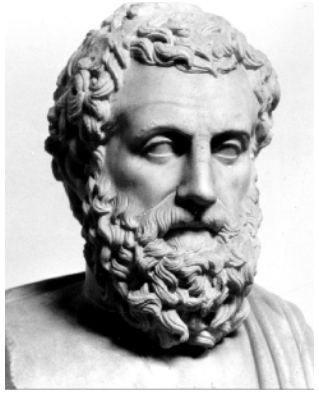
Artificial Intelligence For NLP Lesson-04



人工智能与自然语言处理课程组

2019.April. 20

Some References

SICP, *Structure and Interpretation of Computer Programming*.《计算机程序设计的解释和构造》

Introduction to Algorithms《算法导论》

Artificial Intelligence A Modern Approach (3rd Edition)《人工智能：一种现代方法》

Code Complete 2《代码大全》

Programming Pearls《编程珠玑》

Deep Learning,《深度学习》

《黑客与画家》

《数学之美》， 吴军

Fluent Python

Hands on Tensorflow

Conference: NIPS*， *ICML*， *ICLR*， *ACL*， *AAAI

AI for NLP PATHROAD

Lesson-01 BSF, Syntax Tree

Lesson-02 Probability Model

Lesson-03, Machine Learning, Dynamic Programming

Lesson-04/05, Basic NLP Methods

Lesson-06 Model, Validation, Test

Logistic Regression, Linear

Lesson-07 KNN,SVM,Bayes

Lesson-09 Unsupervised Leanring

Lesson-10 Word Embedding Advanced

Lesson-11 Backprogation , Softmax, Crossentropy

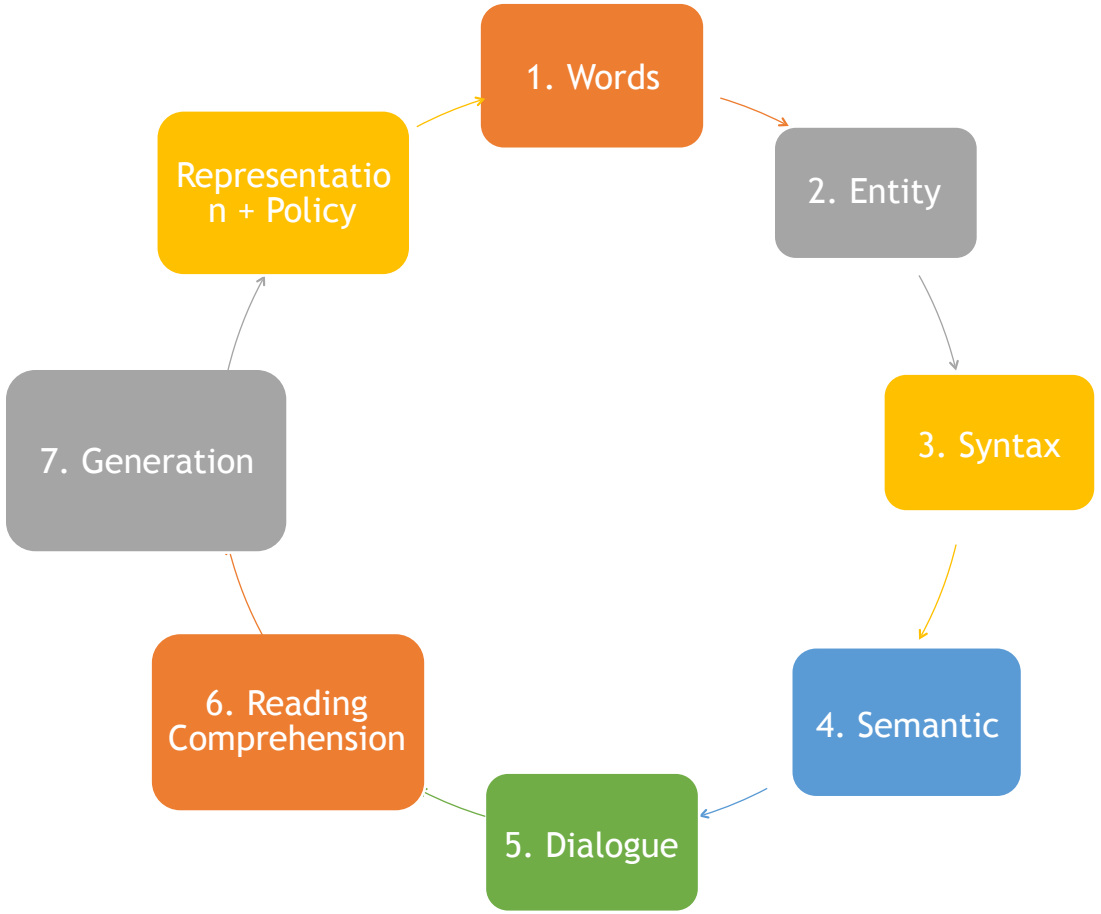
Lesson-12 Dense Neural Netwokrs

Lesson-13 Tensorflow, Keras

Lesson-14 Recurrent Neural Networks, LSTM, GRU

Lesson-15 Convolutional Neural Networks

Lesson-16 Sequence2Sequence , Attention, Transformer, BERT



What NLP concerns

Why NLP?

- There are so many sub-fields of *Artificial Intelligence*
  - Computer Vision
  - Predication and Data Mining
  - Optimization
  - Self Driving
  - Recommend System
  - etc

Why NLP?

Information Chaos

Unstructured

Discrete

Unconventional

OOV

Why NLP?

Language is the representation of mind.

The most “classical” field of AI

The most “immature” field of AI

The most “sophistical” field of AI

- 1. How to flatten the parenthesizes:
  - ((1, 2), (3, 4), (5, 6), (((8, 9), 10), 11))
- 2. Remove the duplication
  - [1, 2, 2, 2, 1, 1, 2, 1, 1, 2, 4, 4, 5, 4, 5, 6, 6, 7, 9, 10, 11, 4, 5, 6, 6, 7, 9, 10, 11]

Some utilities for NLP

- 1. Similarity: Edit Distance, Word Distance
- 2. Key words
- 3. Name Entity Recognition
- 4. Dependency Parsing
- 5. Topic Model

Similarity: Word2Vec

