CS 330 MIP – Lecture 08

文本信息处理 4

Text Information Processing 4

Jimmy Liu 刘江

2025-04-09

几个面向中文的命名实体识别和关系抽取工具

功能		工具	网址
中文分词	1	jieba	https://github.com/fxsjy/
命名实体识别	1	LTP	http://ltp.ai/
	2	PyHanlp	https://github.com/hankcs/pyhanlp
	3	BosonNLP	http://static.bosonnlp.com/
	4	Lac	https://github.com/baidu/lac
	5	fnlp	https://github.com/FudanNLP/fnlp
	6	StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/
关系抽取	1	DeepKE	https://github.com/zjunlp/deepke
	2	Jiagu	https://github.com/ownthink/Jiagu
	3	DeepDive	http://www.openkg.cn/dataset/cn-deepdive

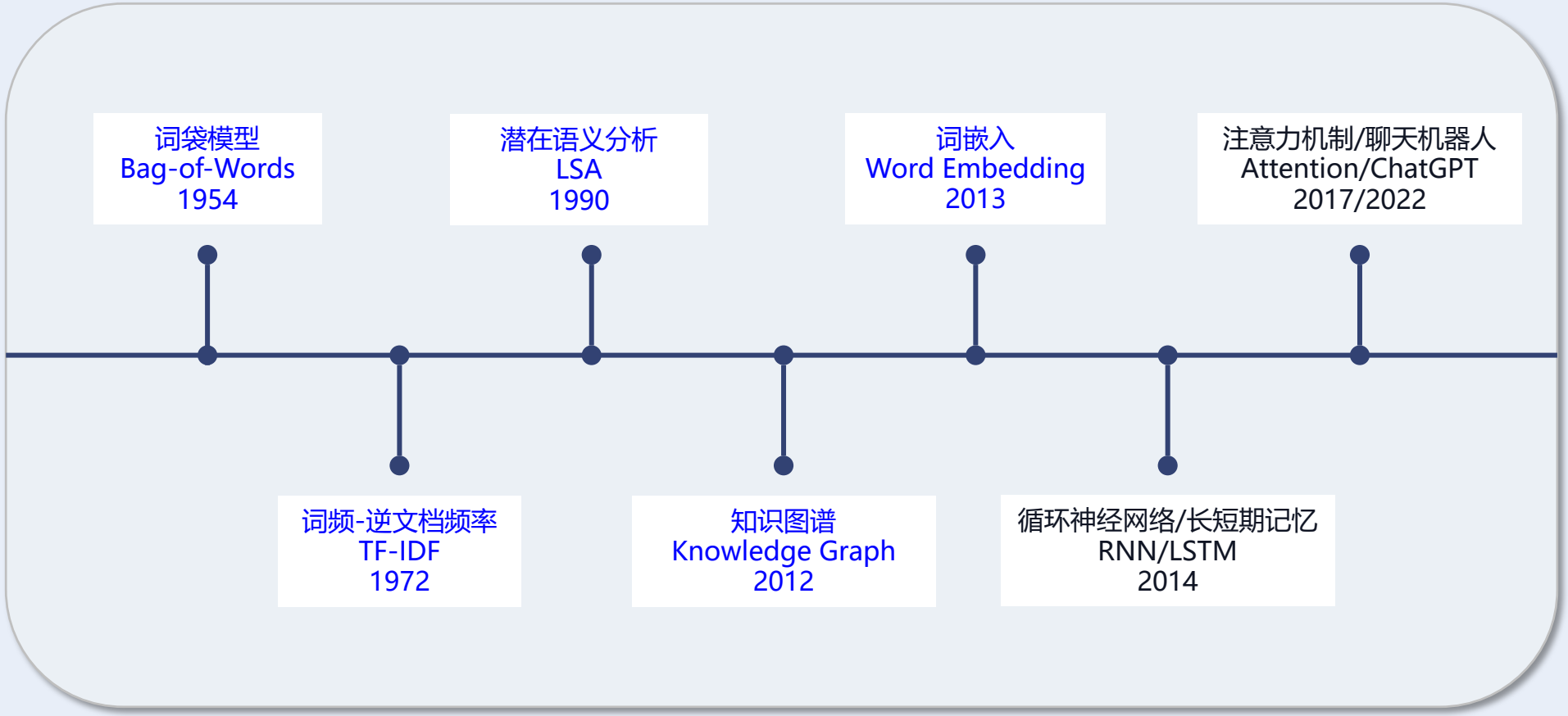
知识图谱开发平台

1	**Neo4j Graph Platform**	https://neo4j.com/
2	**TigerGraph Platform**	https://www.tigergraph.com.cn/
3	**KG_华为云**	https://support.huaweicloud.com/kg/index.html
4	**Maji**	https://magi.com/
5	**OpenKG**	http://openkg.cn/

