南京大学

本 科 毕 业 论 文

院　　系　计算机科学与技术系

专　　业　计算机科学与技术

题　　目　基于历时词嵌入模型的语义演变研究

年　　级　2019　学 号 191220154

学生姓名　张涵之

指导教师　黄书剑　职称　副教授

提交日期　2023 年 5 月 26 日

南京大学本科毕业论文（设计）
诚信承诺书

本人郑重承诺：所呈交的毕业论文（设计）（题目：《基于历时词嵌入模型的语义演变研究》）是在指导教师的指导下严格按照学校和院系有关规定由本人独立完成的。毕业论文（设计）中引用他人观点及参考资源的内容均已标注引用，如出现侵犯他人知识产权的行为，由本人承担相应法律责任。本人承诺不存在抄袭、伪造、篡改、代写、买卖毕业论文（设计）等违纪行为。

作者签名：张洒洒

学号：191220154

日期：2023年5月17日

南京大学本科生毕业论文（设计、作品）中文摘要

题目：基于历时词嵌入模型的语义演变研究

院系：计算机科学与技术系

专业：计算机科学与技术

本科生姓名：张涵之

指导教师（姓名、职称）：黄书剑 副教授

摘要：词汇语义演变是语言整体进化的一部分，在传统语言学研究中已有近百年的历史，由于语言的进化往往折射出社会和大众心理的变化，在社会学领域也引起了广泛的重视。自从 Word2Vec 为代表的词嵌入模型问世以来，人们得以从大规模历时语料中挖掘词义的演变。由于词向量能反映词的含义，通过对比词语在不同时间的嵌入向量，就可以观察和评估词义的变化。本文在《人民日报》语料库上使用历时词嵌入模型，寻找 1946-2022 年间语义发生改变的中文词。在定性分析方面，本文首先以余弦相似度为标准筛选出含义可能有较大变化的词汇，然后通过单词邻域来判断这些词的语义变化，并且绘制了单词向量的演变轨迹，最后将语义变化大致归类到词性、词义、领域和搭配、感情色彩变化几个集合。在定量分析中，本文对部分单词语义变化的速度进行分析，确定其含义变化最大的关键时期，得出的结果与轨迹图一致。此外，本文分别从语言内部变化和语言对外部世界变化的反映两个角度设计了角色和概念的跨时间类比任务，证实了历时词表征可以捕捉包括科技发展、生活方式变迁、传染病大流行、武装冲突等各类现实事件。这些从历时词向量中挖掘的知识可以用于研究人类社会的发展。

关键词：历时词嵌入；语义演变

南京大学本科生毕业论文（设计、作品）英文摘要

THESIS: Analysing Lexical Semantic Change Using Temporal Word Embeddings

DEPARTMENT: Department of Computer Science and Technology

SPECIALIZATION: Computer Science and Technology

UNDERGRADUATE: ZHANG Hanzhi

MENTOR: Professor HUANG Shujian

ABSTRACT: Lexical semantic shift is a part of the general language evolution, which has been a subject of linguistic research for nearly a century, and has attracted a great deal of attention in the field of sociology, as the evolution of language often reflects changes in society and mass psychology. Since the invention of word embedding models such as Word2Vec, it has been possible to explore the evolution of word meanings from large-scale diachronic corpora. Since word embeddings can capture the meaning of words, changes in word sense can be detected and evaluated by comparing the embedding vectors of words at different times. This paper applies a temporal word embedding model to the *People's Daily* corpus, to find out Chinese words that have changed semantically between 1946 and 2022. In the qualitative analysis, words with potentially significant changes in meaning are first filtered out according to cosine similarity. We then identify semantic changes through a word’s neighbourhood, plot the trace of word vectors, and group semantic changes into clusters – part of speech, word sense, domain and collocation, etc. In the quantitative analysis, the speed of semantic change is analysed to identify key periods when their meaning changes the most, which is consistent with the results presented in the trace map. In addition, changes within the language and changes in the external world reflected on the language are discussed separately, as a temporal analogy task of roles and concepts is applied to demonstrate that temporal word representations can capture real-life events, such as technological developments, changes in people's lifestyles, infectious diseases and epidemics, armed conflicts, through which we may study the development of human society.

KEYWORDS: Temporal Word Embeddings; Semantic Shift

目　录

第一章 导论

1.1 研究背景及意义

在传统语言学和社会学领域，对语义演变的研究起源较早。词汇语义演变指的是词汇随着时间推移发生的含义结构变化，包括词义的扩大、缩小、转移、转类、虚化等。语义演变是人类语言进化的副产品之一。

传统语言学对语义演变的研究通常为定性分析，即根据某词在不同历史时期的使用举出例句，并针对观察到的词义变化进行描述、分类、归纳成因等几种工作。随着语料库语言学的兴起，开始出现使用一定统计工具的定量分析，如对不同时期同一词语不同语义的使用频率进行统计。然而，对于语料库中词语在每个句子里所表示的含义，需要语言学家预先规定分类标准，逐条进行人工标注，这就大大限制了数据的规模，且准确性严重受制于标注者的经验和判断。随着历时语料库的建设和计量语义学的发展，研究人员试图以大规模数据集来推动对历时语义演变的研究。词嵌入模型的出现使人们可以从共现数据中提取词向量，词向量可用于表示词义，并且可以比较和衡量不同词汇之间的相似度。

词语搭配模式的变化反映了词义的变化[1]。通过词嵌入和相关模型研究词义的演变不但可以验证传统语言学观测到的现象，而且可以提出新的发现。近期的许多工作除了聚焦于特定词汇的演变，还从大规模数据中总结出变化规律，如频繁使用的词语变化较慢，多义词的含义变化更快，同一个单词的不同含义在演变过程中可以相互竞争与合作，等等。这些发现与前人基于定性分析和小规模语料库定量分析人工推出的语义演变规律相互印证，互为补充[2-4]。

词汇语义演变与社会文化的变迁息息相关。典型的例子包括词语核心意义的变化（比如英文中的 gay 一词在 20 世纪从“无忧无虑”转变为“同性恋”）和文化关联的微妙变化（如伊拉克或叙利亚在武装冲突开始后逐渐与“战争”的概念联系在一起）[5]。该方向在跨领域理论研究（如社会语言学、计量语言学、数字人文等）有很大潜力，对基于文本挖掘的事件发现和预测也有所贡献。

1.2 国内外研究现状

1.2.1 词汇语义演变

语义演变（lexical semantic shifts 或 semantic change）常用的定义在 1933 年由 Boomfield 提出，称为一种改变词语含义而非语法功能的创新[6]。有关语义演变的大部分理论工作都致力于记录和分类各种类型的语义演变，如 Boomfield 将语义演变分为九个类别，包括词义的扩大（例如英语中的 dog 源自中古英语单词 doge，前者泛指狗，后者则限制于某一特定品种的狗），词义的缩小（如英语中的 meat 表示可食用的肉，其词源系古英语中的 mete，泛指一切食物），以及词汇感情色彩的变化（如英语中的 knight 表示身份较崇高的“骑士”，而古英语中的同源词 cniht 则表示地位较低微的“男仆”）等[6]。另一种受到普遍认可的分类方法将语义演变划分为语言学层面的转移（单词核心含义缓慢而有规律的变化）和社会文化层面的转移（单词在特定文化背景下引发联想的变化）[5]。

1.2.2 历时词嵌入模型

以 Word2Vec 为代表的词嵌入模型（有时也称为词向量）[7]基于 Firth 的分布假设，即一个词的含义可以由该词的上下文来定义。词向量属于静态模型，认为每个单词的含义在时间上是固定的，没有考虑到单词的语义变化。自从词向量问世以来，就被广泛应用于各种自然语言处理（NLP）场景，如词性标注、信息检索、问答、情感分析等。近年来，将词向量与时间动态结合的方法在计量语言学和社会学领域引发了广泛的兴趣。在此之前，学界有根据词频和共现频率研究语义演变的例子[8]，对事件检测的研究中也有用到语义向量的分布式模型[9]。通过从共现关系中提取词汇表征（密集的嵌入词向量），分布式模型在检测语义转变任务上的表现优于基于频率的方法，这一点 Kulkarni 等人进行了证明[10]。

基于词向量的语义演变研究最早着眼于较长时间跨度上的变化，如 Sagi 等人将英语按中世纪早期、中世纪晚期和现代英语进行划分，研究英文单词在较大尺度上的含义变化[11]。时间跨度更小的语料切片可以更为敏锐地捕捉到与社会文化有关的语义演变，如 Kim 等人（2014）、Liao 和 Cheng（2016）的研究都是按年分割的[12-13]。除了把数据按时间划分成独立的块，Rosenfeld 与 Erk 于 2018 年首次把时间作为连续变量引入，从而进一步刻画了语义演变的连贯性[14]。

Kim 等最早基于预测的词嵌入模型研究[12]用到增量更新和负采样的连续跳格（Skip-gram with Negative Sampling，简称 SGNS）。Hamilton 于 2016 年证实了 SGNS 相对基于正点互信息（PPMI）的显式分布模型的优越性[4]，Tsakalidis 则比较了基于计数（Temporal Random Indexing）和基于预测（Word2Vec）的模型性能[15]。从那时起，相关领域内的大多数研究都使用密集的词向量表征，其中最常用的两种分别是 SVD 因子 PPMI 矩阵，以及基于预测的浅层神经模型。

获得词汇语义在不同时期的向量表征之后，需要对其进行比较。由于神经网络训练过程的随机性，不能直接计算不同模型中同一个词的嵌入之间的余弦来获得语义的相似性。对此，Kulkarni 提出对历时词向量进行分析之前应该先用线性变换对模型进行调整，使其处于同一个向量空间[10]。Hamilton 使用二阶嵌入和正交 Procrustes 变换来调整异时空模型，这种做法基于两个假设，一是大多数单词的含义随着时间的推移近似不变，二是如果嵌入维度足够大（则优化问题不那么非凸），那么这些含义近似不变的词在不同时间片上的嵌入向量相差一个全局旋转，通过计算这个旋转可以实现向量空间的强制对齐[4]。这一实现受到 Bamler 与 Mandt 的批评，因为旋转的虚像与实际存在的语义变化很难严格区分[16]。

在此基础上，Zhang 等人引入“本地锚”（local anchor），即用线性投影对小的近邻集进行映射，将被查询的单词映射到对应的时间片[17]。类似地，Di Carlo 等使用被冻结的局部向量作为“指南针”（compass），以确保在不同时间片上分别训练的词向量天然地处于同一个共享的坐标系内[18]。Rosenfeld 与 Erk 则将词向量设计为应用于连续时间变量的线性变换，实现单词 w 在时间 t 的嵌入[14]。

然而，上述模型具有一个共同的缺陷，即在每个时间段将单词表示为单个向量，无法区分一词多义的情况。单一词向量可以视为同一词语不同语义根据使用频率的加权平均，虽然能反映一部分语义变化，但不够精确。最近，大规模预训练语言模型（如 ELMo[19]和 BERT[20]）的蓬勃发展在 NLP 领域引起了广泛的关注。这些模型可以捕捉更细粒度的词义在不同语境中的变化，即不同上下文的单词可以产生不同的表征。基于这类模型的语义演变研究不再对具体的单词，而是对含义（senses）计算嵌入向量，如 Giulianelli 等 2020 年的工作使用 K-Means 将词语在具体上下文中的用法划分到不同的 sense，并选择最接近质心的五个向量来识别聚类中最具代表性的用法[21]。这一作法的进步性在于能够正确区分多义

词的含义，包括字面义和比喻、引申义等，局限性包括聚类数量 k 的选择比较随机，难以解释原因，以及对每个聚类所包含的 sense 仍然需要进行额外的人工归纳和总结。对此，Hu 等结合了权威词典（如牛津英语词典）中的例句，首先计算出目标词的含义表示，再将不同语境下的具体用法划分到与之余弦相似度最高的 sense[22]。与单一词表征相比，基于上下文的嵌入模型可以更精确地统计词语含义在不同时期的使用频率，从而衡量和比较词义变化的幅度和速度。

1.2.3 评估方法

理想情况下，评估历时词嵌入模型需要一个人工标注的语义转移词表作为参照系（如根据变化的程度进行定量分析和排序）。然而，这样的黄金标准数据很难批量获得，目前大部分相关工作只包含少量手动挑选的例子。

