**文章编号**：1003-0077(2020)08-0010-13

# 基于长时间跨度语料的词义演变计算研究

孙琦鑫[1,2]，饶高琦[1,2,3]，荀恩东[1,2]

(1. 北京语言大学 信息科学学院，北京 100083；
2. 北京语言大学 语言资源高精尖创新中心，北京 100083；
3. 北京语言大学 汉语国际教育研究院，北京 100083)

**摘 要**：该文收集了自晚清到 21 世纪间长达 144 年的连续历时报刊语料，通过统计分析和词语分布式表示两类方法展开研究，计算并辅助识别汉语词语的词义历时演变现象。采用 TF-IDF、词频比例等多种统计分析的评价指标和目标词语在文段中的共现实词及其重合度挖掘出现词义演变的词语。针对历时语料上不同时间段的词向量对齐，采用 SGNS 训练词向量加正交矩阵投影、SGNS 递增训练和“锚点词”二阶词向量表示三种方法，其中以 SGNS 递增训练效果最佳。针对自动发现的词义演变现象，采用目标词历时自相似度和锚点词历时相似度的分析方法，并利用近邻词来明确目标词变迁前后的词义。

**关键词**：词义演变；历时语料；分布式表示

**中图分类号**：TP391 **文献标识码**：A

## A Study on Semantic Evolution Computation with Diachronic Corpus

SUN Qixin[1,2], RAO Gaoqi[1,2,3], XUN Endong[1,2]

(1. School of Information Science, Beijing Language and Culture University, Beijing 100083, China;
2. Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China;
3. Institute of International Chinese Language Education, Beijing Language and Culture University, Beijing 100083, China)

**Abstract**: This paper collected a diachronic corpus of Chinese newspapers and periodicals for the past 144 years dated back to the late Qing Dynasty. A study on word semantic evolution computation is conducted for Chinese via statistical analysis and word distributed representation. Chinese word with potential semantic evolution is first discovered by context overlapping of content words via TF-IDF, word frequency ratio and other statistical indicators. Then, to align the word embeddings derived from corpus of different time periods, three methods are examined: orthogonal matrix alignment after SGNS training, second-order word vector representation and SGNS incremental training (which bears top performance). Finally, the word semantic evolution is identified by the diachronic self-similarity of the candidate word and the diachronic similarity of anchor words, with neighboring words as the description of the word meaning in the evolution.

**Keywords**: word semantic evolution; diachronic corpus; distributed representation

## 0 引言

随着社会的飞速发展，语言也处在一个不断发生变化的过程中。词是语言中具有意义且能独立运用的最小单元，也是最能直接体现语言变迁的语言单位。其中，词的变化表现为新词语的产生、新词义的产生、旧词语的消亡、旧词义的消亡等。在科技飞速发展的今天，随之产生的词语变化也给自然语言理解相关任务造成了巨大的障碍。例如，“苹果”一词早期只表示一种水果，而随着科技发展，智能手机、电脑的出现，“苹果”一词也有了新的含义，特指

**收稿日期**：2019-09-16 **定稿日期**：2019-10-09

**基金项目**：教育部人文社科基金(20YJC740050)；北京语言大学青年英才培养计划(1090/501321102)；北京语言大学中央高校基本科研业务费(19YJ130005)

美国的苹果公司及其产品。

从中国近代以来，中国社会经历了晚清、民国和共和国等重要历史时期，期间中国与世界都发生了重大的社会变革。政治、科技、文化、经济等与人民生活密切相关的要素都发生着日新月异的变化，新鲜事物层出不穷，语言也发生了重要的改变。词语正是我们研究语言历时演变、分析社会变化的重要工具。研究词义变化是研究词语变化的直观方式，词义的变化通过词语的使用情况体现，因而研究历时语料中词语的分布和频率等形式化特征，是探究词义变化的良好切入点。

从研究历史上来看，许多语言学上的理论研究都投入大量精力在记录、探索不同类别词语的语义变化上。传统语言学中的训诂就关注词语语义在漫长历史长河中的演变，并有丰富的成果，其方法是传统的对特定作品（小规模语料）进行分析研究[1-3]。王惠[4]从《现代汉语词典》收录的条目中提取出现代汉语多义词列表，利用词频统计分析，找出词频与词义之间的内在反向关联。贾佳[5]通过对《儒林外史》的词频统计及相关定量统计，分析《儒林外史》词汇的发展演变轨迹。金观涛和刘青峰[6]更是利用词频具体描述和分析意识形态更替的语言学痕迹。饶高琦等[7]基于历时 70 年的报刊语料，使用多种统计方法计算词语的历时使用情况，并对词语的稳定性、词语的覆盖度以及词语的时间区分性能等进行逐一考察，筛选出稳定词表。

然而，词义演变并不总是直接伴随词频变化的。有时候这种变化并不直接，因此如果能构建模型直接表示词语含义，效果会优于统计分析的研究方法。近几年词语分布式表示[8]技术日趋成熟，并被广泛应用于自然语言处理的各个研究领域中。词义演变研究中也有不少学者基于分布式表示做出了一些尝试，并证明该方法在解决词义演变任务上十分有效[9-10]。Kulkarni 等人[11]证实用向量表示词语的方法在检测词义演变问题上要优于基于词频的方法。该方法能更精确地追踪词义的演变过程，并给出解释。Kulkarni 等人还提出在对不同时间段上训练的词向量计算相似度之前，首先应将各个时间段上训练出的词向量结果放到同一个矢量空间中，并利用线性变化来保护各个词向量空间的结构稳定性。Zhang 等人[12]在同一时期也提出在保持距离的基础上投影，并在随后的研究[13]中将该方法加以扩展应用，同时利用整个模型的线性投影和近邻词集合来将待查询词语映射到对应时间段上。

通过调研已有工作，我们发现可以进行词义演变的历时语料主要包括 Google Books Ngrams① 数据集、Corpus of Historical American (COHA②)以及人民日报历时 60 年数据集等，多为英语语料，中文语料数量少，时间跨度短。Google Ngrams 数据集中虽包含中文语料，但皆为图书类文本，其词语的时间敏感性差，从作者写书到图书出版所经历时间跨度大，在反映词语的历时演变上存在一定的滞后性。

由于中文语料的限制，词义演变的相关研究中针对汉语词语进行的研究分析方法单一，且耗费大量人力物力，缺乏全局高效的研究手段。

本文基于长时间跨度的中文报刊语料，计算并辅助识别汉语词语的词义历时演变现象，结合语料的特点采用统计分析和分布式表示两类方法来进行研究。

## 1　长时间跨度报刊语料建设

### 1.1　现代历时语料

《人民日报》是中国第一大报，是中国对外进行文化交流的一个重要窗口和展现中国发展的舞台。本文收集了 1946 年到 2015 年历时 70 年的《人民日报》，经过统计，其包含约 12 亿字（包括中文单字、英文单字母、空格），约 2 650 万句，约 3.5GB。年度数据分布情况如图 1 所示。

