HMM&NER (realization of simplicity)

superhy

SCUT

2016.6

Catalog

- NER
- HMM
- Solutions
- Expectation

Named Entity Recognition

- Recognize the Entities that has a specific meaning in the text

- Usually has:
  - Names of persons
  - Organizations
  - Locations

Some example (In Chinese)

- <START:Person>李彦宏<END>是<START:Organizations>百度<END>的创始人
- Some simple situation we can use POS tagging directly ~~~
- 李彦宏/nr 是/v 百度/n 的/uj 创始人/n

- Some situation ~~~
- 小明/nr 硕士/n 毕业/n 于/p 中国科学院/nt 计算所/n
- 著名/a 的/u 北京/ns 协和/nz 医学院/n

Problem representation for computer

- NER can be represented as a sequence prediction problem

- 小明/nr 硕士/n 毕业/n 于/p 中国科学院/nt 计算所/n
- pre pre pre pre org org

- 著名/a 的/u 北京/ns 协和/nz 医学院/n
- ? ? ? ? ?
- Prediction “?” : is org or not?

Hidden Markov Model

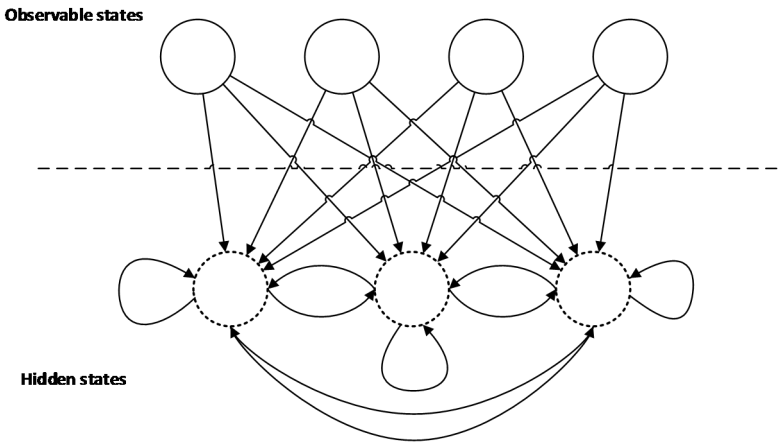
- Representation the generation relationship of sequence elements by statistical probability
- Generative model & Probabilistic graph model
- Hypothesis: only related to previous state

- Common algorithm:
  - Forward-backward algorithm
  - Viterbi algorithm

Usage of HMM

- Give a solution of sequence prediction problem

- Training: get the emit probability from observable states to hidden states, and trans probability between hidden states
- Testing: give the observable states, predict the hidden states



Classic solution

- Proposed by ZHANG Hua-ping(The Chinese Academy of Sciences)
- Main idea: Tagging the sematic role for Chinese words

表 2　地名识别角色简表

角色	意义	示例
A	地名的上文	我/来到/中/关/园
B	地名的下文	刘家村/和/下岸村/相邻
C	中国地名的首部	石/河/子/乡/
D	中国地名的中部	石/河/子/乡/
F	中国地名的末部	石/河/子/乡/
G	中国地名的后缀	海/淀区
X	连接词	刘家村/和/下岸村/相邻
Z	其它非地名成分	

表 3　机构名识别角色表

角色	意义	例子
A	上文	参与/亚太经合组织/的/活动
B	下文	中央/电视台/报道
X	连接词	北京/电视台/和/天津/电视台
C	特征词的一般性前缀	北京/电影/学院
G	特征词的地名性前缀	交通/银行/北京/分行
H	特征词的机构名前缀	中共中央/顾问/委员会
I	特征词的特殊性前缀	中央/电视台
D	机构名的特征词	国务院/侨务/办公室
Z	其它非机构名成份	