另一种可能的评估方法是跨时间类比，即在不同时间片上寻找特定概念的对等词汇（例如，针对“美国总统”这一概念实体，2015 年的“奥巴马”对应 2017 年的“特朗普”；而对于“音乐播放器”，2000 年的 iPod 相当于 20 世纪 80 年代的随身听）。在英语中已经有一些这样的历时等价词汇数据集[23-24]，但在其他语言上还没有太多相关资源。还有一种评估策略是使用检测到的历时语义变化来追踪或预测武装冲突等现实世界事件[25]。所有这些评估方法仍然依赖于人工标注的语义演变数据，且历史语言学家也很难准确地描述语义的变化程度。

1.2.4 中文领域现状

Kutuzov 等在 2018 年的一篇综述中提出基于词嵌入模型的语义演变研究所面临的几点挑战[5]，包括在英语之外的其他语料库上实践不足。尽管 Hamilton 使用了 4 种语言（英语、德语、法语和汉语）的 6 个语料库，但中文数据相对其他语种（200 年）跨度较小（1950-1999），且分词工具效果不佳（从公开模型中看到“红楼梦”切分为“红”、“楼”、“梦”三个单字，“毛主席”分为“毛”和“主席”，“吃饱”分为“吃”和“饱”等），粒度过细的分词导致向量空间中位置接近的往往并非近义词，而是常见词语和命名实体中的字符组合[4]。

国内大多数对中文词义变化的研究仍使用传统语言学方法，近年也有词向量模型的应用[26-27]，但由于这方面工作起步较晚，且缺乏大规模中文历时语料，多数研究只涉及少量例子的定量讨论和分析，相比英文上的研究还较为初步。

1.3 本文主要工作

1.3.1 研究内容

本文使用基于预测的词汇表征 Word2Vec，参考 Di Carlo 的 TWEC (*Temporal Word Embeddings with a Compass*)模型[18]，以 1946-2022 年的《人民日报》新闻数据为历时中文语料，分别对按 5 年和 1 年划分的两种时间片训练词向量。得到词义嵌入表征之后，利用余弦距离筛选出可能存在明显语义变化的词，首先依据被查询词的近邻对词义变化定性分析，并对演变路径进行了可视化。本文随后尝试定量分析了部分单词含义的变化速度，其中包括与自身词义对比和与相邻词语的对比，并将二者进行交叉验证。最后，本文仿照英文词汇跨时间类比任务[23-24]设计了有限的历时等价中文词表，并对词向量进行了初步的测试。

1.3.2 研究目的

由于现有的基于历时词嵌入模型的中文词汇语义演变研究较少，本文尝试进行一些启发式的探索。本文尝试提出一些可能性，如类似模型不但可以被用于验证和量化语言学家已经观察到的现象，而且可以创新性地发现人类尚未留意过的变化，然后再交由语言学家进行分析和讨论。这方面的具体应用还有待发掘和检验。本文还初步探索了几种衡量词汇语义演变方向和速度的方法在中文历时词向量上的应用，提出了将跨时间类比任务迁移到中文的可能性，以此来呼吁中文历时词表的进一步建设和完善。此外，本文希望对词义变化进行分类和归纳成因并统计各种类型所占的比例，但由于语言学方面的知识有限，在此只是举出了一些例子，并没有进行较完备的研究，期待未来在这方面有更多探讨。

第二章 实验

2.1 实验数据集

过去在英文语料上的相关工作主要使用的是带有时间戳的小说和杂志集（如Google N-Gram 和 COHA），或者网络社区（如 Facebook 和 Twitter）的帖子。Yao 等人使用了《纽约时报》作为数据集，并提出新闻比起其他体裁的语料更容易保持语言风格的统一，而且更有利于从社会时事的角度研究语言的进化[24]。考虑到这两方面因素，本文也选择中文的新闻语料作为实验数据。

2.1.1 数据来源

本文收集了 1946-2022 年的全部《人民日报》新闻文本，资料来源于人民日报官方的电子版图文数据库网站（http://paper.people.com.cn）。

2.1.2 时间片划分

通常情况下，如果时间片太大，则只能实现词汇语义随着时间变化的极粗粒度表示。然而，在细粒度的时间片上，语料库的大小有限，训练得到的词向量受随机噪声的影响较为明显，很难从微妙的语义变化中排除这种噪音干扰。本文分别按照 5 年和 1 年进行时间切片，在两种时间片上分别进行训练。

2.1.3 数据预处理

本文使用开源工具 jieba 进行分词，根据哈工大停用词表去除停用词，处理后的数据为 4.12GB。将全部语料合并到一个文档中，称为 compass。5 年的时间切片按 0-4 和 5-9 结尾的年份划分，除了首尾（1946-1949 和 2020-2022）。

2.2 模型选择

考虑到时间和资源有限，本文选用基于 Word2Vec 的 TWEC 模型[18]，其优势在于训练过程简化，能够直接将不同时间片上的词向量嵌入同一坐标系，而无需额外的对齐操作，效率较高。比起之前流行的在不同时间片上先分别训练再旋转强制对齐的两步式方法，这一模型较好地避免了近似计算的旋转角度和现实中微小语义偏移相互干扰的问题，在小规模数据集上也能保持稳定的性能。

2.2.1 原理介绍

TWEC 的作者使用了“指南针”（compass）这一形象的比喻为模型命名。总的来说，该模型和先前的两步式模型基于同一个假设，即大多数单词的含义不会随着时间的推移而改变[15]。在此基础上，作者进一步提出新假设，即对于一个含义随着时间发生了变化的单词来说，构成它的上下文的大多数单词含义也是近似不变的。这个假设允许在训练过程中启发式地将目标（target）嵌入视为静态的并将其冻结，上下文（context）嵌入则根据特定时间片上的共现频率动态变化。在这里，目标嵌入是与时间无关的，而得到的上下文嵌入就是历时词表征。读者可以想象一个有关地图的比喻[28]，取一个城镇在不同历史时期绘制的地图来研究该地的地理或行政区划演变，则很难保证所有的制图师都按相同的方位作图。要将这些地图叠在一起比较，则需要进行一定的旋转对齐。然而，如果给每个制图师指南针并规定方位，就可以实现自动对齐，本模型依据的就是这一思路。

对于一个使用 CBOW 的 Word2Vec 模型，上下文嵌入 $\vec{c}_j$ 被编码在神经网络的输入权重矩阵 $\mathbf{C}$ 中，而目标嵌入 $\vec{u}_k$ 被编码在输出权重矩阵 $\mathbf{U}$ 中。假设有历时语料库 D 分为 n 个时间片 D^{t_i}，其中 $1 \leq \mathrm{i} \leq \mathrm{n}$，训练分为两个阶段：首先，对整个历时语料库 D（即所有新闻语料合并得到的 compass 文件）应用原始的 CBOW 模型，构建两个与时间无关的矩阵 $\mathbf{C}$ 和 $\mathbf{U}$，分别表示全局的上下文嵌入和目标嵌入集合。其次，对于每个时间片 D^{t_i}，使用先前训练的目标嵌入矩阵 $\mathbf{U}$ 初始化神经网络的输出权重矩阵，再应用修改过的 CBOW 算法来构造历时的上下文嵌入矩阵 $\boldsymbol{C^{t_i}}$（第二阶段用到的 CBOW 只对输入矩阵 $\boldsymbol{C^{t_i}}$ 的上下文嵌入进行更新，而输出矩阵 $\mathbf{U}$ 的目标嵌入保持不变）。

对每个时间片 $\langle w_k, \gamma(w_k)\rangle \in D^t$ 的更新可视为下面的优化问题：

$$\max_{C^t} \log P(w_k|\gamma(w_k)) = \sigma(\vec{u}_k \cdot \vec{c}^t_{\gamma(w_k)}) \qquad (1)$$

其中 $\gamma(w_k) = \langle w_{j_1}, \dots, w_{j_M}\rangle$ 表示时间片 D^t 上 w_k 上下文的 M 个词（在这里窗口大小为 $\frac{M}{2}$），$\vec{u}_k \in \mathbf{U}$ 是 w_k 与时间无关的目标嵌入，而

$$\vec{c}^t_{\gamma(w_k)} = \frac{1}{M}(\vec{c}^t_{j_1} + \cdots + \vec{c}^t_{j_1})^T \qquad (2)$$

是上下文中单词 w_{j_m} 的历时上下文嵌入 $\vec{c}^t_{j_m}$ 的均值。

模型中用到的 softmax 函数 σ 根据负采样（Negative Sampling）计算。训练过程是符合直觉的：第二阶段的 CBOW 实际上是根据全局的目标嵌入 **U** 和局部的（时间片 D^t 的）上下文来预测单词 w_k。在训练中，由于我们假定大多数词的含义近似不变，那么将在该时段经常与 w_k 处在邻近上下文的 w_{j_m} 的历时上下文嵌入 $\vec{c}_{j_m}^t$ 不断向着 w_k 的全局目标嵌入 $\vec{u}_k$ 移动，最终得到的 $\boldsymbol{C^{t_i}}$ 就是在时间 t_i 的词义表征。矩阵 **U** 起到的就是所谓“指南针”的作用。

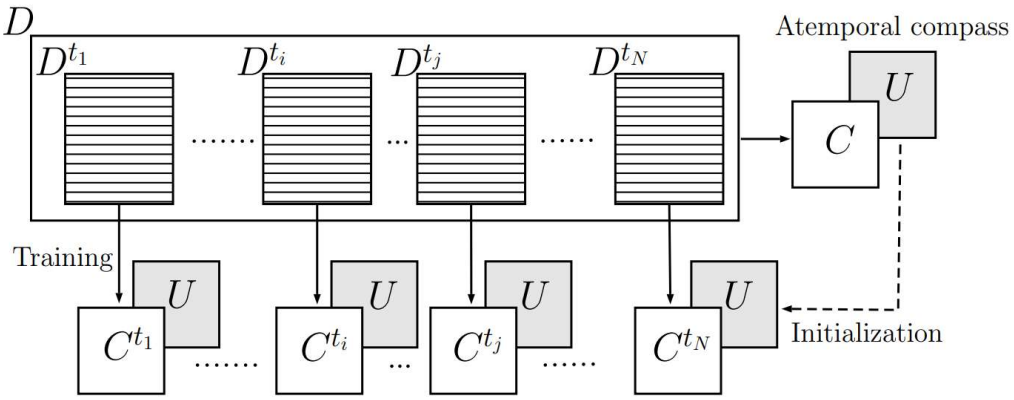

图 1 TWEC 模型示意图[1]

2.2.2 参数设置

本文参照 Di Carlo 等的工作[18]，设置嵌入向量维度为 30，窗口大小为 5，负采样的数量为 10。由于时间和能力有限，只作了有限的训练，没有对参数进一步调整，多数其他参数（如学习率）仍沿用 TWEC 原版模型中列举的缺省值。

[1] Di Carlo, V., Bianchi, F., & Palmonari, M. 2019. Training Temporal Word Embeddings with a Compass. *AAAI Conference on Artificial Intelligence*.

