DOI:10.19689/j.cnki.cn10-1361/h.2016.06.010

词汇语义变化与社会变迁定量观测与分析*

刘知远　刘　扬　涂存超　孙茂松

（清华大学计算机科学与技术系　北京　100084）

提　要　随着社会的发展和科技的进步，人们交流的内容与方式发生着翻天覆地的变化，交流所使用的词汇和语义也发生了显著变化。在过去的研究中，研究者主要通过词汇的使用频度变化来观测和分析词汇语义的变迁，取得了很多重要发现。但是这些词频统计方法无法考虑词汇的语义内涵。为了更精确地捕获词汇语义变化及其反映的社会变迁，我们利用分布式词表示方法，提出将词汇的多个词义用不同的低维向量表示。利用该方法，研究者可以根据词义使用频度的变化情况，定量观测与分析词义变化与社会变迁。这将为语言演化、社会语言学乃至语言规划研究提供重要量化工具。

关键词　词汇语义；社会变迁；时序信息；分布式表示；词向量

Lexical Semantic Variation and Social Change: Quantitative Observation and Analysis

Liu Zhiyuan, Liu Yang, Tu Cunchao and Sun Maosong

Abstract　With social and technological developments, the contents and means of human communication have undergone tremendous changes, which, in turn, lead to the evolution of word forms and their meanings in human language. In literature, much scholarship has been devoted to the semantic dynamics of words from the perspective of usage frequency, yet this frequency-based method cannot explain clearly the lexical-semantic change due to its failure to cover word senses. In this paper, a large-scale Chinese newspaper text corpus is employed and the distributed representations of some words and their senses are elicited in order to observe the diachronic evolvement of word semantics. The semantic change of the words in the timeline suggests that the distributional method proposed in this paper is effective for the exploration of lexical semantic dynamics. The implication of this study is that the corpus-based distributional method can become a useful tool for studies in other fields, such as language evolution, sociolinguistics and language planning.

Key words　lexical semantics; social change; temporal information; distributed representation; word representation

一、研究背景

词汇语义变化是指随着时间发展，一个词的使用方式发生了较大程度的变化（Traugott & Dasher 2001）。当时间跨度较大时，词汇语义变化现象尤其显著。词汇语义变化是一种非常普遍的现象，与人类进步与社会发展等有密切联系，是认知语言学和社会语言学等学科的重要研究课题。

研究词汇语义变化的方法之一是观测使用词频随时间变化的情况（Michel *et al*. 2011），有很多重

作者简介：刘知远，男，清华大学计算机科学与技术系助理教授，主要研究领域为自然语言处理和社会计算。电子邮箱：liuzy@tsinghua.edu.cn。刘扬，男，美国伊利诺伊大学香槟分校博士生，主要研究领域为机器学习。电子邮箱：largelymfs@gmail.com。涂存超，男，清华大学计算机科学与技术系博士生，主要研究领域为自然语言处理和社会计算。电子邮箱：tucunchao@gmail.com。孙茂松（通讯作者），男，清华大学计算机科学与技术系教授，主要研究方向为信息检索、人工智能和自然语言处理。电子邮箱：sms@tsinghua.edu.cn

*　该论文得到北京成像技术高精尖创新中心（BAICIT-2016006）、国家社科基金重大项目（13 & ZD190）和国家科技支撑计划（2014BAK04B03）的资助，特此致谢。

要发现，特别是在语言演化和人类文化研究等方面（具体介绍见第二部分）。但是，基于词频的方法无法考察词汇的不同词义，也无法考察词汇之间的语义关联，从而极大地限制了利用这类方法探索词汇语义变化的深度与精度。

近年来随着深度学习的发展，分布式表示成为自然语言语义表示的新兴技术。该表示技术通过机器学习技术，自动学习语义空间，并将语言对象（如词汇、短语、句子等）表示为该空间中的一个稠密、实值的低维向量（一般只有几百维）。语言对象在该语义空间中的相对距离代表它们之间的语义关联度。以面向词汇的分布式表示为例，如果两个词语的语义越相近，那么它们在该语义空间中对应的词向量夹角（即余弦相似度）就越小。由于该技术能够有效缓解大规模文本中的数据稀疏问题，因此在很多自然语言处理任务中取得了显著效果。

