	0.928	$cats2, cat_{s0}, dog_{s0}, dog_{s2}, dog_{s1}$ $cat_{s2}, cat_{s0}, dog_{s0}, dog_{s2}, dog_{s1}$	0.779 0.811
[18]	cat_{s0} cat_{s1}	0.876	$cat_{s1}, cat_{s2}, dog_{s0}, dog_{s1}, dog_{s0}$ $dog_{s1}, cat_{s2}, cat_{s0}, dog_{s0}, dog_{s1}$	0.831 0.784

表 4.1 经典方法 [18, 22, 33] 中的伪多义现象。

值得注意的是，通过方法 [18, 22] 学习出来的向量中，相同词的几乎所有词义向量都呈现出了“堆在一起”的形态。这表明，尽管这些工作在词义级别相似度的任务上取得了较好成果，它们仍然是“单义词向量”的一个变种，而没有如工作 [33] 一样能够较为自动地识别出词义。同样地，[12] 也报告了类似的重复 (duplication) 现象，该文章作者们也通过一些类似与伪多义消除中算法二的方案对伪多义进行消除后再去应用到下游任务中。

因此，在接下来的实验中，我们将着重考虑通过模型 [13] 和模型 [33] 训练出的公开向量。这两份向量的“伪多义”区分程度较大而非呈现一种“堆在一起”的形态，这种性质是本文提出的识别和消除算法的进行的前提——本文所定义的“伪多义”也更倾向于这种形态的同义、不同词向量。

4.2 对于伪多义检测算法一的直观评估

STAR

[13]	princess, series, cast, serial, midway, sparkle, 1940s, leo, closet, co-star	01
	silver, boy, cat, version, adventures, stars, emerald, destroyer, terrace, planet	02
	energy, disk, wheel, disadvantage, block, puff, radius, diamond, chord	03
	version, bronze, standard, colors, ring, emblem, silver, wear, shoulder, red	01
	workshop, shop, paper, merchandise, plain, corporation, stock, likeness	03
	guard, baseball, starter, tennis, basketball, brazil, class, world, morocco, ncaa	01
	appearance, entertainer, pat, alumnus, freelance, brother, session, receiver	01
	fictional, ongoing, manga, super, japanese, silver, interactive, asian, fiction	01
	die, express, ride, opera, spanish, musical, hour, disaster, sun, blue	01
	galaxy, spiral, variable, guide, magnitude, companion, satellite, crater	02
[33]	blue, dragon, acbl, diamond, purple, legion, arrow, mercury, eagle, cross	01
	fan, legend, show, moesha, heroes, guest-star, flicka, lassie, tv-movie	01
	stars, sun, constellation, galaxy, eridani, pegasi, supergiant, ceti, starburst	02

01: person.n.01 02: celestial_body.n.01 03: whole.n.02

ROCK

[13]	blur, indulgence, pop, noise, bands, lacuna, reformed, wave, genre, taster	01
	energy, silver, cat, song, cd, planet, dawn, hero, video, terrace	02
	metal, classic, legendary, dubbed, american, hard, belgian, short-lived, debut, da	01
	soft, shifting, disappear, fill, crystalline, false, pitch, expanse, heat, pile	03
	vinyl, concert, limited, box, summer, double, dance, enhanced, gold, inch	04
	hop, well-known, folk, occasional, jazz, music, concert, array, hard, pop	01
	morris, miami, wood, ghost, silver, pearl, chase, corner, oak, thousand	03
	hard, pop, cm, jazz, hip, hop, r&b, gutter, wave, subculture	01
	hard, hip, short-lived, classic, jazz, raw, metal, ep	01
	jazz, rally, star, roll, live, entertainer, appearance, session, pop, cover	01
[33]	metal, rippling, dense, swirling, chirping, blues, punk, psychedelia, bands, pop	01
	sand, rocks, butte, ash, sandy, little, cedar, rocky, sugarloaf, spring-fed	03
	hip, alternative, indie, progressive, hop, reggae, roll, rock/metal, post-hardcore	01

01: popular_music.n.01 02: person.n.01 03: material.n.01 04: whole.n.02

表 4.2 经典英文多义词 star(明星人物/天体星星) 和 rock(摇滚乐/人名绰号/岩石) 的词向量根据其在向量空间中的近邻向量对 WordNet[31] 中 Synset 的映射。第一列表示模型，第二列表示每个词义向量的前 10 名近邻，第 3 列展示了检测出的该词义向量所对应的 Synset 的上位 Synset 编号，上位 Synset 的名称附于每个子表格下。whole.n.02 表示没能通过词向量中的近邻词发现上位词(Synset)。为了公平地比较两个模型，这里采用了 [33] 中的固定词义数目为 3 的 50 维多义词向量(MSSG-50D)。两个模型都在维基百科语料上进行了训练。