近日，有网友质疑“四川省达州市市政工程管理处”官方微博账号长期发布一些与其身份不符的内容，形同虚设。11月30日，四川新闻网记者向达州市市政工程管理处了解相关情况，达州市市政工程管理处相关负责人表示，他们已经看到了网友所反应的情况，但对于该微博究竟是谁注册，又是谁在发布信息，他们还不太清楚情况和原因，对于该微博发布的一些信息他们也觉得很奇怪。

- Edit Distance
  - Looks Like but not means same
  - Talk: What’s the advantages and disadvantages of Edit Distance?

Word Embedding

1

1. What is embedding

2

2. Why we need word embedding

3

3. How do we implement word embedding;

- One-hot; PCA, SVD

4

4. What is the word embedding

Representation

Why do we need to represent words and text?

- The only could be computing is number for computers.

How to represent words?

- ASCII: a, b, c : 97, 98, "abc" : 979899
- Unicode:
- etc: UTF-8

Why do not represent words as numbers?

- ‘你今天真好看’: \u12412, \u32121, \u5542,\u189301 => [12412, 32121, 5542, 189301]
- Is words numeric or categorical?
- How can we represent it as one-hot?

- Are words merely categorical?
  - () A: Yes, there are no mathematical relations between different words;
  - () B: No, Some words are much ‘closer’ than some others.

If we treat we words as mere categorical

[0, 0, 0, 1]

[0, 1, 0, 0]

[1, 0, 0, 0]

[0, 0, 1, 0]

Cannot keep similarity: $(v1 + v2) \cdot v3 == v1 \cdot v3$

- The problem of representing words as one-hot.
  - [ ] A. Cannot represent the relation of ‘similar' words and some ‘not similar words’;
  - [ ] B. The are space consuming;
  - [ ] C. It’s difficult to get the values;
  - [ ] D. If we add new words, we need to re-calculate all the words;

- 1. If we do PCA of this of this one-hot matrix;
  - What is PCA
  - How do we get PCA: https://www.wikiwand.com/en/Principal_component_analysis
- 2. If we do SVD of this one-hot matrix;
  - What is SVD
  - How do we get SVD: *https://www.wikiwand.com/en/Singular-value_decomposition*

- Problems of PCA and SVD:
  - [ ] A. When adding new words, need recalculate all the words;
  - [ ] B. It’s computing consuming.
  - [ ] C. This algorithm it’s hard to implement.
  - [ ] D. Cannot Solving Polyseme(多义词)

What features do our vectors need ?

