Lecture 1:
Course Introduction

CMPSCI 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor

UMASSCS
SCHOOL OF COMPUTER SCIENCE
50 YEARS

What

- Learn fundamental principles and methods in natural language processing
    - Hands-on implementation experience
    - Appreciation of basic linguistic issues
    - Know about useful open-source NLP tools
    - Know when NLP works and when it doesn’t
    - Be able to read current research papers in NLP
- “AI systems”

How

- Lots of math!
    - This course: more than traditional CS, but less than traditional physics, engineering
- Lots of data!
- Lots of code that *implements* math & algorithms
    - Skill: translating from math to code
    - Skill: debugging math/linguistic/algorithm code
- A little bit of linguistics (goes a long way)

Requirements

- (20%) Participation and short exercises
- (30%) Problem sets
  - Written: math and concepts
  - Programming: Python programs
- (20%) Midterm (in-class, Oct 20)
- (30%) Final projects (groups of 1-2)
  - Choose a topic, or select a suggested topic
  - Project Proposal
  - Progress Report
  - In-class presentations (last week)
  - Final Report

Logistics

- Main course website for assignments and links to everything. This is ground truth if there are contradictions http://people.cs.umass.edu/~brenocon/inlp2015/
- Piazza for announcements, discussions and restricted access files
- Moodle for homework submissions (sometimes) and to see grades

- Waitlist situation

- To check:
  - SPIRE-registered students should have Piazza invites. Check @umass.edu email if you don't!
  - Email me if you can't access Piazza.
- Homework #0 due Thursday!

Readings

- Readings will be provided as PDFs on Piazza
  - Draft chapters of Jurafsky and Martin, *Speech and Language Processing*
  - We will use selections from both
    - 2nd edition (published)
    - 3rd edition (unpublished)
  - Other readings on occasion as well

Related courses at UMass

- Computational Linguistics: Ling 409, 492B (Bhatt, Dillon)

  https://sites.google.com/site/umasslx409/home

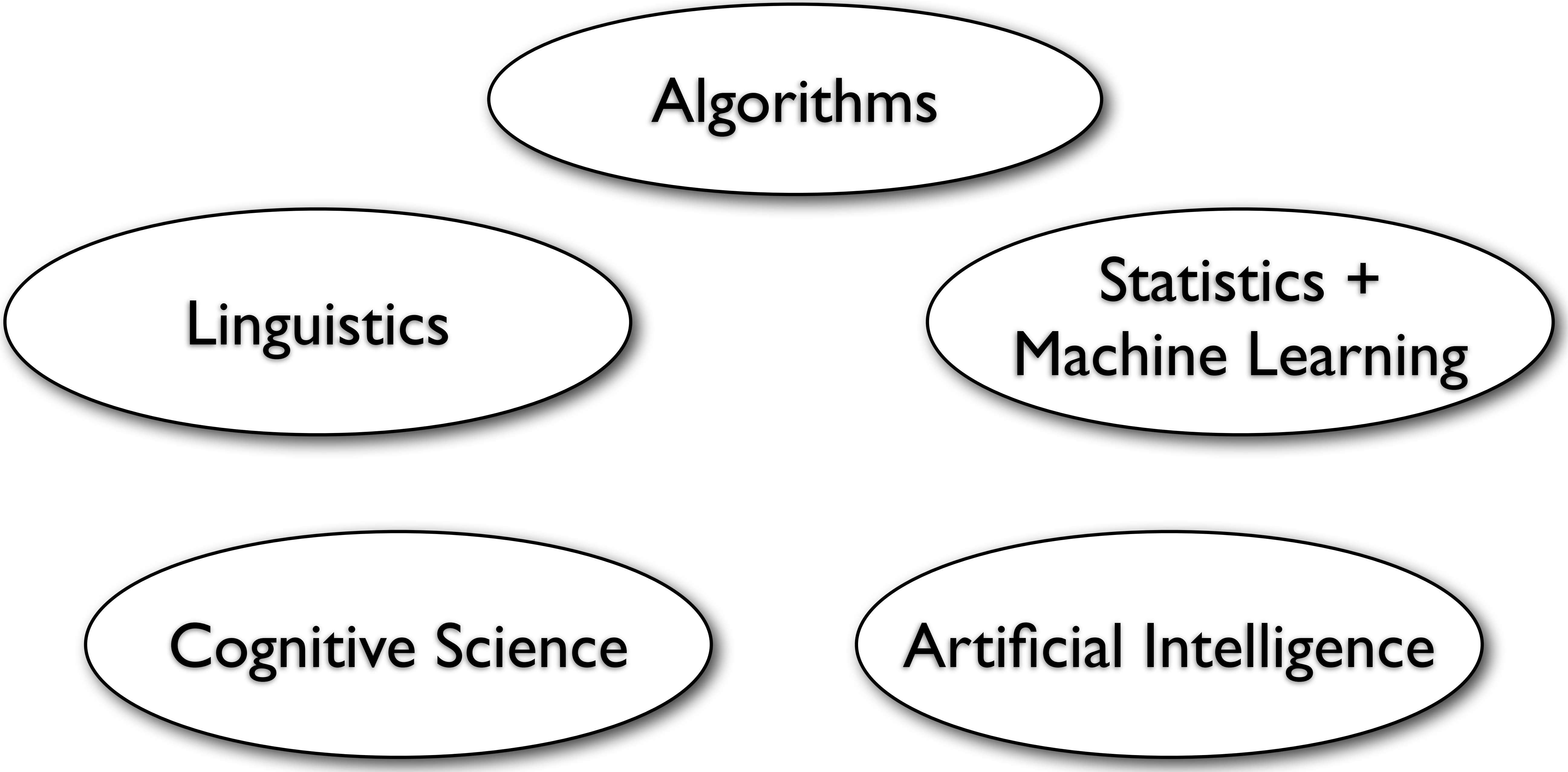
  http://www.umass.edu/linguist/courses/detail.php?cid=571
- Speech: Ling 592B (Yu)

  http://courses.umass.edu/linguist592b-kmyu/category/info.html
- Information Retrieval: CS 446, 646 (Allan, Croft)

  http://ciir.cs.umass.edu/cmpsci446/

  http://ciir.cs.umass.edu/~allan/cs646/

NLP is interdisciplinary



Algorithms

Linguistics

Statistics +
Machine Learning

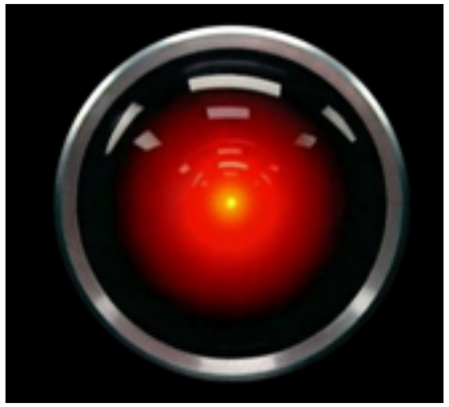
Cognitive Science

Artificial Intelligence

“Can Machines Think?”

- British mathematician and founding figure in computer science
- Alan Turing (1950)
- How do we know when we have AI?
- “Imitation Game”

NLP imagined



NLP today

- Speech interfaces
- Machine translation
- Sentiment analysis
- Search engines
- ...