表4.2直观地展示了利用算法 1 进行伪多义检测的结果。可以看到，尽管 [13] 对每个词义根据上下文聚类学习出了多个词向量，但大多都是冗余的——对每个词只学习 3 个词义向量的 MSSG 模型也能够达到同样的效果。通过肉眼观察，算法一识别出的

WordNet 上位 Synset 与人类的先验知识比较吻合，可以作为一个较好的基础方案对无监督多义词向量的词义进行映射。

4.3 对于伪多义检测算法二的直观评估

对于伪多义检测算法二，我们用式 (3.2)定义两个词义项的伪多义程度，并人工设定阈值 1 进行真伪多义的判定。在判定伪多义之后，我们在原语料中将对应的伪多义词强制标为同一含义。回顾 [33] 所提出的多义词 skip-gram 向量训练方案，该方案需要对语料中的每个词根据其上下文估计其含义。在强制将伪多义标记为同一含义后，我们省略这一步骤，直接将语料中的词分配给特定的词义进行多义词 skip-gram 的训练。

NORWAY	
NP-MSSG [33]	Denmark, Troms, Sogn, Hedmark Denmark, Sweden, Finland, Iceland Denmark, Sweden, Finland, Netherlands Denmark, Austria, Germany, Belgium
+ Tag	Denmark, Norwegian, Sweden, Trondheim

STAR	
NP-MSSG [33]	stars, movie, song, heart stars, award, eagle, two-time supergiant, constellation, aurigae
+ Tag	stars, movie, superstar, MVP supergiant, stars, g5v, white main

ALGORITHM	
NP-MSSG [33]	hash, algorithms, quick sort, recursive algorithms, optimization, public-key
+ Tag	algorithms, computation, iteratively

表 4.3 加入标记前的（带伪多义的）词的词义向量中不同词义的近邻，以及加入标记后重新训练的词向量中各个词的近邻。

表4.3展示了 300 维 NP-MSSG 模型在标记前和标记后的表现。可以看出，这种标记的方案不仅能够基本去除伪多义，还能够保留真多义词之间的区别，是一种比较有效的伪多义去除手段。

4.4 对于伪多义检测算法三的直观分析

我们回顾考虑伪多义检测算法三中对于词内部意义对差矩阵 M 进行分解的方式:

$$M = L + E + S \tag{4.1}$$

其中，L 为低秩矩阵, E 为高斯噪声矩阵, S 为稀疏噪声矩阵。直观上讲，低秩矩阵 L 代表了伪多义空间的所在方向，而稀疏噪声矩阵 S 则在对应真多义的词对的位置上比较显著。本小节将直观说明这一点。

#	代表词对
1	after$_{1,2}$, eventually$_{0,1}$, whilst$_{0,1}$, again$_{1,2}$, finally$_{1,2}$
2	although$_{1,2}$, well$_{2,3}$, initially$_{1,2}$, more$_{1,4}$, both$_{0,2}$
3	Brian$_{1,2}$, February$_{0,2}$, Daniel$_{1,2}$, September$_{2,7}$, Frank$_{0,2}$

表 4.4 给定伪多义空间秩为 3 时的算法 3 提取出的每一个主成分方向上投影比率最大的词义对差向量，每个方向用 5 个词义对差表示。

词	近邻词	$\|\|\mathbf{S_p}\|\|_2$
prime$_{s0}$	minister, cabinet, parliament	
prime$_{s1}$	modulo, space, equivalently, real	3.35
yard$_{s0}$	touchdown, interception, kickoff	
yard$_{s1}$	lawn, backyard, garden, porch	2.75
engine$_{s0}$	jeep, truck, wheel, vehicle, car	
engine$_{s1}$	camshaft, turbine, gearbox	0.61
cat$_{s0}$	dog, pet, wolf, bird, animal	
cat$_{s1}$	dog, pets, puppy, cats, fox	0

