Statistical NLP

CSE 256, Spring 2019

Lecture 1: Introduction

Ndapa Nakashole, UCSD

1 April 2019

Plan for today

1. What is Natural Language Processing?

2. A Sample of NLP applications

3. Why is language understanding difficult?

4. Examples of state-of-the-art NLP methods

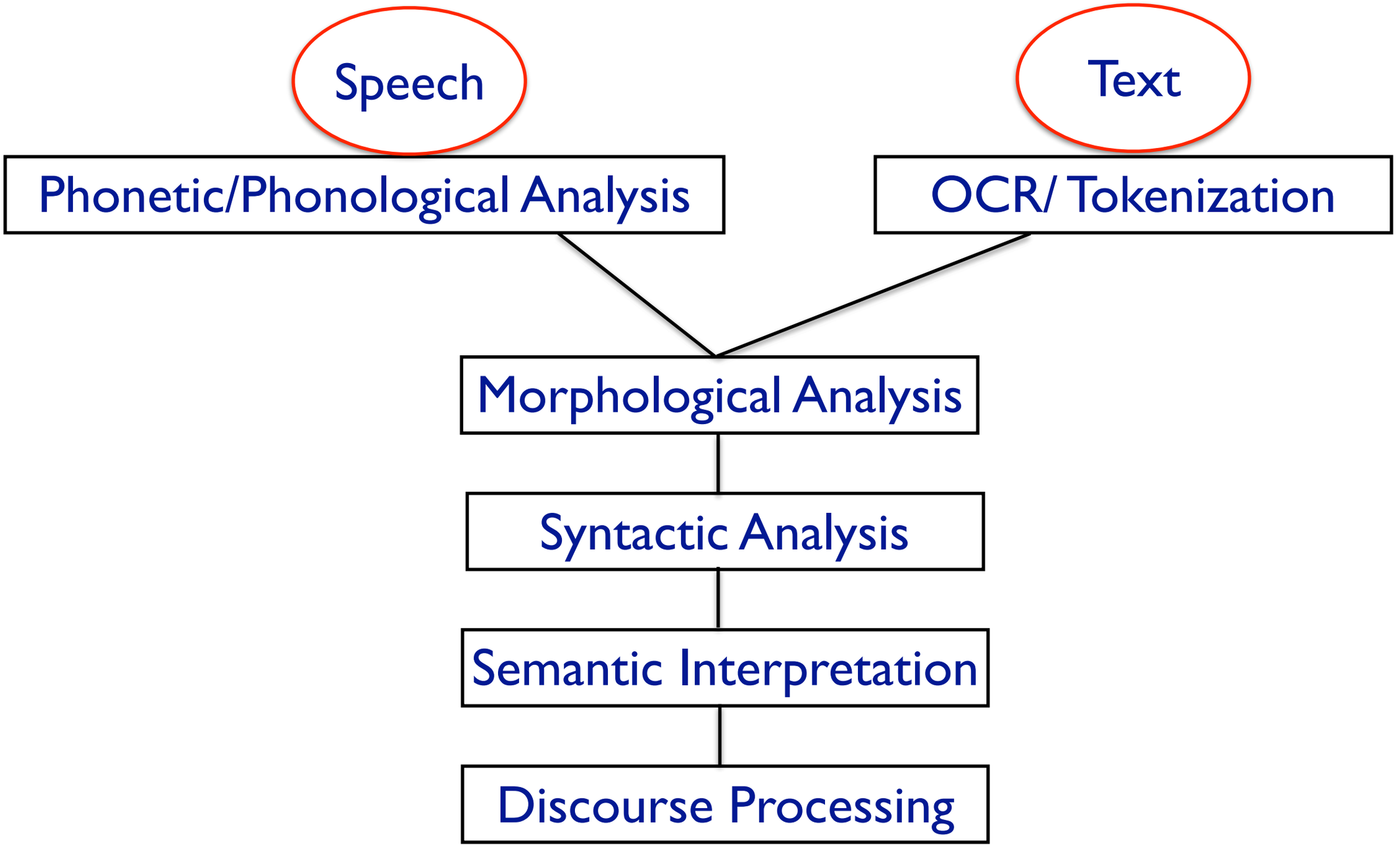
5. Course logistics

1. What is Natural Language Processing (NLP)?

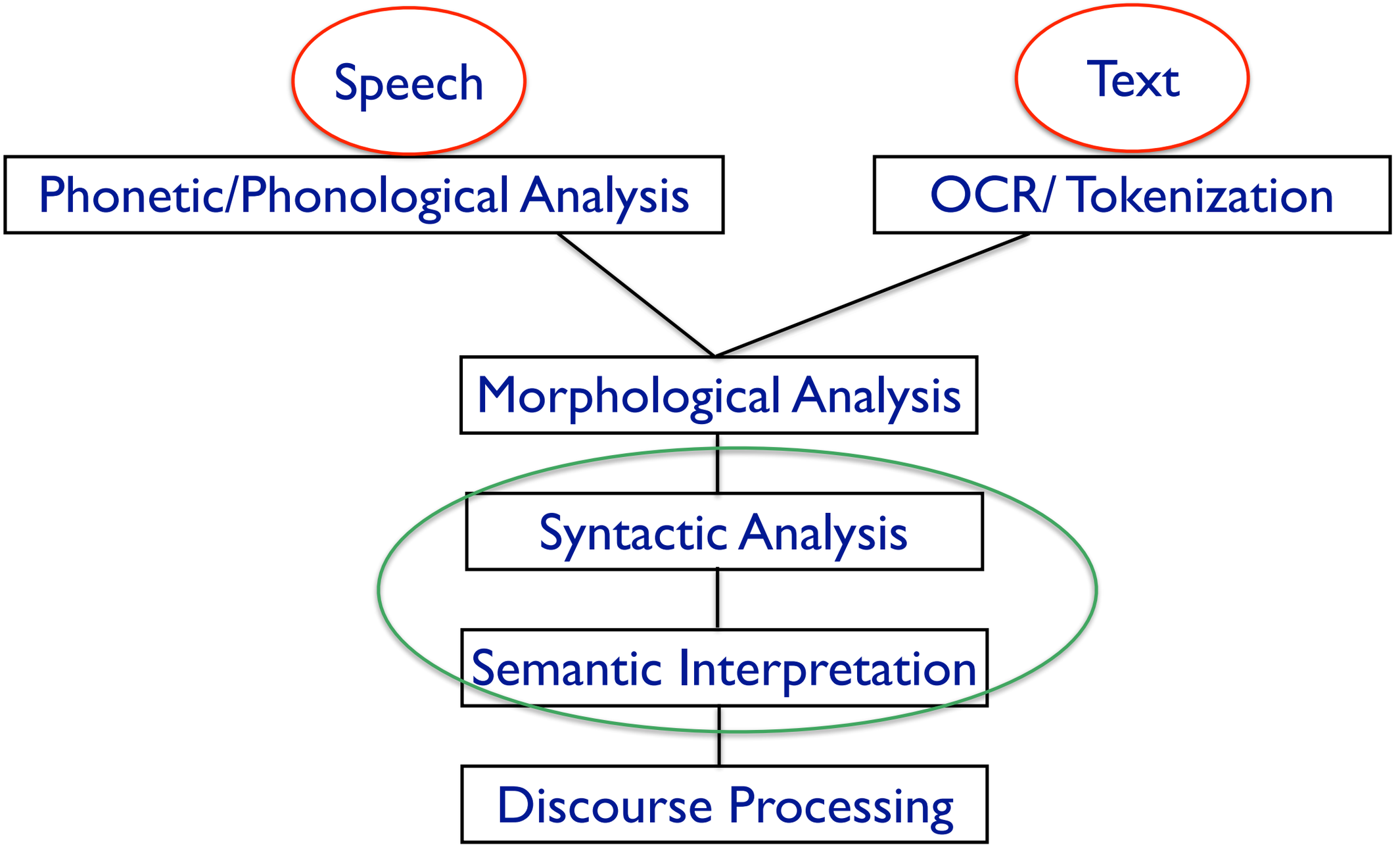
- Natural languages
  - Eg., English, Mandarin Chinese, Zulu, Spanish, …

- NLP develops algorithms for
  - **Analysis of language** (NL to some useful output): e.g., text classification, question answering, …
  - **Generation of language** (NL to NL; image to NL, etc): e.g., summarization, image captioning, MT
  - **Acquisition of a representation** (data to some representation): e.g., learning word embeddings

Levels of Language Analysis

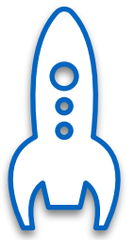


Levels of NLP



What is special about human language?

