

Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心



跨语言词向量

博士研究生 赵一飞

2018年02月04日

- 基础回顾
- 背景
- 跨语言词向量简介
- 基于词对齐平行语料的CWE训练
- GAN-CWE
- Q&A



基础回顾

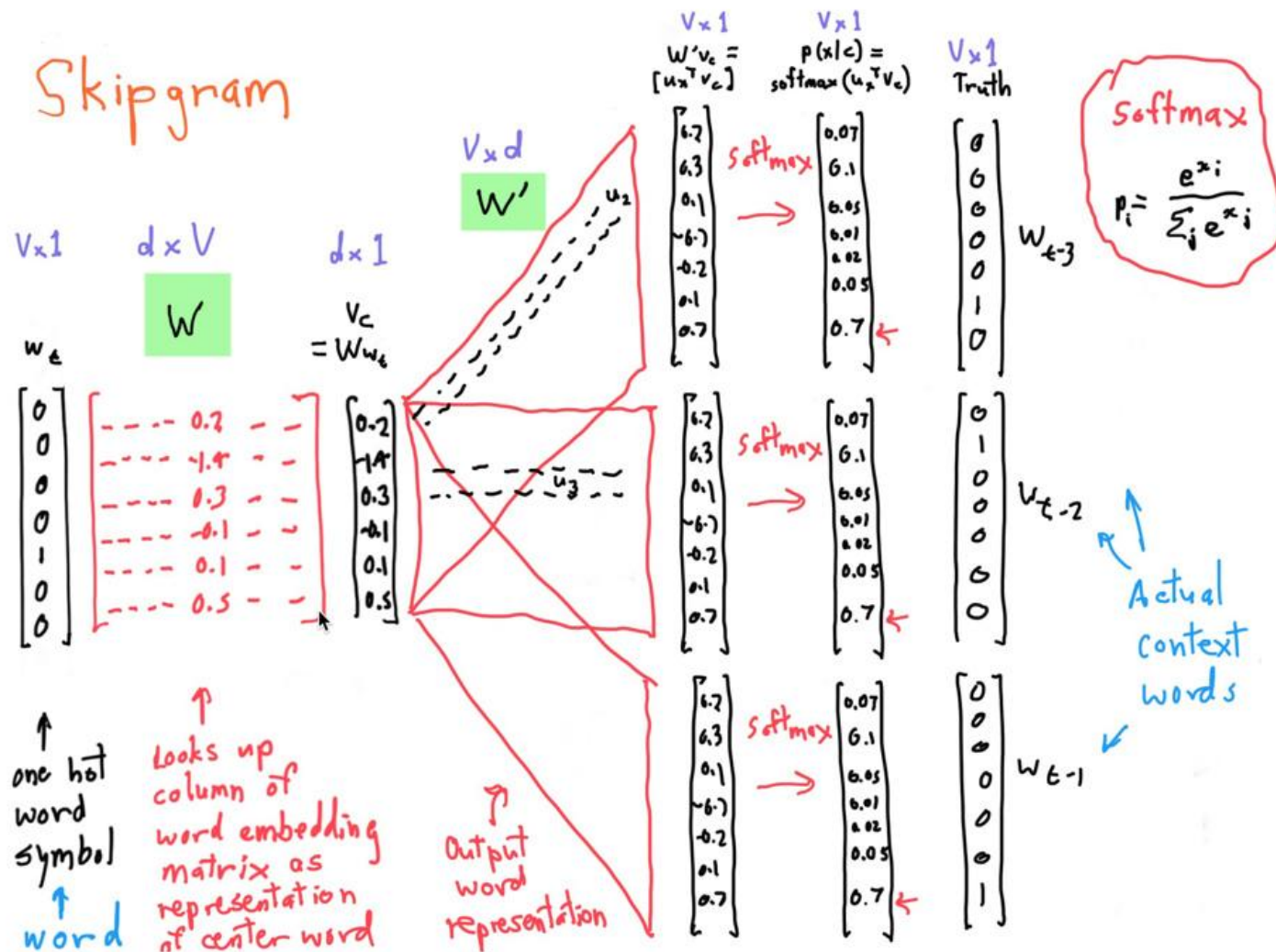
- 词向量回顾

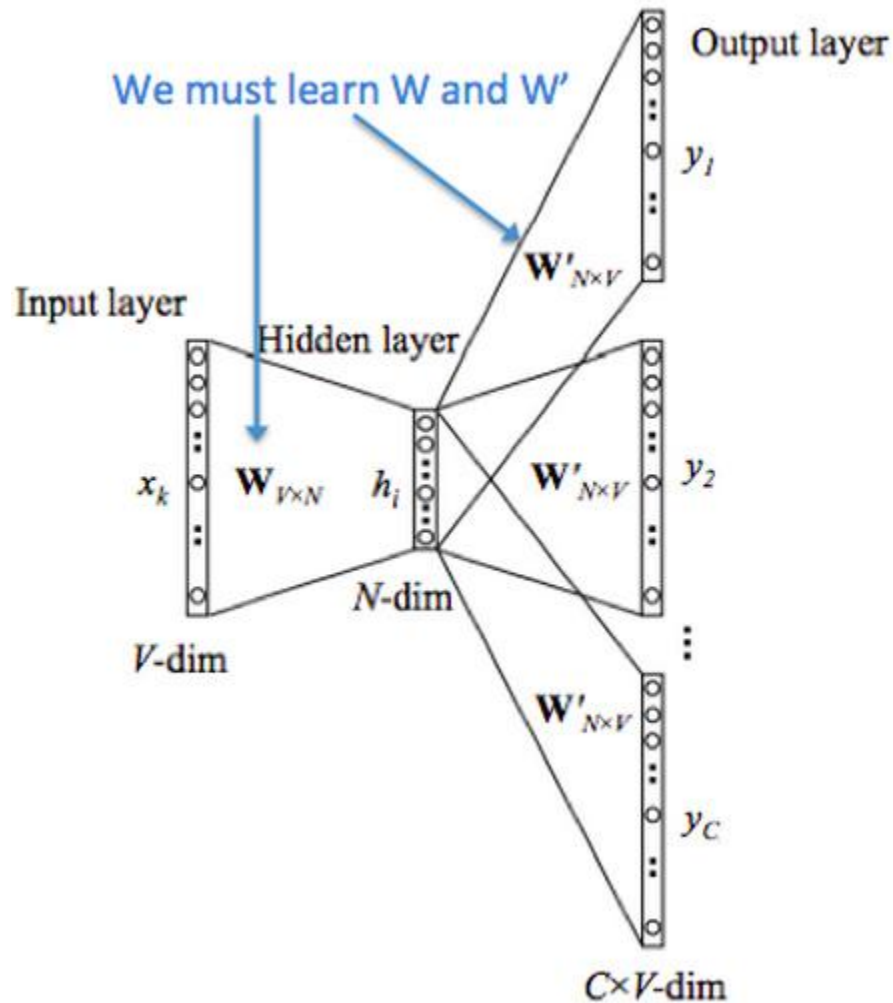
- 词向量、词嵌入、word vector、word embedding
- 理论基础：词的意思就是其上下文

government debt problems turning into banking crises as has happened in
saying that Europe needs unified banking regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