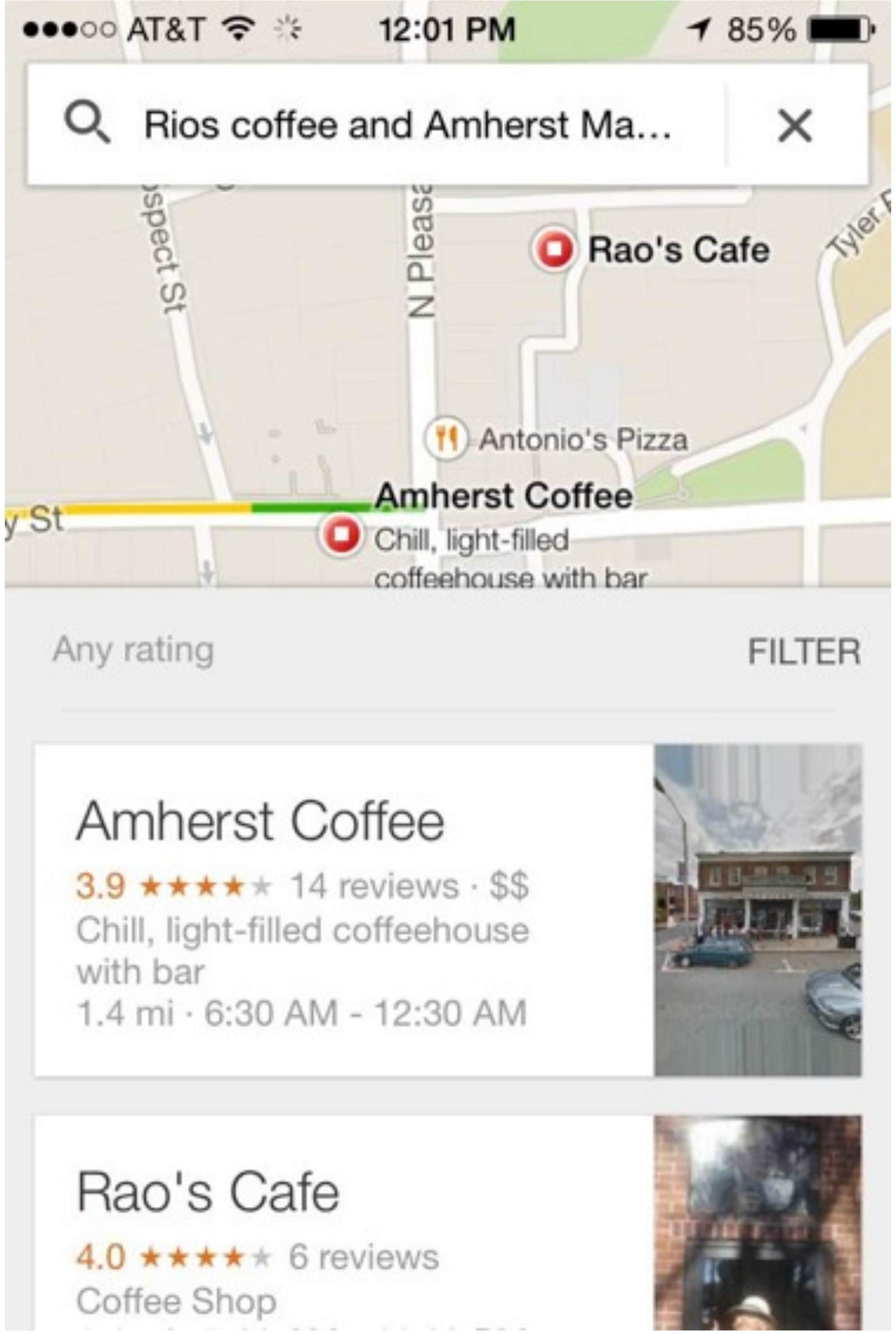
- [This course: document text analysis]

NLP today: Speech interfaces

What can I help you with?

“Rao’s coffee in Amherst, Massachusetts”

“Rao’s coffee in Amherst, Massachusetts”



