法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

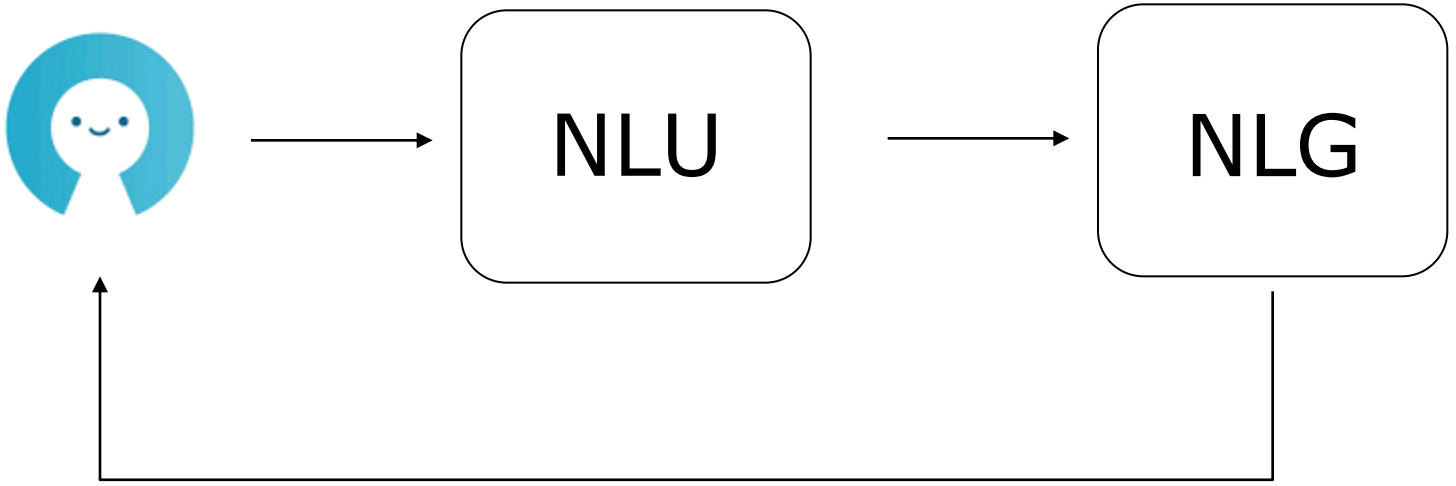
- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

第一节课：自然语言处理基础知识

- Semantics
- 词向量
- RNN

自然语言处理中的语义信息

SEMANTICS



对话系统需要理解自然语言的（抽象的）意义，即**semantics**

- While the models are able to generate reasonable responses, they are often **generic** or lack a semantic understanding of the context.

——— Training End-to-End Dialogue Systems with the Ubuntu Dialogue Corpus 2017

- 今天好玩的事是我睡午觉时女儿百无聊赖，不知怎么她想起和 Siri 聊天，聊着聊着两人竟然吵起来，还有模有样的吵了十几分钟。女儿躺在床上急得直跺脚，“你不懂文明啊？你听不懂我的话啊？你会不会采蓝莓？”Siri 总是淡定的贱答：“**这是个有趣的问题**”。女儿快气死了，我蒙在被子里快笑出声来。

——— douban 广播

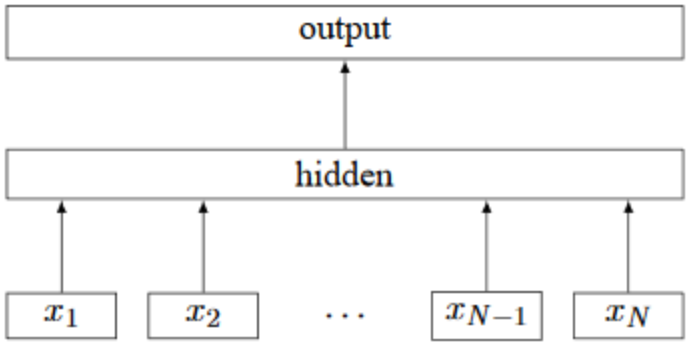
语义分析案例研究

- 自然语言的话题(topic)分析
  - 非监督学习案例:NMF提取文本话题

语义分析案例研究

自然语言的话题(topic)分析

监督学习案例：使用fasttext进行文本分类



Bag of Tricks for Efficient Text Classification （2016）

Dataset	Classes
AG's News	4
Sogou News	5
DBPedia	14
Yelp Review Polarity	2
Yelp Review Full	5
Yahoo! Answers	10
Amazon Review Full	5
Amazon Review Polarity	2