第三章 分析

3.1 寻找目标词

由于本文的定位是发现可能的词汇语义变化，获得历时词表征后，首先对在1946-2022年间向量变化较大的词进行了筛选。基本思路是对于5年尺度的时间切片分别统计词频，对达到一定词频（此处定为5年间出现次数大于100）的单词取交集，计算这些词在1946-1949和2020-2022一头一尾两个切片上向量的余弦相似度并从小到大排序，最后按照排序得到的词表依次筛选。

3.2 定性分析

3.2.1 单词邻域

在人工筛选的过程中，可通过观察单词在向量空间中的“邻域”（同样依据余弦距离定义）理解词义的变化，如英文中“特朗普”（Trump）一词的邻居从“房地产”（real estate）先后变为“电视”（television）和“共和党”（republican），反映了人物从经商、进入演艺圈到从政的角色变化。类似地，“苹果”一词在英文的语义空间里逐渐从草莓、芒果移动到iPhone、iPad的附近[24]。

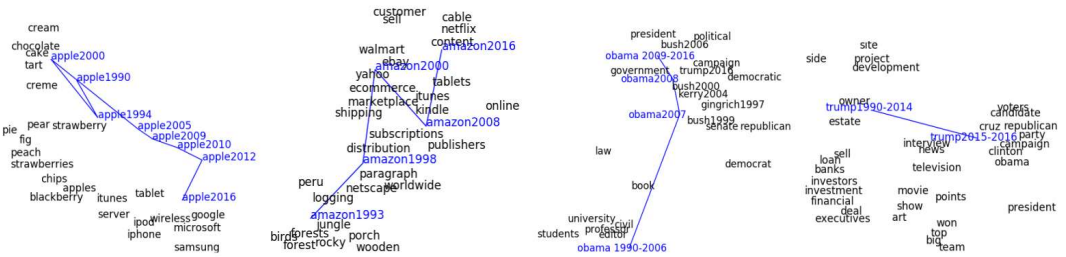

图 2 通过邻域追踪英语中品牌名和人名的语义变化[1]

下面展示了一些词语的最近邻变化，并利用t-SNE对词向量进行降维，从而将词义的迁移可视化为轨迹，以求提供变化幅度和速度的直观印象。

[1] Yao, Z. et al. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

表1 “提升”在5年切片上的最近邻（topn=5）

年份	最近邻
1946-1949	升任，见习，下士，领工，斌
1950-1954	架工，机长，机工，领班，升为
1955-1959	回柱，提绞，工等，机械动力，截煤机
1960-1964	充填，副井，主井，维修工，穿孔机
1965-1969	五级，电工，顶替，王兆龙，科
1970-1974	凿岩，回采，副井，吊装，绞车
1975-1979	工程师，晋升为，机电，技师，会计师
1980-1984	晋升为，提为，晋升，破格提拔，双肩挑
1985-1989	不胜任，现职，不称职，降免，晋升
1990-1994	晋升，提拔，评聘，不称职，师级
1995-1999	提高，大大提高，全面提高，明显提高，努力提高
2000-2004	提高，进一步提高，增强，明显提高，大大提高
2005-2009	提高，增强，进一步提高，明显提高，全面提高
2010-2014	提高，增强，进一步提高，明显提高，努力提高
2015-2019	提高，进一步提高，增强，明显提高，努力提高
2020-2022	提高，增强，进一步提高，明显提高，努力提高

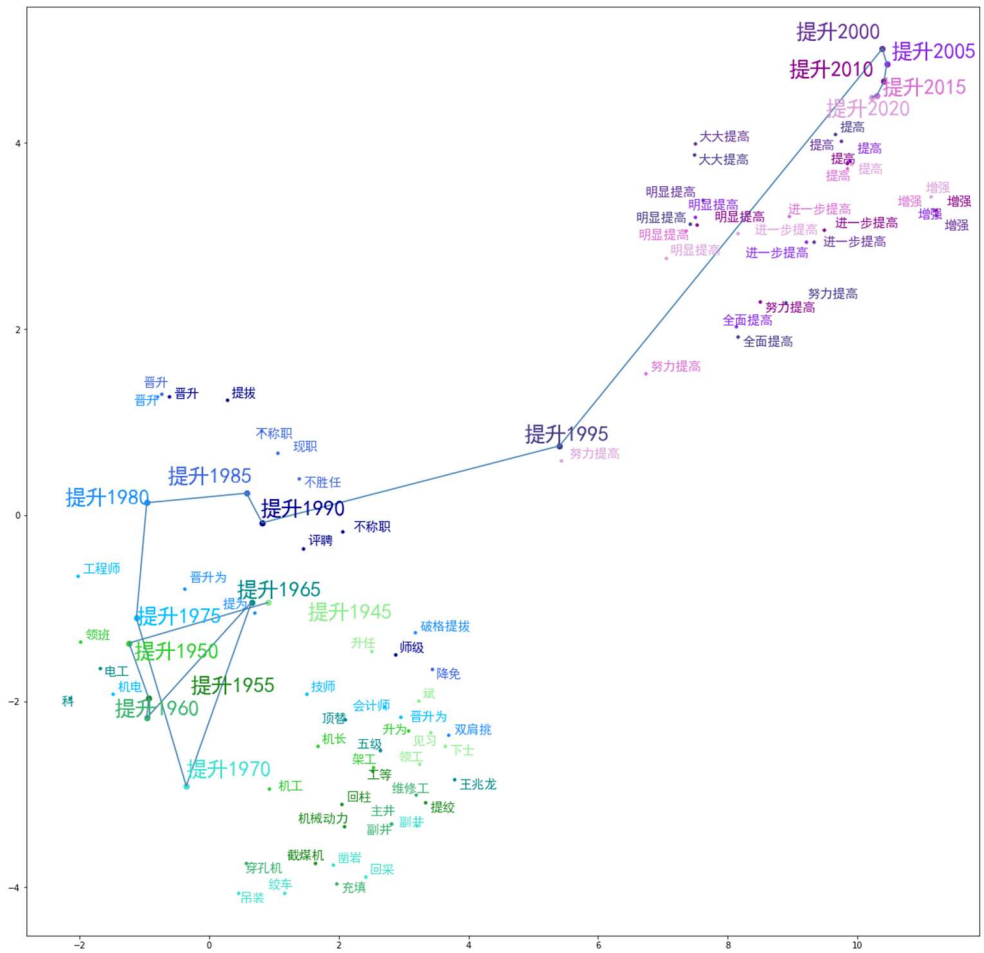

图 3 “提升”在 5 年切片上的迁移轨迹

表2 “释放”在5年切片上的最近邻（topn=5）

年份	最近邻
1946-1949	自首，开释，驱逐出境，被捕，赔偿损失
1950-1954	日俘，遣送回国，获释，被释，开释
1955-1959	获释，引渡，驱逐出境，赦免，受审
1960-1964	获释，赔偿损失，惩办，处决，撤消
1965-1969	无理，判罪，勒令，拘留，横蛮无理
1970-1974	关押，拘留，拘禁，囚禁，被俘
1975-1979	获释，逮捕，扣押，被捕，处决
1980-1984	获释，处死，扣押，逮捕，处决
1985-1989	获释，在押，驱逐出境，处死，扣押
1990-1994	获释，处死，关押，被害，驱逐
1995-1999	获释，关押，人质，处死，战俘
2000-2004	关押，作出反应，获释，在押，还击
2005-2009	地使，释放出来，施加，掌控，激起
2010-2014	释放出来，激活，激发，地使，促使
2015-2019	释放出来，激发，迸发，激活，涌流
2020-2022	释放出来，显现，激活，激发，迸发

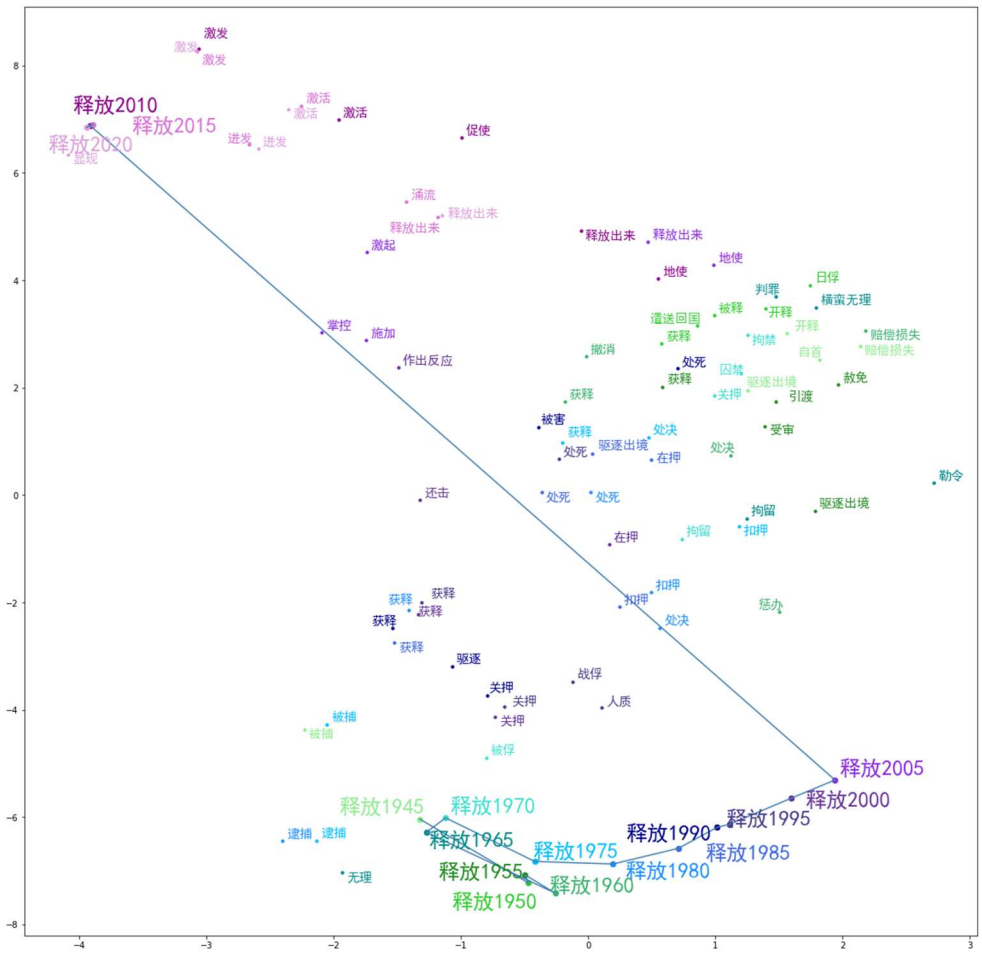

图 4 “释放”在 5 年切片上的迁移轨迹

表3 “清洁”在5年切片上的最近邻（topn=5）

年份	最近邻
1946-1949	喷雾，洗澡，清洁卫生，室内外，防蝇
1950-1954	清洁卫生，保持清洁，室外，清扫，打扫
1955-1959	清洁卫生，室内外，打扫，个人卫生，清扫'
1960-1964	清洁卫生，食具，防蝇，漱口，防尘
1965-1969	刷洗，澡，痰盂，打扫，清扫
1970-1974	清扫，打扫，整洁，浴室，食堂
1975-1979	清洁卫生，清扫，打扫，浴室，街巷
1980-1984	整洁，清洁卫生，车容，室内外，洁净
1985-1989	采光，清洁卫生，通风，杂用，无烟
1990-1994	清洁卫生，保暖，洁净，通风，淋浴
1995-1999	清洁卫生，照明，净水，垃圾处理，室内环境
2000-2004	节约能源，节能，节能型，环境友好，再生资源
2005-2009	能源开发，洁净煤，节地，节能型，低碳
2010-2014	能源开发，节能，低耗，高碳，低碳型
2015-2019	能源开发，再生能源，低碳，电能，可再生
2020-2022	能源开发，生物质能，绿色，低碳，能源

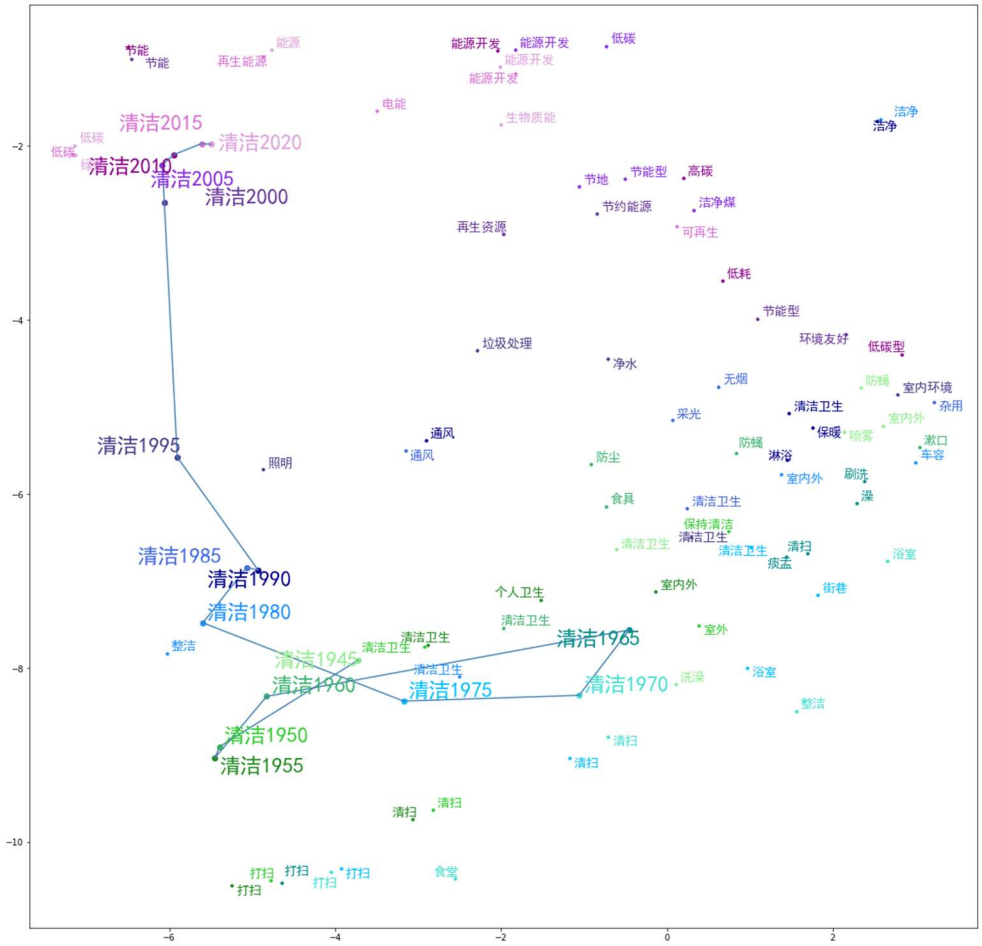

图 5 “清洁”在 5 年切片上的迁移轨迹

表4 “功夫”在5年切片上的最近邻（topn=5）

年份	最近邻
1946-1949	工夫，十来天，个把月，三回，十几天
1950-1954	工夫，干完，七八天，三两天，二十多天
1955-1959	工夫，琢磨，要费，干完，七八天
1960-1964	苦功，苦功夫，工夫，练得，下功夫
1965-1969	练得，精，硬功夫，武艺，省劲
1970-1974	气力，没用，脑子，力气，劲
1975-1979	力气，气力，工夫，脑子，管用
1980-1984	苦功，真功夫，力气，唱功，气力
1985-1989	气力，几手，力气，硬功夫，劲
1990-1994	苦功夫，硬功夫，劲，劲儿，脑子
1995-1999	苦功夫，真功夫，力气，唱功，气力
2000-2004	真功夫，力气，苦功夫，唱功，一招一式
2005-2009	真功夫，一招一式，做功，唱念，绝技
2010-2014	真功夫，做功，苦功夫，一招一式，绝招
2015-2019	苦功夫，硬功夫，真功夫，力气，耐性
2020-2022	苦功夫，真功夫，动脑筋，气力，绣花

