Natural Language Processing

Lecture #1

Outline

- Course Introduction
- Brief Introduction of NLP

Course Introduction

- Course Name: Natural Language Processing
  - Language Modelling
  - POSTagging
  - Parsing
- Course Code: IFP31963
- Credits : 3

- 14 weeks of each:
  - 3 x 50’ class
  - 3 x 50’ structured tasks
  - 3 x 50’ self study

Rules

- Attendance -> please be aware and follow institution rules
- Cheating and plagiarism -> E
- No additional assignments or exams to improve final grade
- 15’ late tolerance

Prerequisite

- Probabilistic and Statistic
- Artificial Intelligence
- Automata and Language Theory

Course Objectives

- Student is able to build and evaluate a Language Modelling and POSTagging based system
- Student is able to build and evaluate a syntactic parsing based system
- Student is able to build and evaluate a semantic based system (semantic vector and word sense disambiguation)
- Student is able to design, build and evaluate an NLP based system for a real world problem

Syllabus

- NLP Introduction
- Language Modelling
- Part of Speech (POS) Tagging, HMM
- Syntactic Parsing: Context Free Grammar, Syntactic Parsing, PCFG
- Semantic similarity: Semantic Vector, Word Sense Disambiguation
- Text Classification: Naïve Bayes, Logistic Regression

