Natural Language Processing

Zhiyuan Liu

liuzy@tsinghua.edu.cn

THUNLP

Course Info

- Natural Language Processing

- For Master Program in Advanced Computing
- Teaching Language: English
- Third Year at Tsinghua

- Content
  - 14 Week Teaching
  - 1 Week Poster Report Session

What to Teach

- Big picture knowledge about NLP
  - The problem, the challenge, the solution
  - Why NLP moves forward from statistical learning to deep learning
  - Remaining open problems and challenges
- Ability to build practical systems for major NLP tasks
- Ability to solve open problems in NLP tasks
  - How to review related works in the given area
  - How to identify the key challenges for the task
  - How to figure out solutions to the challenge

About Me

- Zhiyuan Liu
- Research Interests
  - Natural Language Processing
  - Knowledge Graph
- Contact Info
  - Email: liuzy@tsinghua.edu.cn
  - Tel/Wechat: 138-1032-5978
  - Office: Room 4-506, FIT Building
- Webpage
  - nlp.csai.tsinghua.edu.cn/~lzy/

刘知远

北京 海淀

扫一扫上面的二维码图案，加我微信

Course Prerequisites

- Programming Skill in Python, C++, Java
  - TensorFlow, PyTorch, ...
- Advanced Mathematics
  - Multivariate Calculus
  - Linear Algebra
  - Probability and Statistics
- Fundamentals of Machine Learning
  - Clustering, classification, loss functions
  - Simple derivatives
  - Optimization with gradient descent

Course Plan

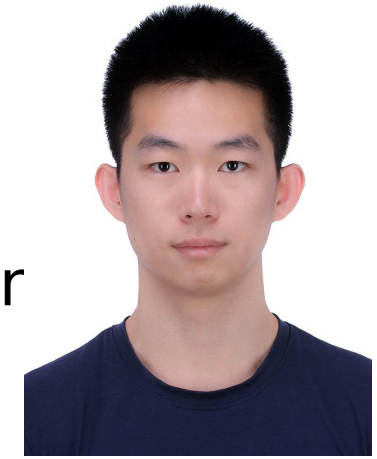
- L1 - Introduction
- L2 - Word Representation & Neural Network
- L3 - Seq2Seq Modeling
- L4 - Pre-Trained Models
- L5 - Knowledge Graph
- L6 - Information Extraction - 1
- L7 - Information Extraction - 2
- L9 - Information Retrieval
- L10 - Question Answering
- L11 - Text Generation
- L12 - Discourse Analysis
- L13 - Interdisciplinary Areas
- L14 - Future of NLP
- W16 - Report Session
- One week for holiday

Grading Policy

- Four Assignments: 15% x 4 = 60%
- Final Course Project (1–3 people): 30%
  - Including: project proposal, milestone, interacting with mentor
  - Final poster session: 10% of the 30%
- Class Presentation: 10%
- Late policy
  - 10% off per day late
  - Assignments not accepted after 3 late days per assignment
- Collaboration policy
  - Encourage collaboration and document it in the report

TA Info

- Leading TA: Yuan Yao
- PhD student
- Research Interest: Relation Extractior
- Email: yaoyuanthu@163.com



- TA: Ganqu Cui
- Master student
- Research Interest: Knowledge Graphs
- Email: cgq19@mails.tsinghua.edu.cn

- Office: Room 4-506, FIT Building

Course Resources

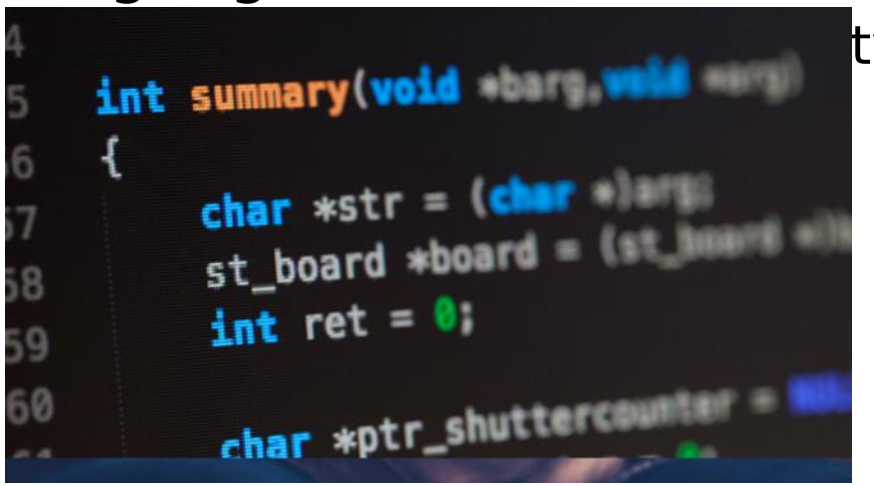
- Wechat Group: Please contact TA to join
- Online QA: https://github.com/thunlp/NLP-THU
- A textbook (in progress): ***Language, Knowledge and Learning***
  - *Proofread, feedback, and suggestions are welcome*
  - *Feedback can be sent to TAs or me via email or wechat*
  - *Contributors will be mentioned in acknowledgement, as well as bonus to course grade*

Brief Introduction to NLP

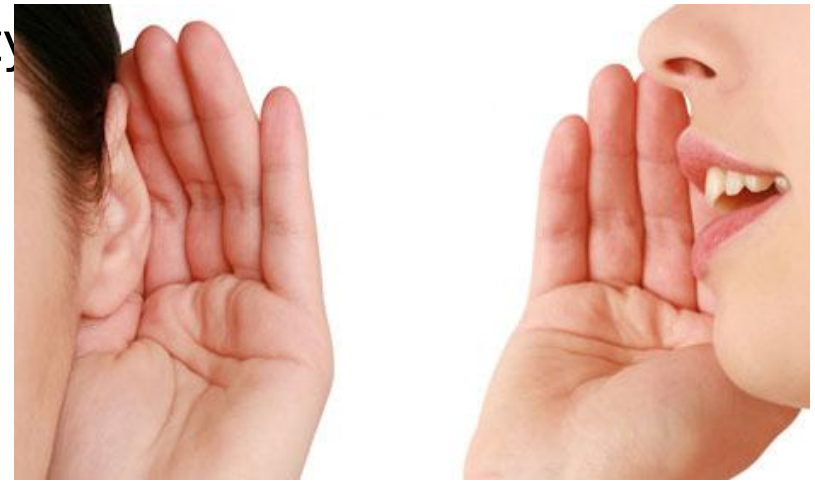
THUNLP

What is Natural Language

- Natural language is used to communicate information, thoughts, and knowledge among humans
- Natural language differs from programming languages



Programming Languages



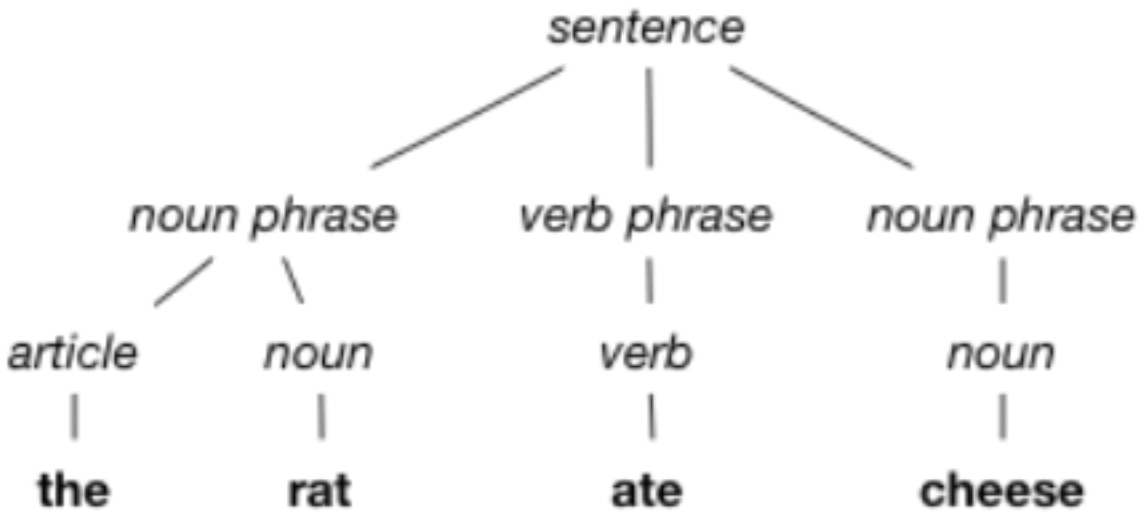
Natural Languages

What is Natural Language Processing?

Input: The rat ate cheese

Structure Prediction

Output:

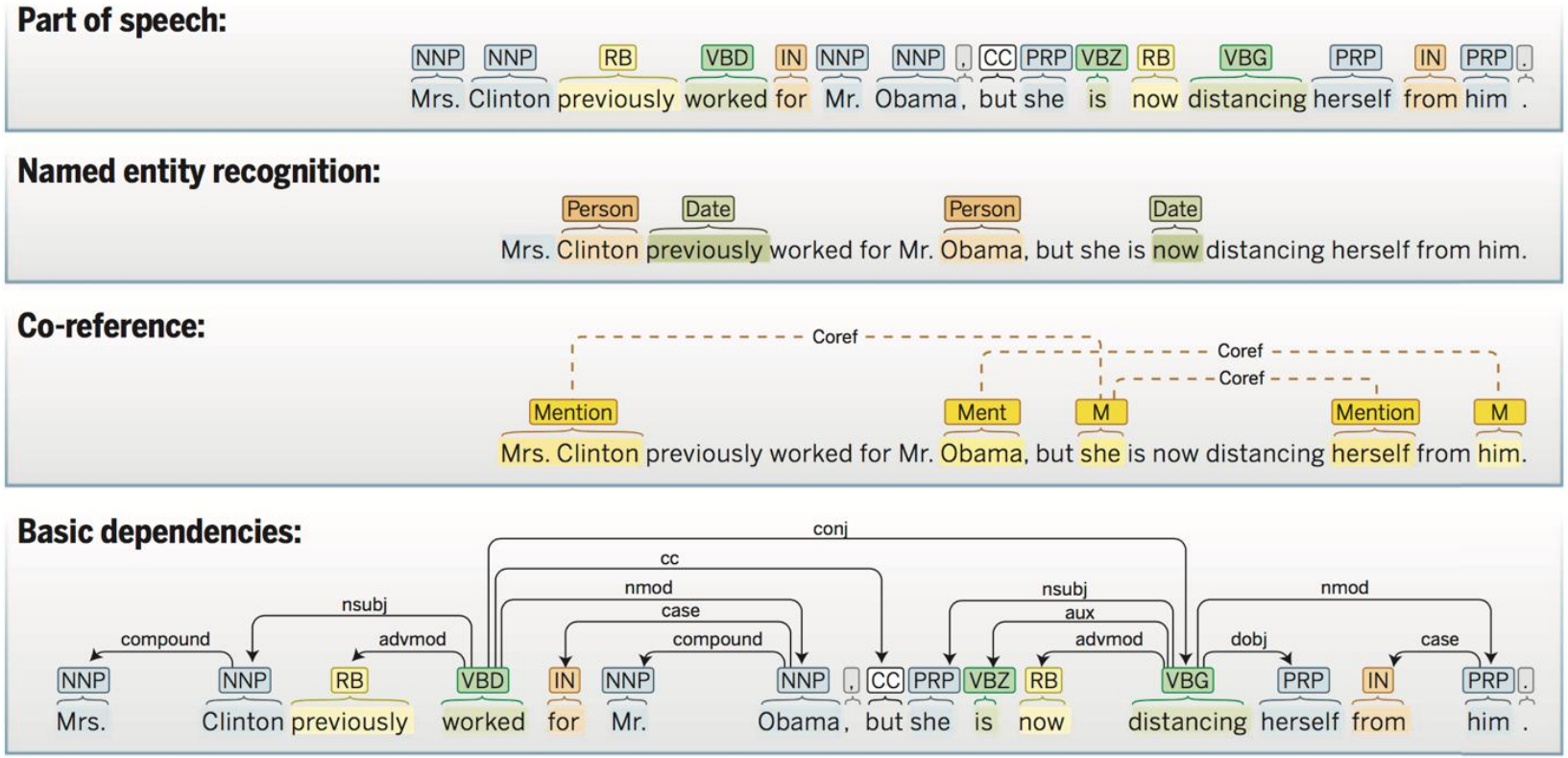


Syntactic Structure

The Nature of NLP is Structure Prediction!

What is Natural Language Processing?

- NLP aims to make computers understand languages
- The nature of NLP is structure prediction

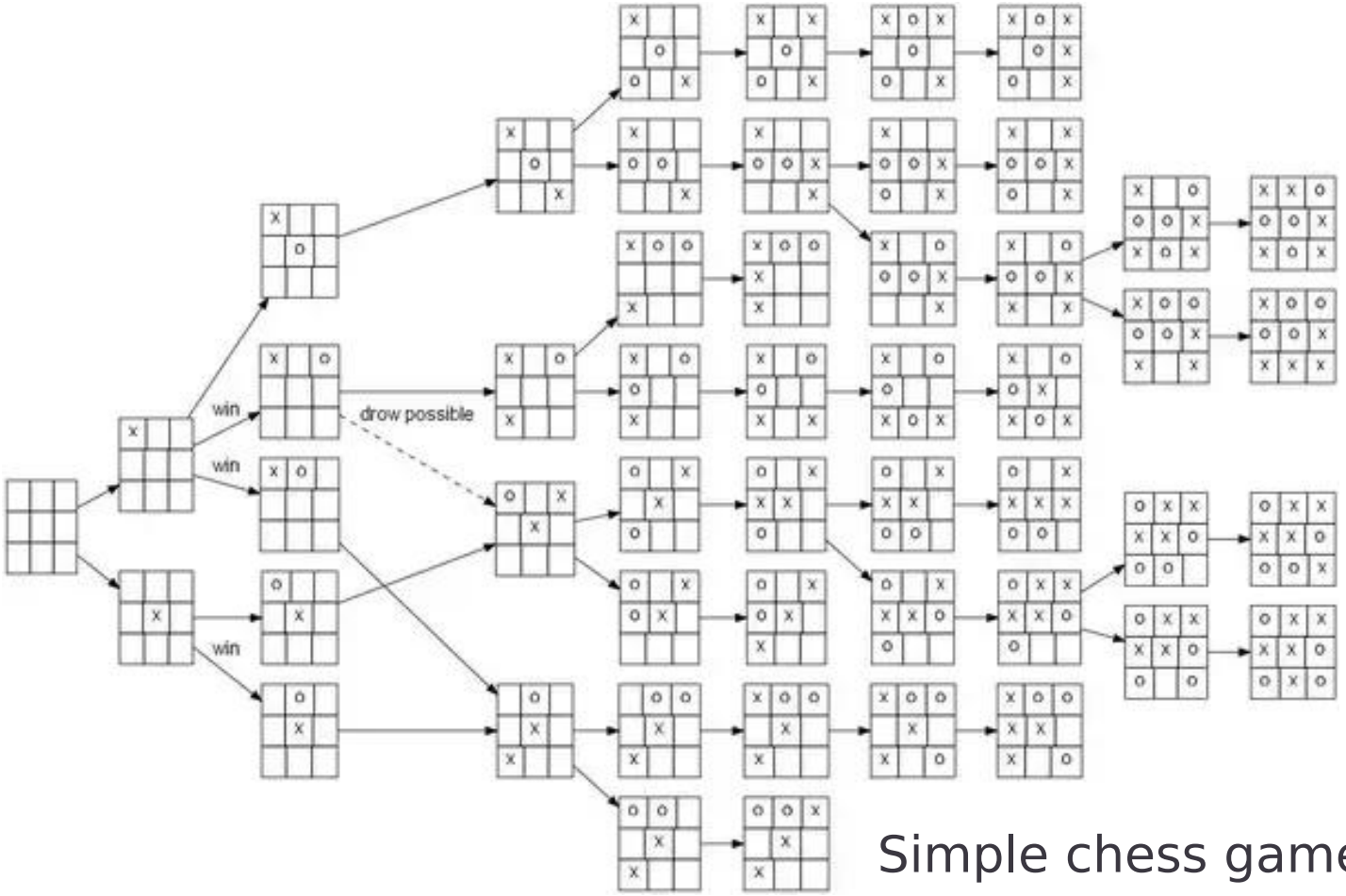


Challenges of NLP

THUNLP

Challenges of General AI

- AI as a Search Process

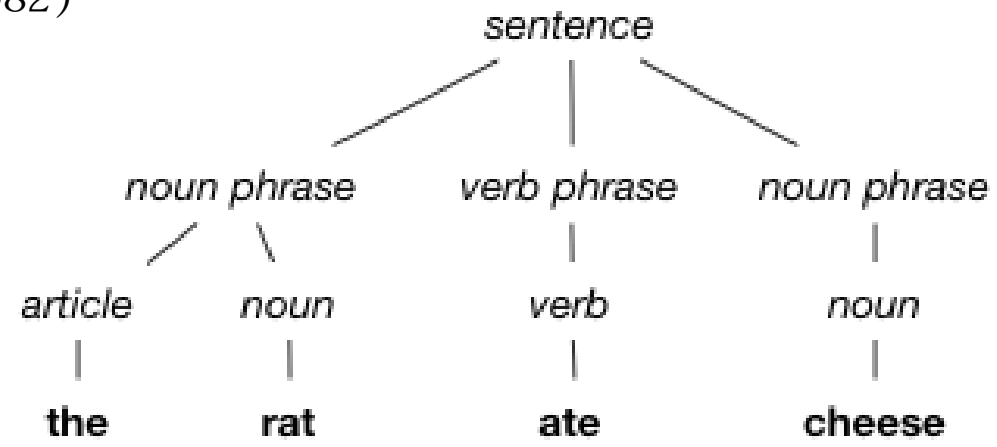


Simple chess game: tic-tac-toe

Challenges of NLP

- The search space of possible syntactic trees of a sentence: exponential growth with sentence length

$$\frac{(2n)!}{(n+1)!(n)!}$$ (Church and Patil, 1982)



Length	Tree Numbers
1	1
2	2
3	2
4	5
5	14
6	42
7	132
8	429
9	1, 430
10	4, 862
11	16, 796
12	58, 786
13	208, 012
14	742, 900
15	2, 674, 440
16	9, 794, 845
17	35, 357, 670
18	129, 644, 790
19	477, 638, 700
20	1, 767, 263, 190

Challenges of NLP

- Find optimal structure regularized with prior syntactic and semantic knowledge

- The regularized search in NLP is difficult, with too many ill-posed issues
  - Variety
  - Recursion
  - Ambiguity
  - ...

Challenges of NLP: Variety

- New words

Filmcholy
noun
The sudden feeling during a movie you're enjoying when you realize it's almost over.

Ploopsie
noun
The phenomenon of a previously broken device suddenly working when you ask someone to help fix it.

