# 自然语言处理中的文本表示和语言模型综述

李凡姝　姚登峰
北京市信息服务工程重点实验室　北京 100101
(13153760846@163.com)

**摘　要**　自然语言处理是计算机科学领域与人工智能领域中的一个重要方向，它研究如何能让计算机去理解和使用人类的语言从而实现人机之间的信息交互，继而从大量的文本信息中提取出有效的信息。但自然语言处理领域仍然还处于起步阶段，其中文本数据的表示一直是机器学习领域的一个重要的研究方向。论文主要对传统文本表示的方法和基于深度学习的文本表示方法进行了阐述和分析，并且研究了最近十年语言模型的发展历程，最后对自然语言处理在文本表示中的应用难点进行展望。

**关键词**：人工智能；自然语言处理；文本表示；语言模型
**中图法分类号**　TP181

## Text Representation and Language Model in Natural Language Processing

LI Fan-shu and YAO Deng-feng
Beijing Key Laboratory of Information Service Engineering, Beijing 100101, China

**Abstract**　Natural language processing is an important direction in the field of computer science and artificial intelligence. It studies how to make the computer understand and use human language, so as to realize the information interaction between human and computer, and then extract effective information from a large number of text information. Data representation is still an important field in the field of text representation. This paper mainly describes and analyzes the traditional text representation method and the text representation method based on deep learning studies the development process of language model in the recent ten years, and finally prospects the application difficulties of natural language processing in text representation.

**Keywords**　Artificial intelligence, NLP, Text representation, Language model

## 1　引言

进行自然语言处理时，首先要按照分词标准分词，如果还想要用于其他的应用，如文本分类或者句子相似度分析等，则需要文本的表示。文本表示的目标是计算语言对象之间的相似度，并且希望能把语言中词的符号转变成机器可以理解的语义。通常，人工构建的词典往往很难捕捉到一些有新含义的单词，并且或多或少会包含人类的主观情感，从而忽略掉近义词之间的细微区别，例如 Apple 和 Amazon 的意思分别是苹果和雨林，而它们的新含义是 IT 公司，人工构建的词典可能捕捉不到 IT 公司这个新含义，而且效率低、耗时费力。为了弥补这个缺点，则采用计算机的词表示，把文字表示成计算机能够运算的数字或者向量。

## 2　传统文本表示的方法

### 2.1　单词表示

语言的最小使用单位是词，词的向量表示决定了机器学习模型的构建方法[1]。One-Hot Repre-

通信作者：姚登峰(tjtdengfeng@buu.edu.cn)

sentation 又称为独热编码，其给每个词赋予独一无二的向量，每一个向量空间只有一个维度是 1，其他均为 0。例如，词典为：[我，要，去，北京]，那么需要构造 4 个维度为词典大小的向量，其中“我”在词典的第一个位置上，所以它的向量上的第一个维度应该赋值为 1，其他维度应该赋值为 0，因此“我”：[1，0，0，0]，“要”：[0，1，0，0]，“去”：[0，0，1，0]，“北京”：[0，0，0，1]。One-Hot Representation 如果采用稀疏方式存储，虽然形式上非常简洁[2]，但却无法衡量任意两个词之间的相似度，易出现语义鸿沟现象。

### 2.2 句子表示

词袋（Bag-of-words）模型简称 BOW 模型，BOW 是统计词典库中的单词在一个句子里出现的次数。

#### 2.2.1 boolean representation(布尔类型的表示)

首先构造一个词典库[“我们”，“去”，“爬山”，“今天”，“你们”，“昨天”，“跑步”]，查找词典库中每个位置上的单词是否出现，若出现，则向量中对应维度赋值为 1，否则赋值为 0，故“我们今天去爬山”的向量是[1，1，1，1，0，0，0]。最常用的就是将文档处理成一个与词典中词汇数量一样长的向量，同样词出现赋值为 1，其余赋值为 0，形成一个文档的高效表示，并且也可以计算两个文档的相似度。