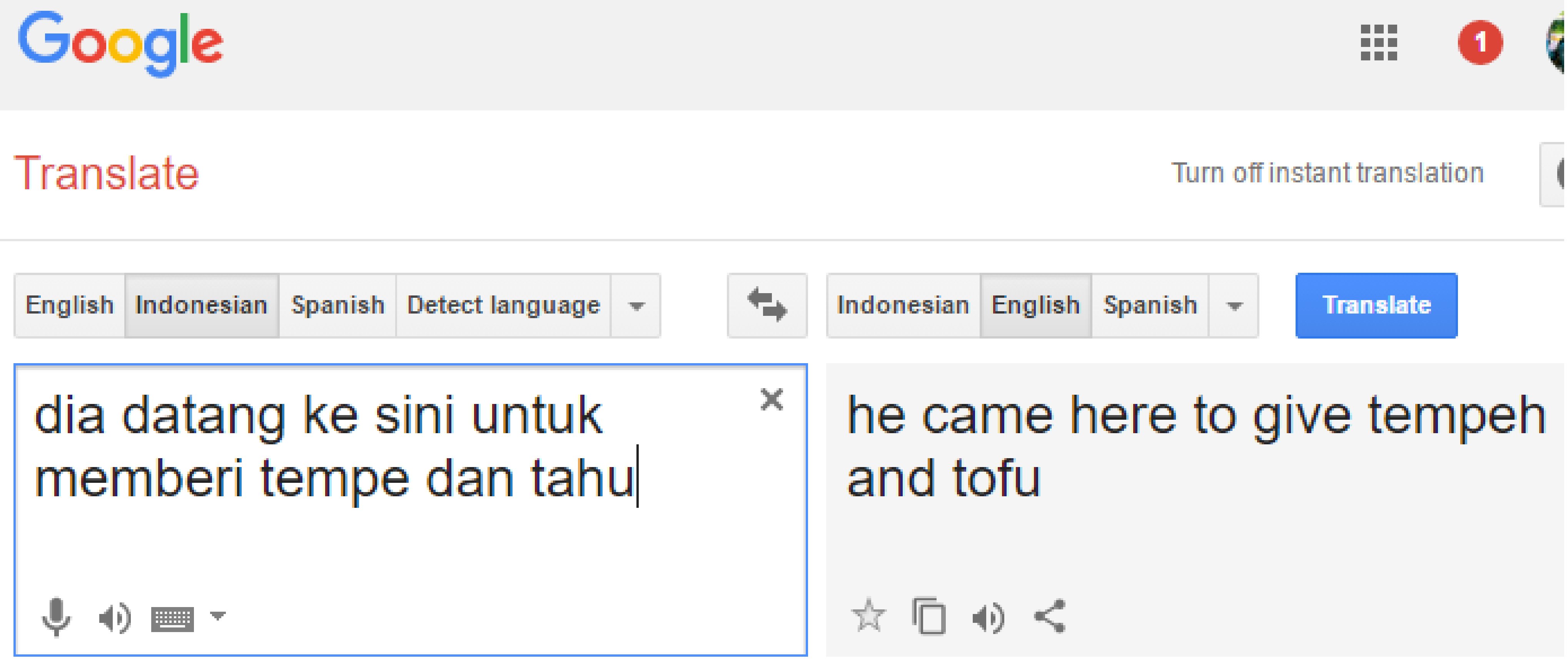
Weekly Assignments

References

- Jurafsky, David, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 2000. ISBN: 0130950696..
  - Second Edition: http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf
  - Third Edition (draft): https://web.stanford.edu/~jurafsky/slp3/
- http://www.nltk.org/book/