此外，我们还收集了 1949 年到 2007 年（其中缺少 2000 年数据）总共 58 年的《贵州日报》作为现代报刊数据的补充。其报道内容贴近百姓生活，具有较好的权威性和公信力。经过统计，其包含约 7 亿字，约 1 493 万句，约 2GB。年度数据分布情况如图 2 所示。

### 1.2　近代历时语料

19 世纪末到 20 世纪初是现代汉语孕育并发展成熟的关键时期。为了在更长的时间跨度上探索词义演变现象，本文还搜集了从 1872 年创刊到 1949 年停刊，历时 78 年的《申报》数据。

本文收集的《申报》数据是在图片基础上，经过图像识别技术获取的《申报》电子版数据。经过统计，《申报》共包含约 9.1 亿字，约 2.6GB，均为繁体字表示。年度数据分布情况如图 3 所示。

① https://books.google.com/ngrams

② https://www.english-corpora.org/coha/

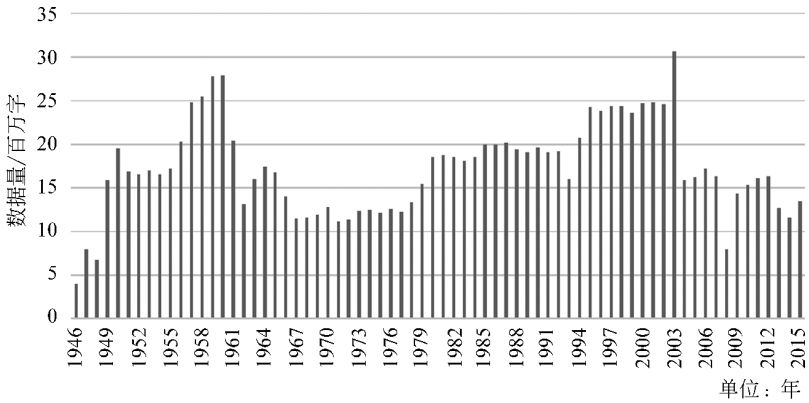


图1 《人民日报》数据量

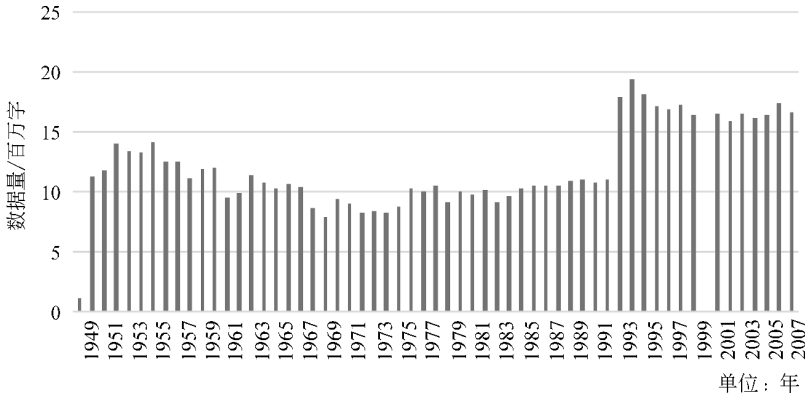


图2 《贵州日报》数据量

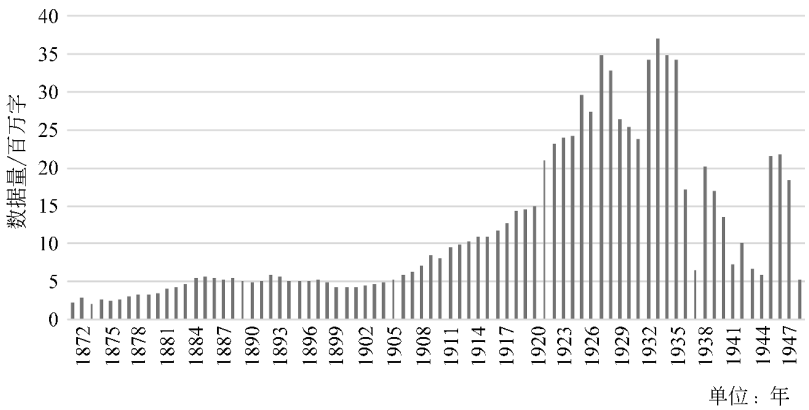


图3 《申报》数据量

针对《申报》语料特点，通过添加少量人工分词，提取出种子词表再对其进行二次分词来提高语料的分词正确率，并将繁体表示转化为简体表示，以统一历时语料的字表示。

## 2 基于统计分析的词义演变研究

### 2.1 基于统计分析的词义演变研究方法

词频是指某给定词语在某给定文件中出现的次数，是最常用的词语量化指标。这个数字通常会被正规化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，但该词频并不能说明该词语在长文件中就有更高的重要性）。另外，词频逆文档频（TF-IDF）可以在全局上利用频率信息观察词义演变情况。词语共现是指以一定频率共现于同一语篇中的一组词，经常被用于词语搭配、词义消歧等研究。

## 2.2　实验及结果分析

本文对收集到的 144 年历时语料，使用 jieba[①] 分词，按照时间顺序排序，划定五年为一个时间段，共划分为 29 个时间段。该划分既保证切分出来的时间段足够多，可以有效反映词义变迁的情况，也尽可能地扩充了单位时间段内的数据量。此外，没有选择按照重大历史事件作为分隔点也是为了尽可能保证划分后各个时间段语料大小分布均匀。其中，第一个时间段较为特殊，为 1872 年至 1875 年，共四年。

### 2.2.1　基于词频的实验及结果分析

词频逆文档频（TF-IDF）由词频（TF，Term Frequency）和逆文档频（IDF，Inverse Document Frequency）两部分组成，是基于词频常用的一种评价指标，如式(1)～(3)所示。

$$\mathrm{TF\text{-}IDF}(w) = \mathrm{TF}(w) \cdot \mathrm{IDF}(w) \tag{1}$$

$$\mathrm{TF}(w) = \frac{F_w}{F} \tag{2}$$

$$\mathrm{IDF}(w) = \log\left\{\frac{D}{(D_w + 1)}\right\} \tag{3}$$

其中 $F_w$ 和 $D_w$ 分别表示词 $w$ 在整个历时语料中出现的次数和整个语料中包含词 $w$ 的文档数，$F$ 和 $D$ 则表示整个历时语料中全部词次数和文档数。根据前文，已知将 144 年的历时语料按照五年一个时间段切分为 29 个时间段，故本文中文档总数为 29。经过计算历时语料上所有词语的 TF-IDF 值，并按照降序排列。表 1 是选取的 TF-IDF 值最大的 10 个词，表 2 选取的是 TF-IDF 值最小的 10 个词。

**表 1　TF-IDF 值大的 10 个词**

| 词 | TF-IDF 值 |
|---|---|
| 毛主席 | 0.000 415 |
| 苏联 | 0.000 175 |
| 社会主义 | 0.000 173 |
| 美帝国主义 | 0.000 104 |
| 四人帮 | 0.000 104 |
| 习近平 | 0.000 057 |
| 三个代表 | 0.000 046 |
| 非典 | 0.000 034 |
| 林彪 | 0.000 025 |
| 法轮功 | 0.000 025 |