部分有代表性的自然语言**分类数据集**
Character-level Convolutional Networks for Text Classification （2015）

语义分析案例研究

- 自然语言的意图(intent)分析
  - 监督学习案例：使用fasttext进行意图分类

语义分析案例

- Semantic network
  - 案例研究：ConceptNet, An open, multilingual knowledge graph

语义分析案例

- 语义相似度
  - 案例研究：SemEval2017

SemEval2017：语义相似度

- 能否准确理解语言表示的意义和细微区别？

Tasks

We are pleased to announce the following exciting tasks in SemEval-2017:

Semantic comparison for words and texts

- Task 1: Semantic Textual Similarity
- Task 2: Multilingual and Cross-lingual Semantic Word Similarity
- Task 3: Community Question Answering

Detecting sentiment, humor, and truth

- Task 4: Sentiment Analysis in Twitter
- Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News
- Task 6: #HashtagWars: Learning a Sense of Humor
- Task 7: Detection and Interpretation of English Puns
- Task 8: RumourEval: Determining rumour veracity and support for rumours

Parsing semantic structures

- Task 9: Abstract Meaning Representation Parsing and Generation
- Task 10: Extracting Keyphrases and Relations from Scientific Publications
- Task 11: End-User Development using Natural Language
- Task 12: Clinical TempEval

单词的语义

- SemEval-2017 任务-2：单词的语意相似和相关度
  - *sunset* - *string*: 0.05
  - *computer science* - *mathematics*: 3.1
  - *automobile* - *car*: 3.82

句子的语义

- SemEval-2017 任务-1：句子的语意相似度
- e.g.
  - I did this **one time** as well. # 这个事我也做过**一次**
  - I have this **habit** as well. #我也有这个习惯
- e.g.
  - How can I connect additional wires to a receptacle? # 如何将额外的电线连接到插座？
  - How do I connect the wires to this USB receptacle? # 如何将电线连接到此**USB**插座？

段落的语义

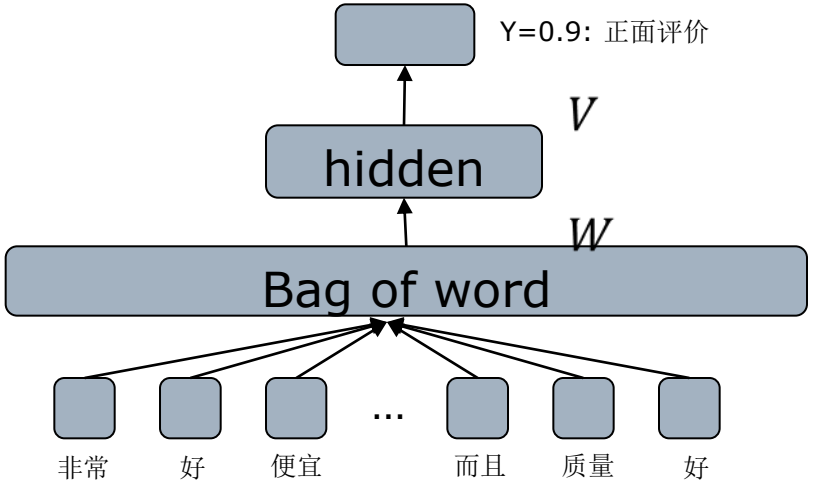
- SemEval-2017 任务-3：段落的语意相关度，可用于QA，客服机器人
  - *Q:Can I drive with an **Australian driver's license** in **Qatar**?*
  - *Q1: How long can i **drive in Qatar** with my international driver's permit before I'm forced to **change my Australian license to a Qatari one**? When I do change over to a Qatar license do I actually lose my Australian license? I'd prefer to keep it if possible...*
  - *Comment to Q1: Thank you for your email! With regards to your query below, **a foreigner is valid to drive in Doha with the following conditions**: **Foreign driver** with his country valid driving license **allowed driving only for one week** from entry date Foreign driver with international valid driving license allowed driving for 6 months from entry date Foreign driver with GCC driving license allowed driving for 3 months from entry"*
    As an Aussie your driving licence should be transferable to a Qatar one with only the eyetest (temporary, then permanent once RP sorted).

