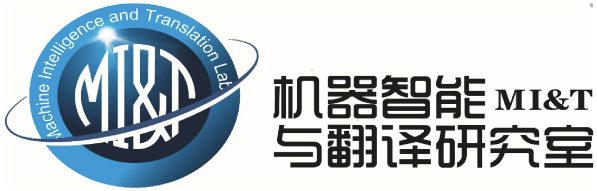


基于 CCA 的跨语言词向量模型

白雪峰

xfbai@hit-mtlab.net

March 12, 2018

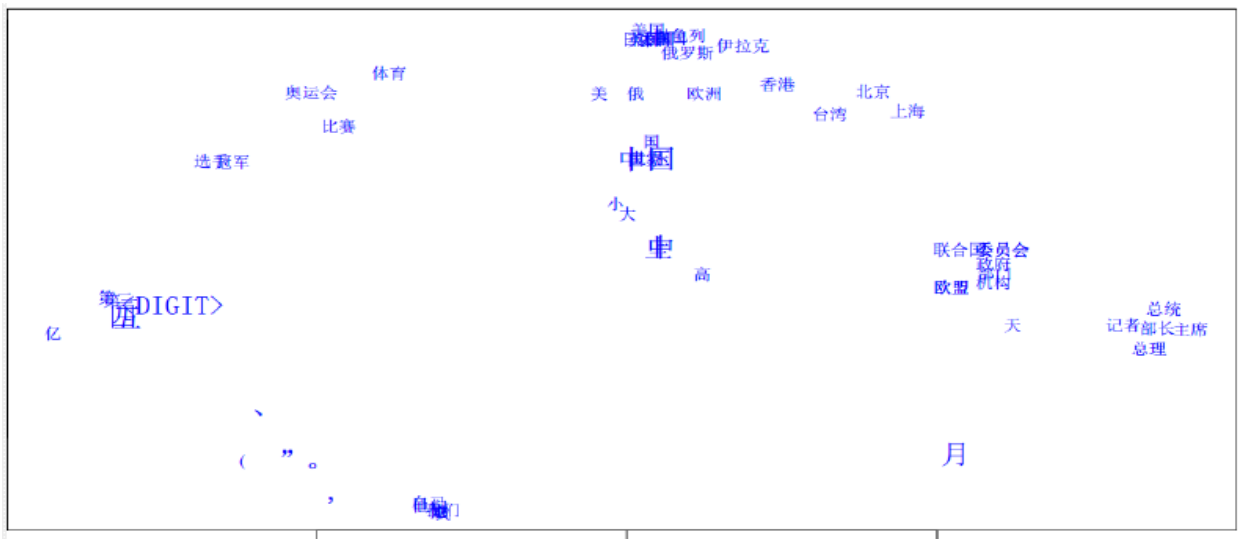
Outline

机器智能 MI&T 与翻译研究室

词向量与跨语言词向量 1

词向量: 一种词在低维、稠密的向量空间的表示方式，又常常被称为 Distributed Word Representation, Word Embedding

性质: 越相似的词距离越近、预训练好的



用途: 序列标注，文本分类，机器翻译，依存分析等

机器智能 MI&T 与翻译研究室

词向量与跨语言词向量 2

跨语言词向量： 将两种或者多种语言的词表示到一个共享的向量空间中，又称 Cross-lingual Word Embedding(Representation)