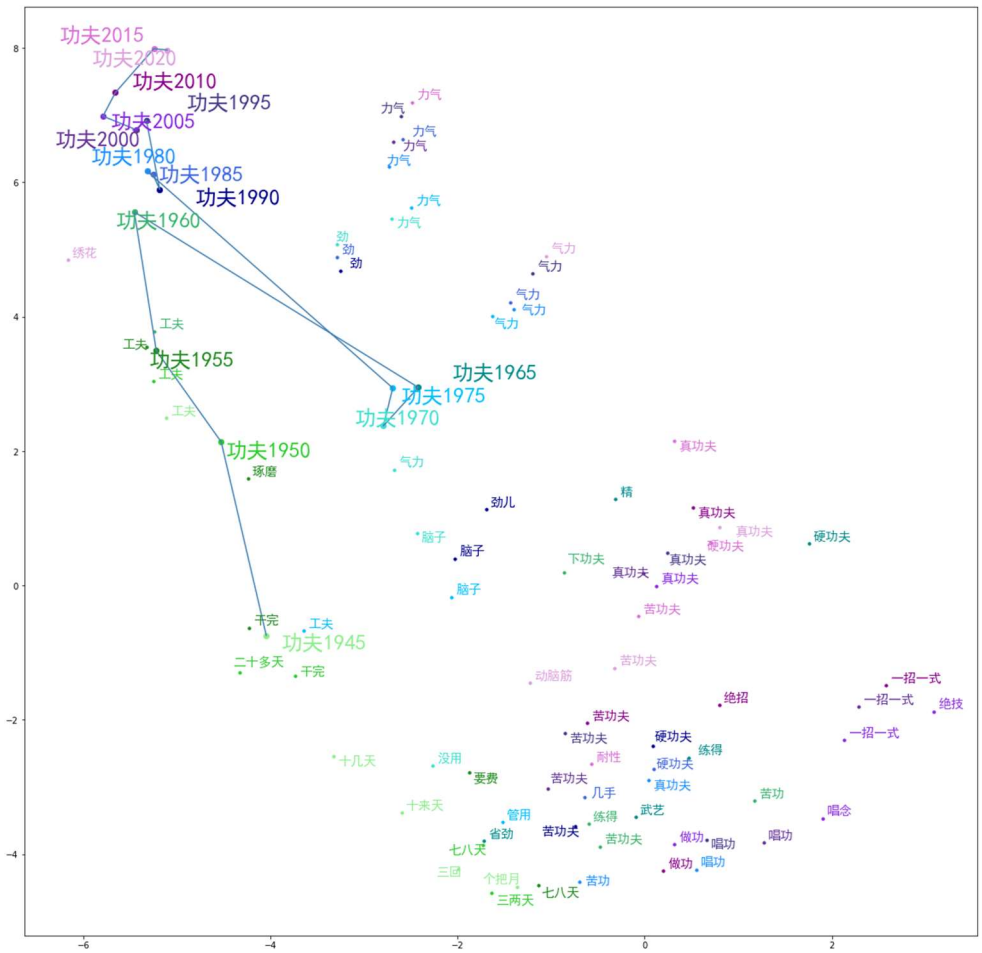

图 6 “功夫”在 5 年切片上的迁移轨迹

表5 “导师”在5年切片上的最近邻（topn=5）

年份	最近邻
1946-1949	伟大领袖，斯大林，继承者，伟大，敬爱
1950-1954	伟大领袖，敬爱，解放者，台尔曼，英明领袖
1955-1959	革命家，先行者，孙逸仙，赫尔岑，革命领袖
1960-1964	革命领袖，巴黎公社，革命家，奠基人，先驱者
1965-1969	领袖，革命领袖，革命家，思想家，马克思列宁主义者
1970-1974	列宁，革命领袖，新民主主义论，巴黎公社，马克思
1975-1979	丰功伟绩，追思，敬爱，朝鲜劳动党，虽死犹荣
1980-1984	教育家，李四光，理论家，平心，思想家
1985-1989	复旦大学，哲学系，芝加哥大学，哈佛大学，陈省身
1990-1994	硕士生，名教授，博士生，数学系，复旦大学
1995-1999	博士生，副教授，吉林大学，破格晋升，物理系
2000-2004	博士生，博导，副教授，哲学系，生物系
2005-2009	博士生，博导，讲师，硕士生，中科大
2010-2014	博士生，博导，硕士生，助教，青年教师
2015-2019	助教，博士生，指导老师，青年教师，讲师
2020-2022	青年教师，博士生，硕士生，指导老师，博士

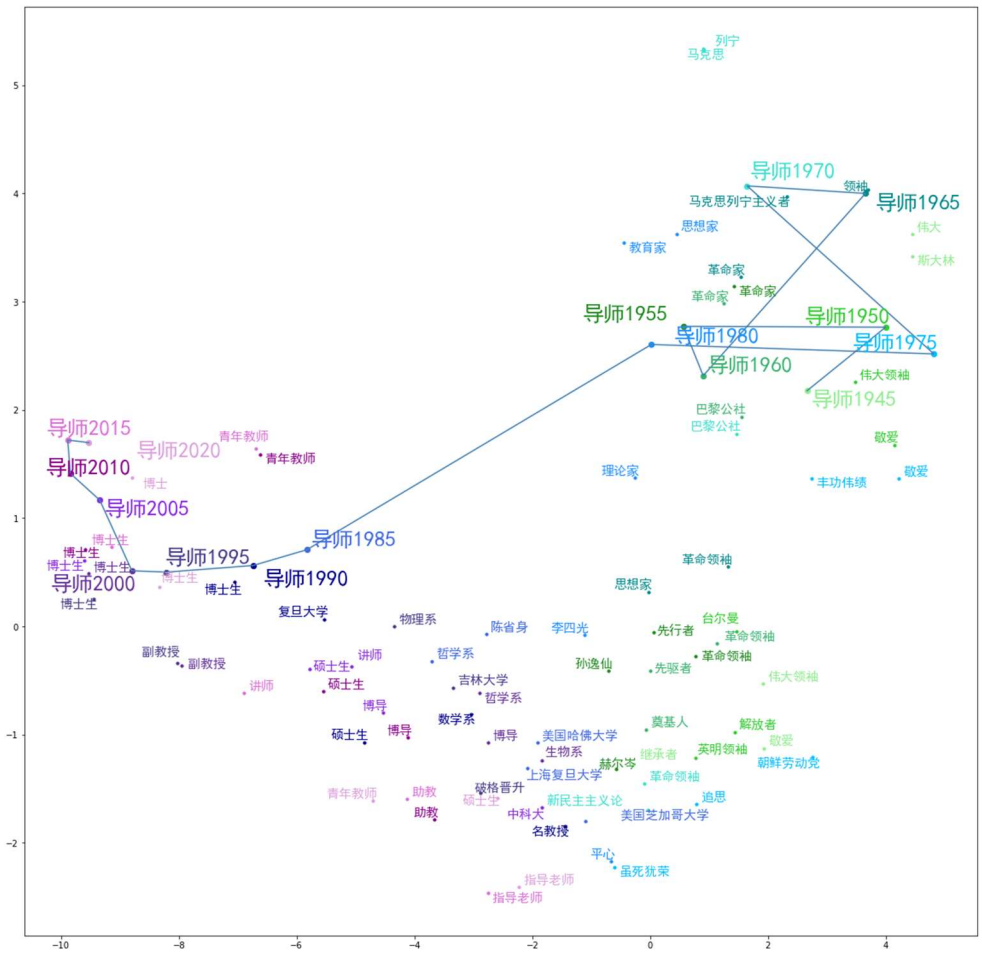

图 7 “导师”在 5 年切片上的迁移轨迹

图中对不同年份的被查询词和近邻词着渐变色（绿——蓝——紫），读者可以较直观地对概念的集群进行区分。本文选择了几个有趣的例子，通过查看和对比示例词语在不同时期的最近邻，很容易看出如下列举的词义演变现象：

(1) “提升”最早的常用含义为职务级别的“提拔”或“晋升”，这一时期的邻居有近义词“升任”、“升为”和表示职位的搭配词“领工”、“领班”等。50-70年代，“提升”有时也表示用机械“向高处运送（矿物、材料等）”，近义词有“提绞”、“吊装”，搭配词有“绞车”等，都是施工用语。90年代中期，“提升”变为抽象的“提高”、“增强”的近义词，常以“大大提高”、“明显提高”、“进一步提高”等搭配出现，此后含义便逐渐稳定下来（见表1、图3）。

(2)“释放”最早为刑事概念“恢复被拘押者或服刑者的人身自由”，邻居有近义词“开释”、“获释”和反义词“被捕”、“拘留”等。大约在2000-2005年间转为表示“把所含物质或能量放出来”，与通常作用于抽象事物的“激发”、“激活”、“激起”等动词位于相邻的语义空间（见表2、图4）。

(3)“清洁”在上世纪作形容词时与“卫生”、“整洁”、“洁净”是近义词，表示器具或环境干净整洁，作动词时则与“刷洗”、“打扫”、“清扫”是近义词，常与“食具”、“浴室”、“街巷”等被清洗和打扫的对象或场所搭配。步入21世纪后，“清洁”最常用的含义为“节能环保”，因此也向“环境友好”、“低碳”、“绿色”、“可再生”这些概念靠近（见表3、图5）。

(4)“功夫”在60年代前通“工夫”，表示做一件事情所需要耗费的时间，常与“十来天”、“个把月”等时间单位连用，后来逐渐扩展到花费精力、力气、努力，如“下功夫”、“苦功夫”，最后则指武术、歌舞等方面的本领和造诣，如“真功夫”、“硬功夫”，近邻也有“唱功”、“招式”等（见表4、图6）。

(5)“导师”早期比喻指引革命方向的领袖人物，近义词有“伟大领袖”、“革命家”、“先行者”等，邻域中出现的人名则包括孙逸仙、马克思和列宁这些在历史上影响较大的政治家。在1975-1985年代，“导师”逐渐转为表示高等院校和研究机构的教师，邻居有“博士生”、“博导”、“教授”等（见表5、图7）。

3.2.2 词义变化归类

本文尝试总结了一部分词汇的语义演变，并归类如下。主要考察了语法功能即词性变化、词义变化如比喻/引申义的产生，领域和搭配变化等。

首先是词性，同时也是词语在句子中所担任语法成分的变化。如“动态”旧义为名词，表示某领域最新近况，新义为形容词，与“静态”相对，指事物不断发展变化的一面。类似地，“封闭”由动词“查封”变为形容词“关闭的”（从而又可引申为“僵化保守”），“地道”由名词“地下通道”（如“地道战”）变为形容词“质量标准”（如“地道的特色小吃”），而“生气”则由动词“发怒”（如“他莫名其妙地生气了”）变为名词“活力”、“朝气”（如“生气勃勃”）。

