Improved Chinese Language Processing for an Open Source Search Engine

Forrest Sun

Agenda

- Introduction
  - Background of Yioop
  - Background of Natural Language Processing
- Design and Implementation
  - Chinese Text Segmentation
  - Chinese Part-of-Speech Tagging
  - Chinese Named Entity Recognition
  - Chinese Question Answering System
- Test and Result
- Conclusion

Background of Yioop

- An open source Search Engine written in PHP
- Provides features:
  - Search Results
    - Crawling, Indexing, Retrieving web pages
  - Media Servers
    - News, Video streaming
  - Social
    - Groups, Blogs, Wikis
  - Websites
    - Build websites, Wikis, RSS, JSON service
  - Monetization
    - Ads, etc.

(311 Results)

San Jose mall shooting threat deemed not credible. San Jose Mercury News - 11/03/2020
www.mercurynews.com/2020/03/11/san-jose-mall-shooting-threat-deemed-not-credible
SAN JOSE – A social media post threatening a mass shooting at Westfield Valley Fair mall in the San Jose is not credible, authorities said.

San Jose: Two stabbed downtown early Sunday morning. San Jose Mercury News - 16/02/2020
www.mercurynews.com/2020/02/16/san-jose-two-stabbed-early-sunday-morning
San Jose police said the stabbings happened at about 2:15 a.m. Sunday morning on East San Salvador Street and 2nd Street

San Jose/Campbell community calendar for the week of Feb. 7. San Jose Mercury News - 07/02/2020
www.mercurynews.com/2020/02/07/san-jose-campbell-community-calendar-for-the-week-of-feb-7
Special Events Rose Garden Farmers Market: Saturdays, 10 a.m.-2 p.m. Lincoln High School parking lot, 577 Dana Ave., San Jose. Farmers Market: Willow

Why the owners of San Jose's historic Hotel De Anza are suing the city. San Jose Mercury News - 24/02/2020
www.mercurynews.com/2020/02/24/owners-of-san-joses-historic-hotel-de-anza-sue-the-city
The owners of San Jose's historic Hotel De Anza are joining forces with preservationists in an effort to thwart the construction of a newly approved,

San Jose/Campbell calendar of events for the week of Feb. 21. San Jose Mercury News - 21/02/2020
www.mercurynews.com/2020/02/21/san-jose-campbell-calendar-of-events-for-the-week-of-feb-21
Special Events Rose Garden Farmers Market: Saturdays, 10 a.m.-2 p.m. Lincoln High School parking lot, 577 Dana Ave., San Jose. Farmers Market: Willow

San Jose/Campbell calendar of events for the week of Feb. 28. San Jose Mercury News - 28/02/2020
www.mercurynews.com/2020/02/28/san-jose-campbell-calendar-of-events-for-the-week-of-feb-28
Special Events Rose Garden Farmers Market: Saturdays, 10 a.m.-2 p.m. Lincoln High School parking lot, 577 Dana Ave., San Jose. Farmers Market: Willow

Search in a Search Engine

- www.worldwidewebsize.com shows:
  - There are at least 5.7 billion pages (Wednesday, 06 May, 2020) on internet.
- Impossible to search all pages after user enter the keywords
- Searching “San Jose State University” in Google:
  - About 303,000,000 results (1.14 seconds)

Inverted Index

- Also referred to as a postings file or inverted file
- A database index storing a mapping from content, such as words or numbers, to its locations in a table, or in a document or a set of documents
- Forward index: document->term vs. Inverted index: term-> document

Forward Index

Document	Words
Document 1	the,cow,says,moo
Document 2	the,cat,and,the,hat
Document 3	the,dish,ran,away,with,the,spoon

Inverted index

Word	Documents
the	Document 1, Document 3, Document 4, Document 5, Document 7
cow	Document 2, Document 3, Document 4
says	Document 5
moo	Document 7