- 训练方法：填词预测



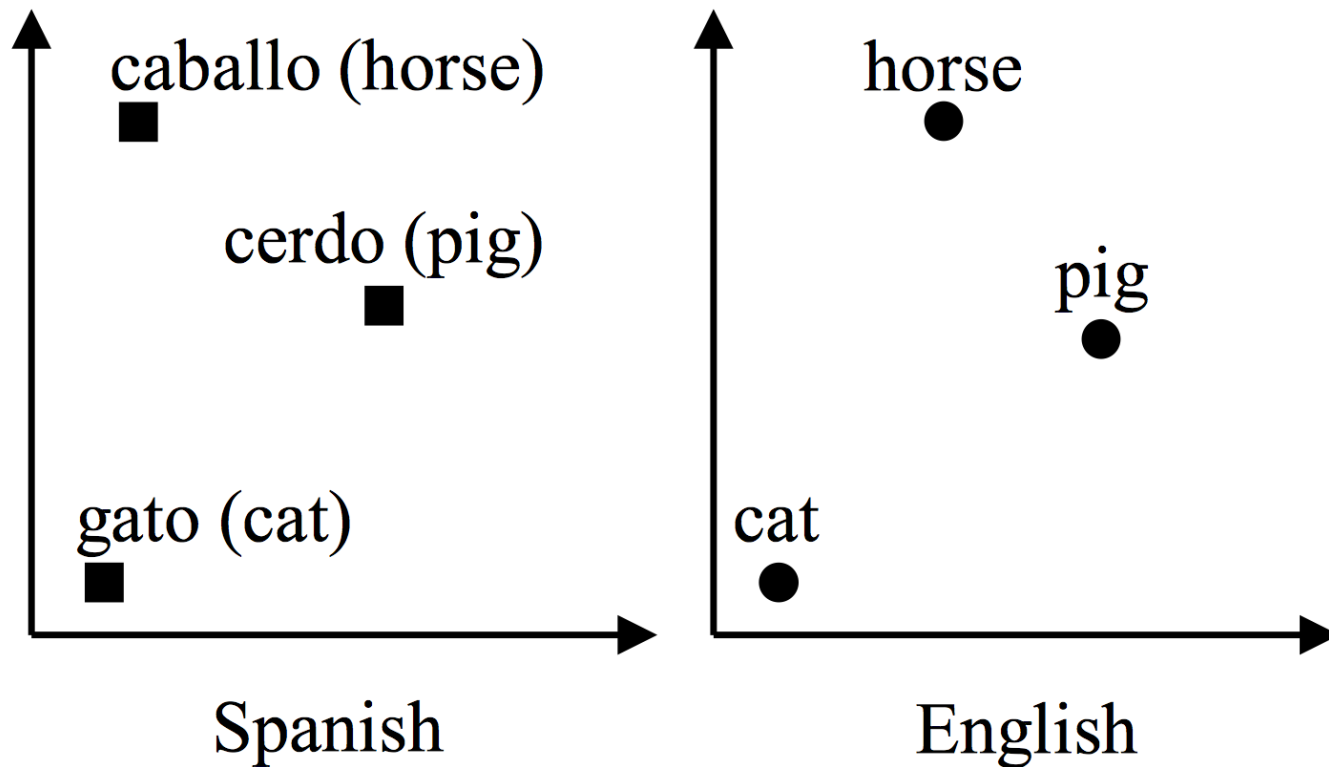


banking[0.792, -0.177, -0.107, 0.109, -0.542, ...]



背景

- 近似同构 (Approximate Isomorphism)
 - Mikolov et al., 2013a



- 跨语言NLP存在的问题
 - 单语言下训练的模型无法适用于其他语言
 - 某些小语种语料匮乏，人工标注成本过高，NLP发展受限
 - 不同语种之间存在特征空间异构
- 灵感
 - 通过探寻不同语种之间词的关系解决上述问题



跨语言词向量简介

- 跨语言词向量/Cross-lingual word embedding(CWE)
- 跨越语种障碍的词向量
- 跨语言词向量按照训练数据分类
 - 单词级别平行
 - 句子级别平行
 - 文档级别平行
 - 无需平行语料

- 单词对齐 (Word-level alignment)
 - Dog/狗
 - Cat/猫
- 单词可比 (Word-level comparable)



—Kitty/猫

- 句子对齐 (Sentence-level alignment)
 - The dog chases the cat.
 - 狗追猫。
- 句子可比 (Sentence-level comparable)



—The dog chases
the cat in the grass.

—猫在逃离狗。

- 文档可比 (Document comparable)
 - There are a lot of dogs in the park. They like to chase cats.
 - 猫咪们在放松。它们一见到狗就会逃跑。

	Parallel	Comparable
Word	Mikolov et al. (2013)	Bergsma and Van Durme (2011)
	Dinu et al. (2015)	Vulić et al. (2016)
	Lazaridou et al. (2015)	Kiela et al. (2015)
	Xing et al. (2015)	Vulić et al. (2016)
	Zhang et al. (2016)	Gouws and Søgaard (2015)
	Artexte et al. (2016)	Duong et al. (2015)
	Smith et al. (2016)	
	Vulić and Korhonen (2016)	6篇
	Artexte et al. (2017)	
	Hauer et al. (2017)	
	Mrkšić et al. (2017)	
	Faruqui and Dyer (2014)	
	Lu et al. (2015)	
	Ammar et al. (2016)	
	Xiao and Guo (2014)	
	Gouws and Søgaard (2015)	
	Duong et al. (2016)	
	Adams et al. (2017)	
	Klementiev et al. (2012)	
	Kočiský et al. (2014)	
	20篇	