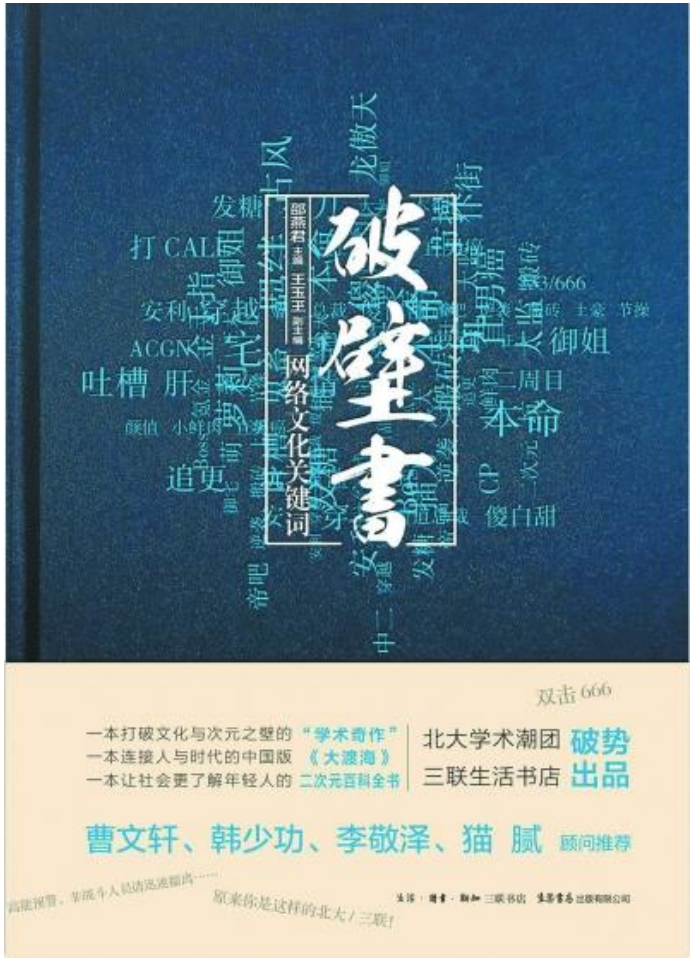
Kurbublin
verb
Typing on a chat, then stopping/deleting, typing again, and repeating this for several minutes causing the recipient to lose their minds.

Swasome
noun
An event where your boss walking by while you have actual work on your computer screen.

Fooldumdums
noun
People who pay for high end HDMI cables

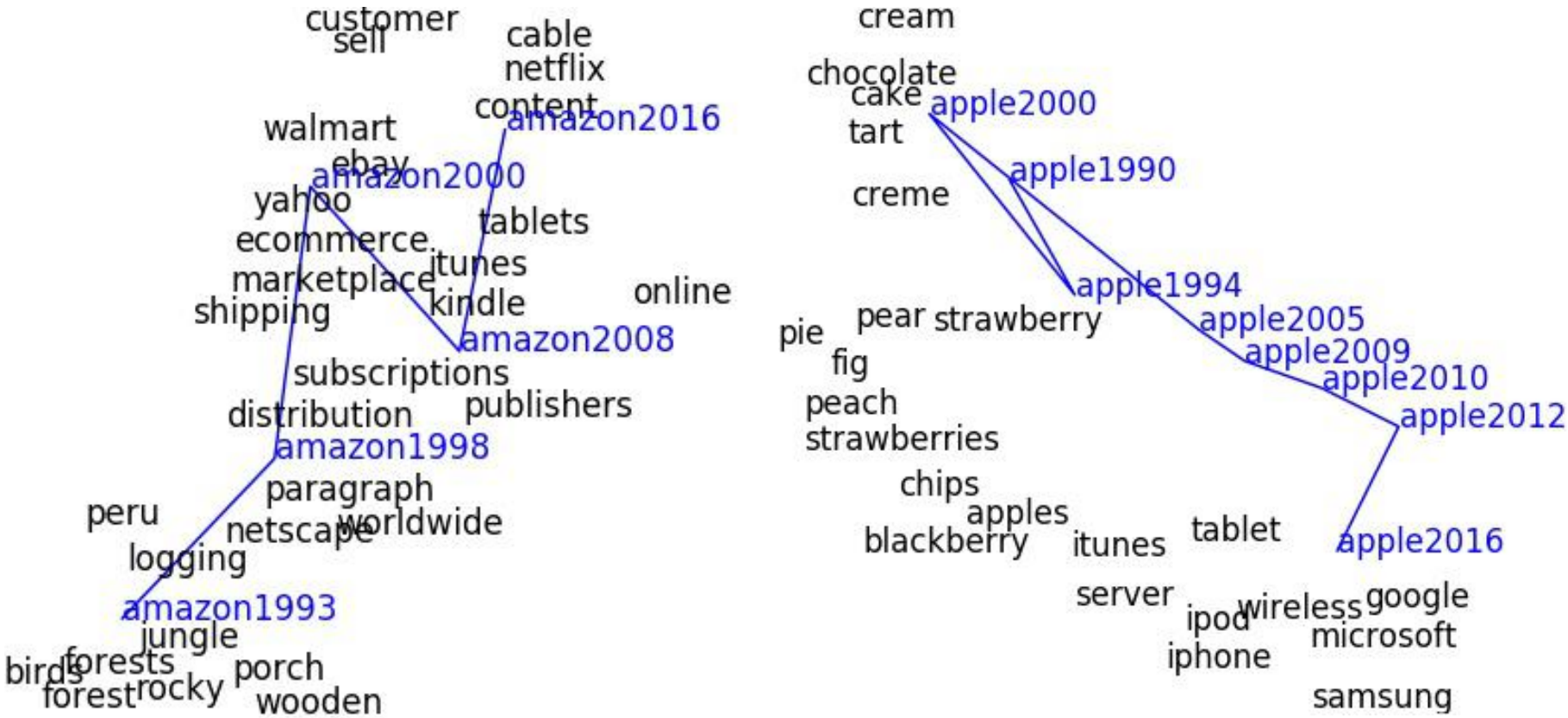
Musicre
verb
Playing a song on repeat until it ceases to produce any form of enjoyment.

DOGHOUSEDIARIES



Challenges of NLP: Variety

- New meanings of existing words



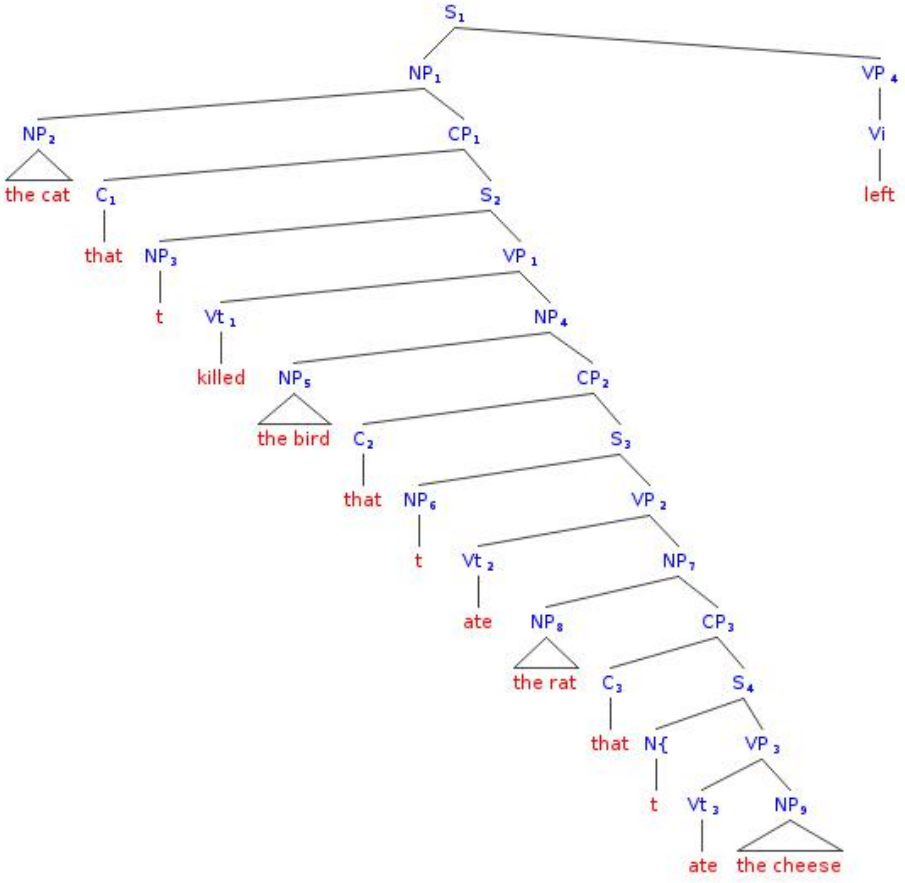
Dynamic Word Embeddings for Evolving Semantic Discovery. ArXiv 2018.

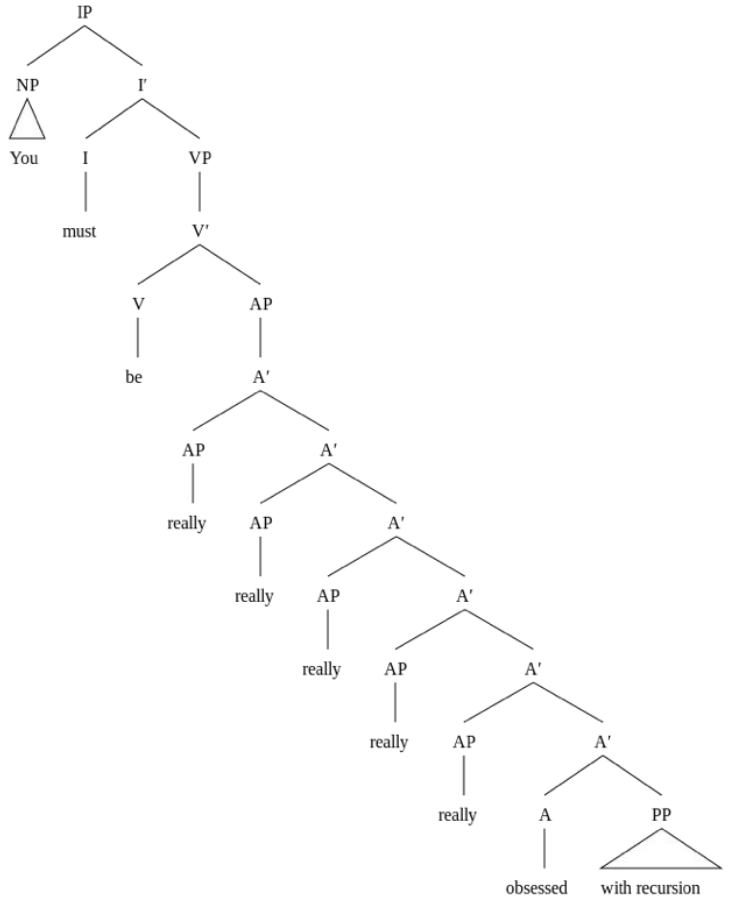
Challenges of NLP: Recursion

- Recursion is everywhere in our world
- Including Natural language

Challenges of NLP: Recursion

- Natural language uses recursive structure to precisely express information





Challenges of NLP: Recursion

Noam Chomsky

The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?

We hypothesize that FLN only includes recursion and is the only uniquely human component of the faculty of language.

Challenges of NLP: Ambiguity

- Ambiguity is a ubiquitous phenomenon from words, sentences to discourse
- People (listener/receiver) disambiguate with context and knowledge

bank

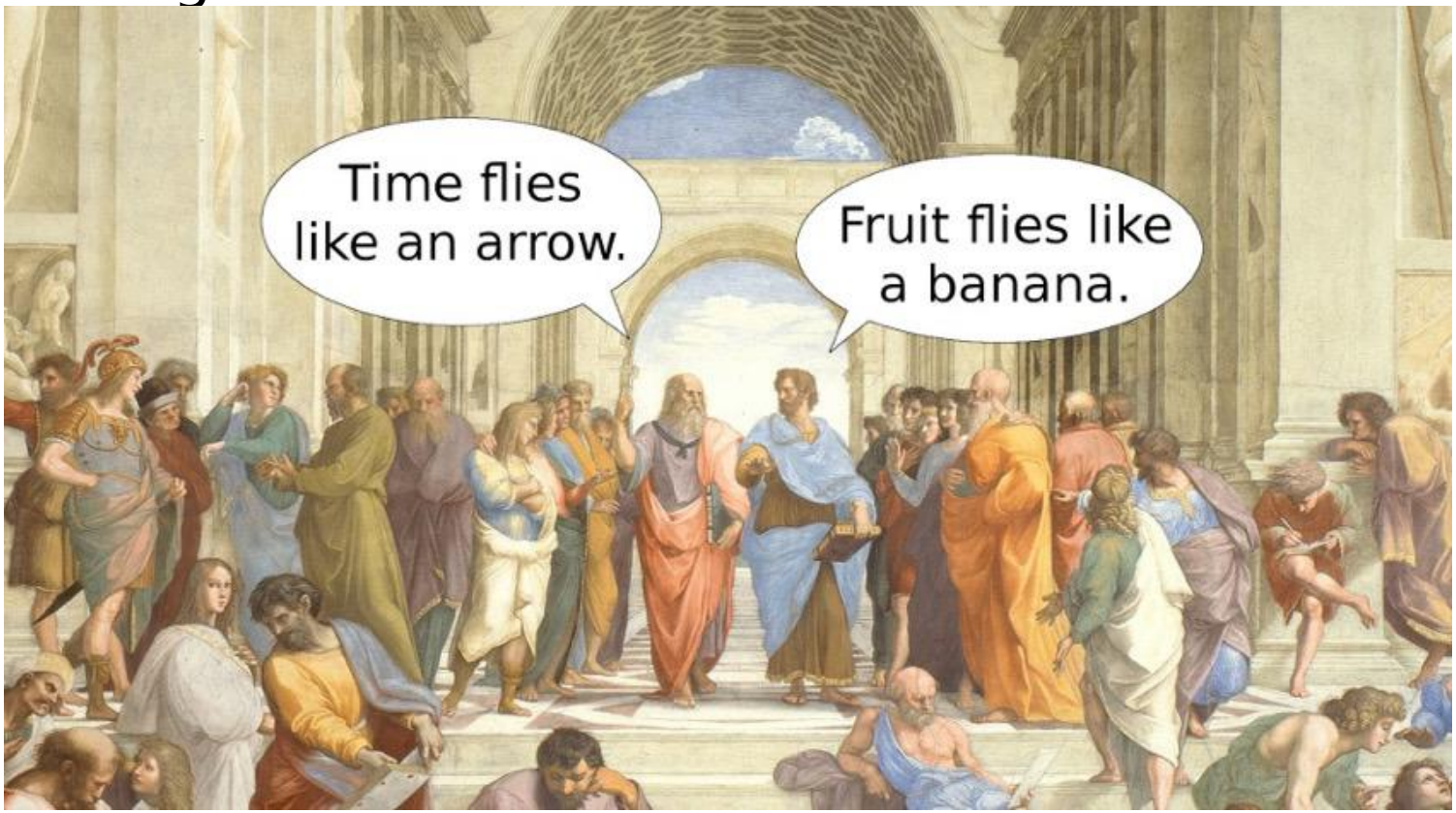
Challenges of NLP: Ambiguity

- Lexical-level ambiguity
- Be disambiguated with sentence context and knowledge

He sat on the bank of the lake

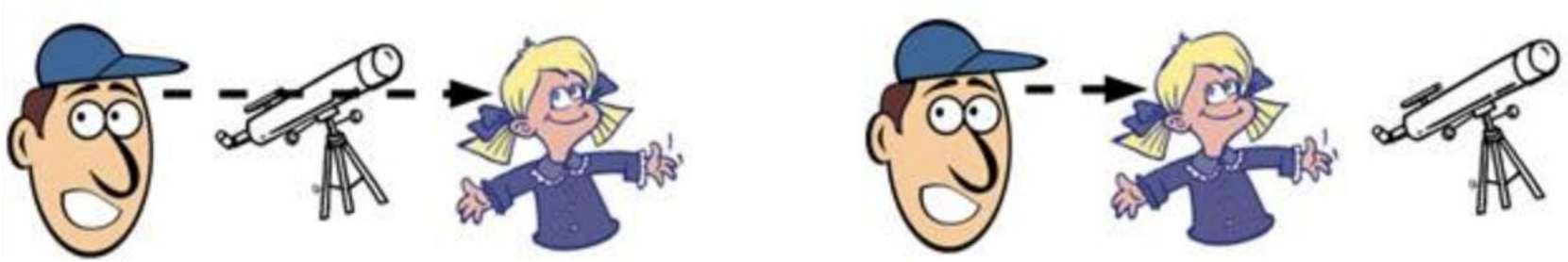
Challenges of NLP: Ambiguity

- Lexical-level ambiguity
- Be disambiguated with sentence context and knowledge



Challenges of NLP: Ambiguity

- Sentence-level ambiguity
- Be disambiguated with external knowledge or information outside the sentence
  - Commonsense knowledge, world knowledge
  - Cross-modal information (vision, speech)



I saw a girl with a telescope.