Introduction to NLP

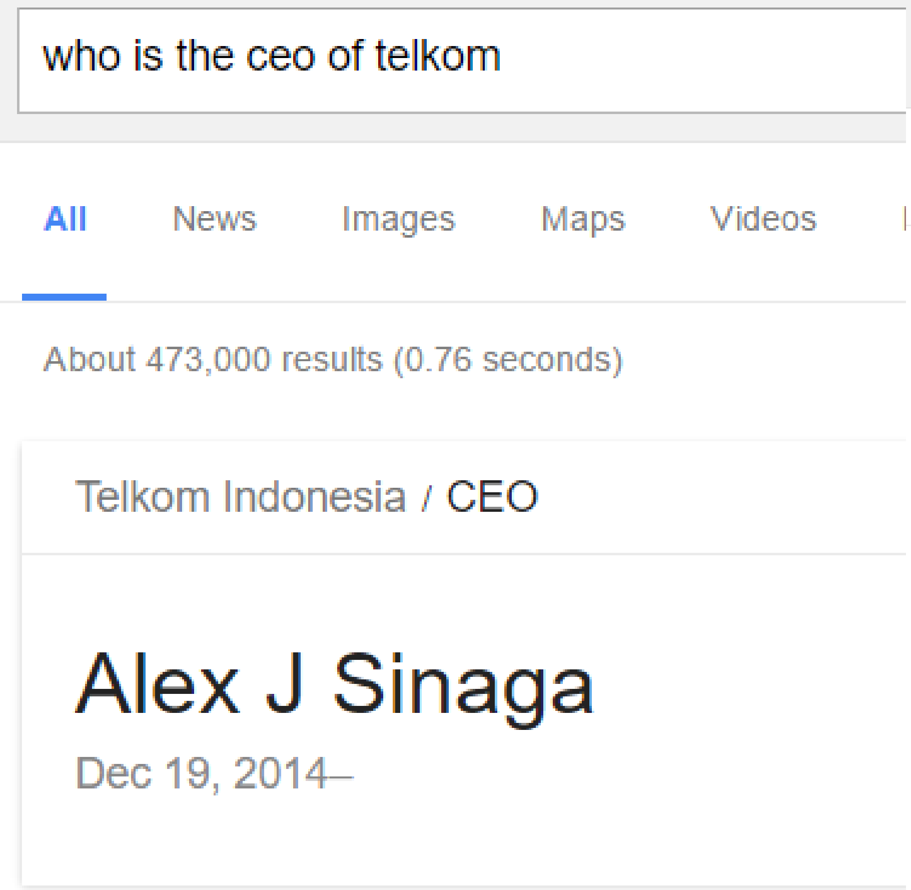
Machine Translation



dia datang ke sini untuk memberi tempe dan tahu

he came here to give tempeh and tofu

QA



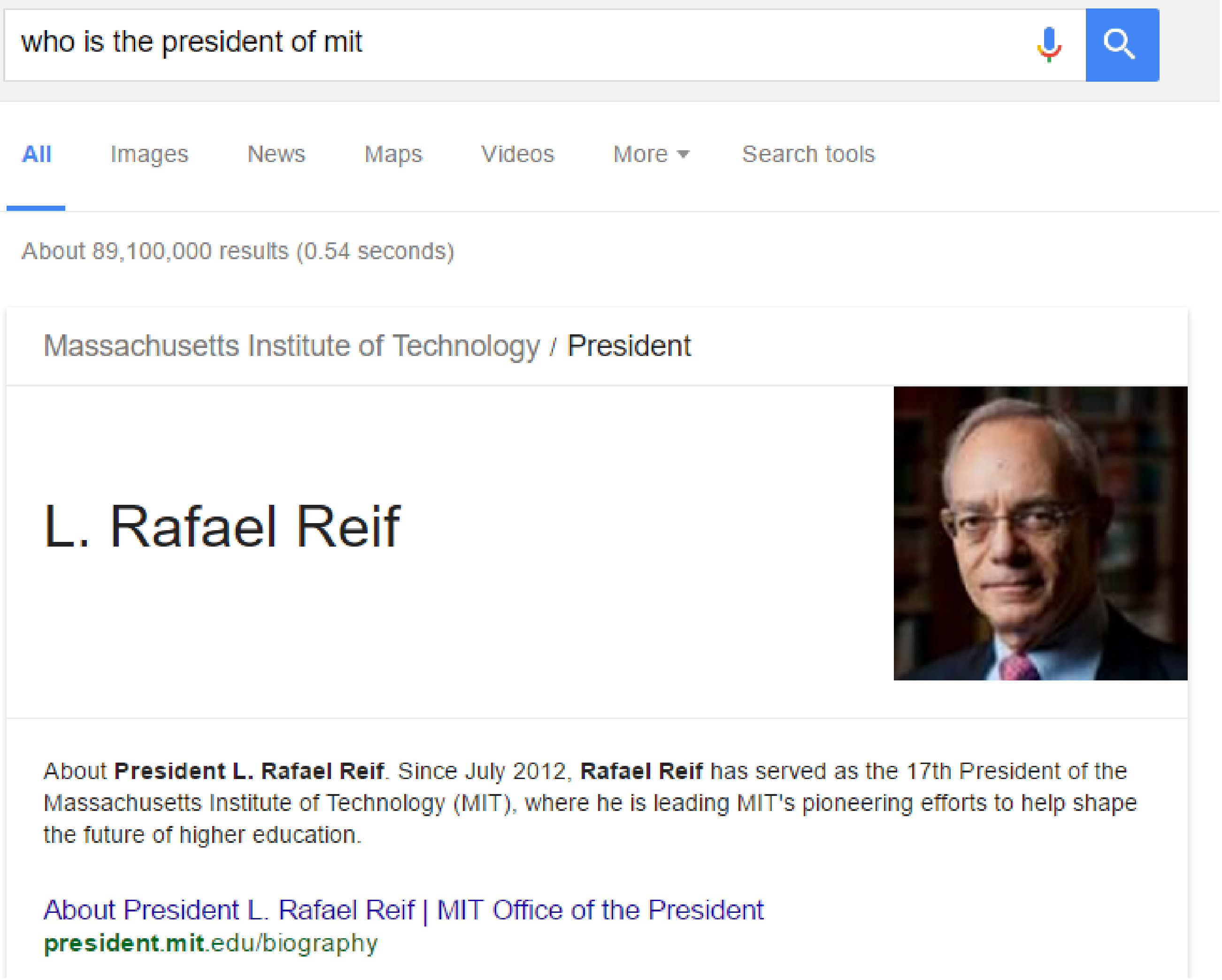
who is the ceo of telkom

About 473,000 results (0.76 seconds)

Telkom Indonesia / CEO

Alex J Sinaga

Dec 19, 2014–



who is the president of mit

About 89,100,000 results (0.54 seconds)