#### 2.2.2 count-based representation(基于计数的表示)

基于次数的文本表示考虑了词频，也就是说，当构建一个向量时，需要在对应位置上查词典库中对应的词出现的总次数，且向量里的值可以是 0 或正数。例如“我们昨天去爬山今天去爬山”的向量是[1，2，2，1，0，1，0]。其特点是会把一篇文章看成一个不考虑顺序的词袋，词只要出现过就置非零。传统的基于计数的向量空间模型依赖于共享的单词和同源词[3]。

### 2.3 N-gram 模型

N-gram 模型是一种语言模型（Language Model，LM）。语言模型是一个基于概率的判别模型，其输入是一句话（单词的顺序序列），输出是这句话的概率，即这些单词的联合概率。为了解决语言模型中的联合概率计算量大这个难题，通常会把 $N$ 限定为 2 或者 3，即 Bi-gram（$N=2$）和 Tri-gram（$N=3$）。但是 $N$-gram 的也有缺点，因为 $N$ 会被简化成 2 或者 3，这样就会存在误差，也不会很好地捕捉到词与词之间的相似度。

## 3 基于深度学习的文本表示方法

### 3.1 用上下文表示单词

为了弥补独热编码的缺点，更好地表示一个词的意思，英国语言学家 J. R. Firth 在此基础上造出“You shall know a word by the company it keeps”的语料库语言学名句，大致意思是由词的结伴可知其词。Hinton 提出了一种用 Distributed representation 表示词的方式[4]，通常称为词向量。首先是 Co-Occurrence Counts 基于同现的统计方法，主要是看这个词在上下文里的分布情况，有两种矩阵表现形式，分别是 Term-Term Matrix 和 Term-Document Matrix。

(1)Term-Term Matrix

词与词在语料中的同现次数所组成的矩阵，如“I like deep learning”“I like NLP”“I enjoy flying”所组成的 Term-Term 矩阵，如图 1 所示。

| counts | I | like | enjoy | deep | learning | NLP | flying | . |
|---|---|---|---|---|---|---|---|---|
| I | 0 | 2 | 1 | 1 | 0 | 1 | 1 | 0 |
| like | 2 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| enjoy | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| deep | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| learning | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| NLP | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| flying | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| . | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

图 1 Term-Term Matrix

Fig. 1 Term-Term Matrix

(2)Term-Document Matrix

Term-Document Matrix 的意思是词在文档中出现的次数所组成的矩阵，如“I like deep learning”“I like NLP”“I enjoy flying”所组成的 Term-Document 矩阵，如图 2 所示。

| counts | D1 | D2 | D3 |
|---|---|---|---|
| I | 1 | 1 | 1 |
| like | 1 | 1 | 0 |
| enjoy | 0 | 0 | 1 |
| deep | 1 | 0 | 0 |
| learning | 1 | 0 | 0 |
| NLP | 0 | 1 | 0 |
| flying | 0 | 0 | 1 |
| . | 1 | 1 | 1 |

图 2 Term-document Matrix

Fig. 2 Term-document Matrix

由图2可以看出，D1这一列表示的是第一篇文章“I like deep learning”，这样文章里的词与词就有了顺序，因此可以得出这样的结论：在一个大规模的集合里面，不管是Term-Term Matrix还是Term-Document Matrix，如果两个词的向量较大，那么这两个词的语义相似度就较大，则两个文档的向量也较为相似。

### 3.2 神经网络语言模型

为了避免出现N-gram的缺点，相关研究者研究出了Neural Language Model。神经网络与语言模型的结合最早源自Xu等[5]提出的一种使用神经网络构建二元语言模型的思想；而Bengio等[6]利用3层神经网络来构建$n$元语法模型的工作，把神经网络与语言模型训练的结合推上了一个新的台阶。