- A human language is a system **specifically constructed to convey the speaker/writer's meaning**
  - Not just an environmental signal, it's a deliberate communication

- A human language is a **discrete/symbolic/categorical signaling system**
  - car =  rocket = 
  - Thus we communicate with other people via symbols
  - With very minor exceptions for expressive signaling ("I loooove it." "Whoomppaaa")

- The large vocabulary, symbolic encoding of words creates a problem for machine learning — sparsity!

2. A Sample of NLP Applications

- Basic language processing tasks
  - spellcheckers
  - keyword search: finding synonyms, etc

Question Answering

- More than search
  - Requires precise answers than just documents

- QA Complexity
  - Can be easy: *Who is the Prime Minister of the UK?*
  - Can be harder: *How many capital cities are also the largest cities of their countries?*
  - Can be open ended: *How did the 2008 financial crisis happen?*

- State of the art
  - Success on factoid questions, even when text is not a perfect match

Question Answering : Watson

IBM Watson

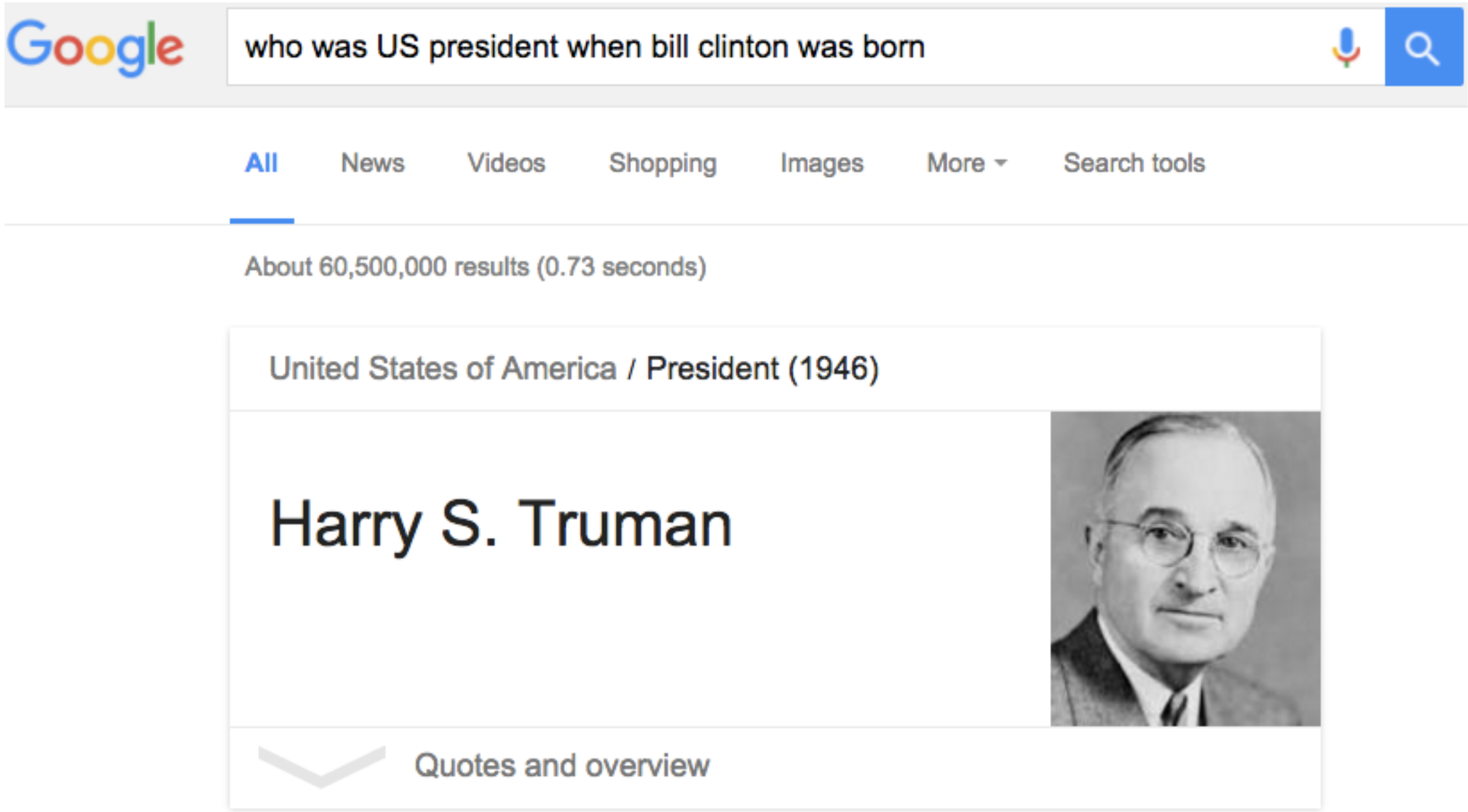
Ken Jennings

Brad Rutter

D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.

IBM Journal of R&D 56(3/4), 2012: This is Watson

Question Answering



Knowledge Discovery and Information Extraction

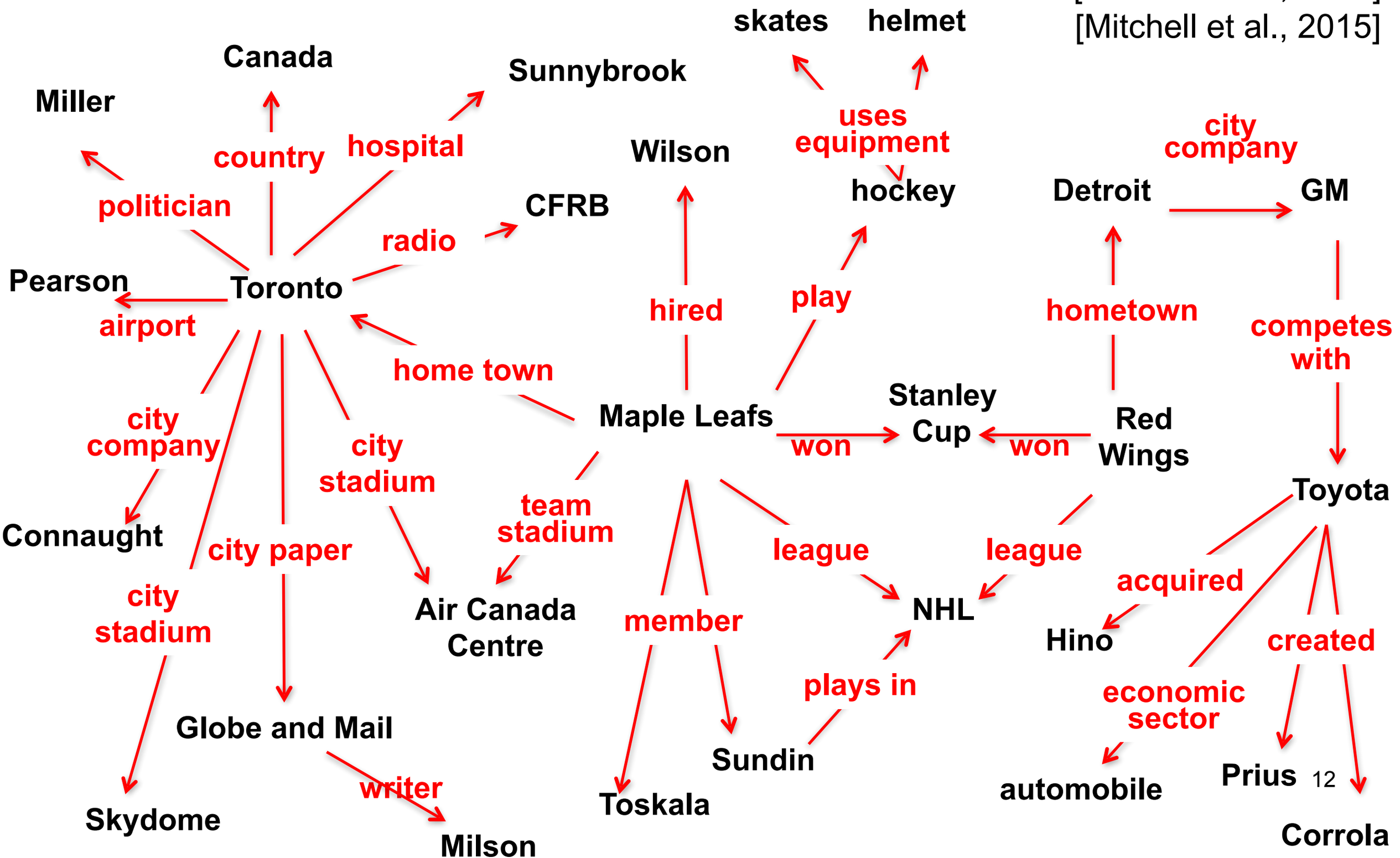
- Generation of knowledge graphs (KGs)
  - Collect and store world knowledge in a format that is suitable for machine inference and reasoning

