中文词向量研究报告

朱俊杰[1]

（杭州电子科技大学媒体智能实验室，浙江 杭州 310018）

摘要：一直以来，词嵌入都是自然语言处理中一个重要的研究方向，通过对每一个词向量化，使得计算机能够理解人类的语言。目前对于英语的词向量研究较多，报告中，结合英文常用的词嵌入方法详细介绍在中文词向量中的两个模型：CWE 和 charCBOW，这两种模型是根据不同的方法提出的，前者是基于单汉字的语义进行建模，后者是基于汉字偏旁的语义进行建模。并在报告中分析两种模型的核心技术以及优缺点，最后将对中文词向量的发展趋势和方法进行展望。

关键词：词嵌入；中文；深度学习；

中图分类号：TN401　　**文献标识码**：A　　**文章编号**：

0 引 言

语言是构成人类社会的一个基础要素，通过语言，人们能够快速便捷的进行沟通，交换思维，理解对方的思想。语言的存在，极大的加速了人类社会的发展，而经过成千上万年的演变，人类的语言本身变得十分抽象和复杂，要明白一句话的含义，就需要理解词本身的意思、语义关系以及上下文关系。

对于计算机而言，直接给定一段文字，计算机并不能理解其中的含义。因此需要找到合适的方法，将文本数据转换为数值型数据，让计算机能够理解文本的含义。这里的将文本数据转为数值数据，用到的方法就是词向量。词向量，又叫词嵌入式，是自然语言处理中的一组语言建模和特征学习技术的统称。从概念上讲，它涉及从每个单词一维的空间到具有更低维度的连续向量空间的数学嵌入。

一个优秀的词向量表示，不仅能够让计算机理解一段文字的含义，而且还能够根据给定的文字，作出相应的回答，在这里又会涉及到将计算机理解的数值数据转化成人类理解的文本数据。 因此，词向量拥有巨大的实际应用前景，是人类与计算机之间进行沟通的一座桥梁，为日后的人机语音对话、机器问答等领域提供了基础。

1 英语词向量发展

作为文本表示的基础，词向量表示的目的是将单词表示为向量，向量既可以用来计算单词之间的语义关联，也可以作为单词特征输入机器学习系统。在自然语言处理中，词向量的表示是一个很重要的研究方向，它能让计算机理解人类的语言，理解词语与词语之间的关系。而在该领域中，学者研究最多的就是英语的词向量表示，相继提出了 One-hot 编码、CBOW、Skip-Gram[1] 和 GloVe[2]等方法，让机器越来越理解人类的语言，并获得一定的成效。

在不少自然语言处理任务中，常使用 One-hot 编码，它将其中每个词表示为一个词汇量向量，只有一个非零条目。由于其表示简单且使用方便，因此 One-hot 表示在自然语言处理中被广泛采用，作为 bag-of-words (BOW)文档模型的基础[3]。但是 One-hot 编码方法它没有

作者简介：朱俊杰（1997-），男，浙江杭州人，研究生，研究方向：TextVQA.

考虑到词与词之间的相互联系，而且得到的特征是离散稀疏的，所以显得并不“智能”。

CBOW 和 Skip-Gram 模型是由 Mikolov 等人在 2013 年提出的从大规模文本语料库中学习词嵌入的有效模型。CBOW 的训练目标是结合上下文词的嵌入来预测目标词;而 Skip-Gram 则是利用每个目标词的嵌入来预测其上下文词。

在 2014 年，Jeffrey Pennington 等人提出了著名的 GloVe 模型，它是基于全局词汇共现的统计信息来学习词向量，可以理解为在 CBOW 和 Skip-Gram 模型的基础上，增加了全局矩阵分解的思想，从而能够将统计信息与局部上下文窗口方法的优点都结合起来，在单词之间一些语义特性，比如相似性（similarity）、类比性（analogy）等方面获得了较大的提升。

2 中文词向量模型

根据中文的语言特性，常见的中文词向量处理思想有基于单汉字的语义、汉字偏旁的语义和汉字笔画的语义等，现就其中的基于单汉字的语义和基于汉字偏旁的语义两种方法，结合具体的论文，详细描述其模型。

2.1 基于单汉字的语义

在 One-hot、CBOW、Skip-Gram 和 GloVe 方法中，更多的考虑是词语与词语之间的联系，却鲜少考虑到组成词语的字符之间的联系。很显然，尤其在中文中，词汇的含义与构成它的汉字有着非常紧密的关系，例如与“菜”这个汉字搭配的词语大多都是能做副食品的植物：菠菜、青菜等。所以对于一个词语中的每个汉字理解也显得尤为重要，能够根据汉字推测出词语的含义。

在论文《Joint Learning of Character and Word Embeddings》中，作者基于 CBOW 方法，提出了一种模型 CWE，以增强词向量的表示[4]。该模型的数学表达可表示为：

$$X_j = W_j \circ \frac{1}{Nj}\sum_{k=1}^{Nj} C_k$$

其中，W_j是通过 CBOW 方法得到的词向量，Nj是构成W_j词的每个汉字数量，C_k是构成W_j词的汉字，○表示一种复合操作，在该论文中指的是相加处理。以“智能时代到来”为例，“智能”的词向量 +（“智”+“能”）两个的字向量=“智能”的综合向量，同理得到“到来”的综合向量，用来预测“时代”这个词，如图 1 所示。