在面向词汇的分布式表示学习模型中，目前最流行的是 Mikolov 等（2013）推出的 word2vec。该模型能够从大规模文本数据中为每个词自动学习低维向量表示，我们将在第三部分介绍该模型的基本思想。但是 word2vec 模型默认只能为每个词学习一个向量，无法处理一词多义的情况。因此，有学者提出为词汇的每个词义学习单独的表示向量，并能够为文本中每个词自动分配最适合的词义向量。

综合上述词汇和词义分布式表示学习的优势，本文为每个词语的不同词义学习表示向量，通过观测该词语不同词义的使用概率分布的变化情况，研究词汇语义变化现象，并探索该现象与社会变迁的关系。

词汇是人类语言中负载信息的基本单位，考察文本大数据中词汇及其词义的时空变化模式，对于语言演化研究具有重要意义。相关分析结果也将为语言政策与语言规划工作（陈章太 2005；刘海涛 2007），如词典编纂、语音规范、术语翻译等，提供重要的量化依据。

二、词汇语义变化的相关工作

利用词汇使用在时空中的变化情况开展社会学研究工作，在国内外都不鲜见。例如，金观涛和刘青峰（2009）整理近代文献数据库，分析其中特定词汇的使用频度的变化情况，探讨了中国现代重要政治术语的形成，被公认为思想史研究的重要流派。近年来，哈佛大学研究团队提出“文化组学”[①]（Culturomics）的学术思想，利用 Google Books 收集的 1800 年到 2000 年间的 500 万种出版物，通过观测关键词使用频度随时间的变化，研究人类文化演进的模式与特点，取得很多突破性成果（Aiden & Michel 2013；Michel *et al*. 2011）。例如他们发现，在过去几百年里英语中越来越多的不规则变化动词转化成了规则变化动词（Lieberman *et al*. 2007）。再如，他们通过观测历年来使用“The United States is”和“The United States are”的频度变化，发现在南北战争后美国才逐渐被作为统一国家的概念为人们所接受（Aiden & Michel 2013）。可见，面向文本大数据词汇使用的定量分析，为社会科学研究提供了全新的视角。

新词语产生后会随着交流中的应用而广泛传播和演化。其流行程度和形式会随着时间而演化，出现爆发和变形。不同新词语的爆发程度和变形情况可能会受到不同因素的影响。同时，新词语使用者的社交网络往往受到地域限制，新词语的传播也会反映在地理位置的扩散上：一个新词可能会首先在某个地域流行，然后逐渐扩散到全国甚至全世界。

在线社会媒体的兴起与广泛应用，为研究者提供了词汇使用时空变化定量分析的重要平台。探索词汇的时空传播与演化具有重要研究意义。斯坦福大学 Leskovec 等（2009）从不同来源收集了约 9000 万篇新闻文章，利用引号从新闻中自动抽取流行语句，建立 MemeTracker 系统跟踪这些语句的使用频度随时间变化的情况，能够及时、有效地把握美国政治、经济和文化生活的热点信息。例如，作者提到“you can put lipstick on a pig”，是 2008 年美国总统大选中奥巴马讽刺竞选对手时引用的谚语，全句是“你就算给猪涂上口红，它也还是只猪”，在民众中引起广泛争议，也让最早出现于 20 世纪 20 年代的这条谚语重新流行起来，一时间成为美国民众很爱使用的谚语。作者还进一步使用聚类算