自然语言处理中的词向量方法

WORD EMBEDDING

从one-hot到词向量

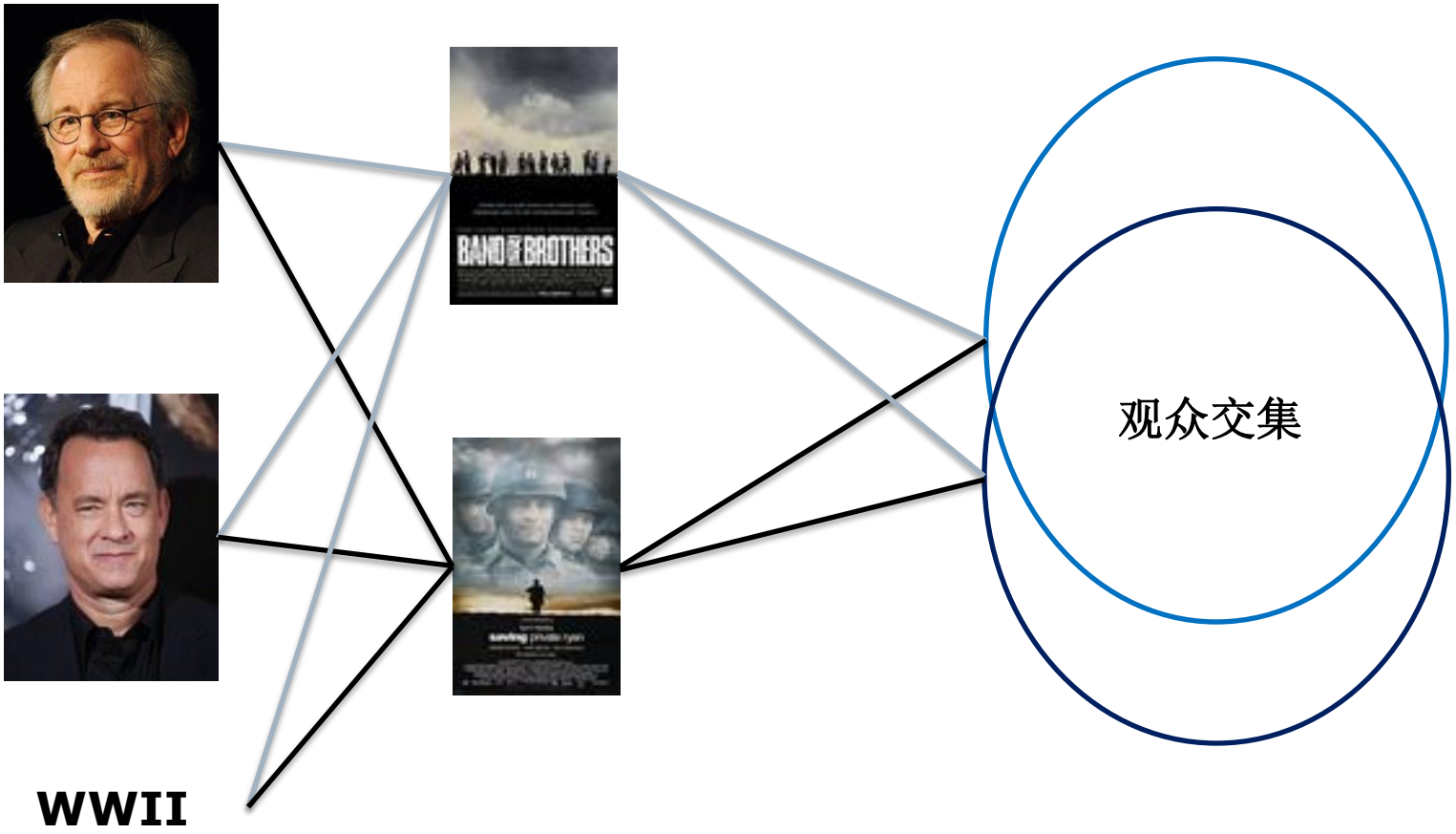
- 在使用NN的分类问题中，每个单词已经从离散的ID转换成为低维度的连续向量
  - $x = [0, 0, 1, 2, 0, 0, \ldots 1, 0]$
  - $h = x \cdot W = W_2 + 2 \cdot W_3 + \ldots + 1 \cdot W_{9999} \in R^{128}$
  - $y = \sigma(h \cdot V) \in [0,1]$
- 文本分析 (e.g. 情感分析 sentiment analysis；话题分类 topic classification) 任务中的词向量
  - 词向量侧重也局限于分类任务
  - 监督学习数据有限