1~3：https://zhuanlan.zhihu.com/p/136584662
4~6：https://zhuanlan.zhihu.com/p/164626545
10-11：https://www.cnblogs.com/jetHu/p/12327629.html

6	**明略科技HAO图谱Open API**	https://www.mininglamp.com/
7	**阿里云知识图谱开放平台DataG**	https://www.aliyun.com/product/datag
8	**联想知识图谱开放平台HyperGraph**	https://dibg.lenovo.com.cn/leapHyperGraph.html
9	**百度大脑AI开放平台**	https://ai.baidu.com/solution/kgaas
10	**中文通用百科知识图谱（CN-DBpedia）**	http://kw.fudan.edu.cn/cndbpedia/download/
11	**思知识图谱**	https://github.com/ownthink/KnowledgeGraphData

文本处理发展里程碑


词袋模型
Bag-of-Words
1954
词频-逆文档频率
TF-IDF
1972
潜在语义分析
LSA
1990
知识图谱
Knowledge Graph
2012
词嵌入
Word Embedding
2013
循环神经网络/长短期记忆
RNN/LSTM
2014
注意力机制/聊天机器人
Attention/ChatGPT
2017/2022

词嵌入

词向量（Word embedding），又叫Word嵌入式自然语言处理（NLP）中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射到实数的向量。 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。

生成这种映射的方法包括神经网络，单词共生矩阵的降维，概率模型，可解释的知识库方法，和术语的显式表示 单词出现的背景。

当用作底层输入表示时，单词和短语嵌入已经被证明可以提高NLP任务的性能，例如语法分析和情感分析。

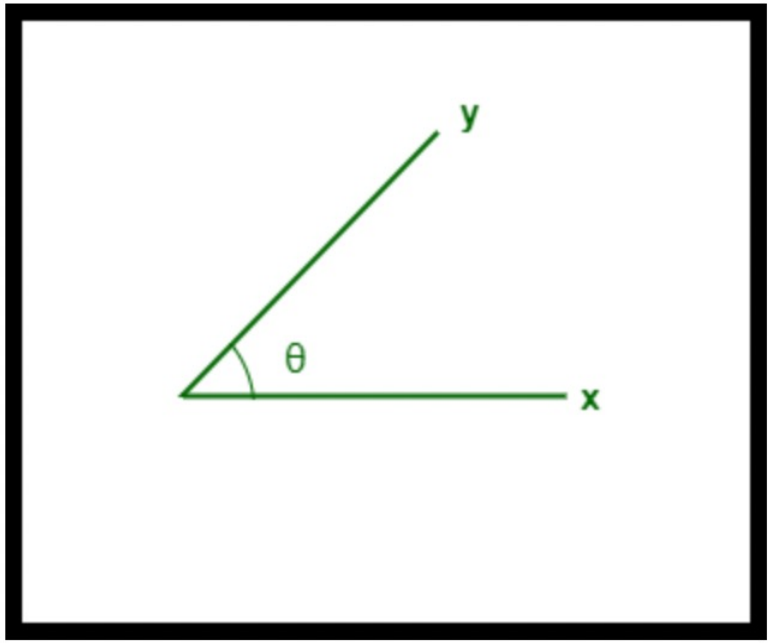
用于培训和使用文字嵌入的软件包括Tomas Mikolov的Word2vec，斯坦福大学GloVe，fastText，Gensim， Indra和Deeplearning4j。主成分分析（PCA）和T分布式随机邻居嵌入（t-SNE）都用于减少单词向量空间的维度，并可视化单词嵌入和集群。

词嵌入的目标是使具有相似含义的词在空间上占据相近的位置

余弦相似性 (Cosine Similarity)

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0度角的余弦值是1，而其他任何角度的余弦值都不大于1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为1；两个向量夹角为90°时，余弦相似度的值为0；两个向量指向完全相反的方向时，余弦相似度的值为-1。

The cosine similarity between two vectors is measured in 'θ'.
If θ = 0°, the 'x' and 'y' vectors overlap, thus proving they are similar.
If θ = 90°, the 'x' and 'y' vectors are dissimilar.