Word	词义 #0	词义 #1
prime	总理	素数
yard	码（长度单位）	院子
engine	火车头/发动机	机械
cat	动物	动物

表 4.5 上：选中词的词义向量 k 近邻，以及对应稀疏噪声矩阵中向量的 2 范数；下：根据近邻词所推测的词义表达内容。

表4.4直观地展示了每一个提取出的主成分方向上的代表词对，我们不难发现，这些词对所对应的词一半都是副词或专有名词，且大多都是出现语境比较复杂的单义词——这正是伪多义最容易出现的情况，也是绪论一章中所介绍的我们要重点关注的伪

多义的产生背景。

此外，我们随机选择了一些比较常见的词，并在表4.5中展示了其两个词义向量的 k 近邻，以及人工根据这些近邻推测出的词义向量的含义，附以对应稀疏噪声矩阵中的向量 2 范数。可以看出，差向量所对应的稀疏噪声的强度某种程度上正比于两个词义表达意义不同的程度。特别地，我们看到，*yard* 和 *prime* 两个词表达了完全不同的含义，所以其对应稀疏向量 2 范数较大；而 *engine* 所对应的两个词义尽管细分时有区别：发动机/机械（统称），但其意义关联较大，表现为上下文相关程度较大，所以对应稀疏向量的模长上也不如前两个词一样大；而 *cat* 的两个词义向量则是明显的伪多义，其对应稀疏向量模长为 0。4.6 小节将更加细致地对 RPCA 分解词内部意义对差矩阵的效果进行量化的说明。

以上三个小节再一次直观地说明了伪多义现象是一种系统性产生的现象而非偶然。在接下来的小节，本文将通过不同的量化手段评估伪多义消除的效果，并证明相差过大的伪多义的存在对于空间的语义表达能力实际上是一种干扰。

4.5 词义级别相似度

词相似度是一个量化评估词向量质量的常见方法。其基本思想是以人工标注好的词对相似度作为标准 (gold label)，将词向量中这些词对的相似度（常见为余弦相似度）作为待打分元素。按照词对，将人工打分和词向量打分分别列为两个数列，每个位置对应同一词对。这两个数列的 Spearman rank correlation($\rho \in [-1, 1]$) 即可作为两个打相似程度的度量，其中 1 为完全正相关，-1 为完全负相关，0 为不相关。

常见的用于词向量评估的相似度数据集主要有 WordSim353[9] 和 Stanford Contextual Word Similarity(SCWS, [13])。WordSim-353 数据集给出了 353 对词对和对每个词对中词汇相似度的人工打分，人工打分的平均值将作为评估标准；而 SCWS 数据集在此之外，还包括了各个词对出现的上下文信息。

对于未给出上下文的评测数据集（WordSim-353, WS-353），我们以 *avgSim*[37] 作为评估指标；对于给出了上下文的评测数据集，我们以 *maxSimC*[33] 作为评估指标。这两个指标定义如下：

$$avgSim(w_a, w_b) = \frac{1}{|S(w_a)||S(w_b)|} \sum_{s_a \in S(w_a)} \sum_{s_b \in S(w_b)} \cos(\mathbf{v}_{s_a}^{w_a}, \mathbf{v}_{s_b}^{w_b}) \tag{4.2}$$

其中，w_a 和 w_b 表示待求词义的两个词，$S(w_a)$ 和 $S(w_b)$ 表示两个词分别对应的词义集合。

$$maxSimC(w_a, w_b) = \cos(\mathbf{v}_{\hat{s}_a}^{w_a}, \mathbf{v}_{\hat{s}_b}^{w_b}) \tag{4.3}$$

其中

$$\hat{s}_a = \text{argmax}_{s_a \in S(a)} P(s_a | C(w_a))$$

$$\hat{s}_b = \text{argmax}_{s_b \in S(b)} P(s_b | C(w_b))$$

表示在给定两个词上下问时最可能取到的词义，其计算方法是与第一章提到的上下文中心向量计算相似度，取到对应的词义的概率与计算出的相似度成正比即可。

简单来说，两个指标的含义分别是：“对两个词所有词义给出的相似度取平均”，和“在两个词所有词义中分别找到一个最适合当前上下文的词义向量并计算它们的相似度”。值得注意的是，许多评测方式都采取了一种加权平均方案 *avgSimC*[12, 18, 22, 33]，即按照选取每个词义的概率对每一对词义的相似度进行加权平均得到最终的词义对相似度。这是一种基于分布式语义假设的评估方案，也具有一定的合理性。但为了衡量词向量某一具体词义的意义表达，本文不采用这种度量方式。