Word

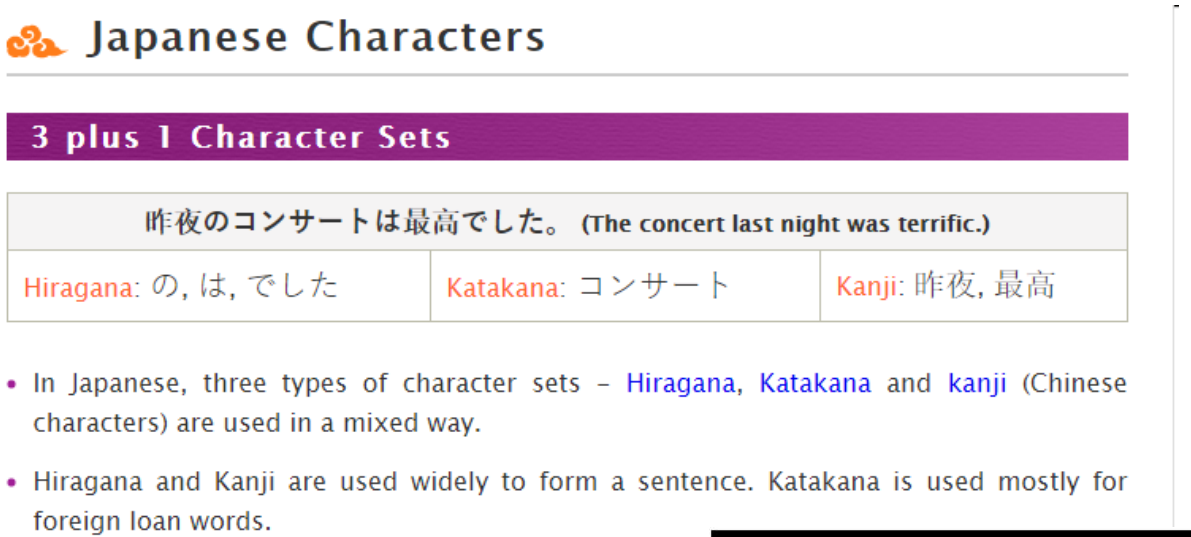
- Leonard Bloomfield introduced the concept of "Minimal Free Forms" in 1928. Words are thought of as the smallest meaningful unit of speech that can stand by themselves.

Introduction of Natural Language Processing

- How Computer analyzes large amounts of natural language data
- A subfield of linguistics, computer science, information engineering, and artificial intelligence
- Has a wide range of research tasks and sub-tasks:
  - Syntax
    - Part-of-speech tagging, Stemming, Word segmentation
  - Semantics
    - Machine translation, Named entity recognition, Optical character recognition (OCR)
  - Speech
    - Speech recognition, Text-to-speech

What is the problem in Chinese Language or some Asian Languages when indexing?

- No delimiter between words
  - English
    - Hello World
  - Chinese
    - 你好世界
  - Japanese
    - こんにちは世界
- Not meaningful to index characters instead of words
- Need help from Natural Language Processing

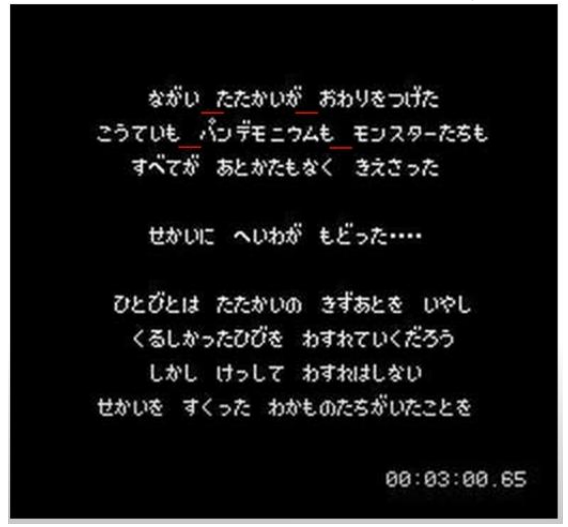


Japanese Characters

3 plus 1 Character Sets

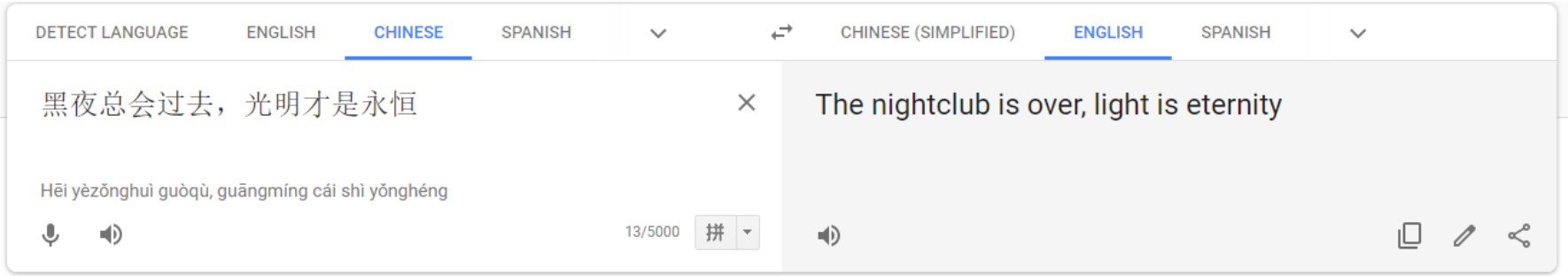
昨夜のコンサートは最高でした。 (The concert last night was terrific.)		
Hiragana: の, は, でした	Katakana: コンサート	Kanji: 昨夜, 最高