**表 2　TF-IDF 值小的 10 个词**

| 词 | TF-IDF 值 |
|---|---|
| 也 | −0.000 093 |
| 等 | −0.000 094 |
| 为 | −0.000 129 |
| 有 | −0.000 137 |
| 之 | −0.000 166 |
| 是 | −0.000 195 |
| 和 | −0.000 282 |
| 了 | −0.000 313 |
| 在 | −0.000 328 |
| 的 | −0.001 351 |

观察表 1 和表 2 中的词可以发现，TF-IDF 值大的词语时间敏感性强，如“四人帮”是文革时期特有的名称，“非典”2002 年在中国广东顺德首发，一度成为当时社会各界密切关注的热点事件。TF-IDF 值小的词语则多为语言中的基础词语，如基础动词“是”、常用助词“的”等。

利用词频比例可有效挖掘演变词语。计算词表中全部词语在不同时间段上的词频比例，该比例称为 $P$ 值，并选择词语在各时间段上的最大 $P$ 值和最小正 $P$ 值，进行除法运算得到商，该商值称为 $U$ 值。根据 $U$ 值大小确定可能发生词义演变的词语。如表 3 所示。

由表 3 可以看出，部分词频变化明显的词是由于近现代用语习惯不同造成的。如“钦此”“禀”等词，旧时对帝王的决定、命令或其所做的事冠以“钦”字，而下级对上级陈述报告多用“禀”字。另有部分名词变化明显，是由于名词所指事物本身的变迁造成的，如“都督”是中国古代军事长官的一种，兴于三国，民国初年各省也设有都督，兼管民政。现代社会已没有这种官位称呼。“军机大臣”一词同理。而“、”等标点变化明显，多是由于人们书写习惯改变造成的。近代语料通篇基本没有标点，而随着社会发展，标点符号的使用越来越规范，标点在语料中也成为高频出现的一类符号。

### 2.2.2　基于共现搭配的实验及分析

在历时语料上，综合考虑语料平均句长等因素，

① https://github.com/fxsjy/jieba

**表 3 $U$ 值最大的 10 个词**

| 词 | $U$ 值 | 最大 $P$ 值和最小 $P$ 值时间段 |
|---|---|---|
| 钦此 | 18 902.60 | 最大 $P$ 值 1876—1880<br>最小 $P$ 值 1931—1935 |
| 饬 | 14 181.00 | 最大 $P$ 值 1896—1900<br>最小 $P$ 值 1976—1980 |
| 旨 | 14 063.00 | 最大 $P$ 值 1876—1880<br>最小 $P$ 值 1996—2000 |
| 、 | 12 891.21 | 最大 $P$ 值 1926—1930<br>最小 $P$ 值 1936—1940 |
| 已 | 12 791.55 | 最大 $P$ 值 1872—1875<br>最小 $P$ 值 1956—1960 |
| 都督 | 11 508.29 | 最大 $P$ 值 1911—1915<br>最小 $P$ 值 1996—2000 |
| 禀 | 9 858.36 | 最大 $P$ 值 1906—1910<br>最小 $P$ 值 2001—2005 |
| 供称 | 8 731.55 | 最大 $P$ 值 1891—1895<br>最小 $P$ 值 1986—1990 |
| 本馆 | 8 675.07 | 最大 $P$ 值 1872—1875<br>最小 $P$ 值 1981—1985 |
| 军机大臣 | 8 118.21 | 最大 $P$ 值 1876—1880<br>最小 $P$ 值 2001—2005 |

确定共现窗口前后大小均为 5。分时间段统计与每个词在指定窗口大小中共同出现的名词、动词和形容词(后文统称实词)及其共现次数,得到每个词在不同时间段的前 100 个高频共现实词。

图 4 是“透明”一词在不同时间段上与前一时间段的前 100 个共现实词重合个数的柱状图。由图可知,在 1971 年到 1975 年时间段上,共现实词重合个数少,且在随后的时间段中,重合个数逐年增多,且趋向稳定。针对此现象进一步查看“透明”在 1971 年到 1975 年时间段上的高频共现实词,以及其前后时间段中共现实词重合度高的时间段上的高频共现实词,分别选取 1961 年到 1965 年和 2006 年到 2010 年,如表 4、表 5 和表 6 所示。

综合表格可知,早期“透明”出现频次低,共现实词个数少,基础词语占比较高,并且词义单一,用于形容(物体)能透过光线,比如形容水是无色透明的液体。随着社会发展,“透明”一词产生新的含义,在报刊语料中高频出现,多用于形容公开、不隐藏,比如政府某项工作透明公开。

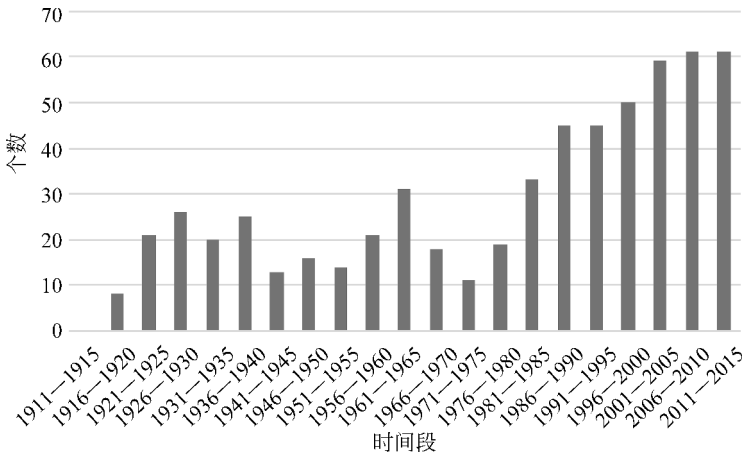


图 4 “透明”前 100 个共现实词历时重合个数

**表 4 “透明”1961 年到 1965 年时间段上的高频共现实词**

| 词/词性 | 共现次数 |
|---|---|
| 是/v | 42 |
| 有/v | 19 |
| 像/v | 18 |
| 玻璃/n | 16 |
| 液体/n | 13 |
| 晶莹/a | 12 |
| 塑料/n | 11 |
| 水晶/n | 9 |

续表

| 词/词性 | 共现次数 |
|---|---|
| 色彩/n | 8 |
| 薄膜/n | 7 |
| 如/v | 6 |
| 看到/v | 6 |
| 半透明/n | 5 |
| 看见/v | 5 |
| 纯洁/a | 5 |
| 单纯/a | 5 |