Challenges of NLP: Ambiguity

- Ambiguity sometimes corresponds to cultures of different areas / countries / languages

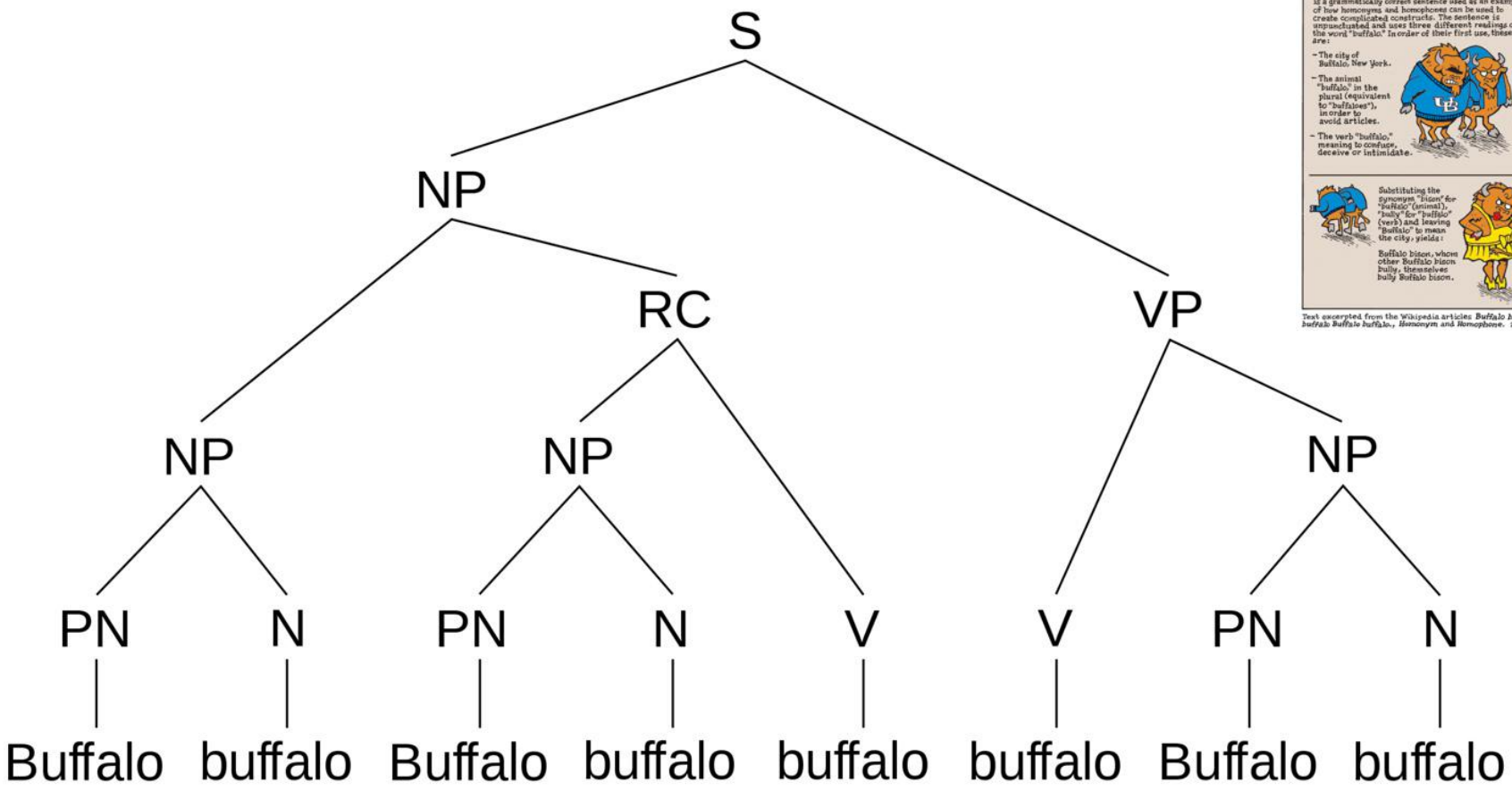
CCTV

Closed-circuit television
监控、闭路电视

CCTV

中国中央电视台

Challenges of NLP



Simplified parse tree S = sentence NP = noun phrase RC = relative clause

Challenges of NLP

1. 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，氏始试食十狮尸。食时，始识是十狮，实十石狮尸。试释是事。

<< Shī Shì shí shī shǐ >>
Shíshì shīshì Shī Shì, shì shī, shì shí 1
Shì shíshí shì shì shì shī.
10 shí, shì 10 shī shì shì.
Shì shí, shì Shī Shì shì shì.
Shì shì shì 10 shī, shì shǐ shì, shī shì 1
Shì shí shì 10 shī shī, shì shíshì.
Shíshì shī, Shì shǐ shì shì shíshì.
Shíshì shì, Shì shí shì shí shì 10 shī.
Shí shí, shǐ shì shì 10 shī, shí 10 shī shī.
Shì shì shì shì.

2. 骑车差点摔倒， 好在我一把把把把住了。

3.小龙女动情地说：“我也想过过过儿过过的生活”。

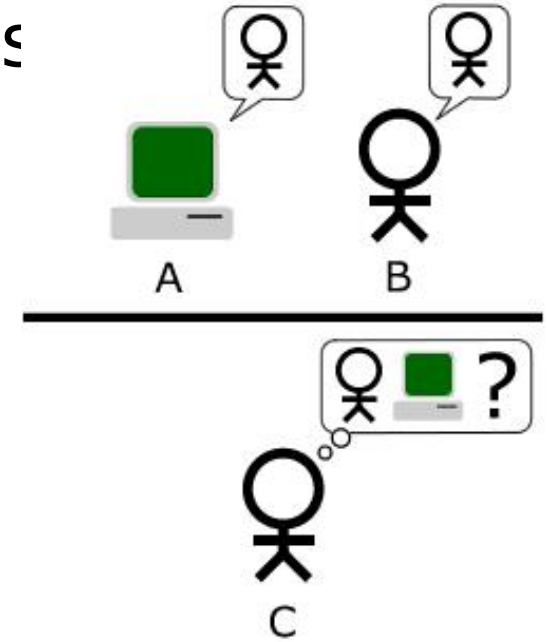
4.校长说衣服上除了校徽别别别的。

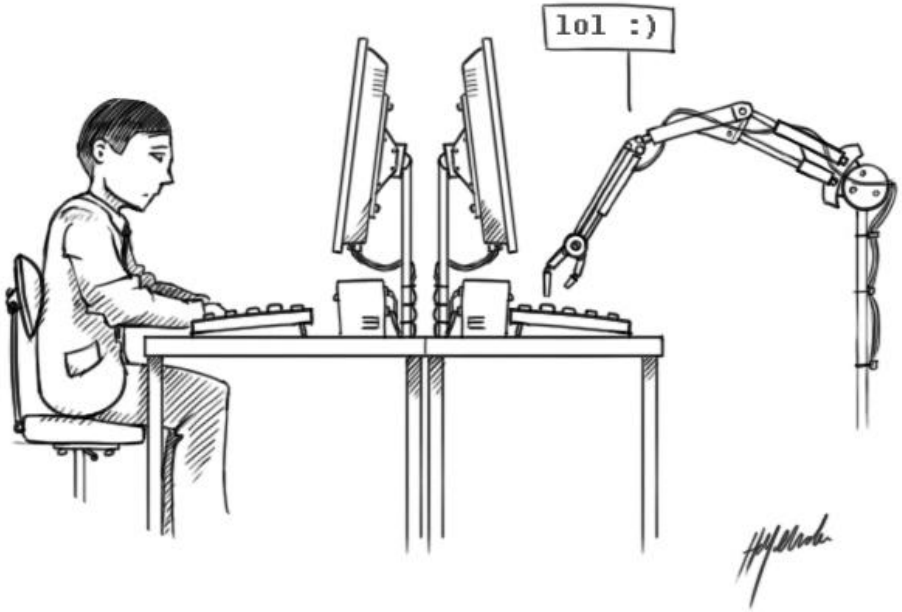
Why is NLP Important?

THUNLP

Scientific Impact of NLP

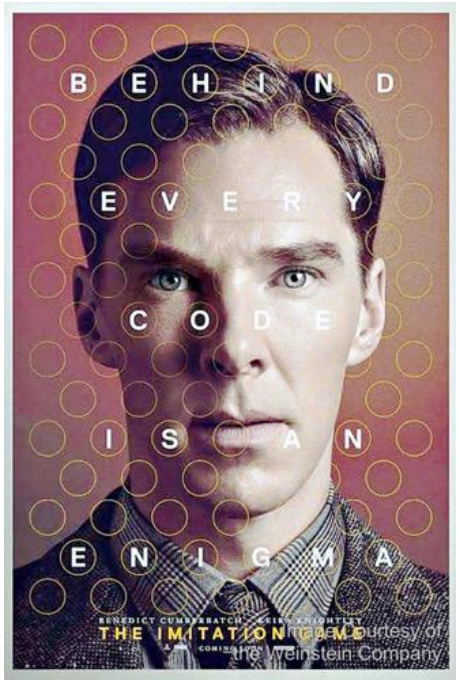
- Turing Test: A test of machine ability to exhibit intelligent behavior indistinguishable from a human
- Language is the communication tool in the tes

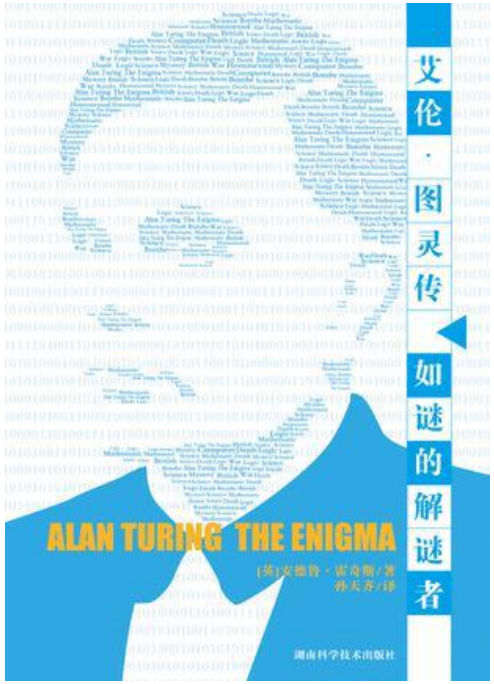


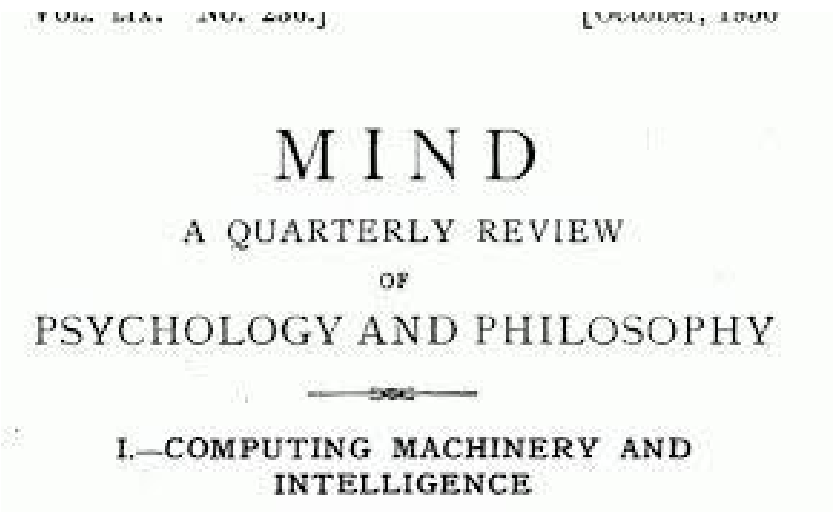


Scientific Impact of NLP

- Origin Version: Imitation Game







MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

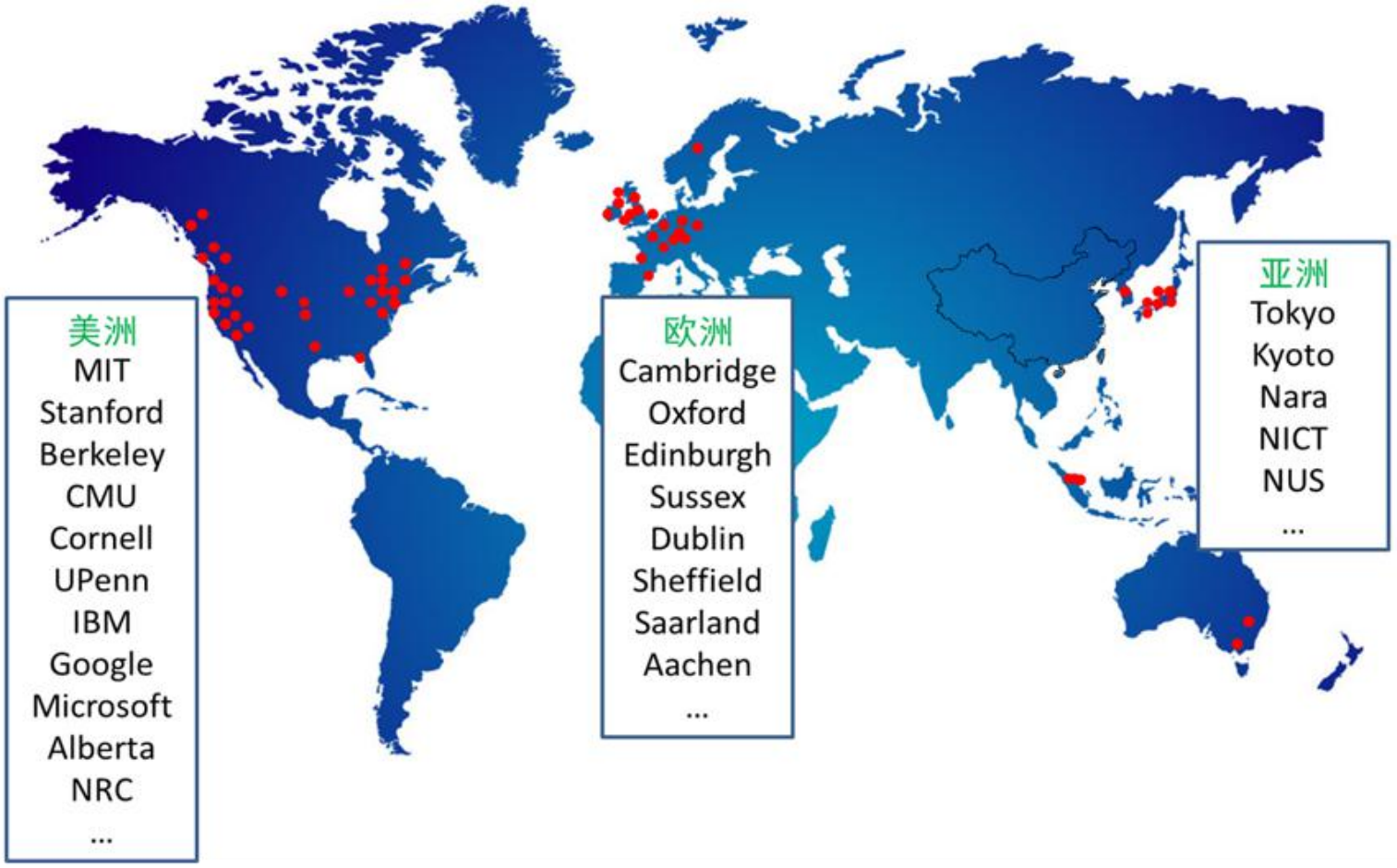
Scientific Impact of NLP

- Natural language question-answering
- 2011: IBM Watson DeepQA system competed on *Jeopardy!* and received the first place
- A new milestone of AI after DeepBlue won world champion of chess in 1997

Q: Who was presidentially pardoned on September 8, 1974?
A: Nixon.

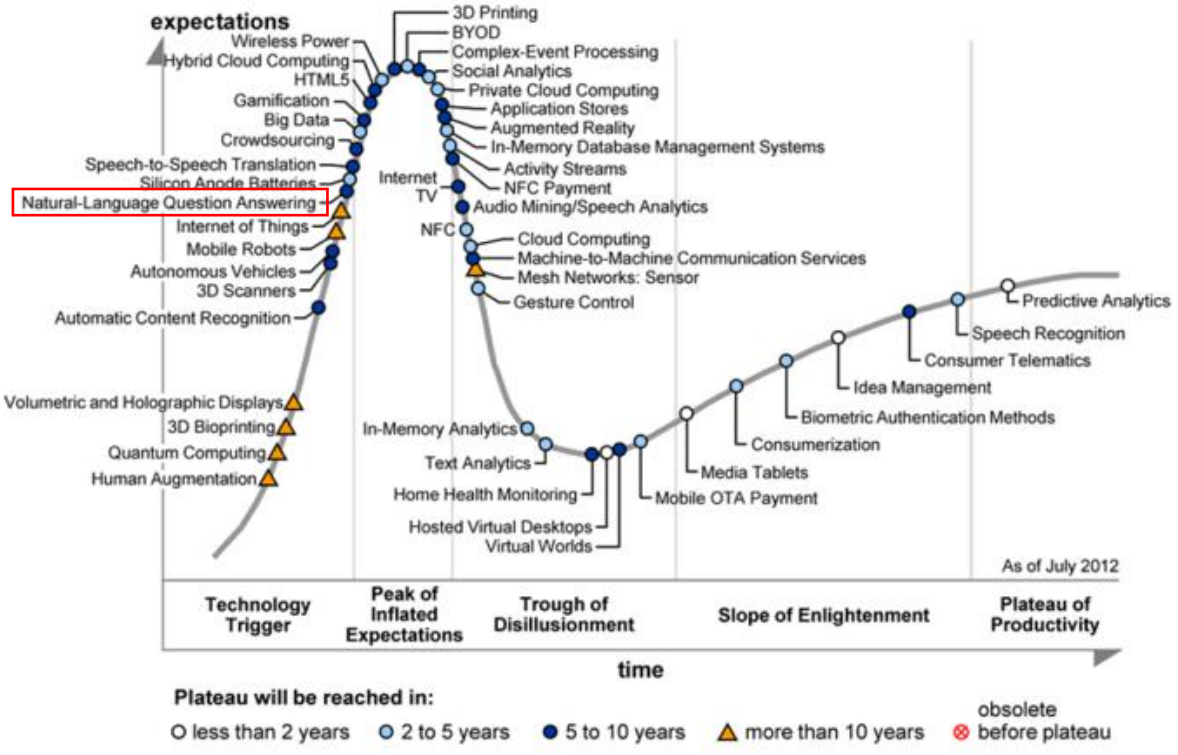
Scientific Impact of NLP

- Many institutes take NLP as key research areas



Application Impact of NLP

- Nature 2011: Natural Language QA will be next-generation search engine
- Gartner Hype Cycle 2012



Application Impact of NLP

- IT giants launch their NLP products

Apple
Siri

Speech
Translator

Sogou Input

Google Knowledge Graphs

Typical Tasks & Applications of NLP

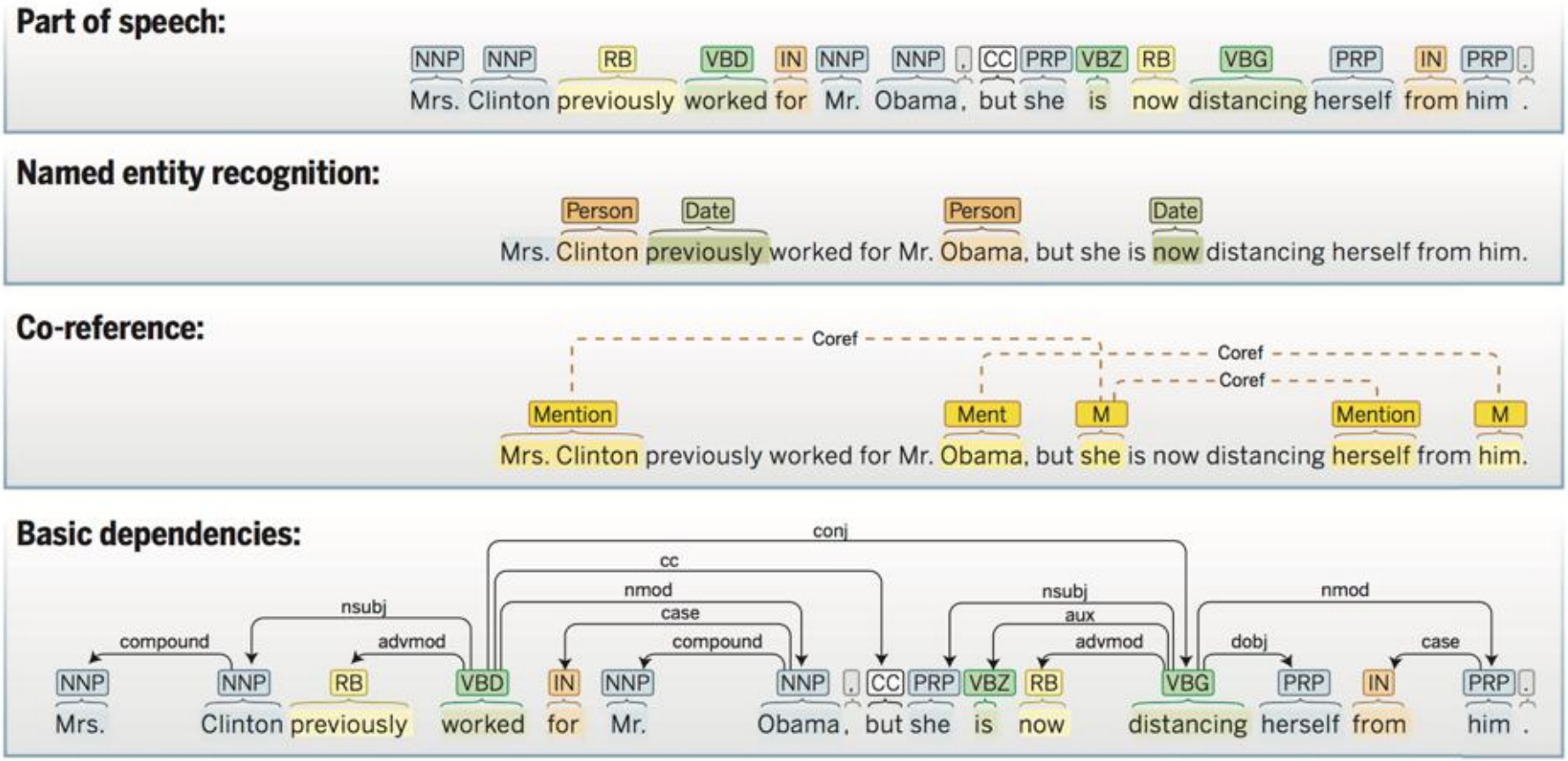
THUNLP

A Nice Review on NLP

- Advances in Natural Language Processing
- Julia Hirschberg, Columbia University
  - AAAI, ACL Fellow
- Christopher Manning，Stanford University
  - ACM, AAAI, ACL Fellow
  - Google Scholar Citation > 50,000