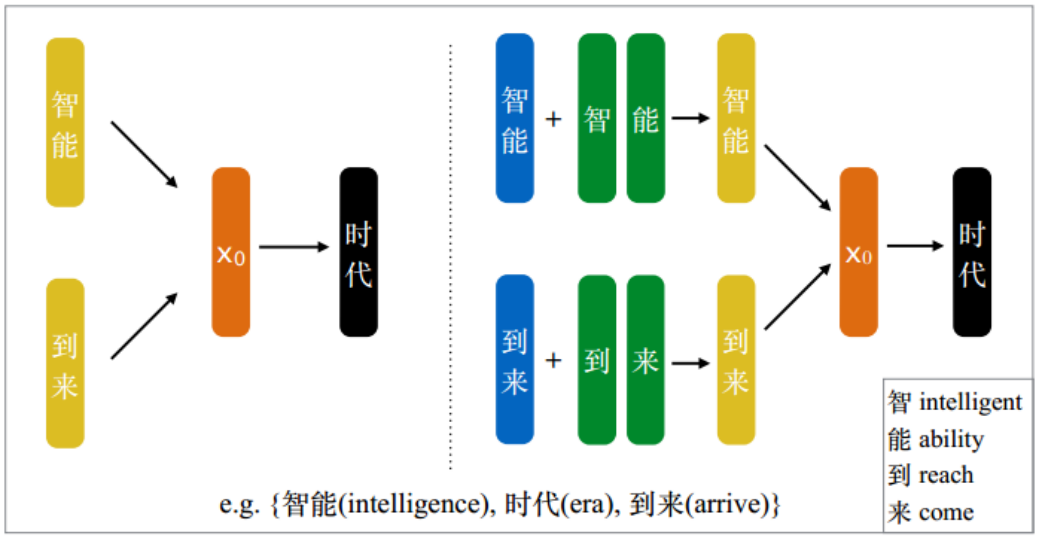


图 1 CBOW 和 CWE

相较于英语单词而言，中文的每个汉字信息更加模糊，也就是说在不同的词语中，同一个字时常会有不同的意思，所以仅用一个向量表示一个字符是不够的。例如“喝酒”和“喝彩”，虽然都有“喝”这个字，但其语义完全不同，所以不可用单一向量表示。为解决此类问题，作者提出了三种方法用来表示相同汉字的多模式向量。

2.1.1 基于位置的多模式向量表示

在一个词的内部，字的位置有很大的作用，不同的位置表达不同的意思。该方法为每个汉字保留了 3 个向量：C^B，C^M，C^E，分别表示汉字位于词语位置中的头部，中部，尾部。这时候词向量的数学表达式为：

$$X_j = W_j \circ \frac{1}{Nj}(C_1^B + \sum_{k=2}^{Nj-1} C_k^M + C_j^E)$$

2.1.2 基于聚类的多模式向量表示

该方法类似于 K-means 算法，对于每一个汉字提前分配 x 个字向量，x 的个数是模型的一个超参数，代表了每个汉字潜在的语义模式。这时候词向量的数学表达式为：

$$X_j = W_j \circ \frac{1}{Nj}\sum_{k=1}^{Nj} C_k^{r_k^{max}}$$

其中的r_k^{max}是与上下文内容中余弦相似度最大值对应的模式向量，即：

$$r_k^{max} = argmax_{r_k}\cos(C_k^{r_k}, V_{context})$$

其中的$V_{context}$为该汉字对应词语前后窗口内的 2k 个词语，这些词语的向量形式同样由词向量和字向量叠加形成，词向量也是通过 CBOW 方法得到，字向量是过去被挑选最多次的模式向量，用数学公式表示为：

$$V_{context} = \sum_{t=j-K}^{j+K} X_t = \sum_{t=j-K}^{j+K} w_t + \frac{1}{N_t} \sum_{c_u \in x_t} c_u^{most}$$

2.1.3 基于非参数的多模式向量表示

该方法与基于聚类的多模式向量表示中的思想相似，唯一不同的是，基于聚类的多模式向量表示中的每一个汉字对应的模型向量的数量是一个预先设定的固定值，也就是作为模型的超参数。而在基于非参数的多模式向量表示的方法中，该值是一个模型自动学习的值。具体方法如下式所示：

$$r_k = \begin{cases} Nc_k + 1, if \cos\left(C_k^{r_k}, V_{context}\right) < \varepsilon \ for \ all \ r_k \\ r_k^{max}, \quad otherwise \end{cases}$$