续表

| 词/词性 | 共现次数 |
|---|---|
| 发亮/v | 5 |
| 白色/n | 5 |
| 清水/n | 3 |
| 纯净/a | 3 |

**表 5　“透明”1971 年到 1975 年时间段上的高频共现实词**

| 词/词性 | 共现次数 |
|---|---|
| 晶莹/a | 6 |
| 清澈/a | 5 |
| 是/v | 4 |
| 塑料薄膜/n | 3 |
| 材料/n | 3 |
| 使/v | 3 |
| 陶瓷/n | 3 |
| 蒸灯/n | 3 |
| 井水/ns | 2 |
| 水/n | 2 |
| 吹动/v | 2 |
| 液压机/n | 2 |
| 塑料/n | 2 |
| 琉化/nz | 2 |
| 塑料管/n | 2 |
| 颜色/n | 2 |
| 清彻/a | 1 |
| 水位/n | 1 |
| 纯净/a | 1 |
| 溶液/n | 1 |

**表 6　“透明”2006 年到 2010 年时间段上的高频共现实词**

| 词/词性 | 共现次数 |
|---|---|
| 公开/ad | 658 |
| 是/v | 140 |
| 开放/v | 123 |
| 规范/n | 119 |
| 高效/a | 110 |
| 信息/n | 92 |
| 公平/n | 88 |

续表

| 词/词性 | 共现次数 |
|---|---|
| 运行/v | 81 |
| 权力/n | 73 |
| 建立/v | 71 |
| 行政/n | 62 |
| 环境/n | 61 |
| 政府/n | 57 |
| 原则/n | 56 |
| 坚持/v | 56 |
| 制度/n | 45 |
| 政策/n | 43 |
| 廉洁/a | 42 |
| 民主/n | 42 |
| 清澈/a | 18 |

## 3　基于分布式表示的词义演变研究

基于统计分析的方法，证实了语言学、社会学中的许多现象，借用数据说明，使其结论更严谨、科学，有助于语言学、社会学的发展。但必须指出，仅仅基于频率信息和共现搭配信息，无法从整体上反映词汇系统中词语的变化情况，如词频和词义变化并非有效关联，高频词语常因用语习惯产生新的词义，而频率信息无法有效获取该变化。而在本节中，则从计算的角度出发，利用分布式表示来研究词义演变现象。

### 3.1　基于分布式表示的词义演变研究方法

#### 3.1.1　锚点词假设

斯瓦迪士核心词列表（Swadesh list）是美国语言学家莫里斯·斯瓦迪士在 1940 年代到 1950 年代提出的一个词表[14]。他认为，基本上所有语言的词汇都应该包含此表中的这 200 词，且学会使用该词表内的词语，可以满足语言的基本沟通交流需要。

在此核心词列表上进行统计，确定出现在本研究所用的历时语料中所有时间段中的 186 个词为本文研究所用的“锚点词”，如表 7 所示。

#### 3.1.2　对齐词向量的方法

本文尝试了三种方法，实现词向量的对齐操作。

表7 本文所用锚点词列表

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 我 | 短 | 草 | 心 | 挖 | 浮 | 烧 | 近 |
| 你 | 窄 | 绳 | 肝 | 游 | 流 | 路 | 远 |
| 他 | 薄 | 肉 | 喝 | 飞 | 冻 | 山 | 右 |
| 我们 | 女 | 血 | 吃 | 走 | 肿 | 红 | 左 |
| 你们 | 男 | 骨 | 咬 | 来 | 日 | 绿 | 在 |
| 他们 | 人 | 蛋 | 吸 | 躺 | 星 | 黄 | 里 |
| 这 | 妻 | 角 | 吐 | 坐 | 水 | 白 | 与 |
| 那 | 夫 | 尾 | 吹 | 站 | 雨 | 黑 | 和 |
| 这里 | 母 | 发 | 呼吸 | 转 | 河 | 夜 | 若 |
| 那里 | 父 | 头 | 笑 | 落 | 湖 | 白天 | 因 |
| 谁 | 动物 | 耳 | 看 | 给 | 海 | 温 | 名 |
| 什么 | 鱼 | 眼 | 听 | 拿 | 盐 | 冷 | |
| 哪 | 鸟 | 鼻 | 知 | 挤 | 石 | 满 | |
| 何时 | 狗 | 口 | 想 | 磨 | 沙 | 新 | |
| 如何 | 蛇 | 牙 | 嗅 | 洗 | 尘 | 旧 | |
| 不 | 虫 | 舌 | 怕 | 擦 | 地 | 好 | |
| 所有 | 树 | 指甲 | 睡 | 拉 | 云 | 坏 | |
| 少 | 森 | 脚 | 住 | 推 | 雾 | 脏 | |
| 其他 | 棍 | 腿 | 死 | 扔 | 天 | 直 | |
| 大 | 果 | 膝 | 杀 | 系 | 风 | 圆 | |
| 长 | 种 | 手 | 斗 | 缝 | 雪 | 尖 | |
| 宽 | 叶 | 腹 | 击 | 计 | 冰 | 滑 | |
| 厚 | 根 | 肠 | 切 | 说 | 烟 | 湿 | |
| 重 | 树皮 | 颈 | 分 | 唱 | 火 | 干 | |
| 小 | 花 | 背 | 刺 | 玩 | 灰 | 对 | |

(1) SGNS+正交 Procrustes 对齐

使用 Word2Vec 中的 Skip-Gram 模型在已经划分好的 29 个时间段上训练词向量,每个词的向量维度是 100,训练的窗口大小为 7,使用负采样,负采样个数设为 9,采样阈值为 1E-4,迭代次数设为 10,频次少于 10 的词丢弃,同时以二进制和普通存储的方式存储词向量。该方法称为 SGNS。

由于每个时间段的训练都是各自随机初始化的,因此每个时间段都处于各自的向量空间中,互不相通。采用 Hamilton 等[15]提出的 Procrustes 对齐的方式映射 29 个时间段上的词向量到同一个向量空间上,方便后期进行相似度计算。

(2) SGNS 递增训练

使用 Word2Vec 的 Skip-Gram 模型在已经划分好的 29 个时间段上训练词向量,优化方法及参数设置同方法(1)。区别之处在于,方法(2)中只有第一个时间段的训练是随机初始化的,之后时间段的词向量都是使用其前一个时间段上训练出的词向量初始化,若该词在前一个时间段中不存在词向量,再对其进行随机初始化。此方法得到的 29 个时间段上的词向量都在同一个向量空间中,故不需要额外对齐操作,直接可以进行相似度计算。

(3) "锚点词"二阶词向量

利用"二阶词向量"的思想和"锚点词"词表,计算历时语料的二阶词向量。首先,参数设置及训练方法均同方法(1),分别训练每个时间段上的词向量。其次,选取"锚点词"列表中的 100 个词用于此方法。在每个时间段中分别计算词表中词和 100 个"锚点词"的余弦相似度,将其和锚点词的余弦相似度拼接成 100 维的向量,用该向量表示该词在该时间段的词向量。由于"锚点词"具有稳定性的特点,可以认为"锚点词"是固定不变的,用词和"锚点词"的位置关系来确定词在向量空间中的位置。使用该"二阶词向量"可以达到对齐效果,使得不同时间段的词向量具有可比性。