NLP Tasks

- Basic Tasks of NLP



Search Engines and Ads

Structural Knowledge

write

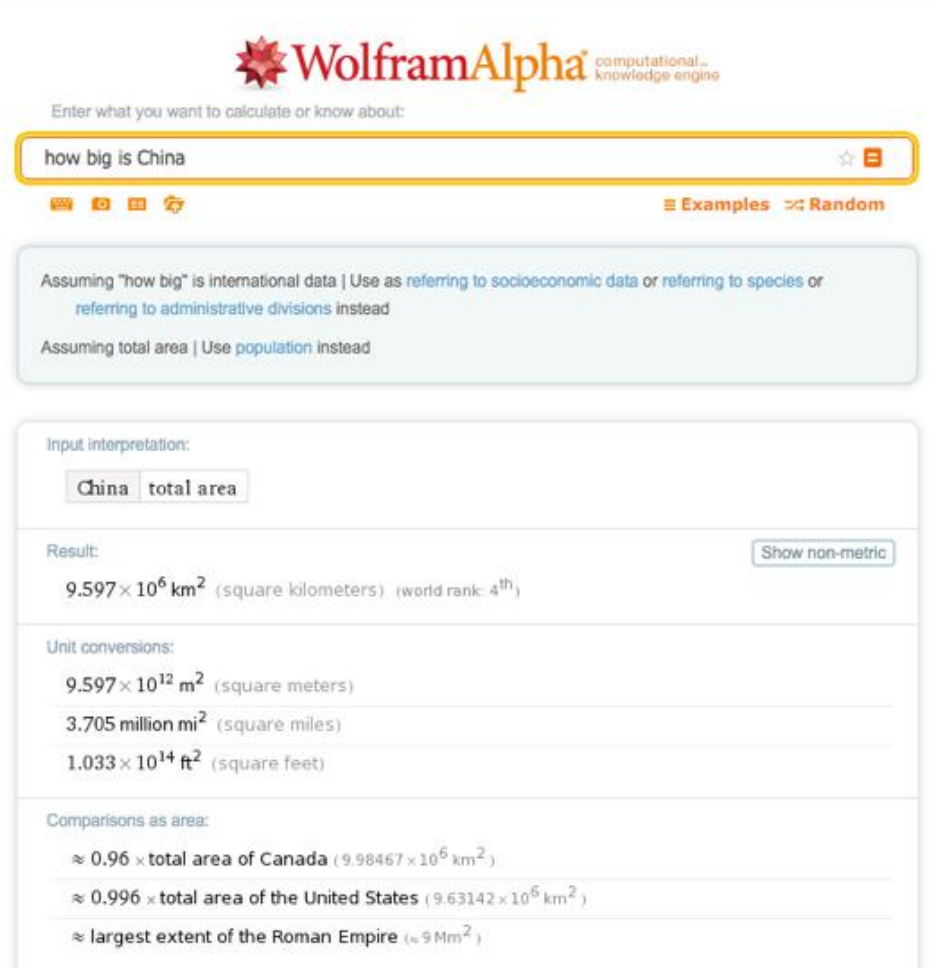
(***William Shakespeare***, book/author/works_written, ***Romeo and Juliet***)

head entity **relation** **tail entity**

Knowledge Graph

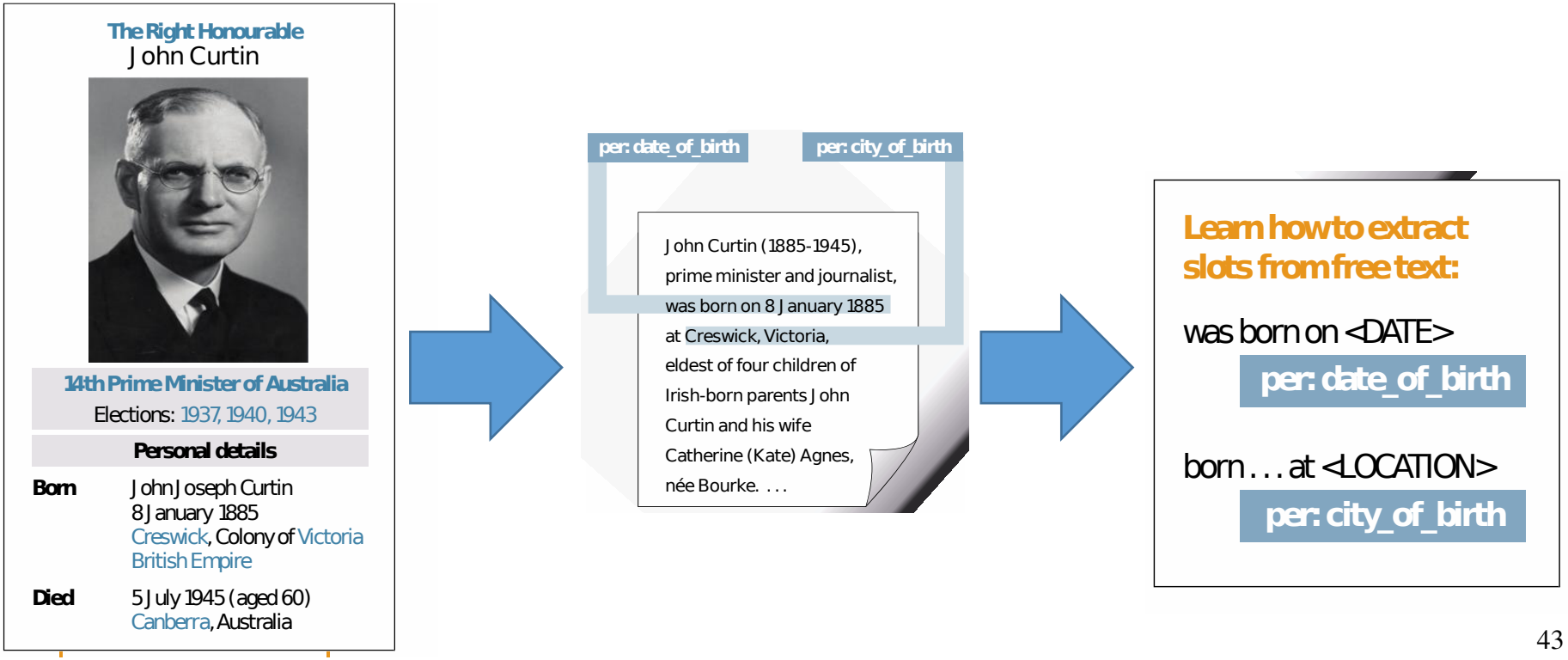
KG Application: Question Answering

- KG provides the knowledge facts of a question



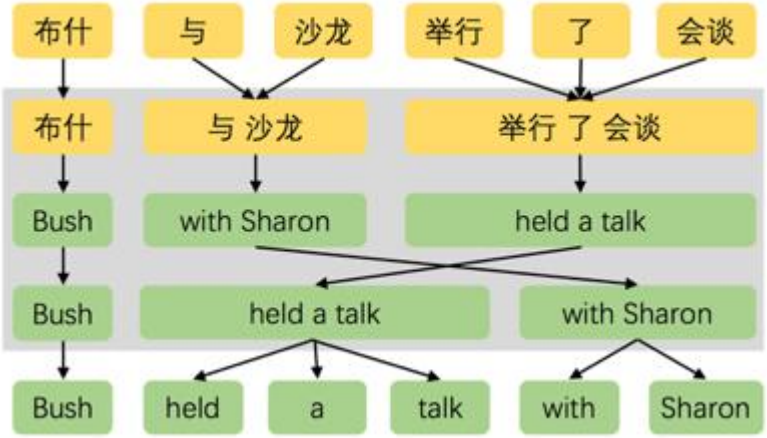
Machine Reading

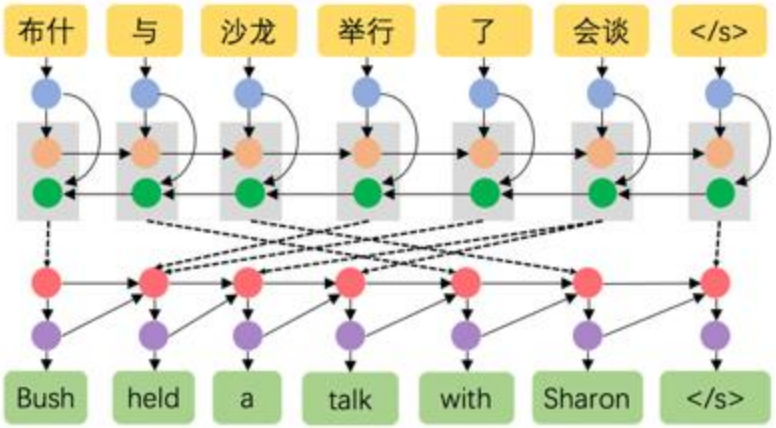
- Machine reading aims to extract structural knowledge (e.g., relational facts between entities) from plain text
- Can expand and update knowledge graphs



Personal Assistant

Machine Translation





Rule-based Phrase-based Neural-based

1960 1990s 1990s 2015

Statistics-based

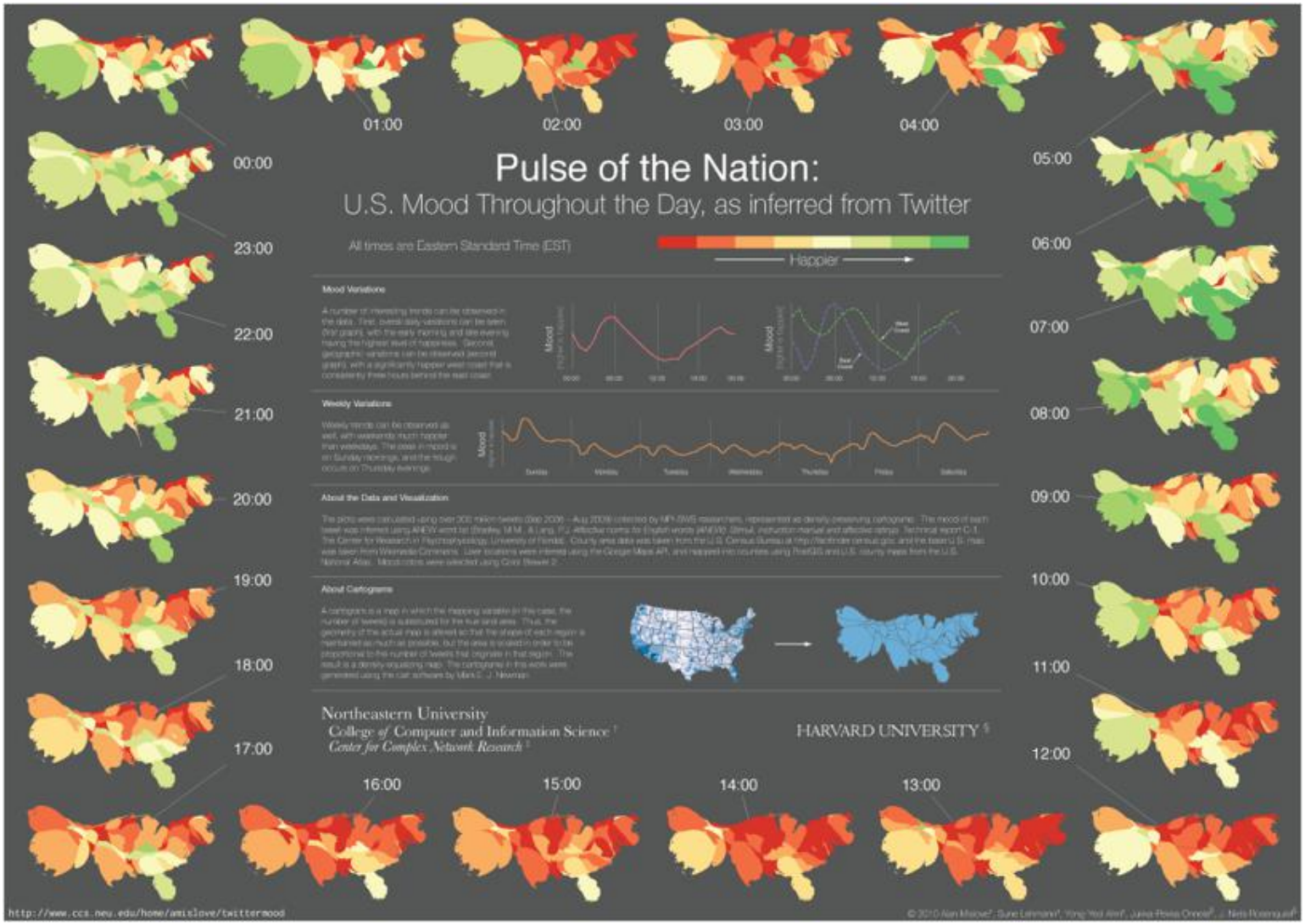
Machine Translation

Sentiment Analysis and Opinion Mining

- Infer personal states via text or speech
  - Including opinions, emotions, …

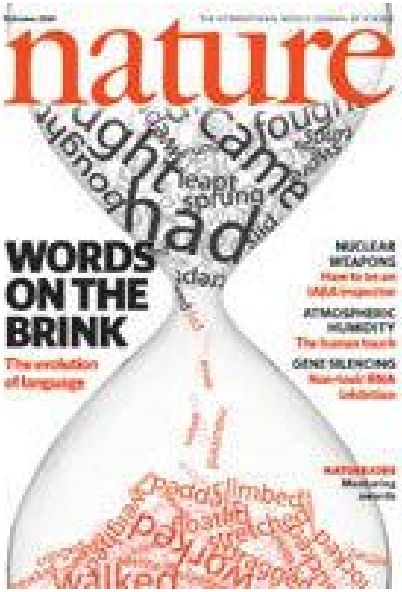
- Detect opinion holders and targets

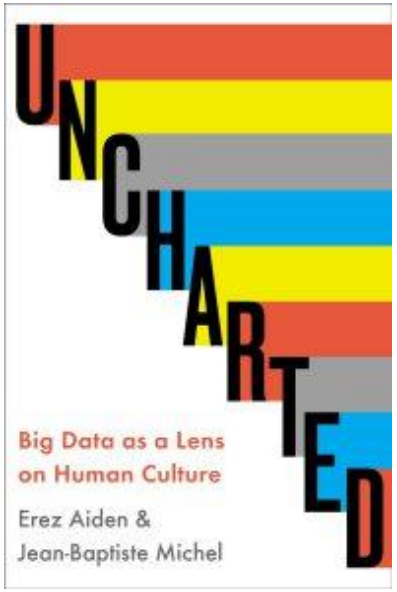
Sentiment Analysis and Opinion Mining



Computational Social Science

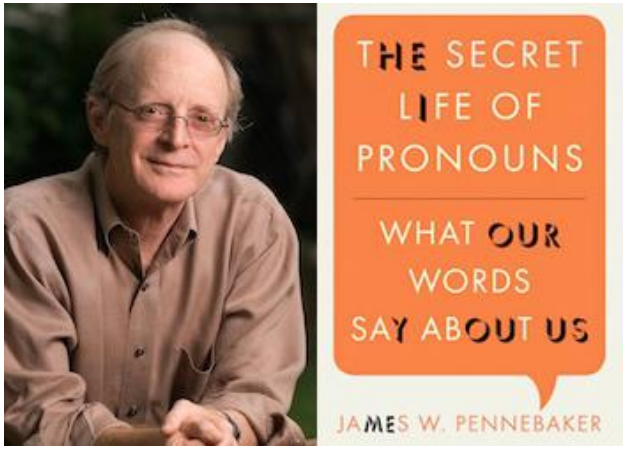
- Culturomics: www.culturomics.org
- Harvard researchers use keywords over Google Books (5 million books from 1800 to 2000) to study the evolution of human culture
- Google Book N-gram: books.google.com/ngrams