- In Japanese, three types of character sets – Hiragana, Katakana and kanji (Chinese characters) are used in a mixed way.
- Hiragana and Kanji are used widely to form a sentence. Katakana is used mostly for foreign loan words.



Chinese Text Segmentation

- Is it easy? No! It's very Ambiguous
  - 黑夜总会过去，光明才是永恒
  - Black Night eventually will be over, light is eternal
- Google translate:
  - 黑夜总会过去，光明才是永恒
  - The nightclub is over, the light is eternity
- Nightclub is a sensitive word in Weibo (Chinese social network) so this post is against law.



What about Human?

- Ambiguous segmentation in Chinese People: Very Common
  - 草帽路飞说要当上海贼王 (Wrong: Shanghai Thief King)
  - 草帽路飞说要当上海贼王 (Correct: Pirate King)
  - Luffy says he will become the Pirate King

- 全国性交易 Nationwide transaction
- 全国性交易 Sex transaction in nation

Some Chinese Text Segment Techniques

- Pure statistic based
- Pure dictionary based
- Statistic and dictionary based
- Machine Learning
  - Maximum Entropy
  - Conditional Random Field

Pure statistic based

- Mutual information
  - Frequency between adjacent chars
- Association measures
  - More measures amount a window of chars
- Non-segmented text

Pure dictionary based

- Greedy Algorithm
- Forward Maximum Matching (FMM)
- Backward Maximum Matching (BMM)
- Yioop can use BMM
- Disadvantages:
  - Ambiguous
  - Cannot segment new words
- Advantages:
  - Fast
  - Standard Accuracy

Statistic and dictionary based

- Use frequency and weight of words
- Newly Implemented in Yioop
- Advantages:
  - Better Accuracy
  - Not too slow
  - No training time (all needs to is to count the words)
- Disadvantages:
  - Cannot segment new words
  - Less accuracy compare to Machine Learning