3.1.3 获取词义变化的方法

采用不同方法解决了词向量所处空间不一致问题后,本文采用两种不同的方法获取词义变化。

(1) 目标词自身历时相似度+$k$ 近邻词列表

利用训练所得到的词向量,计算目标词在不同时间段上词向量之间的余弦相似度,研究目标词自身余弦相似度的变化,并结合目标词的近邻词重合度来分析目标词的词义演变现象。

在不同时间段上,对目标词提取余弦相似度最高的 $k$ 个词,这些词即为 $k$ 近邻词。在本文中,近邻词只提取实词,个数 $k$ 设为 100。根据目标词的 100 个近邻实词的重合情况分析目标词的词义变化。

(2) 目标词与锚点词的历时相似度

基于"锚点词"原理,在经过对齐处理的词向量上,计算目标词在每个时间段上和"锚点词"的余弦相似度,研究其余弦相似度的变化曲线,根据变化曲线的波动程度来衡量目标词的词义变化情况。

### 3.2 实验及结果分析

由于 144 年历时语料词语数量大,且包含许多虚词等对于考察词义演变现象存在干扰的词汇。因

此在本节实验中，主要针对在前一节中筛选出来的易发生词义演变的动词、名词和形容词进行实验，共约 7 000 词。

为提高后期训练词分布式表示的质量，使用 jieba 进行分词和词性标注操作（《申报》数据在进行分词和词性标注操作时需要使用种子词表），将词性标注为数词的词语，全部替换为“M”。其次，基于现代和近代各语料数据量大小的分析和词频统计，对数据中的低频词进行处理，替换低频的人名、地名、机构名为“NR”“NS”和“NT”。

同统计分析方法的时间划分，确定五年为一个时间段。其中，第一个时间段为 1872 年至 1875 年，共四年。经过划分后共为 29 个时间段。

3.2.1　SGNS＋正交 Procrustes 对齐

在 SGNS 加正交投影训练出的历时词向量中，选取名词“导师”作为目标词进行分析。首先，计算目标词在历时词向量上的自身变化情况。这里采取两种方法比较自身相似度变化情况：一种是计算每个时间段上目标词和前一个时间段上目标词的相似度；另一种是计算每个时间段上的目标词和某指定时间段上的目标词的相似度。由于部分词语在近代并没有广泛使用，本文选择 29 个时间段中最后一个时间段作为指定时间段，即计算目标词在每个时间段上的词向量和在最后一个时间段上的词向量的相似度。此外，随机从本文的锚点词列表中选择词作为后续分析所用锚点词，这里选择“手”，计算目标词和锚点词的历时相似度变化。折线图如图 5 所示。

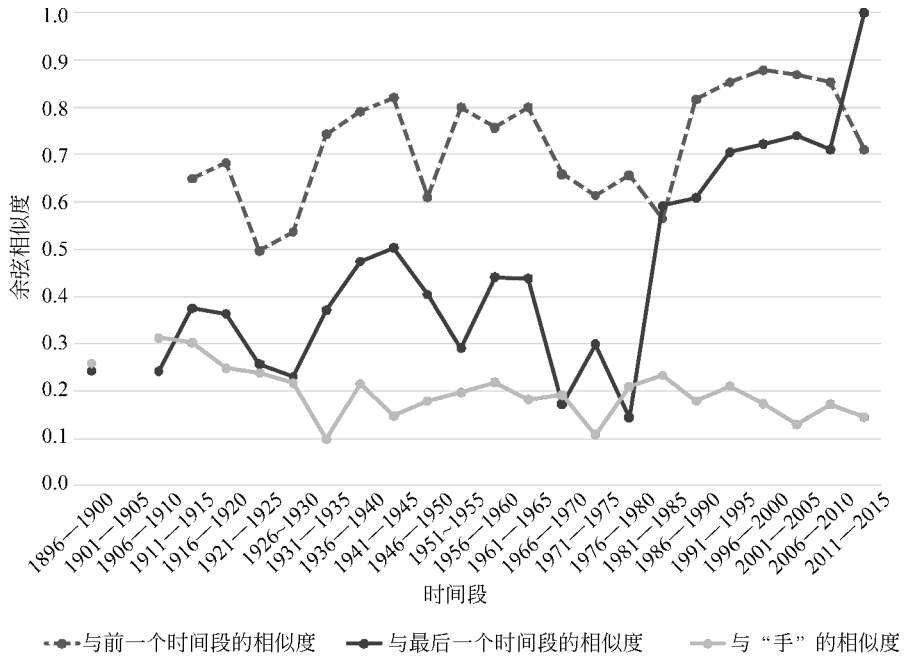


图 5　SGNS＋正交 Procrustes 对齐历时词向量“导师”历时相似度变化

另外，考查了目标词在每个时间段上的前 100 个近邻实词和前一个时间段上的前 100 个近邻实词的重合个数。如图 6 所示。

综合图 5 和图 6 可知，“导师”一词发生明显变化的三个关键时间段，分别是 1921 年到 1935 年、1951 年到 1955 年、1966 年到 1980 年。为查明具体词义的变化情况，查看上述时间段“导师”的邻近实词，如表 8 和表 9 所示。联系社会历史事件，中国民主革命伟大先行者孙中山，投身于革命事业，被尊称为“导师”，逝世于 1925 年。故这个时期的报刊内容中，“导师”一词多与孙中山、革命密切联系。1953 年苏联领导人斯大林逝世，他被当时的共产主义者

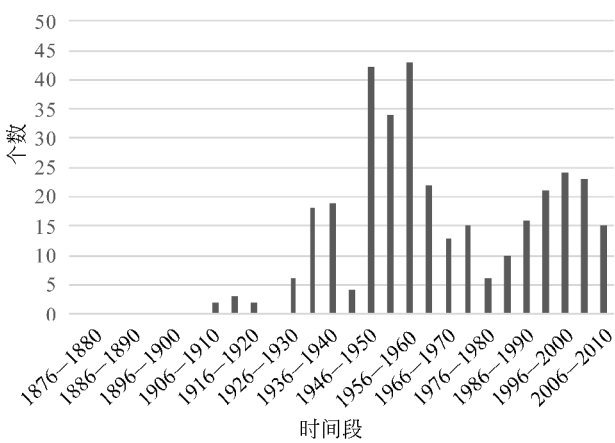


图 6　“导师”前 100 个近邻实词历时重合个数

**表8 “导师”不同时间段上的近邻实词及相似度**

| 1916—1920 | | 1921—1925 | | 1951—1955 | |
|---|---|---|---|---|---|
| 近邻词 | 相似度 | 近邻词 | 相似度 | 近邻词 | 相似度 |
| 研究法 | 0.755 | 痛悼 | 0.715 | 伟大领袖 | 0.826 |
| 教育学 | 0.723 | 节哀顺变 | 0.690 | 大元帅 | 0.760 |
| 模楷 | 0.716 | 尊翁 | 0.655 | 斯大林 | 0.748 |
| 常职 | 0.692 | 节哀 | 0.654 | 逝世 | 0.741 |
| 智德 | 0.690 | 遗志 | 0.650 | 科学共产主义 | 0.712 |
| 诲人不倦 | 0.689 | 景仰 | 0.647 | 列宁 | 0.694 |
| 融会贯通 | 0.686 | 三民主义 | 0.645 | 马克思 | 0.688 |
| 青年人 | 0.685 | 国父 | 0.637 | 敬爱 | 0.684 |
| 培育 | 0.685 | 革命事业 | 0.620 | 永别 | 0.678 |
| 圭臬 | 0.682 | 革命领袖 | 0.619 | 悼念 | 0.667 |