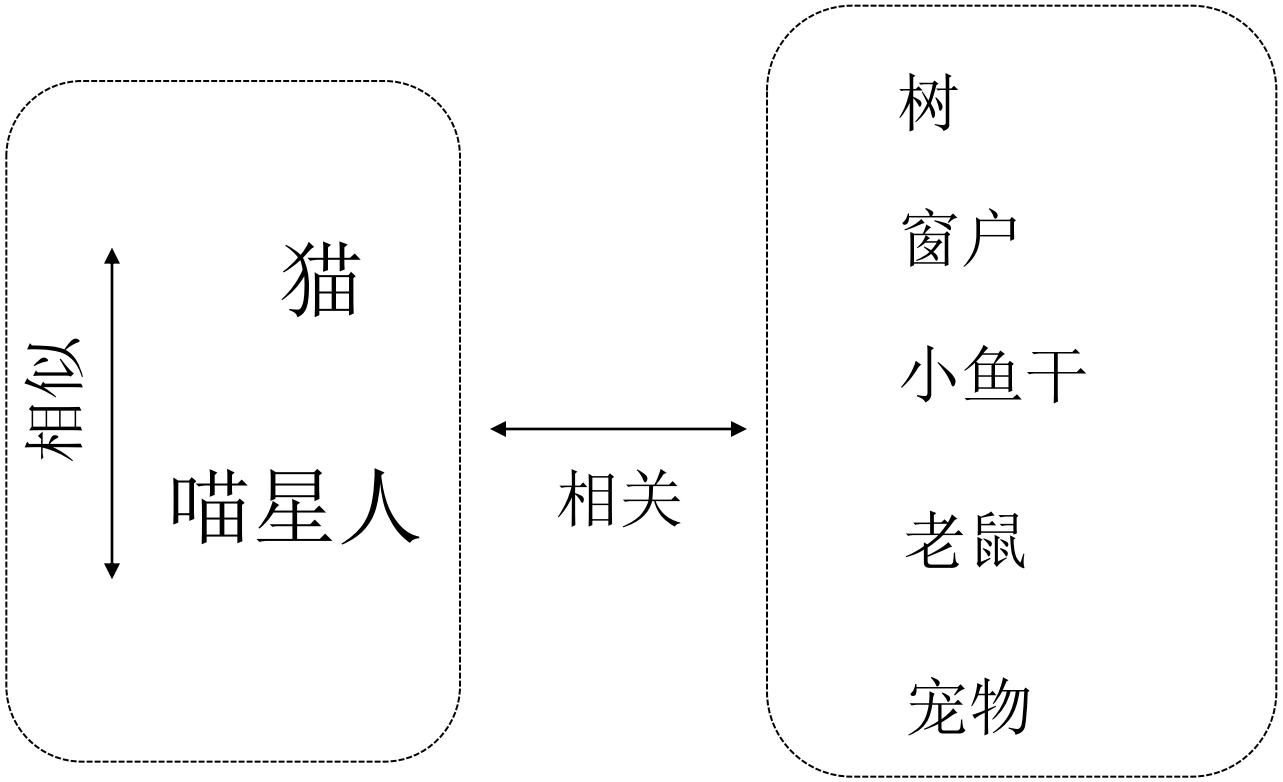
语意词向量

- 语意词向量(semantic word embedding):
  - 我们希望词向量包含一些语义信息
  - e.g. 椅子会让我们想到如下意义:
    - 家具的一种；
    - 用来坐；
    - 和沙发有相似的功能；
    - 常见材料是木头，塑料，金属，纤维；
    - 椅子品牌
    - 等等