表6 词性变化

词语	解释	示例近邻
动态	旧：n.（事情）变化发展的情况	概况, 时事, 概论, 情报资料, 书讯
	新：adj. 运动变化状态的	静态, 动态分析, 实时, 分层, 分类
封闭	旧：v. 查封	查封, 捣毁, 关闭, 拆除, 捣毁
	新：adj. 关闭的	旧有, 半封闭, 僵化, 孤立, 关起门来
地道	旧：n. 地下通道	工事, 壕沟, 山洞, 暗堡, 入口
	新：adj. 质量够标准	味儿, 味道, 纯正, 风味, 正宗
生气	旧：v. 不高兴，发怒	脾气, 怪, 莫名其妙, 难为情, 别扭
	新：n. 活力，朝气	痛快, 得意, 爽快, 神气, 快活

其次是词性不变前提下的核心词义变化。这一类含义变化有的受到时代发展的影响，如“大洋”仅在建国前作为民国时期货币“银洋”的俗称，建国以后就只表示海洋了。有的含义是在现代化和全球化的进程中随着外语中对应概念的演化一同引进的，如用“中心”表示机构和单位，是由英文中的“centre”一词化用而来（如 information centre 信息中心，shopping centre 购物中心）；“清洁”表示节能环保，也与英文中“clean”一词的引申对应（如 clean energy 清洁能源）。另外有些不太规范的用法（如“功夫”和“工夫”的混用）随着汉语的规范化逐渐被纠正。还有由字面义到比喻义、从具体事物到抽象概念的引申，如“推出”原表示“推”的动作，现表示产品上市和推广；“增添”原表示添置采办物品，现表示抽象事物的“焕发”；“缺口”原表示物体缺失一部分形成的孔洞，现多用于资金、资源短缺；“渠道”原指排水用的沟渠，现比喻做成事情所需的门路、途径；“行列”原表示行军或游行中的队列，现可表示各种抽象意义上的群体，如某校或某国“跻身”、“一跃”加入“一流大学”、“经济大国”的行列。

表7 词义变化

词语	解释	示例近邻
大洋	旧：n. 银元的俗称	洋, 银洋, 七千元, 冀钞, 净赚
	新：n. 海洋	港湾, 航行, 游弋, 科考, 深海
中心	旧：n. 事物的主要部分	重心, 重点, 当前, 围绕, 主要
	新：n. 机构	信息中心, 情报中心, 下设, 学院, 研究院
提升	旧：v. 提拔	升任, 晋升, 晋升为, 提为, 评聘
	中：v. 用卷扬机等向高处运送	提绞, 吊装, 绞车, 回柱, 机械动力
	新：v. 提高、增强	提高, 大大提高, 进一步, 全面, 增强
清洁	旧：adj. 干净卫生	清洁卫生, 整洁, 洁净, 刷洗, 通风
	新：adj. 节能环保	节能, 环境友好, 低耗, 低碳, 绿色
功夫	旧：n.（做事）耗费的时间	工夫, 十来天, 个把月, 干完, 要费
	中：n. 力气、努力	苦功, 气力, 力气, 脑子, 劲儿
	新：n. 本领、造诣	真功夫, 硬功夫, 招式, 绝技, 唱功
释放	旧：v. 恢复被拘押者的人身自由	开释, 获释, 赦免, 拘留, 关押
	新：v. 把所含物质或能量放出来	释放出来, 激活, 激发, 迸发, 显现
推出	旧：v. 向外推	推, 一推, 推开, 下推, 拖
	新：v. 向社会大众介绍推展	问世, 面世, 启动, 出炉, 发布
增添	旧：v. 添置物品	添置, 添购, 购置, 置备, 采办
	新：v. 增加、焕发（抽象）	焕发, 展现出, 注入, 充满, 洋溢
缺口	旧：n. 物体缺掉一块形成的空隙	城墙, 堵, 大洞, 豁口, 口子
	新：n. 物质供应的空档	原材料, 供求矛盾, 财力, 资金, 短缺
渠道	旧：n. 用来排灌的水道	河道, 河渠, 水渠, 沟渠, 灌溉渠
	新：n. 门路、途径	途径, 手段, 方式, 机制, 平台
行列	旧：n. 队伍	队伍, 列队, 游行, 红场, 夹道
	新：n. 群体（抽象）	生力军, 新一代, 一跃, 跻身于, 前列

然后有词语领域或搭配方面的变化。比如“线上”曾经表示铁路沿线，如今则表示“网上”，“数字”曾经仅指数学、统计上的数字，现有与信息技术、互联网相关的“digital”义。“重创”、“前沿”、“基地”、“攻克”这些词早期常与战争联系在一起，如重创敌军，阵地前沿，军事基地，攻克某座城市等，后期则用于经济、科技、制造业等领域，如经济受到重创，科技前沿，产业基地，攻克技术难题等。“采集”早期表示收集作物或植物标本，现表示生物信息如指纹、血样的采集；“加盟”原先常与国家和国际组织搭配使用，现在则更多用于体育运动员的转会。此外，“内地”原先是相对于新疆、西藏等边远地区的概念，后来则与港澳台地区对应；“挑战”最早用于生产队之间的竞赛活动，后来逐渐转移到中国在国际政治经济方面遇到的挑战，自 90 年代以来则大量出现在全球化变革的语境中，与“机遇”并列。这些变化是中国内外部环境发展的体现。

表8 领域/搭配变化

词语	领域/搭配	示例近邻
线上	旧：铁路“线”	青藏, 穿越, 沿线, 全线, 枢纽
	新：网络“线”	线下, 在线, 网上, 网上店铺, 网络营销
数字	旧：统计	统计数字, 百分比, 数目, 数据, 比率
	新：互联网（digital）	因特网, 互联网, 5G, 数字化, 人工智能
重创	旧：战争	痛击, 痛歼, 全歼, 歼灭, 击溃
	新：政治经济	难民潮, 经济危机, 经济衰退, 沉重打击, 巨大损失
前沿	旧：战争（阵地前线）	前沿阵地, 工事, 驻守, 哨所, 炮兵阵地
	新：科技	科技前沿, 学术前沿, 前沿技术, 制高点, 新兴学科
基地	旧：军事	军事基地, 空军基地, 海军基地, 导弹基地, 军港
	新：产业建设	重化工, 农牧, 垦区, 产业基地, 示范园区
散发	旧：传单	张贴, 传单, 宣传品, 散布, 传阅
	新：气味	散着, 散发出, 弥漫着, 扑鼻, 馥郁
采集	旧：植物	采, 采挖, 采制, 药材, 标本
	新：各类信息	收集, 录入, 指纹, 血样, 数据库
挑战	旧：生产竞赛	组与组, 各队, 比赛, 竞赛, 应战
	中：国际政治经济	舆论, 讹诈, 战争贩子, 孤立主义, 经济危机
	新：全球化机遇	世纪之交, 变革, 全球化, 机遇, 当今世界
加盟	旧：加入国际组织	拉脱维亚, 俄罗斯联邦, 乌克兰, 苏维埃, 共和国
	新：加入体育队	转投, 转会, 上海东方, 重庆力帆, 国际米兰
支配	旧：意识形态的控制	依赖于, 依附, 制约, 垄断, 资本主义
	新：调度安排财富	家庭财产, 城镇居民, 人均, 恩格尔系数, 实际收入
防护	旧：植被/荒漠化	植造, 森林, 防沙, 沙荒, 林区
	中：尘/毒/污染	防毒, 防尘, 防污, 放射, 环境监测
	新：防疫	防疫, 消毒, 卫生防护, 洗消, 消杀
部署	旧：兵力	布署, 调兵遣将, 诱敌深入, 作战方案, 军事演习
	中：核弹	导弹, 核导弹, 中程导弹, 巡航导弹
	新：政策	指示精神, 周密安排, 战略部署, 决策, 文件精神
攻克	旧：攻下、攻占	收复, 攻占, 攻陷, 攻下, 攻破
	新：解决难题	难关, 闯过, 技术难题, 突破, 卡脖子
出土	旧：植物从土中生长	出苗, 发芽, 生苗, 幼苗, 栽下
	新：古器物被发掘	窑址, 汉墓, 陶俑, 商代, 青铜器
内地	旧：相对新疆西藏地区	边远地区, 边疆, 阿勒泰, 伊宁, 喀什
	新：相对港澳台地区	港澳台, 两地, 祖国大陆, 大陆, 转口贸易

最后，部分修饰词本身的含义不变，但其常见修饰的对象褒贬色彩发生了较为明显的变化。如“十足”旧时常修饰表示贬义的形容词，如“十足伪善”、“十足可笑”、“十足愚蠢”等，现在则多为褒义，如“潇洒十足”、“霸气十足”、“十足的风度”。而“纯粹”过去也有比较贬义的搭配，如“虚伪”、“无知”、“低级趣味”等，现在则是“低调”、“率性”、“正直”这类更正面的组合。

表9 感情色彩变化

词语	解释	示例近邻
十足	旧：常与贬义的形容词连用	伪善，可笑，愚蠢，傲慢，圆滑
	新：常与褒义的形容词连用	潇洒，霸气，风度，灵气，才气
纯粹	旧：常与贬义的形容词连用	庸俗，虚伪，无知，浅薄，低级趣味
	新：常与褒义的形容词连用	低调，率性，善良，正直，朴实无华

注意以上只是从余弦相似度排序词表中人工挑选和归纳的一部分例子，旨在提出历时词向量可以用于捕捉和刻画这几类词汇语义演变现象的初步结论。由于本文使用的语料仅包含《人民日报》这一份报纸，这些现象也有可能受到新闻所关注的题材和语体、语言风格的影响，具体还需要进一步研究。

3.3 定量分析

3.3.1 量化语义变化

两个单词之间的语义相似度/距离可以由其向量之间的余弦相似性/间距来近似刻画[29]。对于历时词嵌入，可以通过两种方式来量化语义变化：（i）单个词的向量如何随时间变化；（ii）成对词之间的相似度随时间的变化 [4]。

对于第一种方法，本文以目标词在最早时间片上的嵌入为基准，绘制该词的语义相似度随时间下降的曲线。对于第二种方法，本文取目标词最具有代表性的四个邻居在最晚时间片上的嵌入（即这些词最近的、最为读者所熟悉和接受的语义表征）为参照，绘制目标词语义与这些邻居之间相似度的变化曲线。两种方法都分别在 5 年和 1 年的时间片上进行了测试，并将获得的结果进行比对。

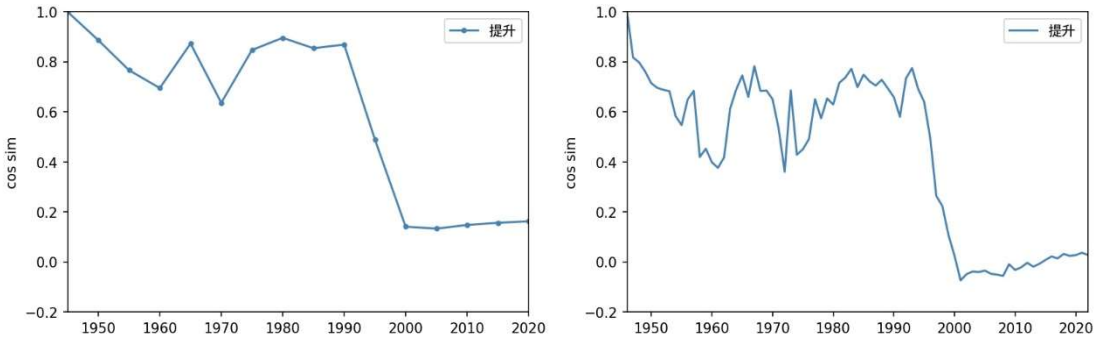

图 8 “提升”在 5 年和 1 年切片上的语义相似度下降曲线

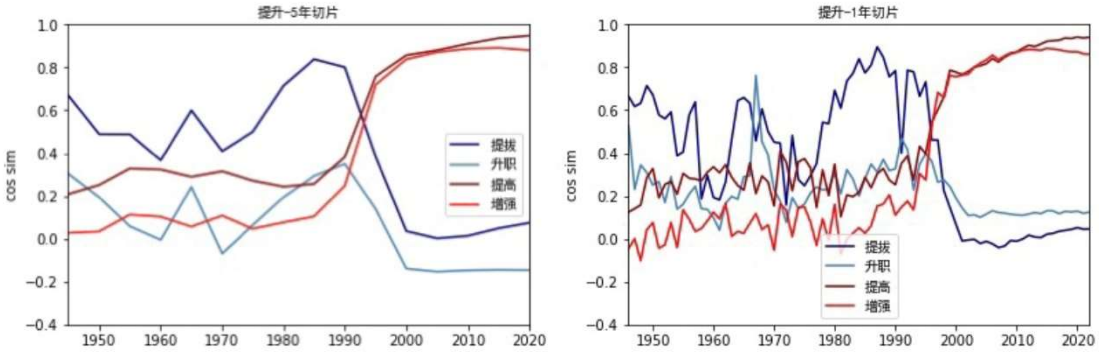

图 9 “提升”在 5 年和 1 年切片与邻居的相似度变化

从曲线可以看出“提升”一词的语义在 60-70 年间有所波动，尤其与“职位晋升”相关的部分。在 90 年代词义变化最快，完全趋向提高、增强的含义，此后便趋于稳定，这和近邻表与迁移轨迹图的直观印象是一致的。