模型	WS-353	SCWS
基于上下文的聚类 [13]	64.2	26.1
多义词 skip-gram (MSSG, 300D)[33]	70.9	57.3
无监督多义词 skip-gram (NP-MSSG, 300D) [33]	69.1	59.8
NP-MSSG + 算法一	68.8	62.2
NP-MSSG + 算法二	69.2	63.7
NP-MSSG + 算法三 (PCA)	69.2	65.3
NP-MSSG + 算法三 (RPCA)	69.2	**65.4**
中餐馆模型 [22]	69.5	62.4
MUSE[18]	69.4	67.9

表 4.6 不同多义词向量训练的模型在词相似度任务上的表现。

表4.6展示了不同模型在不同数据集下的表现。可以发现，在 WordSim-353 任务上，各个模型的表现都比较相似；但在 SCWS 任务上，伪多义消除算法比较有效地提

升了词义相似度的表现。值得注意的是，尽管 MUSE[18] 在 SCWS 上效果更好，但该词向量实际上是将所有同一个词的不同词义嵌入到了词向量空间当中的“一小团”，这种嵌入实际上不具有多义词向量的属性，也就和已知对于多义词向量区分度较高的 NP-MSSG[33] 没有可比性。

4.6 PCA 和 RPCA 的性能对比

在4.4中，我们提到 RPCA 对词内部意义对差矩阵的分解所产生的强噪声向量可以作为一个真多义词的指向。然而，相对地，如果我们将 RPCA 替换为 PCA 的话，意义对差对于 PCA 所得到的伪多义空间的投影所占原向量的比例也能够通过每个词向量差的伪多义程度反向地作为一个真多义的表示方式。然而，这两个表达方式是等价的吗? 我们设计了如下的 WordNet Synset 召回实验对这一点进行分析。

我们预先指定一个 PCA 和 RPCA 所分解出的伪多义空间维度 d。对于 RPCA，我们对每一个词内部意义对按照对应的强噪声向量模长从大到小进行排序，排名越靠前表明我们认为这是一对真多义；同样地，对于 PCA，我们按照词内部意义对向量在提取出的伪多义空间的投影长度比例从小到大排序，也是排名越靠前我们认为这是真多义的概率越高。

对于每一对真多义，我们采用4.2中提到的根据近邻估算 Synset 的方案对词内部意义对中的两个词估计出对应的 Synset，如果估算出的两个词义相同，则记为一次正确，否则记一次错误。如果在前 n 个词内部意义对中，某种方案都维持了较高的正确率，则说明这种方案实际上是更有效的。

图4.1展示了在不同伪多义空间维度设定下的词对数-准确率曲线。每个子图中，蓝色实线都在橙色虚线上面。根据图中的展示，RPCA 的稀疏强噪声矩阵在不同的设定下，对于 PCA 所提取出的伪多义空间都有一定的优势。

此外，表4.7展示了 PCA 和 RPCA 所挖掘出的主成分所张成的空间的各个方向对比。可以看到，RPCA 所挖掘出的第 3 个主成分和 PCA 所挖掘出的第 2 个主成分方向基本是一致的。而对于另外两个方向，RPCA 更为倾向于比较广泛的词性方向：副词；而 PCA 则转而关注了话题方向。直观上讲，前者可能出现的上下文更为丰富，这一点可以直观地展示 RPCA 在词义挖掘上有一定帮助。