1. SPACE ECONOMICAL

2. ADAPTIVELY UPDATE

3. SEMANTIC SIMILARITY

BUT HOW?

“You shall know a word by the company it keeps”

“每天早上我都要去门口那家早餐店吃___”

“昨天早上我是在家做的___，味道还不错”

- Assuming a density vector <v1, v2, .. vN>
- If occurrence position of some words always is same of the position of this vector;
- We get the similar vector;

Embedding

- What is embedding?
  - Graph embedding, node embedding, graph embedding, etc
  - Importance of representation.

Assuming a vector v1, depend on some linear project, to a new space v1', how can we evaluate this v1' is good or bad?

We test if the co-occurrence feature could keep.

v1 —> v1' —> v1'M ==> (word1, word2, word3)

v2 —> v1' —> v2'M ==> (word1, word2, word4)

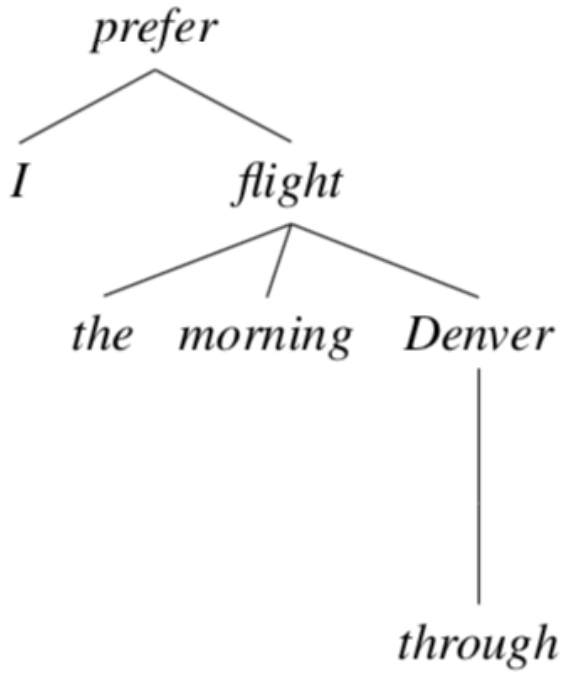
v1 is similar as v2

- *The more detail we will talk on lecture 10 or lecture 11*

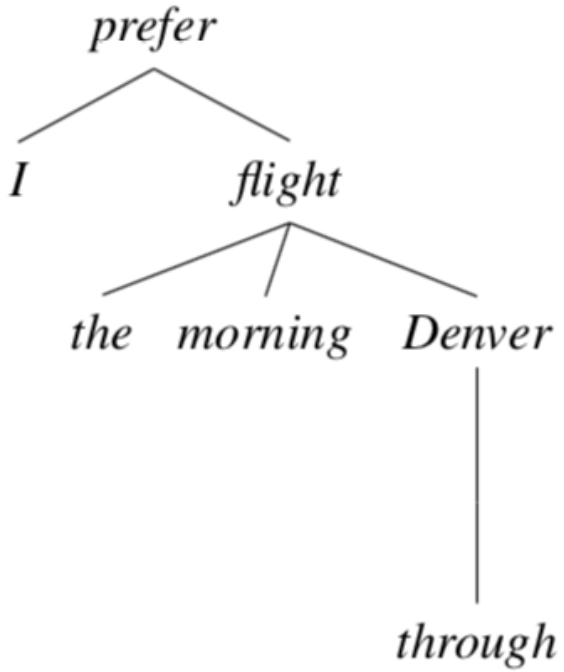

Gensim is our friend
Jieba is our friend
Wikipedia is our friend

- TF-IDF Keywords
- Words Cloud
- Based on Graph and Word Embedding
- Text-Rank (We will talk in future)
- Based on Machine Learning(We will talk in future)

- (on line coding presentation)