Cosine Similarity between two vectors

Word2Vec

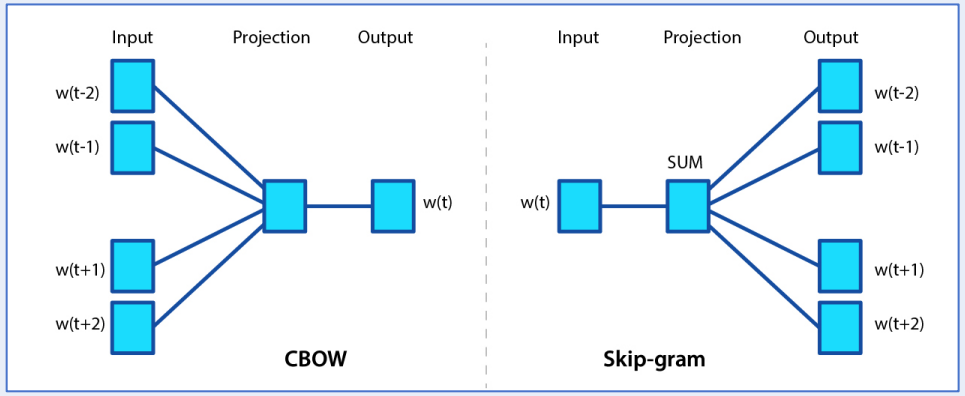
Word2vec是一种用于自然语言处理的技术，于2013年发表。Word2vec算法利用神经网络模型从大型文本语料库中学习词汇关联。经过训练后，该模型可以检测同义词或为不完整的句子提供额外的词汇建议。

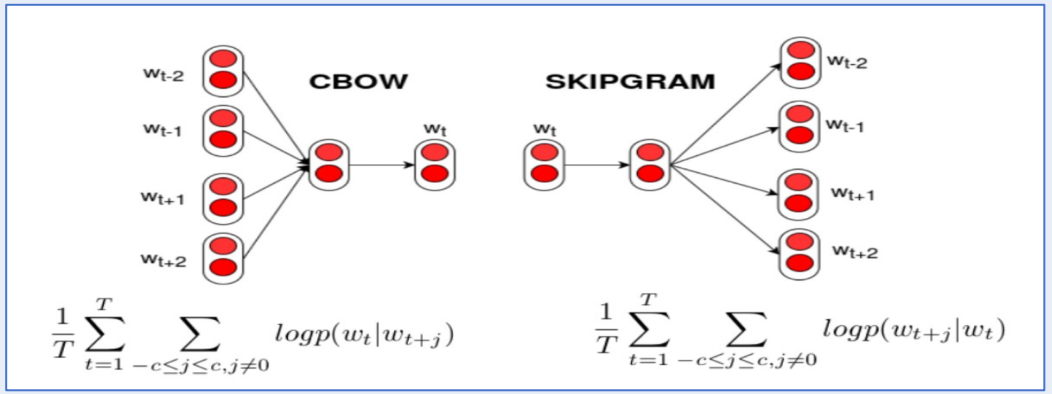
- Word2vec是由Google的Tomas Mikolov领导的研究团队引入的一种计算词向量表示的方法：
  - 谷歌发布了一个以Apache 2.0许可发布的Word2vec开源版本。
  - 在2014年，Mikolov离开Google加入了Facebook。
  - 在2015年5月，谷歌获得了该方法的专利，但这并不废除其发布时采用的Apache许可证。

1. Tomas Mikolov. (2013). Distributed Representations of Words and Phrases and their Compositionality.
2. Tomas Mikolov. (2013). Efficient Estimation of Word Representations in Vector Space.
3. Xin Rong. (2013) word2vec Parameter Learning Explained.