**表9 “导师”不同时间段上的近邻实词及相似度**

| 1966—1970 | | 1976—1980 | | 2011—2015 | |
|---|---|---|---|---|---|
| 近邻词 | 相似度 | 近邻词 | 相似度 | 近邻词 | 相似度 |
| 领袖 | 0.814 | 伟大领袖 | 0.893 | 博士生 | 0.800 |
| 革命领袖 | 0.755 | 毛主席 | 0.820 | 研究生 | 0.696 |
| 敬爱 | 0.725 | 敬爱 | 0.802 | 硕士 | 0.676 |
| 毛泽东 | 0.689 | 领袖 | 0.784 | 数学系 | 0.675 |
| 深孚众望 | 0.670 | 毛泽东 | 0.768 | 外语系 | 0.673 |
| 敬仰 | 0.670 | 伟大 | 0.752 | 美国麻省理工学院 | 0.667 |
| 解放者 | 0.665 | 崇敬 | 0.744 | 主修 | 0.659 |
| 天才 | 0.664 | 缅怀 | 0.729 | 教授 | 0.652 |
| 杰出 | 0.662 | 逝世 | 0.688 | 南昌大学 | 0.647 |
| 伟大 | 0.651 | 追思 | 0.685 | 助教 | 0.646 |

尊称为“导师”,故该时间段中,“导师”的近邻实词出现“斯大林”“逝世”“科学共产主义”等词。文化大革命时期,毛泽东被称为“伟大的领袖,伟大的导师,伟大的舵手”,在这一时期,“导师”特指毛泽东,故具有较高的相似度。现代“导师”更多指老师,特别是带领硕士生和博士生学习和研究的老师。

追本溯源,查看语料中句子,也可以证实此分析。

**例1** “本党总理孙中山先生是国民革命的领袖世界被压迫的民众解放的导师”——摘自1926年到1930年语料。

**例2** “本期为纪念全世界劳动人民的伟大领袖和导师斯大林逝世,发表题为‘斯大林的事业永垂不朽!’”——摘自1951年到1955年语料。

**例3** “毛主席是我国各族人民最伟大的领袖,是全国青年最敬爱的导师。”——摘自1966年到1970年语料。

**例4** “当以优异成绩完成课题准备回国时,导师一再挽留,将为他提供高档住所和先进的实验室。”——摘自2011年到2015年语料。

3.2.2 SGNS递增训练

与3.2.1节类似,选择名词“导师”作为目标词进行分析。作出“导师”在历时语料中的自身相似度变化曲线图、与锚点词“手”的相似度变化曲线,以及在不同时间段上“导师”和前一个时间段上前100个近邻实词的重合个数柱状图,如图7和图8所示。

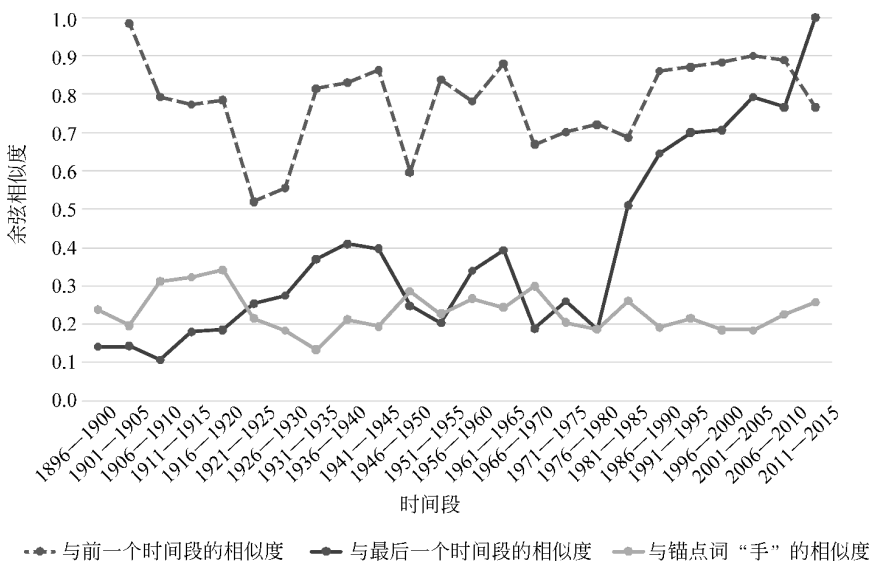


图7 SGNS递增历时词向量“导师”历时相似度变化

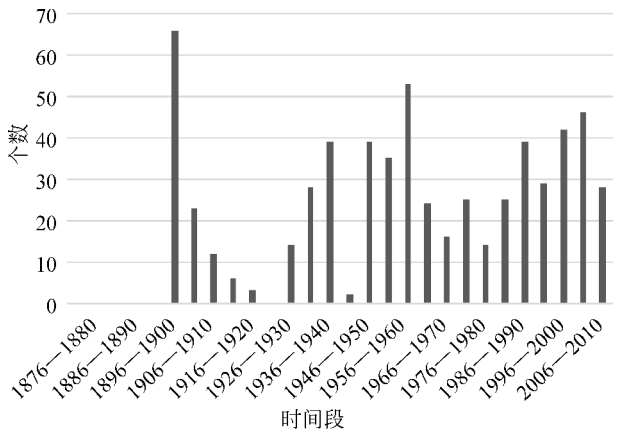


图 8　“导师”前 100 个近邻实词历时重合个数

由图可以发现，根据此方法得出变化趋势和 3.2.1 节中的结果基本一致。

3.2.3　二阶词向量

本节的实验数据仅为经过前期处理的 70 年《人民日报》数据，利用方法(3)得到 100 维的“二阶词向量”。仍然选择名词“导师”作为目标词进行分析。

图 9 是“导师”二阶词向量表示下的逐年相似度变化趋势图。根据前两组实验已知，“导师”一词在特定时间段内是存在明显变化的，而根据本实验所做出的折线图，却没有明显的波动。