2.2 基于汉字偏旁的语义

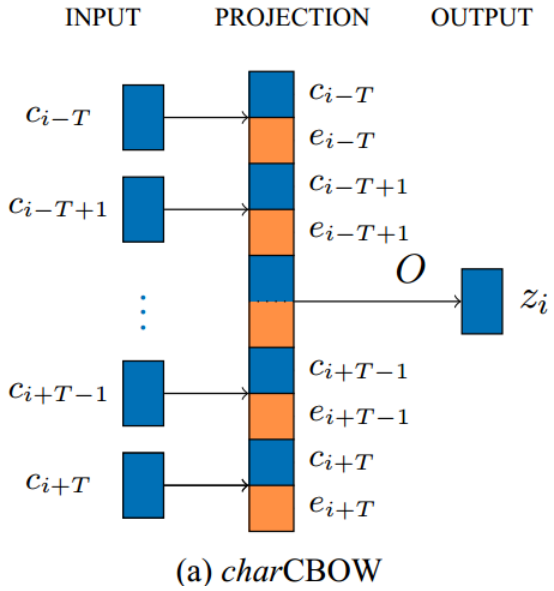
在英语中，词是是进行语义表达的最小单位，即使再细分，英语单词也只能分为 26 个字母。而对于中文而言，词语不仅可以细分为字，还可以继续细分为偏旁部首，甚至笔画。因此，可以认为在中文汉字内包含了不少信息。在论文《Component-Enhanced Chinese Character Embeddings》中，作者认为汉字的组件（部首）包含了大量的语义信息，某些偏旁部首相同的字(如“江”和“河”、“潮”和“湿”)所代表的语义有很大的相似性和关联，基于此提出了两个词向量模型，分别是基于 CBOW 方法改进的 charCBOW 模型和基于 Skip-Gram 方法改进的 charSkipGram 模型，对中文字向量进行了改善，提高了词向量的信息表示[5]。这里以 charCBOW 模型为例，具体描述。

论文中，作者将汉字中的每个组件组成了一个 component 列表，其中的部首信息要比其他的组件信息包含更加丰富的语义信息，如图 2 所示。

transform	meaning	transform	meaning
艹 → 艸	grass	扌 → 手	hand
亻 → 人	human	氵 → 水	water
刂 → 刀	knife	車 → 车	vehicle
犭 → 犬	dog	攵 → 攴	hit
灬 → 火	fire	纟 → 糸	silk
钅 → 金	gold	耂 → 老	old
麥 → 麦	wheat	牜 → 牛	cattle
饣 → 食	eat	飠 → 食	eat
礻 → 示	memory	忄 → 心	heart
罒 → 网	nest	王 → 玉	jade
讠 → 言	speak	衤 → 衣	cloth
月 → 肉	body	辶 → 辵	walk

图 2 component 列表

由于部首蕴含的信息更多，因此将部首放在了 component 列表的前部进行训练，其训练策略和 CBOW 相似，即结合上下文词的嵌入来预测目标词。如图 3 所示，其中，c 代表的是上下文词，e 代表的是 component 列表中的组件，z 代表的是目标词。



(a) *char*CBOW

图 3　charCBOW 模型

根据 charCBOW 模型，其目标函数即对数似然函数可表示为：

$$L = \sum_{z_i^n \in D} log\, p(z_i|h_i)$$

其中，h_i表示上下文组件的拼接，即：

$$h_i = cat(c_{i-T}, e_{i-T}, \dots, c_{i+T}, e_{i+T})$$

3 分析与总结

词向量中的 One-hot、CBOW、Skip-Gram 和 GloVe 等方法大多都是基于英语，对于中文的支持并不是很友好，但是可以看到目前对中文的词向量研究越来越多。由于中文是一种强表意文字，汉字中的一笔一划，一个部首都有其含义，因此相对于英语等语言来说，提高中文词向量表示的一个研究方向就是对于汉字内部语义的分析。在这篇研究报告中，详细介绍了基于单汉字的语义和基于汉字偏旁的语义两种方法，但除此之外，还有别的方法，如阿里巴巴在 2018 年提出的 cw2vec 模型是基于汉字的笔画特性来提升中文词向量表示的[6]。

该研究报告中提到的这两种方法虽然提升了词向量的表示，但是仍旧有进一步提升的空间。CWE 模型中，仅根据汉字在词语中的位置来确定字的模式向量并不十分准确，例如“喝彩”和“喝酒”两个词语中，“喝”虽然都位于词语的首部，但在这两个词语中，

“喝”并不是同一个意思，因此不可用同一模式向量表示。此外，CWE 模型中，词向量和字向量的复合处理只是简单的相加，或许可以考虑别的方法，例如增加权重等方法，让词向量和字向量的表示更符合中文的特性。

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop Papers.

[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532－1543.[3] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to information retrieval, volume 1. Cambridge university press Cambridge, 2008.

[3] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to information retrieval, volume 1. Cambridge university press Cambridge, 2008.

[4] X. Chen, L. Xu, Z. Liu, M. Sun, and H. Luan, “Joint learning of character and word embeddings,” in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[5] Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced chinese character embeddings. arXiv preprint arXiv:1508.06669.

[6] Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li.2018. cw2vec: Learning chinese word embeddings with stroke n-gram information