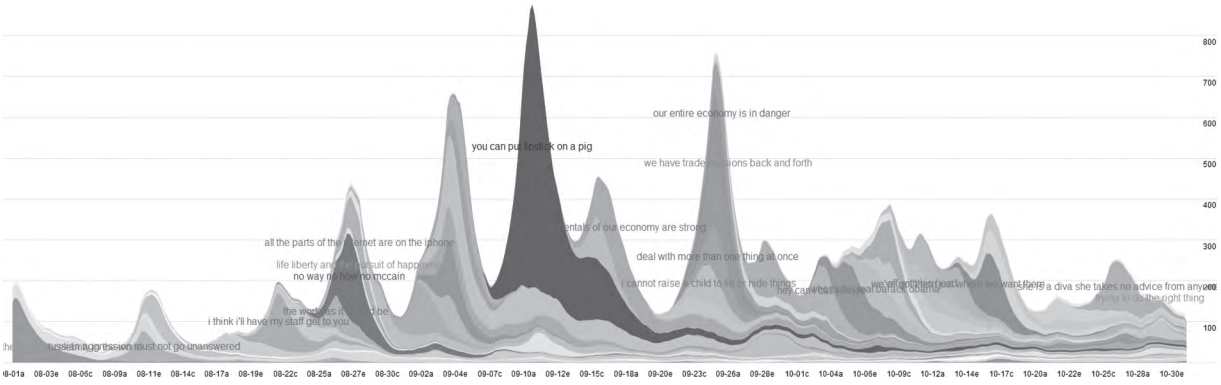

图 1　MemeTracker 提供的模因时序变化趋势②

法研究这些流行语扩散的时序特征，总结出六种时序类型（Yang & Leskovec 2011），这对探索词汇语义传播模式具有重要启发意义。

以上研究主要针对流行语使用和扩散的时序变化开展研究。由于不同地域在文化风俗、地标建筑和方言俗语等方面有显著差别，词汇使用也有明显的地域特色。因此，很多学者聚焦于定量分析词汇与地域的关系。例如，Eisenstein 等（2010）发现，同样的话题在不同地域会以不同的方式提出和讨论，为了探索词汇与使用者所处地域的关系，他们建立级联模型来分析词汇变化是如何受到话题和地域的双重影响的。他们还把地理空间按照语言学意义的群体进行划分，能够成功地通过所用词汇来预测用户的地域信息，验证了两者的关联关系。

词汇语义变化是语言演化的典型现象。许多研究者通过观测与指定词汇共同出现的其他词汇的频度变化来探索词汇语义变化，考察社会学现象与规律（Bamman & Crane 2011；Wijaya & Yeniterzi 2011；Mihalcea & Nastase 2012）。也有研究者将指定词分解为若干词义，通过观测这些词义的使用频度变化来探索词汇语义变化。例如，“苹果”有两个典型词义，分别是“水果”和“苹果公司”，1990 年“水果”词义所占的比例较高，而进入 2000 年后“苹果公司”词义开始占据更高比例，这反映了“苹果”这个词从传统的指称“水果”到指称“IT 公司”的语义迁移。由于第二种方式更符合人们对于语言演化的直观认知，本文选择该方式开展词汇语义变化研究。

三、分布式词汇和词义表示学习模型

传统的自然语言处理和信息检索一般采用与词表规模相同的向量表示词汇，每个词对应的向量中只有一个位置值非零，因此被称为独热表示（one-hot representation）。为了区分不同的词语，词与词的非零位置均不同。这种表示方案简单有效，但是忽略了词汇之间固有的语义相关信息，而且在处理大规模文本时面临严重的数据稀疏问题。

为了解决独热表示的缺陷，随着深度学习技术的兴起，有学者提出分布式词汇表示模型，将词汇语义信息表示为稠密、实值的低维向量，词义越相近的词语，它们向量的余弦距离越近。Bengio 等（2003）提出基于人工神经网络的语言模型，利用文本中前 N-1 个词的向量预测第 N 个词的向量③，是较早的分布式词表示的成功尝试。

后来，Mikolov 等（2013）提出 CBOW 和 Skip-Gram 两个简单高效的分布式词汇表示学习模型，并推出 word2vec 工具，引起学术界与产业界的广泛关注。以 Skip-Gram 模型（简写作 SG）为例，该模型旨在用文本序列 $\{w_1, w_2, \ldots, w_T\}$ 中每个词的向量（w_t）预测该词上下文词的向量（w_{t+j}），通过最大化全局预测概率来学习词向量：