Google
Knowledge Graph

Microsoft
Satori

NELL Knowledge Graph Fragment

[Carlson et al., 2010]
[Mitchell et al., 2015]



Names vs Entities

- Reference resolution: find mentions that refer to the same entity

Input: Twitter is losing the head of its Asia Pacific business, Aliza Knox, after she announced the end of her tenure with the company. —TechCrunch, April 2017

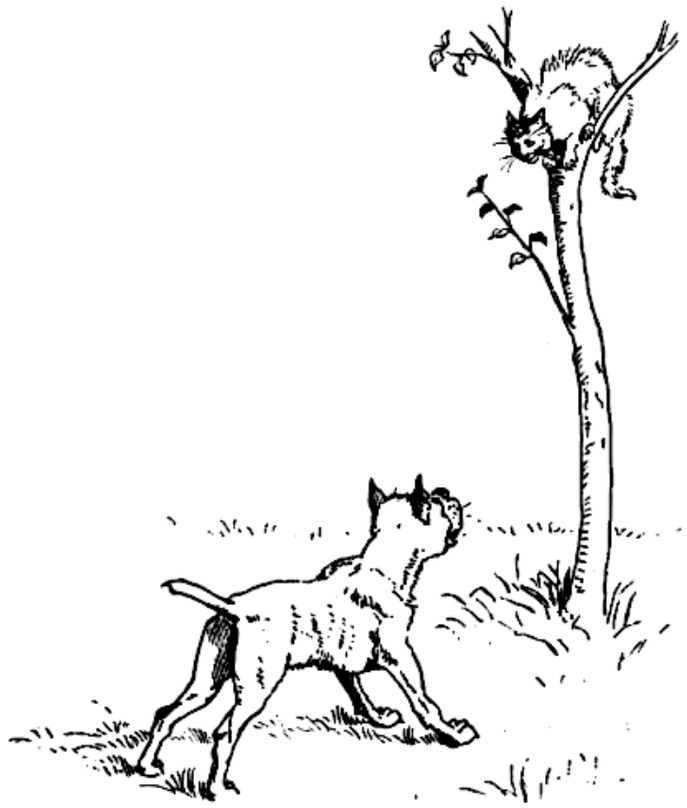
Correct:

Twitter	its	the company
Aliza Knox	she	her

Guess:

Twitter	its	
the company		
Aliza Knox	she	her

Anaphora



The **dog** *chased the* **cat**, *which ran up a tree.* **It** *waited at the top.*

The **dog** *chased the* **cat**, *which ran up a tree.* **It** *waited at the bottom.*

Relation Discovery

Category Pair	Frequent Instance Pairs	Text Contexts	Suggested Name
MusicInstrument Musician	sitar, George Harrison tenor sax, Stan Getz trombone, Tommy Dorsey vibes, Lionel Hampton	ARG1 master ARG2 ARG1 virtuoso ARG2 ARG1 legend ARG2 ARG2 plays ARG1	Master
Disease Disease	pinched nerve, herniated disk tennis elbow, tendonitis blepharospasm, dystonia	ARG1 is due to ARG2 ARG1 is caused by ARG2	IsDueTo
CellType Chemical	epithelial cells, surfactant neurons, serotonin mast cells, histomine	ARG1 that release ARG2 ARG2 releasing ARG1	ThatRelease
Mammals Plant	koala bears, eucalyptus sheep, grass goats, saplings	ARG1 eat ARG2 ARG2 eating ARG1	Eat
River City	Seine, Paris Nile, Cairo Tiber river, Rome	ARG1 in heart of ARG2 ARG1 which flows through ARG2	InHeartOf

Machine Translation

- Der Spiegel (German newspaper)

SPORT SIEG GEGEN NADAL

Roger Federer setzt seinen unfassbaren Siegeszug fort

Stand: 02.04.2017 | Lesedauer: 2 Minuten

Federer (r.) gewann das vierte Aufeinandertreffen gegen Nadal in Serie, der Spanier liegt mit 23:14-Siegen im Gesamtvergleich aber noch vorn

Quelle: AP

Der 35-jährige Roger Federer spielt eine beeindruckende Tennissaison. Nun hat er in Miami seinen langjährigen Rivalen Rafael Nadal erneut bezwungen und zum dritten Mal dieses ATP-Turnier gewonnen.

Machine Translation

Roger Federer continues his incredible victory

Last update: 02.04.2017 | Lesedauer: 2 minutes

Federer (r.) Won the fourth encounter against Nadal in a row, but the Spaniard is still ahead with a 23: 14 victory in the overall comparison

Source: AP

The 35-year-old Roger Federer plays an impressive tennissaison. Now he has defeated his longtime rival Rafael Nadal in Miami and won this ATP tournament for the third time.

2 comments

display

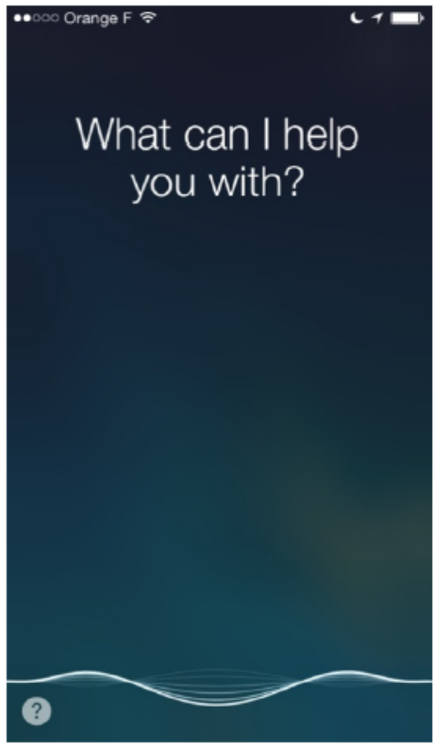
oger Federer won the eternal duel

- Chrome in-browser translation

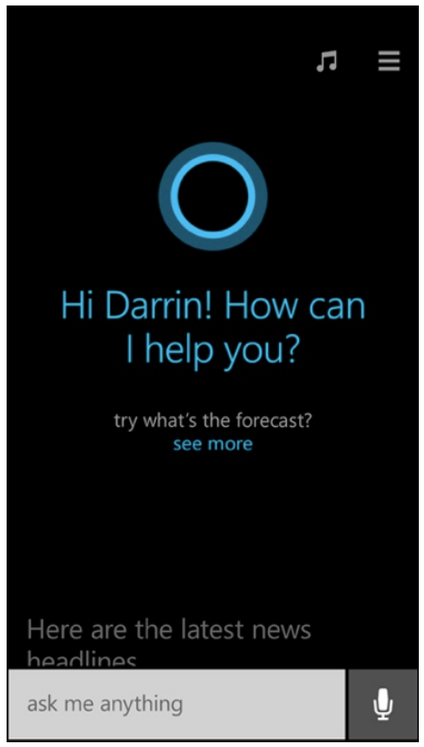
End-to-End Example

- Personal assistants contain
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

Apple
Siri



Microsoft
Cortona



NLP in industry is taking off

- Search (written and spoken)
- Online advertisement matching
- Automated/assisted translation
- Sentiment analysis for marketing or finance/trading
- Speech recognition
- Chatbots / Dialog agents
  - Automating customer support
  - Controlling devices
  - Ordering goods
  - …