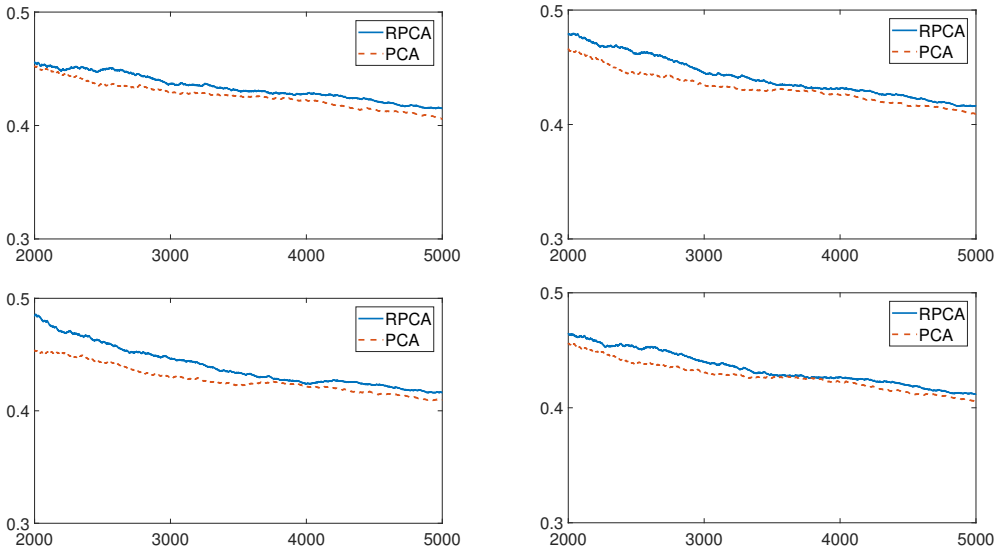

图 4.1 词对-准确率曲线。上左：$d = 1$，上右：$d = 2$, 下左：$d = 3$, 下右：$d = 5$.

	#	代表词对	特征
RPCA	1	after$_{1,2}$, eventually$_{0,1}$, whilst$_{0,1}$, again$_{1,2}$, finally$_{1,2}$	副词
	2	although$_{1,2}$, well$_{2,3}$, initially$_{1,2}$, more$_{1,4}$, both$_{0,2}$	副词
	3	Brian$_{1,2}$, February$_{0,2}$, Daniel$_{1,2}$, September$_{2,7}$, Frank$_{0,2}$	专有名词
PCA	1	income$_{2,4}$, campaigns$_{1,5}$, age$_{6,7}$, development$_{4,5}$, goals$_{2,6}$	政治话题
	2	Berlin$_{0,6}$, Martin$_{3,4}$, Greek$_{0,3}$, Jan$_{0,4/0,6}$, name$_{1,3}$	专有名词
	3	quarterback$_{3,9}$, playoff$_{3,9}$, NBA$_{0,1}$, Houston$_{1,3}$, mayor$_{0,6}$	体育话题

表 4.7 RPCA 和 PCA 的代表词对（下标表示词义对中两个词义的编号）。

4.7 句意理解

最后，本文在句意理解的下游任务上评测句子表示的效果。对于每个句子中的词，如果其在多义词向量中对应多于一个向量，我们通过 *maxSimC* 指标选出最可能的一个。我们采用对词袋子中的词直接求平均 (BoW) 的方式求出每个句子的表示。

以下实验采用 Facebook Research 发布的 SentEval[7] 框架进行评测。该框架支持若干个基于句子表示的分类任务，我们采用了 SUBJ（主客观句子分类）、TREC（问句分类）和 MSRP（段落识别）三个任务作为评测标准，并在表4.8中展示了原空间和增强（经过伪多义消除的）空间上这三个任务的表现。可以看出，增强后的空间对于提升句意理解下游任务上的表现有一定的帮助。

模型	**SUBJ**	**TREC**	**MSRP**
原空间	91.0	78.4	70.0
+ 算法一	91.2	83.2	70.6
+ 算法二	91.0	84.2	70.0
+ 算法三 (RPCA)	**92.3**	**85.4**	**71.0**

表 4.8 原空间和增强后的空间在下游任务上的表现。

第五章 总结与展望

本文对已有多义词向量的共性问题——伪多义现象进行了全方位的分析。提出检测多义词向量中的伪多义、保留真正多义是一种提升多义词向量表达能力的方式。这种方式不仅保留了多义词向量“对每个义项学习一个单独的向量以避免词义混淆”的优势，也在可能的范围内尝试去除同一词义被学习出多个词向量表示的劣势。增强后的多义词向量空间明显在许多基于理解的下游任务上表现更优；此外，本文提出的新的对于健壮主成分分析解法适用于词向量问题。本文的伪多义检测和消除算法对无监督地同时进行词义挖掘和词向量学习提供了一个新的思路。

词义 (Word Sense)，是人类理解自然语言的一个细粒度的单位，是一个离散的概念。而词向量 (Word Embeddings)，则是对于词语或词义的一个基于分布式语义假设的表征，从空间上讲是连续的。如何处理这种连续与离散之间的关系也是在机器理解过程中的一个重要问题。期待看到未来机器学习领域能够诞生更多对于平衡连续与离散之间关系的有效模型。

参考文献

[1] Amir Beck and Marc Teboulle. "*A fast iterative shrinkage-thresholding algorithm for linear inverse problems*". *SIAM journal on imaging sciences*, **2009**, *2*(1): 183–202.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent *et al.* "*A neural probabilistic language model*". *Journal of machine learning research*, **2003**, *3*(Feb): 1137–1155.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou *et al.* "*Man is to computer programmer as woman is to homemaker? debiasing word embeddings*". In: *Advances in Neural Information Processing Systems*, **2016**: 4349–4357.