Amherst Coffee

3.9 ★★★★★ 14 reviews · $$

Chill, light-filled coffeehouse with bar

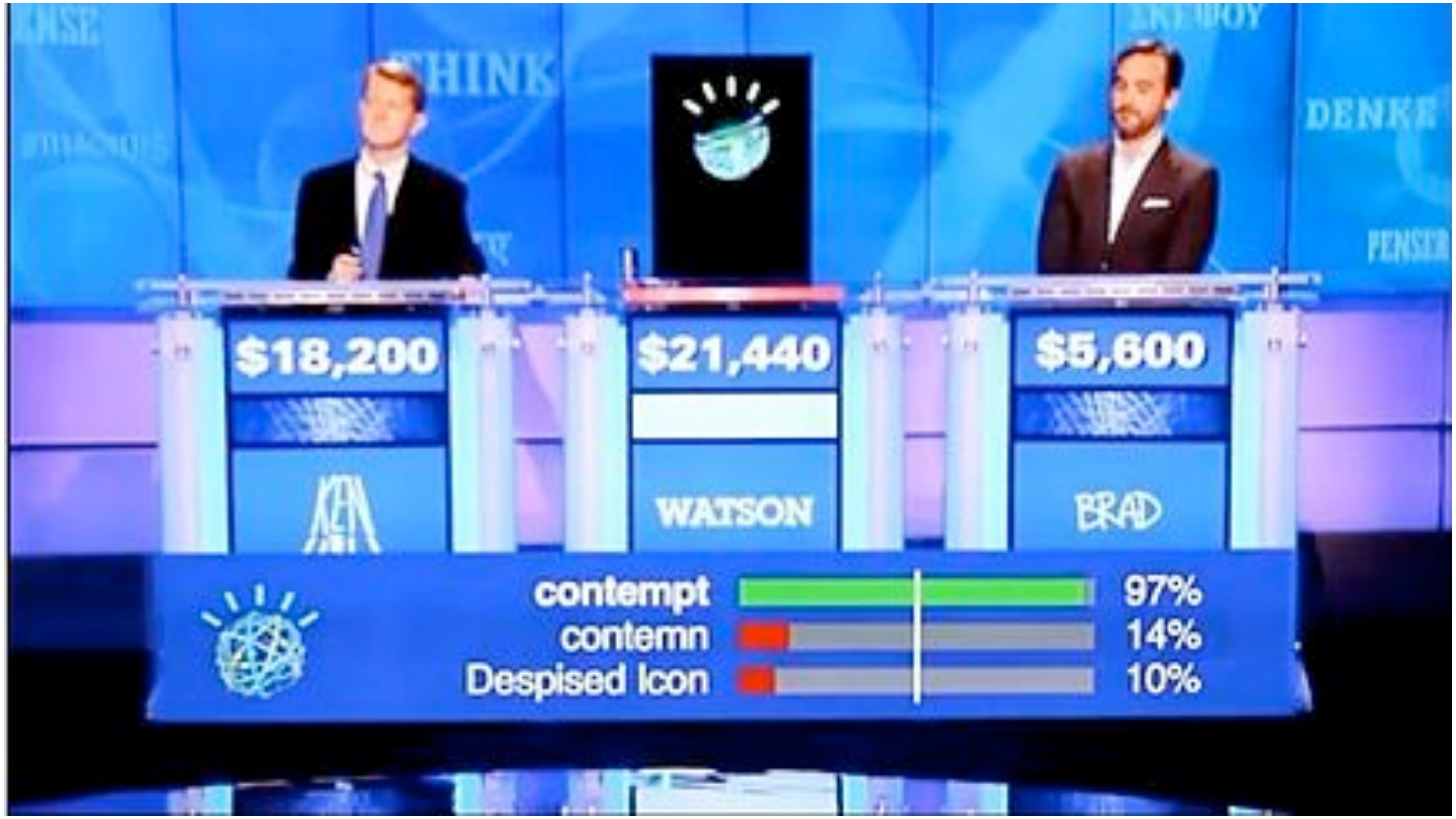
1.4 mi · 6:30 AM - 12:30 AM

Rao's Cafe

4.0 ★★★★★ 6 reviews

Coffee Shop

NLP today: Question answering

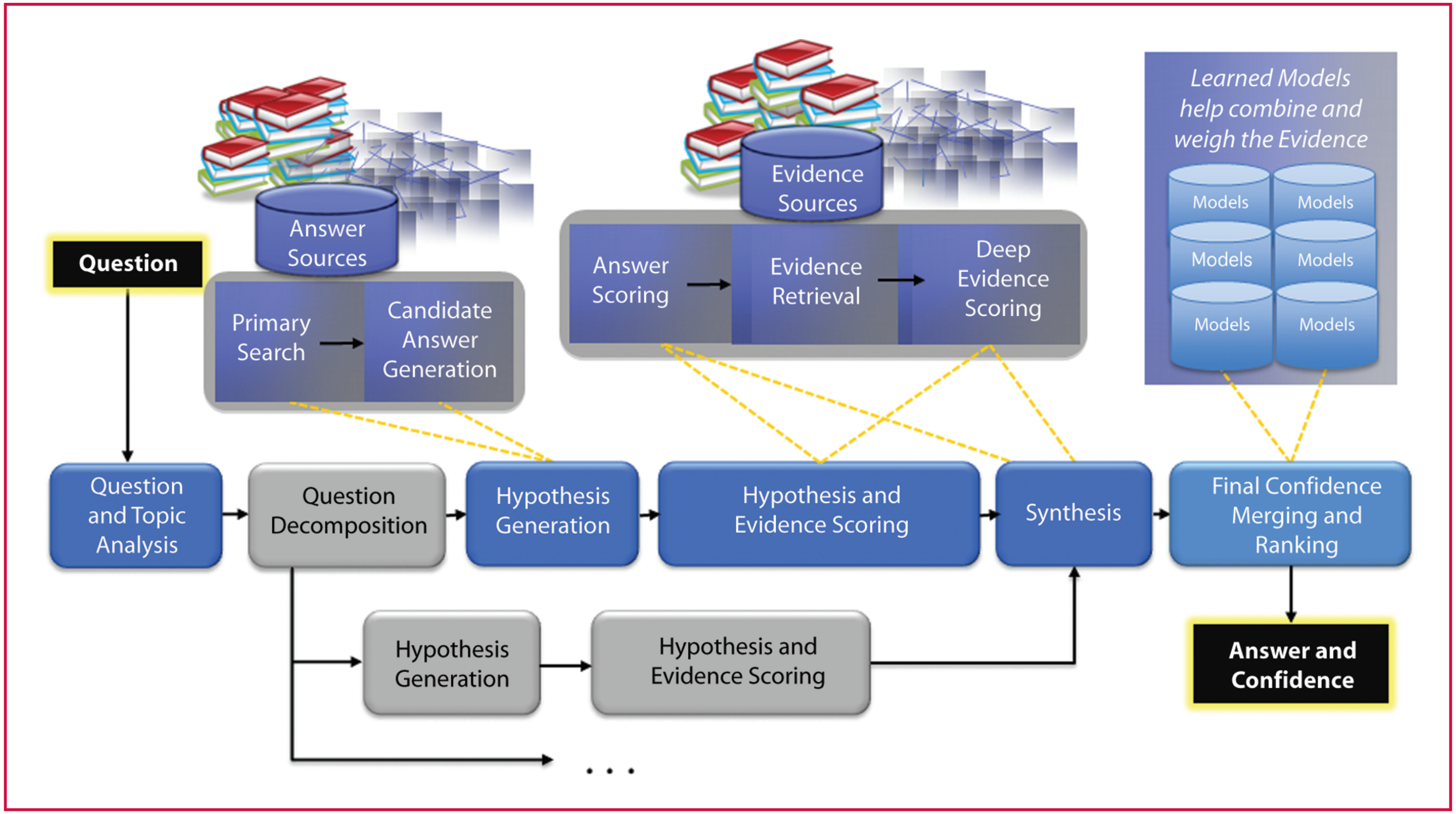


IBM Watson

Wanted for general evilness, last seen at the Tower of Barad-Dur. It's a giant eye, folks, kinda hard to miss

NLP today: Question answering

From IBM Journal of Research and Development, 2012



IBM Watson

25 engineers, 4 years, 200 subsystems,
2,880 CPU cores, 15 TB storage

NLP today: Question answering

From IBM Journal of Research and Development, 2012

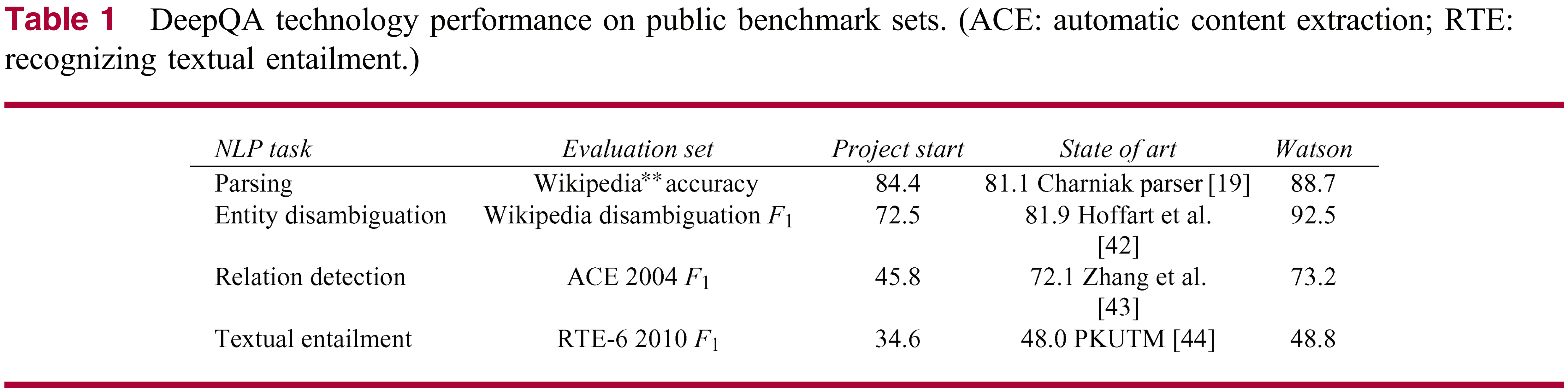


Table 1 DeepQA technology performance on public benchmark sets. (ACE: automatic content extraction; RTE: recognizing textual entailment.)

NLP task	*Evaluation set*	*Project start*	*State of art*	*Watson*
Parsing	Wikipedia** accuracy	84.4	81.1 Charniak parser [19]	88.7
Entity disambiguation	Wikipedia disambiguation F_1	72.5	81.9 Hoffart et al. [42]	92.5
Relation detection	ACE 2004 F_1	45.8	72.1 Zhang et al. [43]	73.2
Textual entailment	RTE-6 2010 F_1	34.6	48.0 PKUTM [44]	48.8

IBM Watson

Imperfect NLP is still useful

Ambiguity: why NLP is hard

Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant

Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors

Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.

Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.

Ambiguity: why NLP is hard

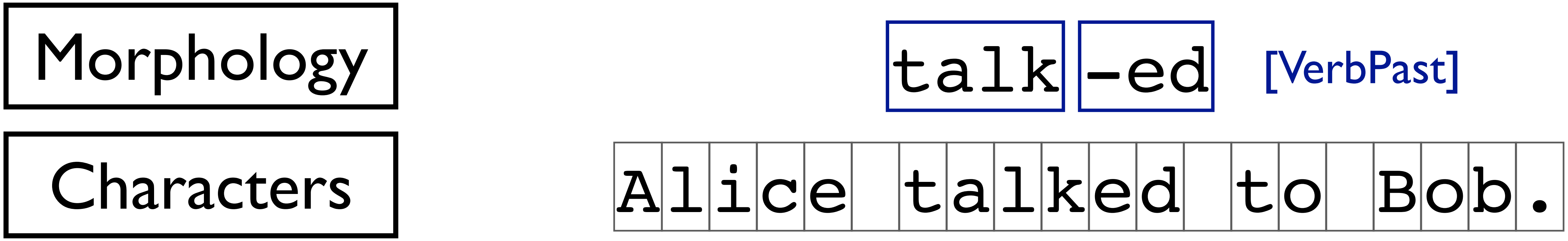
- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.
- They found that in order to attract settlers -- and make a profit from their holdings -- they had to offer people farms, not just tenancy on manorial estates.

Levels of linguistic structure

Characters

A	l	i	c	e		t	a	l	k	e	d		t	o		B	o	b	.

Levels of linguistic structure



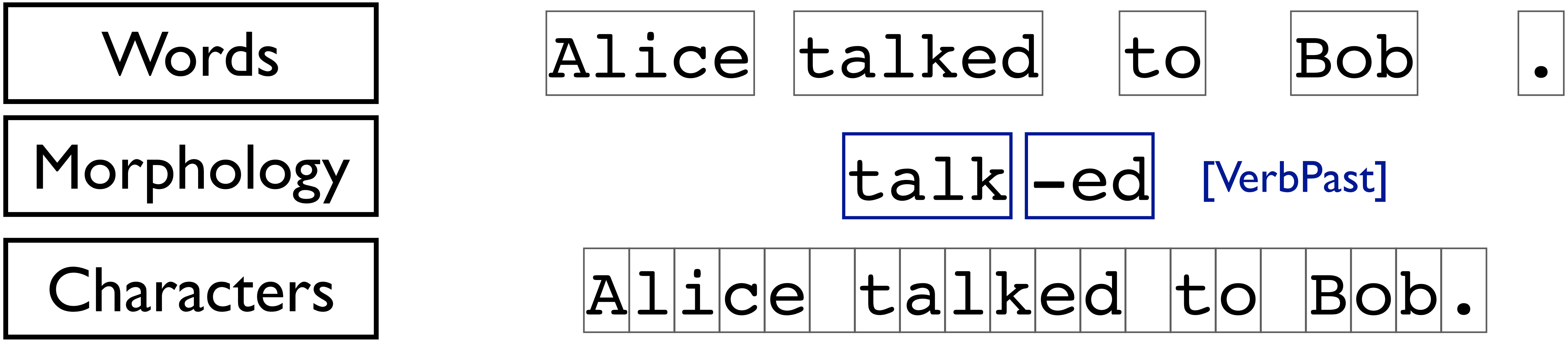
Morphology

talk -ed [VerbPast]

Characters

A	l	i	c	e		t	a	l	k	e	d		t	o		B	o	b	.