	Parallel	Comparable
Sentence	Zou et al. (2013)	Calixto et al. (2017)
	Shi et al. (2015)	Gella et al. (2017)
	Gardner et al. (2015)	2篇
	Vyas and Carpuat (2016)	
	Guo et al. (2015)	
	Hermann and Blunsom (2013)	
	Hermann and Blunsom (2014)	
	Soyer et al. (2015)	
	Laully et al. (2013)	
	Chandar et al. (2014)	
	Gouws et al. (2015)	
	Luong et al. (2015)	
	Coulmance et al. (2015)	
	Pham et al. (2015)	
	Levy et al. (2017)	
	Rajendran et al. (2016)	

16篇

	Parallel	Comparable
Document		Vulić and Moens (2016) Vulić and Moens (2013) Vulić and Moens (2014) Søgaard et al. (2015) Mogadala and Rettinger (2016)

5篇



CWE训练示例

- 基于词级别对齐的平行语料训练方法
 - 使用英语语料训练一组词向量
 - 使用中文语料训练一组词向量
 - 使用词典或者词对齐学习一个两者之间的转移矩阵

$$J = \mathcal{L}_{\text{SGNS}}^s + \mathcal{L}_{\text{SGNS}}^t + \Omega_{\text{MSE}}$$

核心思想：

使用平行语料作为纽带，使对齐的词的词向量尽可能相似。

- Skip-gram negative sample(SGNS)

$$P(w_{t+j} | w_t) = \log \sigma(\tilde{\mathbf{x}}_{t+j}^\top \mathbf{x}_t) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n} \log \sigma(-\tilde{\mathbf{x}}_i^\top \mathbf{x}_t)$$

- Ranks true sentences above noisy sentences
 - Sigmoid function $\sigma(x) = 1/(1 + e^{-x})$
- Mean squared error(MSE)

$$\Omega_{\text{MSE}} = \sum_{i=1}^n \|\mathbf{W} \mathbf{x}_i^s - \mathbf{x}_i^t\|^2$$

- Max-margin hinge loss(MMHL)

$$\mathcal{L}_{\text{MMHL}} = \sum_{sent \in \mathcal{S}} \sum_{w \in V} \max(0, 1 - f(sent) + f(sent^w))$$

- Training a model to output a higher score for a correct word sequence than for an incorrect $f(\cdot)$
- $f(\cdot)$ is a neural network that outputs score of original sentences and reconstructed sentences

- 平行语料平行语料并不能包含所有的词。
 - 比如：囍在平行语料中没有对应的
- 对于某些小语种来说没有相应的平行语料。
 - 比如：克里奥尔语和阿非利堪斯语
- 如何不适用任何平行语料训练跨语言词向量？
即，非监督方法？

GAN-CWE



GAN-CWE



GAN-CWE

- Adversarial Training for Unsupervised Bilingual Lexicon Induction, Zhang et al.(2017)
- 基于生成对抗网络（GAN）训练跨语言词向量方法（一种无需平行语料的方法）

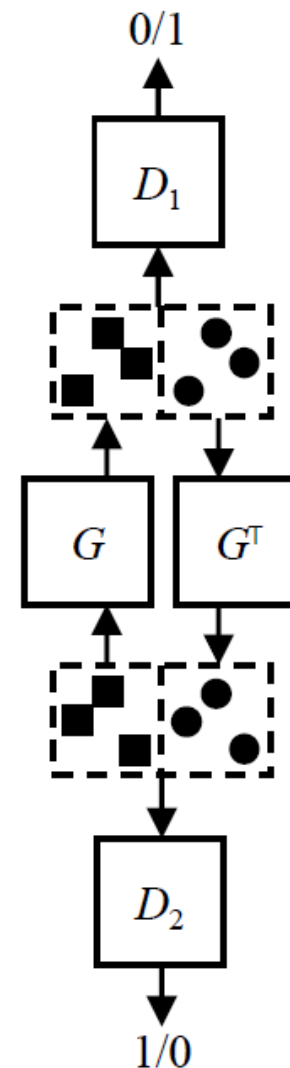
- Generative adversarial networks (Goodfellow et al., 2014).