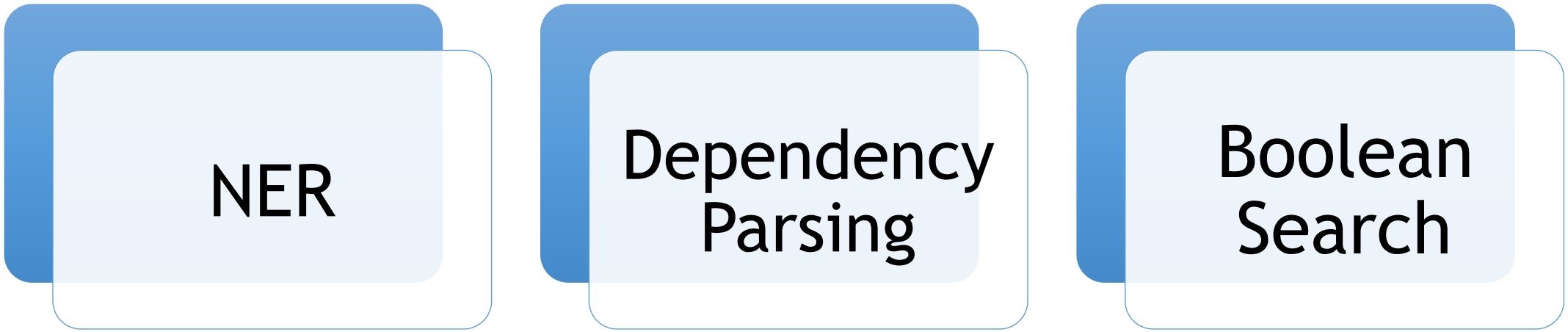


Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction



Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.
	We *booked* her the first **flight** to Miami.
IOBJ	We *booked* **her** the flight to Miami.
NMOD	We took the **morning** *flight*.
AMOD	Book the **cheapest** *flight*.
NUMMOD	Before the storm JetBlue canceled **1000** *flights*.
APPOS	*United*, a **unit** of UAL, matched the fares.
DET	**The** *flight* was canceled.
	Which *flight* was delayed?
CONJ	We *flew* to Denver and **drove** to Steamboat.
CC	We flew to Denver **and** *drove* to Steamboat.
CASE	Book the flight **through** *Houston*.

Natural Language Processing: Chapter-14, this book is saved on our github repository


NER
Dependency Parsing
Boolean Search

- Extracting the Person's talk from New Corpus.
  - Dataset: News Corpus
  - Toolset: Pandas, Matplotlib, Numpy, Jieba, Gensim
  - Applicatoin:
    - Trending Analysis
    - Knowledge Graph
    - Semantic Analysis
    - Risk Predication

"我本来是想通过微博向他们反映一下问题，结果发现他们的微博居然全是和身份不相符的内容，甚至还有些广告，真的有点可笑。"11月29日，有网友通过达州本地论坛发表《达州市市政工程管理处微博形同虚设为何无更新》贴文。今日上午，四川新闻网记者联系到了发帖人杜某某，杜某某称他今年上高三的表弟，11月28日晚下自习路过达一中附近，因疑似市政工程安全问题导致腿部受伤，缝合了8针，本想通过微博平台向达州市市政工程管理处反应一下，结果发现其官方微博发布的信息都是一些与其身份不相符的信息。

四川新闻网记者在新浪微博平台上看到，微博号为"@达州市市政工程管理处"其官方微博认证为"四川省达州市市政工程管理处官方微博"，从2013年11月5日到2018年5月31日共发布微博258条，关注度180，粉丝数356。该微博曾在2014年、2015年发布、点赞过5条与其身份相符的信息(其中发布信息4条，点赞1条)，发布信息主要内容大致为介绍达州市市政工程管理处成立的时间、职责职能范围以及办公地点等;唯一1条点赞出现在2015年，有网友反映达州惊现"趵突泉"，该微博为其点赞。同时，今年8月、10月相继有网友@达州市市政工程管理处，欲通过微博向达州市市政工程管理处反映相关情况，但该微博均无回应。

11月30日，四川新闻网记者联系上了达州市市政工程管理处相关负责人，该负责人表示，他们已经从网上了解到了网友反映的情况。但对于该微博究竟是谁注册，又是谁在发布信息，他们还不太清楚情况和原因，对于该微博发布的一些信息他们也觉得很奇怪。"我是2016年才到的该岗位，之后我们也一直没有注册和运营过官方微博。"该负责人表示，他们将尽快与平台联系，核实相关情况。