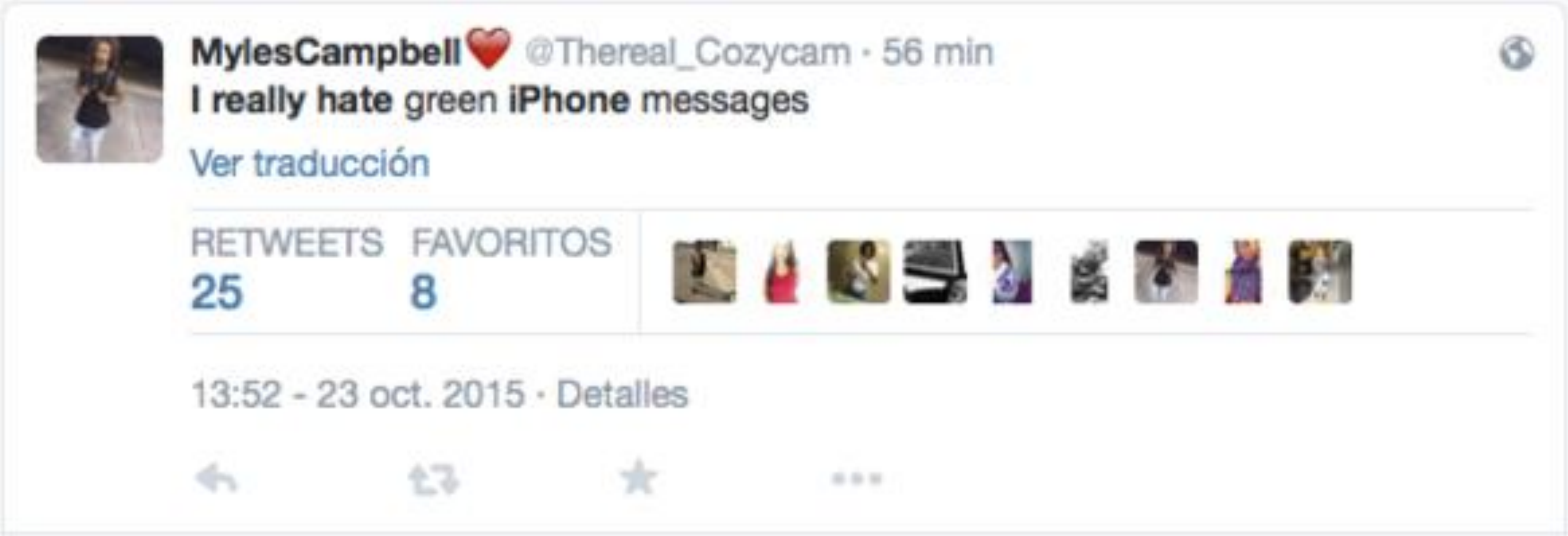
Massachusetts Institute of Technology / President

L. Rafael Reif

About **President L. Rafael Reif**. Since July 2012, **Rafael Reif** has served as the 17th President of the Massachusetts Institute of Technology (MIT), where he is leading MIT's pioneering efforts to help shape the future of higher education.

About President L. Rafael Reif | MIT Office of the President
president.**mit**.edu/biography