[4] Elia Bruni, Gemma Boleda, Marco Baroni *et al.* "*Distributional semantics in technicolor*". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, **2012**: 136–145.

[5] Jianpeng Cheng and Dimitri Kartsaklis. "*Syntax-Aware Multi-Sense Word Embeddings for Deep Compositional Models of Meaning*". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, **2015**: 1531–1542.

[6] Ronan Collobert and Jason Weston. "*A unified architecture for natural language processing: Deep neural networks with multitask learning*". In: *Proceedings of the 25th international conference on Machine learning*, **2008**: 160–167.

[7] Alexis Conneau and Douwe Kiela. "*SentEval: An Evaluation Toolkit for Universal Sentence Representations*". In: *Proceedings of the 11th Language Resource and Evaluation Conference*, **2018**.

[8] TM Cover and JA Thomas. "*Elements of information theory: Wiley Online Library*". **1991**.

[9] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias *et al.* "*Placing search in context: The concept revisited*". In: *Proceedings of the 10th international conference on World Wide Web*, **2001**: 406–414.

[10] Xavier Glorot, Antoine Bordes and Yoshua Bengio. "*Domain adaptation for large-scale sentiment classification: A deep learning approach*". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, **2011**: 513–520.

[11] Gene H Golub and Christian Reinsch. "*Singular value decomposition and least squares solutions*". *Numerische mathematik*, **1970**, *14*(5): 403–420.

[12] Fenfei Guo, Mohit Iyyer and Jordan Boyd-Graber. "*Inducing and Embedding Senses with Scaled Gumbel Softmax*". *arXiv preprint arXiv:1804.08077*, **2018**.

[13] Eric H Huang, Richard Socher, Christopher D Manning *et al.* "*Improving word representations via global context and multiple word prototypes*". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, **2012**: 873–882.

[14] Ian T Jolliffe. "*Principal component analysis and factor analysis*". In: *Principal component analysis*. Springer, **1986**: 115–128.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski *et al.* "*Bag of Tricks for Efficient Text Classification*". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, **2017**: 427–431.

[16] George Lakoff. "*The contemporary theory of metaphor*". **1993**.

[17] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, **2008**.

[18] Guang-He Lee and Yun-Nung Chen. "*MUSE: Modularizing Unsupervised Sense Embeddings*". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, **2017**: 327–337.

[19] Omer Levy and Yoav Goldberg. "*Dependency-based word embeddings*". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, **2014**: 302–308.

[20] Omer Levy and Yoav Goldberg. "*Linguistic regularities in sparse and explicit word representations*". In: *Proceedings of the eighteenth conference on computational natural language learning*, **2014**: 171–180.

[21] Omer Levy and Yoav Goldberg. "*Neural word embedding as implicit matrix factorization*". In: *Advances in neural information processing systems*, **2014**: 2177–2185.

[22] Jiwei Li and Dan Jurafsky. "*Do Multi-Sense Embeddings Improve Natural Language Understanding?*" In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, **2015**: 1722–1732.

[23] Yitan Li, Linli Xu, Fei Tian *et al.* "*Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective.*" In: *IJCAI*, **2015**: 3650–3656.

[24] Zhouchen Lin, Risheng Liu and Zhixun Su. "*Linearized alternating direction method with adaptive penalty for low-rank representation*". In: *Advances in neural information processing systems*, **2011**: 612–620.

[25] Li Liu, Douglas M Hawkins, Sujoy Ghosh *et al.* "*Robust singular value decomposition analysis of microarray data*". *Proceedings of the National Academy of Sciences*, **2003**, *100*(23): 13167–13172.

[26] Yang Liu, Zhiyuan Liu, Tat-Seng Chua *et al.* "*Topical Word Embeddings.*" In: *AAAI*, **2015**: 2418–2424.

[27] Thang Luong, Richard Socher and Christopher Manning. "*Better word representations with recursive neural networks for morphology*". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, **2013**: 104–113.

[28] Tomas Mikolov, Kai Chen, Greg Corrado *et al.* "*Efficient estimation of word representations in vector space*". *arXiv preprint arXiv:1301.3781*, **2013**.