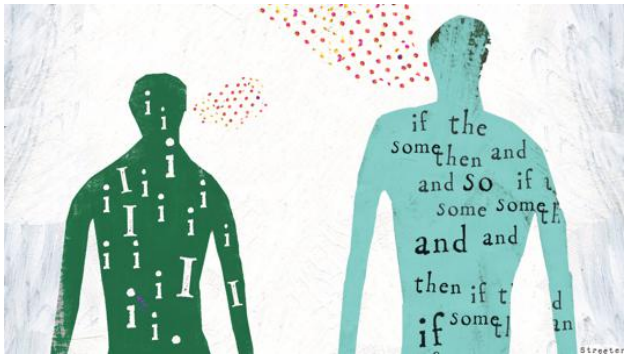


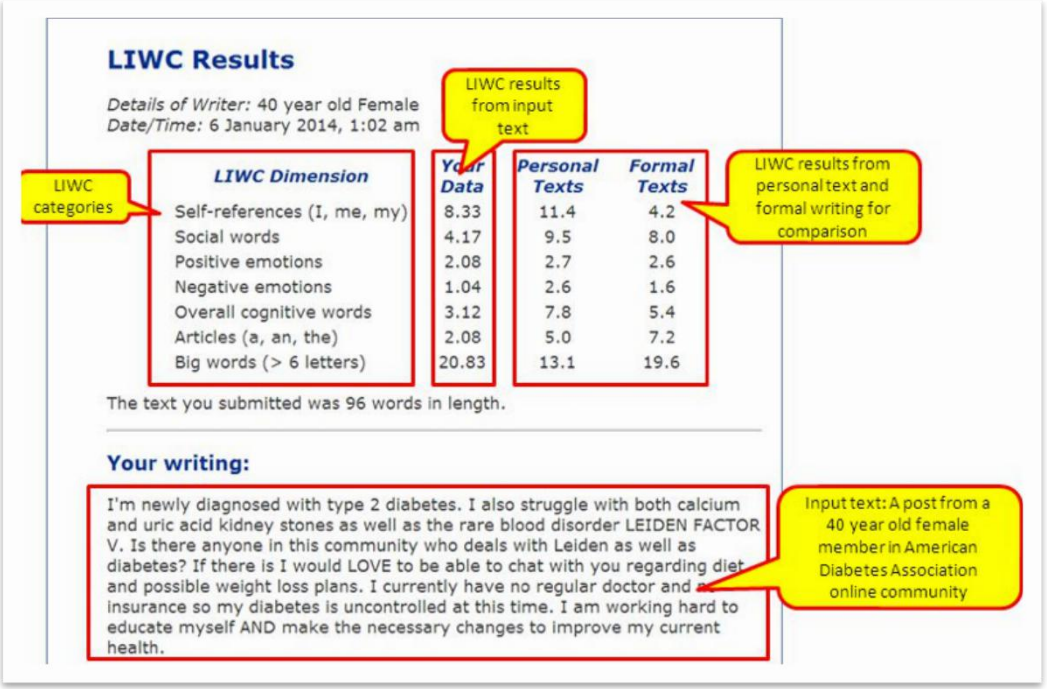


Computational Social Science

- Use language usage to study psychology states of humans



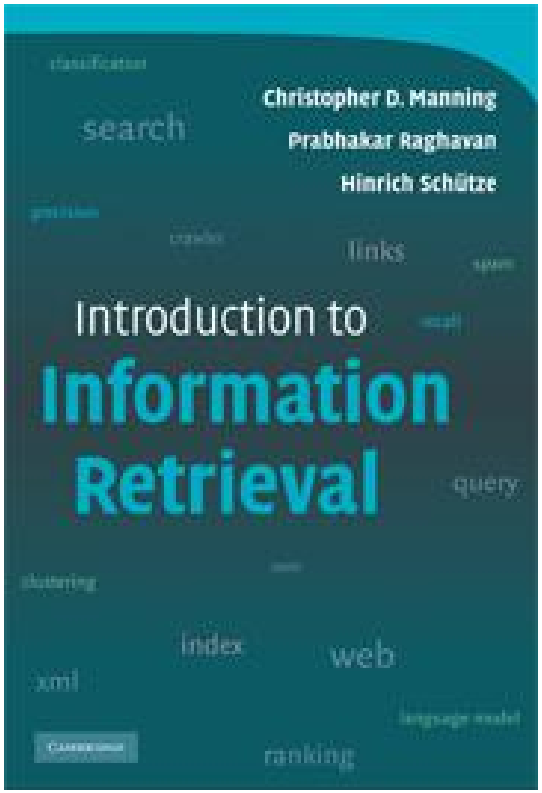


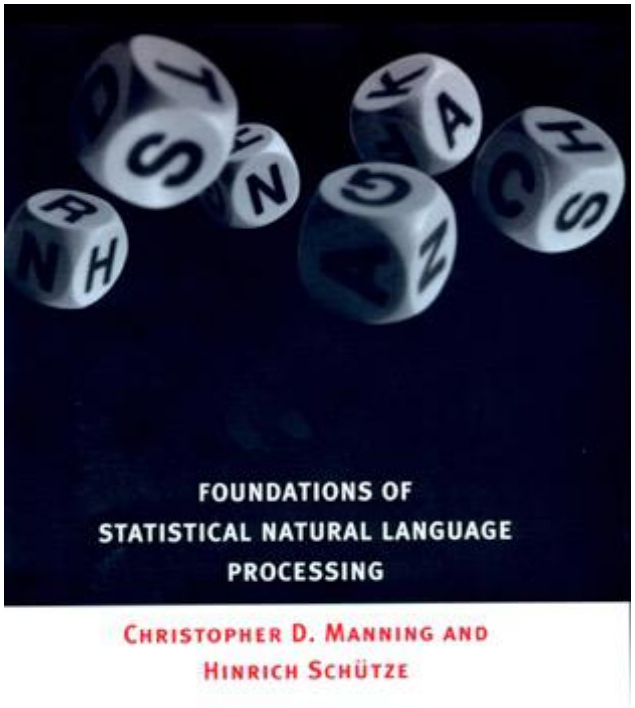


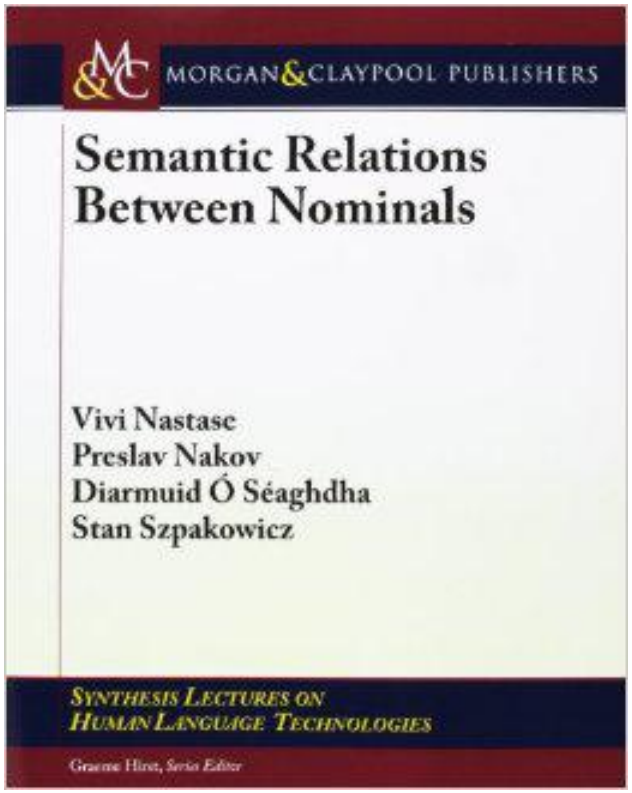
Recommended Readings

THUNLP

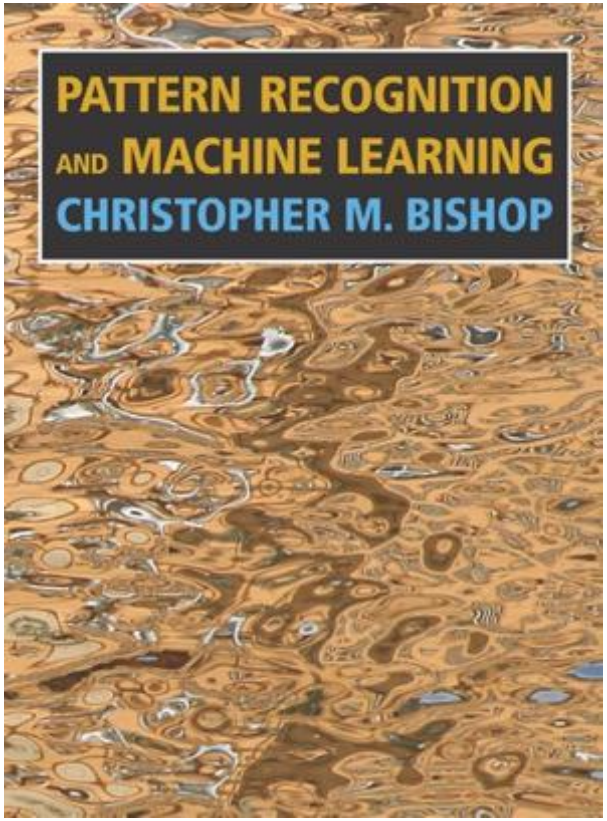
Natural Language Processing

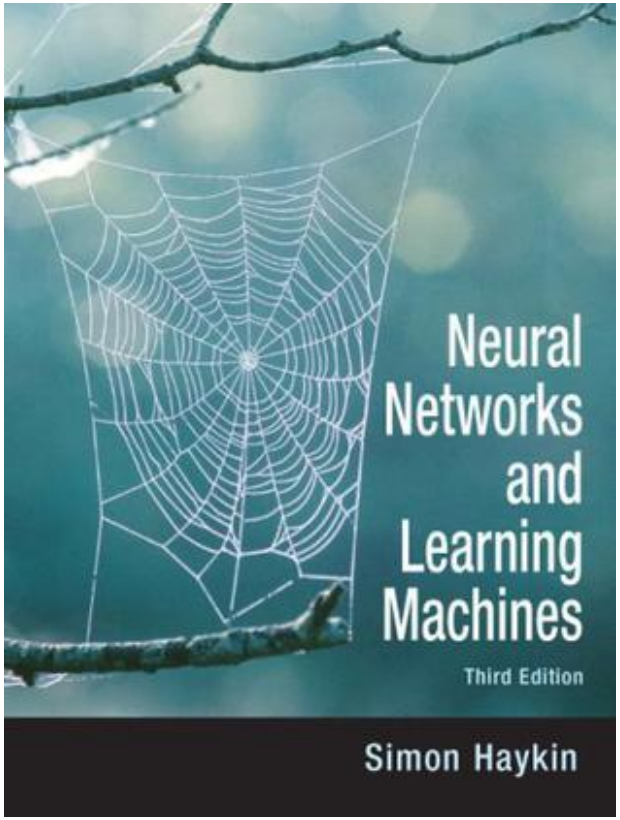


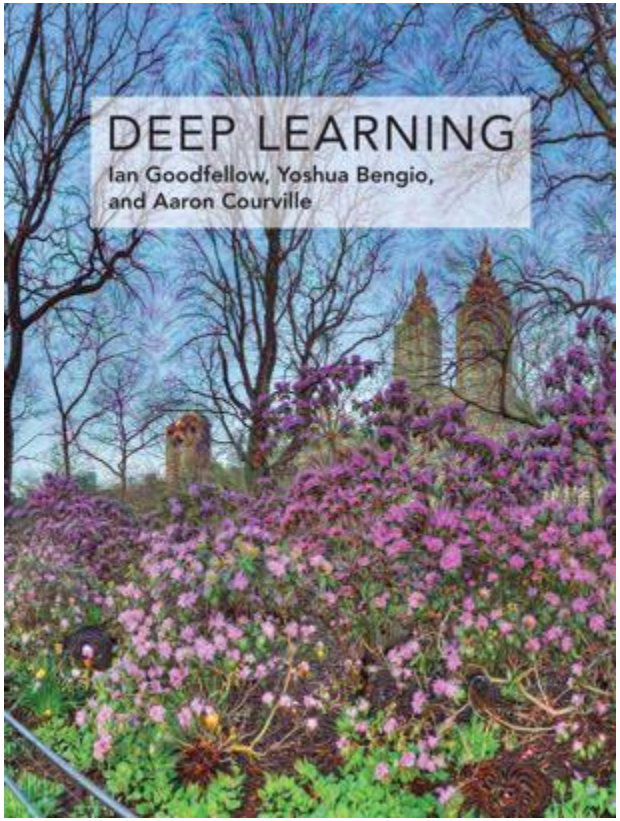




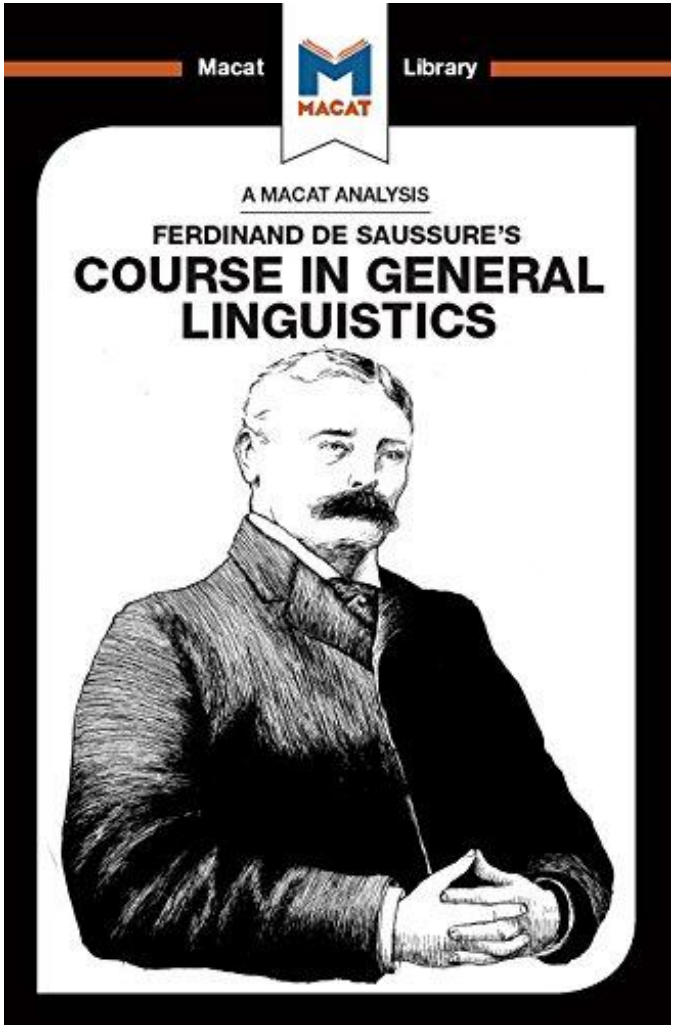
Machine Learning

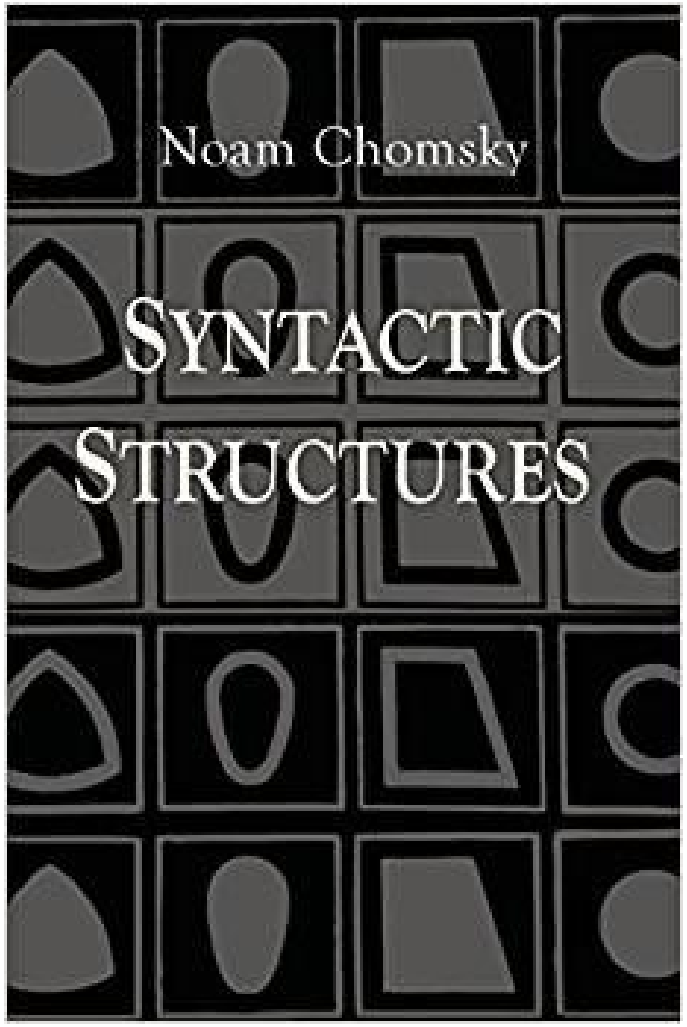






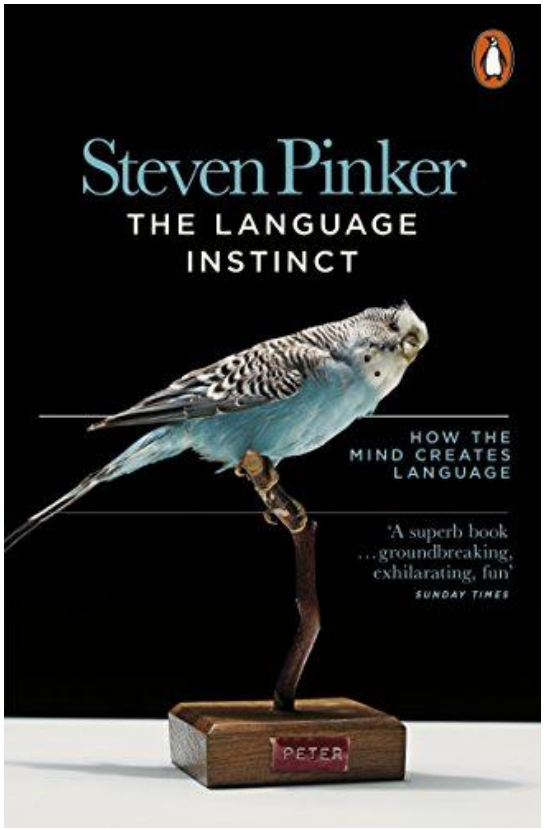
Linguistics

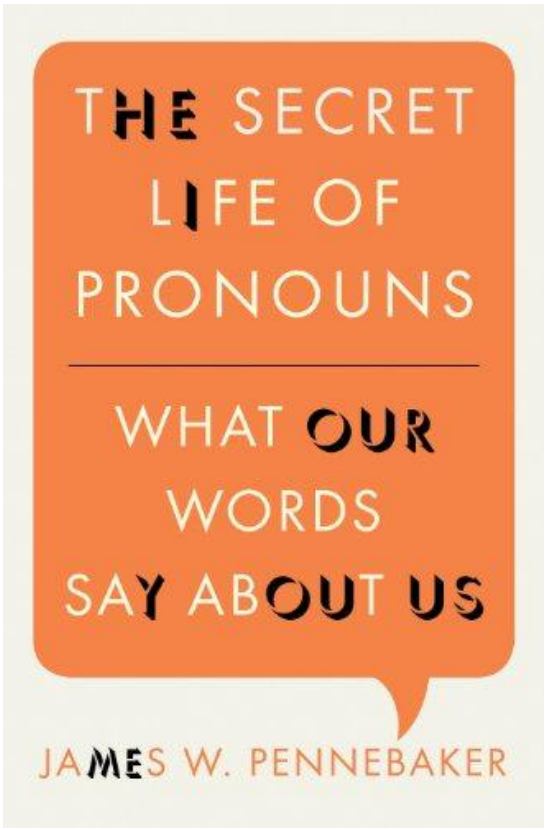


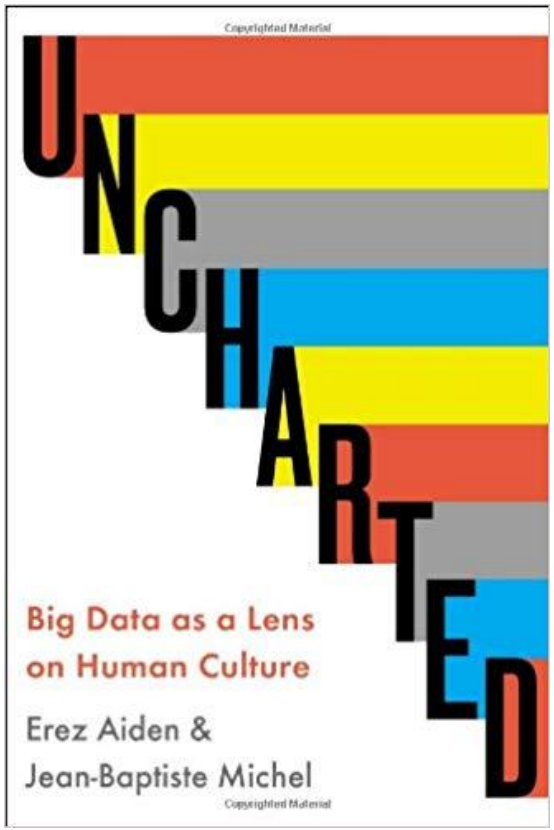


Misc

- Other popular books about languages from the perspectives of Cognition, Social Psychology, and Computational Social Sciences







Academic Websites

- Google Scholar: http://scholar.google.com/
- ACM Portal: http://dl.acm.org/
- Association of Computational Lingusitics
  - ACL Anthology: http://www.aclweb.org/anthology/
  - ACL、EMNLP、NAACL、COLING

Thanks! Demo Time.