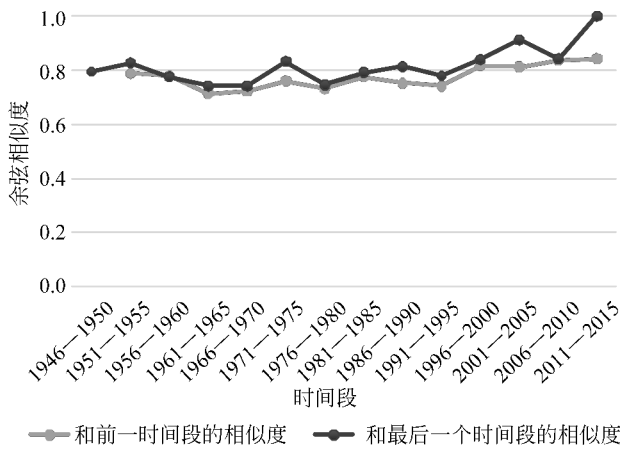


图 9　二阶词向量“导师”历时相似度变化

随后，作出图 10，该图是“导师”在 100 维二阶词向量表示下的前 100 个近邻实词的重合程度。

与相似度变化趋势不同，重合词个数的变化较为明显。再分析近邻实词列表，如表 10 所示。

由表 10 得知，近邻实词符合前两组实验中得出的“导师”一词的变化结论，但仔细观察相似度发现，数值差均非常小，这也是导致词语自相似曲线图没有明显波动的原因。由于用与 100 个锚点词的相似度来代表原本词向量，每一维的数值表示都相差不大，无法通过计算有效区分不同词义的词语。

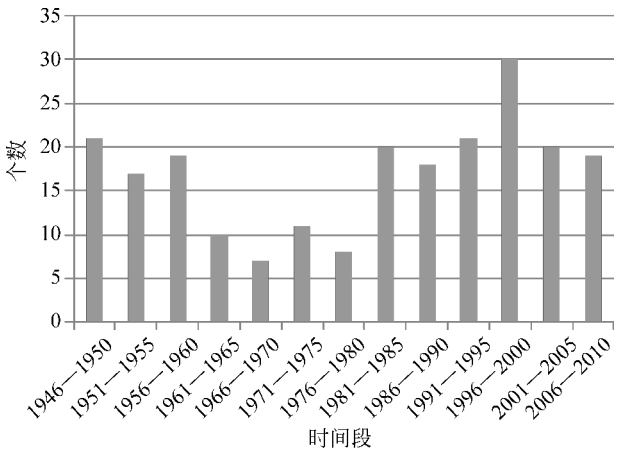


图 10　“导师”前 100 个近邻实词历时重合个数

**表 10　“导师”不同时间段上的近邻实词及相似度**

| 1966—1970 | | 2006—2010 | |
|---|---|---|---|
| 近邻词 | 相似度 | 近邻词 | 相似度 |
| 领袖 | 0.965 | 教授 | 0.979 |
| 近邻词 | 相似度 | 近邻词 | 相似度 |
| 革命领袖 | 0.964 | 副教授 | 0.977 |
| 毛泽东 | 0.958 | 博士生 | 0.976 |
| 解放者 | 0.957 | 讲师 | 0.973 |
| 敬爱 | 0.949 | 青年教师 | 0.964 |
| 马克思列宁主义者 | 0.948 | 名教授 | 0.964 |
| 伟大领袖 | 0.945 | 研究生 | 0.964 |
| 敬仰 | 0.945 | 苏步青 | 0.963 |
| 英明领袖 | 0.944 | 院士 | 0.963 |
| 主席 | 0.940 | 杨振宁 | 0.961 |

### 3.3 实验方法对比分析

基于3.2节对各个方法的实验，可以发现，三种训练历时词向量的方法得到的历时词向量，在时间段内都有较为不错的词义区分效果，可以将词义相关的词聚类到一起，词的近邻实词也可以有效反映该时间段内词语的词义。

Hamilton等[15]在考察英文语料上的词义演变现象时，主要比较PPMI、SVD+正交Procrustes对齐、SGNS+正交Procrustes对齐三种历时词向量计算方法，缺少对中文语料以及SGNS递增训练、二阶词向量的比较。在中文历时词向量的词义相似度计算上，SGNS加正交Procrustes对齐方法和SGNS模型递增训练的方法都有不错的表现，如图11所示。特别是SGNS模型递增训练得到的词向量可以有效消除因为某些年份词语频次低导致词向量缺失的问题。而二阶词向量法，由于最终的词向量差别小，难以通过计算有效区分不同词义。

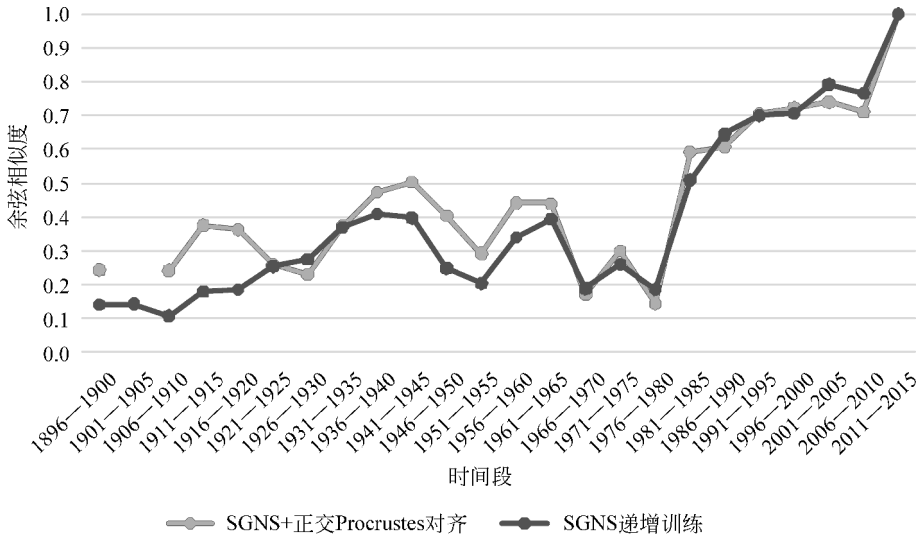


图11 不同对齐方式下“导师”不同时间段上词向量和最后一个时间段词向量的相似度

在发现词义演变现象的基础上，为了更好地比较几种方法的有效性，在此选择本文锚点词列表中的名词“动物”进行对比实验，已知锚点词均没有词义变化。如图12是“动物”在历时语料中的自身相似度变化曲线、与锚点词“手”的相似度变化曲线。图13是不同时间段上“动物”和前一个时间段上前100个近邻实词的重合个数的柱状图。