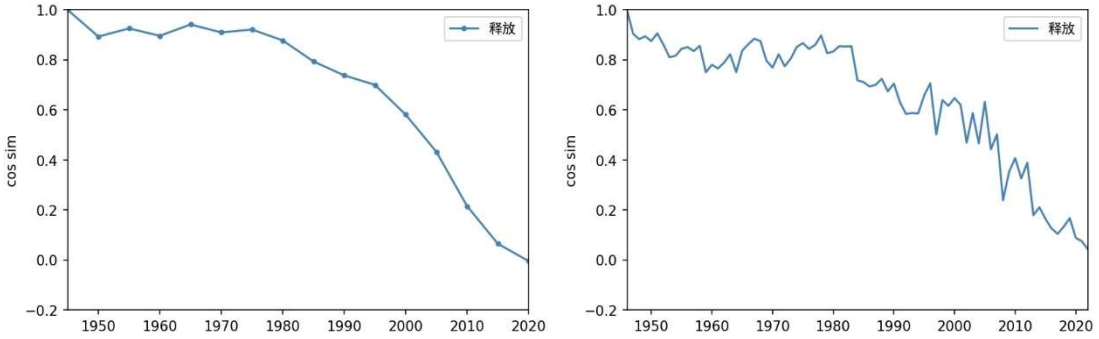

图 10 “释放”在 5 年和 1 年切片上的语义相似度下降曲线

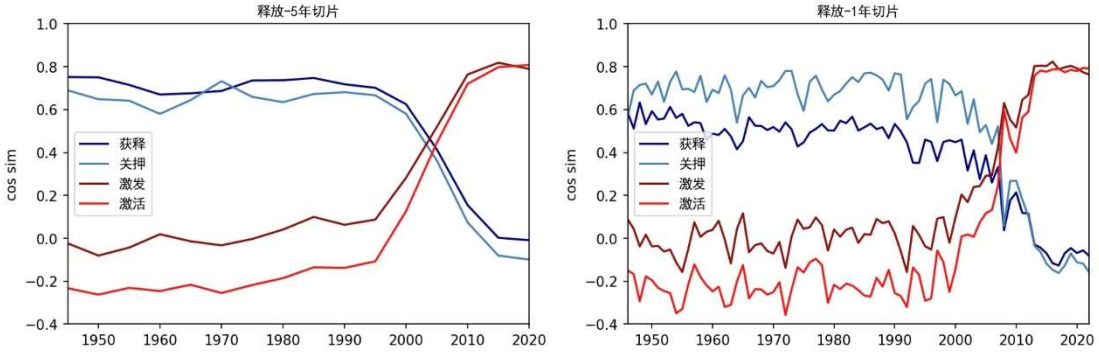

图 11 “释放”在 5 年和 1 年切片与邻居的相似度变化

也可以看出“释放”从刑事用词到抽象的激发、激活过渡，其中变化最明显的时期是 2000-2010 年间，其他时间则相对较平缓，这与轨迹图中 2000 年以前及 2005 年以后相邻时间片间距短，仅这两个时间片距离较长也是一致的。

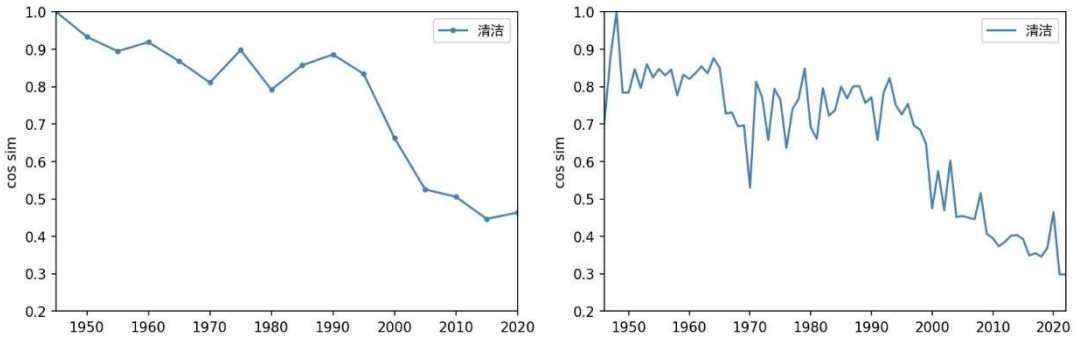

图 12 “清洁”在 5 年和 1 年切片上的语义相似度下降曲线

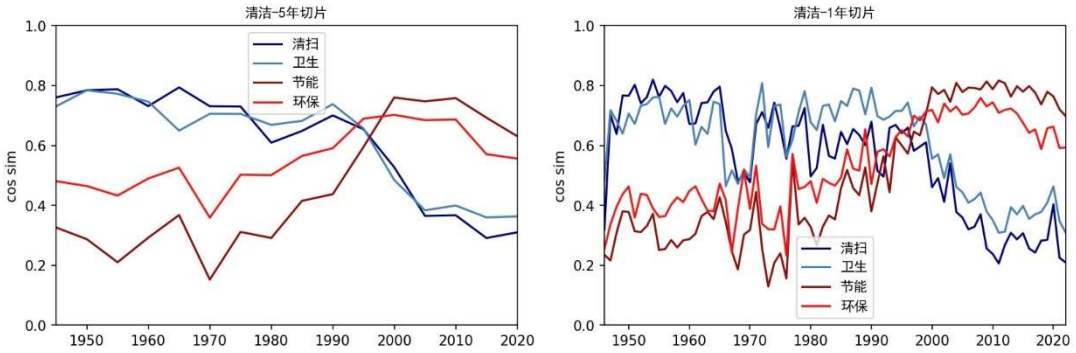

图 13 “清洁”在 5 年和 1 年切片与邻居的相似度变化

同样可见“清洁”从形容词干净、卫生和动词清扫、打扫到“节能环保”的变化是在 2000 年左右最为显著，与之前的结论相同。此外，两种时间切片的曲线总是大致相似，其中 5 年切片更平滑，而 1 年切片波动干扰较多。可见较细粒度的时间片切分可能捕捉到更幽微的语义变化，但同时受制于语料库的大小，随机噪声的影响较为明显。不妨将不同尺度的切分方式结合使用，互为对照。

3.3.2 跨时间类比任务

众所周知，嵌入模型可以捕捉概念之间的复杂关系，如单词类比任务要求模型通过向量加减法“求解”一系列形如“单词 w_1 对单词 w_2 就像单词 w_3 对单词 w_4 一样”的问题。一个经典的例子是英语中“男人”和“女人”之间的关系与“国王”和“王后”之间的关系[7]，这一事实可以通过简单地加减相应的词向量来验证（king – man = queen – woman）。类似地，除了人类的两种性别，词向量还可以对单词之间的其他多种关系进行建模，如对某种动物的统称和对该动物幼崽的昵称的关系（kitten – cat = puppy – dog），国家和首都、国家和法定货币的关系（Italy – Rome = France – Paris，Japan – yen = US – dollar）等。

事实证明，词向量的比较可以在不同的向量空间之间进行（当然，前提是这些向量空间是对齐的），从而实现不同方言之间甚至跨语言的关系建模。对英式英语和美式英语的语料，可以找到一个概念在两种方言中的惯用说法，比如“英式英语中的 lift 相当于美式英语中的 elevator”[23]。Mikolov 于 2013 年就极富创造性地提出了单词类比任务可以跨语言进行，如英语和西班牙语中数字和动物在向量空间中的相对位置关系是非常接近的，基于这个发现，可以通过大量的单语语料和少量平行语料获得语言之间的单词和短语映射，用于机器翻译[30]。

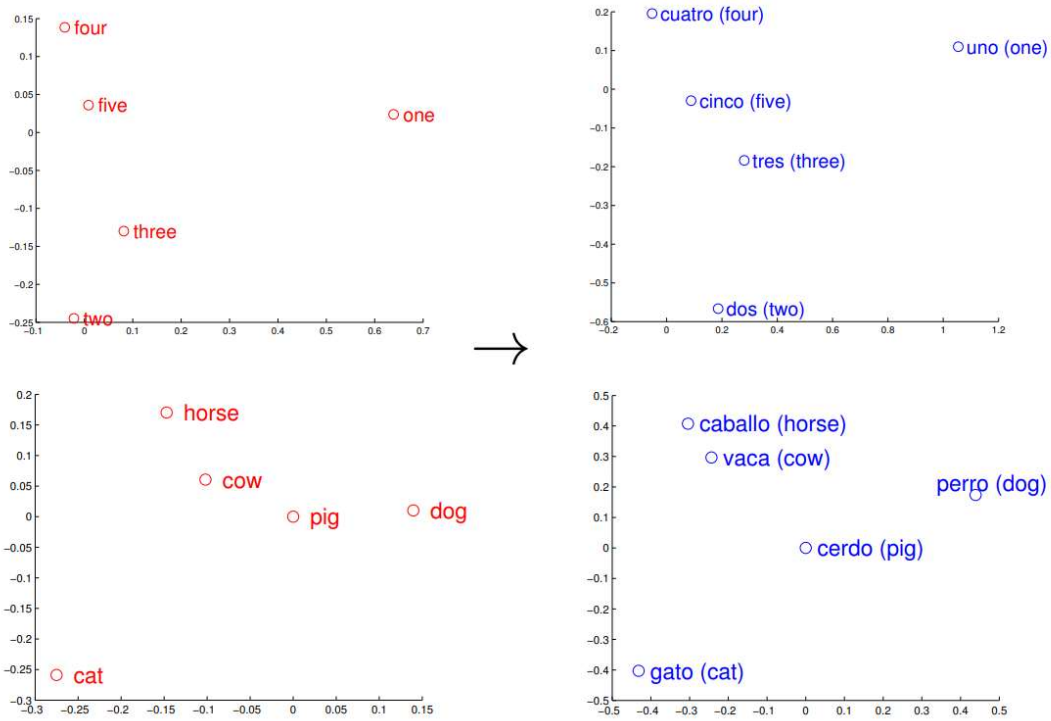

图 14 英语和西班牙语中的数字和动物[1]

除了跨方言和跨语言的单词关系建模，Szymanski 提出了跨时间类比，即“时间 t_α 的单词 w_1 就像时间 t_β 的单词 w_2 一样”，比如“1987 年的罗纳德·里根就像 1997 年的比尔·克林顿”，“1987 年的随身听就像 2007 年的 iPod” [23]。

词语含义的变化可粗略地分为两种，一种是词义本身改变，即语言系统内部的变化。另一种则是外部世界变化导致词语或命名实体的内涵（connotation）和外延/指称（extension/denotation）的变化。前一种例如英文单词 awful 的含义近

[1] Mikolov, T. et al. 2013. Exploiting Similarities among Languages for Machine Translation.