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log P(w_{t+j} | w_t)$$

由于 word2vec 中的词汇表示学习模型取消了非线性操作，并且不考虑文本中的词语顺序，因此极大地提升了训练效率。

虽然 SG 等模型获得广泛应用，但它们均只用一个向量表示一个词语，没有考虑一词多义现象（Navigli 2009），极大地限制了应用空间。因此有学者提出各种词义表示学习模型（Reisinger & Mooney 2010；Tian *et al*. 2014；Chen *et al*. 2014），为每个词的不同词义建立不同的表示向量。这些模型可以根据词语出现的具体上下文选择某个特定的词义，并用该词义的向量与上下文词向量建立预测关系。我们将基于这类词义表示学习模型，在历时文本数据集合上学习时序敏感的词义表示。

在历时文本数据集合中，我们利用整体数据集学习每个词的不同词义表示，然后将数据集按照时间划分为不同的片段（如按年份划分）。在不同片段上，我们可以统计一个词语所有出现位置的词义分配情况，从而得到该词语在该片段上的词义分布概率。针对两个不同的时间片段，我们可以根据词义分布概率的变化情况，来观测和分析这个词语的词义变化。假设：时刻在前，时刻在后，那么词义会出现以下几种情形：

1. 词义产生：在 j 时刻该词义出现的概率高于某阈值，而 i 时刻该词义概率低于该阈值，我们认为这种情况说明该词义产生。

2. 词义消亡：在 j 时刻该词义出现的概率低于某阈值，而 i 时刻该词义概率高于该阈值，我们认为这种情况说明该词义消亡。

3. 词义分裂：在 i 时刻有一个词义 s_i^k，在 j 时刻有两个词义 $s_j^{k_1}$ 和 $s_j^{k_2}$，如果 $(s_i^k, s_j^{k_1})$ 和 $(s_i^k, s_j^{k_2})$ 的相似度大于某阈值，而且 $(s_j^{k_1}, s_j^{k_2})$ 的相似度低于某阈值，我们称在 j 时刻，词义发生了词义分裂现象，s_i^k 产生了两个新的词义 $s_j^{k_1}$ 和 $s_j^{k_2}$。

四、在《人民日报》（1950—2003）上的定量观测与分析

为了验证分布式词汇表示学习模型的有效性，我们选用在 1949 年后持续出版的、与中国社会变迁息息相关的《人民日报》文本作为训练数据。我们收集了 1950 年至 2003 年的所有《人民日报》文本建立数据集，进行词义表示学习模型的训练，并进行词汇语义变化的定量观测，探究其反映出的社会变迁。

（一）词义表示学习模型的参数影响

模型的主要参数是 α，控制词义学习过程使用局部时间内的文本信息还是全局信息，α=0.0 表示只使用局部信息，α=0.8 表示同时使用局部信息和全局信息，α=1.0 表示只使用全局信息。

如图 2 所示，我们以词语“红色”为例，考察不同参数设置（从左到右依次是 α=0.0，α=0.8，α=1.0）对词义学习的影响。这里每个词义用 #0 到 #4 表示。

其中每个词义在语义空间中对应的最近邻词语可用表 1 显示：

“红色”通常有两个词义：（1）红的颜色；（2）象征革命或政治觉悟高。通过该案例，我们还可以看到，该模型能够有效地识别词语的不同词义，甚至可以得到更细粒度的、具有鲜明时代特色的特殊用法。通过观察我们可以得到以下结论：过于依赖局部信息，会出现难以分辨的词义以及重复的词义；而过于依赖全局信息，则有可能无法准确学习某个时刻特有的词义，无法获得局部突变的词义变化。因此，我们应充分结合两种信息，既可以降低局部噪声的影响，又能够避免过于依赖历史信息，