Levels of linguistic structure

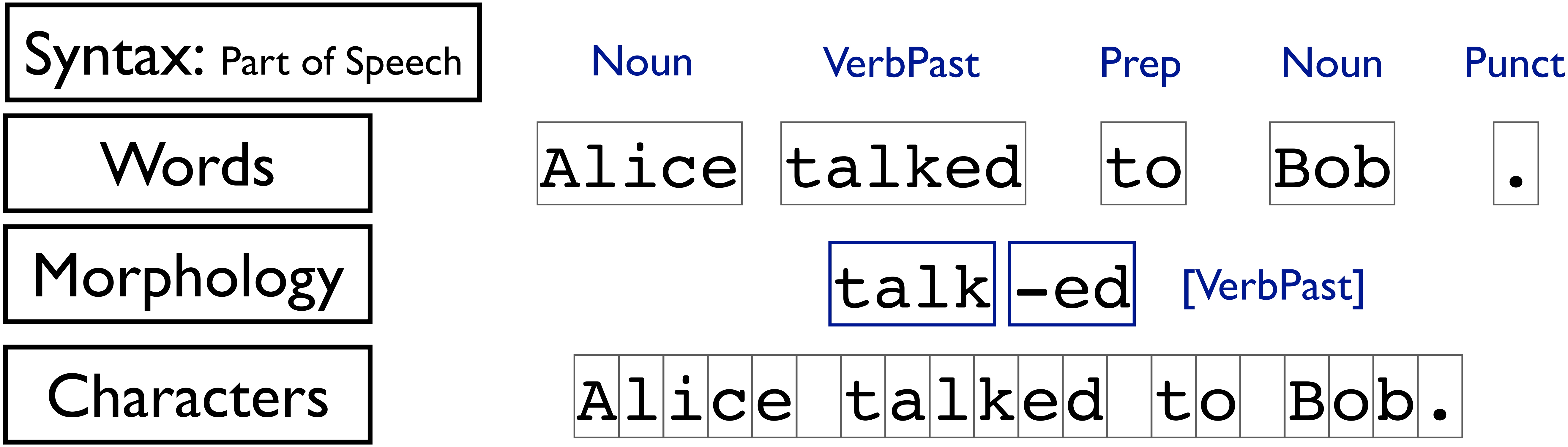


Words — Alice talked to Bob .

Morphology — talk -ed [VerbPast]

Characters — Alice talked to Bob.

Levels of linguistic structure



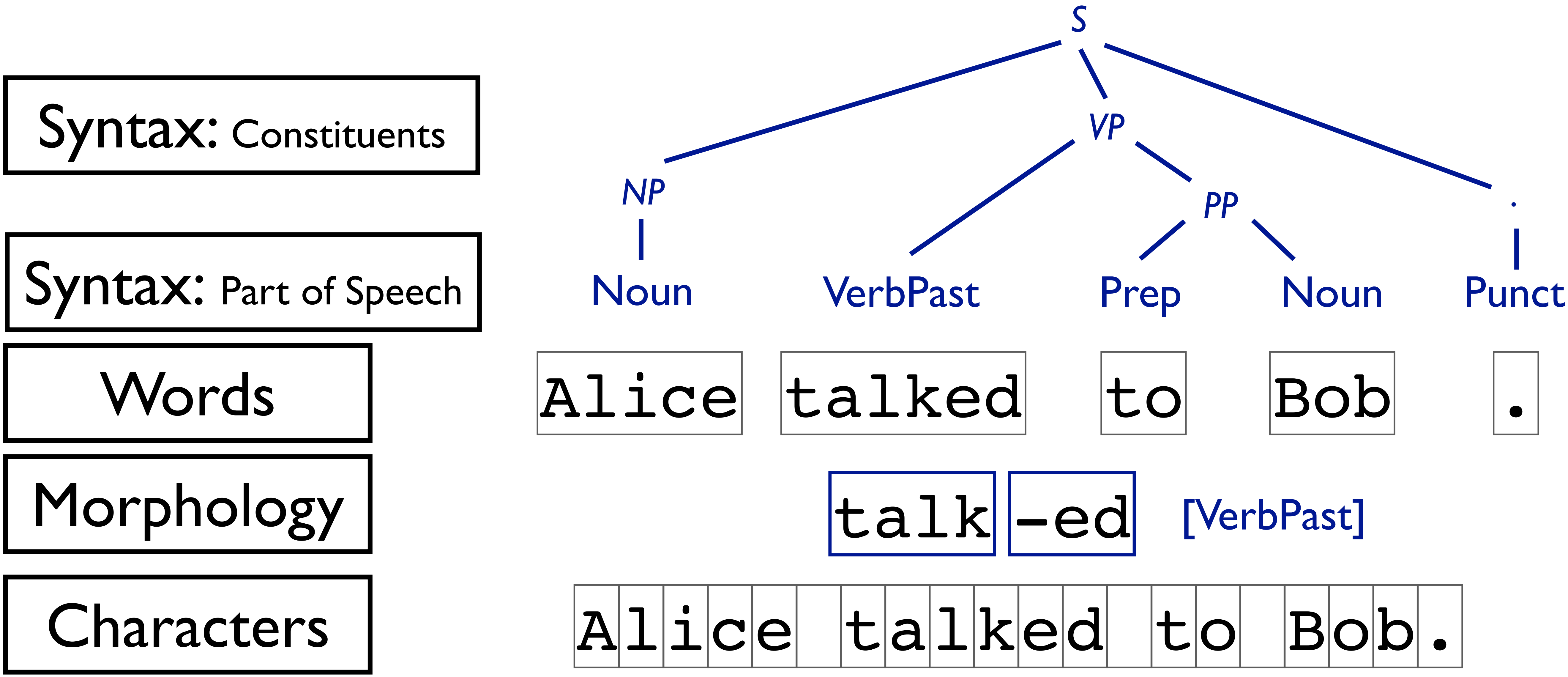
Syntax: Part of Speech — Noun VerbPast Prep Noun Punct

Words — Alice talked to Bob .

Morphology — talk -ed [VerbPast]

Characters — Alice talked to Bob.

Levels of linguistic structure



S

NP VP .

PP

Syntax: Constituents

Syntax: Part of Speech

Noun VerbPast Prep Noun Punct

Words

Alice talked to Bob .