百年来从“令人敬畏的”（awe-inspiring）逐渐变成“糟糕的”（very bad）[23]，而后一种则包括“美国总统”一词在不同时期指向不同的对象，因为这一职务在现实中也由不同的人担任。第二种变化有时并不被语言学家视为语义演变，但在历时词嵌入模型的质量评估中有很大作用，且相对第一种更易于获得。

Szymanski 将跨时间的类比词（Temporal Word Analogies）定义为在不同时间点位于语义空间（近似）同一位置的词[23]。例如，假设存在与“美国总统”相关的语义空间，这个空间在 20 世纪 80 年代被罗纳德·里根占据，在 20 世纪 90 年代则被比尔·克林顿占据。“1987 年的罗纳德·里根”和“1997 年的比尔·克林顿”就是一对跨时间的类比词。英语中已有两种针对跨时间类比词的测试集，一种基于公开记录的知识，主要包括了客观上确实在变化的角色（例如美国总统、纽约市长、超级碗冠军等）[23-24]，另一种则是更依赖人主观判断的概念迭代（如音乐的载体从 CD 逐渐过渡到 MP3）。显然，尽管第一个测试集比起语言本身的变化更像是对现实世界的反映，但在模型质量评估方面更可靠；第二个测试集不确定因素较多，但可以探索更微妙的概念变化，反映社会环境对语言的影响，如新兴技术、品牌和重大事件（疾病爆发、金融危机等）。在 Szymanski 的工作中可以看到 MP3 在 20 世纪 90 年代取代光盘和磁带成为音乐消费的主要形式，后又被在线流媒体 Napster 和 iTunes 取代。21 世纪的 iPhone 和智能手机在 90 年代的类比词是台式机、PC 而非“电话”，可见智能手机的用途不止接打电话，定位更接近便携式电脑。这些是从历时词嵌入中发掘科技发展趋势的例子。Yao 等还探讨了政治事件和流行文化相关词汇，如 1992 年前的“苏联”相当于之后的“俄罗斯”，1987 年的“雅皮士”（yuppie）相当于 2003 年的“潮人”（hipster）[24]。这样的测试集很难大量获得，往往只包括少量人工挑选的例子。一种退而求其次的方法是先让模型自动生成类比词，然后再由人类对其准确性进行打分。

本文提出这两种测试集在中文上的一些实例。对于客观变化的角色，本文针对美国总统进行了测试。对于概念变化，本文主要考虑与科技有关的词汇，如手机、平板、微信。由于这些概念的语义空间变化较快（如总统的换届和电子产品的换代往往以年为单位），需要对更小的跨度进行考察，故只在 1 年时间片上进行了测试。下面取 2022 年的词嵌入作为基准，寻找不同年份的类比词。

3.3.3 现实角色变化

对照 2022 年的“拜登”在各年份的类比词和美国总统的任期可见，词嵌入模型基本捕捉了现实的政治角色变化，模型的性能在一定程度上是有保障的。

表12 2022年的“拜登”的类比词

年份	最近邻
1946-1952	杜鲁门
1953-1959	艾森豪威尔
1961	肯尼迪
1963	约翰逊
1973-1974	尼克松
1975-1976	福特
1977-1981	卡特
1982-1988	里根
1989-1992	布什, 布什总统
1993-2001	克林顿
2002-2008	布什, 布什总统
2009-2016	奥巴马
2017-2020	特朗普
2021-2022	拜登, 特朗普

3.3.4 社会变迁

表13 2022年的“手机”的类比词

年份	最近邻
1946	电报
1953-1993	电话, 电话机
1994-1996	大哥大
1998-1999	BP 机
2000-2002	呼机, 寻呼机, 传呼机
2003-2022	手机

表14 2022年的“电脑”的类比词

年份	最近邻
1947-1955	收音机
1956-1979	收音机, 电视机
1980-1982	打字机, 收音机
1983-1991	打字机, 电脑
1992-2010	电脑
2011-2012	平板, 电脑
2013-2022	平板

表15 2022年的“CD”类比词

年份	最近邻
1949-1956	唱片，留声机
1957-1980	唱片
1981-1998	磁带，影碟，唱片，唱机
2000-2003	CD, VCD, ROM, MP3
2005-2007	随身听，播放器
2009-2022	CD, VCD, DVD, MP3, MIDI

表16 2022年的“电动车”的类比词

年份	最近邻
1946-1948	三轮车，人力车
1949-1950	人力车，自行车
1951-1977	自行车
1978-1998	自行车，摩托车
2002-2007	摩托车
2011-2022	电动车，摩托车

表17 2022年的“微信”的类比词

年份	最近邻
1946-1950	信件，电报
1951-1979	信函，书信，电话
1980-1993	电话号码，电话，号码
1994-1999	传真，热线，热线电话
2001	传真，电子邮件，电子信箱
2002-2007	短信，邮箱
2008-2010	邮箱，短信, QQ
2011-2013	QQ
2014-2022	微信, QQ

表18 2022年的“微博”的类比词

年份	最近邻
1946-1962	书报，报，报纸
1964-1974	收听，广播电台
1976-1996	栏目，专栏，版面
1997-1998	网址
2001-2009	搜狐网，新华网，人民网
2010-2014	微博，新浪
2015-2022	微博，公号，新浪

从上面的表格可以看出：

(1) 手持移动电话取代之前的电报和电话，具体产品从大哥大（94-96）依次过渡到 BP 机（98-99）、呼机/传呼机（00-02）再到手机（03-22）（表 13）；

(2) 可供人们休闲娱乐的家用电器从收音机、电视机（1947-1982）逐渐变为电脑（1992-2010）和近年来更流行的平板（2011-2022）（表 14）；

(3) 音乐媒介从留声机、唱片（1949-1980）变为磁带（1981-1998），最后再到 CD、MP3、MIDI 等（2000-2022）（表 15）；

(4) 代步工具从人力车、三轮车（1946-1948）到自行车（1951-1977）再到摩托车（1978-2007）、电动车（2011-2022）（表 16）；

(5) 通信方式从信函、电报（46-50）到电话（51-93）、传真（94-99）、电子邮件（2001）、短信（02-07）、QQ（08-13）和微信（14-22）（表 17）；

(6) 获取新闻资讯的渠道从书报（1946-1962）、广播电台（1964-1974）到新闻网站（2001-2009），最后到微博、公号（2010-2022）（表 18）。

这些词的对应关系是建国以来人们的生活方式在出行、社交、娱乐等各个层面上变化的真实反映，由此可见，跨时间类比词可以捕捉到新事物何时出现和取代旧事物，在研究人类社会的发展方面具有一定潜力。

此外，本文还试图通过跨时间类比任务来研究两个更加微妙的话题，其一是流行病在中国的发展历程，其二是世界范围内大的武装冲突。众所周知，2022 年在中国最受关注的传染病关键词是“新冠”，而国际上影响最大的武装冲突关键词则是“俄乌”。因此，本文分别取 2022 年的“新冠”和“俄乌”两个词作为搜索的基准向量，在 1946-2022 的每一年上分别搜索距离排名前十的类比词，对其中出现的疾病名和国家/地区组合分别进行了统计。

在流行病方面，由于这些疾病的分布较为零散，很难对相邻的年份进行合并处理，从而直观地提炼出变化规律，本文采用可视化方法，首先将同一病种和病毒合并（如“乙肝”、“乙肝病毒”和“乙型肝炎”，“艾滋病”和“HIV”等），然后筛选出这 77 年间出现次数在 4 次以上的病种，最后对每个病种出现的年份进行画图，则图中点的位置分布可视为对某病在中国流行的时间。

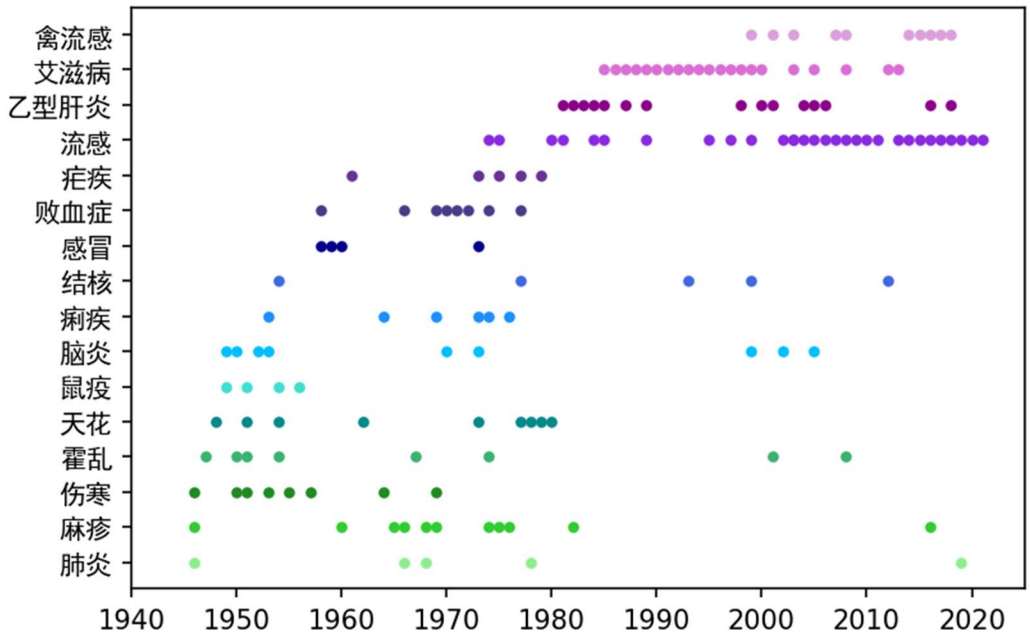

图 15 1946-2022 年间作为 2022 年的“新冠”类比词出现的病名

根据中国疾病预防控制中心[1]的官方网站资料，我国于 80-90 年代基本实现伤寒发病率的控制。过去民间有俗语“孩子出过疹和痘，才算解了阎王扣”，其中疹指麻疹，痘指天花。1965 年我国开始接种液体麻疹疫苗，随着 EPI 和冻干苗的发展，麻疹发病率在 80 年代得到控制。1952 年，我国开始进行有组织的牛痘疫苗接种活动，80 年代天花在全世界范围内都基本绝迹。诸如霍乱、痢疾、疟疾、鼠疫、败血症这些由各种细菌造成的血液和肠道传染病，随着解放后经济社会的发展、生活条件的改善和卫生知识的普及，在 80 年代也逐渐销声匿迹。除去逐渐消失的传染病，近年也有一些新病种，如 80 年代传入的艾滋病，21 世纪几次爆发的禽流感等。此外，尽管乙型肝炎在中国由来已久，但在 80 年代通过大规模的不规范采血迅速扩大传播，才逐渐引起重视。

本文对个别出现频率较低的病种也进行了考察。如上表所示，模型正确地发现了 2003 年非典的爆发、2005 年 H2N2 流感病毒误发的事件，2015 年 MERS 在韩国的爆发（由于韩国是中国的邻国，国内媒体对此关注较多），2016-2017 年 H5N6、H5N8、H7N9 等不同亚型禽流感病毒及甲型 H1N1 流感的先后流行，以及 2021-2022 年新型冠状病毒毒株 Omicron 和 Delta 的变异。

[1] https://www.chinacdc.cn/

表19 病名作为2022年的“新冠”类比词出现的时间

病名	年份
SARS	2003, 2004
H2N2	2005
MERS	2015
H5N6	2016
H5N8	2016
H7N9:	2016, 2017
H1N1	2017
奥密克戎	2021, 2022
德尔塔	2021, 2022

对世界范围内的武装冲突，筛选 77 年间出现次数在 2 次以上的关键词，根据每个关键词出现的年份进行画图，得到重大武装冲突发生的时间。

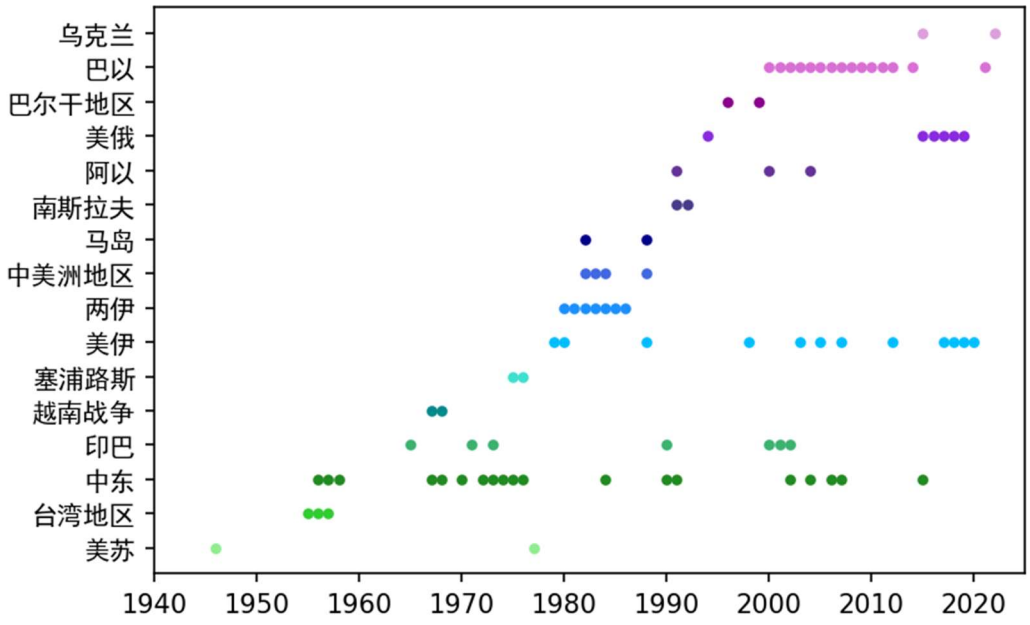

图 16 1946-2022 年间作为 2022 年的“俄乌”类比词出现的国家和地区

可见模型准确地捕捉到了 1950 年代下半叶的台海地区紧张局势，70 年代中期的塞浦路斯政变与希土之争，80 年代持续长达八年之久的两伊战争，以及马岛战争，南斯拉夫内战，从 90 年代起就持续不断的巴以冲突等，且成功发现了乌克兰上一次卷入战争的时间（2014 年克里米亚并入俄罗斯联邦）。

