上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

New Media
Data Analytics and Application

Lecture 9: Natural Language Processing
A Brief Introduction

Ting Wang

Outlines

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

a brief introduction to natural language processing

What is Natural Language Processing

What is NLP

Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.

- NLP is related to the area of human–computer interaction.
- NLP involves natural language understanding and natural language generation.
- Also called as Computational Linguistics

Objectives

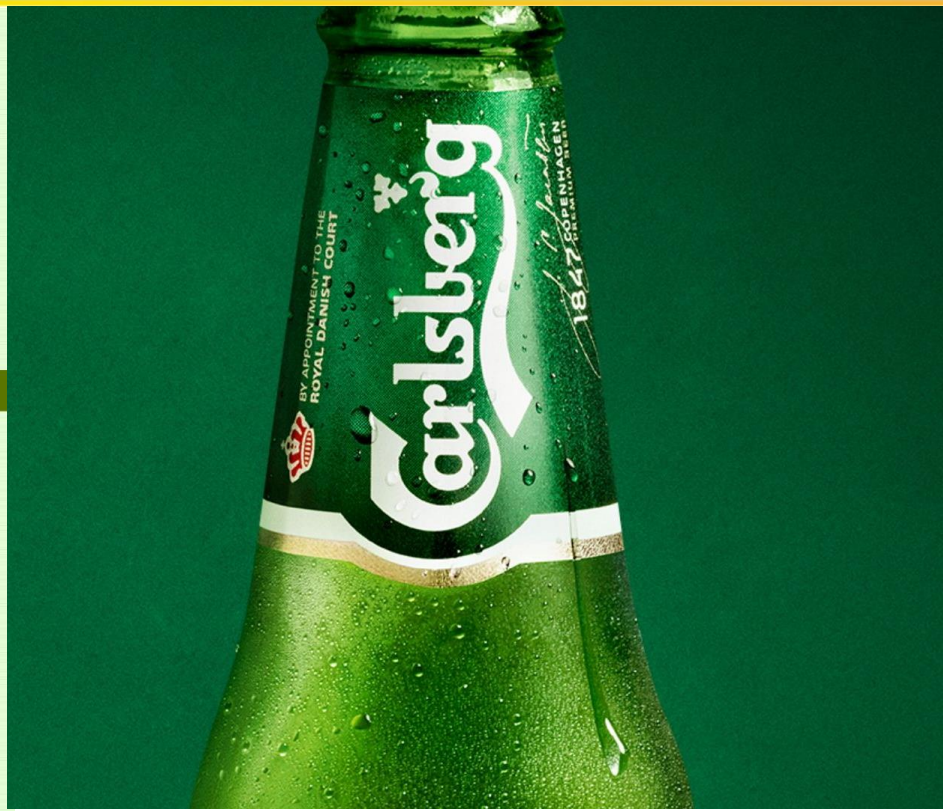
Let your computer know you, and let you know the world

Qualitative Data

Quantitative Data

Natural Language Processing

Review Examples:
INOHERB and GARLSBERG



Natural Language Processing

Tasks (1)

- Automatic summarization 自动总结
- Machine translation 机器翻译
- Named entity recognition 命名实体识别
- Natural language generation 自然语言生成
- Natural language understanding 自然语言理解
- Optical character recognition 光学字符识别
- Part-of-speech tagging 词性标注
- Parsing 语法解析
- Question answering 问答系统
- Relationship extraction 关系提取（主体之间的关系）

Tasks (2)

- ***Sentence breaking*** *断句（文言文，语音）*
- ***Sentiment analysis*** *情感分析*
- ***Speech recognition*** *语音识别*
- ***Speech segmentation*** *语音切分（词汇）*
- ***Topic segmentation and recognition*** *主题切分与 识别*
- ***Word segmentation*** *分词 （中日韩等）*
- ***Word sense disambiguation*** *词汇歧义削减*
- ***Information retrieval*** *信息检索、信息过滤*
- ***Information extraction*** *信息抽取*
- ***Speech processing*** *语音处理 （文字语音互转）*

Natural Language Processing

Approaches

1. Symbolicism 符号主义
   - Regulation 规则（显性规则）:决策树DT
   - Statistics 统计（隐形规则）:贝叶斯、HMM、PCA

2. Connectionism 链接主义
   - Neural Networks 神经网络 : Deep Learning

3. Actionism 行为主义
   - Evolutionism 进化主义 : 遗传算法GA、PSO

Natural Language Processing

NLP using Python

NLTK (http://www.nltk.org/)

- Current Version : NLTK 3
- Installation (http://www.nltk.org/install.html)
- import nltk