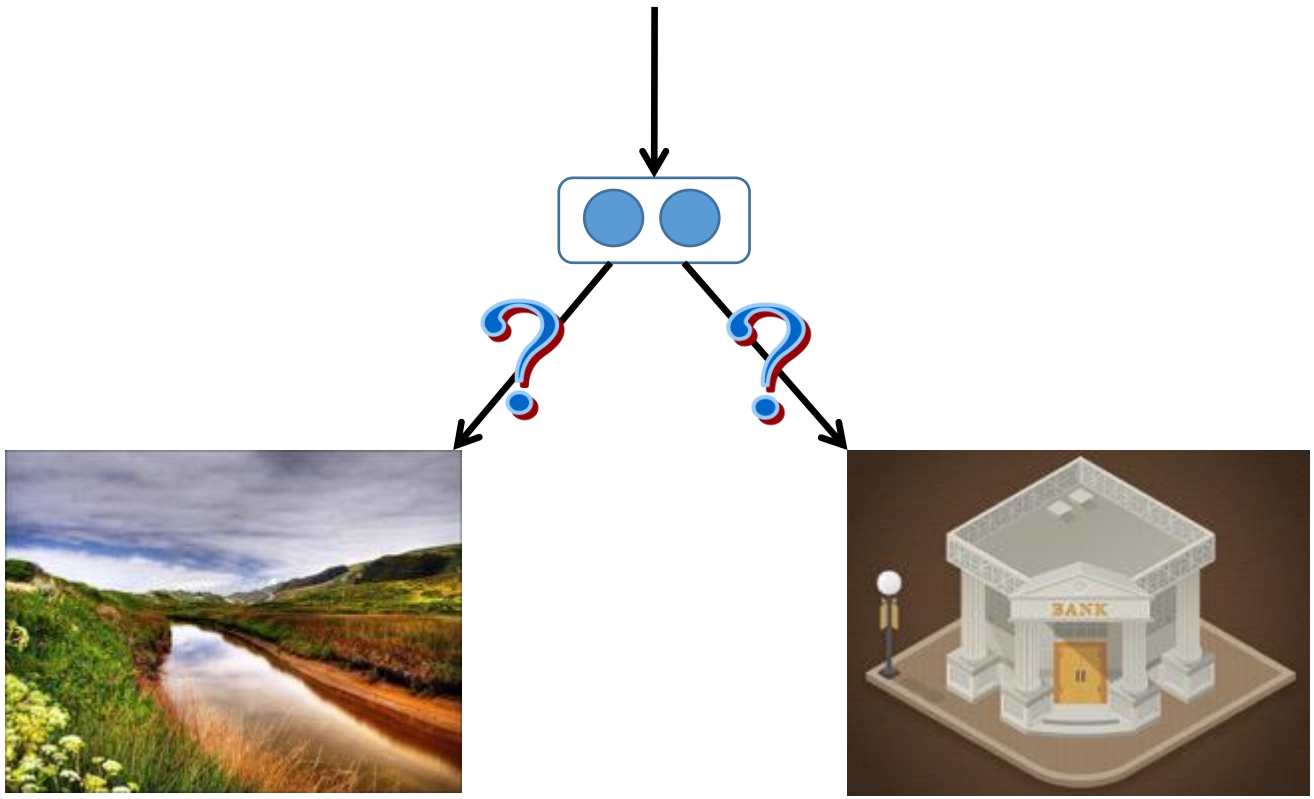
THUNLP

Appendix Applications of Word Embedding

THUNLP

Word Sense Representation

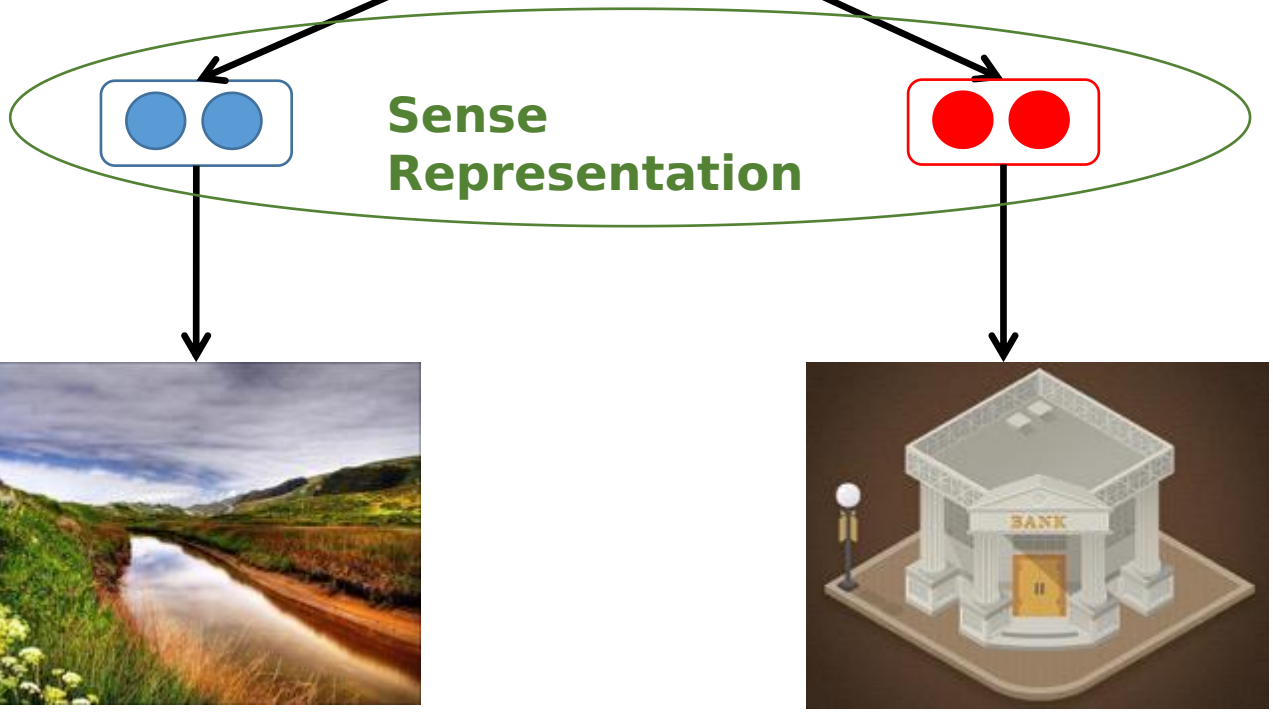
He sat on the bank of the lake



Single vector is not enough

Word Sense Representation

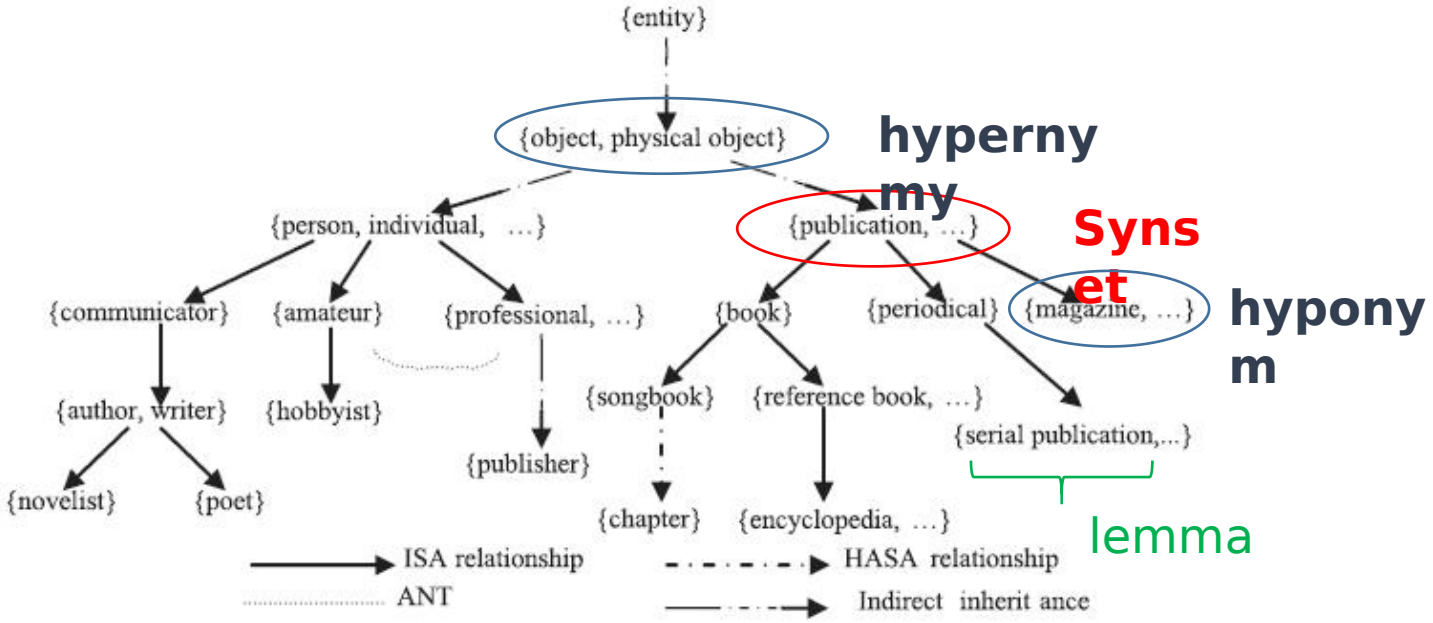
He sat on the bank of the lake



Represent senses with different representations

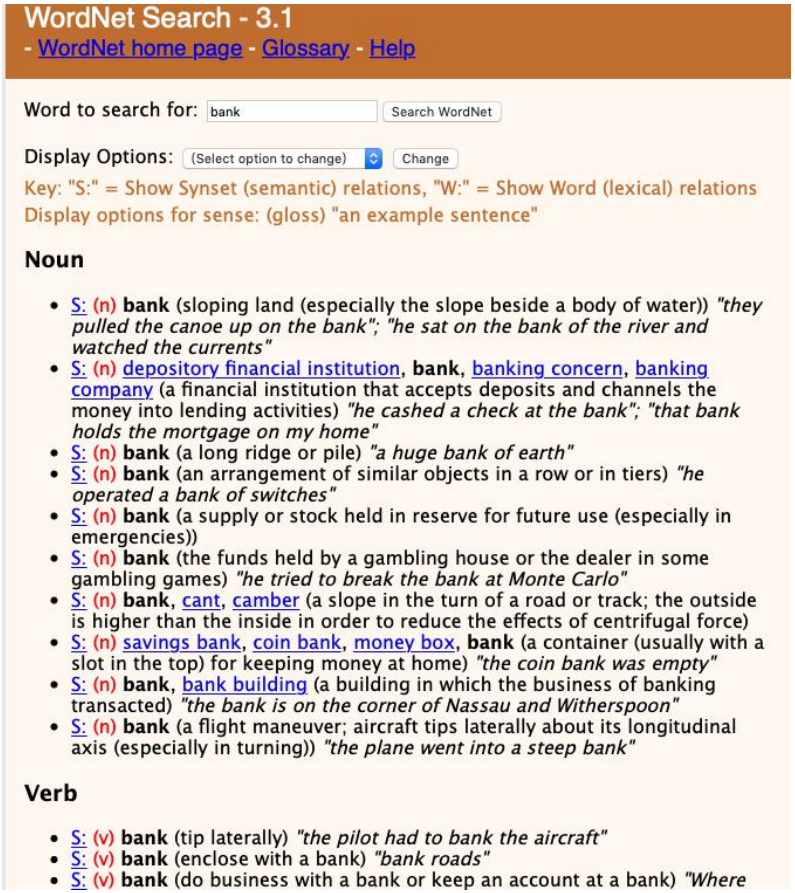
Word Knowledge

- WordNet
  - Machine readable semantic dictionary interlinked by semantic relations
  - A large lexical database for English by Princeton
  - Free for public



Word Knowledge

- WordNet



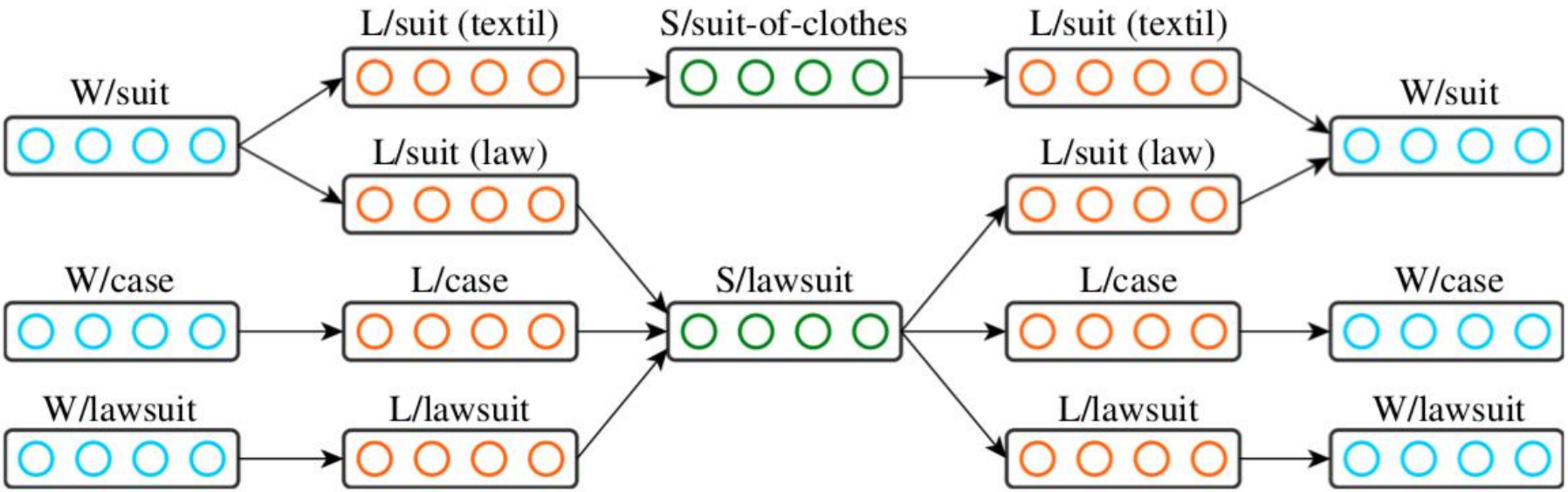

```
>>> from nltk.corpus import wordnet as wn
>>> wn.synsets('bank')
[Synset('bank.n.01'), Synset('depository_financial_institution.
n.01'), Synset('bank.n.03'), Synset('bank.n.04'), Synset('bank.
n.05'), Synset('bank.n.06'), Synset('bank.n.07'), Synset('savin
gs_bank.n.02'), Synset('bank.n.09'), Synset('bank.n.10'), Synse
t('bank.v.01'), Synset('bank.v.02'), Synset('bank.v.03'), Synse
t('bank.v.04'), Synset('bank.v.05'), Synset('deposit.v.02'), Sy
nset('bank.v.07'), Synset('trust.v.01')]
>>> wn.synsets('bank')[0].examples()
['they pulled the canoe up on the bank', 'he sat on the bank of
 the river and watched the currents']
>>> wn.synsets('bank')[0].definition()
'sloping land (especially the slope beside a body of water)'
>>> wn.synsets('bank')[0].hypernyms()
[Synset('slope.n.01')]
>>> wn.synsets('bank')[0].lemmas()
[Lemma('bank.n.01.bank')]
>>>
```

http://wordnetweb.princeton.edu/perl/webwn

http://www.nltk.org/howto/wordnet.html

Sense Representation with WordNet

- Improve sense representations with WordNet architecture
  - Words are sums of their lexemes
  - Synsets are sums of their lexemes



Sascha Rothe and Hinrich Schu¨tze. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. IJCNLP 2015.