第四章 结论和展望

4.1 总结

本文使用基于 Word2Vec 的 TWEC 模型，在 1946-2022 年的《人民日报》新闻数据上对不同粒度的时间片训练中文历时词嵌入模型。获得词表征后，本文筛选出可能存在语义变化的词表，分别进行了定性和定量分析。

在定性分析中，本文通过单词邻域分析了部分单词的含义变化，对变化的类型作初步归纳，并提供了词义变化轨迹的可视化结果。在定量分析中，对单词自身和与相邻词语的历时相似度进行了对比，可以确定词义变化的幅度和关键时间点，与定性分析阶段的直观感受一致。此外，本文分别从角色变化和概念变化两个方面设计了中文的跨时间类比任务，对词向量进行测试，获得的结果与现实情况能够匹配。最后，尝试用跨时间类比词捕捉流行病的发展趋势，并提出历史词嵌入模型用于与时事有关的社会学研究的一种可能路径。

4.2 展望

4.2.1 对上下文的建模

由于大多数基于上下文的历时词嵌入模型需要人工设定聚类的数量[21]（而这个数量如何确定尚不存在确切的依据和标准），或者需要足够多的权威例句来计算目标词的含义表示[22]，本文采用的仍然是更早的 Word2Vec 模型。然而，将不同语境中的单词表示为同一向量存在含义混淆的缺陷，不利于对语义进行准确的建模，比如英文单词 mouse 有“老鼠”和“鼠标”这两个相去甚远的含义，用单一向量来表示难免显得力不从心。在语义空间中，mouse 向量被“鼠标”的含义“拉动”而逐渐脱离动物的邻域向计算机配件移动，但这一运动轨迹是十分模糊并难以测量的。相比之下，基于上下文的建模方式解决了一词多义的问题，在语义变化的测量上优势格外明显。假设模型能分辨 mouse 具体在哪些上下文中表示老鼠，哪些上下文中表示鼠标，也就可以精确地定位“鼠标”的语义在何时何处第一次被使用，使用的频率如何随着时间变化等。

Hu 等人提出了一个建设性的方法，即使用预训练语言模型（如 BERT）和权威词典数据来构建含义（sense）的表示[22]。这样的权威词典需要（1）包含每个词语在不同时期含义和用法的全面记录；（2）每种含义附带足够数量的例句。遗憾的是，包括《现代汉语词典》在内的中文词典通常不提供太多例子，且习惯对词语的含义使用示例词组、短语而非例句来说明。如果要构建中文里 senses 的类似表示，除了参照词典条目之外，还需要获得全面、丰富的例句库。

获得目标词的含义表示后，就可以将语料库不同时间片上包含这个词的句子输入到 BERT，以获得该词的上下文嵌入，计算这个嵌入与目标词含义表示的相似度，取其中最接近的为词语在句中的含义即可。这种方法精度很高，可以识别词典列出了但现实中使用极少的词义，而在单一向量的表征中，不常用的词义对整个向量影响几乎可以忽略不计。这个模型可以精密检测和量化每个 sense 在历史上使用频率的变化，比如对 please 一词用法的研究显示，该词“对不合理的事物表示愤怒”这一含义逐渐消亡，“（使对方）高兴和满足”及“情愿做某事”的含义相对稳定，而“表示礼貌性请求或询问”的用法在增加。总的来说，如果能够获得合适的数据集，从而把蕴含上下文的词嵌入模型应用到历时词义变化的研究中，对于先前的单一词表征方法将会是很好的扩展和补充。

4.2.2 语言学研究

对语义演变的研究是传统语言学的重要分支。历时词嵌入模型不但可以定位和描述语义演变的现象，而且有助于验证语言学理论，甚至可能从数据中发现新的理论。如 Hu 将语义演变与生物界的进化理论结合起来，从生态学的角度对单词的变化进行建模，类比种群之间的竞争与合作提出语言“进化”过程中两种有趣的现象，即含义的竞争与合作[22]。对于多义词来说，新含义的出现和传播往往挤占旧含义的“生态位”，使之逐渐被淘汰，这是符合直觉的。此外，某些相近或关联的含义表现出共同消长的趋势，比如 gay 作为名词的“同性恋者”和作为形容词的“同性恋的”这两个含义是密切相关的，增长曲线也十分一致，它们“合作”淘汰了 gay 的“欢快、无忧无虑”义，这与语言使用者的心理有关，即相似或关联的含义在使用中加深彼此的印象，逐步形成与词汇的绑定。基于词嵌入模型的语义研究是一个有趣的交叉领域，相信未来还会有更多奇妙的发现。

4.2.3 社会学研究

本文尝试用历时词表征来探测传染病的流行。在人文社科领域有许多类似的研究，从文本中挖掘反映现实事件的文化语义变化，如跨时间的信息检索[31]，预测社会动乱[25]，通过词向量范数追踪实体的流行度[24]等。这些技术在商业运营中可以被用于分析大众心理、提高用户体验，还可以预测民意的变化趋势，为政府机构制定政策提供参考，可以说具有相当的应用潜力。

4.2.4 中文历时语料建设

由于中文的发展历程中有白话文运动这一特殊阶段，导致中文在短时间内词汇、语法和语言风格都发生了巨大的变化，文言文和白话文各方面差别较大，不像古英语到现代英语的转变时间较长，速度较慢，故中文的历时语料通常只收录白话文，最早追溯到民国时期。相比之下，文言文历史更加悠久，留下的文字记录也很丰富，或许可以对文言文另建历时语料库，分析古汉语的词义演变。

在体裁方面，现有的中文历时语料库多为报纸、新闻语料，且通常使用官方报纸，如本文所用的《人民日报》。新闻的优点在于语言风格比较统一，而且包含大量的时事信息，缺点则在于题材单一，且为官方的书面话语，对民间口语中的词义变化无法触及。在此基础上，可以增设更多不同领域的文本语料库，如文学作品的语料库，学术论文的语料库等。此外，尽管网络平台上的贴文比较碎片化，而且可能包含大量不规范的缩写、谐音、省略等，但语言学和社会学家对网络流行词的发展很有兴趣，认为它们更日常，比起正式、书面的文体更能反映大众心理的微妙变化，因此用建立中文历时网络语料库也是很有意义的。

参考文献

[1] M. Hilpert. 2008. Germanic future constructions: A usage-based approach to language change. Benjamins, Amsterdam, Netherlands.

[2] Xu, Y. and Kemp, C. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX, USA.

[3] Eger, S. and Mehler, A. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

[4] Hamilton, W. L., Leskovec, J, and Jurafsky, D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

[5] Kutuzov, A. et al. 2018. Diachronic word embeddings and semantic shifts: A survey. ArXiv. https://doi.org/10.48550/arXiv.1806.03537.

[6] Bloomfield, L. 1933. *Language.* Allen & Unwin.

[7] Mikolov, T. et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems.*

[8] Heyer, Gerhard., Holz, F. & Teresniak, S. 2009. Change of topics over time: tracking topics by their change of meaning. In *Proceeding of the International Conference on Knowledge Discovery and Information Retrieval*, Madeira, Portugal.

[9] Jurgens, D. and Stevens, K. 2009. Event detection in blogs using Temporal Random Indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16, Borovets, Bulgaria.

[10] Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy.

[11] Sagi, E., Kaufmann, S., and Clark, B. 2011. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, pages 161–183.

[12] Kim Y. et al. 2014. Temporal analysis of language through neural language models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 61–65, Baltimore, USA.

[13] Liao, X. and Cheng G. 2016. Analysing the semantic change based on word embedding. In *Natural Language Understanding and Intelligent Applications*, pages 213–223. Springer International Publishing.

[14] Rosenfeld, A. and Erk, K. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484, New Orleans, Louisiana, USA.

[15] Tsakalidis, A. et al. 2021. DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data, 8*.

[16] Bamler, R., & Mandt, S. 2017. Dynamic Word Embeddings. *International Conference on Machine Learning*.

[17] Zhang, Y. et al. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, October.

[18] Di Carlo, V., Bianchi, F., & Palmonari, M. 2019. Training Temporal Word Embeddings with a Compass. *AAAI Conference on Artificial Intelligence*.

[19] Peters M. et al. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

[20] Devlin J. et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[21] Giulianelli, M., Tredici, D., and Fernández, R. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

[22] Hu, R., Li, S., and Liang, S. 2019. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

[23] Szymanski, T. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. Association for Computational Linguistics.

[24] Yao, Z.; Sun, Y.; Ding, W.; Rao, N.; and Xiong, H. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 673–681. ACM.

[25] Kutuzov, A., Velldal, E., and Øvrelid, L. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop at ACL 2017*, pages 31–36, Vancouver, Canada.

[26] 刘知远,刘扬,涂存超,孙茂松. 词汇语义变化与社会变迁定量观测与分析. *语言战略研究*, 2016, 1(6): 47-54.

[27] 孙琦鑫,饶高琦,荀恩东. 基于长时间跨度语料的词义演变计算研究. *中文信息学报*. 2020, 34(8): 10-22.

[28] Smith, S. et al. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

[29] Turney, P., D. and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37(1):141–188.

[30] Mikolov, T., Le, Q.V., & Sutskever, I. 2013. Exploiting Similarities among Languages for Machine Translation. *ArXiv, abs/1309.4168*.

[31] Rosin, G., D., Adar, E., and Radinsky, K. 2017. Learning word relatedness over time. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189, Copenhagen, Denmark.

致谢

我在我的主修和辅修专业——计算机和英语系，都算不上聪明的学生。高等数学很难，编程很难，理解那些晦涩的文学理论也很难。我感到我在两个专业的夹缝里挣扎，好像什么都会一点但都会得不多，总是努力地想要把兴趣和专业缝合到一起又总是失败。好在还有语言学，以及它的诸多交叉学科。我至今不敢说我学到了什么，或者未来必定从事什么事业，但至少这四年的学习——如饥似渴地学习，就像一块吸水的海绵，我很充实也很快乐，相信一切都是有意义的。

我经常和人开玩笑说在南京大学念书是我的人生巅峰——好像有点儿妄自菲薄，但也是真心话。高考发挥超常，我总觉得我来了一个我配不上的地方，这里的老师真的很好，我想我一旦走出校门就再也见不到这么好的老师们了，我还能怎么办呢，我只有拼命地上课，把课程表上每一个能塞的空隙都填满。

我说不出来，但是我真的感谢各位老师，你们让我看到了一个新的世界，感觉就好像高雅的什么殿堂，我进不去，但是趴在墙头看了一眼，很知足。我想感谢冯新宇老师，您的 SICP 让我对这个专业产生兴趣，断了转系的念头；感谢梁红瑾老师，形式语义真的让人大开眼界，虽然这门课我修到第二遍才懂……感谢黄书剑老师，让我对自然语言处理这个领域有了初步了解。

除此之外，还要感谢英语系的 Tony Sansotta，您严酷的写作课让我在雅思考试中得心应手；感谢 Shawn Burtoft，您的语言学概论和语义学非常有趣，让我接触到很多新的概念；感谢俞希老师，让我学到一些社会学研究方法；感谢文院的张安琪老师，您是天使；感谢刘润泽和黄鑫宇老师，很遗憾到大四才上了你们的课，虽然目前能力还有限，但是我真的很喜欢研究语言和翻译这些东西。

感谢我的家人和朋友，在我身体和心理最脆弱的时期陪我渡过难关，最后要感谢的是我自己，我知道我通常不说自己什么好话，但是能坚持到现在已经很不错了，至于还有什么不够的地方，再继续努力吧，人生还很长呢。