参赛队员姓名：吴可越

中学：上海平和双语学校

省份：上海

国家/地区：中国

指导教师姓名：严骏驰（教授）

指导教师单位：上海交通大学计算机系

论文题目：词中词——英文语言中单词与子词重要性分析

词中词——英文语言中单词与子词重要性分析

Word in word：Who tells more?

吴可越　上海平和双语学校

摘要：英文单词往往由词根（root）、词干（stem）和词缀（affix）等元素构成，因此表现出子词（sub-word）组合（compositionality）特点，如 congressman 包含 congress（指代职业）和 man（指代性别）两个子词、thermodynamics（热力学）包含 therm（热）、o（连接词）以及 dynamics（动力学）三个子词。本文应用人工智能等方法及相关开源工具来分析单词及其包含子词在表达上下文语义时各自重要程度以及重要程度历史演化，探讨语言在文化发展中的变迁轨迹。为此，本文从美国历史英语语料库（The Corpus of Contemporary American English，COHA）所收集的 1810 年至 2010 年 200 年间文本语料中选取包含体现性别色彩、刻画个体意识和群体意识以及自然科学研究等单词，基于词向量（word2vec）方法进行改进，以 10 年为一个周期，学习单词及其所包含子词在表达上下文内容时的不同重要性，细化词向量表征的颗粒度，进而在得到所选取单词及其包含不同子词的词向量的同时，还能够得到单词与子词的不同权重。本文计算结果表明具有性别色彩单词中分别表示职业角色和性别角色的子词权重随时间推移和社会发展发生了较大变化；相比之下，自然科学领域复合专有名词中单词和子词的权重取值接近且保持稳定。进一步的定量实验结果展示，本文所提出的方法相比于传统词向量表示方法，在英文词性标注数据集 WSJ 上取得了 1.06%的精度提升，展示了本项目方法的先进性和创新性。

关键词：组合性；子词；词向量；权重；方差

目录

一、 研究背景与意义

1.1 研究背景与概要

语言是社会约定成俗的一种符号，与社会发展相互依存。社会行为规范了语言使用准则，语言使用又从另外一个侧面反映社会思想、态度和文化。

英语是世界上词汇量最大、语源成分最复杂的语言，其属于日耳曼语系，由法语、拉丁语、希腊语等相互影响而形成。语言史学家一般把英语的历史分为古英语（Old English，公元 449-1150 年）、中古英语（Middle English，公元 1150-1500 年）和现代英语（Modern English，公元 1500 年至今）三个历史阶段。古英语的词汇有着浓厚的日耳曼语族的特点，主要表现为以复合法为主的构词方法，导致复合词在古英语词汇中占有显著地位[Dai,2019]。有些复合词中不重读的部分，渐渐失去独立地位，演变为词缀（Affix），如 for-，in-等前缀，以及-dom，-hood，-ship，-ness，-the，-ful 等后缀。

虽然英语与其他文化在交融过程中产生了较为复杂的单词构词模式，但是英文单词构成具有一定基本规律：每个单词包含词根（root）、词干(stem)或词缀(affix)等元素。词根是一个词的核心部分，表示去除单词所有其他附加成分以后留下来的、无法再进一步删减的成分。如“unfriendliness”这一单词，去掉“un-”、“-li（由‘-ly’变化而来）”以及“-ness”后，剩下没法再分解的 “friend”就是词根。词干表示去掉曲折词缀（inflectional affix）所剩下的部分，如 friend、friendship 和 unfriendly 都是自己的词干，因为这些单词中都没有屈折词缀（如 s、ing、ed、er 和 est 等）。

为了方便理解，本文**将单词中有意义的字符组合称为子词（sub-word）**，如 autobiography（自传）这一单词由 auto-（前缀，指自己的）、bio（词根，指生命）和-graph（词根，指写）等三个子词构成。一个有趣的问题是，是否可以使用数据驱动的机器学习方法来计算一个词及其包含的子词的权重，从而知道该词及其包含的子词分别扮演着什么角色。

语言是人类生活中重要的沟通工具，语言的产生是为了服务于人们的生活，且会随着社会的不断发展而产生变化。只有语言不断地发生变化，才能跟上社会的发展脚步，才能满足人们在交际中沟通的需求。因此语言与社会之间是相互依存，相互蕴含的。

在此般情景下，若给定一个包含了若干子词的词语，该单词所包含子词权重可刻画单词的语义组合性，即单词在上下文语境中所传递语义在多大程度上由其所包含子词的语义来决定，这一研究可揭示语言进化过程。本文将单词和单词所包含子词所起作用放在一段较长时间周期内进行考查，试图从语言学角度对社会文化发展中若干侧面进行理解。

基于上述动机，本文选取了三种代表性单词来探讨单词及其所包含子词在不同历史阶段过程中表达上下文内容过程中所起不同作用，三类单词分别是：（1）带有性别色彩的词汇；（2）刻画个体、群体意识的单词；（3）以及用于自然科学研究（化学、物理和生命科学）的专有词汇。

为了实现相关的定量研究，本项目结合现有相关语言学文献，在计算机学科背景的教师指导下，采用了人工智能，特别是机器学习的相关信息科学技术进行大范围（COHA 语料库包含 200 年的数据并全采样）、全覆盖（研究了性别色彩、科学技术和个体/群体意识三个角度）、细粒度（以子词为原子单位，通过权重分析）的文本智能分析，利用了既有的开源工具、大数据计算平台等进行了设想的验证和分析，给出了相关的发现。**本文亮点总结如下：**

1) 问题新颖性和意义：本项目是利用人工智能技术来研究性别色彩单词在文献

中影响力变迁的少数工作乃至最早工作之一，如以计算量化方法挖掘了具有性别色彩单词 congressman 中性别角色子词（congress）和社会分工角色子词（man）重要程度的历史变迁。

2) 技术先进性：本项目提出学习单词及其所包含子词的向量和权重的思路，细化词向量表征的颗粒度，在词性标注任务上验证了这一思路的有效性，并使得研究者可以根据此权重进行进一步的研究。

3) 实验结果完善性和发现：本文在性别色彩、社会科学、个体与群体意识演变领域广泛选择了具有代表性的单词，研究其权重在时间维度的变化趋势，并结合相应时间段发生的标志性事件或时代趋势相互印证，发掘出语言在不同时段的演进规律。