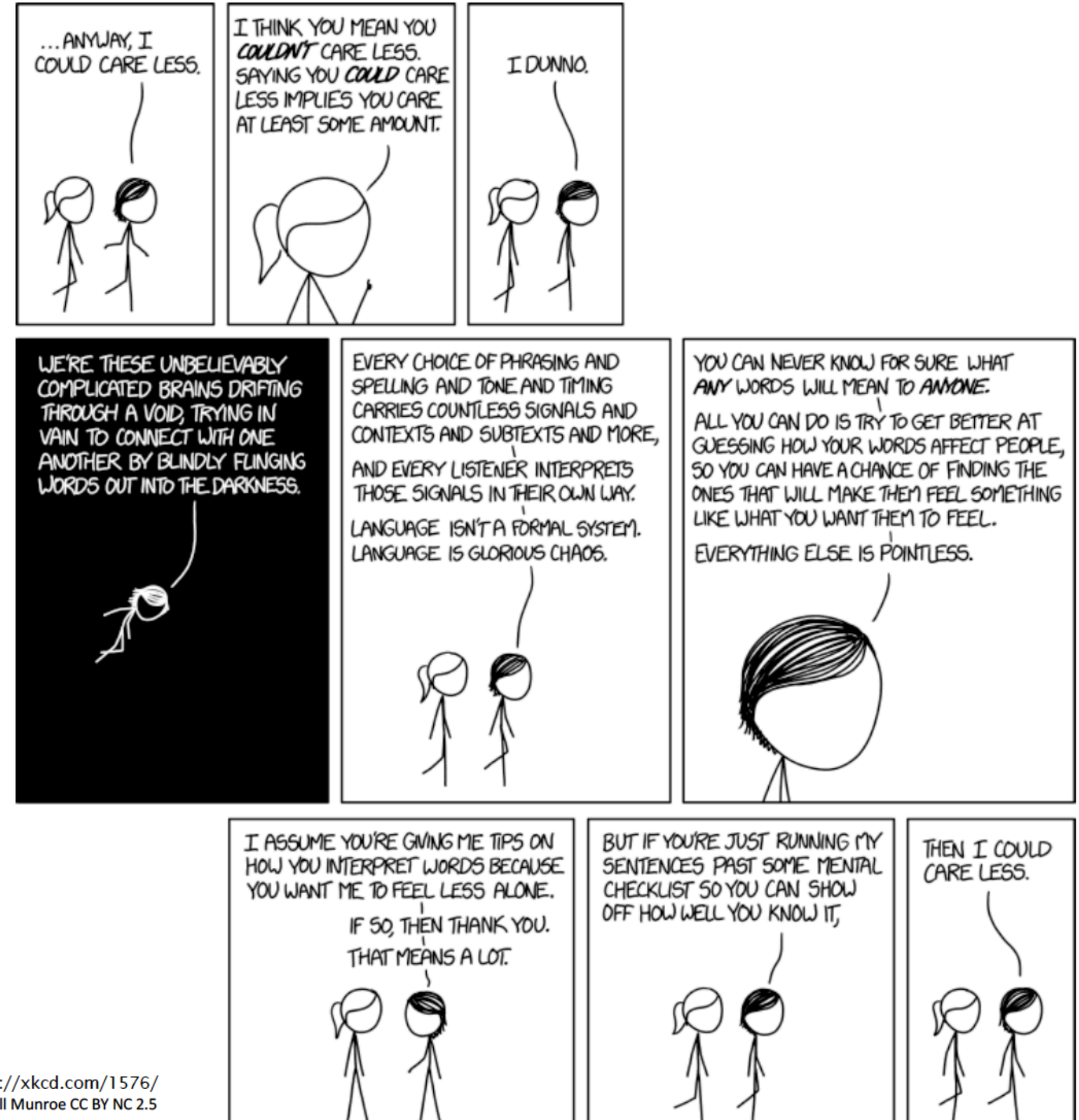
3. Why is language understanding difficult?

- Human languages are **ambiguous** (unlike programming and other formal languages)

- **Richness**: any meaning may be expressed many ways

- Human language interpretation depends on **real world, common sense, and contextual knowledge**

- Linguistic **diversity** across languages, dialects, genres, styles, …

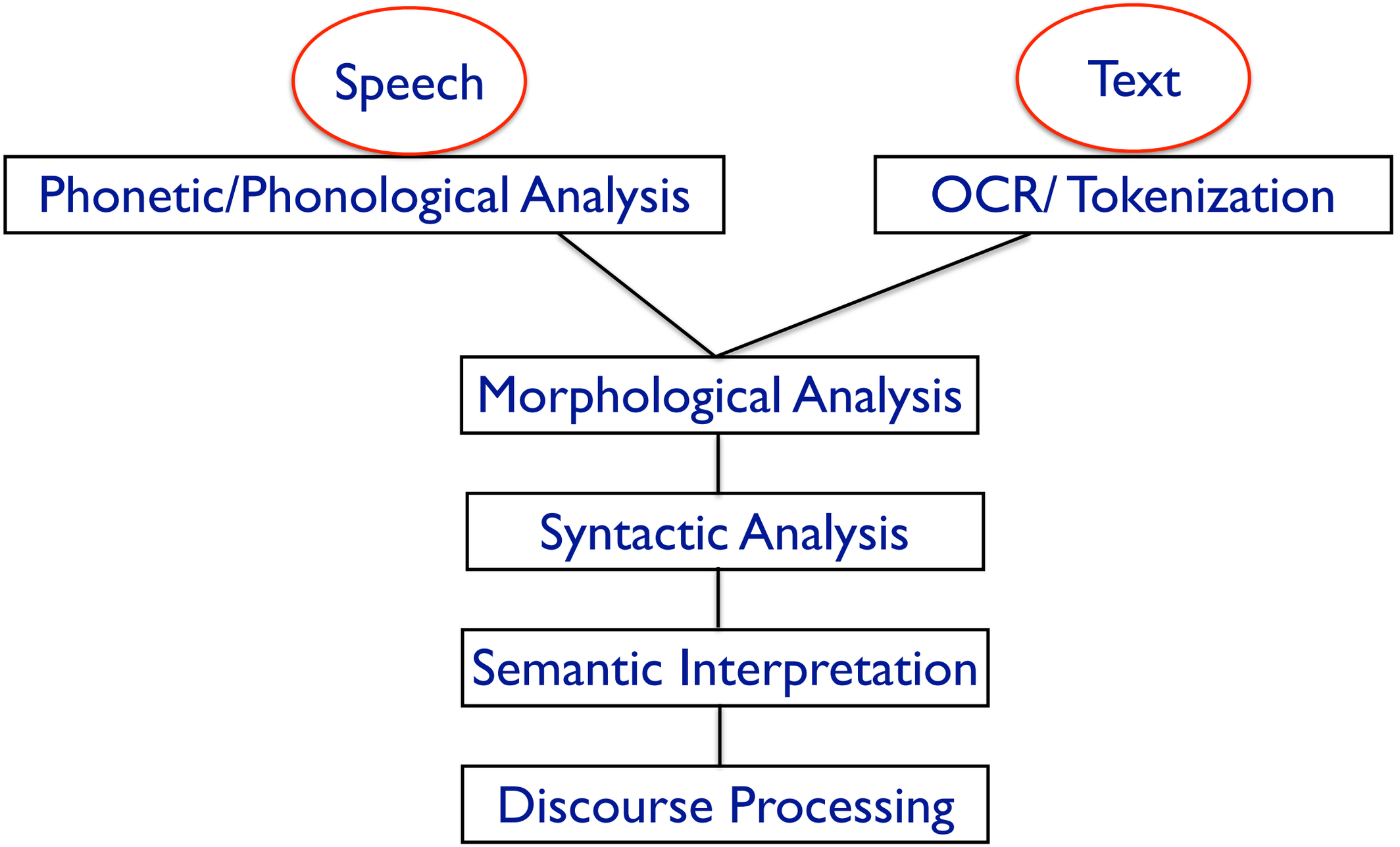


YOU CAN NEVER KNOW FOR SURE WHAT *ANY* WORDS WILL MEAN TO *ANYONE*.

ALL YOU CAN DO IS TRY TO GET BETTER AT GUESSING HOW YOUR WORDS AFFECT PEOPLE, SO YOU CAN HAVE A CHANCE OF FINDING THE ONES THAT WILL MAKE THEM FEEL SOMETHING LIKE WHAT YOU WANT THEM TO FEEL.

EVERYTHING ELSE IS POINTLESS.