Sentiment Analysis



MylesCampbell❤️ @Thereal_Cozycam · 56 min

I really hate green iPhone messages

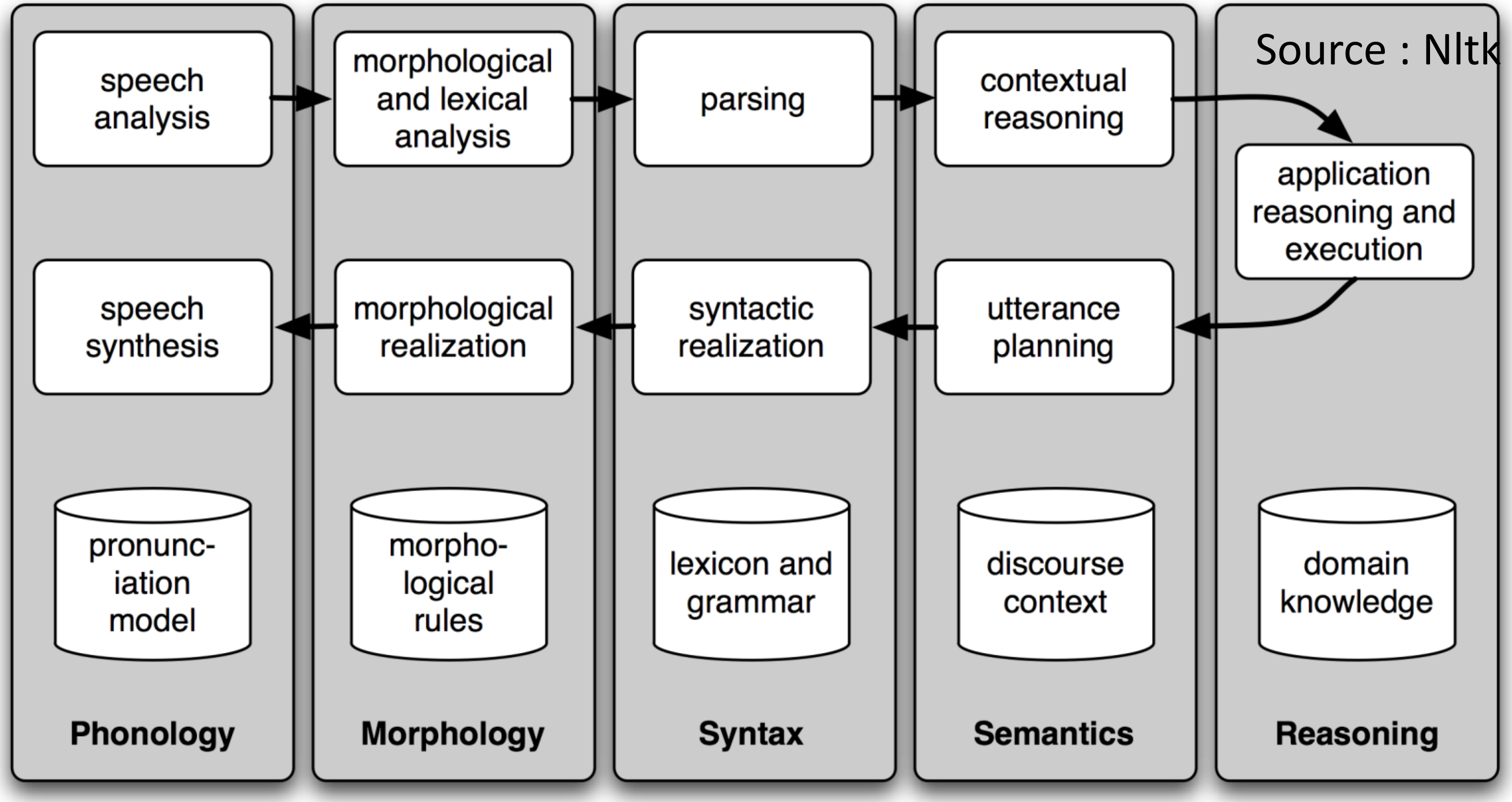
Ver traducción

RETWEETS 25 FAVORITOS 8

13:52 - 23 oct. 2015 · Detalles

Ultimate Dream: Conversational Agent JARVIS ?

Ultimate Dream: Conversational Agent JARVIS ?