Sense Representation with WordNet

- Evaluation
  - 5 nearest Word (W/), Lexeme (L/) or Synset (S/) of W/suit, W/lawsuit and S/suit-of-cloths

nearest neighbors of W/suit
S/suit (businessman), L/suit (businessman), L/accomodate, S/suit (be acceptable), L/suit (be acceptable), L/lawsuit, W/lawsuit, S/suit (playing card), L/suit (playing card), S/suit (petition), S/lawsuit, W/countersuit, W/complaint, W/counterclaim
nearest neighbors of W/lawsuit
L/lawsuit, S/lawsuit, S/countersuit, L/countersuit, W/countersuit, W/suit, W/counterclaim, S/counterclaim (n), L/counterclaim (n), S/counterclaim (v), L/counterclaim (v), W/sue, S/sue (n), L/sue (n)
nearest neighbors of S/suit-of-clothes
L/suit-of-clothes, S/zoot-suit, L/zoot-suit, W/zoot-suit, S/garment, L/garment, S/dress, S/trousers, L/pinstripe, L/shirt, W/tuxedo, W/gabardine, W/tux, W/pinstripe

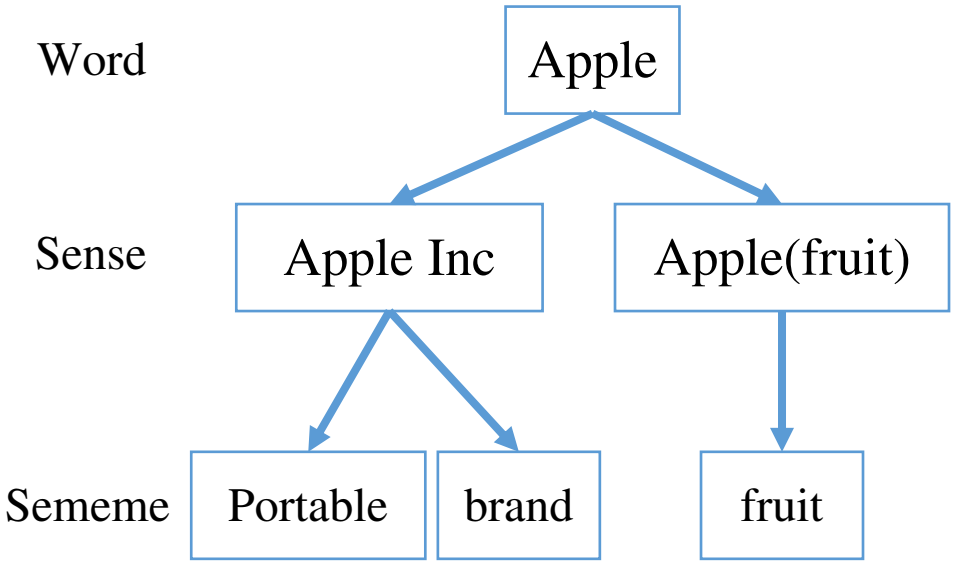
Word Knowledge

- HowNet
  - HowNet adheres to the concept of reductionism and considers the meaning of words can be described in smaller semantic units (Sememes).
  - 银行 (Bank)

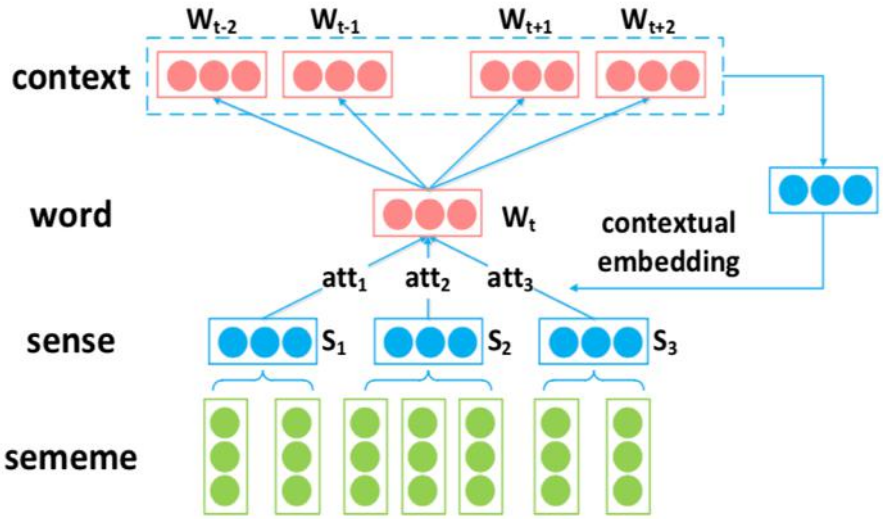
https://openhownet.thunlp.org/home

Sense Representation with HowNet

- Improve sense representations with HowNet



HowNet



Method

Yilin Niu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. Improved Word Representation Learning with Sememes.

Sense Representation with HowNet

- Evaluation
  - Dataset: Google analogy test dataset

Model	Accuracy				Mean Rank			
	Capital	City	Relationship	All	Capital	City	Relationship	All
CBOW	49.8	85.7	**86.0**	64.2	36.98	1.23	62.64	37.62
GloVe	57.3	74.3	81.6	65.8	19.09	1.71	3.58	12.63
Skip-gram	66.8	93.7	76.8	73.4	137.19	1.07	2.95	83.51
SSA	62.3	93.7	81.6	71.9	45.74	1.06	3.33	28.52
MST	65.7	95.4	82.7	74.5	50.29	1.05	2.48	31.05
SAC	79.2	97.7	75.0	81.0	28.88	1.02	2.23	18.09
SAT	**82.6**	**98.9**	80.1	**84.5**	**14.78**	**1.01**	**1.72**	**9.48**

SAC、SAT are models in the paper

Sense Representation with HowNet

- Evaluation
  - Word disambiguation with context

Context Word	SEMEME "Capital"	SEMEME "Cuba"
Cuba	0.39	0.42
Russia	0.39	-0.09
cigar	0.00	0.36

Context word attention to sememe "Havana"

Example	Sense1 : probability	Sense2 : probability
Apple is known as the king of fruits	Brand : 0.28	Fruit : 0.72
Apple does not start properly	Brand : 0.87	Fruit : 0.13
Eight **teams** entered the second stage team competition	Group : 0.90	Army : 0.10
Organizational Construction of Public Security Grassroots **Team**	Group : 0.15	Army : 0.85

Disambiguation examples

Conclusion

- WordNet
  - Vocabulary Matrix: Map between words and synonyms
  - Organized with a set of synonyms as the basic unit
- HowNet
  - A web based knowledge system indicates relations of concepts
  - Organized with a set of sememes as the basic unit
- Commons
  - Construct from top to bottom
  - Include hyponymy, synonym and antonym relations

Word Sense Disambiguation

- Word Similarity in Context (SCWS)
  - Word sense disambiguation dataset with context
  - Extend Word Similarity Task, such as WordSim353, to context based version
  - Evaluation: Spearman's correlation coefficient

Context1	Context2	Human (mean)
...Located downtown along the east bank of the Des Moines River...	...This is the basis of all money laundering , a track...	2.75
A shallow bank , little more than high	developed a road while restoring the riverbank and reducing pollution	3.85

E Huang, et al. Improving word representations via global context and multiple word prototypes.

Word Sense Disambiguation

- SemEval (Semantic Evaluation), SemCor and SenseEval
  - Annotation word sense base on WordNet
  - Predict sense according to context
  - e.g.
    - Source sentence:
      - ...interference affecting the integrity of the **air** navigation system...
    - **air**: air.n.1 air.n.2 air.n.3 ...
  - Evaluation: F1

Corpus with NAF format: https://github.com/rubenIzquierdo/wsd_corpora.git

Word Sense Disambiguation

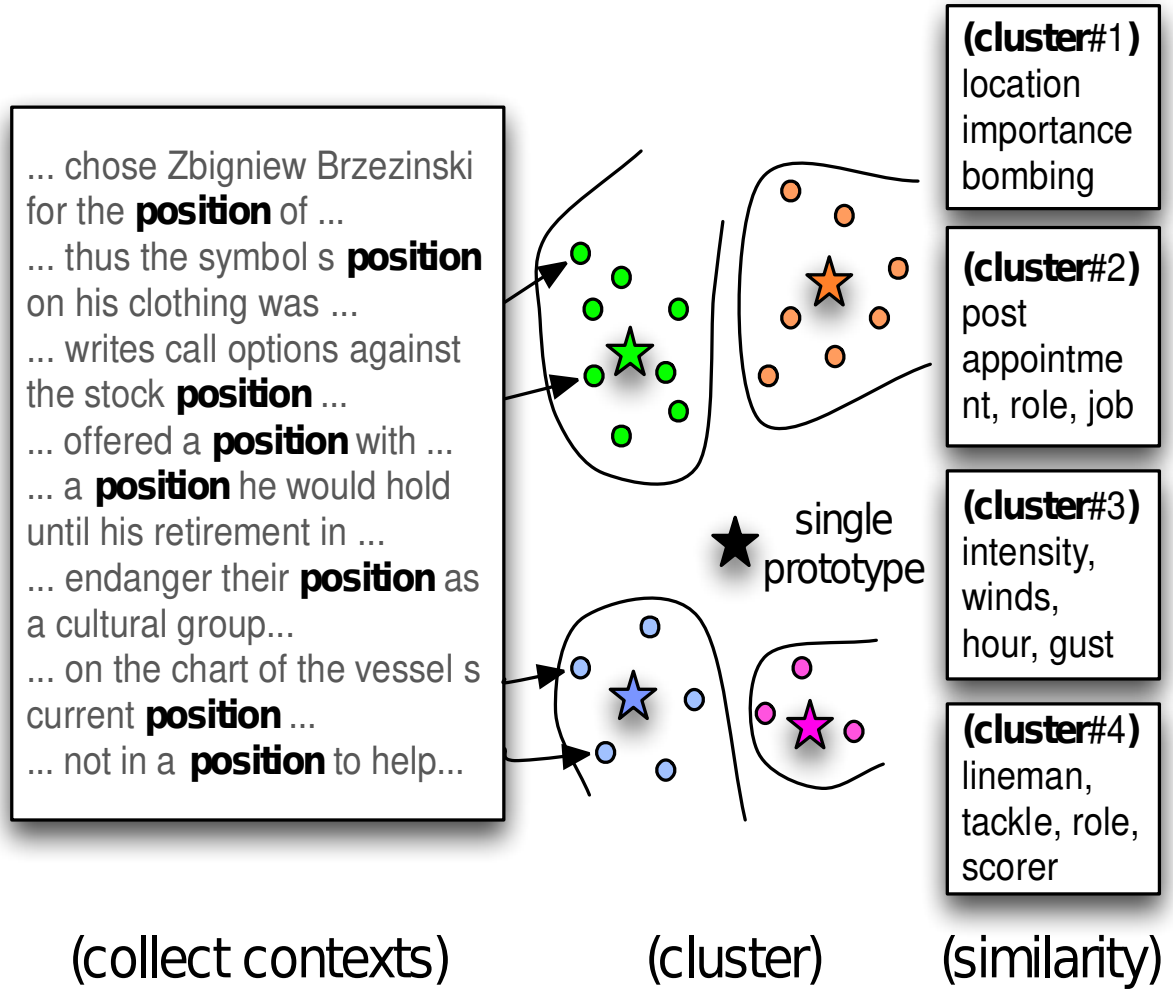
- SemEval (Semantic Evaluation), SemCor and SenseEval
  - Evaluation: F1
    - Precision: $\mathrm{P} = \frac{TP}{TP+F}$
    - Recall: $\mathrm{R} = \frac{TP}{TP+FN}$
    - Consider P and R, F1: $\frac{(\alpha^2+1)PR}{\alpha^2 P+R}$

		True condition	
		Condition positive	**Condition negative**
Predicted condition	**Predicted condition positive**	True positive (TP)	False Positive (FP)
	Predicted condition	False negative (FN)	True negative (TN)

Solution

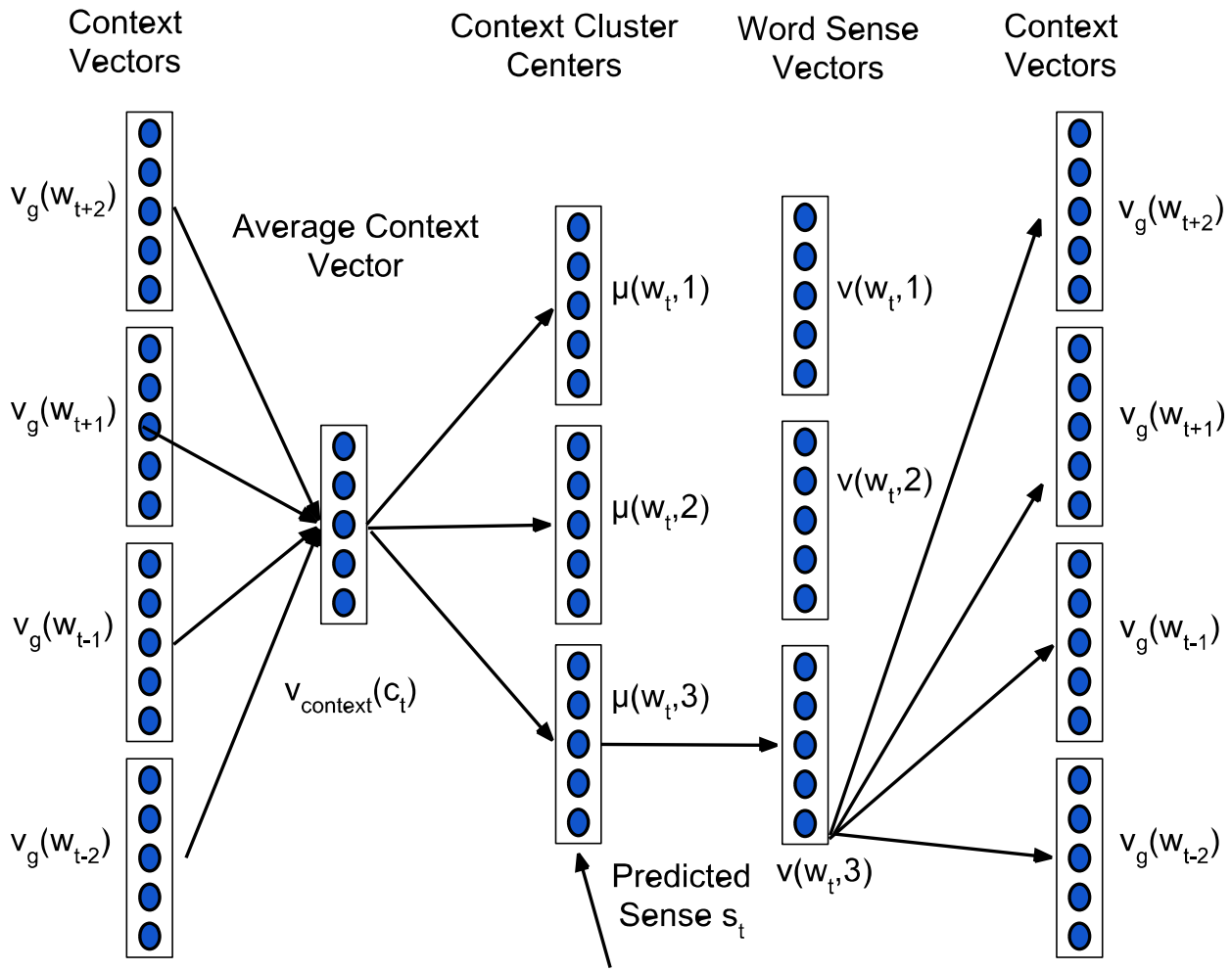
- Multiple Prototype Methods
- Nonparametric Methods
- Word Sense Representation (WSR) and Word Sense Disambiguation (WSD) Unified Model

Multiple Prototype Methods



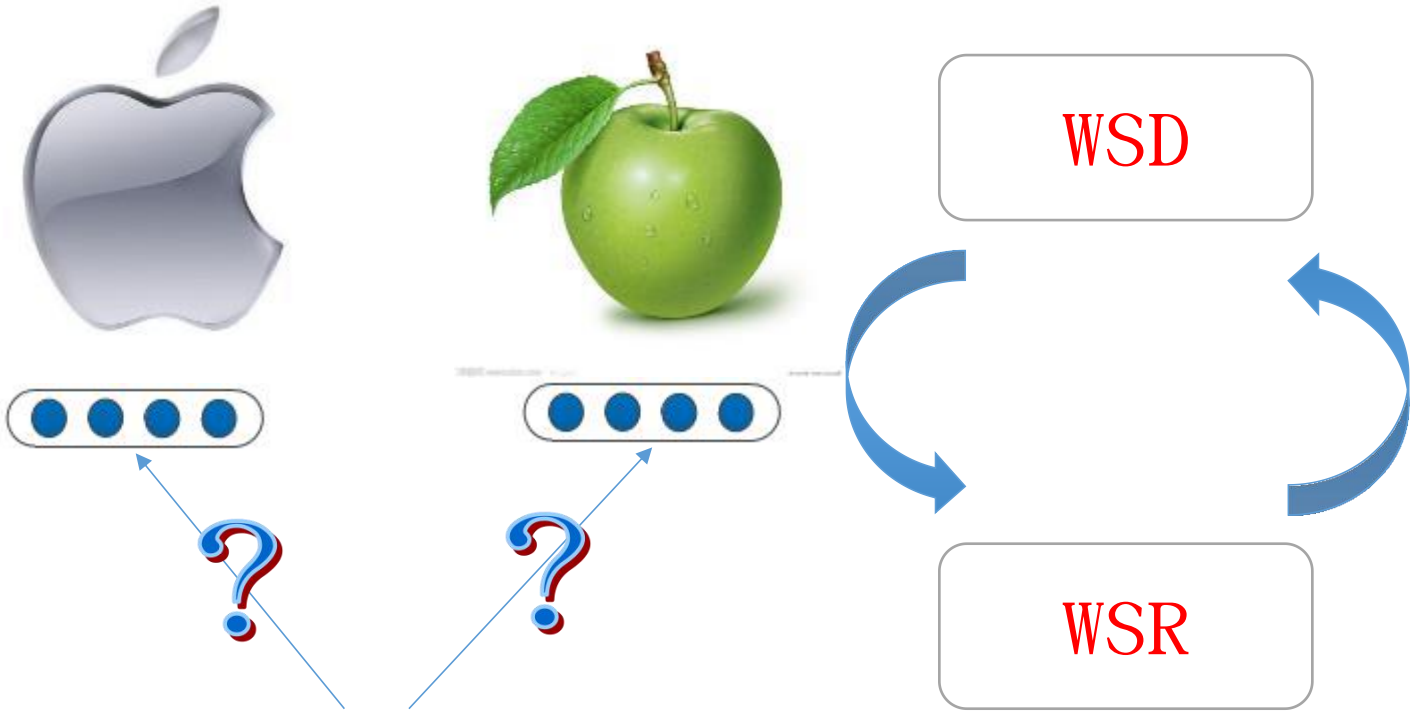
J. Reisinger and R. Mooney. Multi-prototype vector-space models of word meaning. HLT-NAACL 2010.
E Huang, et al. Improving word representations via global context and multiple word prototypes. ACL

Nonparametric Methods



Neelakantan et al. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. EMNLP 2014.

WSD and WSR Unified Model



Chen Xinxiong, et al. A Unified Model for Word Sense Representation and Disambiguation.