Ambiguity is at all Levels of NLP



Ambiguity

- Every level of linguistic analysis comes with its own ambiguities

- Syntax
  - For part-of-speech tagging: multiple POS tag *(bark - noun, bark- verb)*
  - For parsing: multiple parse trees

- Semantics
  - Synonymy: different words same meaning:
  - Polysemy: same word different meanings
  - Multiword expressions: *make a decision*, take out, make up, …

Our focus: Syntax & Semantics

- Syntax
  - What is grammatical
- Semantics
  - What does it mean

- Programming languages analogy
  - Syntax: no compiler errors
  - Semantics: no implementation bugs

Stanford Parser

Please enter a sentence to be parsed:

The boy wants to go to San Diego

Language: English Sample Sentence Parse

Your query

The boy wants to go to San Diego

Tagging

```
The/DT  boy/NN  wants/VBZ  to/TO  go/VB  to/TO  San/NNP  Diego/NNP
```

Syntax annotations (11): phrase structure

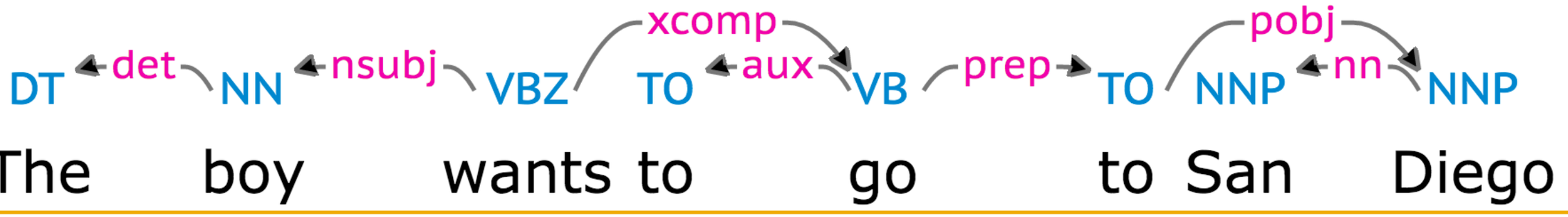
Parse

```
(ROOT
  (S
    (NP (DT The) (NN boy))
    (VP (VBZ wants)
      (S
        (VP (TO to)
          (VP (VB go)
            (PP (TO to)
              (NP (NNP San) (NNP Diego)))))))))
```

Universal dependencies

```
det(boy-2, The-1)
nsubj(wants-3, boy-2)
root(ROOT-0, wants-3)
mark(go-5, to-4)
xcomp(wants-3, go-5)
case(Diego-8, to-6)
compound(Diego-8, San-7)
nmod(go-5, Diego-8)
```

Syntax annotations (111): dependency structure



- Basic dependency representation forms a tree
  - Exactly one word is the head of the sentence (wants)
  - All other words depend on another word in the sentence

- Dependency relations
  - nsubj: subject (nominal)
  - pobj: object (of a preposition)

Syntax uses

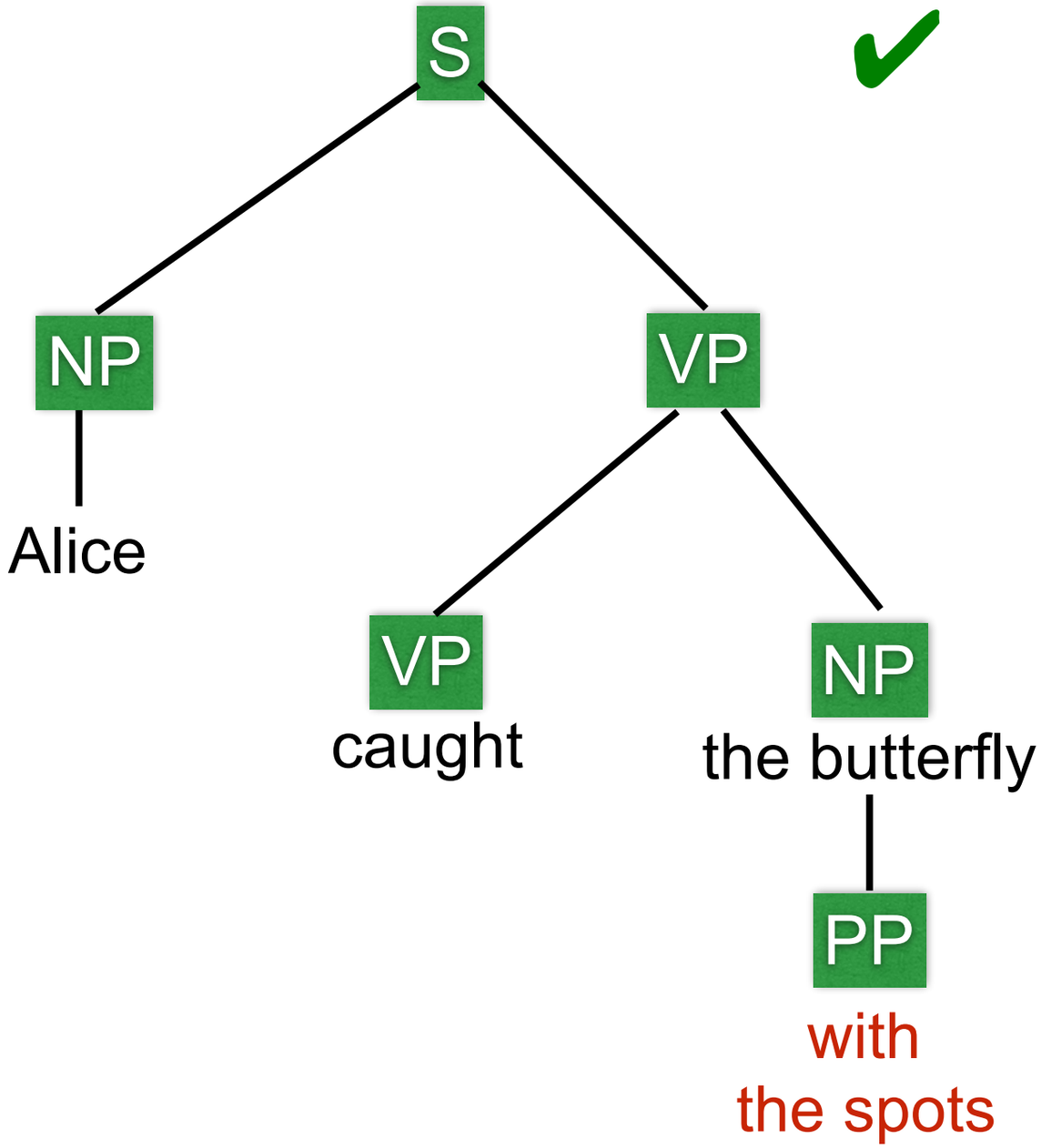
- Parts of speech:
  - Named entity recognition, relation extraction, MT, …

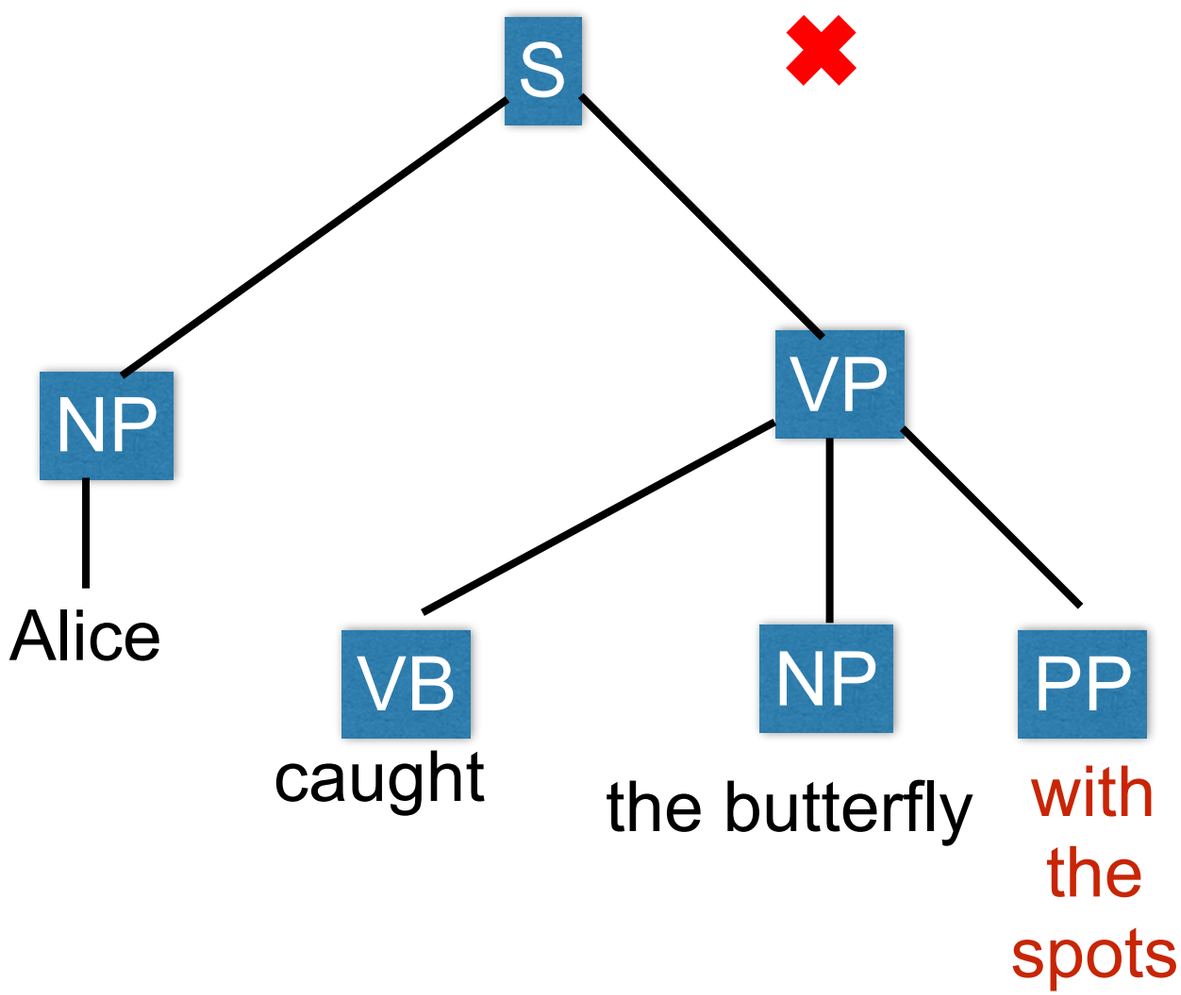
- Dependency relations
  - Relation extraction, sentiment analysis
  - …