Machine Learning

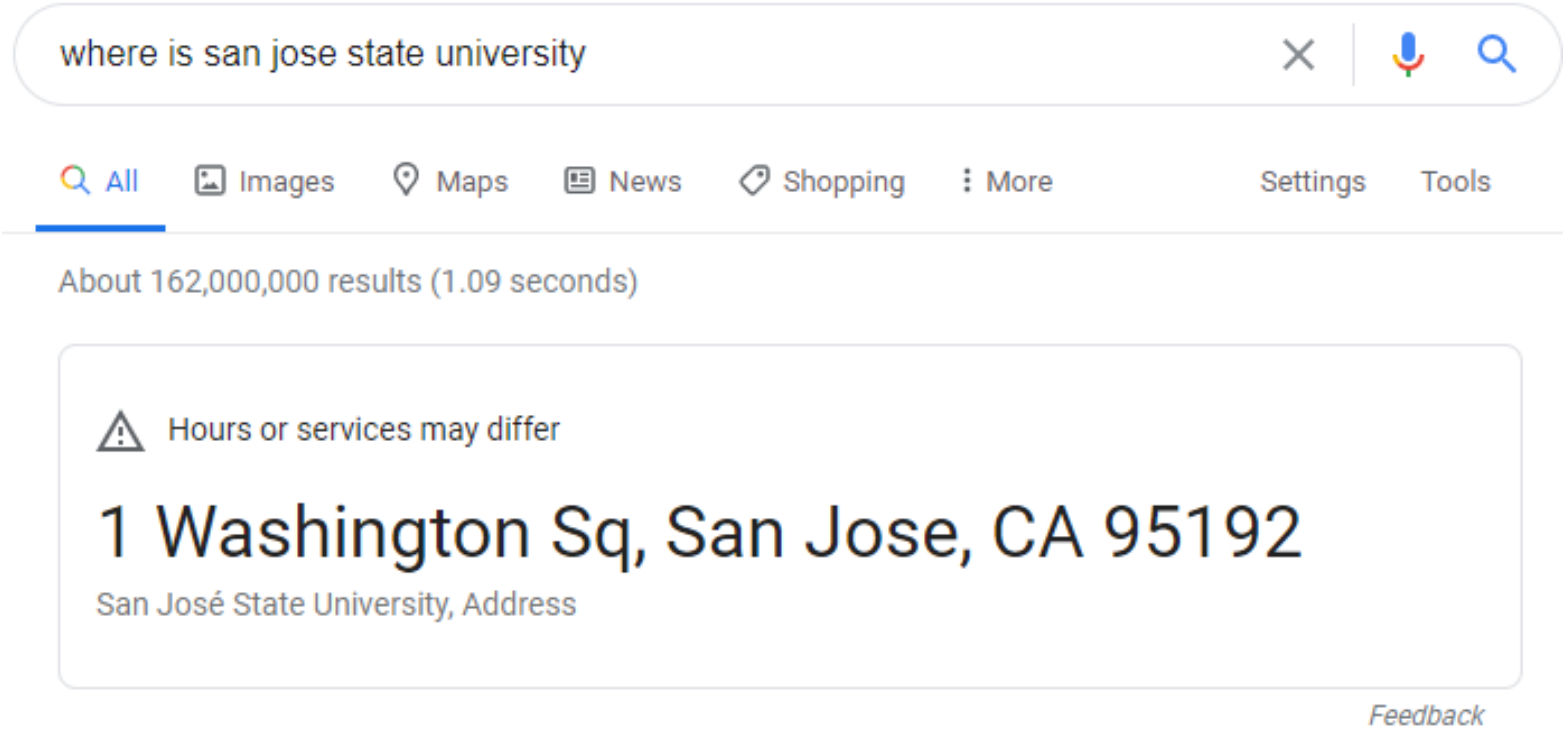
- Maximum Entropy
- Conditional Random Field
- Long Short-Term Memory
- B stands for beginning of the word and E otherwise
  - Ex. 计(B)算(E)机(E)是(B)人(B)类(E)的(B)伟(B)大(E)发(B)明(E)。
- Use Features:
  - Chars around chars, tags around the chars, combination of chars, etc.

Machine Learning. cont'd

- Advantages:
  - Very Accurate
  - Can segment new words
- Disadvantages:
  - Slow
  - Needs a lot of memory
  - High storage space for weights
  - Long training time