性质： 1. 同种语言中语义相近的词距离也很近 2. 多个语言中语义相近的词距离也很近

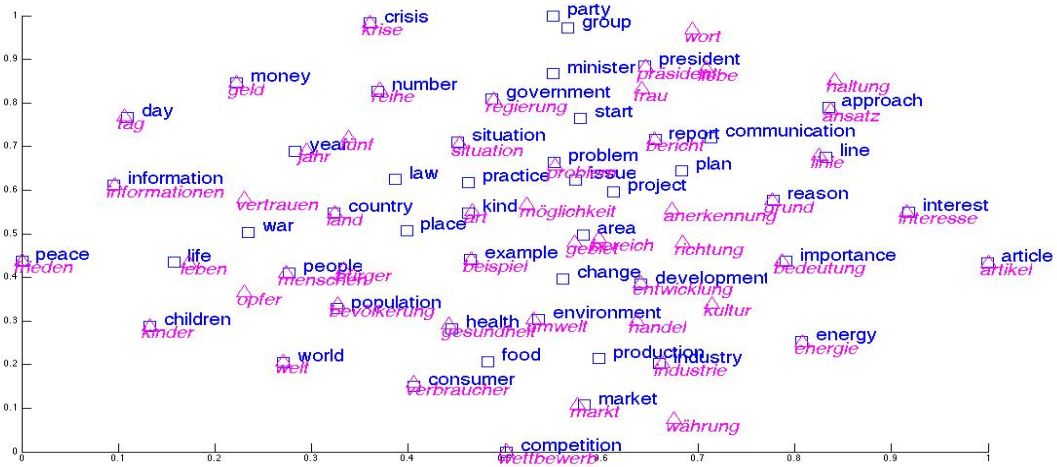


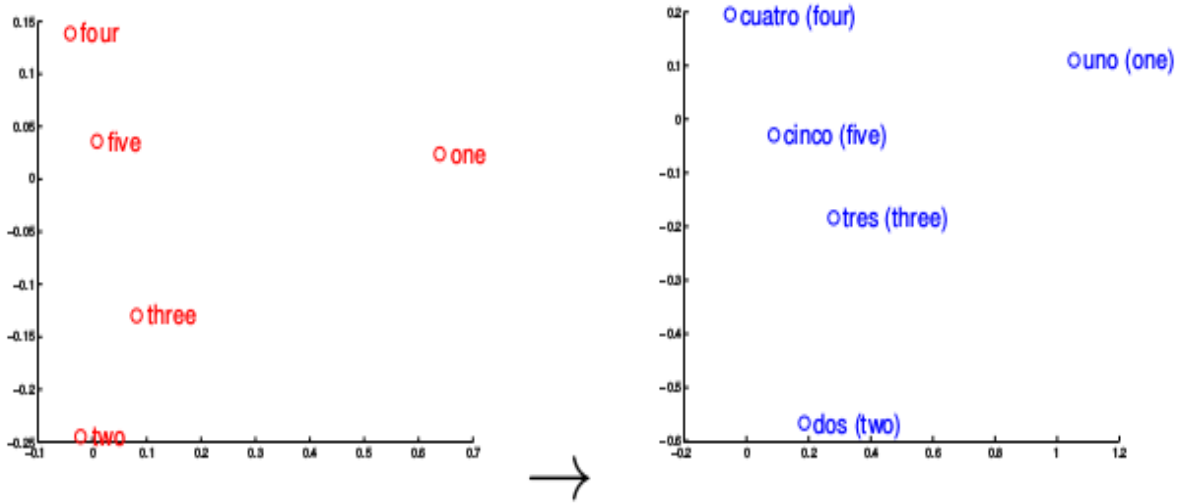
Figure: 两个语言的共享词向量空间 (Luong et al., 2015)

用途： 1. 普通单语任务的提升 2. 跨语言任务（词典抽取，机器翻译

机器智能与翻译研究室 MI&T

基于 mapping 的跨语言词向量

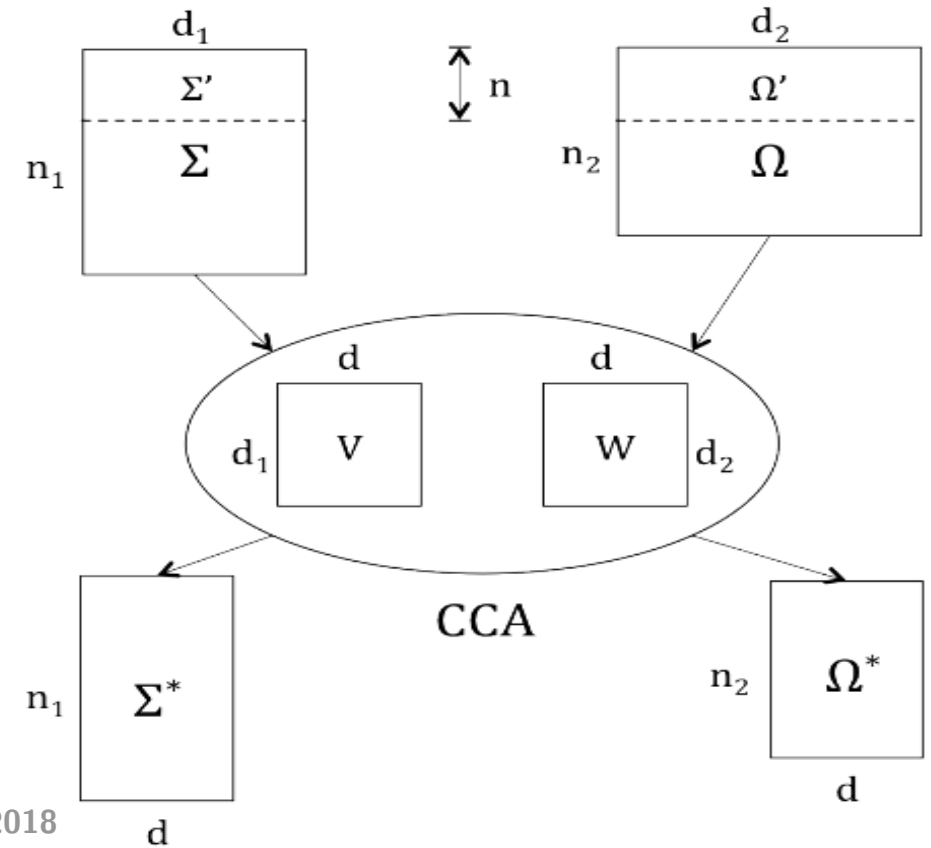
较早的发现: 两种语言词向量分布的几何相似性 (Mikolov 2013b)