1.2 相关研究工作

人工智能，特别是机器学习等相关信息科学技术的发展，不仅为语言学的研究提供了新的技术手段和方法，还引入了新的研究对象和领域。使用词向量(word2vec)模型，研究者可以对大规模语料库进行统计学分析，在自然语言理解领域的多个任务上获得更优异的表现。而通过对词向量的分析，研究者可以发现语言中的模式和规律，深入了解语言的结构和规则。

词向量(word2vec)模型

在基于规则和统计的自然语言传统方法中，通常将单词视为独立符号。为了表达文本，“词袋”（Bag of Words）模型被采用，但是这一模型忽略了文本单词之间的依赖关系，仅仅将文本看作是单词的集合（忽略了单词之间的先后次序）。

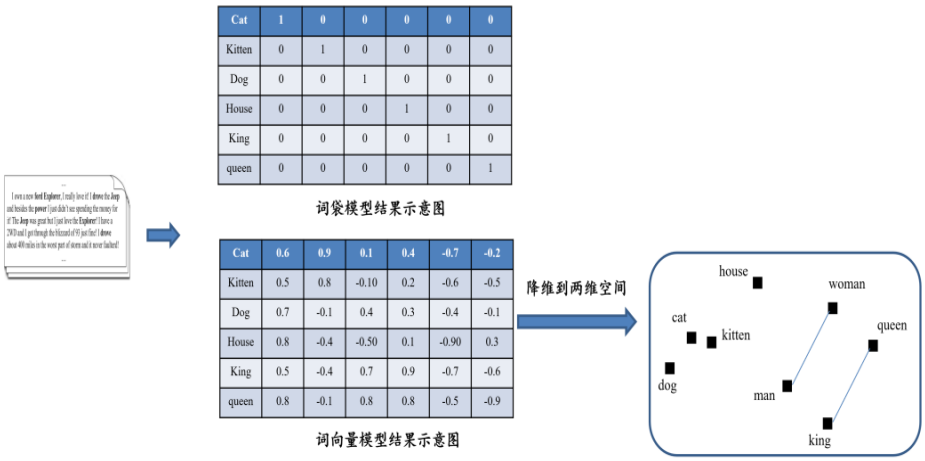


图 1 两种表达单词的不同方法（词袋模型与词向量模型）

在词袋模型中，一个单词按照词典序被表示为一个词典维数大小的向量（被称为 one-hot vector）。在这个向量中，只有一个维度上取值为 1、其他向量取值均为 0。具体而言，每个单词取值为 1 的维度位置是其在词典中排序位置。然而，

随着文本数目增加，单词词典大小也会增加，使得词袋模型中每个单词维数不断增加。因此，使用词袋模型来表示单词时，往往会遭遇维度灾难的问题。除此之外，这种表达方法无法有效计算单词与单词之间的语义相似度。比如，“contentment”与“satisfaction”两个英文单词的向量表达会很不相同，虽然其具有很强的语义相似性。

为了刻画不同单词之间的语义相关性或语义差异性，研究人员希望使用更有效方法对不同单词进行表达。利用深度神经网络模型，可将每个单词表征为N维大小的实数向量。这样可把对文本内容分析简化为N维向量空间中的向量运算，如两个单词在向量空间中之间的夹角（更精确的是余弦距离）可用来衡量这两个单词之间的相似度。用深度神经网络生成每个单词的向量表达，词向量（Word2Vec）是经典的模型[Bengio, 2003]。

在图 1 中，一旦通过词向量模型对单词进行表达，在将单词降维到两维空间时，词义相近的单词在两维空间中距离接近、而词义相差较大的单词在两维空间中距离较远。同时可见，woman 和 man 之间的连线平行于 queen 和 king 之间的连线，这一现象被称为“线性单词类比”(Linear Word Analogies)，即“a 之于 b，相当于x之于 y”，也就是说a和b经过相同的变换后会分别得到x和 y，反之亦然[Kawin, 2019]。于是下述公式计算是成立的：

$$(\overrightarrow{king} - \overrightarrow{man}) + \overrightarrow{woman} \approx \overrightarrow{queen}$$

英文单词与社会文化和科学技术相互影响

社会的经济、文化、政治、娱乐等方面的快速发展使得语言随之发生变化。同时，作为人类最为主要的交流和思维工具，语言是折射社会现象的镜子，它见证着时代的发展和变革。**一个重要的事实是语言会随着社会发展而发生变化。**

现代语言学之父、瑞士语言学家弗迪南·德·索绪尔（Ferdinand de Saussure）认为语言分为两个方面：一方面是通常研究的语言系统形成或语言使用规律，如语法、句法和词法等；另一方面便是言语（Parole），即社会语言学，表示语言在当代的使用偏好，这是与当代社会相关联的研究方向。

与索绪尔观点相一致，建立于 20 世纪 60 年代的社会语言学学科运用语言学和社会学等学科的理论和方法，从不同的社会科学的角度去研究语言的社会本质和差异，其最基本的出发点就在于把语言看成是一种社会现象，主张把语言放到其得以产生和运用的人类社会的广大背景中去研究和考察。

中国社会语言学的主要奠基人之一、为中国社会语言学学科建立了理论框架的陈原先生认为社会语言学不仅应从变动的社会发展中来发现语言的变异，也要从语言的变异中去研究社会发展的变动以及未来的图景[Chen, 2004]，指出社会发展与语言的变异之间有着密不可分的联系，语言的研究因此成为洞悉社会变化的科学视角。

因此，本文选取了若干体现性别色彩的单词及其所包含子词权重变化来研究女权运动对英文词汇的影响。众所周知，世界妇女运动可划分为三个时期：18 世纪妇女运动的兴起到二战结束为第一期，它主要是争取妇女外在的权益；

第二期是从战后本世纪 70 年代初，主要关注性别内在的价值与权益；从 70 年代至今的第三波女权运动则关心第一、二波运动未关注到的宗教、种族、多元文化等问题。而从 congressman、chairman、spokesman 和 policeman 等到 congressperson、chairperson、spokespersons 和 police officer 等单词与子词权重变化可以分析社会文化发展中性别意识的演变。

同时，本文也通过研究刻画个体意识和群体意识的代表性单词来分析社会发展过程中个体意识和群体意识的此消彼长，如研究 egocentric（个人为中心，字根子词 ego 和 centric）和 egomaniac（字根子词分别是 ego 和 maniac）、individualism（个体主义，字根子词 indiv）等单词和其所包含子词权重变化来揭示群体意识和个体意识的演变。