词的语意相关和语意相似

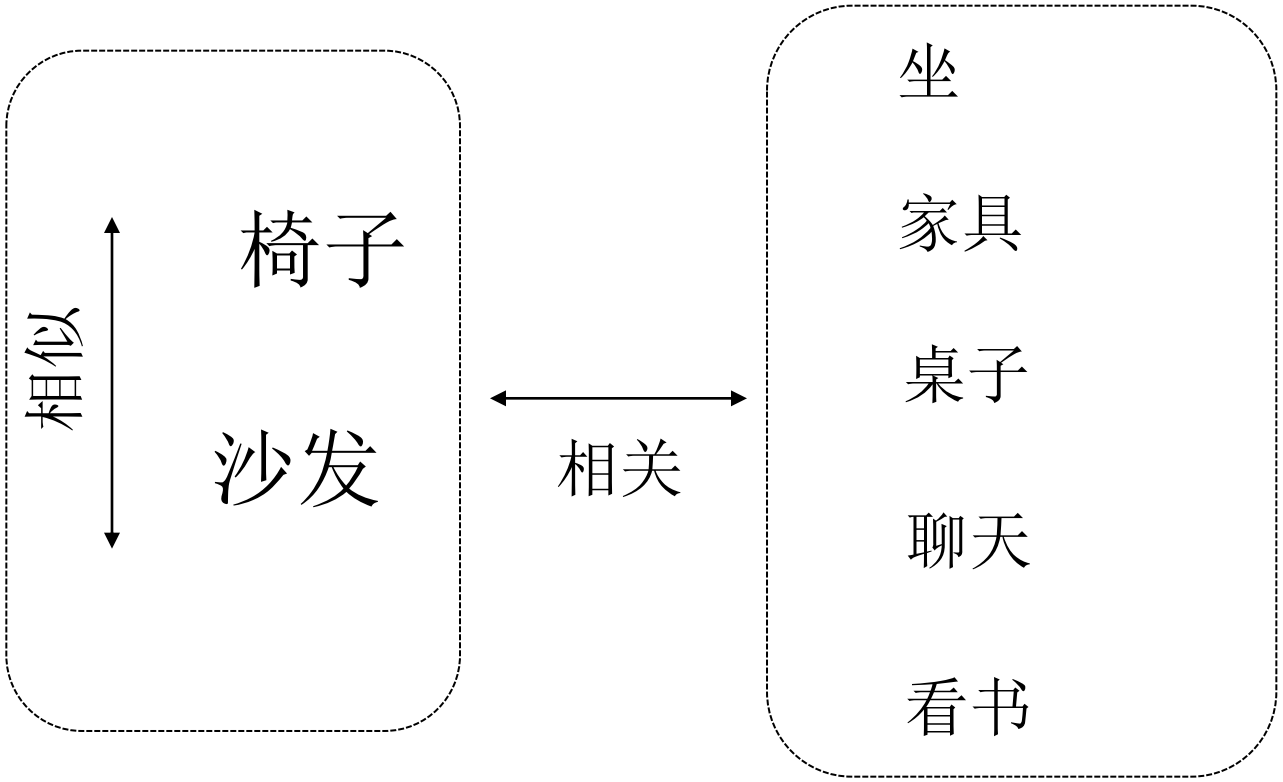


词的语意相关和语意相似



如果两个词和相近的一组词语相关(relatedness)，
那么这两个词的语意很可能是相近(similarity)的

词的语意相关和语意相似



如果两个词和相近的一组词语相关(relatedness)，
那么这两个词的语意很可能是相近(similarity)的

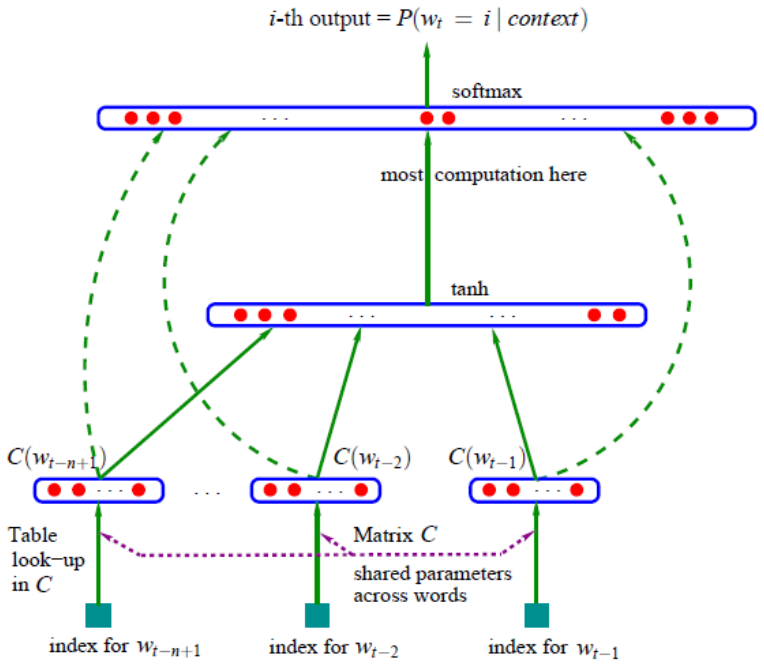
统计语言模型

- $P(w)$
- $P(w_t|w_{t-1})$
- $P(w_t|w_{t-1}, w_{t-2}, w_{t-3},)$
- $P(w_{t-1}, w_t)$