- Sematic role: hidden states, POS or some others: observable states

Some limitations

- Special type of named entities need set special role tagging strategy

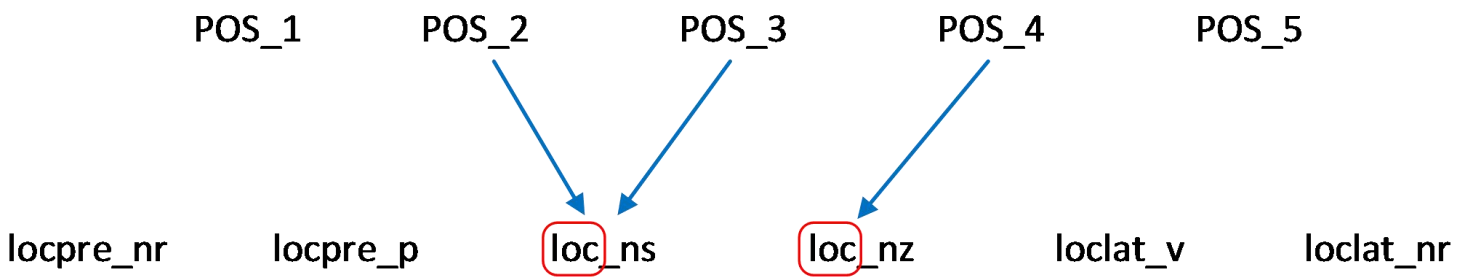
- We can set the role tagging strategy for Names of persons, organizations, locations
- But, how about disease, medicine, Chinese dishes(西红柿炒鸡蛋), and more and more entity types

- We want a unified solution

Our plan

- Use the POS-position mixed tagging strategy

- Sematic role look like: “locpre_p”, “loc_ns”, “loclat_v”
- Set a window to get the states trans probability and emit probability(POS to role)



Probability calculation

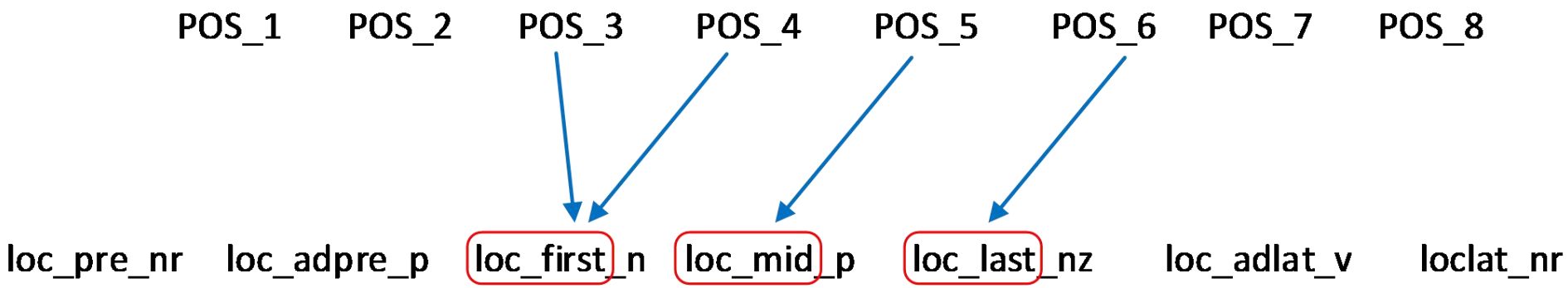
- Using frequency to calculate the probability

- trans_probability: $$\frac{prob(a \to x)}{\sum_{i=1}^{n} prob(a \to i)}$$

- emit_probability: $$\frac{prob\left(POS_a \to role_x^{POS-a}\right)}{\sum_{i} prob\left(POS_a \to role_i^{POS-a}\right)}$$

Finer granularity

- Give more detailed tags for sematic role, like this…



How pity!

- Still in coding ~~~
- Need more experiments to measure the results ~~~

Some APIs

- In Java:
  - OpenNLP:
    - Example code(Chinese NER): https://github.com/Ailab403/ailab-mltk4j/tree/master/src/org/mltk/openNLP by superhy
    - Blog: http://blog.csdn.net/qdhy199148/article/details/51038637 & http://blog.csdn.net/qdhy199148/article/details/51051321 by superhy
- In Python:
  - NLTK: trained model, for English default
  - Stanford NLP: interface in NLTK, trained model for English default
  - Some more?

Some new solutions

- Use deep learning methods
- Use output layer neuron represent entity tagging sequence

- New tools: Word2Vec by Google
- Interface of Gensim called example: https://github.com/superhy/graph-mind/tree/master/src/org_ailab_seg/word2vec

- Readings: http://www.csdn.net/article/2015-06-21/2825013

Reference

- [1] 俞鸿魁,张华平,刘群等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87-94.
- [2] G D Zhou, J Su. Named Entity Recognition using an HMM-based Chunk Tagger. 40th Annual Meeting of the Association for Computational Linguistics,2002.
- [3]张晓鑫. 基于深层神经网络的命名实体识别技术,http://www.csdn.net/article/2015-06-21/2825013. 2015.

Thanks!

Q&A