建立量子论的物理学家普朗克曾说“科学是内在的整体，被分解为单独的部门不是取决于事物的本质，而是取决于人类认识能力的局限性。实际上存在着由物理学到化学、通过生物学和人类学到社会科学的链条，这是一个任何一处都不能被打断的链条”。因此，本文选取了来自物理、化学和生命等领域，体现学科交叉的复合专有名词，从单词和子词权重变化来探讨学科交叉。

英文单词组合现象以及词汇结构

英语是一种抽象的表音文字，其词形变化可体现各种语义和语法现象。对英语词根深有研究的语源学家约·肯尼迪(John Kennedy)指出：词干及其所包含的意义是整个英语的基础。剑桥大学出版社 2001 年出版的《英语单词：历史与结构》一书对英语单词的起源进行了分析 [Robert, 2001]。针对“词是从哪儿来的，词的起源在哪里”这一问题，该书分析和描述了 10 种不同的新词产生方式：继承（inheritance）、创造或使用新词（neologism）、混成（blending）、首字母缩写（acronym）、缩写（shortening）、派生（derivation）、转换（conversion）、组合（compounding）、人名法（eponyms）和拟声法（onomatopoeia）。

如前所述，单词和其所包含子词在表达上下文语义中会发挥不同作用。[Xu, 2019]分析了英语及其他印欧语言中前缀和后缀等子词与单词之间各自权重，以便分析语义组合性。该文发现现代中文的子词语义权重较低，而古代中文则较高；英语等 5 门印欧语言则呈现相反趋势。

本文将代表性单词中有意义字符组合而形成的子词提取出来，分析子词构词刻画出的社会发展。

二、 算法描述

本文从美国历史英语语料库（COHA）所收集 1810 至 2010 年文本语料中选取了包含体现性别色彩、刻画个体意识和群体意识以及自然科学研究（化学、物理和生命科学）等单词，构造了进行单词及其所包含子词权重计算的实验语料库。

为了计算单词及其所包含子词在上下文文本所起语义重要程度，本文应用文本机器学习模型得到单词和子词的词向量（word2vec），以 10 年为一个周期计算代表性单词库中单词和其包含的子词在上下文语境中的权重大小，分析计算结果背后刻画的社会发展和自然科学的演变。

下面对本文中主要算法进行介绍（见图 2）。

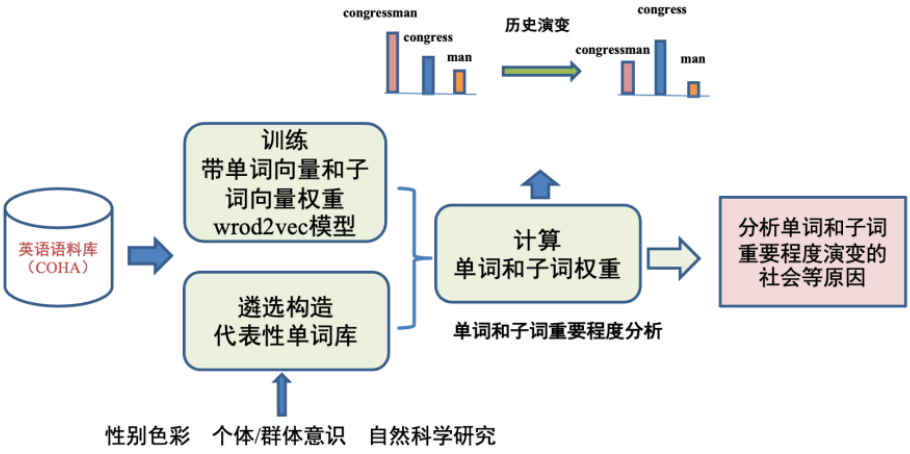
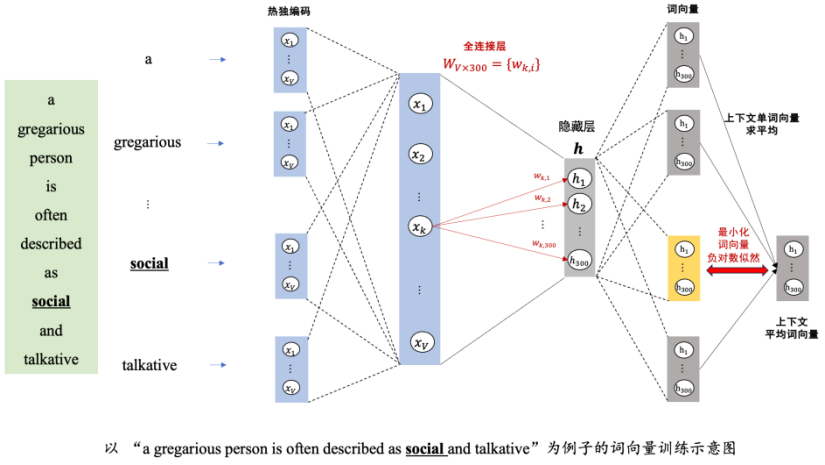


图 2 本文算法主要流程

2.1 单词词向量的预训练



以 “a gregarious person is often described as social and talkative” 为例子的词向量训练示意图

图 3 以 “a gregarious person is often described as social and talkative”为例子的词向量训练示意图

在得到子词的词向量及其权重大小前，为了保证训练的稳定性，我们首先不考虑子词的影响，使用 Word2Vec 预训练得到单词词向量。

通过机器学习算法 Word2Vec 来生成单词词向量表达有两种模型训练模式：一种是连续词袋模型（Continuous Bag-of-Words，CBoW），即根据某个单词所处的上下文单词来预测该单词，另一种则是 Skip-gram，即利用某个单词来分别预测该单词的上下文单词[Tomas, 2013]。本文采用 CBoW 来计算每个英文单词的实数值词向量。

图 3 给出了以“a gregarious person is often described as social and talkative”为例子的词向量训练示意图。这个例子中，用 social 这个单词所在句子中上下文单词（如 gregarious、person 以及 talkative 等）来预测 social 单词，使 social 这一单词

相对于其他单词而言，得到的预测概率最大。在 Word2Vec 的实现中，使用了负对数似然作为损失函数对模型参数进行更新。

假设训练语料库中包含 T个句子和V个单词，CBoW 的学习目标是使得每个单词w_i在其上下文单词出现时，单词w_i出现的概率最大，即使得如下对数函数取值最大：$L_{CBoW} = \sum_{i=1}^{T} \log(p(w_i|C_i))$，这里$C_i$代表$w_i$的上下文单词。