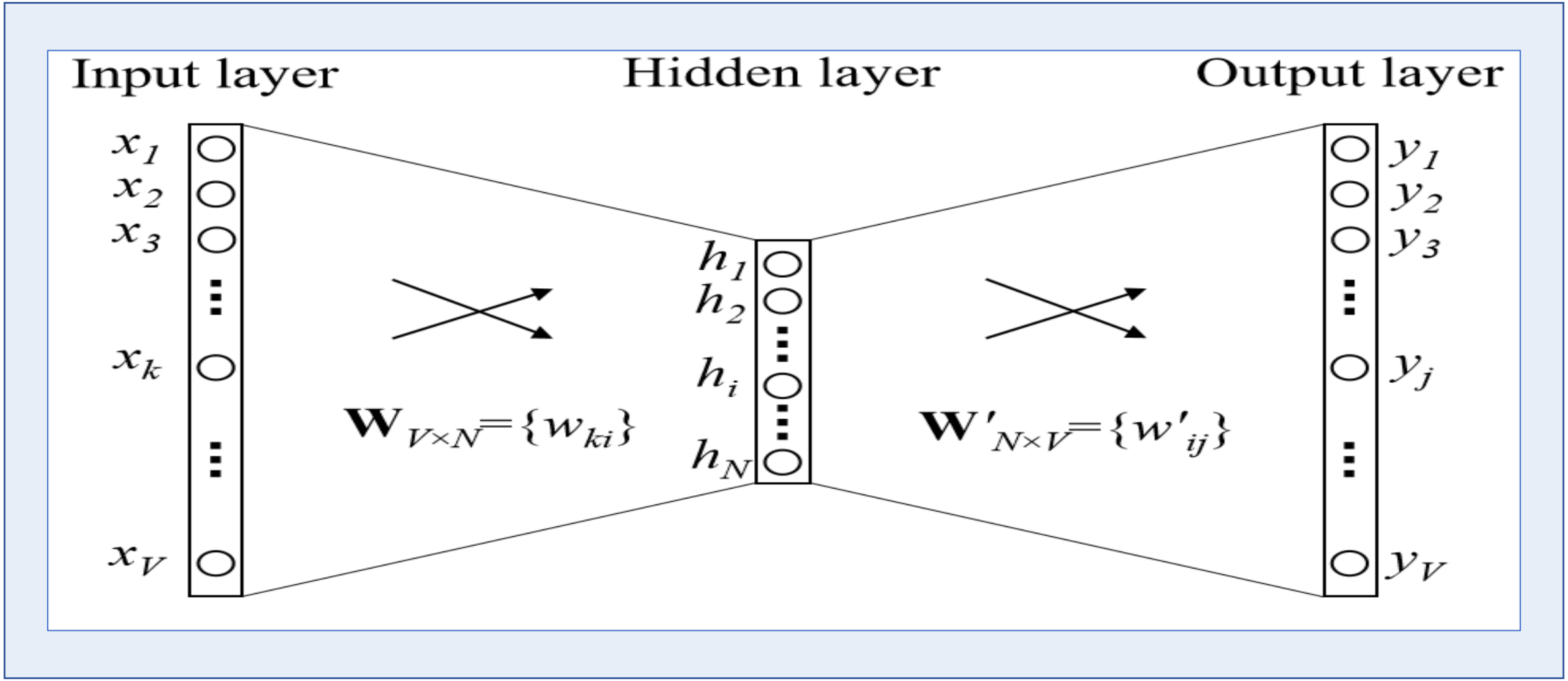
Word2vec – CBOW 和 Skip-Gram

- Word2Vec架构：
  - i) 连续词袋模型 (Continuous Bag of Words, CBOW).在CBOW架构下，模型通过周围上下文单词来预测当前单词。
  - Ii) Skip-Gram模型.神经网络被训练为在给定当前单词的情况下预测周围的上下文单词。