后人的改进: Faruqui and Dyer 2014, Xing et al. 2015, Zhang et al. 2016, Artetxe et al. 2016, Artetxe et al. 2017

BiCCA 模型 (Faruqui and Dyer 2014) 原理简介:

已知两个语言的单语词向量矩阵 $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ 和 $\Omega \in \mathbb{R}^{n_2 \times d_2}$，拿出对齐的一部分 Σ', Ω' 来训练 CCA，得到两个转换矩阵 V, W，再用这两个转换矩阵将原始的单语词向量投影到新的空间中



机器智能 MI&T 与翻译研究室

BiCCA 形式化描述：

首先构造训练词向量矩阵 Σ', Ω' ，训练 CCA 模型，可以得到两个线性变换（投影）矩阵 V, W:

$$V, W = CCA(\Sigma', \Omega') \tag{1}$$

其中 $V \in \mathbb{R}^{d_1 \times k}$ ，$W \in \mathbb{R}^{d_1 \times k}$，$k \leq min\{d_1, d_2\}$，CCA 将随后做出介绍

然后利用训练得到的投影向量矩阵，对源词向量矩阵进行投影：

$$\Sigma^* = \Sigma V, \Omega^* = \Omega W \tag{2}$$

其中 $\Sigma^* \in \mathbb{R}^{n_1 \times k}, \Omega^* \in \mathbb{R}^{n_2 \times k}$ 是用双语知识“丰富”后得到的双语词向量。

机器智能 MI&T 与翻译研究室

CCA 原理介绍:

对于多维随机变量 $X = (X_1 X_2 .... X_m), Y = (Y_1 Y_2 .... Y_n)$，CCA 寻找两个投影向量 a, b, 使得投影后得到的结果 u, v 间的 Pearson 相关系数 $\rho(u, v)$ 最大

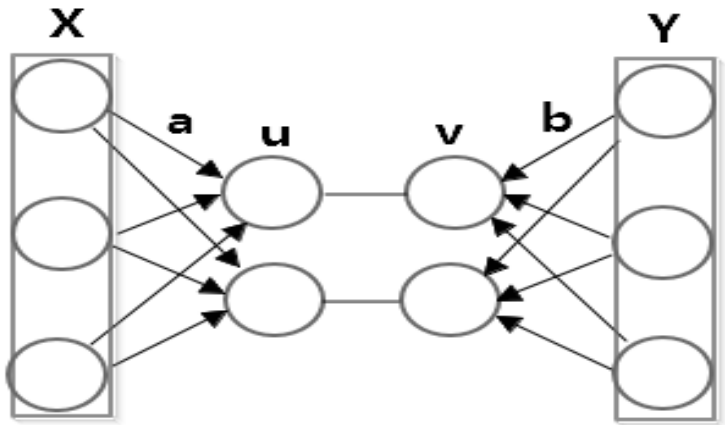


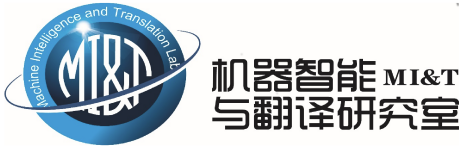
Figure: CCA 工作原理

关于 Pearson 相关系数: $\rho(X, Y) = \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$

机器智能 MI&T 与翻译研究室

CCA 数学求解：

见 Note



Thanks!