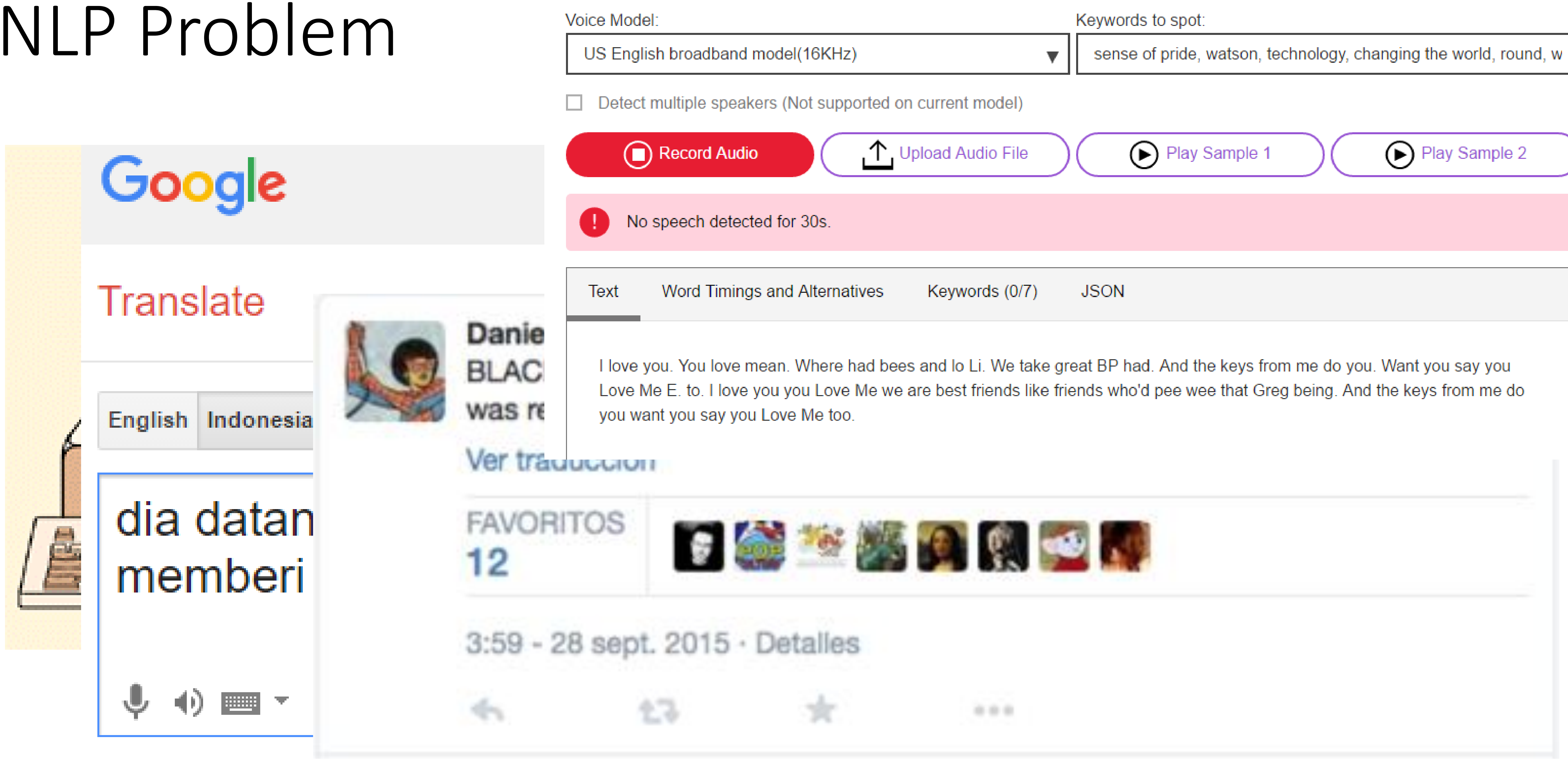
Source : Nltk book

Read Chapter 1, subchapter 1.1 – 1.5 SLP Book

Terminologies

- Phonetics and Phonology – The study of linguistic sounds.
- Morphology – The study of the meaningful components of words.
- Syntax – The study of the structural relationships between words.
- Semantics – The study of meaning.
- Pragmatics – The study of how language is used to accomplish goals.
- Discourse – The study of linguistic units larger than a single utterance

NLP Problem



dia datan
memberi

Danie
BLAC
was re

Ver traducción

FAVORITOS
12

3:59 - 28 sept. 2015 · Detalles

Voice Model: US English broadband model(16KHz)

Keywords to spot: sense of pride, watson, technology, changing the world, round, w

□ Detect multiple speakers (Not supported on current model)

No speech detected for 30s.

I love you. You love mean. Where had bees and lo Li. We take great BP had. And the keys from me do you. Want you say you Love Me E. to. I love you you Love Me we are best friends like friends who'd pee wee that Greg being. And the keys from me do you want you say you Love Me too.

Why is NLP Hard ?

AMBIGUITY AT ALL LEVEL OF ANALYSYS !!!