图 2　词语“红色”词义学习受到参数设置的影响情况

表 1　“红色”词义分析

参数值	词义	对应词义近邻	词义分析
α=0.0	红色 #0	革命派，无产阶级，红卫兵	“文革”相关
	红色 #1	红旗，电影，共和国	革命词义
	红色 #2	新纪录，大海，机遇	无法看出词义
	红色 #3	绿色，彩色，颜色	颜色词义
	红色 #4	绿色，特区，井冈山	革命词义
α=0.8	红色 #0	绿色，象征，颜色	颜色词义
	红色 #2	自由，共产党人，阶级斗争	革命词义
	红色 #4	革命派，工作人员，文化部	“文革”相关
α=1.0	红色 #0	绿色，颜色，色彩	颜色词义
	红色 #1	双方，社员，高等学校	无法看出词义
	红色 #2	战斗，共产主义，共产党人	革命词义

有效捕捉在短时间内突变的词义变化。

我们还将文本数据按照每 5 年划分为一个片段，用来训练词义表示向量。在四种不同参数设置下统计语义产生的样例，对词义产生的准确率进行人工标注，结果如表 2 所示：

表 2　“词义”产生准确率

参数值	α=0.0	α=0.5	α=0.8	α=1.0
词义产生准确率	48.0%	60.2%	58.1%	44.9%

从表 2 可以看到，α=0.5 和 α=0.8 时的准确率明显高出其他设置，这进一步说明综合考虑全局信息和局部信息的效果更好。

我们还定量考察了新词义出现的数量变化。以 α=0.8 为例，对于不同的起始时间（第 1 列表示不同起始时间）和终止时间（第 1 行表示不同终止时间），新词义出现的数量变化如表 3 所示。

我们得到以下结论：（1）对于两段时间，如果起始时间相同，那么时间跨度越大，会产生越多新词义。（2）对于终止时间相同的两段时间（T_{i_1}, T_j），（T_{i_2}, T_j），其中 $i_1<i_2$，一般来讲前者出现的新词义数目较多。

（二）基于词义向量的词汇语义变化观测与社会变迁分析

接下来，我们利用学习得到的词义向量，进行一些案例的定量观测与分析。值得注意的是，以下所谓的“词义”是由算法从文本数据中自动学习发现的，并不严格对应这些词语的语言学意义上的

表 3　新词义出现的数量变化

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
T_0	78	86	110	137	220	265	290	328	375
T_1	0	38	65	92	173	215	241	278	326
T_2	0	0	55	92	186	230	256	294	344
T_3	0	0	0	86	181	225	253	289	339
T_4	0	0	0	0	117	162	188	226	275
T_5	0	0	0	0	0	50	76	116	164
T_6	0	0	0	0	0	0	31	69	119
T_7	0	0	0	0	0	0	0	47	98
T_8	0	0	0	0	0	0	0	0	56

“词义”。但为了表述方便，我们仍然称其为“词义”。在这里，我们仍然用这些“词义”向量的近邻词语来反映该“词义”的内涵。

首先，国家名称是一类特殊的词语。通过观测不同国名的相邻词语变化，可以解读该国名的社会意蕴变化，将为了解中国与不同国家的政治关系提供一种新的视角。

图 3 是“美国”相邻词语变化示意图。其中词义 0 对应的最近邻词语是“英国、白宫、法国、布什、国务卿、克林顿”，词义 2 的近邻是“战争、美英联军、发动”等。可以总结出，词义 0 是与政治有关的，而词义 2 是与战争有关的。折线图中可以看到，《人民日报》提到美国时的“政治”词义比例越来越高，而“战争”词义的比例大大降低，很好地反映了中美关系从对抗到合作的历史现实。

有些特定的时间词也蕴含着丰富的意义。例如“一九五四年”，其词义分布如图 4 所示。这个词的词义 1 的近邻词语是“日内瓦、巴黎、国际法、协议、公约”。这与一九五四年签订印度支那停战协定的政治事件有密切关系，该事件在 20 世纪 60 年代到 70 年代经常提及，然后出现了下降。这表明中国当时对东南亚政局的高度重视。