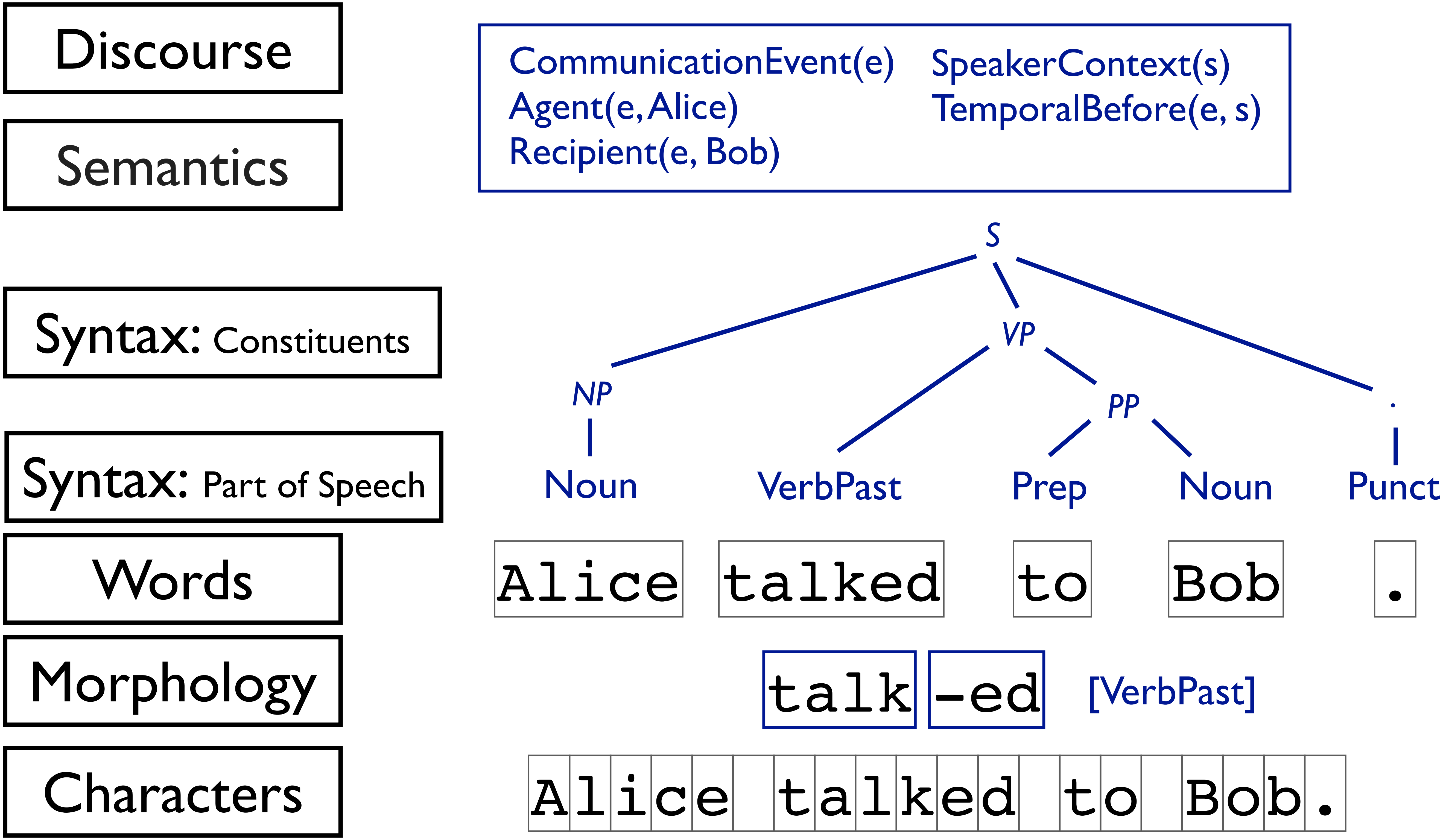
Morphology

talk -ed [VerbPast]

Characters

A l i c e t a l k e d t o B o b .

Levels of linguistic structure



Discourse

Semantics

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)

SpeakerContext(s)
TemporalBefore(e, s)

Syntax: Constituents

S
NP
VP
PP
.

Syntax: Part of Speech

Noun VerbPast Prep Noun Punct

Words

Alice talked to Bob .

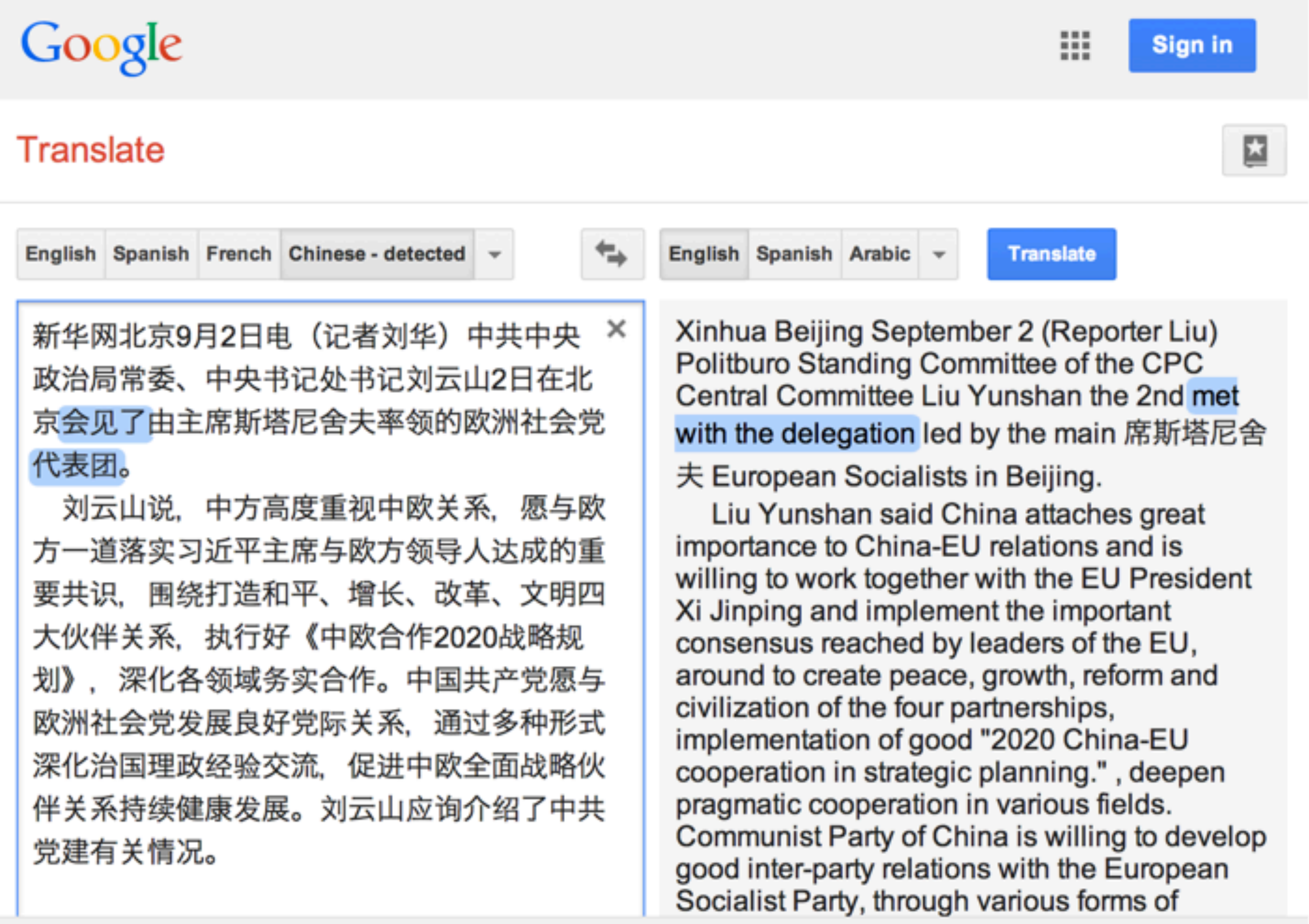
Morphology

talk -ed [VerbPast]

Characters

Alice talked to Bob.

NLP today: Machine translation



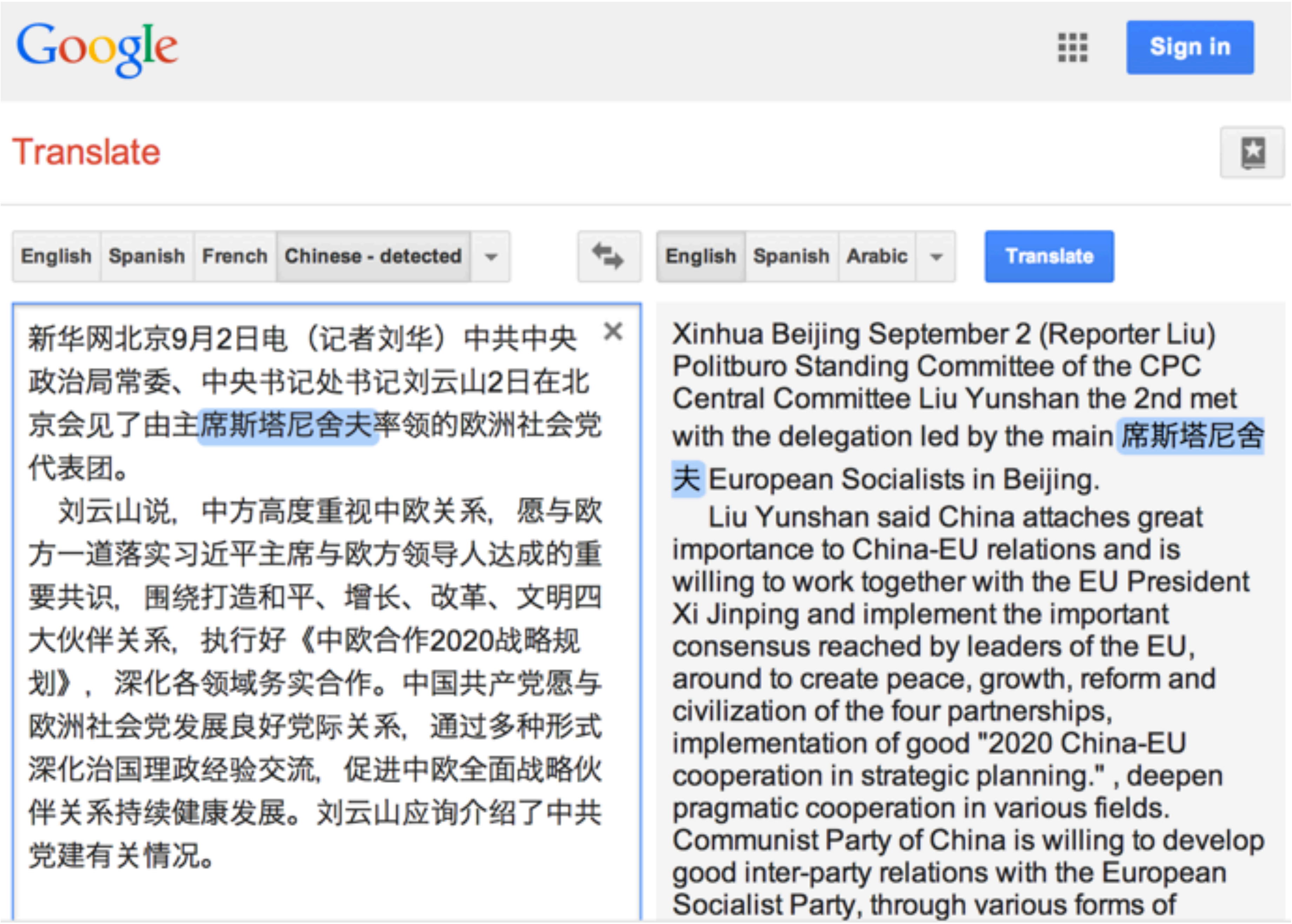
新华网北京9月2日电（记者刘华）中共中央政治局常委、中央书记处书记刘云山2日在北京会见了由主席斯塔尼舍夫率领的欧洲社会党代表团。