Question Answering

- Return answers directly without going into the documents

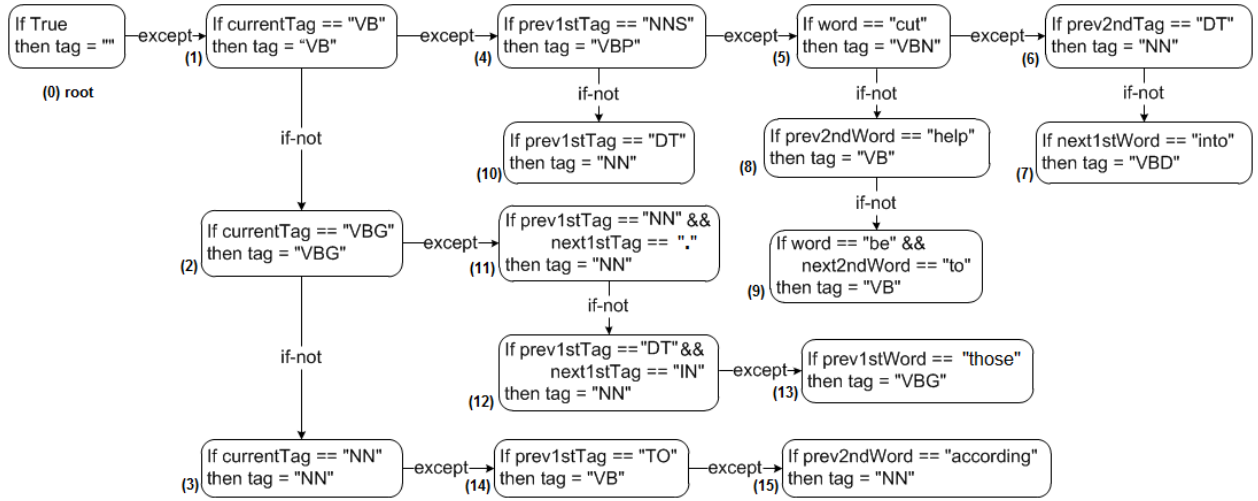


Question Answering

- Knowledge-based
  - Gramma Parser
    - Semantic Triple (triplet): subject–predicate–object
    - Part-of-speech tagging (POS)
  - Named Entity Recognition
    - Location, person name, organization name
    - Ex. San Jose State University
- Information retrieval-based
  - Named Entity Recognition
  - Question detection
  - Searching and ranking documents
  - Ranking paragraphs and answering extraction

POS tagging

- Rule-based
  - Ripple down rule



- Machine Learning
  - Maximum Entropy
  - Conditional Random Field
  - Features:
    - Words surrounded
    - Tags surrounded

Named Entity Recognition

- What / who is the main target in the content?
  - Person
    - Forrest Sun, Dr. Pollett
  - Location
    - California, San Jose
  - Organization
    - San Jose State University
- Encoding
  - IO encoding
  - IOB encoding

	IO encoding	IOB encoding
Forrest	PER	B-PER
lived	O	O
in	O	O
Shanghai	LOC	B-LOC
Mainland	LOC	B-LOC
Of	LOC	I-LOC
China	LOC	I-LOC

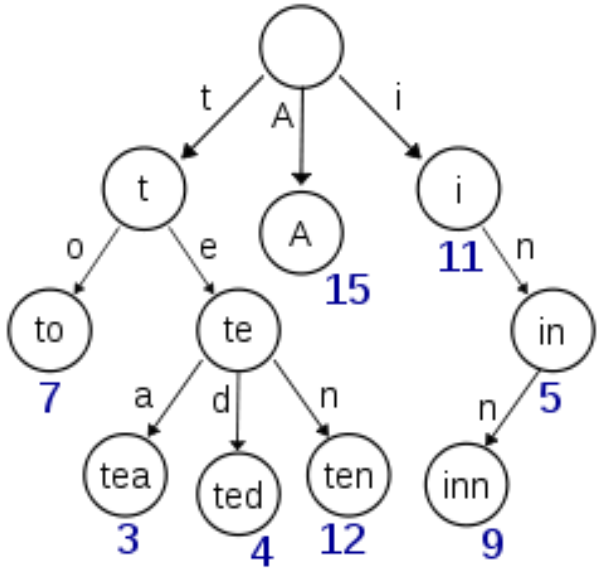
Implementation Details

- Chinese Word Segmentation
- POS tagging
- Named Entity Recognition
- Question Answering

Chinese Word Segmentation

- Use statistic and dictionary-based approach
  - Not too hard to implement
  - Faster than Machine Learning
  - Light in storing weight files
  - Good Accuracy
- Trie Array
  - Good for finding words one char by one char
  - Extremely flexible
  - Consume huge memory when it is very deep

    Especially in Dynamic languages such as PHP



Chinese Words Segmentation cout.

- Frequency (f)
  - How frequent the words appear in the training data
- Probability based on frequency
  - We have a string of characters $C_1C_2C_3$, where both C_1C_2 and C_2C_3 can be a word
    - P_1=Frequency(C_1C_2)*Frequency(C_3)
    - P_2=Frequency(C_1)*Frequency(C_2C_3)
    - P_3= Frequency(C_1)* Frequency(C_2)* Frequency(C_3)
    - Compare and pick the highest choice
  - max $\prod frequency$
  - If we have a long string, P might underflow because each frequency is a very low number

Chinese words segmentation cout.