Foundations of NLP: Semantic Resource 语义资源

- Dictionary 字典
- Stop Word List 停用词表
- Knowledge Graph 知识图谱
  - Knowledge Base 知识库
  - Semantic Networks 语义网络
- Regulation Base 规则库
- Corpus 语料库

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

the foundation of NLP

Semantic Resource

Semantic Resource

Dictionary 字典

	ID	Chinese	English	Memo
1	1	α值	Alpha	在股票收益方面，α值衡量某种证券或基金经风险调整后的回报。α值是代表证券收益率超出风险/收益模型所...
2	2	《美国破产法》第七章	Chapter 7	《美国破产法》第七章是关于非自愿清盘的法规，债权人据此请求法庭颁令判决债务人破产。该章赋予由法庭...
3	3	《美国破产法》第十一章	Chapter 11	按《美国破产法》第十一章的安排，无力偿债的债务人若成功申请破产保护，将可保住企业的财产及经营的控...
4	4	3A等级	Triple A Rated	参见AAA/Aaa（3A等级）。 债券
5	5	3A等级（最高信用评级）	AAA	给予优质债券的最高评级。由标准普尔、穆迪和惠誉国际等主要评级机构评定。参见Credit Rating（信用评级...
6	6	J-曲线	J-Curve	经济学上的一个概念，指一个变量在受到某种刺激时，有时可能会继续按原先的方向发展，然后才出现明显的...
7	7	Vega值	Vega	里度期权标的资产价格波动率的变动如何影响期权价值的指标。参见Option（期权）。The measure of change...
8	8	β值	Beta	贝塔系数是里度股票投资系统风险的指标。所谓系统风险是指股票投资中没有办法通过分散投资来减低的风险...
9	9	㏀系数	NULL	参见：㏀值
10	10	θ值	Theta	里度期权价值如何随着期权有效期的缩减而变动的指标。期权的价值会随着时间过去——即期权日益接近到期...
11	11	λ值	Lambda	指里度期权杠杆水平的一个比率，显示标的资产的价格每变动一个百分点，可导致期权价格变动的百分比. 标的...
12	12	阿尔法系数	NULL	参见：α值
13	13	阿拉伯石油输出国组织	OAPEC	英文Organization of Arab Petroleum Exporting Countries的缩写。该组织的宗旨是促进阿拉伯产油国之间...
14	14	阿历山大过滤器	Alexander's Filter	指技术分析的一种方法，以涨跌百分比来衡量特定时间内价格上涨或下跌的速度。升速很快为买进讯号，反之...
15	15	阿姆斯特丹	ARA	英文Amsterdam/Rotterdam/Antwerp 的缩写。石油货品若称cost and freight ARA，是指可将阿姆斯特丹/鹿特...
16	16	艾略特波浪理论	Elliott Wave Theory	技术分析的一种理论，认为市场走势不断重复一种模式，每一周期由5个上升浪和3个下跌浪组成。艾略特波浪...
17	17	安特卫普地区	NULL	参见：阿姆斯特丹
18	18	按比例偿债基金	Pro Rata Sinking Fund	偿债基金是在债券到期前提前偿还部分债务的一种安排。债券发行时若附设偿债基金条款，发债人必须定期将...
19	19	按揭证券	MBS	英文Mortgage-backed Security的缩写，由一篮子住房抵押贷款提供担保的证券，该抵押贷款组合每月收到的还...
20	20	按面值	At Par	指证券的售价与其面值相等。When a security is selling at a price that is equal to face value.期货...

Stop Word List
停用词表

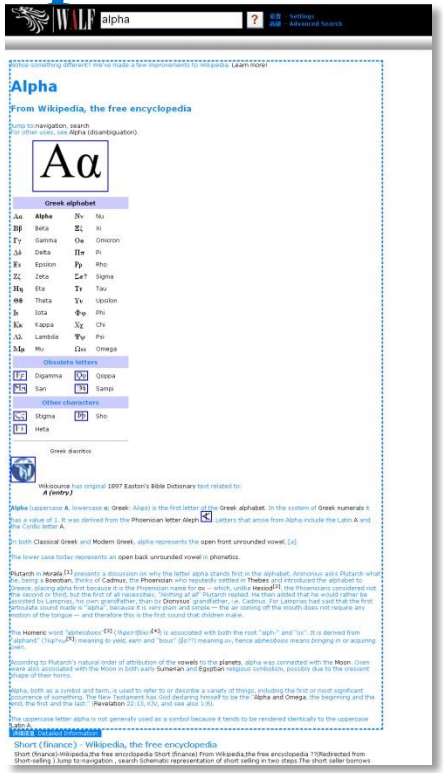
- *Punctuation 标点*
- *Symbol 符号*
- *Function Word虚词*
- *Interjection 叹词*
- *Empty Word无意义的词*
- *Ambiguous Word 引起歧义的词*