WSD and WSR Unified Model

- Methodology
  - Learn word representations (word2vec)
  - Initial sense embedding with average of word embeddings from WordNet gloss

$bank_1$

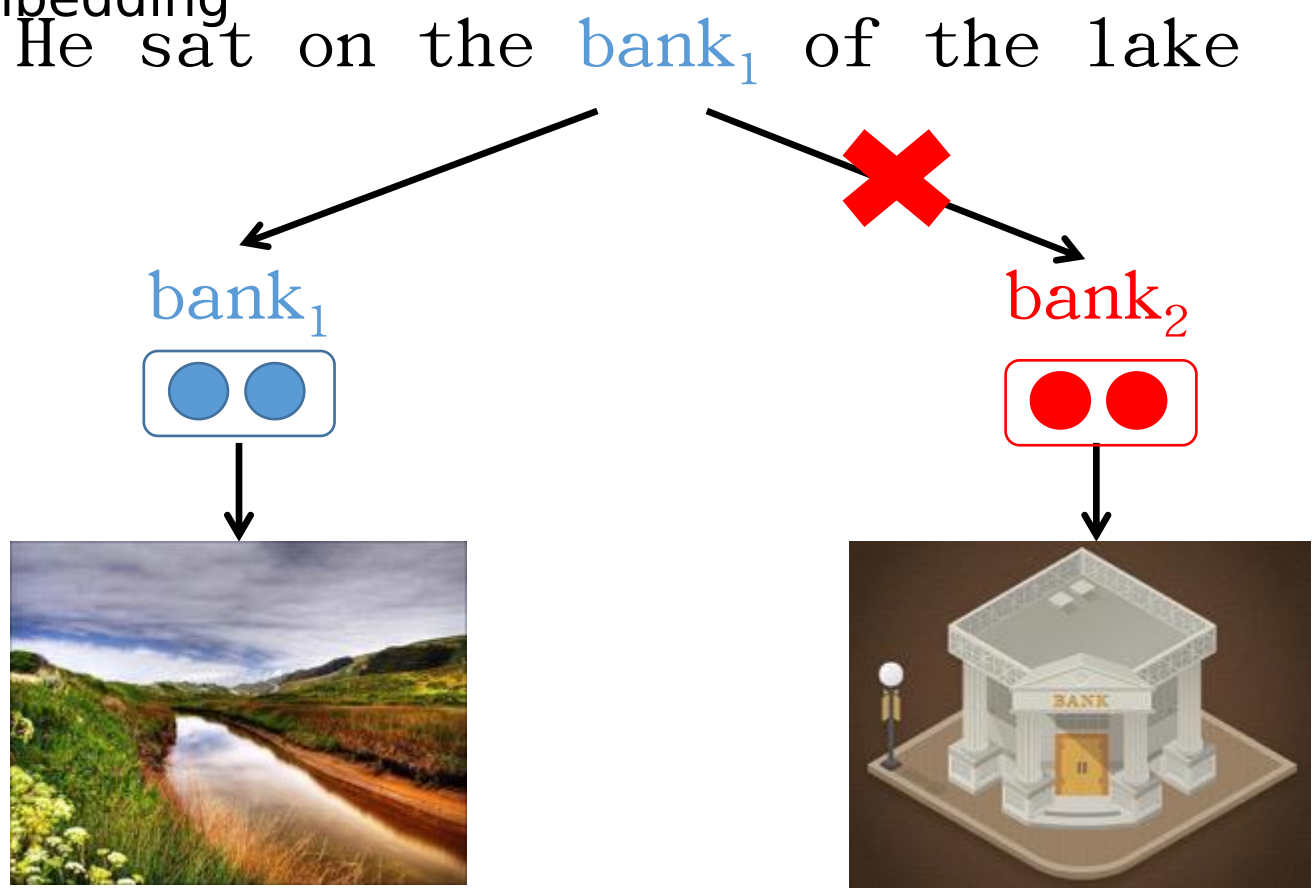
sloping land (especially the slope beside a body of water)

$bank_2$

a financial institution that accepts deposits and channels the money into lending activities

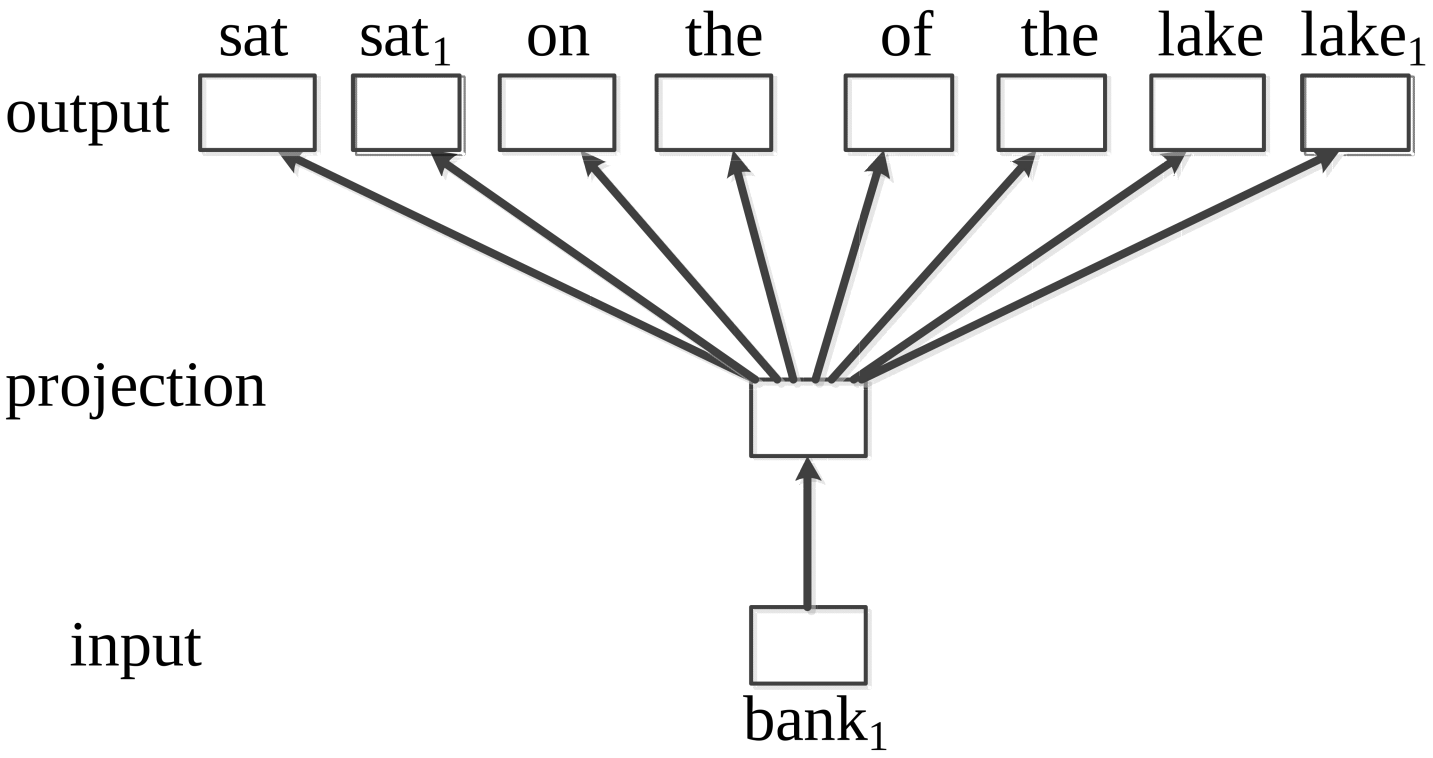
WSD and WSR Unified Model

- Methodology
  - Disambiguation with word embedding and sense embedding



WSD and WSR Unified Model

- Methodology
  - Update word and sense embedding



WSD and WSR Unified Model

- Examples

Word or Sense	Nearest neighbors
bank	banks, IDBI, CitiBank
$bank_1$	river, slope, Sooes
$bank_2$	mortgage, lending, loans
star	stars, stellar, trek
$star_1$	photosphere, radiation, gamma-rays
$star_2$	someone, skilled, genuinely
plant	plants, glavaticevo, herbaceous
$plant_1$	factories, machinery, manufacturing

WSD and WSR Unified Model

- Evaluation
  - Dataset: Sports and finance[1]
    - 41 words and 100 sentences per word

Algorithm	Sports Recall	Finance Recall
Random BL	19.5	19.6
MFS BL	19.6	37.1
k-NN	30.3	43.4
Static PR	20.1	39.6
Personalized PR	35.6	46.9
Degree	42.0	47.8
Our Model	**57.3**	**60.6**

[1]Rob Koeling and Diana McCarthy. Sussx: Ws- d using automatically acquired predominant senses. In Proceedings of SemEval 2007.

WSD and WSR Unified Model

- Evaluation
  - Dataset: SemEval2007

Algorithm	Type	Nouns F1	All F1
Random BL	U	63.5	62.7
MFS BL	Semi	77.4	78.9
SUSSX-FR	Semi	81.1	77.0
NUS-PT	S	82.3	**82.5**
SSI	Semi	**84.1**	**83.2**
Degree	Semi	**85.5**	**81.7**
Our Model	U	81.6	75.8
Our Model	Semi	**85.3**	**82.6**

WSD and WSR Unified Model

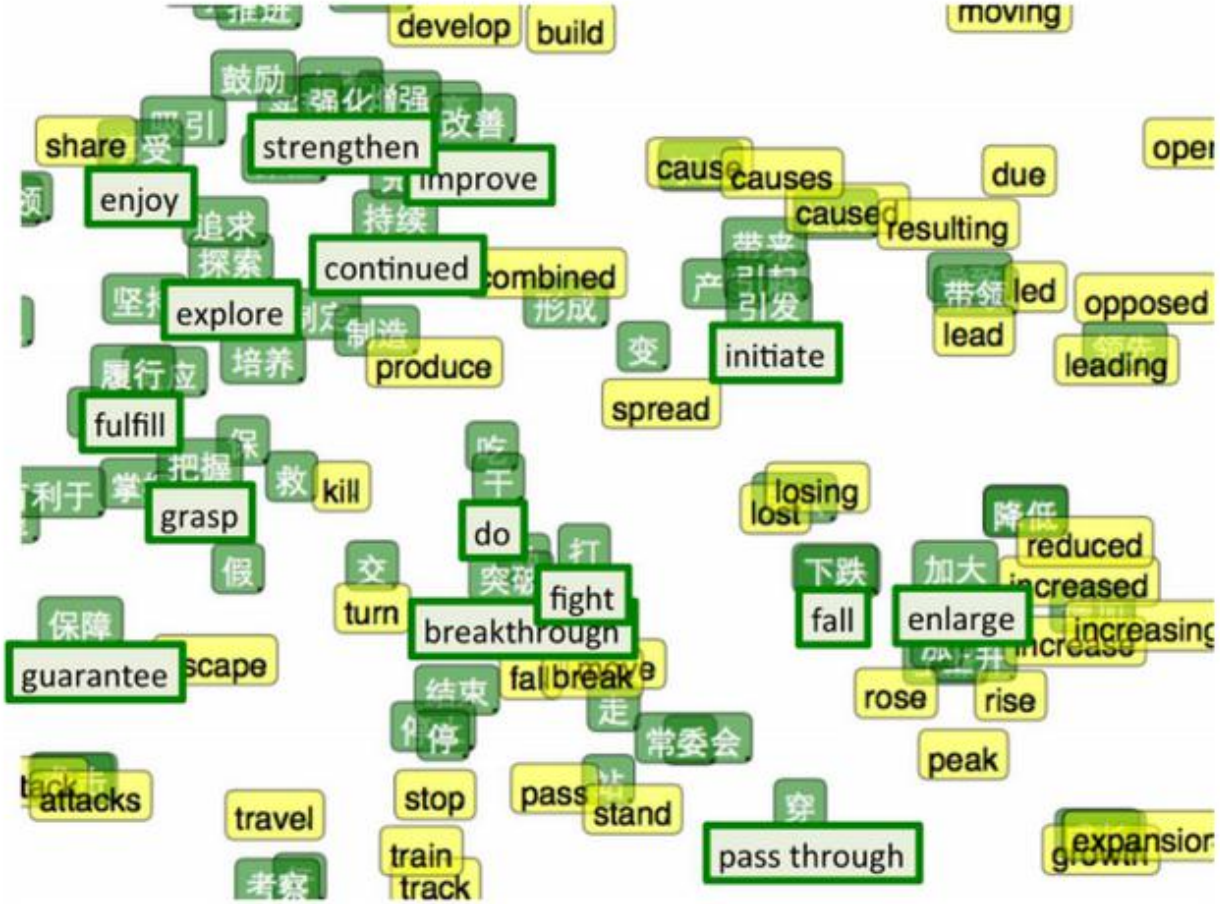
- Evaluation
  - Dataset: SCWS

Model	
C&W-S	57.0
Huang-S	58.6
Huang-M AvgSim	62.8
Huang-M AvgSimC	65.7
Our Model-S	64.2
Our Model-M AvgSim	66.2
Our Model-M AvgSimC	**68.9**

Conclusion

- Unsupervised WSD Methods
  - Multiple Prototype Methods
    - The number of word sense is set to constant
  - Nonparametric Methods
    - Dynamic sense number according to different words
    - Slower than Multiple Prototype Methods
  - WSD and WSR Unified Model
    - Disambiguation with WordNet glosses
    - Fine-tuning with large corpus for further improvement

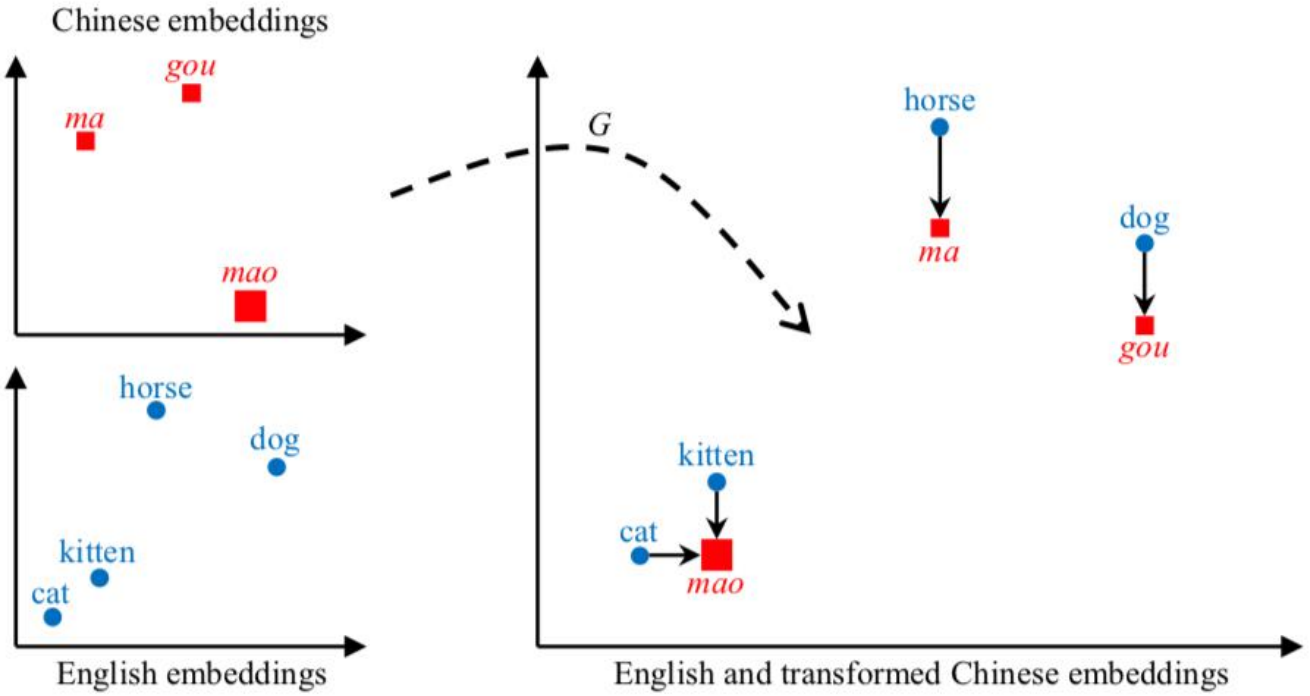
Bilingual Word Embeddings



Zou, Will Y., et al. Bilingual word embeddings for phrase-based machine translation.

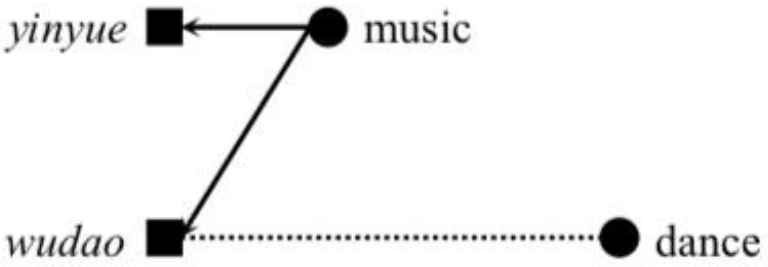
Problem Statement

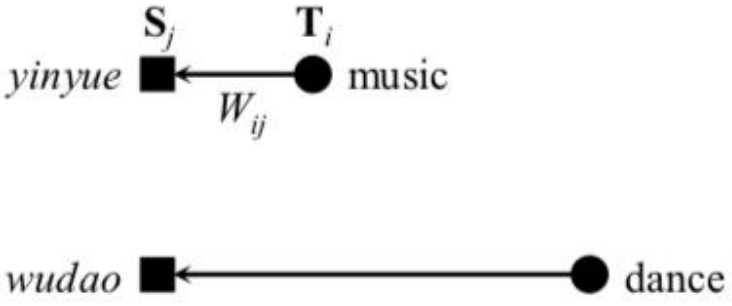
- Bilingual word alignment
  - Find word pairs in different languages with the same semantic
  - Challenge: many-to-many mappings



Bilingual Word Alignment

- Earth Mover's Distance
  - Nearest neighbor (left)
  - Earth Mover's Distance (right)





wudao is closer to music

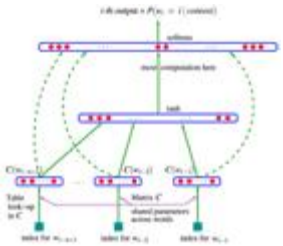
Zhang et al. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation. AAAI 2016.

Resources

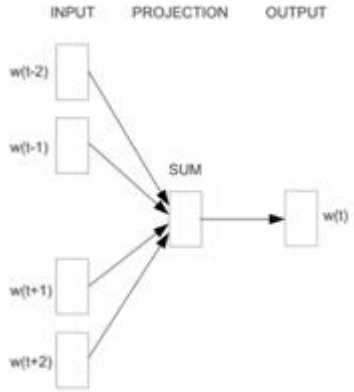
- Knowledge base
  - WordNet:
    - https://wordnet.princeton.edu
  - HowNet:
    - http://www.keenage.com
- Tutorials of word2vec
  - Principle tutorial:
    - https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html#context-based-skip-gram-model
  - Tensorflow implementation:
    - https://www.tensorflow.org/tutorials/representation/word2vec
  - PyTorch implementation:
    - https://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html
  - C/C++ implementation:
    - https://code.google.com/archive/p/word2vec/
  - Python implementation:
    - https://radimrehurek.com/gensim/models/word2vec.html
- State-of-the-art Model
  - Analogy:
    - https://aclweb.org/aclwiki/Analogy_(State_of_the_art)

Re-search, Re-invent

word2vec ≃ MF

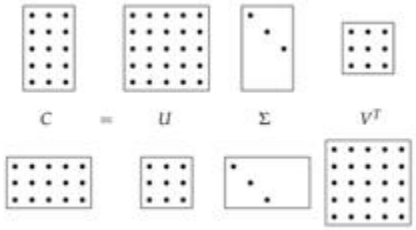


Neural
Language
Models



Distributional
Repre

he curtains open and the stars shining in on the barely
ars and the cold , close stars " . And neither of the w
rough the night with the stars shining so brightly , it
made in the light of the stars . It all boils down , wr
surely under the bright stars , thrilled by ice-white
sun , the seasons of the stars ? Home , alone , Jay pla
n is dazzling snow , the stars have risen full and cold



SVD

Levy and Goldberg. Neural word embedding as implicit matrix factorization. NIPS 2014.