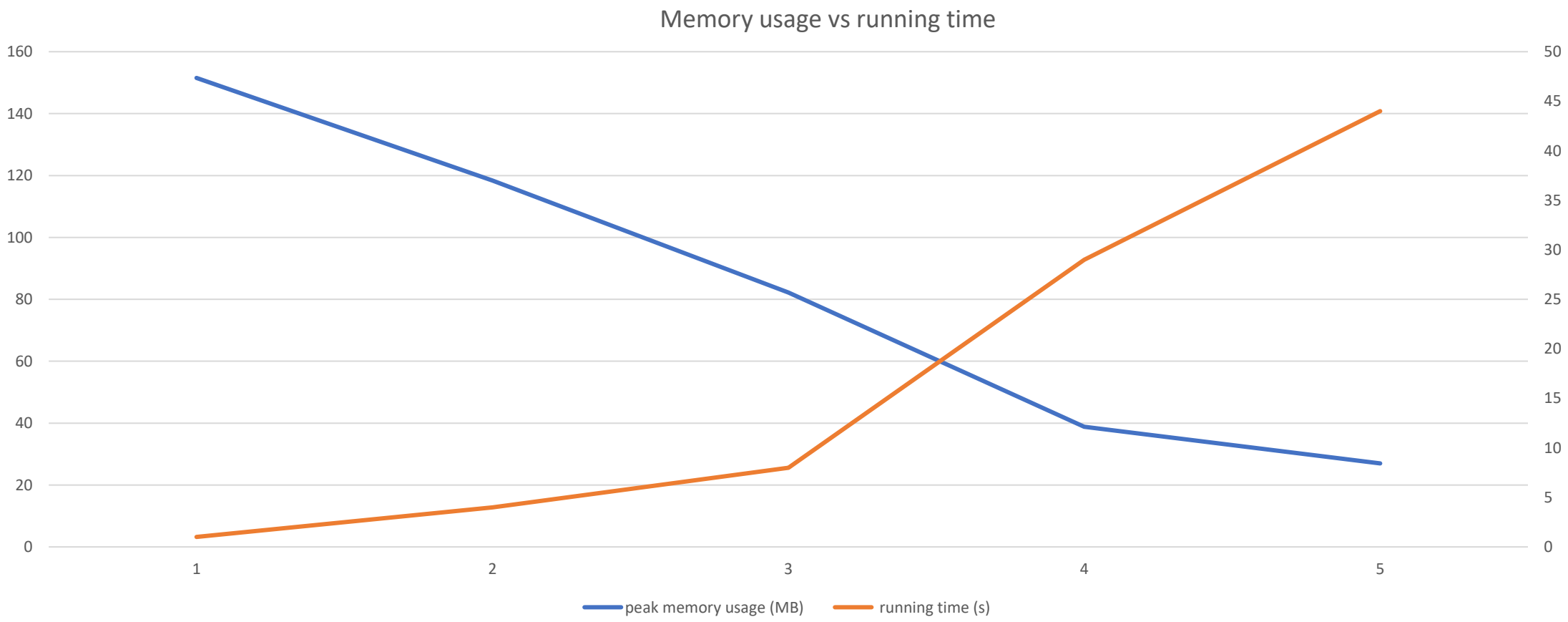
- Weight:
  - $C = -\log(frequency)$
  - min $\sum C$
- Viterbi algorithm / Dynamic programming find the sum

- Enhancement on Memory, caching
  - Trie array consumes too much memory in PHP (about 8 times higher than in C)
  - Caused by the depth of the Array
  - Caching only words in use and throw away outdated
  - 80/20 rule, Pareto principle

Chinese Segmentation Result

Dataset	Stochastic Segmentation (ours)	Reverse Max Match	Neural with Multi-Criteria Learning (Current Highest)
AS	94.3%	89.42%	96.6%
PKU	86.8%	83.10%	96.6%

Segmentation time vs. memory



Maximum Entropy Model / Multinomial logistic regression

- Use logistic regression
- Many classifiers instead of just one
- Can have many features
- Select the one with highest Probability

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\beta_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

Maximum Entropy Model cont'd

- $\mathrm{p}'(\mathrm{h},\mathrm{t}) = \mathrm{w}_0 + \sum_{j=1}^{k} \mathrm{w}_j f_j^{(h,t)}$
  - h is the contexts
  - t is tag
  - f is indicator function
  - w is the weight to be trained
  - k is the number of the indicator functions
- s=sigmoid(p') = $\frac{1}{1+e^{-p'}}$
- Loss function: −(y log(s)+(1−y)log(1− s))