Phonetics and Phonology

- I Scream vs Ice cream

Morphology

- Unionized = union + ized vs un+ionized

Syntax

- Squad helps [dog bite victim] vs [Squad helps dog] bite victim

Semantics

- Jack invited Mary to the Halloween ball

Why is NLP Hard ?

AMBIGUITY AT ALL LEVEL OF ANALYSYS !!!

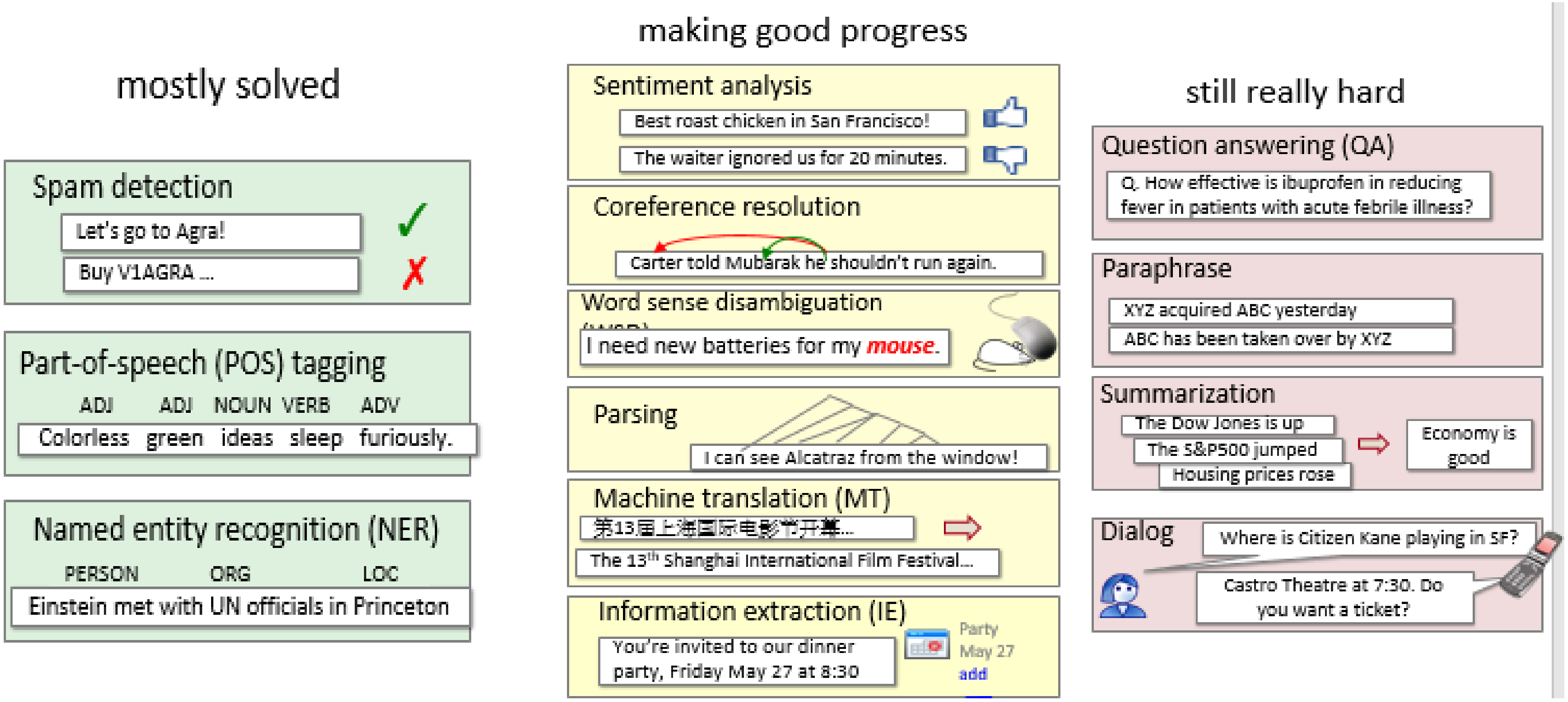
Discourse

- Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will be called Prodome Ltd

Pragmatics

- Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.
  “I just came from New York”
  - Would you like to go to New York today?
  - Would you like to go to Boston today?
  - Why do you seem so out of it?
  - Boy, you look tired.