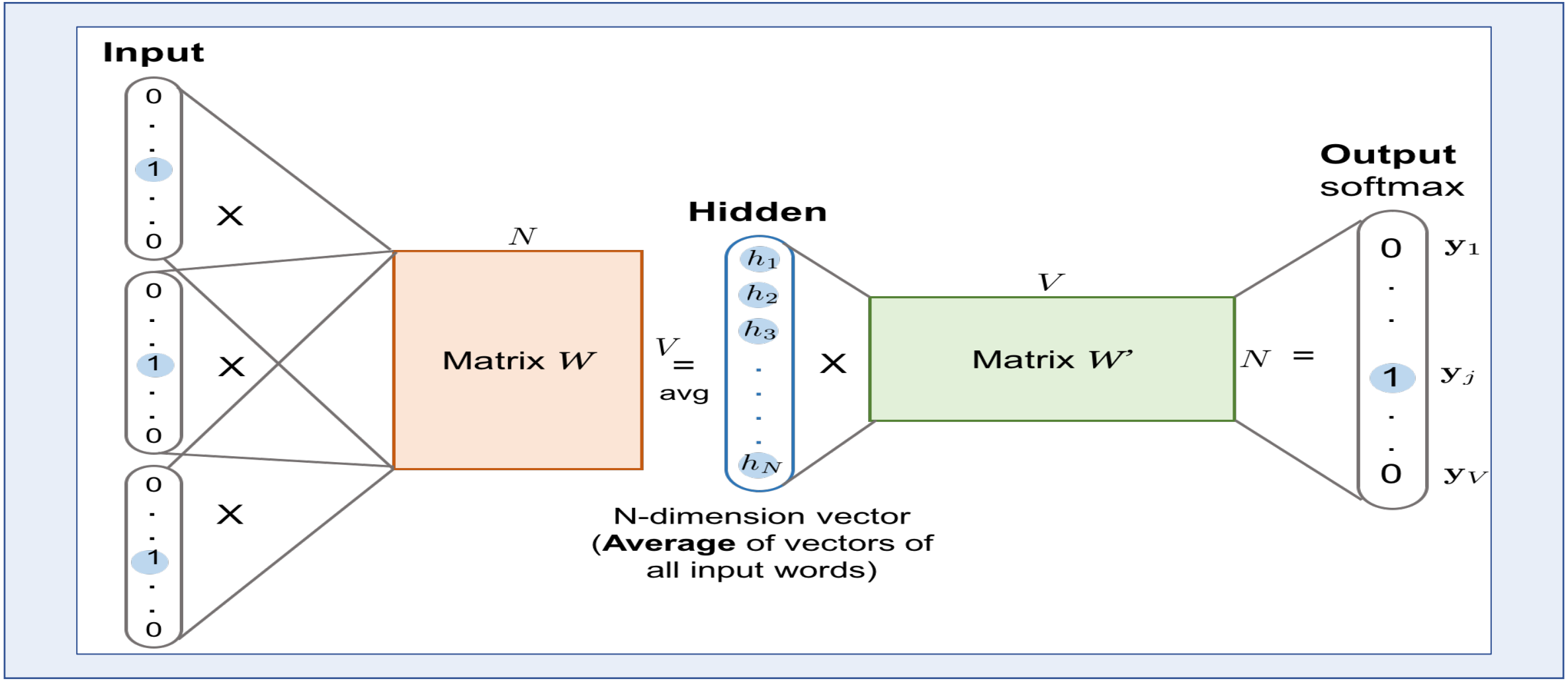




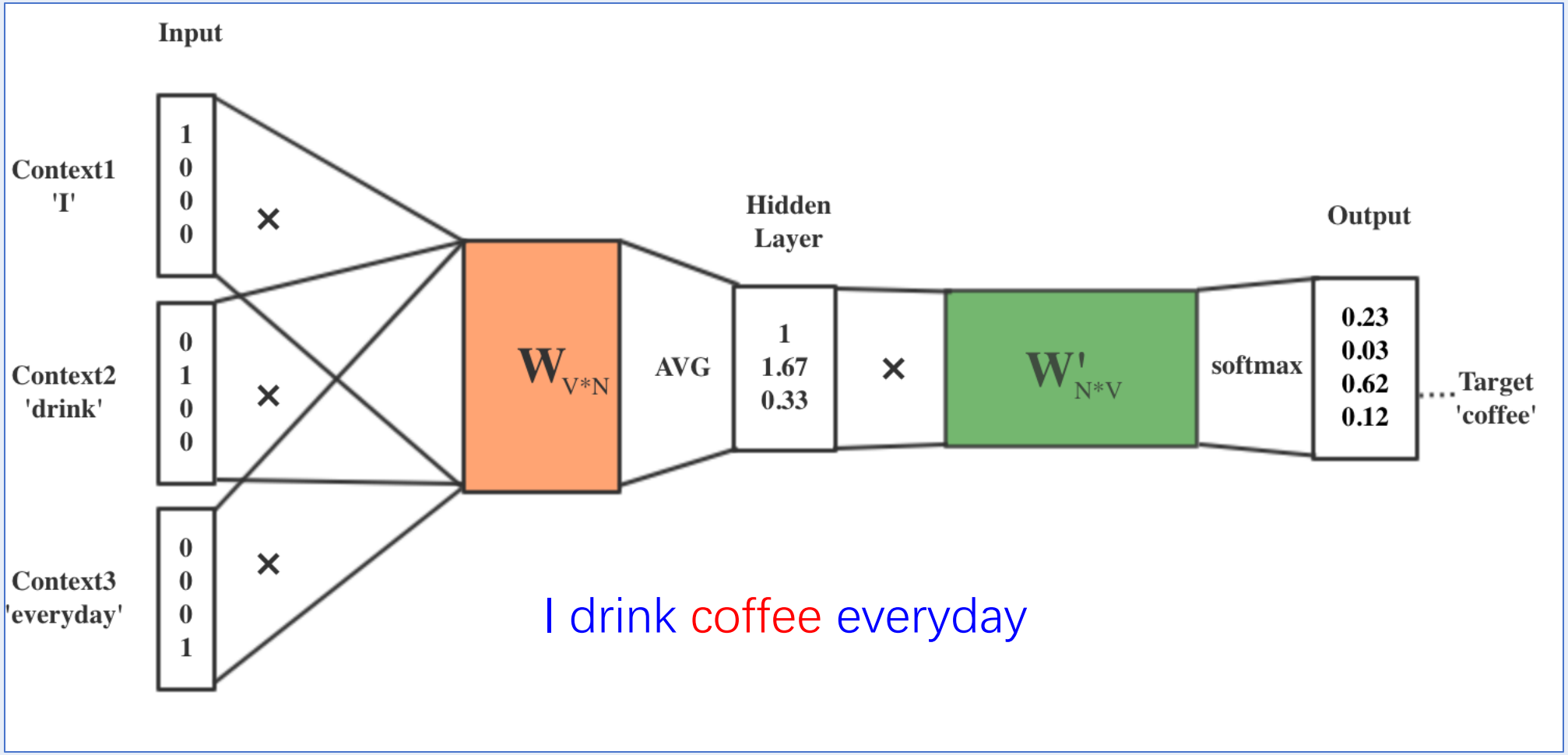
CBOW – 基于嵌入的预测
输入 和 输出 长度 V， 向量长度 N



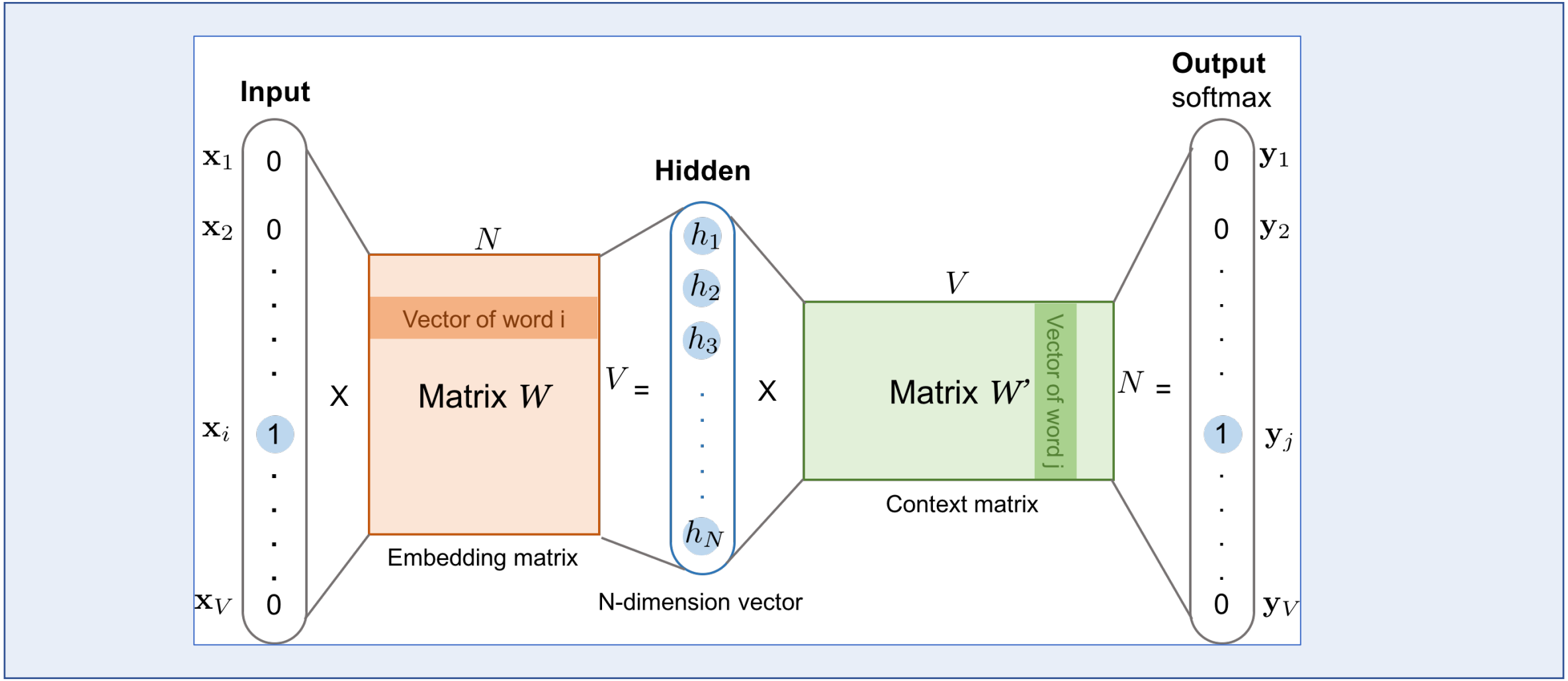
CBOW – 输入层多个词库



CBOW 举例 – 隐含层用Average操作，输出前用Softmax