ex. Forrest likes video games

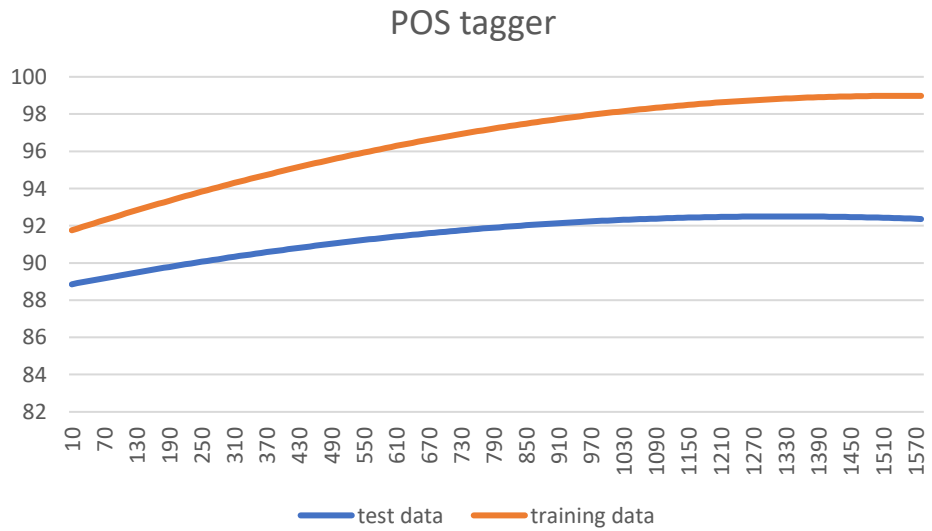
(noun) (???) (noun) (noun)

POS Tagging

- Maximum Entropy Model
- Features:
  - Current word
  - 2 words before current word
  - 2 words after current word
  - 2 tags before current word

POS tagging result

- Data: Chinese Three Bank
  - 30% for training
  - 10% for testing
  - Accuracy: 92.5%



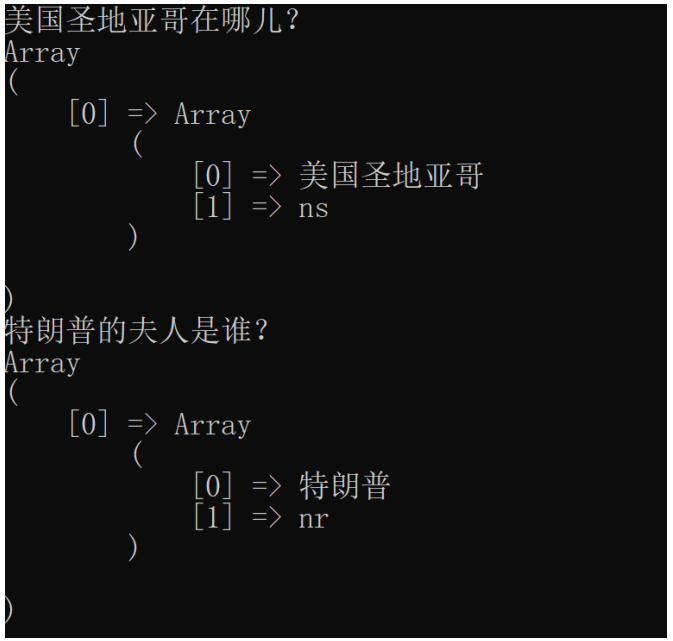
迈向/v 充满/v 希望/v 的/u 新/a 世纪/n ——/nx 一九九八年/t 新年/t 讲话/n （/w 附/v 图片/n 1/m 张/q ）/w
中共中央/nt 总书记/n 、/w 国家/n 主席/n 江/nr 泽民/nr
（/w 一九九七年/t 十二月/t 三十一日/t ）/w
１２月/t ３１日/t ，/w 中共中央/nt 总书记/n 、/w 国家/n 主席/n 江/nr 泽民/nr 发表/v １９９８年/t 新年/t 讲话/n 《/w 迈向/v 充满/v 希望/v 的/u 新/a 世纪/n 》/w 。/w （/w 新华社/nt 记者/n 兰/nr 红光/nr 摄/Vg ）/w
同胞/n 们/k 、/w 朋友/n 们/k 、/w 女士/n 们/k 、/w 先生/n 们/k ：/w
在/p １９９８年/t 来临/v 之际/f ，/w 我/r 十分/t 高兴/a 地/u 通过/p 中央/n 人民/n 广播/vn 电台/n 、/w 中国/ns 国际/n 广播/vn 电台/n 和/c 中央/n 电视台/n ，/w 向/p 全国/n 各族/r 人民/n ，/w 向/p 香港/ns 特别/d 行政区/n 同胞/n 、/w 澳门/ns 和/c 台湾/ns 同胞/n 、/w 海外/s 侨胞/n ，/w 向/p 世界/n 各国/r 的/u 朋友/n 们/k ，/w 致以/v 诚挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/v ！/w
１９９７年/t ，/w 是/v 中国/ns 发展/v 历史/n 上/f 非常/d 重要/a 的/u 很/d 不/d 平凡/a 的/u 一/m 年/q 。/w 中国/ns 人民/n 决心/n 继承/v 邓/nr 小平/nr 同志/n 的/u 遗志/n ，/w 继续/v 把/p 建设/v 有/v 中国/ns 特色/n 社会主义/n 事业/n 推向/v 前进/v 。/w 中国/ns 政府/n 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/w 并/c 按照/p “/nx 一国两制/j ”/nx 、/w “/nx 港人治港/l ”/nx 、/w 高度/n 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/v 稳定/v 。/w 中国/ns 共产党/n 成功/a 地/u 召开/v 了/u 第十五/m 次/q 全国/n 代表大会/n ，/w 高举/v 邓小平理论/n 伟大/a 旗帜/n ，/w 总结/v 百年/t 历史/n ，/w 展望/v 新/a 的/u 世纪/n ，/w 制定/v 了/u 中国/ns 跨/v 世纪/n 发展/v 的/u 行动/vn 纲领/n 。/w