- $P(w_{t-3}\ w_{t-2}\ w_{t-1}\ w_t)$

使用统计语言模型量化词和词的相关程度

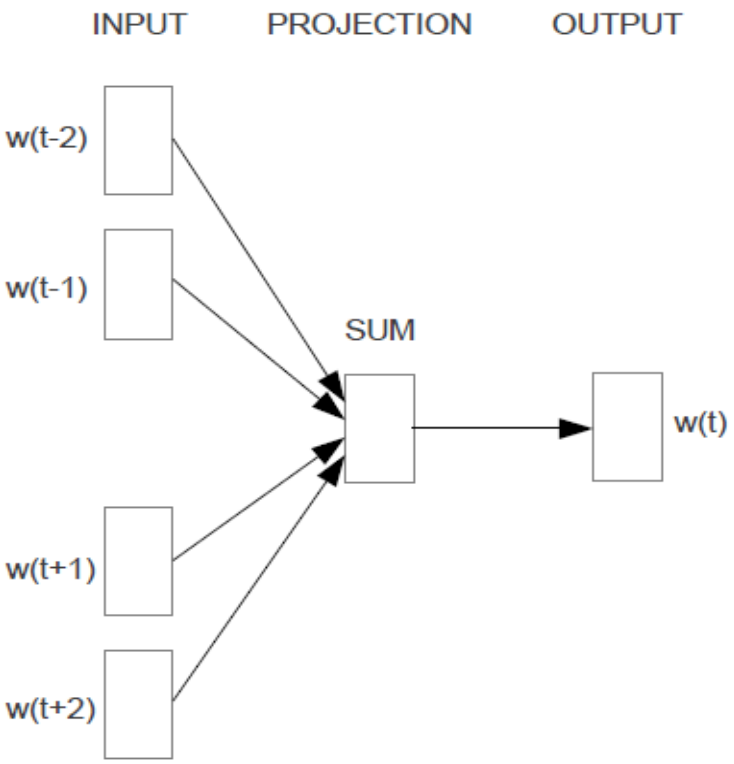
非监督学习训练语言模型

- 给定语境(context), 预测观测到一个词语的概率
  - $P(w_t|w_{t-1}\ ,\ w_{t-2}\ldots w_{t-n+1})$: semantically informative
- CBOW:
  - 给定语境，模型预测观测到各个词语的概率
  - 预测到实际词语的概率大于其他词语的概率
  - 例如：中国 男足 很 X
    - P(X=烂)>P(X=好)
    - P(X=烂)>P(X=椅子)
    - P(X=水)>P(X=强)