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen *et al.* "*Distributed representations of words and phrases and their compositionality*". In: *Advances in neural information processing systems*, **2013**: 3111–3119.

[30] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. "*Linguistic regularities in continuous space word representations*". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **2013**: 746–751.

[31] George A Miller. "*WordNet: a lexical database for English*". *Communications of the ACM*, **1995**, *38*(11): 39–41.

[32] Andriy Mnih and Geoffrey Hinton. "*Three new graphical models for statistical language modelling*". In: *Proceedings of the 24th international conference on Machine learning*, **2007**: 641–648.

[33] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos *et al.* "*Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space*". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, **2014**: 1059–1069.

[34] Yu Nesterov. "*Smooth minimization of non-smooth functions*". *Mathematical programming*, **2005**, *103*(1): 127–152.

[35] Adam Paszke, Sam Gross, Soumith Chintala *et al. Pytorch*, **2017**.

[36] Jeffrey Pennington, Richard Socher and Christopher Manning. "*Glove: Global vectors for word representation*". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, **2014**: 1532–1543.

[37] Joseph Reisinger and Raymond J Mooney. "*Multi-prototype vector-space models of word meaning*". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, **2010**: 109–117.

[38] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. "*Learning representations by back-propagating errors*". *nature*, **1986**, *323*(6088): 533.

[39] Holger Schwenk. "*Continuous space language models*". *Computer Speech & Language*, **2007**, *21*(3): 492–518.

[40] Richard Socher, Cliff C Lin, Chris Manning *et al.* "*Parsing natural scenes and natural language with recursive neural networks*". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, **2011**: 129–136.

[41] Joseph Turian, Lev Ratinov and Yoshua Bengio. "*Word representations: a simple and general method for semi-supervised learning*". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, **2010**: 384–394.

[42] Peter D Turney. "*Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase*". *Transactions of the Association of Computational Linguistics*, **2013**, *1*: 353–366.

[43] Peter D Turney, Yair Neuman, Dan Assaf *et al.* "*Literal and metaphorical sense identification through concrete and abstract context*". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, **2011**: 680–690.

[44] Peter D Turney and Patrick Pantel. "*From frequency to meaning: Vector space models of semantics*". *Journal of artificial intelligence research*, **2010**, *37*: 141–188.

[45] Jason Weston, Samy Bengio and Nicolas Usunier. "*Wsabie: Scaling up to large vocabulary image annotation*". In: *IJCAI*, **2011**: 2764–2770.

[46] Svante Wold, Kim Esbensen and Paul Geladi. "*Principal component analysis*". *Chemometrics and intelligent laboratory systems*, **1987**, *2*(1-3): 37–52.

[47] John Wright, Arvind Ganesh, Shankar Rao *et al.* "*Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization*". In: *Advances in neural information processing systems*, **2009**: 2080–2088.

[48] MK Stephen Yeung, Jesper Tegnér and James J Collins. "*Reverse engineering gene networks using singular value decomposition and robust regression*". *Proceedings of the National Academy of Sciences*, **2002**, *99*(9): 6163–6168.

[49] 伍胜健. 数学分析. 北京大学出版社, **2010**.

[50] 梅家驹. 同义词词林. 商务印书馆; 上海, **1984**.

本科期间的主要工作和成果

会议论文

1. **Haoyue Shi**[1*], Jiayuan Mao[1*], Tete Xiao[1], Yuning Jiang, Jian Sun. *"Learning Visually-Grounded Semantics from Contrastive Adversarial Samples"*. 2018. International Conference on Computational Linguistics (COLING 2018). Santa Fe, New-Mexico, USA. August 2018. (1：Equal Contribution, *: Corresponding Authors)

2. **Haoyue Shi**, Xihao Wang, Yuqi Sun, Junfeng Hu*. *"Constructing High Quality Sense-specific Corpus and Word Embedding via Unsupervised Elimination of Pseudo Multi-Sense"*. Language Resources and Evaluation Conference (LREC 2018). Miyazaki, Japan. May 2018. (*: Corresponding Author)

3. **Haoyue Shi**, Jia Chen*, Alexander G. Hauptmann. *"Joint Saliency Estimation and Matching using Image Regions for Geo-Localization of Online Video"*, 2017. ACM International Conference on Multimedia Retrieval (ICMR 2017). Bucharest, Romania. June 2017. (*: Corresponding Author)

4. **Haoyue Shi**, Caihua Li, Junfeng Hu*. *"Real Multi-Sense or Pseudo Multi-Sense: An Approach to Improve Word Representation"*. Workshop on Computational Linguistics for Linguistic Complexity (associated with COLING 2016). Osaka, Japan. December 2016. (*: Corresponding Author)

5. Shan Xu, **Haoyue Shi**, Xiaohui Duan*, Tiangang Zhu, Peihua Wu and Dongyue Liu. *"Cardiovascular Risk Prediction Method Based on Test Analysis and Data Mining Ensemble System"*. 2016. IEEE International Conference on Big Data Analysis. Hangzhou, China. March 2016. (*: Corresponding Author)

6. **Haoyue Shi**, Junfeng Hu*, Yuqi Sun, Xihao Wang. *"Improving Sense Discovery via Robust Principal Component Analysis"*. In Submission to EMNLP 2018. (*: Corresponding Author)

7. **Haoyue Shi**, Hao Zhou*, Jiaze Chen, Lei Li. *"On Tree-Based Neural Sentence Modeling"*. In Submission to EMNLP 2018. (*: Corresponding Author)

致谢

这项工作是在胡俊峰老师的指导下完成的，是对我本科阶段关于多义词向量方向研究成果的一个总结和提升。我希望在此向我的导师、也是我踏入研究大门的领路人胡俊峰老师致以最诚挚的谢意，感谢他四年以来的教导、鼓励和鞭策。

我想特别感谢孙雨奇同学和王希豪同学，他们对本工作的实验做了不少帮助；感谢吴先同学为本工作的实验环境提供的帮助；感谢张悦眉同学和沈澈同学，他们对本工作的数学基础提供了许多有用的知识和讨论。

胡俊峰老师、Prof. Alex Hauptmann、陈佳老师以及 Prof. Sam Bowman 在我成长的路上起到了不可替代的作用，尽管我现在还并不是一个成熟的研究者——那是一个博士生毕业时该做到的。在此，我想对我的论文合作者们一并表示感谢，他们是：胡俊峰老师、Prof. Alex Hauptmann、段晓辉老师、朱天刚老师、孙剑老师、陈佳老师、李磊博士、周浩博士、姜宇宁前辈、陈家泽师兄、许珊师姐、刘冬月师姐、茅佳源同学、黎才华同学、肖特特同学、孙雨奇同学、王希豪同学，与他们的合作让我学到了许多；也感谢 Google 公司, 尤其是我的两位 host 任晓祎和毛杰帮助我养成了较为规范的代码习惯，让我能够高效而有条理地进行实验。

值得一提的是，在本科最后一年里，我十分幸运地认识了茅佳源——他是我遇到的同龄人中最富活力、想象力和创造力的研究者，也是与我在研究上“口味”最为一致的同学。与他的讨论和合作让我非常愉快。我同样看到，世界上有许多以 Dr. Omer Levy 和 Dr. Sebastian Riedel 为代表的出色研究者与我们拥有一样的兴趣和理想。我对他们致以我衷心的感谢：他们不仅是我的榜样和先驱，也是我愿意在研究的道路上继续向前的不竭动力。

我要感谢刘鸣赫师兄和张悦眉、李芊、田菁曳、吴先、张馨元等好朋友。在我最难过的日子里，是他们的微笑、鼓励和陪伴给了我直面生活的勇气。感谢我的室友徐梓楠、郭秭含、张爽、于晓凡、吴萤西、姜宛彤，以及古琴社、跑协和信科女篮的伙伴们，与他们一同度过本科生涯这段美好的时光是我莫大的幸运。

我也必须感谢刁如心。如果没有认识她，我或许到现在都对艺术和美抱着强烈的理性主义偏见；如果没有与她道别，我或许仍然是一个长不大的孩子。她是一个再美

好不过的人，她在我的生命里书写了一段再美好不过的故事。

很长一段时间里，我都以为自己再也无法找回快乐，但事实却不是这样。我能重新快乐起来并全力以赴地投入研究工作，全赖李煜东一路相伴。未来的路还很长，愿能携手走过每一分、每一秒。

感谢我的家人，尤其是我的妈妈和爸爸。他们无私地给予我平等的爱和温暖，也一直为我的学业提供着最大的支持。尽管我一直追寻的理解是可遇而不可求的完美，而无理由的关爱则是同样难得的温馨。我的妈妈爸爸是也将永远是我最爱的人。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：　　日期：　　年　　月　　日