Named Entity Recognition

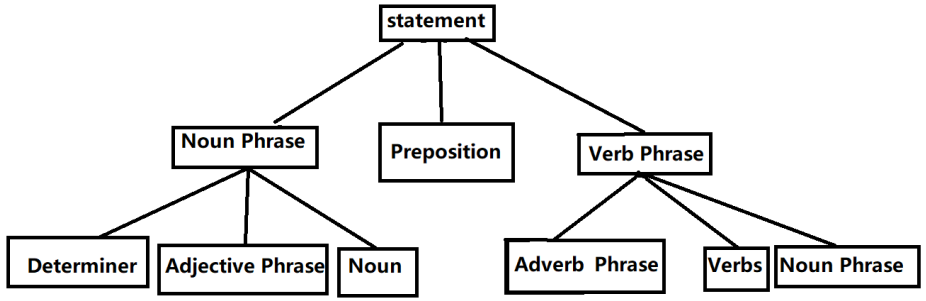
- Maximum Entropy Model
- Features:
  - Current character
  - 2 characters before current character
  - 2 characters after current character
  - 2 tags before current character

Named Entity Recognition Result

- MSRA dataset
  - With its own Training / Test dataset
  - 83.6%



Question Answering System



- Chinese syntax
  - Subject predicate object
  - Very flexible compare to English
  - Lots of particle words. Usually has no meaning. Used for tense
    - 的，地，得，了，着
- One sentence consists of many sub-sentences.
  - Sub-sentence does not necessarily contain subject
    - 珠穆朗玛峰是喜马拉雅山脉的主峰，同时是世界海拔最高的山峰，位于中国与尼泊尔边境线上。
    - Mount Qomolangma is the main peak of the Himalayas, at the same time is the highest mountain in the world, is located on the border between China and Nepal.

Chinese syntax Parser result

- The subject, predicate and object are parsed

```
[QUESTION_ANSWER_LIST] => Array
    (
        [qqq 是 主峰] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 主峰] => 是
        [珠穆朗玛峰 是 qqq] => 世界最高峰
        [qqq 是 喜马拉雅山脉 的 主峰] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 喜马拉雅山脉 的 主峰] => 是
        [qqq 是 世界海拔] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 世界海拔] => 是
        [qqq 是 世界海拔  最高 的 山峰] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 世界海拔  最高 的 山峰] => 是
        [qqq 位于 中国与尼泊尔边境线上] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 中国与尼泊尔边境线上] => 位于
        [珠穆朗玛峰 位于 qqq] => 中国与尼泊尔边境线上
        [qqq 是 世界最高峰] => 珠穆朗玛峰
        [珠穆朗玛峰 qqq 世界最高峰] => 是
    )
```

Conclusion

- Implemented and improved many features in Yioop
  - Chinese Segmentation
  - POS tagging
  - NER
  - QA
- Many different approaches are used in this project
  - Rule based
  - Statistics based
  - Machines learning

Future work

- Compare to latest techniques in Natural language processing, the accuracy in my project is not great.
- Needs more machines learning library support for NLP.
- Question Answering System can have IR-based feature.