Syntax ambiguity

Prepositional Attachment Ambiguity

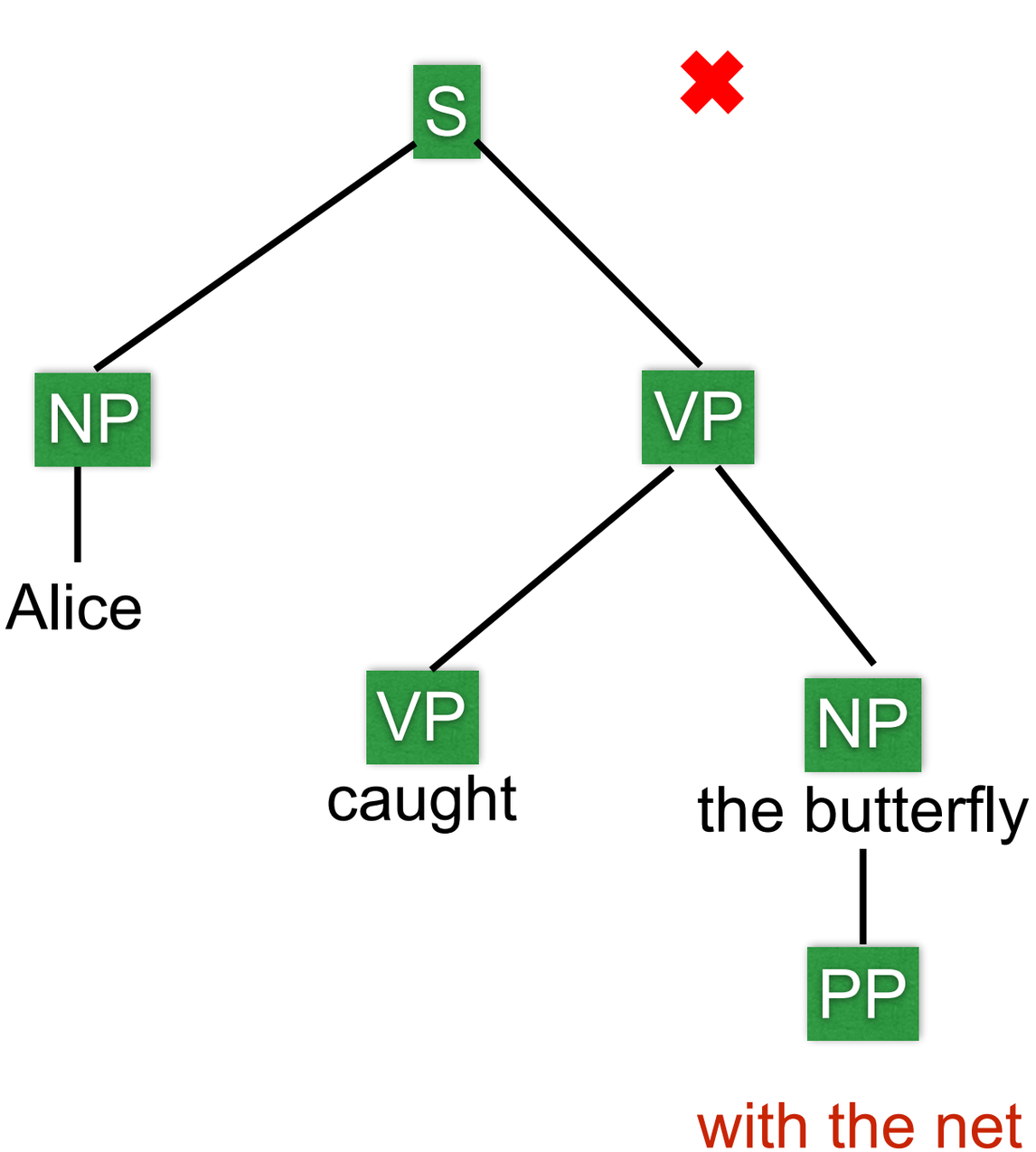
- s1) Alice caught the butterfly with the spots