通过对比一组相同类型词语的词义变化，能够更好地看到社会的显著变化。例如，图 5 显示了“农民”“工人”“知识分子”和“解放军”四类群体与“政治”有关的使用分布变化情况。

可以看到，“农民”和“工人”两个词语与“政治”有关的使用在 20 世纪 70 年代出现明显的峰值，这主要受到“文化大革命”的影响，此后则很少再被作为“政治”概念使用。“知识分子”这个词则从“文革”开始，始终被作为重要的与“政治”相关的概念使用。“解放军”在新中国建立初期被作为重要的“政治”概念，而如今更加单纯地使用“军队”的意义，而不再以“政治”概念出现在语言生活中。

还可以通过观察特定词语不同相邻词义的分布

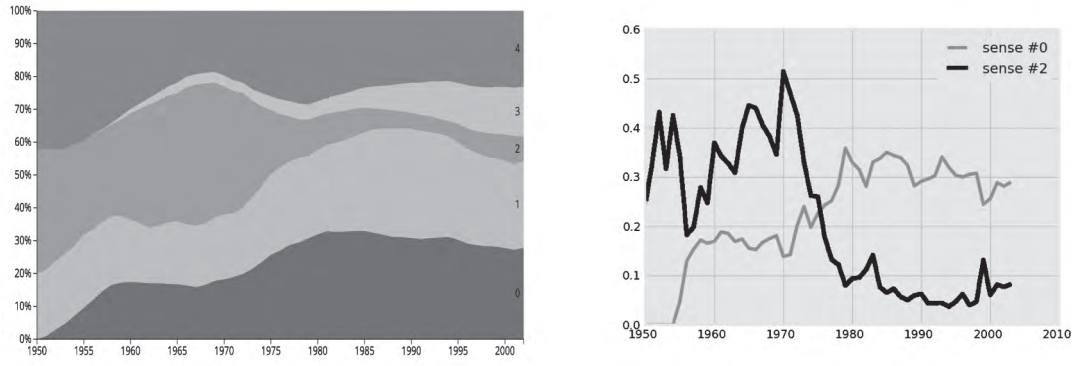

图 3 “美国”相邻词语变化

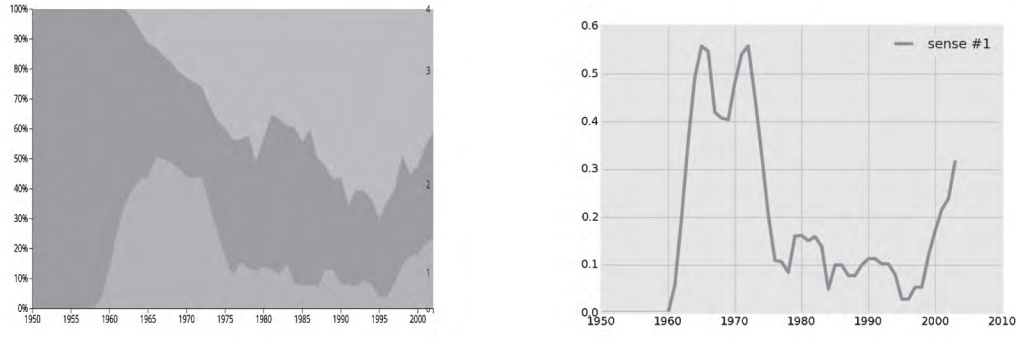

图 4 “一九五四年”相邻词语变化

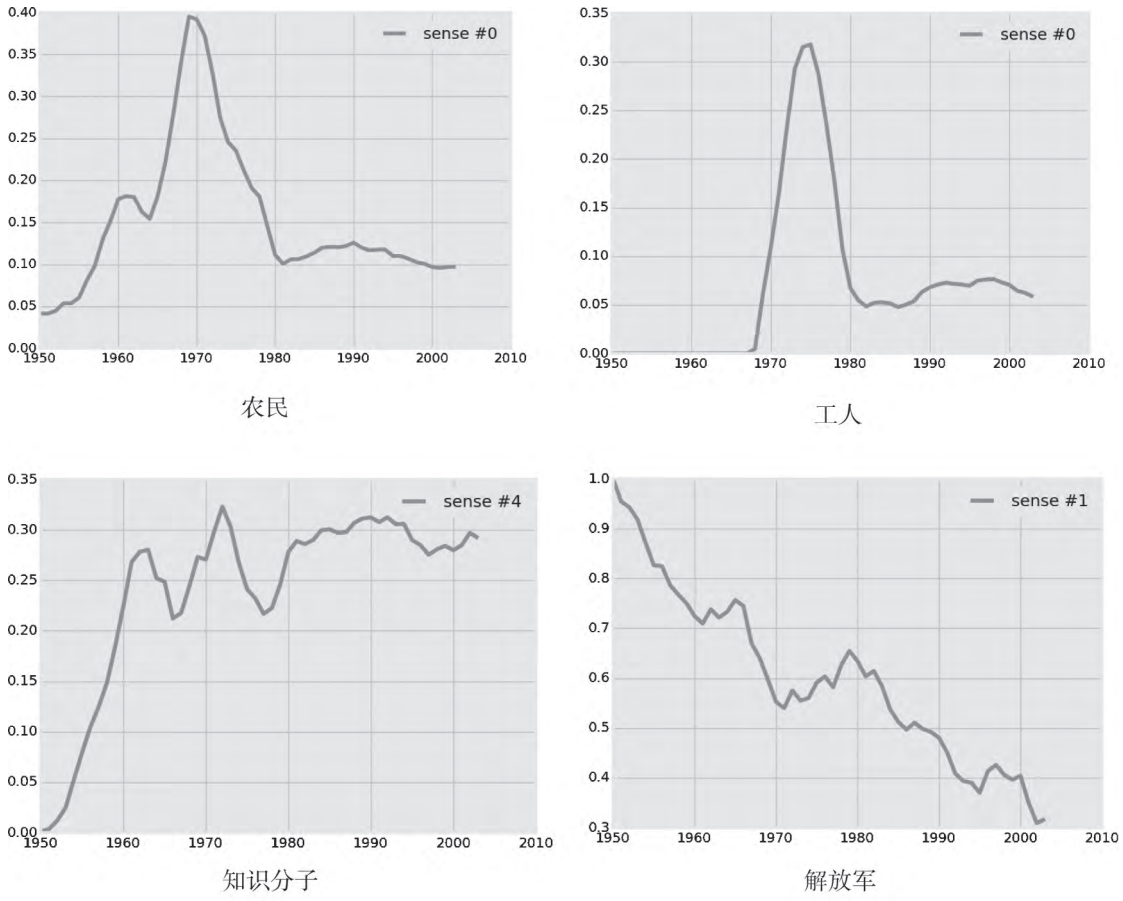

图 5　四个词语中与“政治”有关的使用变化

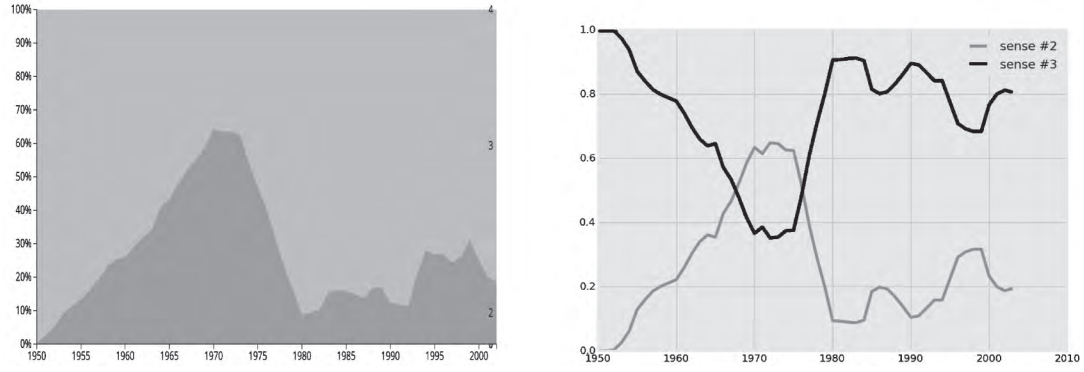

图 6　“印度支那”相邻词语变化

变化来考察不同历史事件的影响。图 6 是词语“印度支那”的使用变化情况。其中词义 2 与官媒动员亚非拉人民共同抗击帝国主义有关；而词义 3 与印度支那战争的客观描述有关。可以看到，词义 2 在 1980 年附近出现了明显的峰值，而词义 3 对应出现明显的下沉，而这正是印度支那战争爆发的时间，这说明：在战争时期，《人民日报》更加注重积极动员对帝国主义的抗争和与其他国家的团结，而在其他时期则更多是对战争的客观描述。这充分反映了《人民日报》作为官方主流媒体对战争报道的特点。

五、结　论

本文针对基于词频统计方案存在的缺点，提出

在历时文本数据集合上学习分布式词义表示模型，通过观测词汇的相邻词语分布随时间变化情况分析词汇语义变化及社会变迁，通过定量实验和案例分析，验证了该方案的有效性。该方案将为在词义级别上的词汇语义变化研究提供有效的定量分析工具，有望对语言政策制定与语言规划研究提供充分的量化依据。

本文工作还比较初步。由于《人民日报》在每个时刻的文本数量比较有限，导致该模型对出现较少词语的语义建模不够精确。未来我们将收集更丰富的历时文本数据集合，包括主流媒体、互联网网页与社会媒体等不同来源的数据，并探索更精准有效的分布式词义表示学习模型，为语言演化和语言规划研究提供重要的数据基础和有效的技术工具。