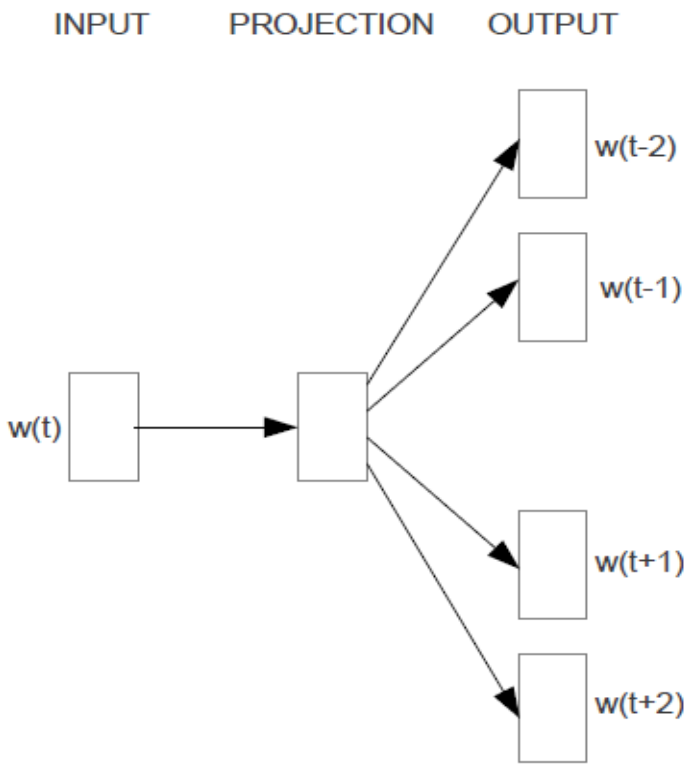


A Neural Probabilistic Language Model. Bengio et al. 2003

非监督学习训练语言模型



CBOW



Skip-gram

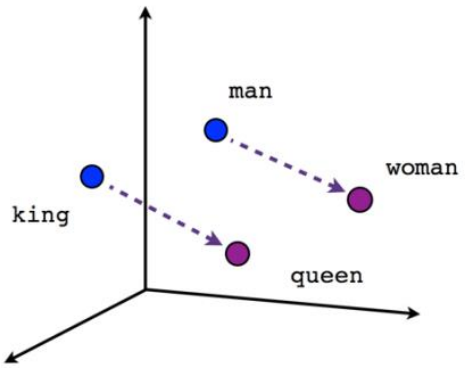
tip：{One-gram;
N-gram;
skip-gram}

代码演示

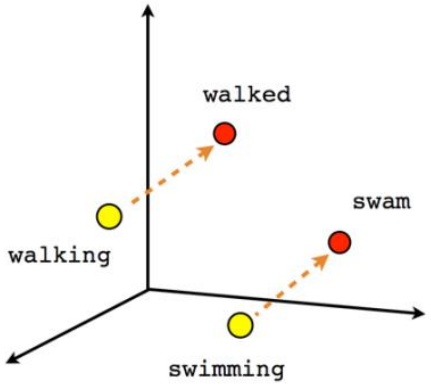
- 使用cbow和skip gram训练词向量
  - 自己动手学习词向量
  - 使用gensim的word2vec和fasttext训练词向量

代码演示

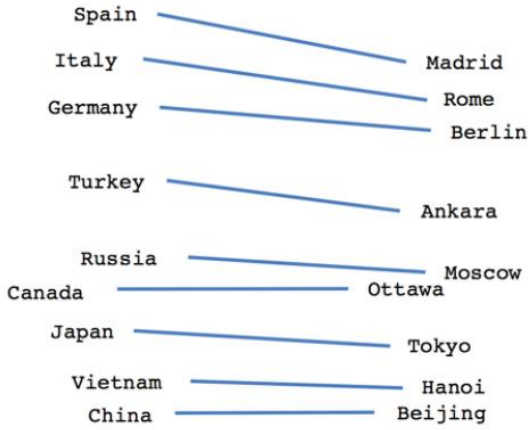
- Google word2vec可视化



Male-Female



Verb tense



Country-Capital

讨论

- 相对于cbow，Skip-gram对于生僻词有更好的效果
  - Yesterday was really […] day
    - beautiful, nice, good
    - delightful

讨论

- 如何处理多义词的embedding?
- 如何识别和学习词组的向量？
- 如何处理未曾出现的新词？

讨论

- 通过非监督学习训练词向量需要多少数据？
- Glove:
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab)
  - Common Crawl (860B tokens, 2.2M vocab)
- Word2vec:
  - First billion characters from wikipedia (use the pre-processing perl script from the bottom of Matt Mahoney's page)
  - Latest Wikipedia dump Use the same script as above to obtain clean text. Should be more than 3 billion words.
  - WMT11 site: text data for several languages (duplicate sentences should be removed before training the models)
  - Dataset from "One Billion Word Language Modeling Benchmark" Almost 1B words, already pre-processed text.
  - UMBC webbase corpus Around 3 billion words, more info here. Needs further processing (mainly tokenization).
  - Text data from more languages can be obtained at statmt.org and in the Polyglot project.