在机器学习中，上述概率取值需要被规整到[0,1]取值空间，因此实际上使用如下 softmax 函数来实现：

$$p(w_i|C_i) = \frac{\exp(w_i^T \cdot v_C)}{\sum_{j \in V} \exp(w_j^T \cdot v_C)}，\text{其中} v_C = \frac{1}{C_i} \sum_{v_k \in C_i} v_k$$

在上面的公式中，C_i表示单词w_i在某个句子中上下文单词的个数，v_C表示单词w_i在句子中所有上下文单词对应向量的平均值，v_k表示单词w_i在句子中第k个上下文单词，w_j表示V个单词中第j个单词。很显然，$p(w_i|C_i)$取值范围在 0 到 1 区间，并且希望当单词w_i的上下文单词出现时，相对于其他单词，单词w_i出现的概率应该最大。

这里出现的计算符号“·”表示内积，指两个向量所对应每个维度相乘，再对乘积求和。比如$[1.3\ \ 5\ \ 2.5] \cdot [2\ \ 4.1\ \ 0.6] = 24.6$。从几何角度而言，两个向量做内积计算与这两个向量之间夹角的余弦值相关，因为空间中向量p和q之间的夹角θ的余弦值为：$\cos\theta = \frac{p \cdot q}{|p| \times |q|}$，这里$||$表示向量的模（即向量的长度），三维空间中一个向量$[x\ y\ z\ ]$的模为$\sqrt{x^2 + y^2 + z^2}$。

要说明的是，虽然$p(w_i|C_i)$取值越大越好。但是在深度学习中，模型对应的损失函数越小越好，因此一般会对 $p(w_i|C_i)$ 这一取值做变化（如取值相反数），使得训练的损失函数越小越好。

比如给定“a gregarious person is often described as social and talkative”这一句子，在计算 social 单词词向量时，先计算 social 周围伴随单词词向量的平均向量，然后希望计算所得 social 这一向量使得如下条件概率取值最大：

$$p(social|\text{平均向量}（gregarious, persons, described, talkative \ldots）)$$

2.2 子词词向量的预训练

如前所述，本文拟分析所选定单词及其包含子词在表达语义时不同权重大小，因此也需要学习每个子词的向量表示（本文称为**子词词向量**）。通过 Word2Vec 预训练得到单词词向量后，本文随即进行子词词向量的训练。

为了训练子词词向量，需要形成子词构成的词库，本文采用字节对编码（Byte Pair Encoding，BPE）来完成[Rico, 2016]。BPE 算法首先将语料库中所有单词拆分为单个字符，用所有单个字符建立最初的词典，并统计每个字符的频率，本阶段子词粒度是字符；然后挑选出现频次最高的符号对，比如 d 和 e 组成 de，将 de 这一新子词加入词表。

在上述每次合并后词表可能出现如下 3 种变化：1）加入合并后的新子词，同时原来的 2 个子词还保留（2 个子词不是完全同时连续出现）；2）加入合并后的新子词，同时原来的 2 个子词中一个保留，一个被消解（一个子词完全随

着另一个子词的出现而紧跟着出现）；3）加入合并后的新子词，同时原来的 2 个子词都被消解（2 个子词同时连续出现）。如此重复操作，直到词表中单词数达到设定量或最高频的两个字节合并后的子词最高频数为 1。

这里要说明的是，本文在进行 BPE 算法的过程中固定了一些子词，以保证后续分析的子词都出现在所构造的子词词库中。比如对于“congressman”这个单词，本文指定分析：“congress”和“man”这两个子词与单词“congressman”的权重大小，以便了解职业和性别的关系与其演化趋势。

子词词向量的学习与词向量学习类似。这一阶段的优化目标与前一阶段相同，仍然为$L_{CBoW}=\sum_{i=1}^{T}\log(p(w_i|C_i))$，其中，$C_i$代表$w_i$的上下文单词。

但是，$p(w_i|C_i)$变为

$$p(w_i|C_i)=\frac{\exp(u_i^T\cdot v_C)}{\sum_{j\in V}\exp(u_j^T\cdot v_C)},\ \text{其中}\ v_C=\frac{1}{C_i}\sum_{v_k\in C_i}v_k\,,u_i=\sum_m^k\lambda_{w_i}^m\times \mathrm{sub}w_m$$

在上面的公式中，单词w_i包含了k个子词$\mathrm{sub}w_m(1\le m\le k)$，单词$w_i$中每个子词权重为$\lambda_{w_i}^m(1\le m\le k)$，单词$w_i$的子词加权向量表示为$u_i$。在这一阶段的训练中，单词词向量$v_k$被固定，仅优化子词词向量$\mathrm{sub}w_m$和子词权重$\lambda_{w_i}^m$。

2.3 单词权重和子词权重的优化训练

一旦预训练得到每个单词的词向量及该单词所包含子词的子词词向量和权重后，可以训练单词权重来对单词词向量和子词词向量进行加权累加，得到单词的最终表达，利用这个最终表达来完成依据单词上下文来预测单词的任务，从而优化得到单词权重大小以及其所包含子词的权重大小。

假设单词w_i包含了k个子词$\mathrm{sub}w_m(1\le m\le k)$，令单词$w_i$的权重为$\lambda_{w_i}$，单词$w_i$中每个子词权重为$\lambda_{w_i}^m(1\le m\le k)$，则在考虑了单词$w_i$本身词向量以及单词$w_i$所包含子词词向量后，单词$w_i$的向量表达$vec_{w_i}$如下计算：

$$\underbrace{vec_{w_i}}_{\text{单词}w_i\text{的词向量}}=\underbrace{\lambda_{w_i}\times w_i}_{\substack{\text{考虑单词}w_i\text{本身词向量}\\(\text{权重为}\lambda_{w_i})}}+\underbrace{(\sum_m^k\lambda_{w_i}^m\times \mathrm{sub}w_m)}_{\substack{\text{考虑了单词}w_i\text{所包含的}k\text{个子词}\\(\text{每个子词的权重为}\lambda_{w_i}^m)}}$$

从上面可以看到，在每个单词w_i的向量表达vec_{w_i}过程中，既考虑了单词w_i本身词向量（权重为λ_{w_i}），又考虑了该单词所包含所有子词的词向量（权重为$\lambda_{w_i}^m$）。在这一阶段的训练中，所有参数都参与优化过程。

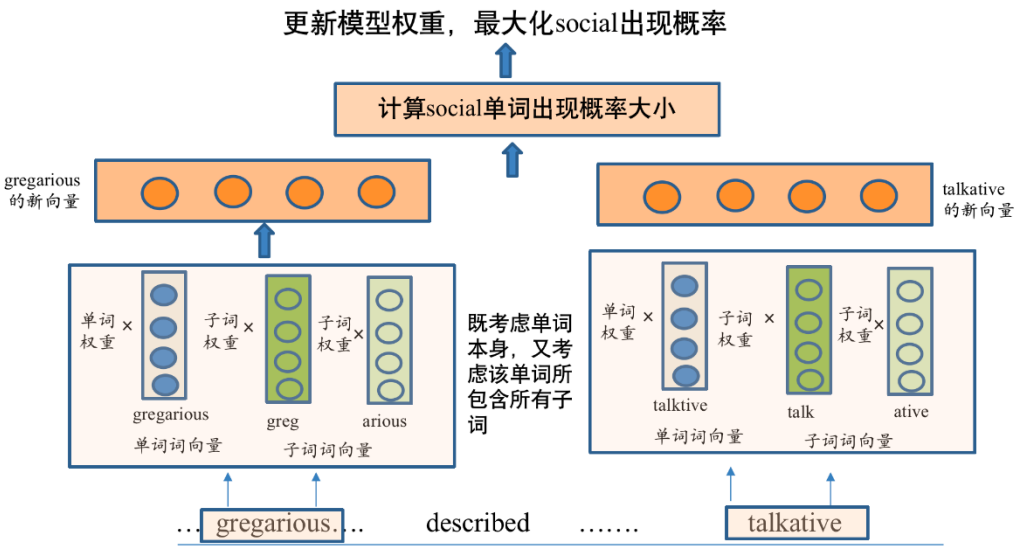


图 4 考虑了单词本身和子词构成的词向量及模型训练流程

在得到每个单词考虑了自身以及子词的新向量基础上，就可以将这个新的词向量引入上下文单词预测过程中，从而计算单词权重和子词权重（如图 4 所示）。

在图 4 中，对于给定的“a gregarious person is often described as social and talkative”句子，拟用 social 这个单词上下文单词来预测其出现概率。于是，上下文单词及其所包含子词均被考虑起来，得到每个单词新的向量，然后再预测 social 单词出现概率大小，从而进行模型训练。

该模型训练的结果可得到单词权重和子词权重的取值。

三、 实验数据

3.1 实验数据集 COHA 以及模型参数描述

美国历史英语语料库（COHA）是最大的历史英语结构化语料库。COHA 包含 1810 年至 2009 年之间 200 年的 4.75 亿单词文本（是其他同类英语历史语料库的 50-100 倍），语料库按体裁逐年平衡。该语料库的创建得益于美国国家人文基金会（NEH）2008-2010 年的资助[coha]。

本文使用 CCOHA 开源工具[Alatrash2020]对 COHA 数据集进行清洗，随后以 10 年为单位，对数据进行了划分，来观测所关注单词及其子词在不同历史阶段的演变。以 10 年为时间单位，分别训练了 20 个单词词向量和子词词向量生成的神经网络模型，来得到单词和子词分别对应的词向量。

本文实验训练的词向量模型中一共包含 341365 个不同单词和 7500 个不同子词，每个单词和子词均映射为 300 维向量。

3.2 所关注单词及其子词

本文选取了包含体现性别色彩、刻画个体意识和群体意识以及自然科学研究（化学、物理和生命科学）等 21 个单词进行分析。具体如下：

3.2.1 性别色彩变化单词及其子词（8 组）

本文选取了如表 1 中所示的八组性别色彩单词和无性别色彩单词进行如下分析：具有性别色彩单词的权重是否降低，而逐渐被性别中立的单词取代，如 congressman 被 congresspersons 取代；性别色彩单词和无性别色彩单词中表示职业性质子词权重（如 congress）和表示性别色彩子词权重（如 man）重要程度的变化。

单词	对应变化单词	职业性质的子词	其他子词
congressman	congressperson	congress	man person officer worker
chairman	chairperson	chair	
spokesman	spokesperson	spoke	
fireman	firefighter	fire	
policeman	police officer	police	
businessman	businessperson	business	
mailman	postal worker	mail, postal	
salesman	salesperson	sales	

表 1 八组性别色彩和无性别色彩单词集合

3.2.2 自然科学研究领域复合专有单词及其子词（7 个）

如前所述，自然科学研究是多学科、多领域交叉结果，出现了大量体现交叉特色复合专有名词。本文从物理、化学和生命等领域选取了如下七个复合专有名词来分析单词及其所构成子词的重要性的演变，以观测在自然科学研究中学科交叉的特色。

单词	子词
chemotherapy（化学疗法）	chemo（化学）、therapy（疗法）
biochemistry（生物化学）	bio(生物)、chem（化学）、istry（名词词尾）
astrophysics（天体物理学）	astro （天文）、physics（物理）、

thermodynamics(热力学)	therm(热)、o(连接词)、dynamics（动力学）
energize（给与能量）	ener（能量）、ize（动词词尾）
invigorate（鼓舞）	vigor（活力）、ate（使）
microbiology（微生物学）	micro（微小的）、biology（生物）

表 2 自然科学领域选取的代表性复合专有单词及子词

3.2.3 个体意识和群体意识色彩的单词及其子词（6 个）

本文选取了刻画个人意识与群体意识的六个单词，计算其单词及子词权重变化。强调个体意识的单词如下：egotist（自高自大的人，字根子词为 ego）、egocentric（个人为中心，字根子词 ego 和 centric）以及 individualism（个体主义，字根子词 indiv）。

包含群体意识的单词如下：gregarious（群居的，字根子词 greg）、collectivism（集体主义，字根子词 collect）、socialism（社会主义，字根子词 social）。

四、 实验结果分析与讨论

本文对给出对性别色彩单词、自然科学研究复合专有名词和个体/群体意识单词的实验结果和相应的分析。

4.1 单词与子词重要性对比

本文以 10 年为单位，对数据进行了划分，观测单词及其子词在不同历史阶段的演变。为了更好呈现单词与子词在表达语义过程中谁更为重要，本文把以 50 年为间隔，先计算每个单词 50 年内权重变化平均值，然后计算单词权重在 1810 年至 2009 年之间的平均值。

单词	每隔 50 年单词权重取值（1810s-2000s） 注：1 减去单词权重就是该单词包含所有子词权重总和				单词权重平均值	单词和子词权重之差的方差
congressman	-	0.503752	0.502536	0.520436	0.508908	0.062854
chairman	0.527394	0.592272	0.892358	0.959892	0.742979	
spokesman	0.501746	0.506245	0.657336	0.877632	0.635740	
fireman	0.504181	0.510641	0.510773	0.509322	0.508729	
policeman	0.502052	0.529401	0.617423	0.619242	0.567029	
businessman	-	-	0.548056	0.622892	0.585474	
mailman	-	-	0.501248	0.505290	0.503269	
salesman	0.500592	0.510531	0.592099	0.596496	0.549929	
上述具有性别色彩单词权重的均值					**0.57525712**	

chemotherapy	-	-	0.500431	0.510458	0.505444	0.000033
biochemistry	-	-	0.501048	0.502084	0.501566	
astrophysics	-	-	-	0.500538	0.500538	
thermodynamics	-	-	0.501112	0.506570	0.503841	
energize	0.500497	-	-	0.501533	0.501015	
invigorate	0.501201	0.500190	-	0.501588	0.500993	
microbiology	-	-	-	0.500275	0.500275	
上述自然科学领域复合专有名词权重的均值					**0.50195314**	
egotist	0.500305	0.500738	0.500381	0.500137	0.500390	0.000743
egocentric	-	-	0.500873	0.501900	0.501386	
individualism	0.500089	0.503623	0.533548	0.512239	0.512375	
gregarious	0.500125	0.500256	0.502052	0.504364	0.501699	
collectivism	-	0.500312	0.502888	0.500537	0.501246	
socialism	0.500421	0.512136	0.546991	0.538394	0.524486	
上述具有个体和群体意识单词权重的均值					**0.50693033**	

表 3　21 个单词的权重变化及单词权重和子词权重之间差异的方差

同时，由于单词权重和所有子词的权重累加之和为 1，因此可计算单词权重和子词权重之差的波动情况，即单词和子词权重之差的方差。

表 3 给出了 21 个单词每隔 50 年的权重取值、单词权重平均值及单词权重和子词权重之间差异的方差取值。从表 3 可得出如下结论：

1、 具有性别色彩的单词权重明显大于其子词权重。比如 chairman 和 spokesman 的单词权重平均值为 0.742979 和 0.635740，明显远远高于它们对应的子词权重（分别为 0.257021 和 0.36426），**这说明具有性别色彩单词中只有社会角色和性别角色组合在一起才会完整表达单词语义，因此整个单词权重远远大于子词权重。**

2、自然科学中专有复合名词的单词权重与其子词权重相差无几。比如 chemotherapy 和 biochemistry 的权重分别为 0.505444 和 0.501566，与其子词的权重十分接近（分别为 0.494556 和 0.498434），这说明在**体现学科交叉特点的自然科学专有复合名词中，单词和子词所起作用几乎一样，每个学科子词权重累加在一起与整个专有复合名词权重在上下文语境中作用一样。**

3、具有个体色彩和群体色彩的单词权重与其子词权重之间差异不是特别接近、也不是特别相远，我们推测个体色彩和群体色彩单词使用与每个人的使用习惯有关，因此这一类单词在个体色彩单词权重、群体色彩单词权重以及单词和子词权重之差等方面没有表现特别的性质。

4、从表 3 中给出的单词权重平均值及单词权重和子词权重之间差异的方差取值可以看出：自然科学中单词和子词之间权重差异微乎其微，**这说明自然科学专有复合名词在表达上下文内容过程中变化程度很微小。**

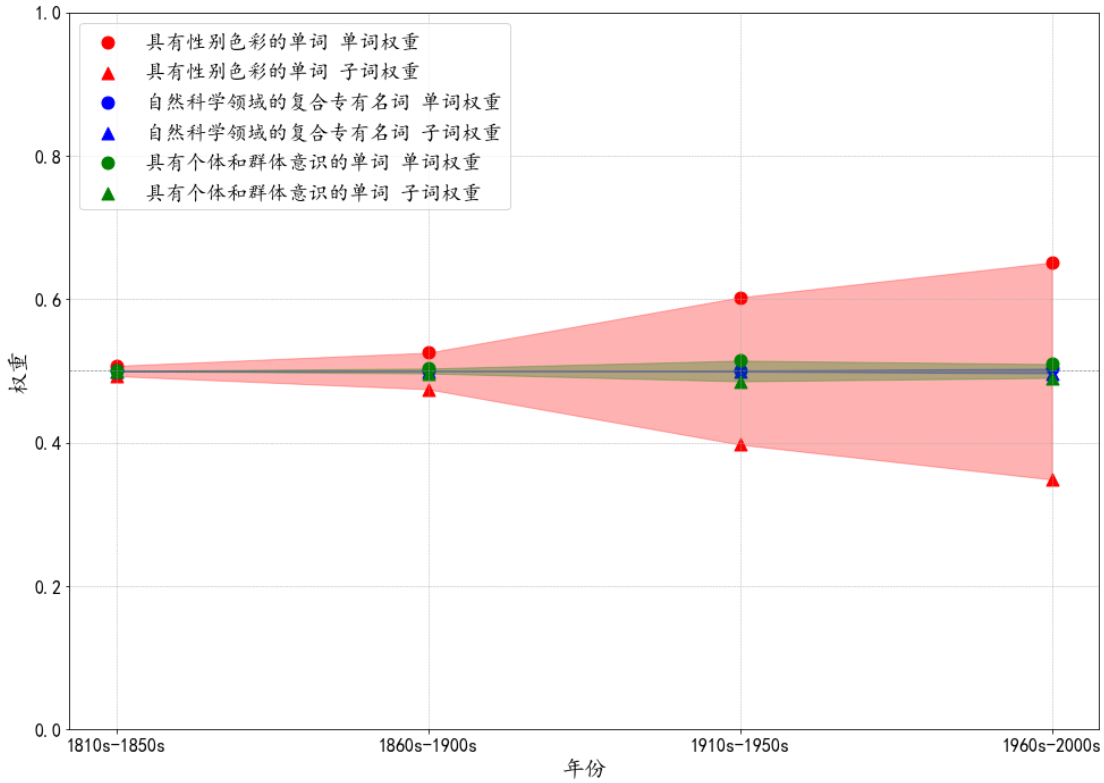


图 5 单词权重和子词权重在中值线（0.5）两侧波动情况

图 5 给出了上述单词权重和子词权重以 50 年为间隔在中值线两侧波动情况，可以看出，所有单词权重均大于 0.5、子词权重均小于 0.5。具有性别色彩单词权重与子词权重波动较为明显、自然科学复合专有名词的单词权重和子词权重接近且保持稳定。

4.2 单词中子词重要性变化分析

本文同时研究单词中子词的重要性变化，为此选取了具有性别色彩的若干单词以及自然科学若干单词，来观察其包含子词的权重变化。

表 4 给出了具有性别色彩单词在第一次出现的 10 年间以及最后一个 10 年间（即 2000-2009）所包含社会角色和性别角色的子词权重变化。令人吃惊的是，这些单词在刚刚出现的 10 年间性别角色子词权重全部大于社会角色子词权重，而在最后 10 年中，性别角色子词权重全部小于社会角色子词权重。这说明了这些单词被发明时具有浓厚的性别色彩，但随着社会发展，性别角色色彩减弱，而其代表的社会角色更为凸显。

单词	子词	首个 10 年权重	最后一个 10 年权重（2000s）
congressman	congress	0.15	0.72
	man	0.85	0.28
chairman	chair	0.50	0.52
	man	0.50	0.48
spokesman	spokes	0.37	0.89

	man	0.63	0.11
fireman	fire	0.42	0.25
	man	0.58	0.75
businessman	business	0.36	0.64
	man	0.64	0.36
mailman	mail	0.36	0.637
	man	0.64	0.363

表 4　具有性别色彩单词中所包含社会角色和性别角色的子词权重变化

分析数据表明，**在从 1970s 进入到 1980s 这一阶段，可能由于受到第二、三波女权运动的影响，大多数性别色彩单词的权重呈现较为明显的下降**，且具有性别色彩单词在缓慢被不具有性别色彩单词所替换。

单词	子词	首个 10 年权重	最后一个 10 年权重（2000s）
thermodynamics	therm(热)	0.575489342	0.03280085
	o(连接词)	0.128645107	0.209167734
	dynamics（动力学）	0.295865536	0.758031428
biochemistry	bio(生物)	0.795628548	0.17835778
	chem（化学）	0.063332722	0.389849603
	istry（名词词尾）	0.141038716	0.431792617
microbiology	micro（微小的）	0.151850283	0.484442532
	biology（生物）	0.848149717	0.515557408

表 5　自然科学中复合专有名词中子词权重变化

表 5 给出了三个自然科学复合专有名词的子词权重变化。可以看出，这些代表不同学科的子词权重变化反映了在学科交叉研究中不同研究方向此起彼伏的“波浪状态（有起有落）”。此外，在 thermodynamics 中，作为连接词的子词“o”权重在增加，说明更多单词运用“o”进行组合连接，生动刻画了不同学科方向交叉的画面。

4.3 不同词向量表达在词性标注任务上的对比

为了验证既考虑了单词权重、又考虑子词权重所形成之单词词向量可更好表达上下文语义内容，本文在英文词性标注任务（Part of Speech Tagging, POS）上进行了对比实验。

英文词性标注任务目标是判断输入句子中每个单词应该具有的词性。例如，输入句子为“it is an interesting movie.”，词性标注任务可输出“it”的词性为代词(Pronoun, PRP)、“is”的词性为系动词 (Be-verb, VBZ)、“an”的词性为不定冠词(Article, DT)、“interesting”的词性为形容词 (Adjective, JJ)、“movie”的词性为名词 (Noun, NN)等。

本文对比了如下三种单词词向量表达在词性标注上的性能：

- 传统单词词向量表达：在这一表示方法中，仅考虑每个单词自身的词向量。
- 单词和子词权重平均化表达：在这一方法中，单词的权重为 0.5，所有子

词的词向量先取平均值，然后将子词平均词向量赋予权重 0.5，再将单词和子词平均词向量在权重分别为 0.5 时进行加权累加。

- 单词和子词权重学习所得：即通过本文图 4 所示计算单词和子词权重，然后按照学习所得权重进行加权累加，得到单词之新的词向量。

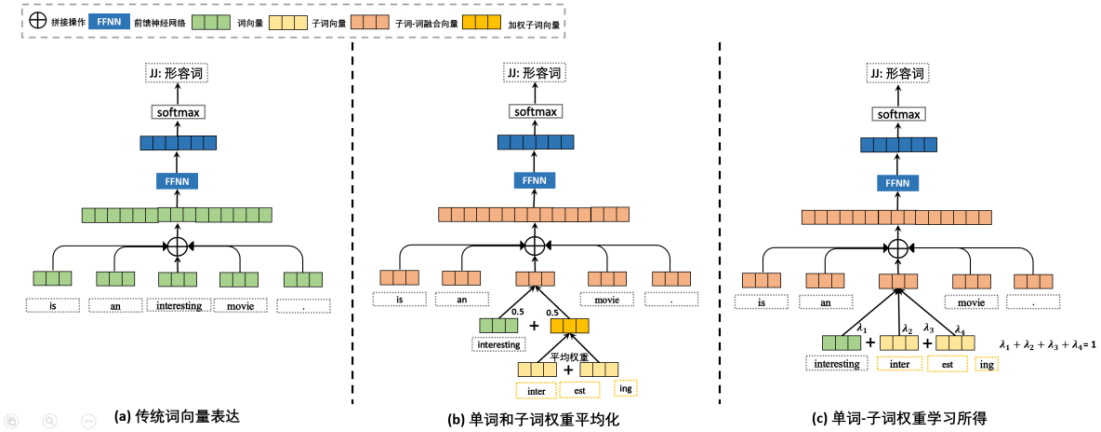


图 6 使用不同单词词向量表达进行单词词性预测

在进行单词词性预测时，将待预测单词前后两个单词的词向量与本单词词向量拼接在一起，将拼接所得向量送入一个前馈神经网络（Feed-Forward Neural Network, FFNN），softmax 函数对 FFNN 的结果进行预测，将概率取值最大的词性作为该待预测单词的词性[Collobert 2011]。

表 6 给出了三种不同词向量表示方法所得的词性预测准确率。

单词词向量表示方法	准确率(%)
仅考虑单词自身	91.24
单词和子词权重平均取值	91.45
即考虑了单词、又考虑了子词的不同权重	92.30

表 6 不同单词词向量表示方法在 WSJ 测试集上词性预测结果准确率

从表 6 可以看出，本文所提出方法在 WSJ 测试集上取得更高的词性预测准确率（92.30），这表明既考虑单词重要性、又考虑子词重要性可以提供更优越的语义表示。

上述实验结果验证了本文的假设，即同时计算单词重要性和子词重要性，不仅可以更好捕获单词内部结构，也可更好析取单词上下文语义关联。

五、 结论

本文从英文单词具有组合性这一特点出发，计算了具有性别色彩单词、自然科学专有复合名词和个体意识及群体意识单词中单词权重和子词权重变化，本文的分析和实验表明：1）所有单词权重大于其包含子词权重总和，这体现了单词之所以为单词的特点，即整体大于部分；2）性别色彩单词所包含的社会角色子词和性别角色子词之间权重因社会发展而发生“互换”，自然科学专有复合名词中单词权重相对于所有子词权重之和而言的重要性几乎一样，反映了学科交叉的特点；3）对于给定的一个单词，在上下文中既考虑了单词重要性、又考虑了其所

包含子词的重要性的表示方法，在自然语言分析任务中可获得更好效果。

应该说本文的研究只是对 COHA 语料库所蕴含丰富内容的一个侧面研究，研究结果还会受到数据完整性以及噪音和其他因素等影响，接下来本文将开展如下后续研究：1）性别色彩单词（如 chairman 等）与无性别色彩单词之间的互动影响；2）是否存在影响单词权重和子词权重发生变化的代表性上下文单词；3）本算法在其他数据集上的性能表现与 COHA 数据集性能表现的对比。

致谢与后记

本文研究的得到了上海交通大学计算机系严骏驰教授及其本科生刘祺和陈梓俊同学在计算平台使用方面的支持和指导。

关于本项目的最初动机，是申请人在备考托福和 SAT 过程中对英文单词构词方法产生了浓厚兴趣。在与具有信息技术专业背景的指导老师多次讨论后，申请人凝练出单词和子词重要程度计算和分析这一问题。同时，也对影响单词和子词变化的社会发展原因产生了兴趣。在进一步文献调研中，我们注意到一篇很有参考意义的论文“Treat the Word As a Whole or Look Inside? Subword Embeddings Model Language Change and Typology（见参考文献[Xu, 2019] ）”，作为一篇人工智能领域的技术性论文，该工作提出了考虑子词影响下单词词向量的编码方法。本文在这篇文章算法基础上，提出了分析性别色彩单词、自然科学复合专有名称以及个体/群体意识单词中单词和子词相互影响的研究问题。

在完成上述工作的过程中，申请人采用了指导老师高校实验室的存量算力资源、大数据平台、人工智能 API 工具集以及相关开源模型，结合自身的编程基础，进行了二次开发工作。在编程和实验的过程中，也得到了严骏驰老师和其课题组两位本科生同学（刘祺和陈梓俊）的无偿悉心指导。具体而言：严骏驰老师对本人所提出思路进行了算法理论的指导；刘祺学长在 COHA 语料库的整理、BPE 算法的实现、机器学习模型和 Jittor 代码的理解上为本人答疑解惑，陈梓俊学长在词向量模型的训练和使用、实验数据的整理、代码和开源框架的调试与修改等方面提供了指导。严骏驰老师和两位上海交大学长对本人所得到实验结果的分析和呈现进行了讨论和指导，同时，他们不耐其烦教我有关计算语言学知识（如词向量和字节对编码等），辅导我实现了相关代码，激励我走入了人工智能与语言学结合的世界，才让我体验了计算语言学的魅力，在此表示深深谢意。

更多单词和子词权重计算结果及其演变分析已经公布在本人个人主页，如有兴趣，可访问下述地址查阅：https://itsallforyou28.wixsite.com/whimsicalme

本文局限与展望：受限于高中生的知识、技能水平及课外时间，当前的工作主要基于既有的开源模型，进行了二次开发和定制，更好地服务本文所研究的实际问题。后续申请人将继续利用课余时间使用本方法分析更多语言的发展，如汉语、法语、德语等其他具有大规模语料的语言；申请人也将使用其他数据集进行分析，以确定文中发现的语言趋势是否是在各来源语料中广泛存在的现象；申请人还计划深入学习前沿的自然语言模型，如 BERT, GPT 等大型语言模型，尝试使用这些规模更大、效果更佳的模型进行语言学研究。

参考文献

1. [Dai,2019] 戴炜华,《语言接触漫语》，上海理工大学学报（社会科学版），ISSN 1009-895X
2. [Bengio,2003]Yoshua Bengio, Rejean Ducharme, Pascal Vincent, Christian Jauvin, A neural probabilistic language model, Journal of Machine Learning Research, 2003, 3:1137-1155
3. [Kawin,2019]Kawin Ethayarajh, David Duvenaud, Graeme Hirst, Towards Understanding Linear Word Analogies, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019
4. [Xu, 2019]Yang Xu, Jiasheng Zhang, David Reitter,Treat the Word As a Whole or Look Inside? Subword Embeddings Model Language Change and Typology, ACL 2019: 136-145
5. [Robert 2001]Robert Stockwell and Donka Minkov,English Words: History and Structure,Cambridge University Press, ISBN 0-521-79362-9,2001
6. [Li, 2020]李平武，英语词根与单词的说文解字（新版），外语教学与研究出版社，2020
7. [Chen,2004]陈原，社会语言学，商务印书馆，2004
8. [Tomas,2013]Tomas Mikolov and Kai Chen,Gregory S. Corrado,Jeffrey Dean,Efficient Estimation of Word Representations in Vector Space,International Conference on Learning Representations,2013
9. [Rico, 2016]Rico Sennrich, Barry Haddow, and Alexandra Birch,Neural Machine Translation of Rare Words with Subword Units, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, German
10. [coha] https://www.english-corpora.org/coha/
11. [Alatrash, 2020] Alatrash R, Schlechtweg D, Kuhn J, et al. Ccoha: Clean corpus of historical american english, Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 6958-6966.
12. [Collobert, 2011]Ronan Collobert,Jason Weston,Léon Bottou,Michael Karlen,Koray Kavukcuoglu, Pavel Kuksa，Natural Language Processing (Almost) from Scratch，The Journal of Machine Learning Research，Volume 12，pp 2493–2537，2011