SW_ID	WORD_NAME	WPS_ID
35	`	1
36	/	1
37	\	1
38	\n	1
39	’	1
40	”	1
41	\t	1
42	:”	1
43	:“	1
44	啊	1
45	阿	1
46	哎	1
47	哎呀	1
48	哎哟	1
49	唉	1
50	按	1
51	按照	1
52	吧	1
53	吧哒	1
54	把	1

Semantic Resource

Knowledge Graph 知识图谱

	WIKI_WORD_ID	WEB_ID	WIKI_WORD
1	1	1	Alpha
2	2	2	Chapter_7
3	3	3	Chapter_11
4	4	4	AAA
5	5	5	Vega
6	6	6	Beta
7	7	7	Theta
8	8	8	Lambda
9	9	9	OAPEC
10	10	10	ARA
11	11	11	Elliott_Wave_Theory
12	12	12	MBS
13	13	13	All_Ordinaries
14	14	14	G8
15	15	15	Paris_Club
16	16	16	MONEP
17	17	17	Pibor
18	18	18	COB
19	19	19	Basel_Committee
20	20	20	White_Knight



	WIKI_RELATIVE_WORD_ID	WEB_ID	WIKI_WORD	RELATION
1	1	1	Greek_alphabet	0
2	2	1	Beta	0
3	3	1	Gamma	0
4	4	1	Omicron	0
5	5	1	Epsilon	0
6	6	1	Zeta	0
7	7	1	Sigma	0
8	8	1	Eta	0
9	9	1	Tau	0
10	10	1	Theta	0
11	11	1	Upsilon	0
12	12	1	Iota	0
13	13	1	Kappa	0
14	14	1	Lambda	0
15	15	1	Omega	0
16	16	1	Digamma	0
17	17	1	Qoppa	0
18	18	1	Sampi	0
19	19	1	Greek_diacritics	0
20	20	1	Wikisource	0

Regulation Base 规则库

X and Y are couples -> Y and X are couples
X and Y are couples, and X is a male-> X is Y's husband
X is Y's husband -> Y is X's wife

	Condition	Result
1	X <and> Y [be] {couple}	Y <and> X [be] {couple}
2	(X <and> Y [be] {couple}) <and> (X [be] {male})	X [be] Y {husband}
3	X [be] Y {husband}	Y [be] X {wife}

Semantic Resource

Corpus 语料库

- http://www.cncorpus.org/
- http://www.corpus4u.org/
- http://bcc.blcu.edu.cn/
- http://corpus.byu.edu/coca/
- http://www.sogou.com/labs/resource/list_yuliao.php

Semantic Resource

- 1.中央研究院近代汉语标记语料库：http://www.sinica.edu.tw/Early_Mandarin/
- 2.中央研究院汉籍电子文献（瀚典全文检索系统）http://www.sinica.edu.tw/ftms-bin/ftmsw3
- 3.国家现代汉语语料库：http://124.207.106.21:8080/
- 4.国家语委现代汉语语料库：http://www.clr.org.cn/retrieval/index.html
- 5.树图数据库：http://treebank.sinica.edu.tw/
- 6.LIVAC共时语料库：http://www.livac.org/s
- 7.北京大学中国语言学研究中心，简称CCL语料库检索系统http://ccl.pku.edu.cn/Yuliao_Contents.Asp
- 8.北京大学《人民日报》标注语料库：http://www.icl.pku.edu.cn
- 9.北京语言大学的语料库：http://www.blcu.edu.cn/kych/H.htm
- 10.清华大学的汉语均衡语料库TH-ACorpus：http://www.lits.tsinghua.edu.cn/ainlp/source.htm
- 11.山西大学语料库http://www.sxu.edu.cn/homepage/cslab/sxuc1.htm
- 12.香港城市大学的LIVAC共时语料库： http://www.rcl.cityu.edu.hk/livac/或http://www.LIVAC.org
- 13.浙江师范大学的历史文献语料库:http://lib.zjnu.net.cn/xueke/hyywzx/xkjj.htm
- 14.中国科学院计算所的双语语料库： http://mtgroup.ict.ac.cn/corpus/query_process.php
- 15.中文语言资源联盟：http://www.chineseldc.org/xyzy.htm
- 16.红楼梦汉英平行语料库：http://score.crpp.nie.edu.sg/hlm/index.htm#
- 17.SKETCHENGINE多语言语料库：www.sketchengine.co.uk