刘云山说，中方高度重视中欧关系，愿与欧方一道落实习近平主席与欧方领导人达成的重要共识，围绕打造和平、增长、改革、文明四大伙伴关系，执行好《中欧合作2020战略规划》，深化各领域务实合作。中国共产党愿与欧洲社会党发展良好党际关系，通过多种形式深化治国理政经验交流，促进中欧全面战略伙伴关系持续健康发展。刘云山应询介绍了中共党建有关情况。

Xinhua Beijing September 2 (Reporter Liu) Politburo Standing Committee of the CPC Central Committee Liu Yunshan the 2nd met with the delegation led by the main 席斯塔尼舍夫 European Socialists in Beijing.

Liu Yunshan said China attaches great importance to China-EU relations and is willing to work together with the EU President Xi Jinping and implement the important consensus reached by leaders of the EU, around to create peace, growth, reform and civilization of the four partnerships, implementation of good "2020 China-EU cooperation in strategic planning." , deepen pragmatic cooperation in various fields. Communist Party of China is willing to develop good inter-party relations with the European Socialist Party, through various forms of

NLP today: Machine translation



新华网北京9月2日电（记者刘华）中共中央政治局常委、中央书记处书记刘云山2日在北京会见了由主席斯塔尼舍夫率领的欧洲社会党代表团。

刘云山说，中方高度重视中欧关系，愿与欧方一道落实习近平主席与欧方领导人达成的重要共识，围绕打造和平、增长、改革、文明四大伙伴关系，执行好《中欧合作2020战略规划》，深化各领域务实合作。中国共产党愿与欧洲社会党发展良好党际关系，通过多种形式深化治国理政经验交流，促进中欧全面战略伙伴关系持续健康发展。刘云山应询介绍了中共党建有关情况。

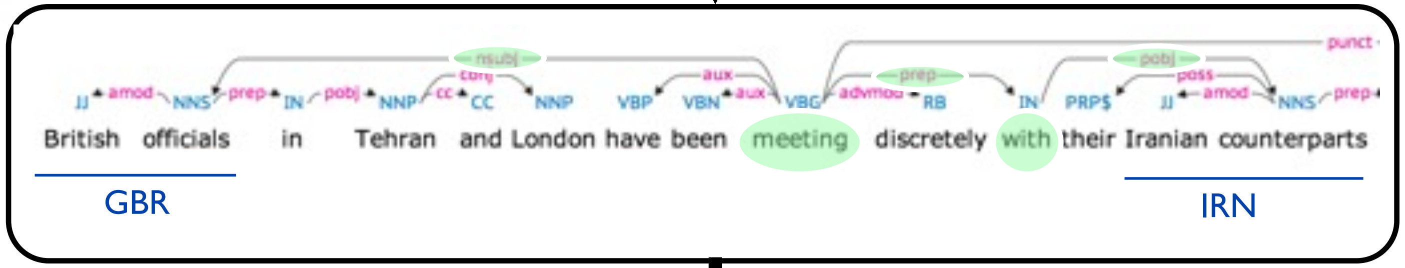
Xinhua Beijing September 2 (Reporter Liu) Politburo Standing Committee of the CPC Central Committee Liu Yunshan the 2nd met with the delegation led by the main 席斯塔尼舍夫 European Socialists in Beijing.

Liu Yunshan said China attaches great importance to China-EU relations and is willing to work together with the EU President Xi Jinping and implement the important consensus reached by leaders of the EU, around to create peace, growth, reform and civilization of the four partnerships, implementation of good "2020 China-EU cooperation in strategic planning." , deepen pragmatic cooperation in various fields. Communist Party of China is willing to develop good inter-party relations with the European Socialist Party, through various forms of

NLP today: Trend analysis

Data: news articles

Dependency parsing to identify events



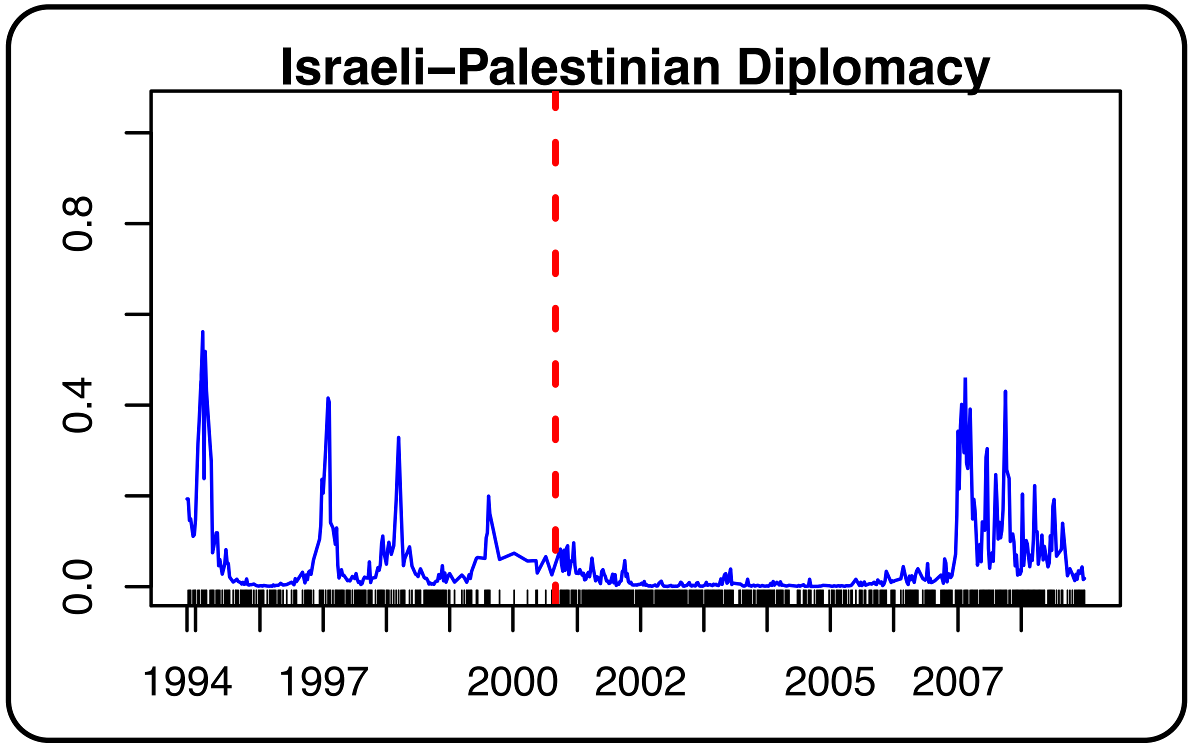
Machine learning from text:

(1) **Event class dictionaries**

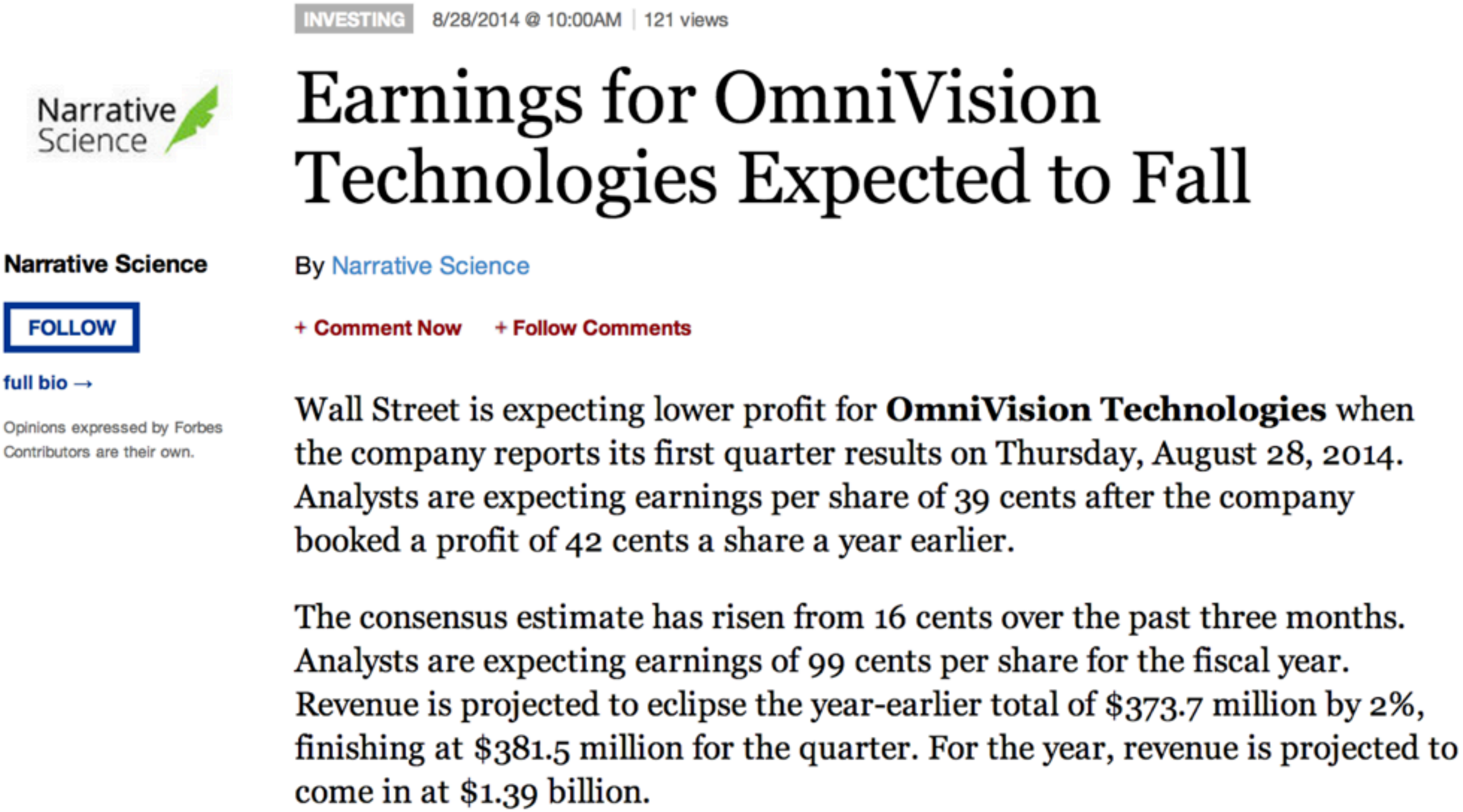
"diplomacy"	arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to,
"verbal conflict"	accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss,
"material conflict"	kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops

↔

(2) **Political dynamics**



NLP today: Story generation



8/28/2014 @ 10:00AM | 121 views

Earnings for OmniVision Technologies Expected to Fall

By Narrative Science

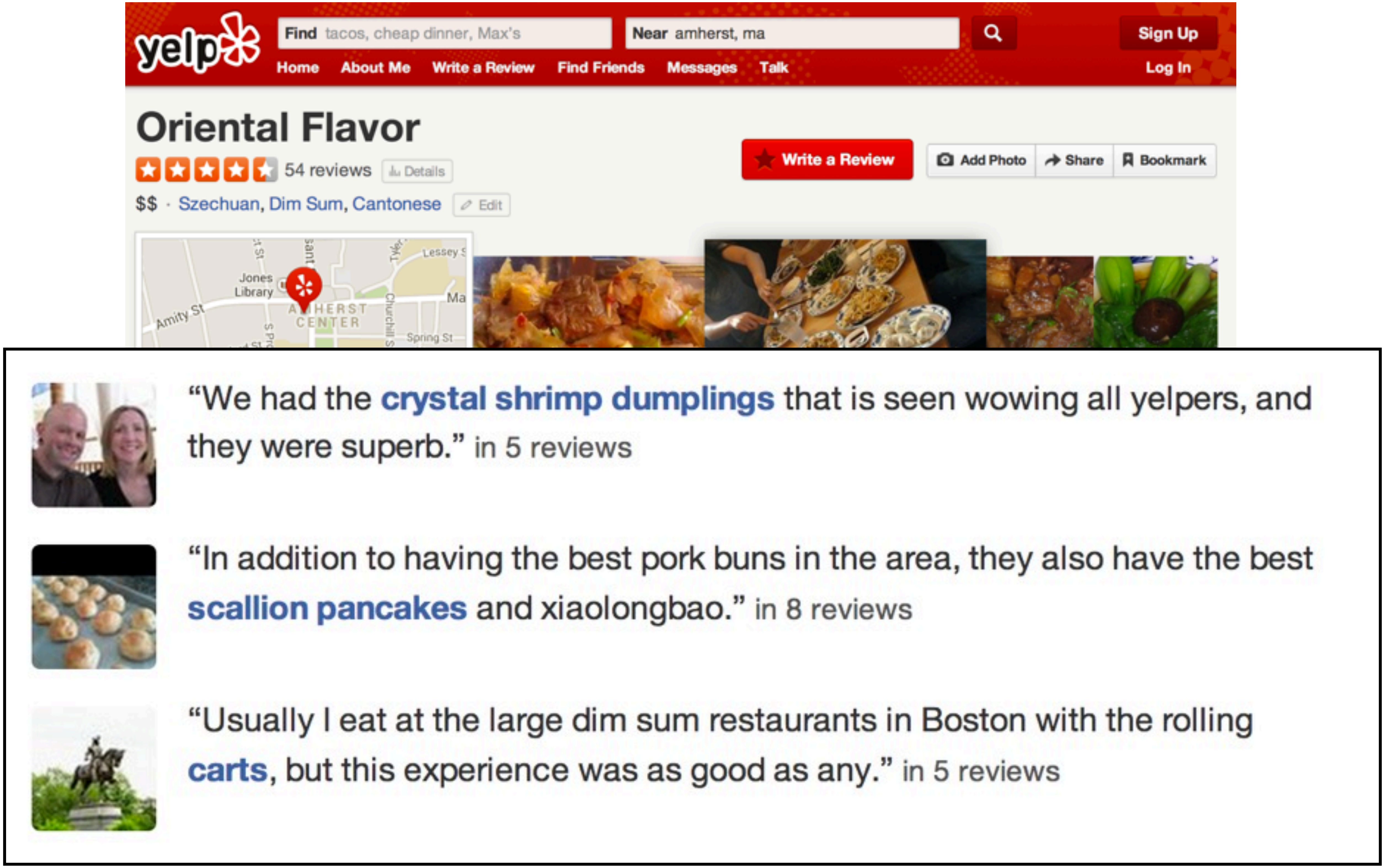
Opinions expressed by Forbes Contributors are their own.

Wall Street is expecting lower profit for **OmniVision Technologies** when the company reports its first quarter results on Thursday, August 28, 2014. Analysts are expecting earnings per share of 39 cents after the company booked a profit of 42 cents a share a year earlier.