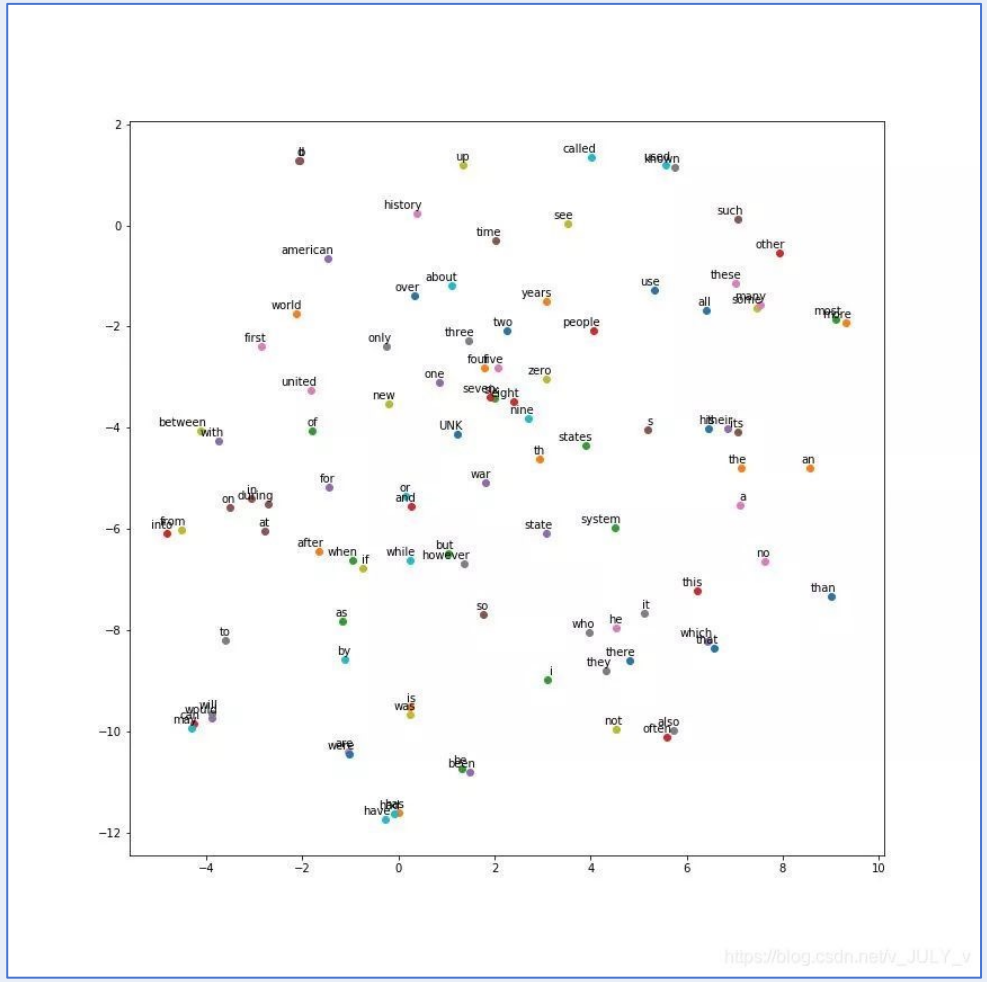
Skip-gram - 输入和输出大小 V，向量大小 N



Word2Vec 应用

- **计算单词相似度**：词嵌入可以用于预测与给定单词语义相似的单词，同时也可以用于预测与给定单词语义不相似的单词。
- **创建相关单词组**：单词嵌入用于语义分组，将具有相似特征的事物归为一组，而将不相似的事物远离归为另一组。
- **文本分类的特征**：将文本映射为向量数组，供模型进行训练和预测。基于文本的分类器模型无法在字符串上进行训练，因此这将把文本转换为可机器训练的形式。此外，它构建语义特征有助于基于文本的分类。
- **文档聚类**：Word2Vec中的词嵌入也被广泛用于文档聚类的应用。
- **自然语言处理**：除了上述的应用，词嵌入在其他很多的应用中也非常有效，取代了特征提取阶段，如词性标注、情感分析和句法分析。