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

using keyword to describe an article

Keyword Analysis

Keyword Analysis

EXAMPLE 2:
华谊 **VS** 万达

Keyword Analysis

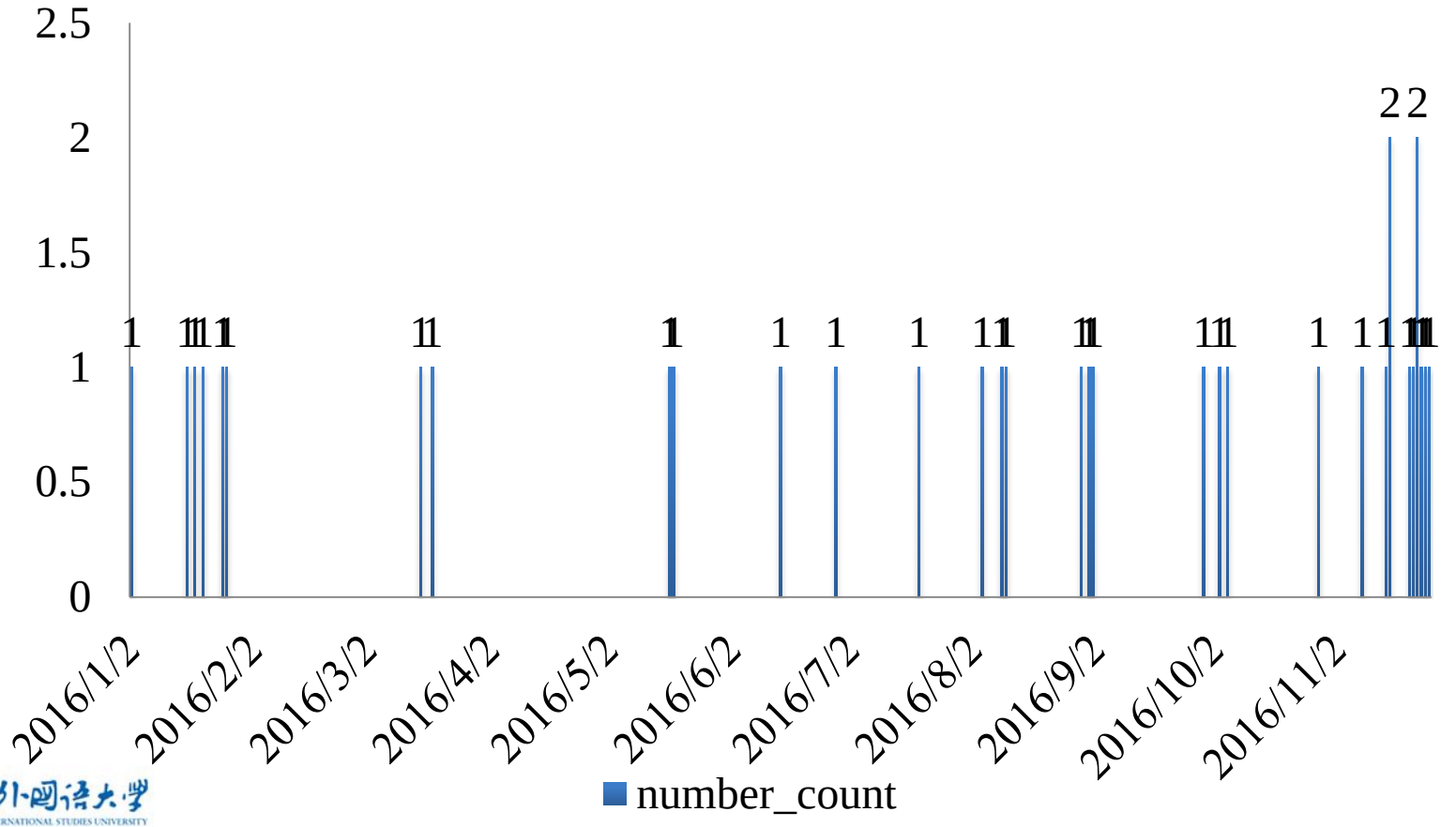
- Step 1: take keywords:
  “冯小刚”，“华谊”，“万达”
- Step 2:
  Get numbers of reports published every day

```
SELECT COUNT(*) AS NUMBER_COUNT, NEWS_TITLE, PUBLISH_DATE FROM
FILM_NEWS WHERE NEWS_ID IN (SELECT NEWS_ID FROM FILM_NEWS WHERE
NEWS_CONTENT LIKE '%万达%' AND NEWS_ID IN (SELECT NEWS_ID FROM
FILM_NEWS WHERE NEWS_CONTENT LIKE '%华谊%' AND NEWS_ID IN (SELECT
NEWS_ID FROM FILM_NEWS WHERE NEWS_CONTENT LIKE '%冯小刚%')) ORDER BY
PUBLISH_DATE) GROUP BY PUBLISH_DATE
```

Keyword Analysis

- Data Description
  - Data Collection :
    - Web crawler gets data from Entgroup
  - The Size of the Dataset
    - About 22000 articles from 2009-2016
    - Totally, about 700 MB data

Keyword Analysis

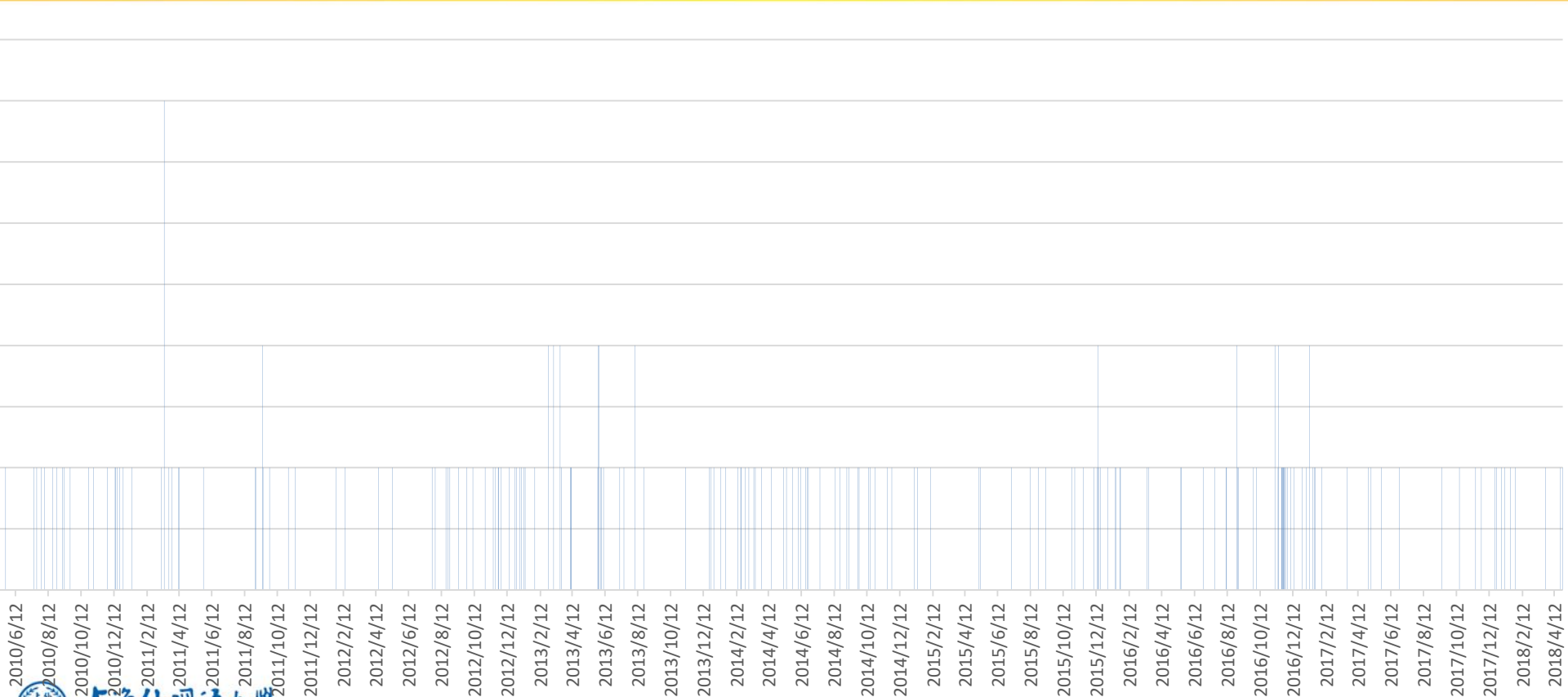


Keyword Analysis

Conclusions

- This event started from Nov. and is now still very hot in these days
- Although this event was taken placed in this Nov. but they have business much earlier.
- Summer and New Year are good seasons for film industry

Keyword Analysis

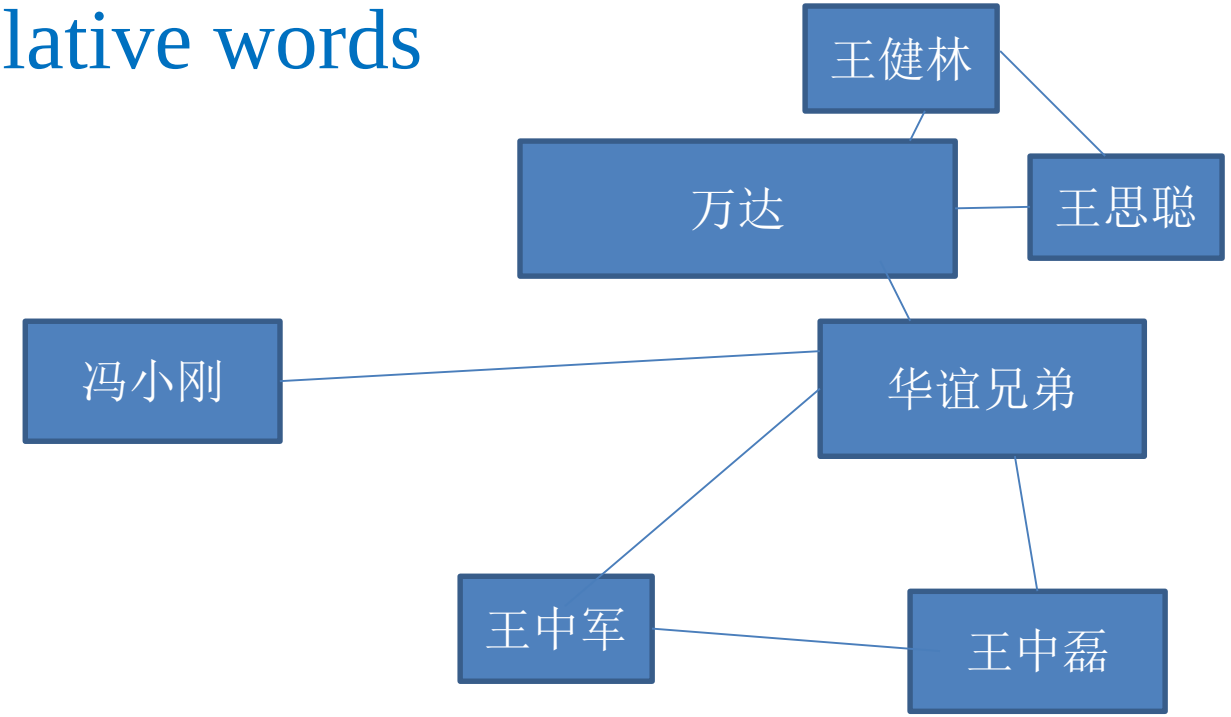


Keyword Analysis

Now, We have data, and results. Are these precise enough?

Keyword Analysis

- Correlative words



one of the most important statistical computational linguistic models

N-gram

Statistical Natural Language Processing

统计自然语言处理

- N-gram N元语法
- Hidden Markov Model(HMM) 隐马尔科夫模型
- Bayes' Theorem 贝叶斯定理

Definition of N-gram N元文法

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n − 1)-order Markov model.

N	N-gram	(N – 1)-order Markov model	Example
1	1-gram(unigram)	Independent from history	One Word
2	2-gram(bigram)	1-order (HMM-1)	Two Words
3	3-gram(trigram)	2-order (HMM-2)	Three Words
…	…	…	…

Unigram 上下文无关文法

- Only consider the probability of the word itself
- Hypothesis: Every word is independent.

$$P(X) = P(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} P(x_i)$$

$$P(x_i) = \frac{Number\ of\ x_i\ in\ the\ artical}{Number\ of\ all\ words\ in\ the\ artical}$$

Bigram 二元文法

The current word is influenced by the previous one word

$$P(X) = P(x_1, x_2, \cdots, x_N) = P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_N|x_{N-1})$$

$$= P(x_1)\prod_{i=2}^{N} P(x_i|x_{i-1})$$

$$P(x_i|x_{i-1}) = \frac{Number\ of\ (x_{i-1}x_i)\ in\ the\ artical}{Number\ of\ all\ x_{i-1}\ in\ the\ artical}$$

Trigram 三元文法

The current word is influenced by the previous two words

$$P(X) = P(x_1, x_2, \cdots, x_N) = P(x_1)P(x_2|x_1)P(x_3|x_2x_1)P(x_4|x_3x_2)\cdots P(x_N|x_{N-1}x_{N-2})$$

$$= P(x_1)P(x_2|x_1)\prod_{i=3}^{N} P(x_i|x_{i-1}x_{i-2})$$

$$P(x_i|x_{i-1}x_{i-2}) = \frac{Number\ of\ (x_{i-2}x_{i-1}x_i)\ in\ the\ artical}{Number\ of\ all\ (x_{i-2}x_{i-1})\ \ in\ the\ artical}$$

N-gram

Tips

1. Previous studies showed that trigram and four-gram often have better performance

2. The larger of *N*, the more complex of the computation

3. N-gram needs training data set, while it is impossible for a training data set to contain all the matches of a word

Smoothing 平滑

- Zero Probability 零概率
- Small Probability 小概率
- Laplace Smoothing 拉普拉斯平滑

$$P(x_i|x_1, x_2, \cdots, x_{i-1}) = \frac{Number\ of\ (x_1 \dots x_i) in\ the\ artical + 1}{Number\ of\ all (x_1 \dots x_{i-1}) in\ the\ artical + Number\ of\ words\ in\ dictionary}$$

Commonly used Smoothing Approaches

- Linear interpolation (e.g., taking the weighted mean of the unigram, bigram, and trigram)
- Good–Turing discounting
- Witten–Bell discounting
- Lidstone's smoothing
- Katz's back-off model (trigram)
- Kneser–Ney smoothing

Ref. https://en.wikipedia.org/wiki/N-gram

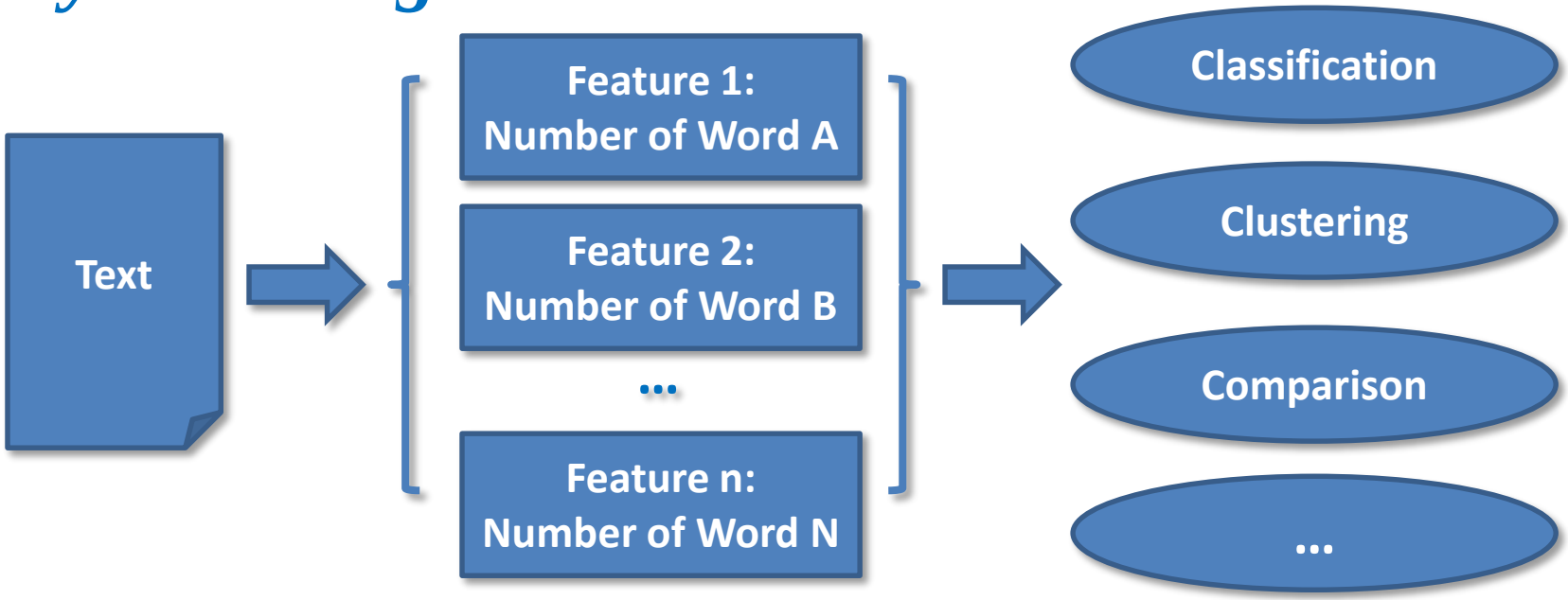
上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

the first step for Chinese information processing

Chinese Word Segmentation

Chinese Word Segmentation

Why Word Segmentation?



However, it is difficult to extract words from Chinese text.

Chinese Word Segmentation

Difficulties: Disambiguation

乒乓球拍卖完了

乒乓|球拍|卖完了

乒乓球|拍卖|完了

N-gram

- Chinese Word Segmentation 分词
  - Forward Max. matching method 正向最大匹配
  - Backward Max. matching method 逆向最大匹配
  - Statistical matching method 统计学方法

Forward Max. matching method, FMM

正向最大匹配

准备工作：需要分词词典D
设MaxLen表示最大词长度
算法：

1. 从生语料N中取长度为MaxLen的字串str，令Len= MaxLen
2. 把str与D中的词相匹配
3. 若匹配成功，则认为str为词，N中去掉str(指针前移Len个单位)，返回1
4. 若匹配不成功，

- 若Len>1则Len--，从生语料N中取长度为Len的字串str返回2；
- 否则，得到单字词， N中去掉str(指针前移1个单位)，返回1

若4中得到的单字不是词，则要进行未登录词处理
若待切分的语料字串长度小于MaxLen，则取str为待切分语料

Chinese Word Segmentation

Backward Max. matching method, BMM

逆向最大匹配

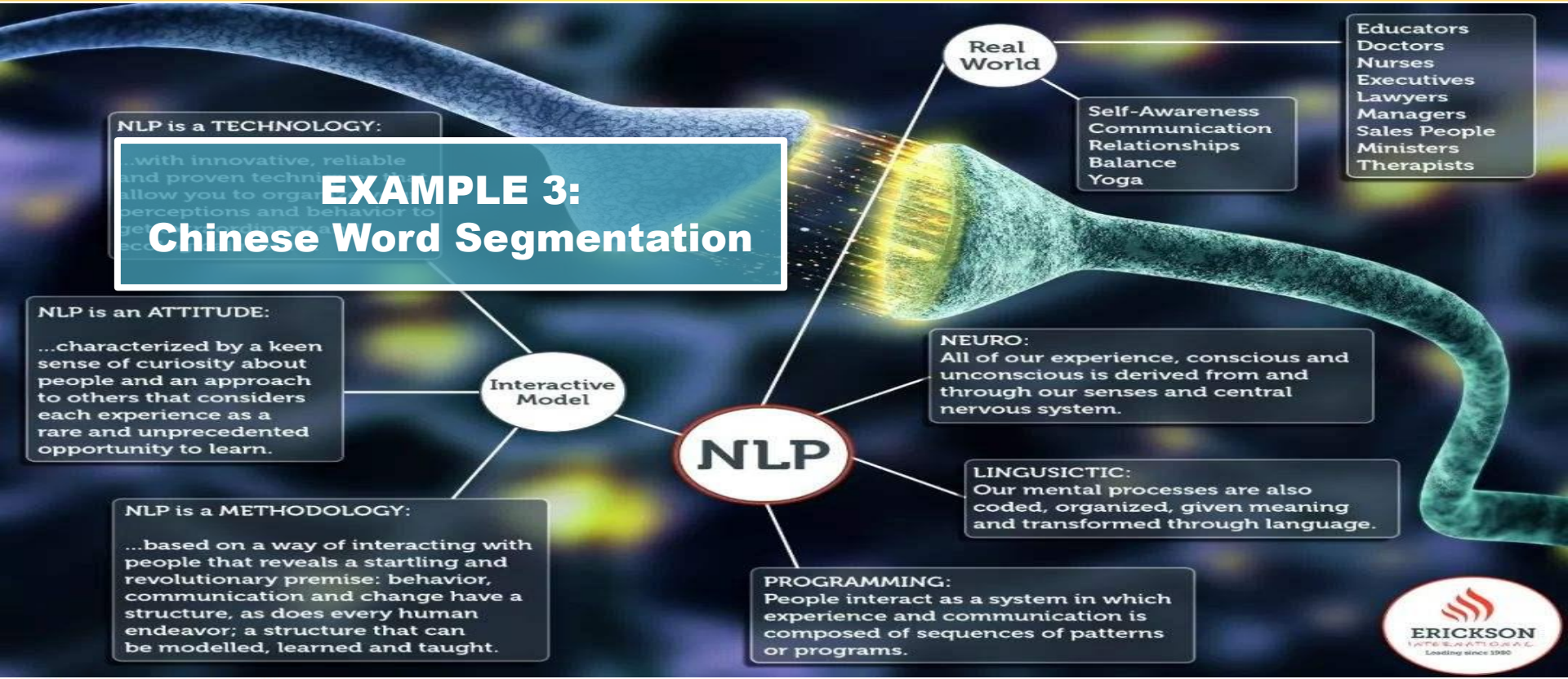
1. Similar to FMM, but the text is scanned from the right side
2. Often jointly use with FMM

Chinese Word Segmentation

- **Statistical Matching Method**

```
FMM and BMM
Begin initialize Path←{},AmbiguousString,SubString←{}
   While (AmbiguousString.Length>0)
   {
       //只考虑以当前HMM第一个状态开始的匹配序列
       SubString←以AmbiguousString中的第一个字为基准，取出所有可能的匹配字符串
       Foreach SubString
       {
           //提供当前情况下所有的概率，为判断歧义作参考
               计算当前每一种可能情况的概率P(SubString)  //unigram, bigram, trigram with smoothing
       }
       //选择概率最大的SubString添加到Path
       将argmax(P(SubString))添加到Path
       //准备考察除去最大概率的SubString后的AmbiguousString，从HMM序列首部开始，除去所有的匹配状态
       AmbiguousString.Remove(0, argmax(P(SubString)).Length)
    }
    Return Path
End
```

Chinese Word Segmentation



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

The End of Lecture 9

Thank You

http://www.wangting.ac.cn