Lecture 1: Introduction and Word Vectors

Representing words as discrete symbols

传统的NLP中，我们将单词视作离散的符号：hotel, conference, motel

单词可以用one-hot向量来表示：

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

向量的维度 = 词汇表中的单词数

Problems with words as discrete symbols

例子：在网络搜索中，如果用户搜索“Seattle motel“，我们乐意看到匹配的文件中包含“Seattle hotel”。但是由于

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

这两个向量是互相正交的，并且one-hot向量表示缺少自然的表达similarity的方式。

解决方案：

- 将similarity编码进词向量中

Representing words by their context（用上下文表示单词）

- 语义分布(Distributional semantics)：一个单词的含义是由经常出现在上下文中的单词所决定的。
- 当一个单词w出现在文本中，它的上下文即是一组出现在w附近的单词（在一个固定大小的窗口中）
- 用许多w的上下文去构建一个w的表示

Word vectors

我们为每一个单词构建一个稠密的向量，使得该词向量相似于它的上下文词向量。如：

$$banking = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Note: 词向量（word vectors）有时候被称为词嵌入（word embeddings）或者词表示（word representations）

Word2vec

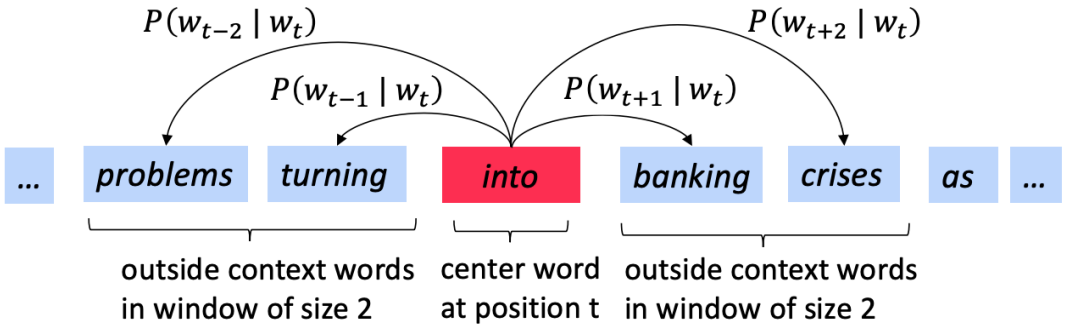
Overview

Word2vec是一个学习词向量的框架。

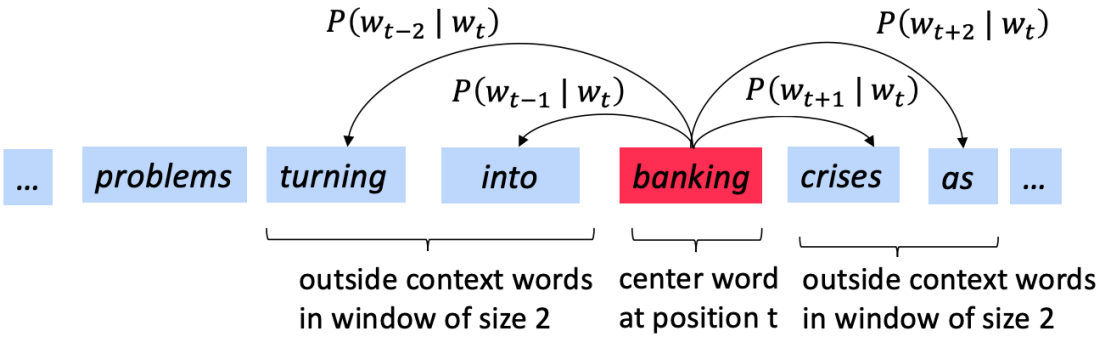
Idea:

- 我们有一个大的语料库
- 一个固定的词汇表中的每个单词都用一个向量来表示
- 遍历文本中的每一个位置t，该位置有一个中心词c和上下文单词o
- 使用c和o之间的词向量相似度去计算P(o|c)
- 持续调整词向量以达到最大化该概率

- Example windows and process for computing $P(w_{t+j} \mid w_t)$



- Example windows and process for computing $P(w_{t+j} \mid w_t)$



Objective function

对于每一个位置t = 1, … , T，在给定中心词Wj的情况下，预测固定窗口为m的上下文单词

$$Likelihood = L(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m, j \neq 0} P(W_{t+j}|W_t;\theta)$$

目标函数是负对数似然函数：

$$J(\theta) = -\frac{1}{T} logL(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} logP(W_{t+j}|W_t;\theta)$$

$$最小化目标函数 \iff 最大化似然函数$$

对每一个单词使用两个向量：

$$v_w 当 w 是中心词$$

$$u_w 当 w 是上下文词$$

那么对于一个中心词c和一个上下文词o：

$$P(o|c) = \frac{exp(u_o^T v_c)}{\sum_{w \in V} exp(u_w^T v_c)}$$

- 指数函数exp()是为了使所有结果保持正数
- 使用点乘比较o和c的相似度，点乘结果更大 = 更高概率
- 分母部分的求和是为了符合概率分布的要求（summation is one）

Training with Gradient Descent

关于中心词c的偏导数推导：

$$\begin{aligned}
\frac{\partial logP(o|c)}{\partial v_c} &= \frac{\partial}{\partial v_c} log \frac{exp(u_o^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial v_c} u_o^T v_c - \frac{\partial}{\partial v_c} log(\sum_{w=1}^{V} exp(u_w^T v_c)) \\
&= u_o - \frac{\frac{\partial}{\partial v_c} \sum_{w=1}^{V} exp(u_w^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} \\
&= u_o - \sum_{w=1}^{V} \frac{exp(u_w^T v_c)}{\sum_{\hat{w}=1}^{V} exp(u_{\hat{w}}^T v_c)} u_w \\
&= u_o - \sum_{w=1}^{V} P(w|c) u_w
\end{aligned}$$

关于上下文词o的偏导数推导：

$$
\begin{aligned}
\frac{\partial log P(o|c)}{\partial u_o} &= \frac{\partial}{\partial u_o} log \frac{exp(u_o^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial u_o} u_o^T v_c - \frac{\partial}{\partial u_o} log(\sum_{w=1}^{V} exp(u_w^T v_c)) \\
&= v_c - \frac{\frac{\partial}{\partial u_o} \sum_{w=1}^{V} exp(u_w^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} \\
&= v_c - \frac{exp(u_o^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} v_c \\
&= v_c - P(o|c) v_c \\
&= (1 - P(o|c)) v_c
\end{aligned}
$$