### 3.3 Word2vec

Word2vec是深度学习[7]在自然语言处理领域取得很大成绩的里程碑，其方法简单有效。Mikolov等提出的word2vec[8]包含两个模型，分别为CBOW[9]和Skip-gram，其基本思路是通过确定中心词和上下文窗口大小来进行预测。Word2vec使用固定大小的滑动窗口沿句子移动。在每个窗口中，中间词是目标词，其他词是上下文词。CBOW是通过上下文来预测中心词，Skipgram是通过中心词来预测上下文，整体来说是通过自监督训练的模型生成词向量，且通过使用大规模的无监督的文本语料数据集训练得到的分布式词向量中包含有更多的语义和语法的信息，可以为模型提供一个较好的初始值[10]。word2vec的主要问题在于其只考虑局部信息，而局部信息的多少取决于上下文窗口的大小.

### 3.4 向量空间模型

向量空间模型是由Salton[11]在20世纪70年代提出的，其主要思想是：将文本内容的处理转换成对向量空间里的向量计算，继而用空间向量之间的相似度来表示句子相似度[12]。

计算两个句子的相似度通常采用余弦相似度，即计算两个句子向量的夹角所对应的余弦值。夹角越小，则相似度越高，余弦值越接近1，表明夹角越接近0度，也就是两个向量越相似，称作“余弦相似性”。例如两个$n$维向量，$A:[A_1,A_2,A_3,\cdots,A_n]$和$B:[B_1,B_2,B_3,\cdots,B_n]$，则$A$与$B$向量的夹角所对应的余弦值公式如下：

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}=\frac{A\cdot B}{|A|\times|B|}$$

**结束语** 本文阐述了各种文本表示的方法模型，分析比较了文本表示中各个模型的核心思想及优缺点。但在现阶段以深度学习方法为主的自然语言研究领域中，依然存在着很多问题，需要进一步的探索和解决。比如，假设给定一个表示对象，我们应该如何学习特征来表示？诸如此类问题都有待解决。超过25年以来，NLP在许多领域中使用定量统计方法来解决问题的次数逐渐增加，并且这些一直是诸如词性标记、解析或分类等任务中的最新技术。随着“深度学习”程序的出现，这种使用定量统计方法来解决问题的趋势再次升温。

## 参 考 文 献


[1] 李枫林，柯佳. 基于深度学习的文本表示方法[J]. 情报科学，2019，37(1)：156-164.

[2] 奚雪峰，周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报，2016，42(10)：1445-1465.

[3] PEIRSMAN Y，PADÓ S. Cross-lingual induction of selectional preferences with bilingual vector spaces[C]// Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010：921-929.

[4] HINTON G E，SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science，2006，313(5786)：504-507.

[5] XU W，RUDNICKY A I. Can artifificial neural networks learn language models? [C]// Proceedings of 2000 International Conference on Spoken Language Processing (ICSLP02000). Beijing，China：Speech Communication Press，2000：202-205.

[6] BENGIO Y，DUCHARME R，VINCENTP，et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research，2003，3：1137-1155.

[7] HINTONG E. Learning distributed representations of concepts[C]// Proceedings of the 8th Annual Conference of the Cognitive Science Society. Amherst，Massachusetts：Cognitive Science Society Press，1986：1-12.

[8] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed Representations of Words and Phrases and their Compositionality.[C]// Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS 2013). South Lake Tahoe, Nevada, USA, 2013: 3111-3119.

[9] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.

[10] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive autoencoders for predicting sentiment Distributions[C]// Proceeding of the Conference on Empirical Methods in Natural Language Processing. EMNLP,2011.

[11] SALTON G. Automatic processing of foreign language documents[J]. Journal of the American Society for Information Science,1970,21(3):187-194.

[12] 崔莹.深度学习在文本表示及分类中的应用研究[J].电脑知识与技术,2019,15(16):174-177.

**LI Fan-shu**, born in 1998, postgraduate. Her research interests include natural language processing and so on.

**YAO Deng-feng**, born in 1979, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include language cognition and computing, information accessibility.