The consensus estimate has risen from 16 cents over the past three months. Analysts are expecting earnings of 99 cents per share for the fiscal year. Revenue is projected to eclipse the year-earlier total of $373.7 million by 2%, finishing at $381.5 million for the quarter. For the year, revenue is projected to come in at $1.39 billion.

http://www.forbes.com/sites/narrativescience/

NLP today: Search/summarization



Oriental Flavor

54 reviews

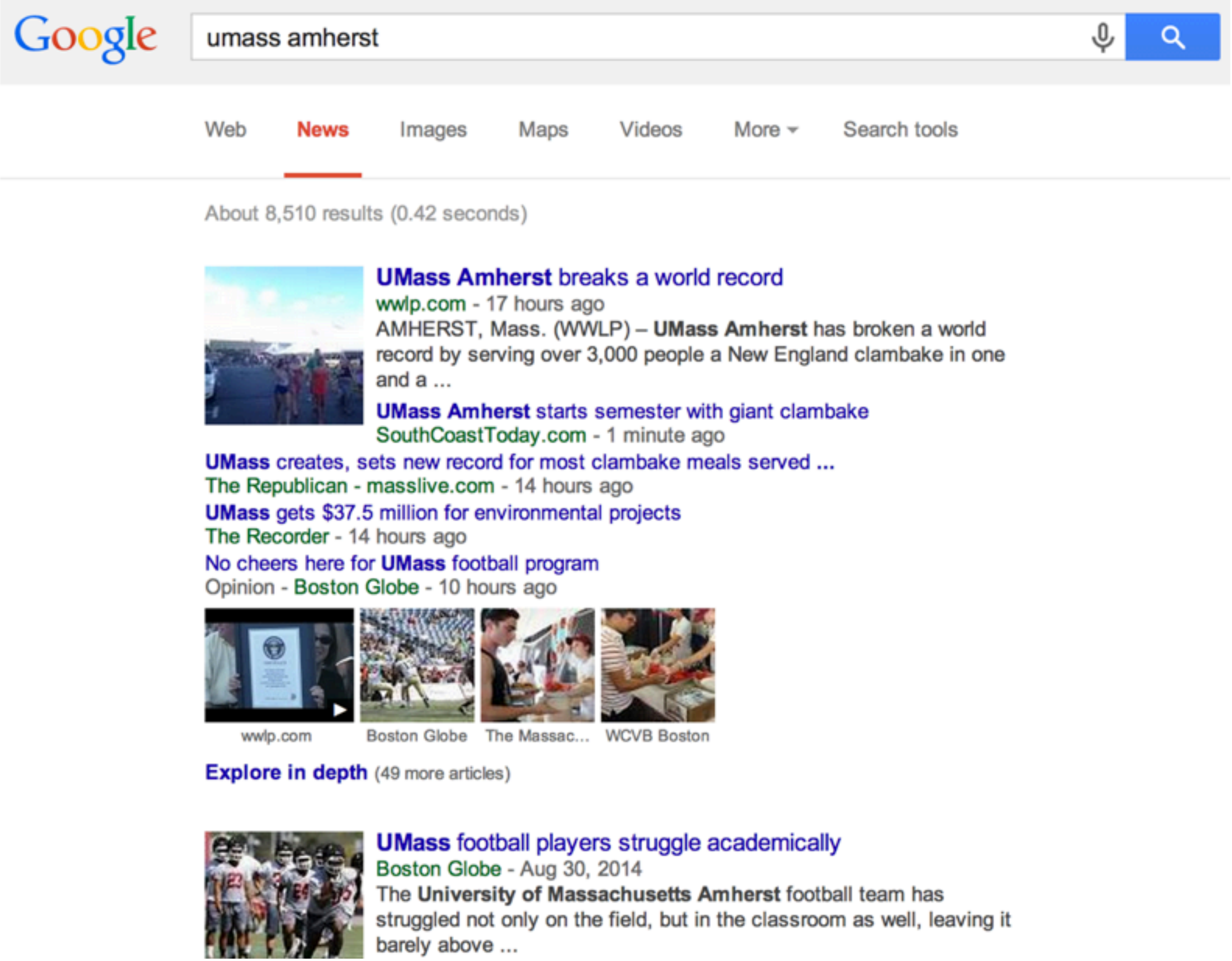
$$ · Szechuan, Dim Sum, Cantonese

"We had the **crystal shrimp dumplings** that is seen wowing all yelpers, and they were superb." in 5 reviews

"In addition to having the best pork buns in the area, they also have the best **scallion pancakes** and xiaolongbao." in 8 reviews

"Usually I eat at the large dim sum restaurants in Boston with the rolling **carts**, but this experience was as good as any." in 5 reviews

NLP today: Search/summarization



umass amherst

About 8,510 results (0.42 seconds)

UMass Amherst breaks a world record
wwlp.com - 17 hours ago
AMHERST, Mass. (WWLP) – **UMass Amherst** has broken a world record by serving over 3,000 people a New England clambake in one and a ...

UMass Amherst starts semester with giant clambake
SouthCoastToday.com - 1 minute ago

UMass creates, sets new record for most clambake meals served ...
The Republican - masslive.com - 14 hours ago

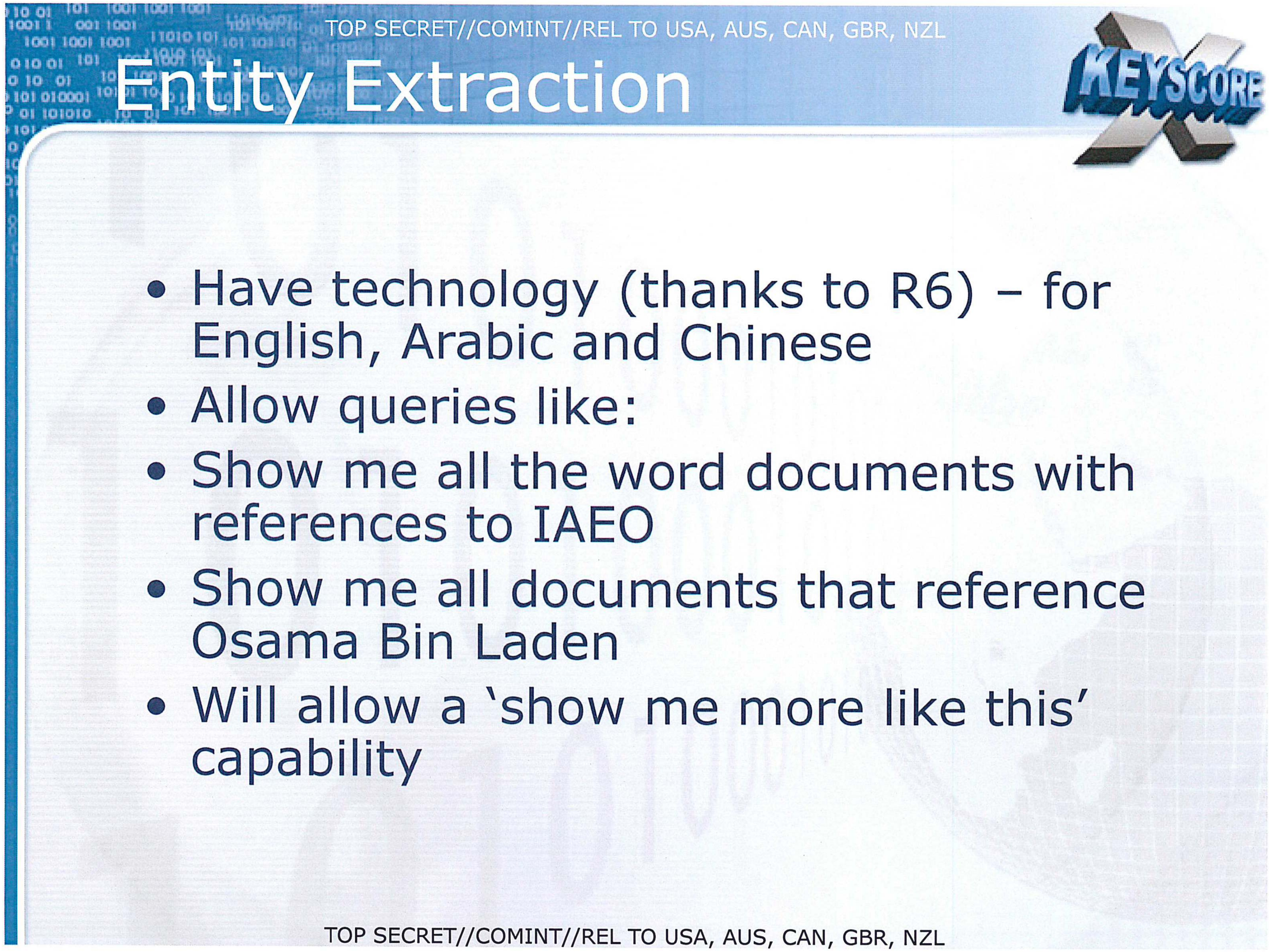
UMass gets $37.5 million for environmental projects
The Recorder - 14 hours ago

No cheers here for **UMass** football program
Opinion - Boston Globe - 10 hours ago

Explore in depth (49 more articles)

UMass football players struggle academically
Boston Globe - Aug 30, 2014
The **University of Massachusetts Amherst** football team has struggled not only on the field, but in the classroom as well, leaving it barely above ...

NLP today: Search/summarization



TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Entity Extraction

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
- Show me all the word documents with references to IAEO
- Show me all documents that reference Osama Bin Laden
- Will allow a 'show me more like this' capability

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

NSA slides from Snowden leaks

http://www.theguardian.com/world/interactive/2013/jul/31/nsa-xkeyscore-program-full-presentation

- HW0
- See you on Thursday