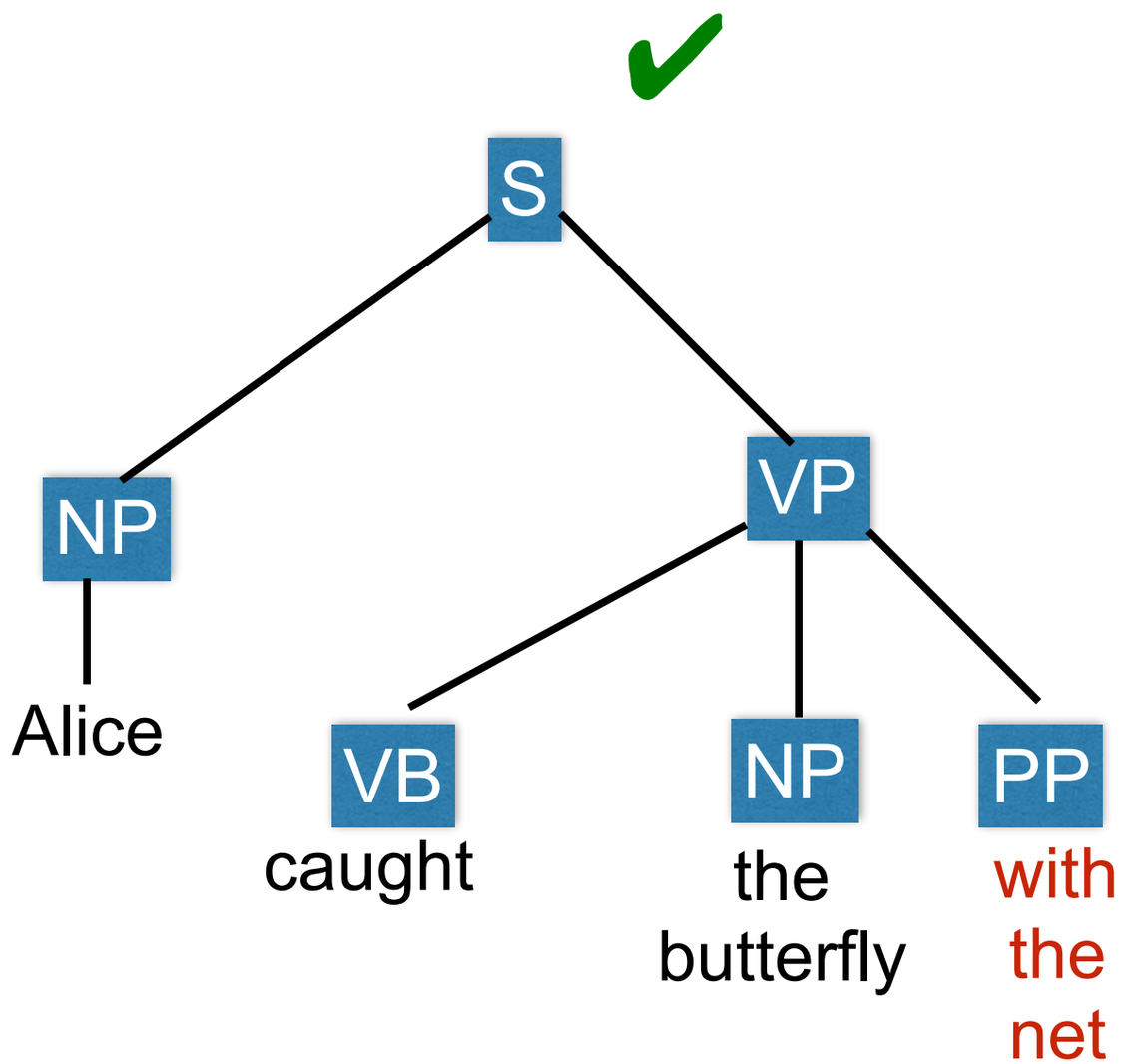




Prepositional Attachment Ambiguity

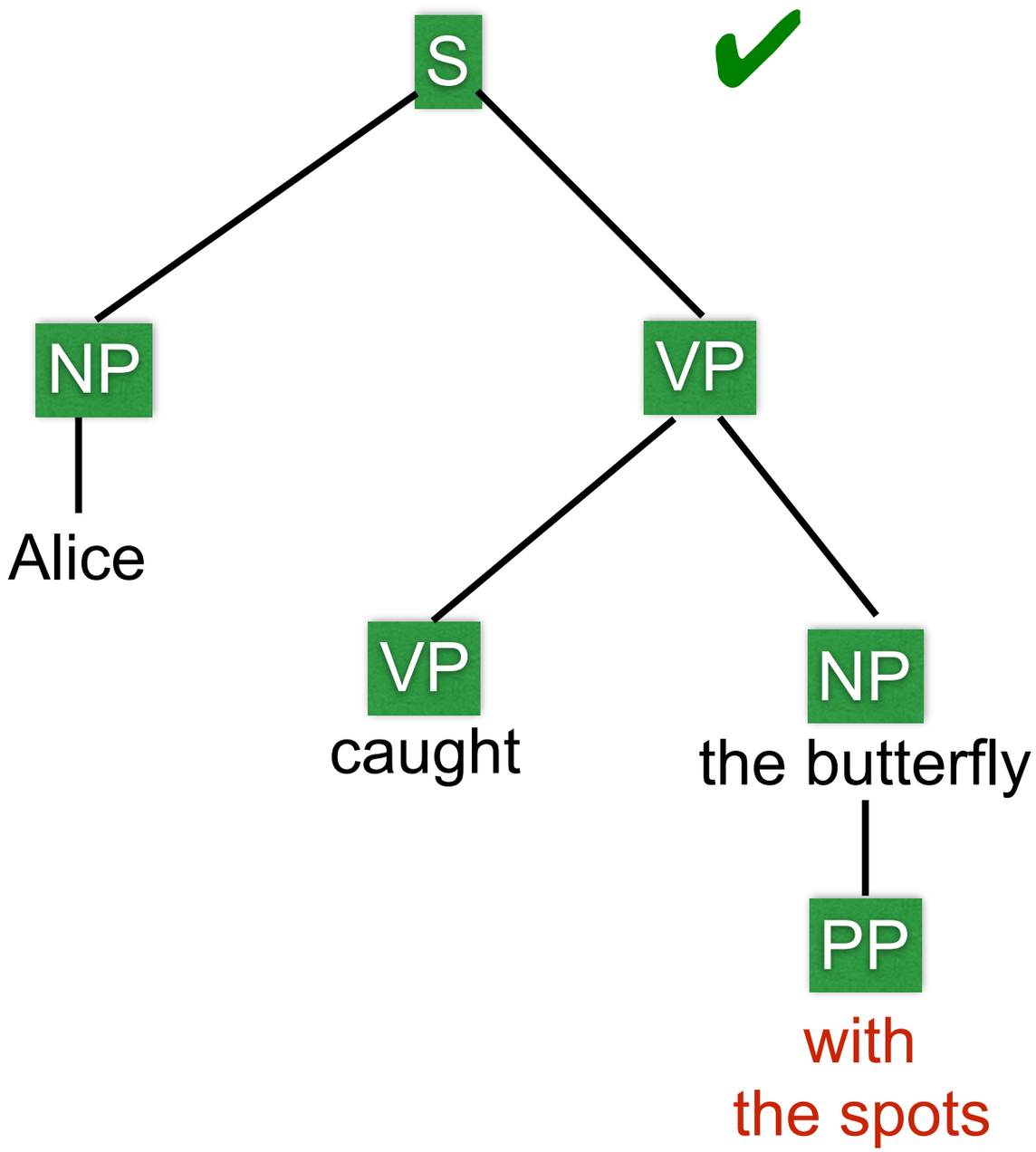
- s2) Alice caught the butterfly with the net

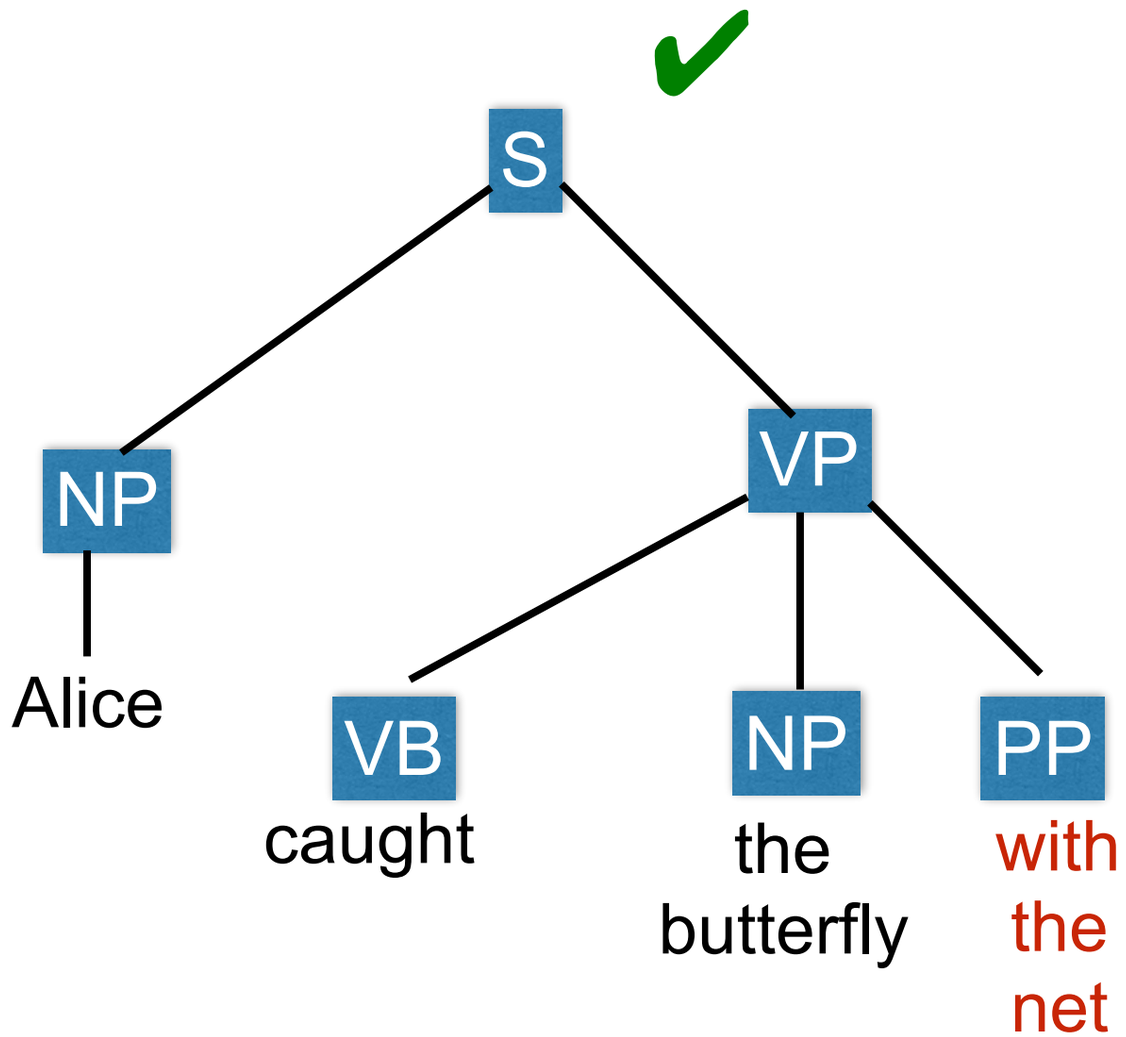




Prepositional Attachment Ambiguity

- Is Syntax enough to disambiguate?