$$\begin{aligned}
 V(D, G) &= \mathbb{E}_{y \sim p_y} [\log D(y)] + \mathbb{E}_{x \sim p_x} [\log (1 - D(G(x)))]
 \end{aligned}$$

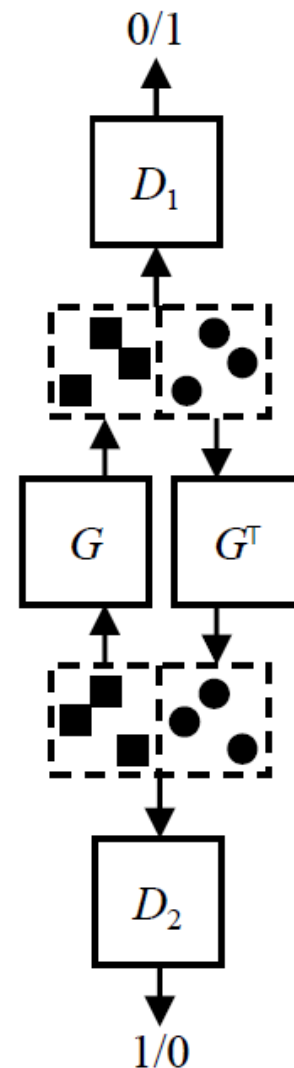
Discriminator

- Objective: Minimax game $\min_G \max_D V(D, G)$

- 生成器G
 - 目的：将源语言的词向量转换成目标语言
 - 是一个转换矩阵 $G \in \mathbb{R}^{d \times d}$
- 判别器D
 - 目的：努力分辨输入是G生成的还是真实数据
 - 是一个二分类器，单隐层前向神经网络
 - 损失函数交叉熵



- 符号表示
 - w_s 表示已训练的源语言的词向量矩阵
 - w_t 表示已训练的目标语言的词向量矩阵
- 两个生成器 G 和 G^T (转置矩阵)
$$Gw_s = w_t$$
$$G^T w_t = w_s$$
- 两个判别器 D_1 和 D_2
 - D_1 判别输入的目标语言词向量真伪
 - D_2 判别输入的源语言词向量真伪



- 两个判别器 D_1 和 D_2

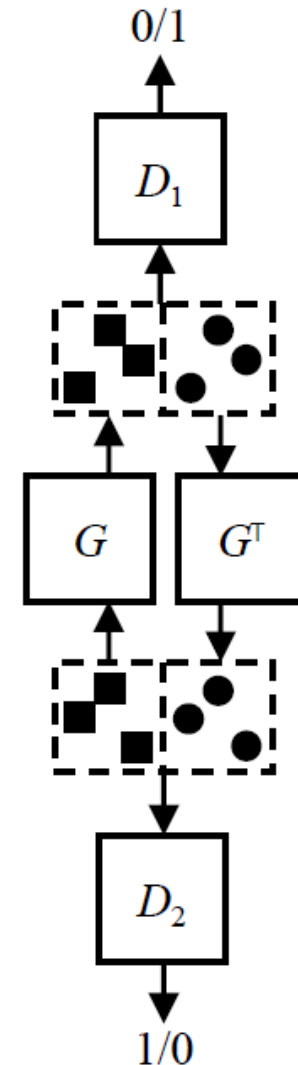
$$L_D = -\log D(y) - \log(1 - D(Gx))$$

- 两个生成器 G 和 G^T

$$L_G = -\log D_1(Gx) - \log D_2(G^T x)$$

- Loss

$$\begin{aligned} V(D, G) \\ = \mathbb{E}_{y \sim p_y} [\log D(y)] + \\ \mathbb{E}_{x \sim p_x} [\log(1 - D(G(x)))] \end{aligned}$$





Q&A

谢谢！

大成若缺，其用不弊。大盈
若冲，其用不穷。大直若屈。
大巧若拙。大辩若讷。静胜
躁，寒胜热。清静为天下正。