50000 Words 和 128 Vectors



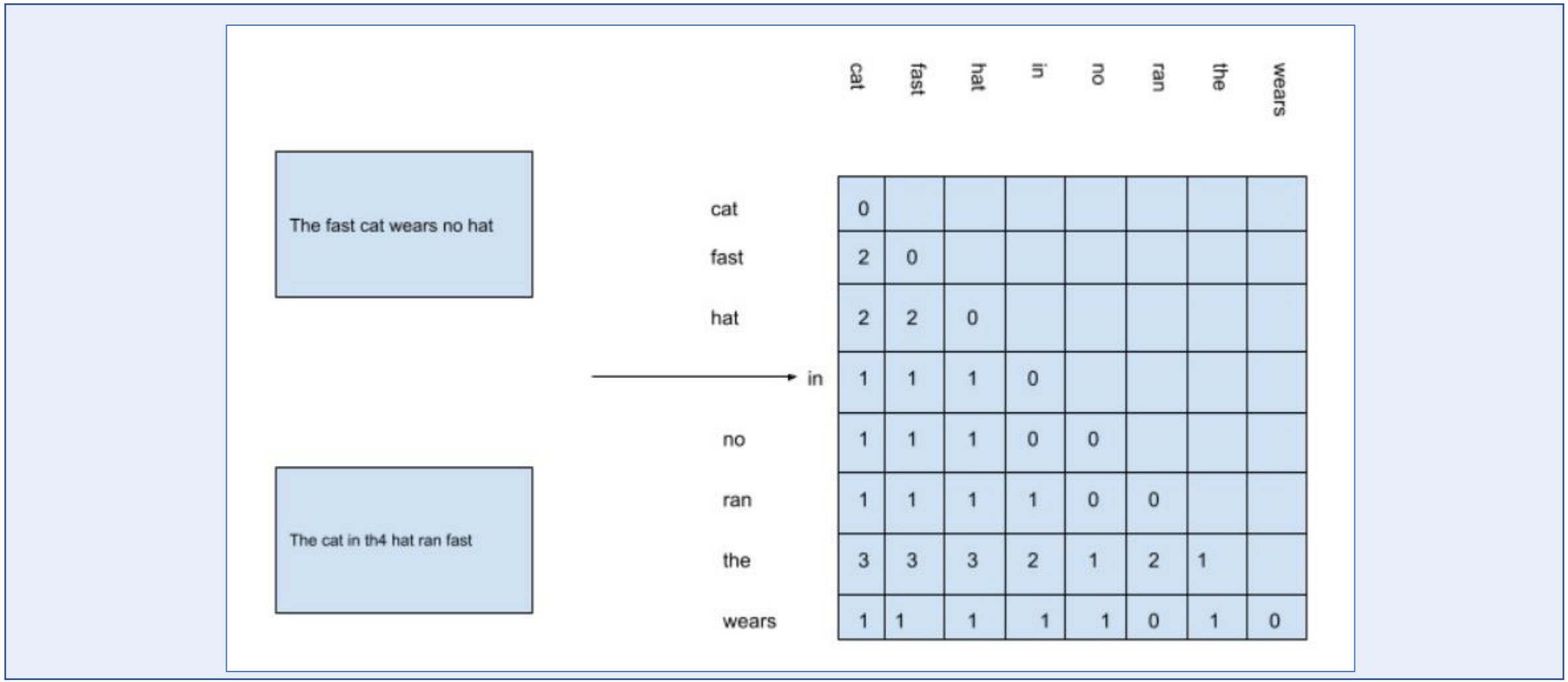
GloVe

GloVe（全称Global Vectors）是一种用于分布式词表示的模型。该模型是一种无监督学习算法，用于获取单词的向量表示。它通过将单词映射到一个有意义的空间中，其中单词之间的距离与语义相似性相关联来实现。训练是基于语料库中的全局单词共现统计数据进行的，并且所得到的表示展示了单词向量空间的有趣的线性子结构。GloVe是在斯坦福大学开源项目中开发的，于2014年推出。作为一种用于无监督学习单词表示的对数双线性回归模型 (log-bilinear regression model)，它结合了全局矩阵分解和局部上下文窗口方法两种模型家族的特点。

双对数模型：因变量和自变量均进行对数变换，表达式为 $\ln y = a + b\ln x + u$。

单对数模型：仅对自变量进行对数变换，表达式为 $y = a + b\ln x + u$。

GloVe – 共现矩阵


The fast cat wears no hat
The cat in th4 hat ran fast
cat
fast
hat
in
no
ran
the
wears
cat fast hat in no ran the wears
0
2 0
2 2 0
1 1 1 0
1 1 1 0 0
1 1 1 1 0 0
3 3 3 2 1 2 1
1 1 1 1 1 0 1 0

GloVe 和 Word2Vec

- Word2Vec利用一个三层神经网络训练，训练网络的副产品是词向量。通过训练神经网络，Word2Vec能够学习到词语的分布式表示，其中相似的词语在向量空间中会有相近的表示。
- GloVe旨在强制模型使用词的共现矩阵来学习单词之间的线性关系。GloVe的特点是能够捕捉到词语之间的全局语义关系，例如词语的语义相似性和关联性。

Word2Vec

无法处理未登录词 (Out-of-Vocabulary-Words, OOV)

Word2Vec的最大问题之一是无法处理未知或超出词汇表的单词。如果模型之前没有遇到过一个单词，它将无法理解该单词或为其构建一个向量。在这种情况下，我们往往只能用一个随机向量替代这些OOV，但是这并不是一个最优解。OOV问题在Twitter或微博这样的环境中尤为成问题，因为在这样的环境中存在很多噪声和稀疏的数据。在这种大体量的数据中，很多单词可能只出现一次或两次。

Homework 08

- 基于以下3篇文章和其他文章写一个WORD2VEC的读书报告（1000字以上）

1. Tomas Mikolov. (2013). Distributed Representations of Words and Phrases and their Compositionality.
2. Tomas Mikolov. (2013). Efficient Estimation of Word Representations in Vector Space.
3. Xin Rong. (2013) word2vec Parameter Learning Explained.

CS 330 MIP – Lecture 08

文本信息处理 4

Text Information Processing 4

Jimmy Liu 刘江

2025-04-09

Transformer