Language Technologies (from Jurafsky’s slide)



mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON				ORG		LOC
Einstein	met	with	UN	officials	in	Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

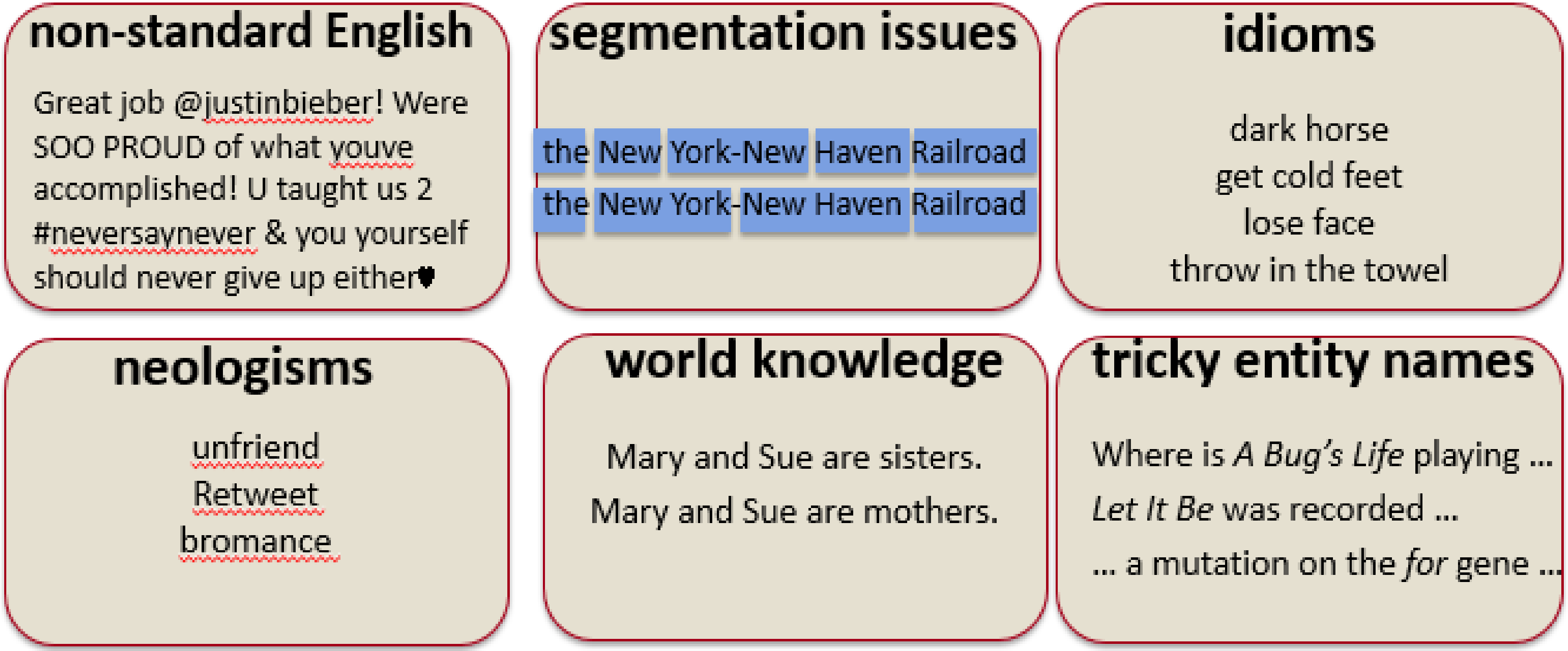
Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Why else is natural language understanding difficult? (from Jurafsky's slide)



non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing ...

Let It Be was recorded ...

... a mutation on the *for* gene ...

But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
  - Luckily, rough text features can often do half the job.

This Course ?

- General introduction to the filed of Natural Language Processing
- Learn techniques to overcome those ambiguities