注　释

① 这是仿“基因组学（Genomics）”而成的新术语。

② 其中最高峰代表“you can put lipstick on a pig”来自网站 http://www.memetracker.org/。

③ 如果有了特定的待考察词语，找到它们直接搭配会更有效。如动词找宾语，修饰语找中心语。后文的“红色”就适合找“N+1”项，以减少很多干扰的噪声。

参考文献

陈章太　2005《当代中国的语言规划》,《语言文字应用》第 1 期。

金观涛、刘青峰　2009《观念史研究：中国现代重要政治术语的形成》，北京：法律出版社。

刘海涛　2007《语言规划的生态观——兼评〈语言规划：从实践到理论〉》,《北华大学学报》(社会科学版）第 6 期。

Aiden, Erez Lieberman and Jean-Baptiste Michel. 2013. *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books.

Bamman, David and Gregory Crane. 2011. Measuring Historical Word Sense Variation. *Proceedings of JCDL*, 1-10.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research* 3, 1137-1155.

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. 2014. A Unified Model for Word Sense Representation and Disambiguation. *Proceedings of EMNLP*.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. *Proceedings of EMNLP*.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. Meme-Tracking and the Dynamics of the News Cycle. *Proceedings of KDD*.

Lieberman, Erez, Jean-Baptiste Michel1, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the Evolutionary Dynamics of Language. *Nature* 449(7163), 713-716.

Mihalcea, Rada, and Vivi Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change Over Time. *Proceedings of ACL*, 259-263.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014), 176-182.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of NIPS*, 3111-3119.

Navigli, Roberto. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)* 41.2, 10.

Tian, Fei, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A Probabilistic Model for Learning Multi-Prototype Word Embeddings. *Proceedings of COLING*.

Traugott, Elizabeth Closs, and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Reisinger, Joseph, and Raymond J. Mooney. 2010. Multi-Prototype Vector-Space Models of Word Meaning. *Proceedings of HLT-NAACL*.

Wijaya, Derry Tanti, and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words Over Centuries. *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, 35-40.

Yang, Jaewon, and Jure Leskovec. 2011. Patterns of Temporal Variation in Online Media. *Proceedings of WSDM*.

责任编辑：金艳艳