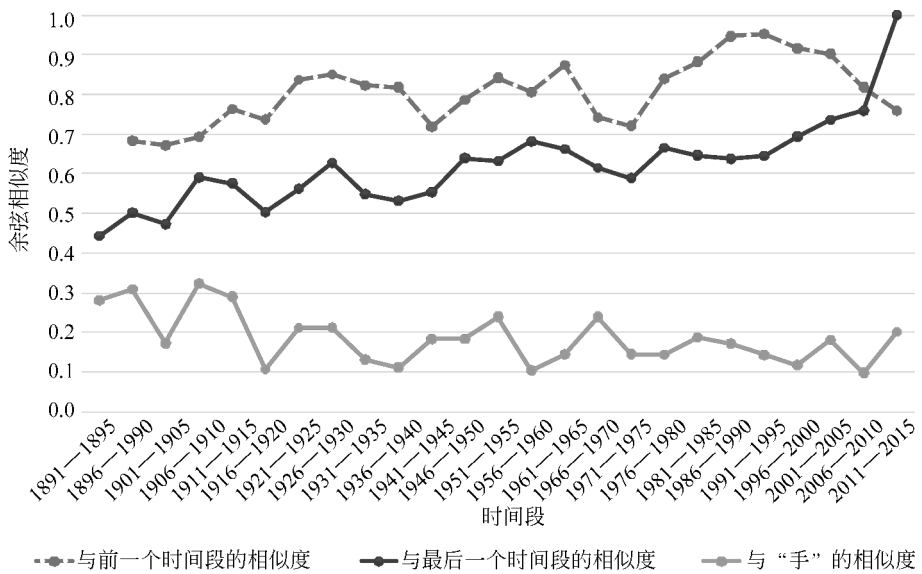


图12 锚点词“动物”历时相似度变化

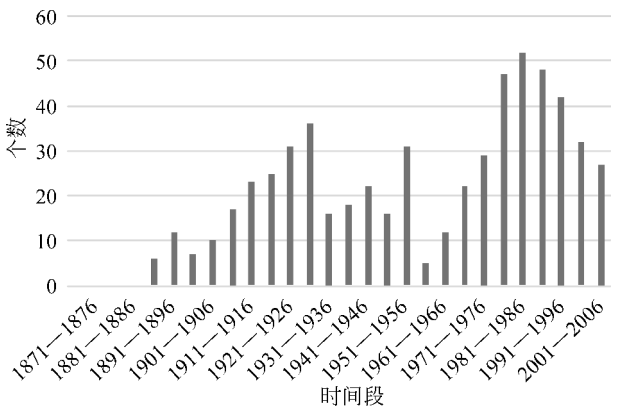


图 13　“动物”前 100 个近邻实词历时重合个数

结合多组实验可知，历时自相似度在发现词义演变现象上效果最好，有明显词义演变的词的自相似度曲线波动大，如“导师”“透明”“落马”等，而无明显词义演变的词的自相似度曲线波动小，如“动物”。其次是与锚点词的相似度，虽然在发生词义演变的时间节点上会出现明显的曲线波动，但由于语料规模有限，语料内容会对锚点词的词义产生一定的偏向性影响，在没有词义演变的词、时间节点上也会出现轻微曲线波动。

而近邻实词的历时重合个数，可以辅助明确词义演变前后的词义差异，并根据近邻实词列表来确定演变前后的明确指向。但仅依靠近邻实词的重合个数，不能有效发现词义演变的词语。在图 13 中，“动物”一词是已知没有词义演变的词语，而其近邻实词重合个数却出现明显波动，具体分析出现波动时间节点的近邻实词列表，如表 11 所示。

**表 11　“动物”不同时间段上的近邻实词及相似度**

| 1956—1960 | | 1961—1965 | |
|---|---|---|---|
| 近邻词 | 相似度 | 近邻词 | 相似度 |
| 鸟类 | 0.806 | 两栖类 | 0.773 |
| 生物 | 0.775 | 生物 | 0.772 |
| 动植物 | 0.765 | 动植物 | 0.766 |
| 松鼠 | 0.751 | 大熊猫 | 0.761 |
| 蠕虫 | 0.750 | 低等动物 | 0.753 |
| 小狗 | 0.747 | 植物 | 0.750 |
| 狗 | 0.740 | 鸟类 | 0.745 |
| 单细胞 | 0.735 | 昆虫 | 0.741 |
| 水貂 | 0.731 | 猿猴 | 0.740 |
| 雄性 | 0.725 | 哺乳类 | 0.730 |

由表 11 可以发现，“动物”一词在 1956 年到 1965 年之间词义是没有明显变化的，都是指生物中的一个种类。但由于动物中包含生物种类繁多，加上语料自身的局限性，导致其出现图 13 中的情况，在某些年份近邻实词重合个数低。

## 4　总结与展望

本文在收集的长达 144 年的历时报刊语料上提出了“锚点词”的概念，并通过多组实验表明，统计分析方法和基于分布式表示的方法都在一定程度上能够发现词义变化的时间、变化的趋势，并能明确变化前后的词语含义。统计分析的方法在新词的发现和旧词的消失上表现更直接，而分布式表示的方法在明确词语含义的变迁上有更出色的表现。在历时词义的区分效果上，SGNS 模型递增训练的方法效果优于其他两种方法。在发现词义演变现象上，历时自相似度的计算方法效果最优，其次是与“锚点词”相似度的计算方法，而近邻实词的历时重合个数仅能辅助明确词义，无法有效识别词义演变现象的发生。

在未来的工作中，我们将探索更为严格的形式化的中文历时词向量模型，加入知识库信息进行词义演变的研究，提高词语自动获取的效率，并对检测出的词义演变进行明确的类型辨别和归类。

## 参考文献

[1]　吴福祥. 汉语语义演变研究的回顾与前瞻[J]. 古汉语研究，2015(04)：2-13，95.

[2]　刘永厚. 汉语社会称谓语的语义演变[M]. 北京：知识产权出版社，2017.

[3]　李邵唐. 古今词义演变举隅[M]. 北京：语文出版社，2017.

[4]　王惠. 词义·词长·词频——《现代汉语词典》(第 5 版)多义词计量分析[J]. 中国语文，2009，(02)：120-130，191.

[5]　贾佳. 《儒林外史》词汇在现代汉语中的变化考察[D]. 保定：河北大学硕士学位论文，2010.

[6]　金观涛，刘青峰. 观念史研究：中国现代重要政治术语的形成[M]. 北京：法律出版社，2009

[7]　饶高琦，李宇明. 基于词汇聚类方法的现代汉语分期与分期体系构建[J]. 中文信息学报，2017，31(06)：18-24.

[8]　Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compo-

sitionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, ACM, 2013: 3111-3119.

[9] Turney P D, Pantel P. From frequency to meaning: Vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37: 141-188.

[10] Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, 1: 238-247.

[11] Kulkarni V, Al-Rfou R, Perozzi B, et al. Statistically significant detection of linguistic change[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 625-635.

[12] Zhang Y, Jatowt A, Bhowmick S, et al. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, 1: 645-655.

[13] Zhang Y, Jatowt A, Bhowmick S S, et al. The past is not a foreign country: Detecting semantically similar terms across time[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2793-2807.

[14] Calude A S, Pagel M. How do we use language? Shared patterns in the frequency of word use across 17 world languages[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2011, 366(1567): 1101-1107.

[15] Hamilton W L, Leskovec J, Jurafsky D. Diachronic word embeddings reveal statistical laws of semantic change[J]. arXiv preprint arXiv:1605.09096, 2016.

孙琦鑫(1993—),博士研究生,主要研究领域为自然语言处理。
E-mail: chasqx1993@163.com

荀恩东(1967—),博士,教授,主要研究领域为计算语言学、语言资源建设、计算机辅助教学。
E-mail: edxun@126.com

饶高琦(1987—),通信作者,博士,助理研究员,主要研究领域为计算语言学、语言政策与规划、数字人文。
E-mail: raogaoqi@blcu.edu.cn