监督学习方法与词向量

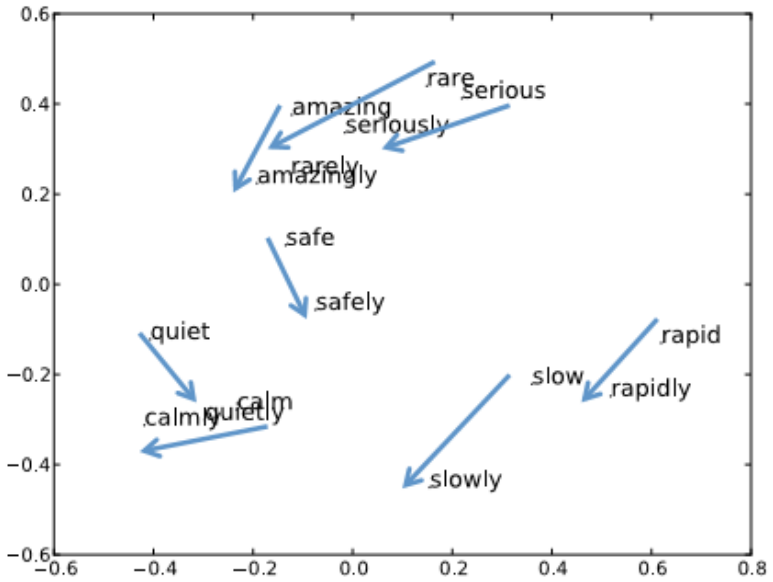
- Cbow, skip-gram以及Glove方法基于词语之间“共存”这种关联学习语意信息
  - 受数据集质量和数量影响
  - 复杂的语意关联
  - 具体应用领域的词向量

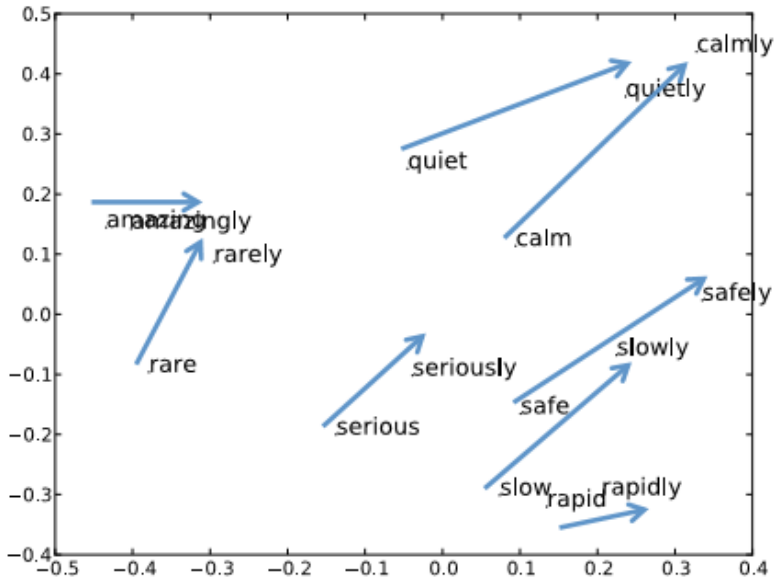
- Semantics derived automatically from language corpora contain human-like biases, by Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, published in *Science*
- Man is to Computer Programmer as Woman is to Homemaker? Debiasing word embeddings, by Tolga Bolukbasi et al., working with Microsoft Research
- Scientists Taught A Robot Language. It Immediately Turned Racist., written by Nidhi Subbaraman for that paragon of science reporting, *BuzzFeed*

监督学习方法与词向量

- 自然语言处理领域已经有很多语意方面的领域知识 (valuable information that is contained in semantic lexicons such as **WordNet**, **FrameNet**, and the **Paraphrase Database**)
  - Retrofit
  - Conceptnet embedding

监督学习方法与词向量


rare
serious
amazing
seriously
rarely
amazingly
safe
safely
quiet
calm
quietly
calmly
slow
rapid
rapidly
slowly


calmly
quietly
quiet
calm
rarely
safely
slowly
seriously
safe
serious
slow
rapid
rapidly
rare

自然语言处理中的深度学习方法

RECURRENT NN

动机

- 自然语言
  - 高维
    - 上万个单词，汉字
  - 长度不一
    - 男足很弱；中国男足很弱
    - 它便宜一半的话，我会觉得它是一个好商品
  - 词语顺序
    - X喜欢Y,但是Y讨厌X

代码演示

- 使用简单的合成(synthetic)数据训练RNN模型

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop