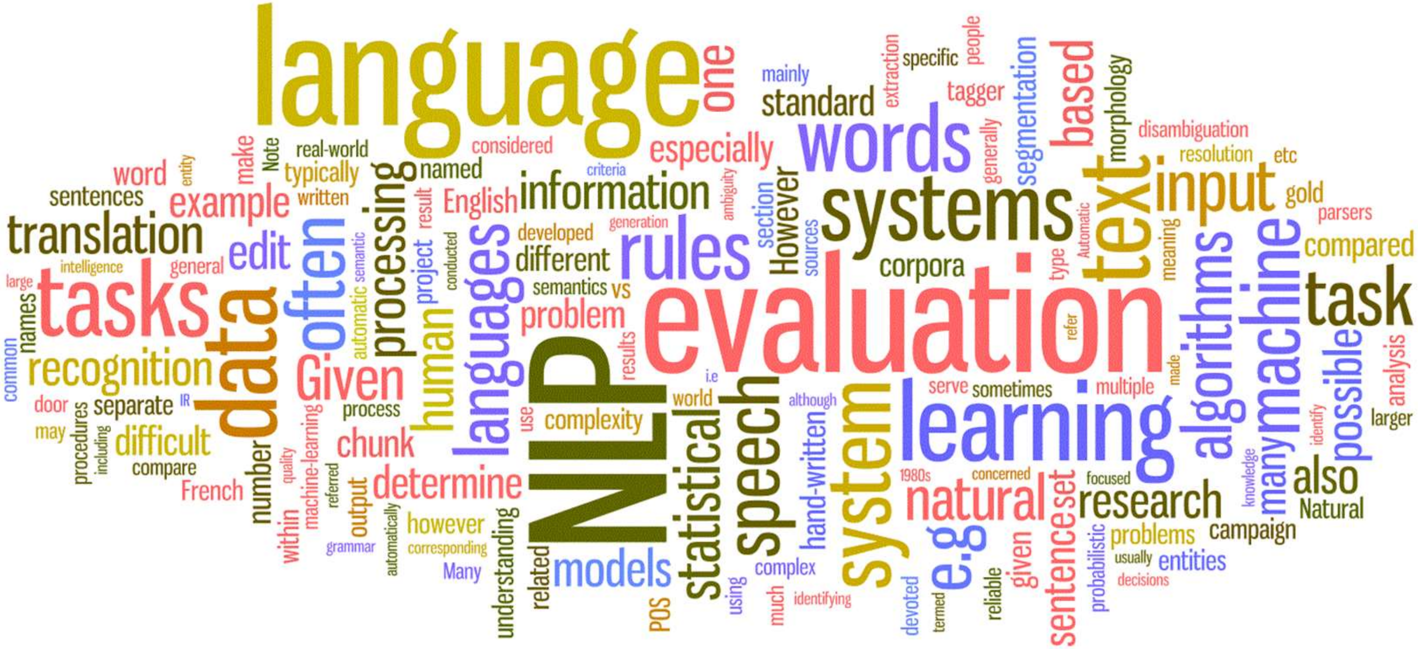


Lecture 1: Introduction to Natural Language Processing

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Course Information

- Lecture
  - Prof. Xu Ruifeng 徐睿峰
    Office: L-1602
    Email: xuruifeng@hit.edu.cn

22自然语言处理-本

扫一扫二维码，加入群聊。

- Teaching Assistant

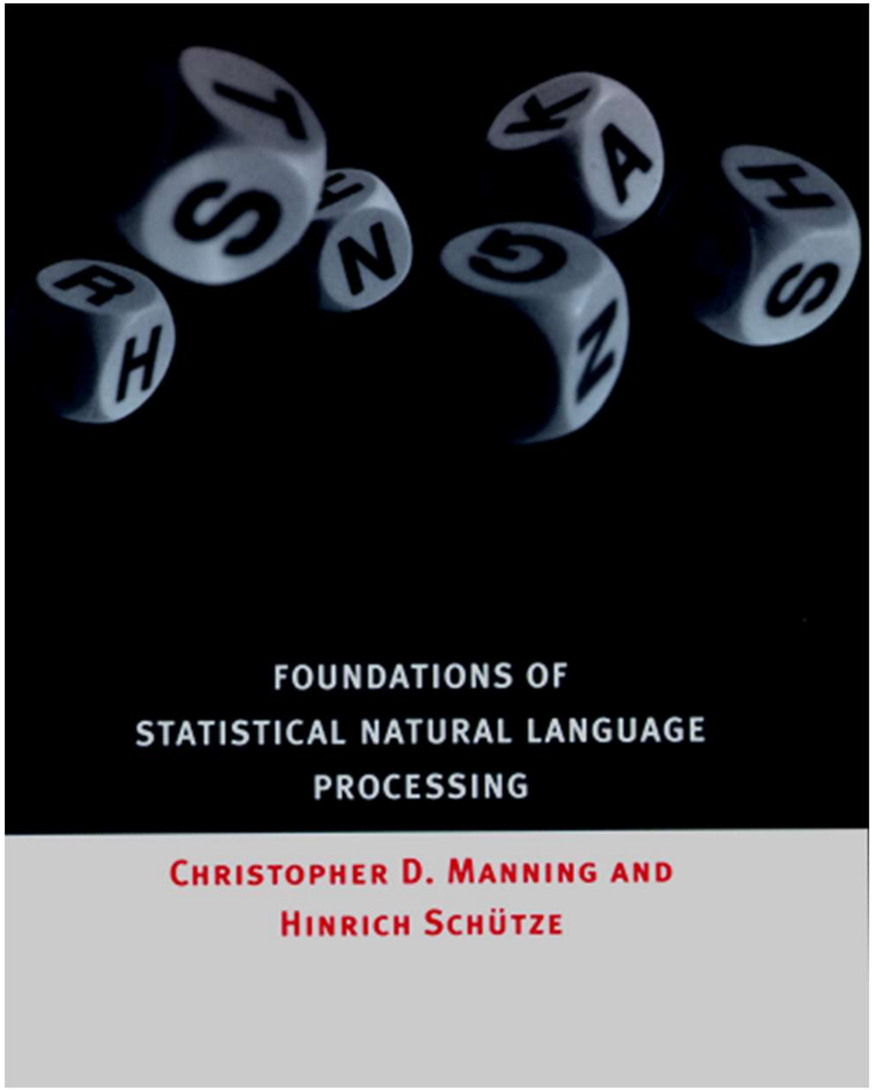
杨才华

陈奕

林琦慧

  - Office: L-1406

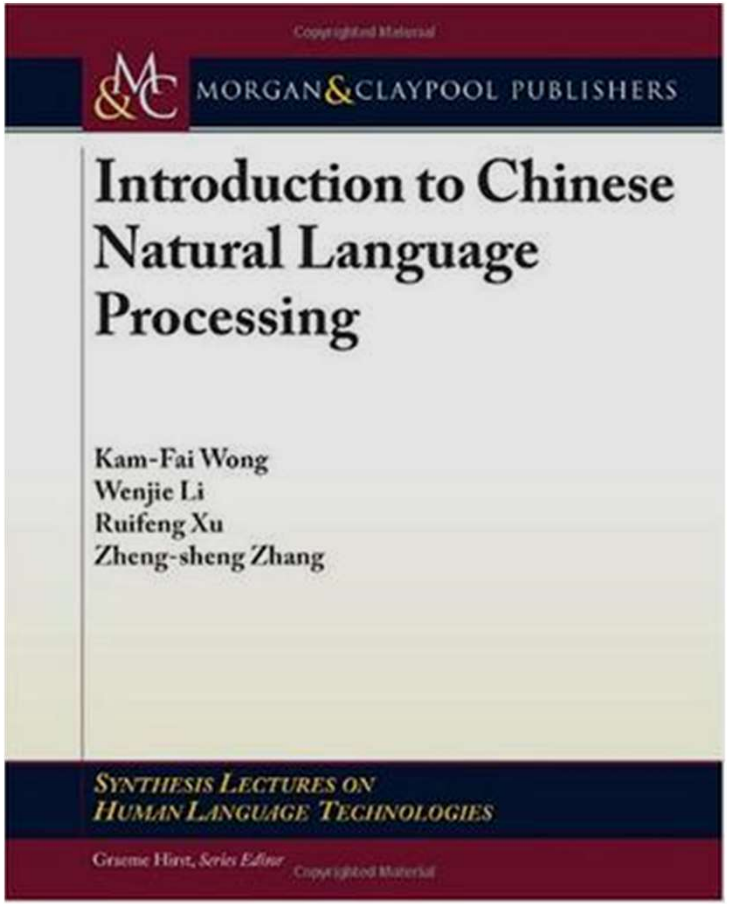
Textbook



FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

- Christopher D. Manning and Hinrich Schutze, **Foundations of Statistical Natural Language Processing**, MIT Press 1999

Textbook



MORGAN&CLAYPOOL PUBLISHERS

Introduction to Chinese Natural Language Processing

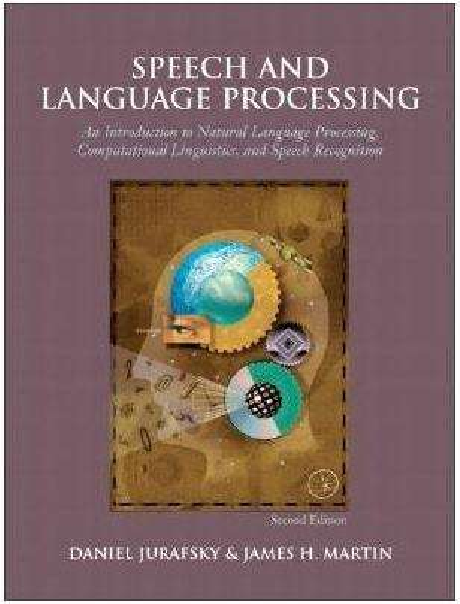
Kam-Fai Wong
Wenjie Li
Ruifeng Xu
Zheng-sheng Zhang

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, *Series Editor*

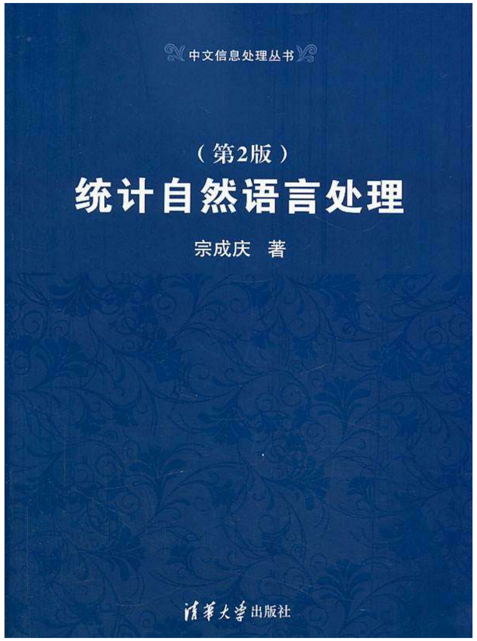
- Kam-Fai Wong, Wenjie Li, Ruifeng Xu and Zheng-sheng Zhang, **Introduction to Chinese Natural Language Processing**, Morgan and Claypool Publisher,2009

Reference book

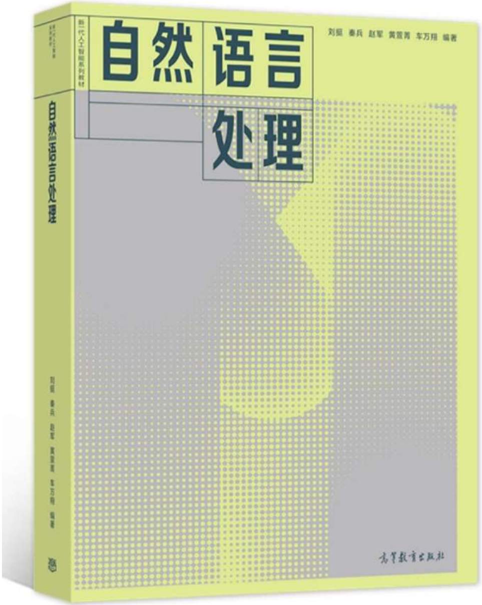

SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Daniel Jurafsky and James H. Martin. **Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.** 2nd edition. Prentice-Hall. 2008.

Reference book



- 宗成庆 《统计自然语言处理》第二版 2013



- 刘挺、秦兵、赵军、黄萱菁、车万翔 《自然语言处理》2021

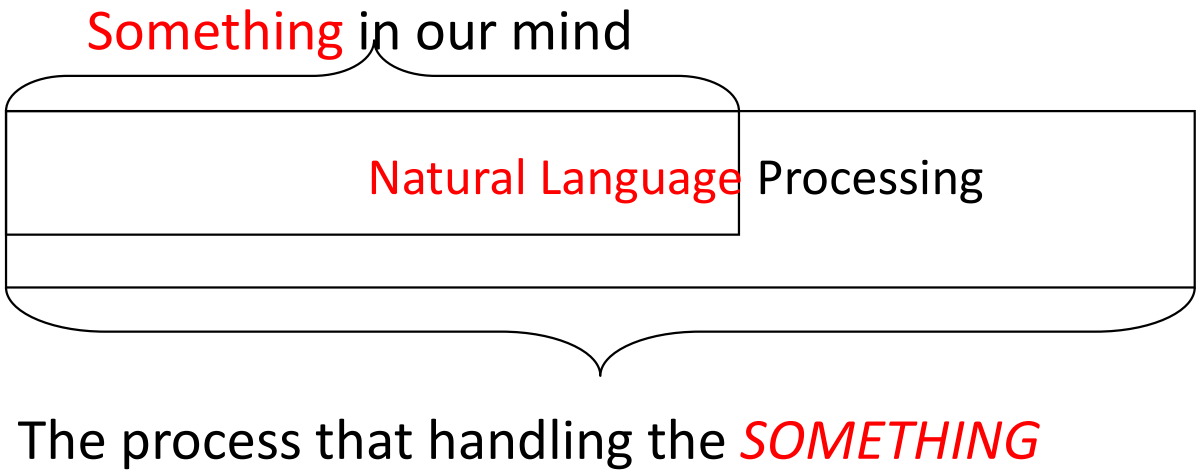
Course Grading

- Grading
  - Course Project 60%
    - System Design and Implementation
    - Grade based on
      - Performance 70%
      - Report and Demonstration 10%
      - Innovative Ideas 20%
  - Final Examination 40%
- Collaboration among students
  - Three students in each group

Questions that today’s class will answer

- What is Natural Language Processing (NLP)?
- What can NLP technique do?
- Why NLP is hard?
- Can we build system that learn from text?
- How to start NLP research
- What will this course be about?

Natural Language



Something in our mind

Natural Language Processing

The process that handling the *SOMETHING*

- **Natural language** is a language that is spoken, written, or signed by humans for general-purpose communication- wikipedia
- What kinds of things do people say?
- What do these things say/ask/request about the world?

Natural Language: Multi-media

- Text form:
  - Plain text, RTF text, Image text
- Audio form:
  - speech
- Video form, a mixed form
  - text, speech

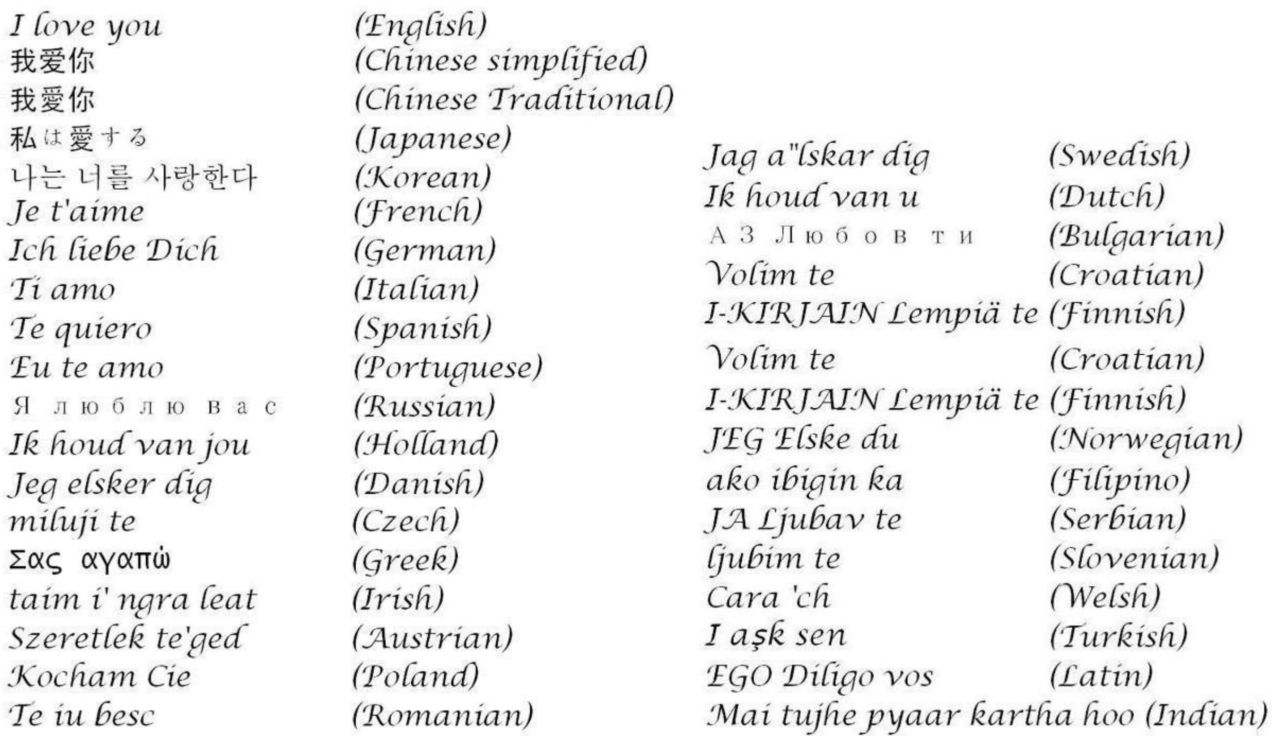
Natural Language : Multi-lingual

- Babel
- Bible Genesis

Natural Language : Multi-lingual

- We have more than 150 languages on the earth

"I love you" in 35 languages



I love you	(English)
我爱你	(Chinese simplified)
我愛你	(Chinese Traditional)
私は愛する	(Japanese)
나는 너를 사랑한다	(Korean)
Je t'aime	(French)
Ich liebe Dich	(German)
Ti amo	(Italian)
Te quiero	(Spanish)
Eu te amo	(Portuguese)
Я люблю вас	(Russian)
Ik houd van jou	(Holland)
Jeg elsker dig	(Danish)
miluji te	(Czech)
Σας αγαπώ	(Greek)
taim i' ngra leat	(Irish)
Szeretlek te'ged	(Austrian)
Kocham Cie	(Poland)
Te iu besc	(Romanian)

Jag a"lskar dig	(Swedish)
Ik houd van u	(Dutch)
АЗ Любов ти	(Bulgarian)
Volim te	(Croatian)
I-KIRJAIN Lempiä te	(Finnish)
Volim te	(Croatian)
I-KIRJAIN Lempiä te	(Finnish)
JEG Elske du	(Norwegian)
ako ibigin ka	(Filipino)
JA Ljubav te	(Serbian)
ljubim te	(Slovenian)
Cara 'ch	(Welsh)
I aşk sen	(Turkish)
EGO Diligo vos	(Latin)
Mai tujhe pyaar kartha hoo	(Indian)

Navajo Code 温州话

Natural Language: Multi-Biological

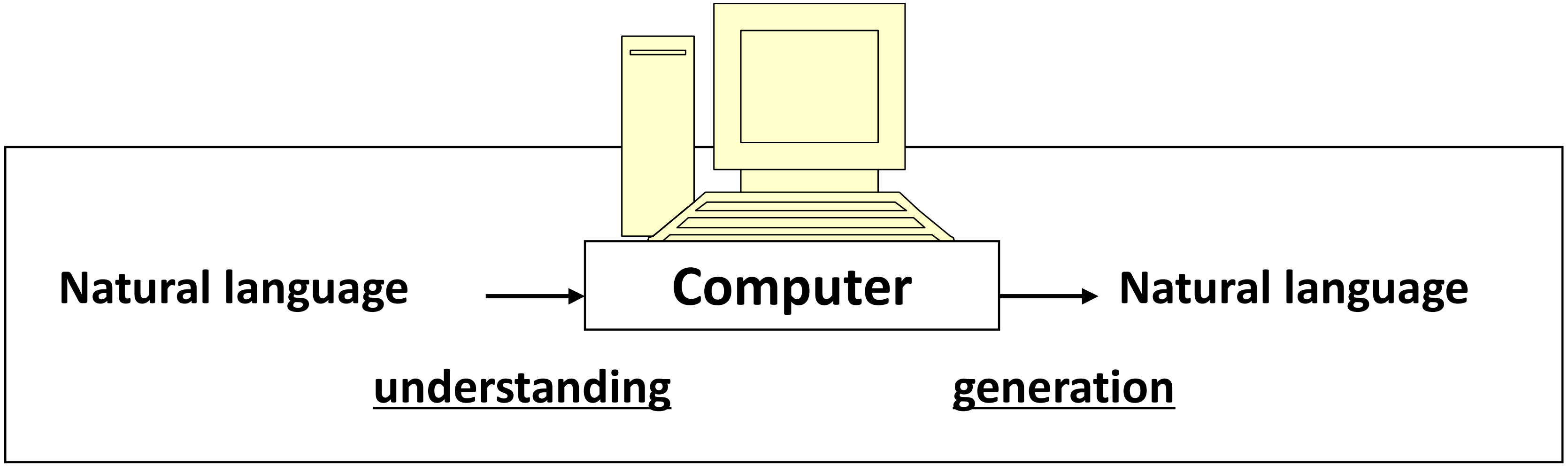
- Written language – text printed/written
- Spoken language – speech signal/oral
- Sign language – body language for the hearing disabled
- Braille – Language for the blind

Natural Language: Historical

- Characters/Words are historical
  - 谏
  - 彩　喝彩
  - 痛哭流涕
- Grammar are historical
  - 人不知而不愠，不亦君子乎？
- Pronunciation
  - 洛阳正音 雅音　东周 （衣冠南渡）
  - 洛阳正音 隋 唐 宋
  - 《洪武正韵》 明 南京话 中原雅音+江淮官话
  - 北京官话　清

Natural Language Processing

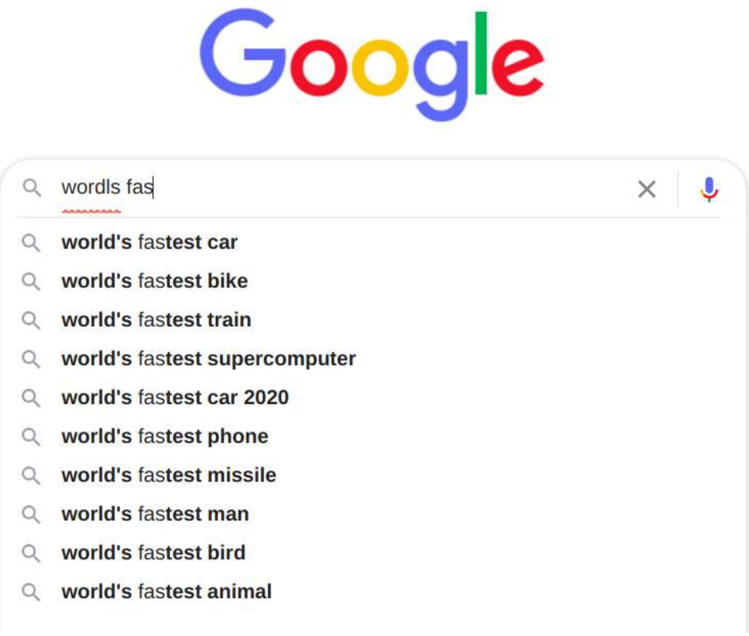
- Natural language processing (NLP) is the study of the problems of automated processing, understanding and generation of natural human languages.



Natural Language Processing Applications

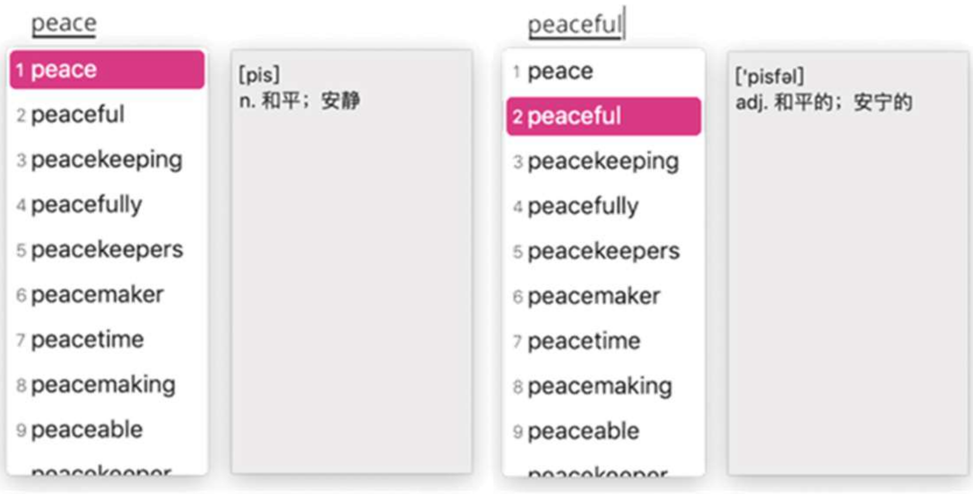
- Machine Translation
- Information Extraction
- Text to Speech System
- Speech Recognition
- Information Retrieval
- Question Answering
- Dialogue System
- Sentiment Analysis
- …

Autocorrect and Autocomplete



wordls fas

- world's fastest car
- world's fastest bike
- world's fastest train
- world's fastest supercomputer
- world's fastest car 2020
- world's fastest phone
- world's fastest missile
- world's fastest man
- world's fastest bird
- world's fastest animal



peace

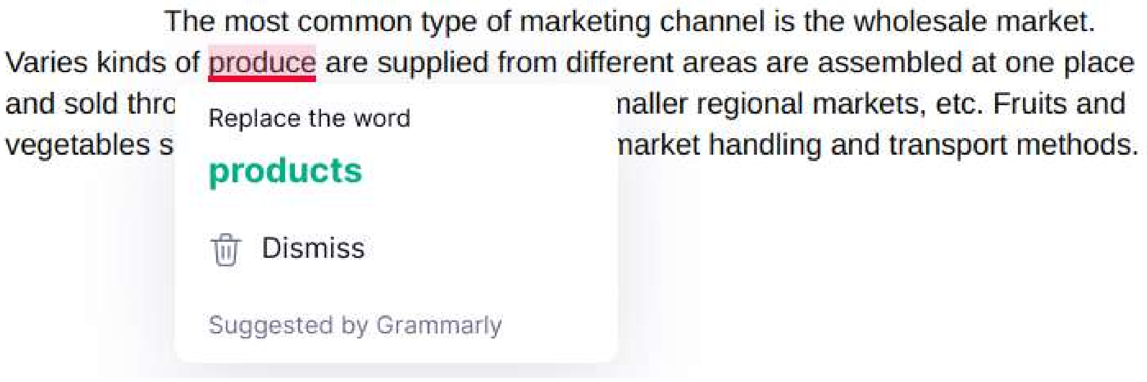
1 peace
2 peaceful
3 peacekeeping
4 peacefully
5 peacekeepers
6 peacemaker
7 peacetime
8 peacemaking
9 peaceable
peacekeeper

[pis]
n. 和平；安静

peaceful

1 peace
2 peaceful
3 peacekeeping
4 peacefully
5 peacekeepers
6 peacemaker
7 peacetime
8 peacemaking
9 peaceable
peacekeeper

['pisfəl]
adj. 和平的；安宁的



The most common type of marketing channel is the wholesale market.
Varies kinds of produce are supplied from different areas are assembled at one place
and sold thro naller regional markets, etc. Fruits and
vegetables s narket handling and transport methods.

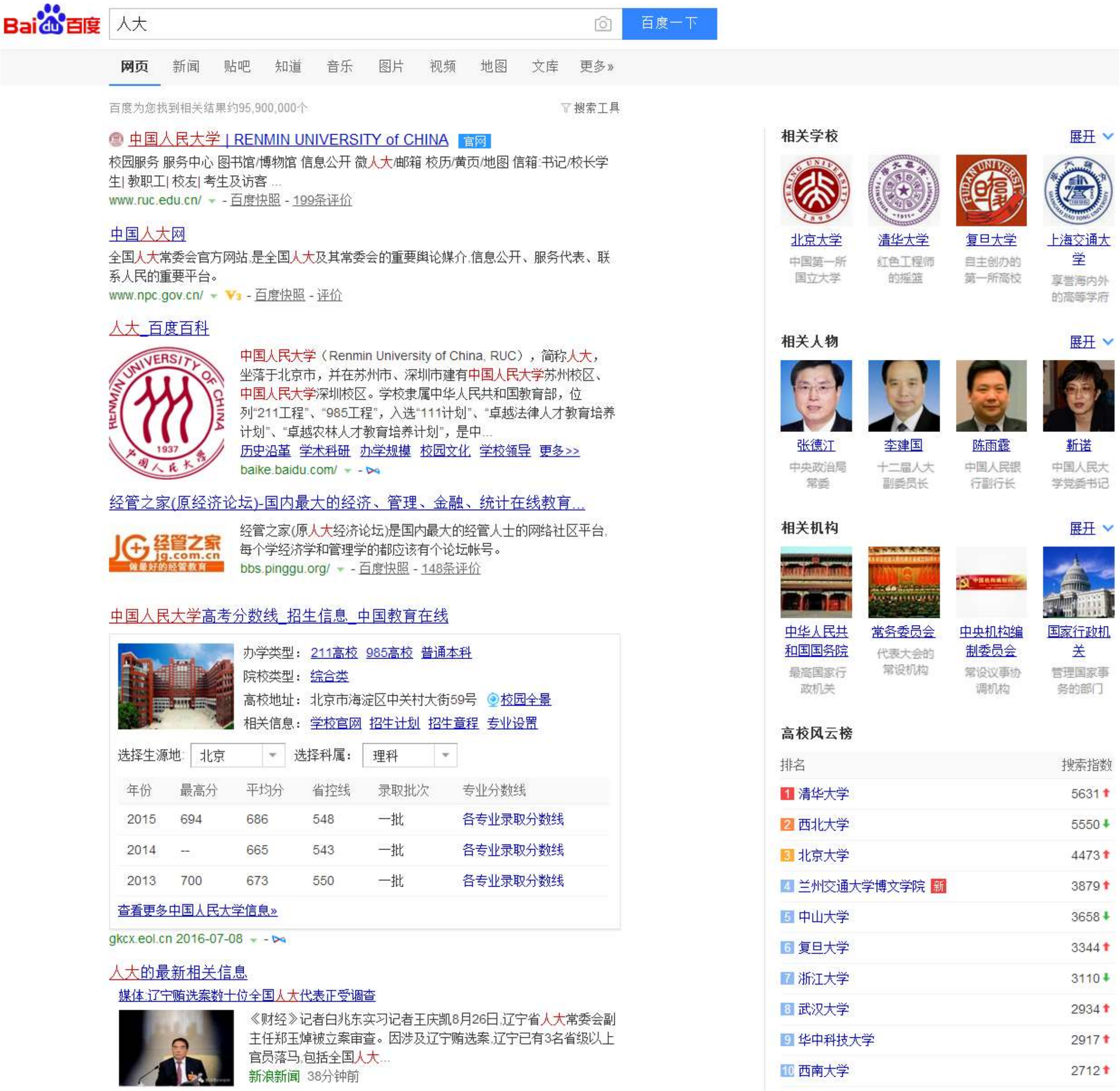
Replace the word

products

Dismiss

Suggested by Grammarly

Information Retrieval



人大

百度为您找到相关结果约95,900,000个

中国人民大学 | RENMIN UNIVERSITY of CHINA 官网

校园服务 服务中心 图书馆/博物馆 信息公开 微人大/邮箱 校历/黄页/地图 信箱:书记/校长学生| 教职工| 校友| 考生及访客 ...

www.ruc.edu.cn/ - 百度快照 - 199条评价

中国人大网

全国人大常委会官方网站,是全国人大及其常委会的重要舆论媒介,信息公开、服务代表、联系人民的重要平台。

www.npc.gov.cn/ V3 - 百度快照 - 评价

人大_百度百科

中国人民大学（Renmin University of China, RUC），简称人大，坐落于北京市，并在苏州市、深圳市建有中国人民大学苏州校区、中国人民大学深圳校区。学校隶属中华人民共和国教育部，位列“211工程”、“985工程”，入选“111计划”、“卓越法律人才教育培养计划”、“卓越农林人才教育培养计划”，是中...

历史沿革 学术科研 办学规模 校园文化 学校领导 更多>>

baike.baidu.com/ -

经管之家(原经济论坛)-国内最大的经济、管理、金融、统计在线教育...

经管之家(原人大经济论坛)是国内最大的经管人士的网络社区平台,每个学经济学和管理学的都应该有个论坛帐号。

bbs.pinggu.org/ - 百度快照 - 148条评价

中国人民大学高考分数线_招生信息_中国教育在线

办学类型：211高校 985高校 普通本科

院校类型：综合类

高校地址：北京市海淀区中关村大街59号 校园全景

相关信息：学校官网 招生计划 招生章程 专业设置

选择生源地: 北京 选择科属：理科

年份	最高分	平均分	省控线	录取批次	专业分数线
2015	694	686	548	一批	各专业录取分数线
2014	--	665	543	一批	各专业录取分数线
2013	700	673	550	一批	各专业录取分数线

查看更多中国人民大学信息»

gkcx.eol.cn 2016-07-08 -

人大的最新相关信息

媒体:辽宁贿选案数十位全国人大代表正受调查

《财经》记者白兆东实习记者王庆凯8月26日,辽宁省人大常委会副主任郑玉焯被立案审查。因涉及辽宁贿选案,辽宁已有3名省级以上官员落马,包括全国人大...

新浪新闻 38分钟前

相关学校 展开

北京大学 中国第一所国立大学

清华大学 红色工程师的摇篮

复旦大学 自主创办的第一所高校

上海交通大学 享誉海内外的高等学府

相关人物 展开

张德江 中央政治局常委

李建国 十二届人大副委员长

陈雨露 中国人民银行副行长

靳诺 中国人民大学党委书记

相关机构 展开

中华人民共和国国务院 最高国家行政机关

常务委员会 代表大会的常设机构

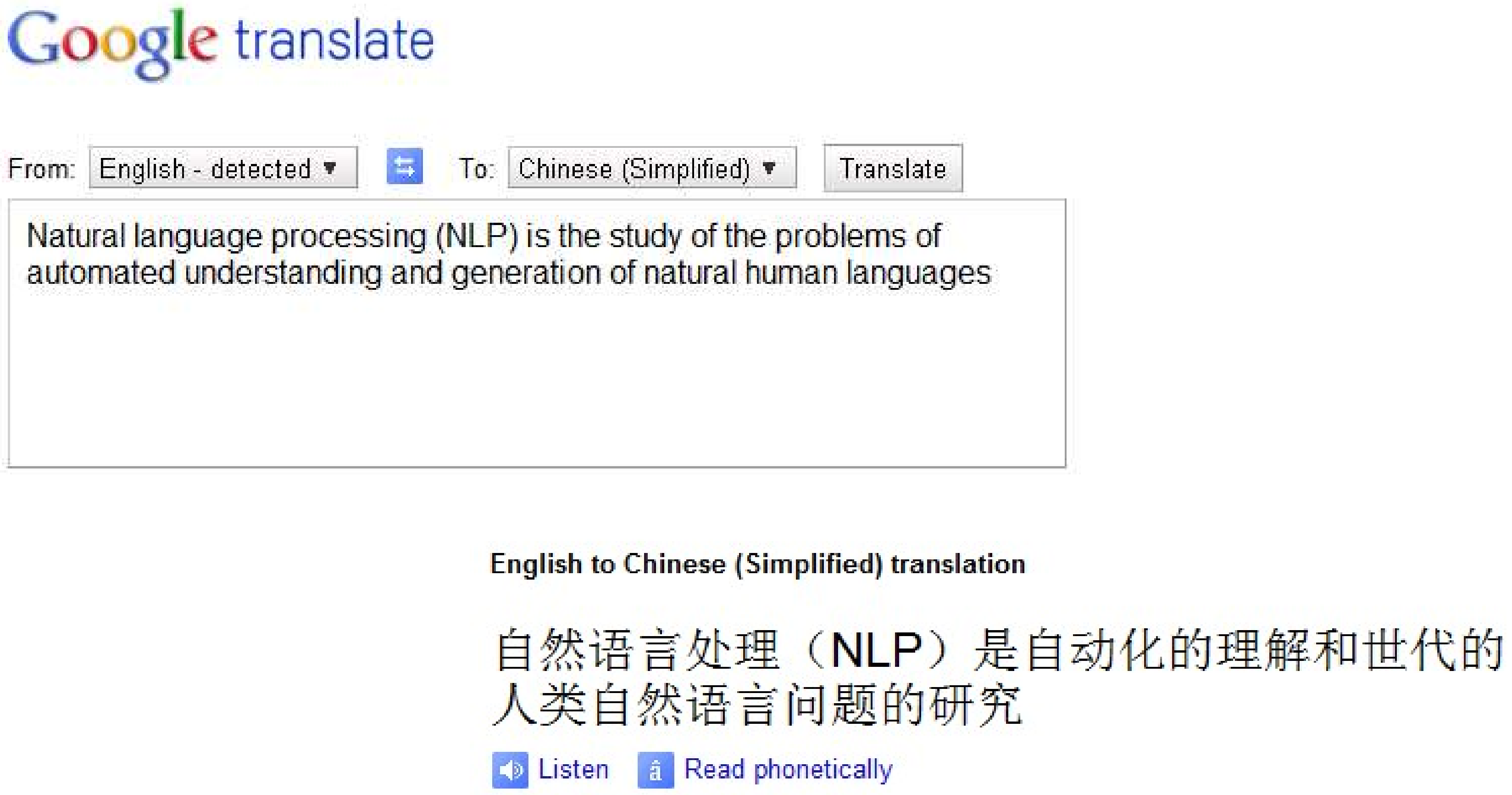
中央机构编制委员会 常设议事协调机构

国家行政机关 管理国家事务的部门

高校风云榜

排名	搜索指数
1 清华大学	5631↑
2 西北大学	5550↓
3 北京大学	4473↑
4 兰州交通大学博文学院 新	3879↑
5 中山大学	3658↓
6 复旦大学	3344↑
7 浙江大学	3110↓
8 武汉大学	2934↑
9 华中科技大学	2917↑
10 西南大学	2712↑

Machine Translation



From: English - detected ▾ To: Chinese (Simplified) ▾

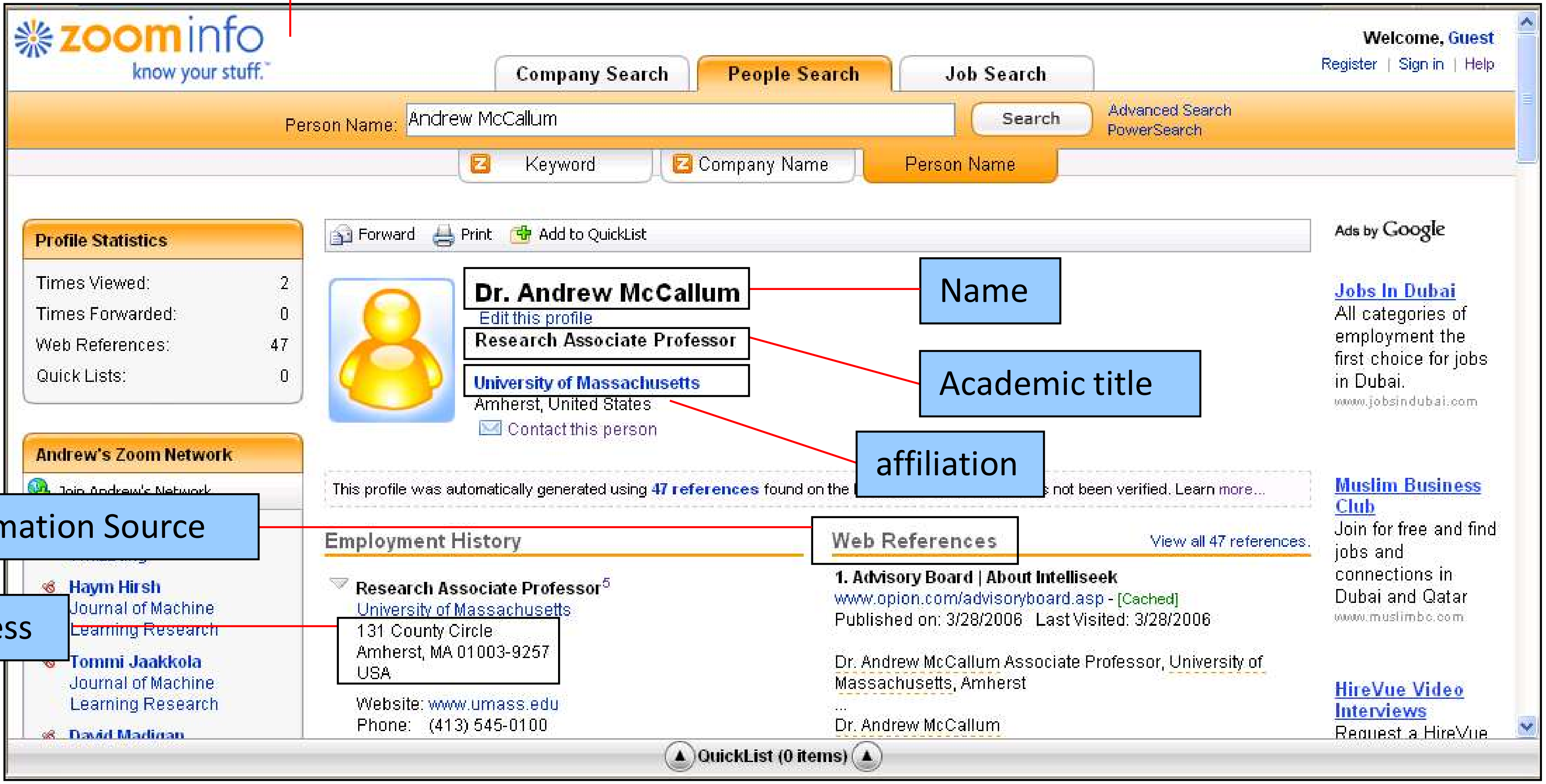
Natural language processing (NLP) is the study of the problems of automated understanding and generation of natural human languages

English to Chinese (Simplified) translation

自然语言处理（NLP）是自动化的理解和世代的人类自然语言问题的研究

Information Extraction: ZoomInfo

Information Extraction based Search Engine



Person Name: Andrew McCallum

Profile Statistics

Times Viewed:	2
Times Forwarded:	0
Web References:	47
Quick Lists:	0

Andrew's Zoom Network

Join Andrew's Network

Haym Hirsh
Journal of Machine
Learning Research

Tommi Jaakkola
Journal of Machine
Learning Research

David Madigan

Dr. Andrew McCallum
Edit this profile
Research Associate Professor
University of Massachusetts
Amherst, United States
Contact this person

This profile was automatically generated using 47 references found on the I… s not been verified. Learn more...

Employment History

Research Associate Professor[5]
University of Massachusetts
131 County Circle
Amherst, MA 01003-9257
USA
Website: www.umass.edu
Phone: (413) 545-0100

Web References View all 47 references.

1. Advisory Board | About Intelliseek
www.opion.com/advisoryboard.asp - [Cached]
Published on: 3/28/2006 Last Visited: 3/28/2006

Dr. Andrew McCallum Associate Professor, University of Massachusetts, Amherst
...
Dr. Andrew McCallum

Jobs In Dubai
All categories of employment the first choice for jobs in Dubai.
www.jobsindubai.com

Muslim Business Club
Join for free and find jobs and connections in Dubai and Qatar
www.muslimbc.com

HireVue Video Interviews
Request a HireVue

Name

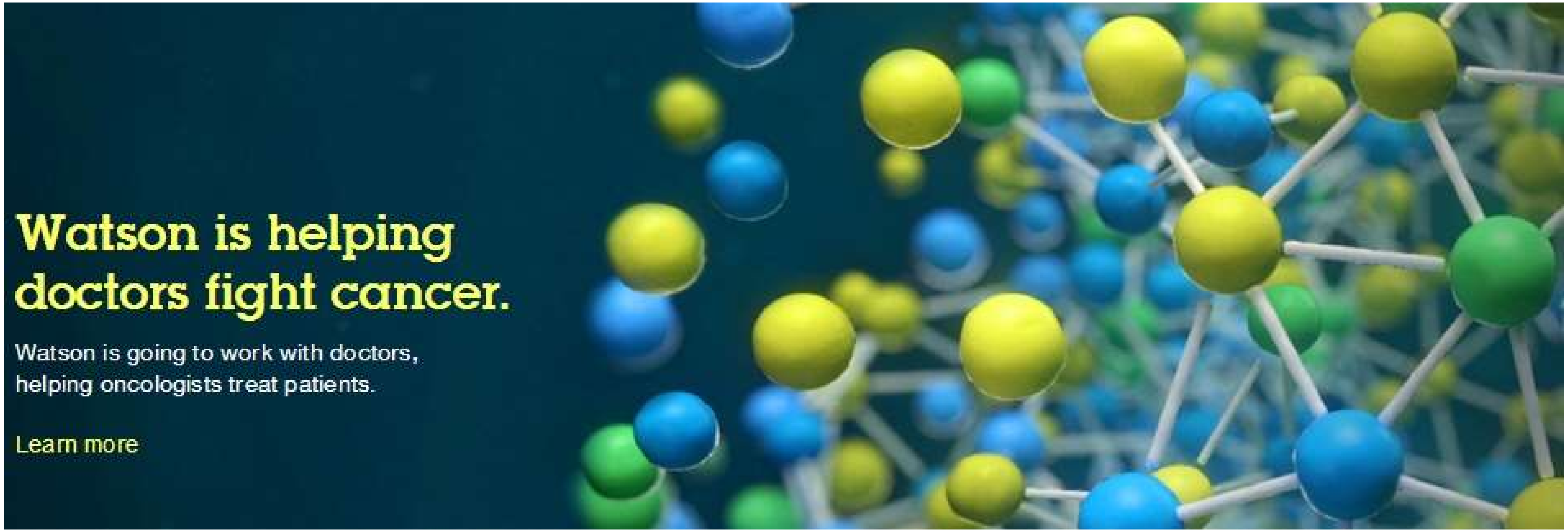
Academic title

affiliation

Information Source

address

IBM‘s Watson: Helping Doctors Fight Cancer

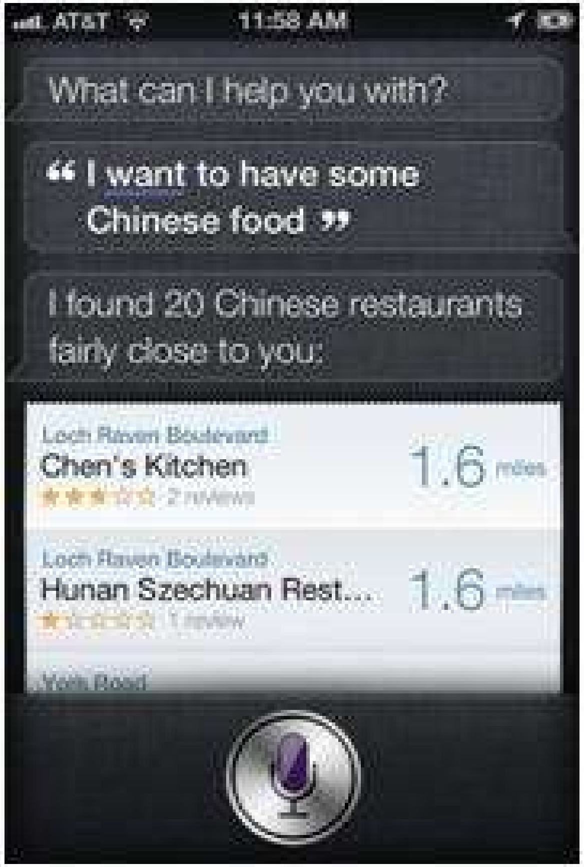


Watson is helping doctors fight cancer.

Watson is going to work with doctors, helping oncologists treat patients.

•One in five diagnoses are incorrect or incomplete and nearly 1.5 million medication errors are made in the US every year.
•Watson uses natural language capabilities, hypothesis generation, and evidence-based learning to support medical professionals as they make decisions.
•Watson will then provide a list of potential diagnoses along with a score that indicates the level of confidence for each hypothesis.

Personalized Intelligent Assistant : Siri



What can I help you with?

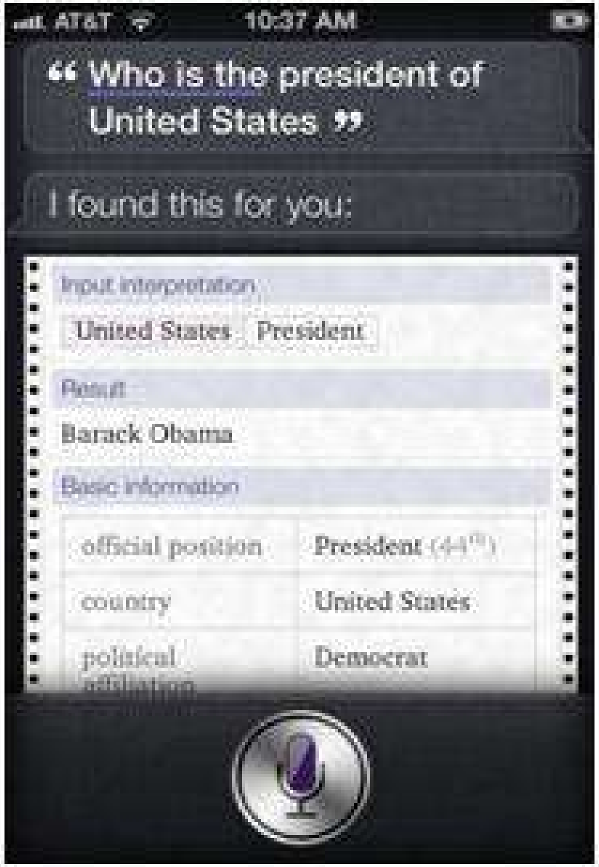
“ I want to have some Chinese food ”

I found 20 Chinese restaurants fairly close to you:

Loch Raven Boulevard
Chen's Kitchen 1.6 miles
★★★☆☆ 2 reviews

Loch Raven Boulevard
Hunan Szechuan Rest... 1.6 miles
★☆☆☆☆ 1 review

York Road



“ Who is the president of United States ”

I found this for you:

Input interpretation

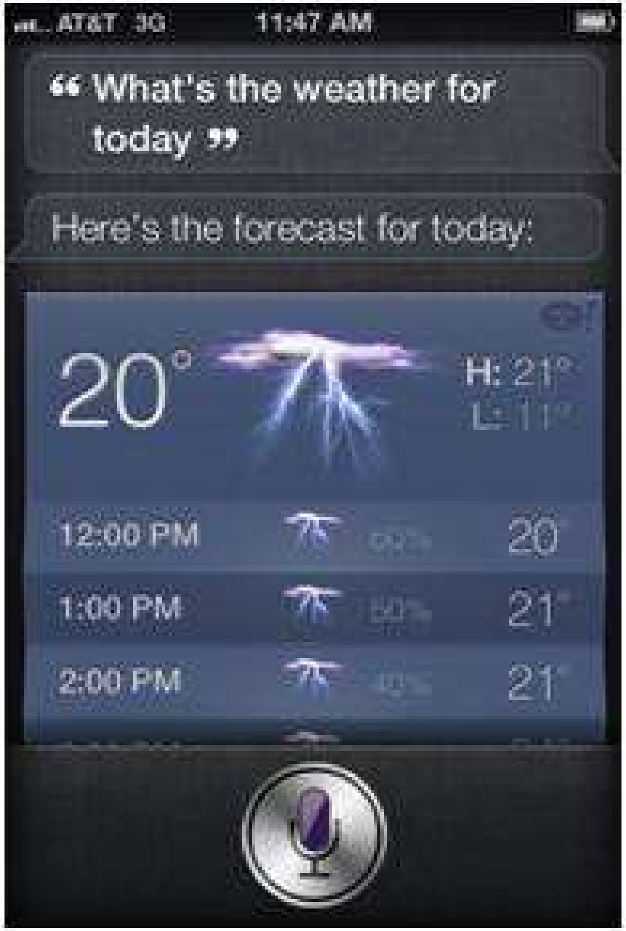
United States | President

Result

Barack Obama

Basic information

official position	President (44th)
country	United States
political affiliation	Democrat



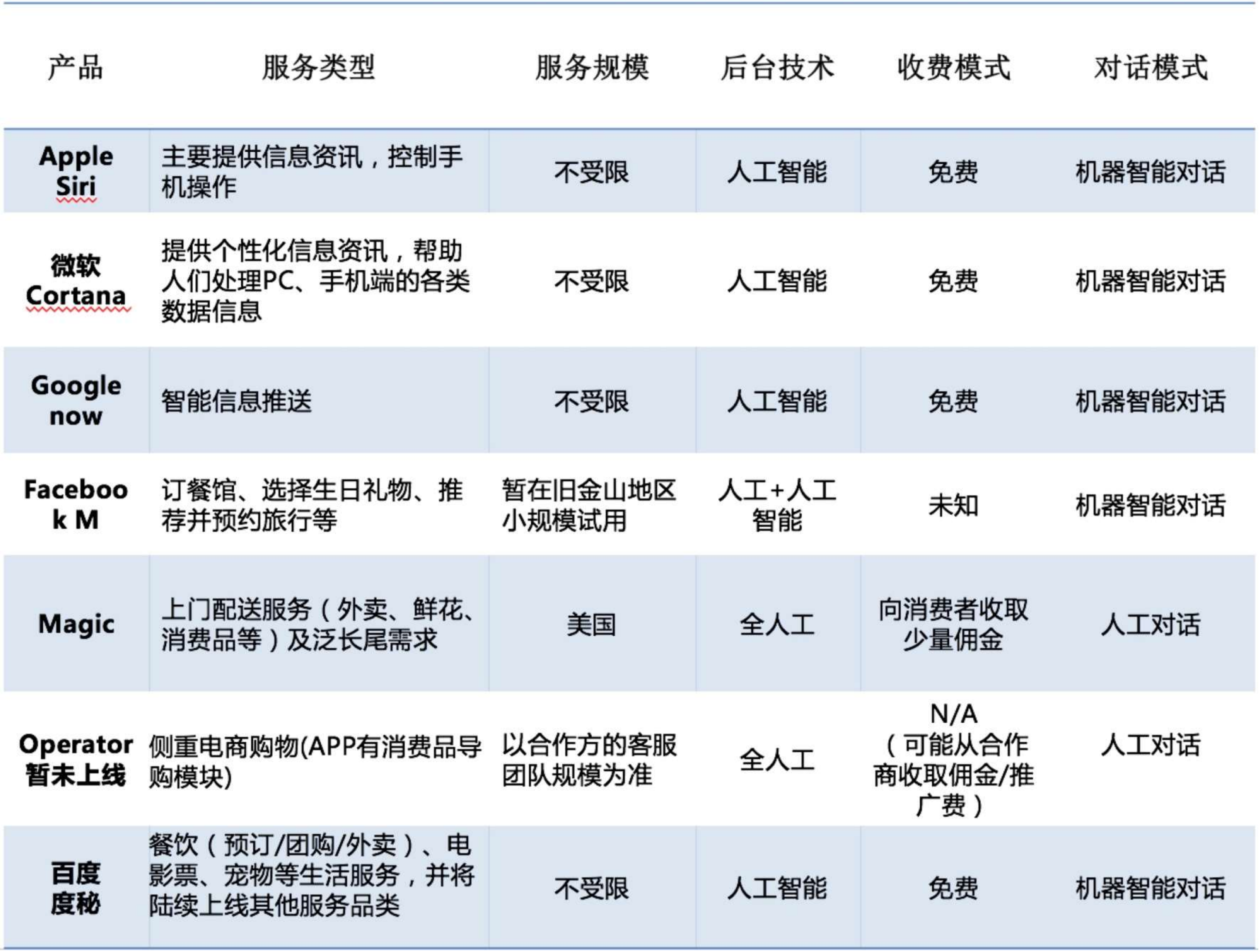
“ What's the weather for today ”

Here's the forecast for today:

20° H: 21° L: 11°

12:00 PM	60%	20°
1:00 PM	50%	21°
2:00 PM	40%	21°

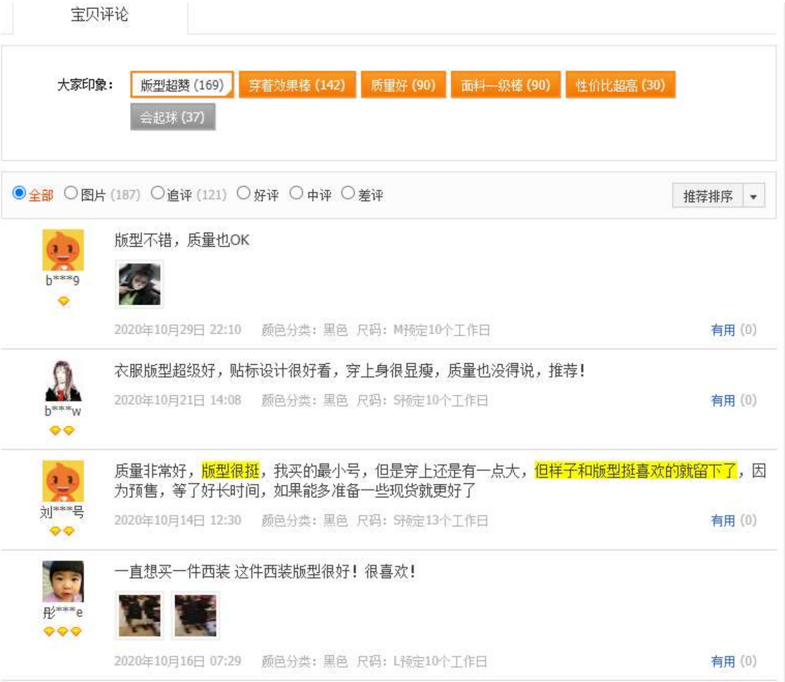
Personalized Intelligent Assistant :



产品	服务类型	服务规模	后台技术	收费模式	对话模式
Apple Siri	主要提供信息资讯，控制手机操作	不受限	人工智能	免费	机器智能对话
微软 Cortana	提供个性化信息资讯，帮助人们处理PC、手机端的各类数据信息	不受限	人工智能	免费	机器智能对话
Google now	智能信息推送	不受限	人工智能	免费	机器智能对话
Facebook M	订餐馆、选择生日礼物、推荐并预约旅行等	暂在旧金山地区小规模试用	人工+人工智能	未知	机器智能对话
Magic	上门配送服务（外卖、鲜花、消费品等）及泛长尾需求	美国	全人工	向消费者收取少量佣金	人工对话
Operator 暂未上线	侧重电商购物(APP有消费品导购模块)	以合作方的客服团队规模为准	全人工	N/A（可能从合作商收取佣金/推广费）	人工对话
百度度秘	餐饮（预订/团购/外卖）、电影票、宠物等生活服务，并将陆续上线其他服务品类	不受限	人工智能	免费	机器智能对话

Comment Keywords Extraction and Categorization

Taobao.com



宝贝评论

大家印象： 版型超赞 (169) 穿着效果棒 (142) 质量好 (90) 面料一级棒 (90) 性价比超高 (30) 会起球 (37)

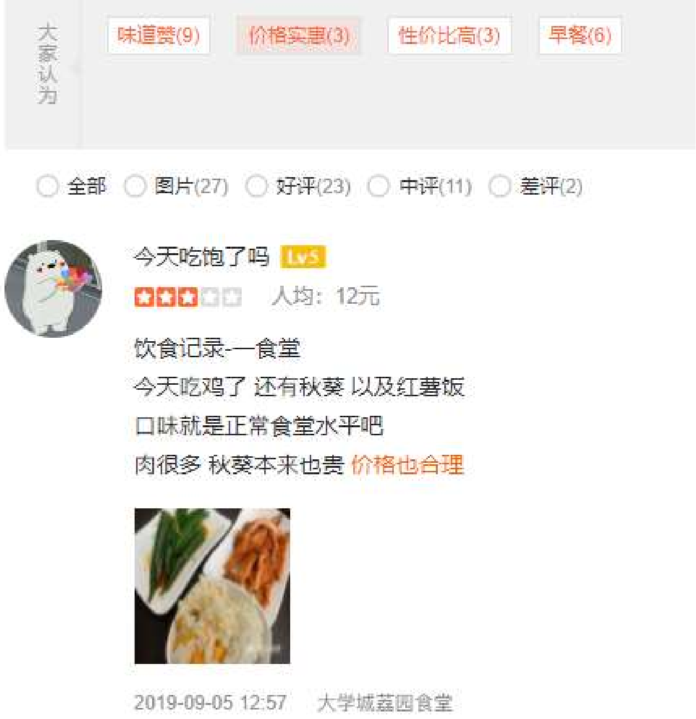
● 全部 ○ 图片 (187) ○ 追评 (121) ○ 好评 ○ 中评 ○ 差评 推荐排序

b***9
版型不错，质量也OK
2020年10月29日 22:10 颜色分类：黑色 尺码：M预定10个工作日 有用 (0)

b***w
衣服版型超级好，贴标设计很好看，穿上身很显瘦，质量也没得说，推荐！
2020年10月21日 14:08 颜色分类：黑色 尺码：S预定10个工作日 有用 (0)

刘***号
质量非常好，版型很挺，我买的最小号，但是穿上还是有一点大，但样子和版型挺喜欢的就留下了，因为预售，等了好长时间，如果能多准备一些现货就更好了
2020年10月14日 12:30 颜色分类：黑色 尺码：S预定13个工作日 有用 (0)

彤***e
一直想买一件西装 这件西装版型很好！很喜欢！
2020年10月16日 07:29 颜色分类：黑色 尺码：L预定10个工作日 有用 (0)



大家认为 味道赞(9) 价格实惠(3) 性价比高(3) 早餐(6)

○ 全部 ○ 图片(27) ○ 好评(23) ○ 中评(11) ○ 差评(2)

今天吃饱了吗 Lv5
人均：12元

饮食记录—一食堂
今天吃鸡了 还有秋葵 以及红薯饭
口味就是正常食堂水平吧
肉很多 秋葵本来也贵 价格也合理

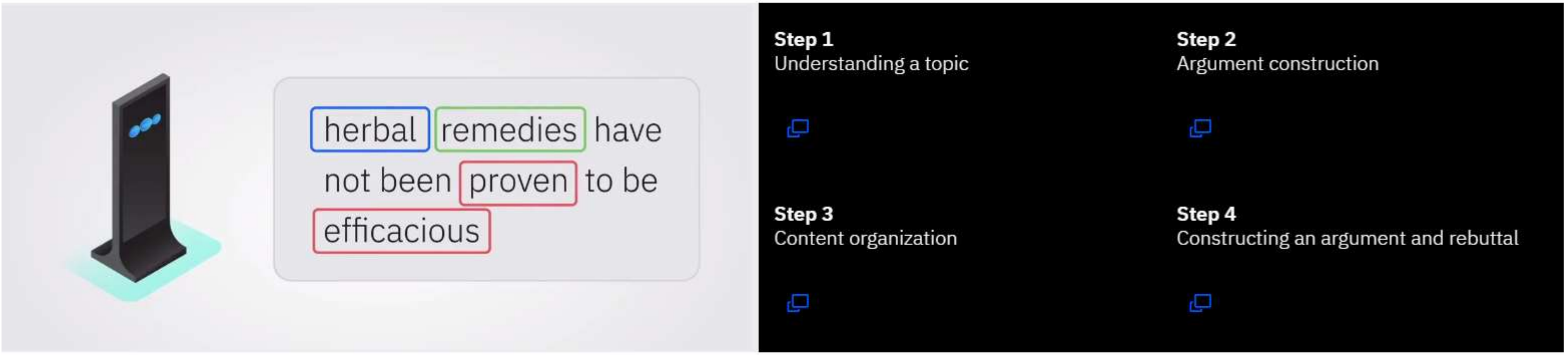
2019-09-05 12:57 大学城荔园食堂

大众点评
发现品质生活

IBM’s Project Debater

How it works?

The basics



herbal remedies have not been proven to be efficacious

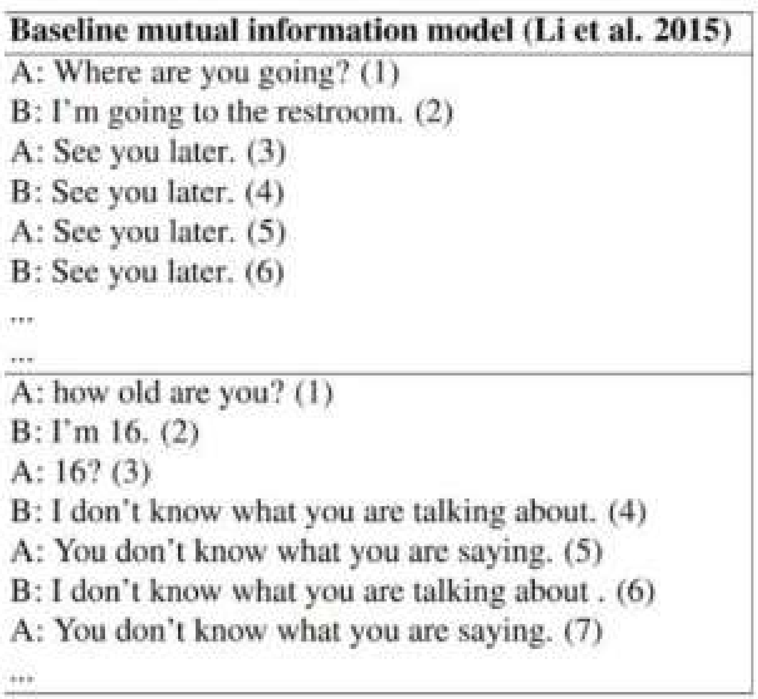
Step 1
Understanding a topic

Step 2
Argument construction

Step 3
Content organization

Step 4
Constructing an argument and rebuttal

Where Are We Now?



Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...

A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about . (6)
A: You don't know what you are saying. (7)
...

Vs.

JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON
her
"HER" DIRECTED BY SPIKE JONZE WRITTEN BY SPIKE JONZE
STARRING JOAQUIN PHOENIX AMY ADAMS SCARLETT JOHANSSON CHRIS PRATT AND ROONEY MARA
UNHOLLY-WOOD.COM

What we have now

What we expect

Why NLP is Hard?

Many “words”, many “phenomena”, many “rules”

- Oxford English Dictionary: 400k words;
- sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!
- irregularity (exceptions, exceptions to the exceptions, ...)
- potato → potato es (tomato, hero,...); photo → photo s, and even: both mango → mango s or → mango es
- Adjective / Noun order: new book, electrical engineering, general regulations, flower garden, garden flower

Basic Problems

- 形态学 (Morphology)
  - Word and morphemes
- 句法(Syntax)
  - 我吃了苹果 苹果我吃了 我苹果吃了
- 语义(Semantics)
  - 这个人真牛
- 语用学（Pragmatics）
  - 我们这边好多了
- 语音学(Phonetics)
  - Hua 花化画

Ambiguity

- "At last, a computer that understands you like your mother"

  1. (*) It understands you as well as your mother understands you
  2. It understands (that) you like your mother
  3. It understands you as well as it understands your mother

  1 and 3: Does this mean well, or poorly?

Ambiguity at Many Levels

- At the acoustic level (speech recognition):

  1. “ ... a computer that understands you like your mother”
  2. “ ... a computer that understands you lie cured your mother”

Ambiguity at Many Levels

- 赵元任《施氏食狮史》

石室诗士施氏，嗜狮，誓食十狮。施氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮。食时，始识是十狮，实十石狮尸。试释是事。

Ambiguity at Many Levels

- Word Level in Chinese

治理解放大道路面积水

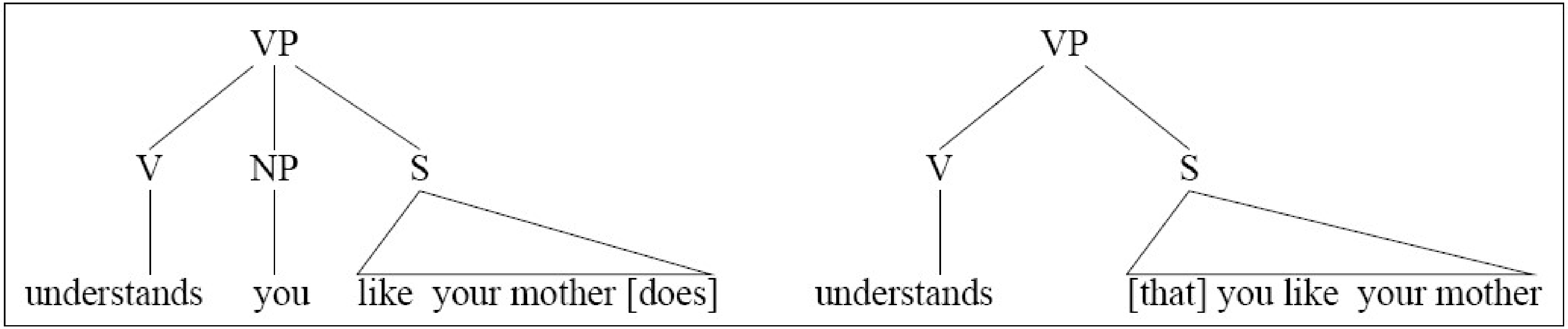
玉米们纷纷发帖支持自己的偶像

伦家酒席酱紫

莓兲想埝祢巳宬儰 1.种漝惯

Ambiguity at Many Levels

- At the syntactic level:



- Different structures lead to different interpretations.

Ambiguity at Many Levels

- At the semantic (meaning) level: Two definitions of “mother”

1. a woman who has given birth to a child
2. a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

- This is an instance of word sense ambiguity

Ambiguity at Many Levels

- At semantic (word sense) level

他说：“她真有意思”。于是人们以为他们有了意思，并让他意思意思。他说：“我根本没那个意思”。她说“你们这么说是什么意思？”，有人说“真有意思”，也有人说“真没意思”

Ambiguity at Many Levels

- At the discourse (multi-clause) level:

中国男足谁都赢不了

美国疫情好多了

Major Challenges

- Universal uncertainties: lexical, syntactic, semantic, pragmatic and phonetic levels and phonetic levels
- Unpredictability of the unknown language phenomena
- Insufficient data
- The complexity of language knowledge expression

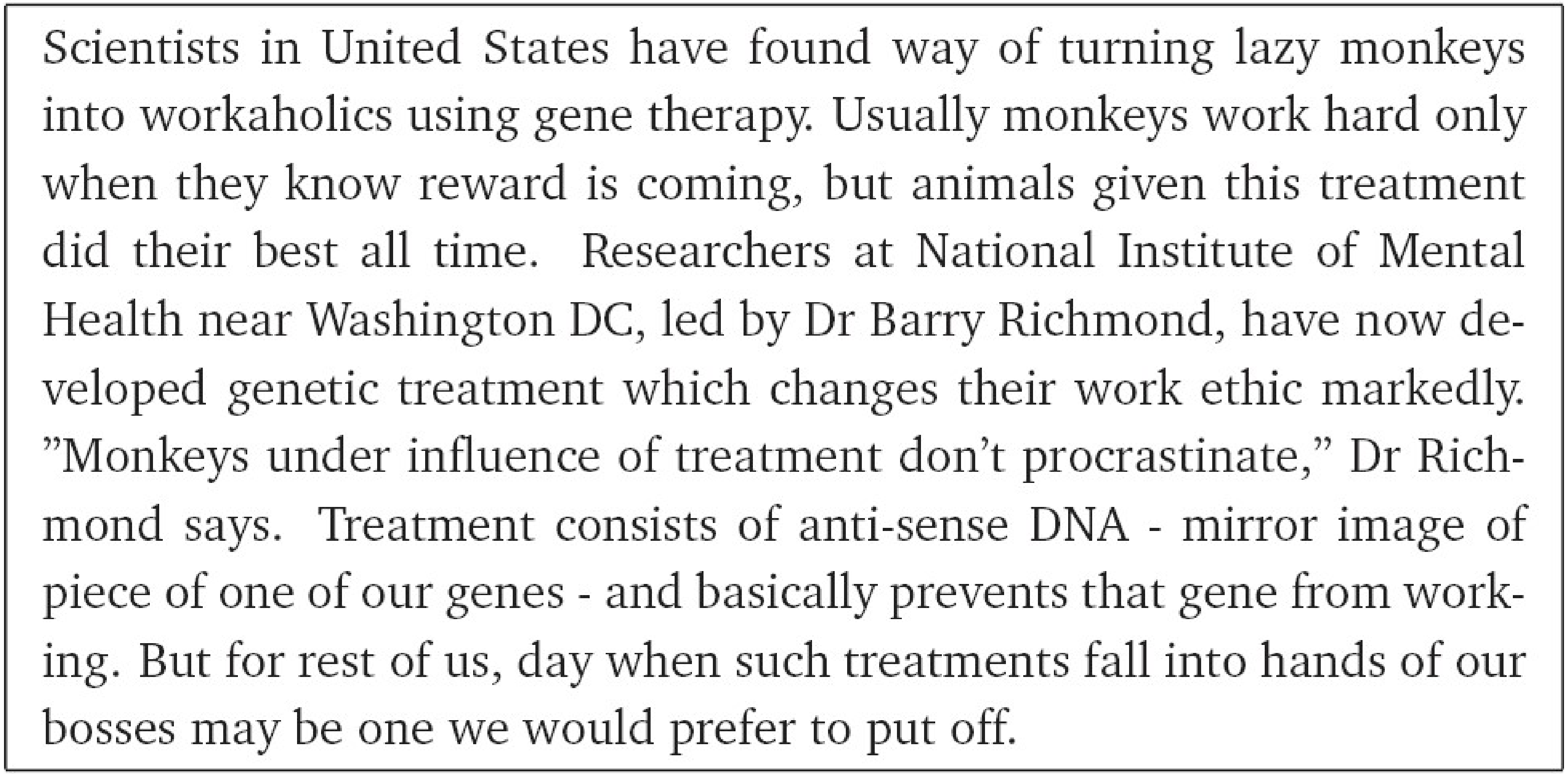
Knowledge Bottleneck in NLP

- Knowledge Required
  - Knowledge about language
  - Knowledge about the world Possible solutions:

- Symbolic approach: Encode all the required information into computer (Rationalism 理性主义)
- Statistical approach: Infer language properties from language samples (Empiricism 经验主义)

Case Study: Determiner Placement

- Task: Automatically place determiners (*a,the,null*)in a text



Scientists in United States have found way of turning lazy monkeys into workaholics using gene therapy. Usually monkeys work hard only when they know reward is coming, but animals given this treatment did their best all time. Researchers at National Institute of Mental Health near Washington DC, led by Dr Barry Richmond, have now developed genetic treatment which changes their work ethic markedly. ”Monkeys under influence of treatment don’t procrastinate,” Dr Richmond says. Treatment consists of anti-sense DNA - mirror image of piece of one of our genes - and basically prevents that gene from working. But for rest of us, day when such treatments fall into hands of our bosses may be one we would prefer to put off.

Relevant Grammar Rules

- Determiner placement is largely determined by:
  - Type of noun (countable, uncountable)
  - Reference (specific, generic)
  - Information value (given, new)
  - Number (singular, plural)

- However, many exceptions and special cases play a role:
  - The definite article is used with newspaper titles (*The Times), but zero article in names of magazines and journals (Time)* …

Symbolic Approach: Determiner Placement

- Linguistic knowledge:
  - Static knowledge: number, countability, ...
  - Context-dependent knowledge: co-reference, ...
- World knowledge:
  - Uniqueness of reference (*the current president of the US*), *type of noun (newspaper vs. magazine), situational associativity between nouns (the score of the football game), ...*

- Hard to encode the knowledge manually

Statistical Approach: Determiner Placement

- Collect a large collection of texts relevant to your domain (e.g., newspaper text)
- For each noun, compute its probability to take a certain determiner

$$p(determiner|noun) = \frac{freq(noun,determiner)}{freq(noun)}$$

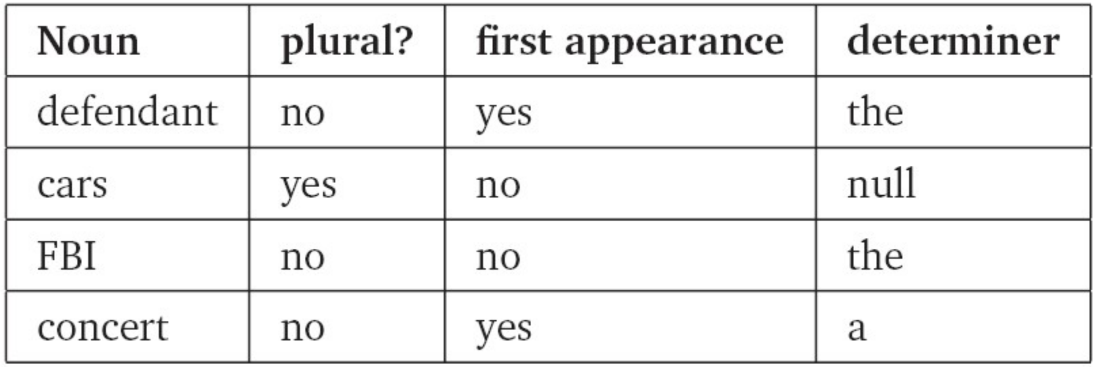
- Given a new noun, select a determiner with the highest likelihood as estimated on the training corpus

Statistical Approach: Determiner Placement

- Implementation
  - Corpus: training — first 21 sections of the Wall Street Journal (WSJ) corpus, testing – the 23th section
  - Prediction accuracy: 71.5%
- The results are not great, but surprisingly high for such a simple method
  - A large fraction of nouns in this corpus always appear with the same determiner
  - *“the FBI”,“the defendant”, ...*

Classification Approach: Determiner Placement

- Prediction: “the”, “a”, “null”
- Representation of the problem:
  - plural? (yes, no)
  - first appearance in text? (yes, no)
  - noun (members of the vocabulary set)



Noun	plural?	first appearance	determiner
defendant	no	yes	the
cars	yes	no	null
FBI	no	no	the
concert	no	yes	a

- Goal: Learn classification function that can predict unseen examples

Classification Approach: Determiner Placement

- Learn a function from $X \rightarrow Y$
  (in the previous example, {−1,0,1})
- Assume there is some distribution $D(X, Y)$, where $x \in X$, and $y \in Y$
- Attempt to explicitly model the distribution $D(X, Y)$ and $D(X|Y)$
- Predication Accuracy : 87.3%

A brief history of NLP

- 1940s: a pre-history
- 1950s: the over-enthusiasm time
- 1960s: the cold-water showering time
- 1970s: the NLP history just started
- 1980s: the time to harvest a bit
- 1990s: the time for statistical approaches to speak
- 2000s: the matter-of-fact re-evaluation of NLP role to human beings: an assistant
- 2010s : deep learning age?

1940s: a pre-history

- 1946: Warren Weaver started the earliest MT project to break enemy codes during World War II.
- 1949: Weaver put the idea of MT to his famous *memorandum*, which inspired many projects to break new ground.
- But people soon realized they were not playing an easy game.

1950s: over-enthusiasm

- Bar-Hillel predicted that FAHQT is impossible without knowledge (1951).

  Then linguists joined the game to explore the linguistic theories.
- Over-enthusiasm dominated the MT researchers
- 1957: Noam Chomsky published his theory on *Syntactic Structures*, which is later referred to as *Generative Grammar*.
- The great debate between rationalists and empiricists:
  - language processing camp dominated by the symbolic generative grammar
  - speech camp dominated by the statistical information theory.
- Introduction of pre- and post-editing

1960s: cold-water showering

- US funding reached $20 million by the mid 1960s.
  - But no encouraging output was made.
- ALPAC report concluded that "there had been no machine translation of general scientific text, and none is in immediate prospect".

  The spirit is willing but the flesh is weak

  English ->Russian-> English

  The wine is good but the meat is spoiled
- US funding for MT stopped, so did the most other NLP work. NLP research entered something of a dormant phase.
- Over-selling stopped, more realistic work on theoretical representation of meaning in late 1960s

1970s: the NLP history just started

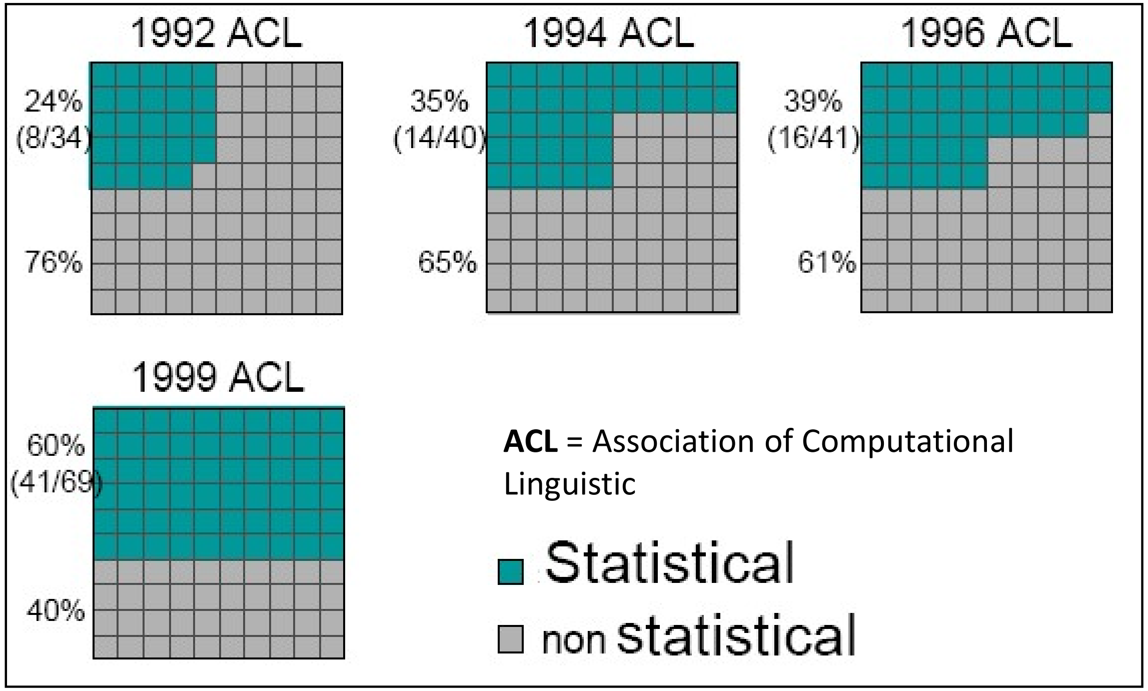
- Theoretical study on computational linguistics continued
- 1965: Chomsky introduced the *transformational model* followed by
  - Fillmore: case grammar;
  - Quillian: semantic networks
  - Schank: conceptual dependency theory;
  - Woods: augmented transition networks
  - Wilks: preference semantics
- Distinguished NLP systems
  - Weizenbaum's ELIZA: about psychologist and a patient
  - Woods' LUNAR: a natural language interface

1980s: the time to harvest a bit

- Symbolic approaches still dominated in significant NLP problems and statistical approaches stayed complementary
- Critical computational resources (hardware, software, data, knowledge base, etc) are available
- Researchers started re-examining non-symbolic approaches
- Successes in speech recognition (IBM) with substantial performance gains over knowledge-based systems

1990s: statistical dominates

- IBM’s statistical machine translation system
- Rise of the Web emphasized the need for language based information retrieval and information extraction



ACL = Association of Computational Linguistic

2000s: a matter-of-fact role

- Machine translation: CAT tools commercialized
- Content management: word segmentation, POS tagging, sentiment analysis, classification
- Search engine: information extraction, a *vertical* solution
- Speech recognition: language model
- Data mining: language text manipulation, ontology

2000s: a matter-of-fact role

- Fred Jelinek (IBM)
- Speech Recognition
- Every time I fire a linguist, the performance of the recognizer improves.

2000s: a matter-of-fact role

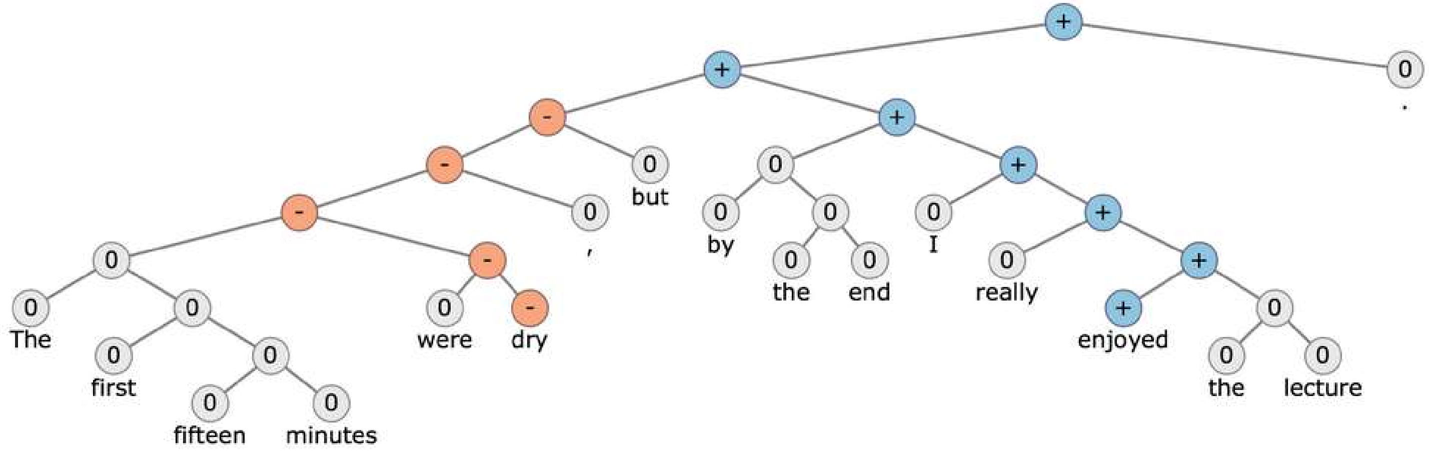
- Machine translation: Franz Josef Och
- “Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.”
- “Parallel corpus is the key for statistical machine translation.”
- Syntax, semantics are few helpful even harmful

2010s: Deep Learning Age

- How to represent the meaning of a word?
  - One-hot representation
  - Distributed representation
- Distributed representation
  - CBOW / Sikp-Gram
  - Transfer data to continuous, low-dimension, dense vectors
  - Word2Vec
    - predicting the surrounding word in a window of length c of each word
    - Maximize the log probability of any context word given the current center word

2010s: Deep Learning Age

- Sentence/Document Composition based on distributed representations
  - – CNN: Convolutional Neural Networks
  - – Recurrent Neural Networks
  - – Recursive Neural Networks



2010s

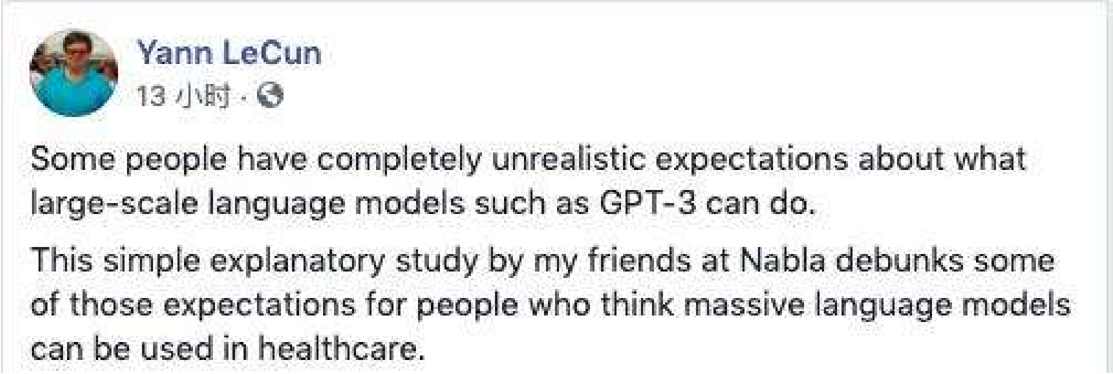
- Natural language processing → Natural language understanding
- Community knowledge
  - Wikipedia
  - SOHU Pinyin
- Semantic Parsing/Abstract Meaning Representation

2010s

- Sequence-to-sequence model
  - Good Progress in Neural Machine Translation System
  - Hard to explain/ Data requirement
  - Not fit in Dialog System
- Manning 2017
  - To a first approximation, the de facto consensus in NLP in 2017 is that no matter what the task, you throw a BiLSTM at it, with attention if you need information flow from Manning,2017
- Compositionally in models
- Neural Symbol Machine

2020 Big Model/Foundation Model

- Pretrain Language Model
  - ElMo
  - GPT
  - Bert
  - Ernie
  - Pangu



Yann LeCun

13 小时 ·

Some people have completely unrealistic expectations about what large-scale language models such as GPT-3 can do.

This simple explanatory study by my friends at Nabla debunks some of those expectations for people who think massive language models can be used in healthcare.

- GPT-3
  - 1750亿 parameters
- Knowledge Distillation
- Knowledge?

2020 Big Model/Foundation Model

1. [05 Jul 2021]   (Baidu) ERNIR 3.0: multi-task learning
2. [02 Jun 2021]   (Alibaba) VECO: cross-attention module
3. [24 Jun 2021] (Tsinghua) CPM 2.0: assemble of SOTA works
4. [18 Aug 2021] (Microsoft) DeltaLM: init transformer by BERT
5. [10 Apr 2021] (Baidu) ERNIE-M: Enhanced multilingual LM
6. [07 Apr 2021] (Microsoft) InfoXLM : cross-lingual constrast
7. [28 May 2021] (Google)ByT5: Token-free multilingual model
8. [03 Jul 2020]   (Google & Facebook)LaBsE & LaSer: score model
9. (PCL)Tongyan: Multilingual translation model

How to Start NLP Research (Zhou Ming)

- https://www.zhihu.com/question/19895141/answer/149475410
- How to grasp first technology in NLP
- Suggestions:
  - Select a open source project
  - Understand the task objective
  - Compile and run the provided demo program
  - Understanding the provided algorithm
  - Implement the provided algorithm
  - Run and compare with the demo program
  - Try to improve the demo program

How to Start NLP Research

- How to select a good research problem
- Suggestions:
  - Find a interesting research area (ACL/AAAI/arxiv)
  - Investigate the status of this research area
    - Mathematic system and machine learning frameworks
    - Available training data and evaluation data
    - Current researchers
  - Look for the available open source project/tools/code
  - Reading papers
  - Try to implement and improve the reported methods
  - Try to find general contribution

How to Start NLP Research

- How to write the first NLP paper
- Suggestions:
  - Determine the title (Clear, Deep, Technical, Interesting)
  - Abstract (problem/method/advantage/evaluation)
  - Introduction
    - Background, Problem Definition
    - Current Approaches and Problem Left
    - Motivation of this study
    - Idea of this study
    - Experimental results
    - Contributions
    - Organization of this paper

How to Start NLP Research

- Related works
  - Camp existing methods into several approaches
  - Typical works for each approach
  - Limitation
- Our approach
  - Framework, Definitions
    - Figure
  - Details of our approach
    - Approach
    - Component

How to Start NLP Research

- Experimental results and discussions
  - Experiment Setting
    - Dataset
    - Parameters
    - Evaluation Metrics
    - Baselines
  - Comparing experimental results
    - Overall performance
    - Component Contribution
    - Case study
    - Negative results
  - Discussions
- Conclusions
- Reference

Academic Organization/Top Conferences

- Association for Computational Linguistics
  - aclweb.org
  - ACL
  - EMNLP
  - NAACL/ECAL/AACL/CoNLL/HLT
- International Committee on Computational Linguistics
  - http://nlp.shef.ac.uk/iccl/
  - COLING
- ACL Anthology
  - http://aclanthology.info/

Other Conferences

- ACM
  - SIGIR, WWW, WSDM
- AI
  - AAAI
  - IJCAI
- Machine Learning
  - ICML
  - NeurIPS (NIPS)
  - ICLR

NLP Organization and Conferences (China)

- CIPS 中国中文信息学会
  - http://www.cipsc.org.cn/
  - CCL
  - CCIR
  - CCKS
  - SMP
  - CWMT
- CCF-TCNLP
  - http://tcci.ccf.org.cn/
  - NLPCC

Course Design

- Post-graduate students on NLP
- NLP
  - Symbolic/Knowledge-based NLP
  - Statistical NLP
  - Neural NLP
- Language
  - English / Chinese
- Three Parts:
  - Preliminaries
  - Fundamental NLP Techniques
  - Applications

Part 1: Preliminaries

- Linguistics Foundations
- Mathematics Foundations
- Machine Learning Foundations
- Deep Learning Foundations

Part 2: Fundamental NLP Techniques

- Language Models
- Word Processing
- Word Meanings and Word Sense Disambiguation
- Phrase Parsing
- Dependency Parsing

Part 3: Applications

- Text Categorization
- Text Clustering
- Information Extraction
- Sentiment Analysis

Tutorials

- Python for NLP

- Deep Learning for NLP with Python

The next lecture

- Lecture 2
  Linguistics Foundation