Semantic ambiguity

- Even correct tree syntactic analyses don't fully nail down meaning.
  - John's boss said he was doing better

- Real newspaper headlines/tweets
  - The pope's baby steps on gays
  - Stolen painting found by tree
  - Kids make nutritious snacks
  - Local high school dropouts cut in half

Factors Changing the NLP Landscape

(Hirschberg and Manning, Science 2015)

- Increases in computing power
- The rise of the web, then the social web
- Advances in machine learning
- Advances in understanding of language in social context

4. Examples of State-of-the-art NLP methods

- Deep NLP = Deep Learning + NLP
  - Combine ideas and goals of NLP with using representation learning and deep learning methods to solve them

- Several big improvements in recent years in NLP with different
  - Levels: speech, words, syntax, semantics
  - Tools: parts-of-speech, entities, parsing
  - Applications: machine translation, sentiment analysis, dialogue agents, question answering

Word meaning as word vector

chair (-0.37, -0.23, 0.33, 0.38, -0.02, -0.37)
on (-0.21, -0.11, -0.10, 0.07, 0.37, 0.15)
dog (0.26, 0.25, -0.39, -0.07, 0.13, -0.17)
...
...
the (-0.43, -0.37, -0.12, 0.13, -0.11, 0.34)
...
...
mouth (-0.32, 0.43, -0.14, 0.50, -0.13, -0.42)
...
...
gone (0.06, -0.21, -0.38, -0.28, -0.16, -0.44)
...

Word similarities

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

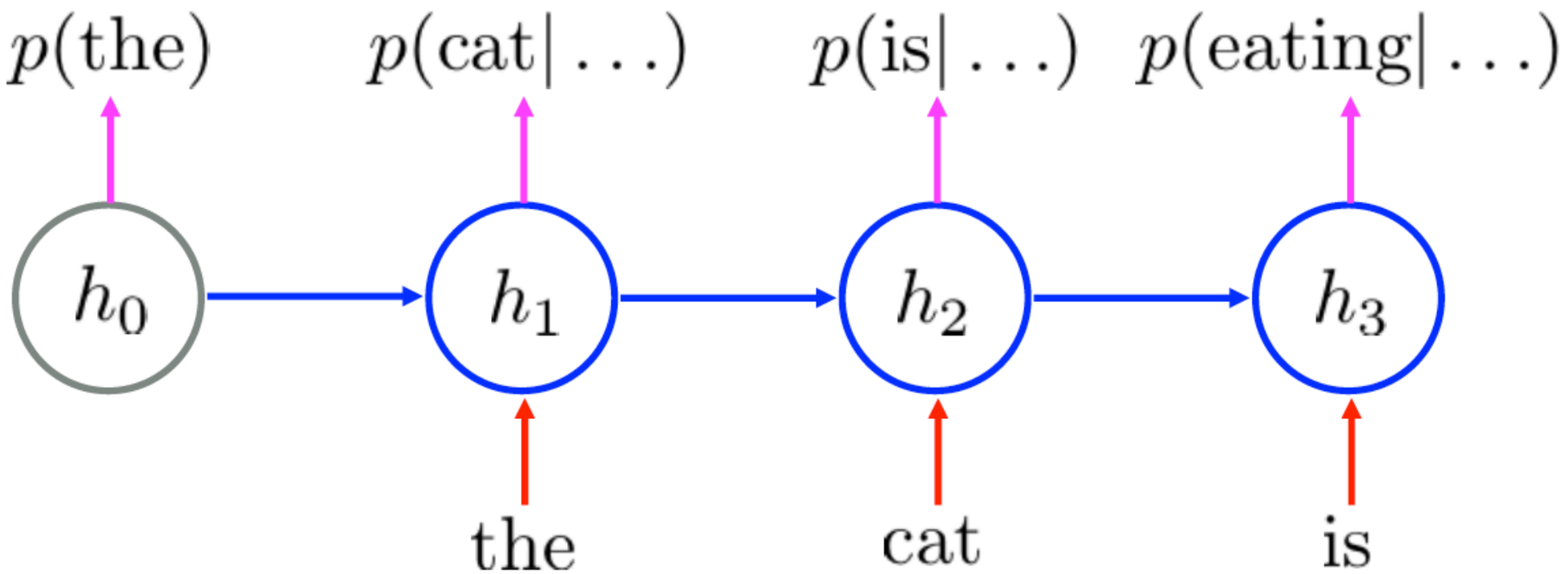
rana

eleutherodactylus

Dialogue agents / Response Generation

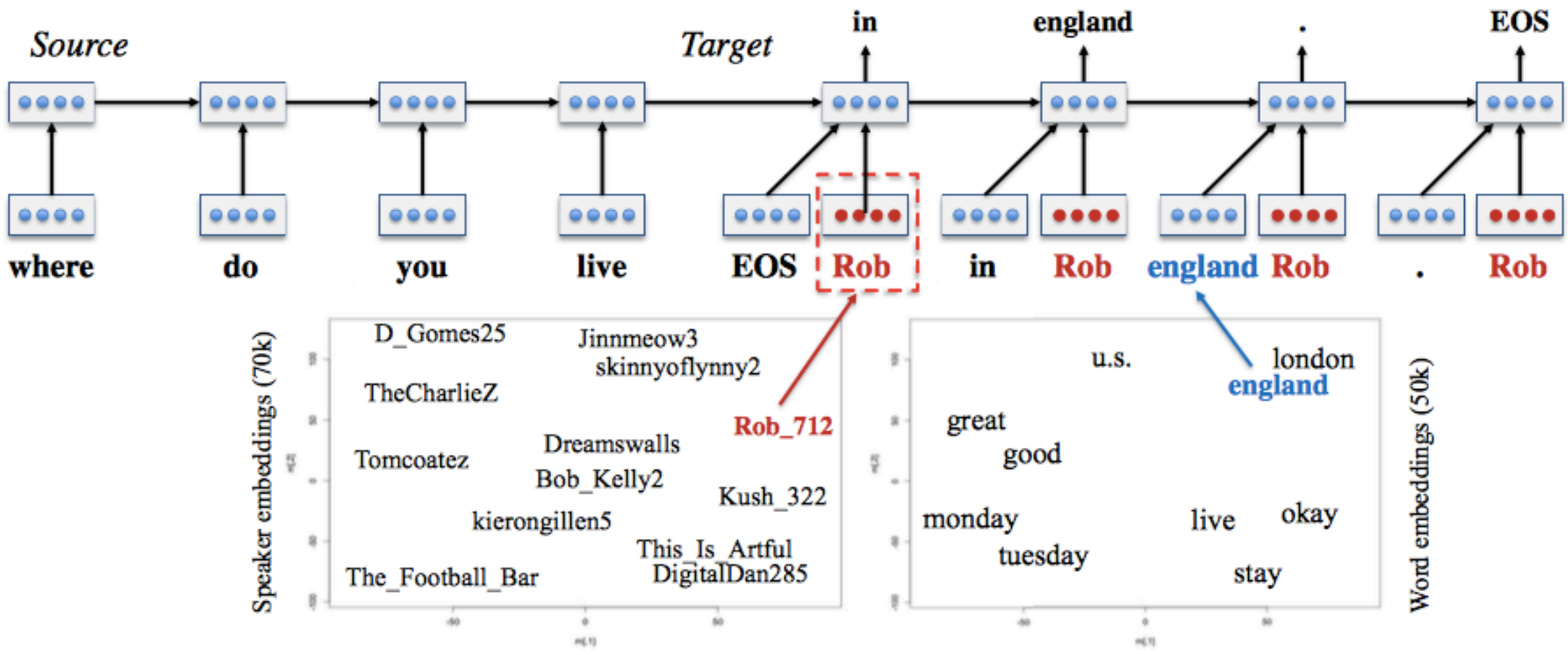
- A simple, successful example is the auto-replies available in the Google Inbox app

- An application of the powerful, general technique of **Neural Language Models**, which are an instance of Recurrent Neural Networks



Persona-based Neural Dialog Model (Li et al 2017)

- Model each speaker in embedding space



- Also model who the speaker is speaking to in speaker-addressee model

Instructable agents

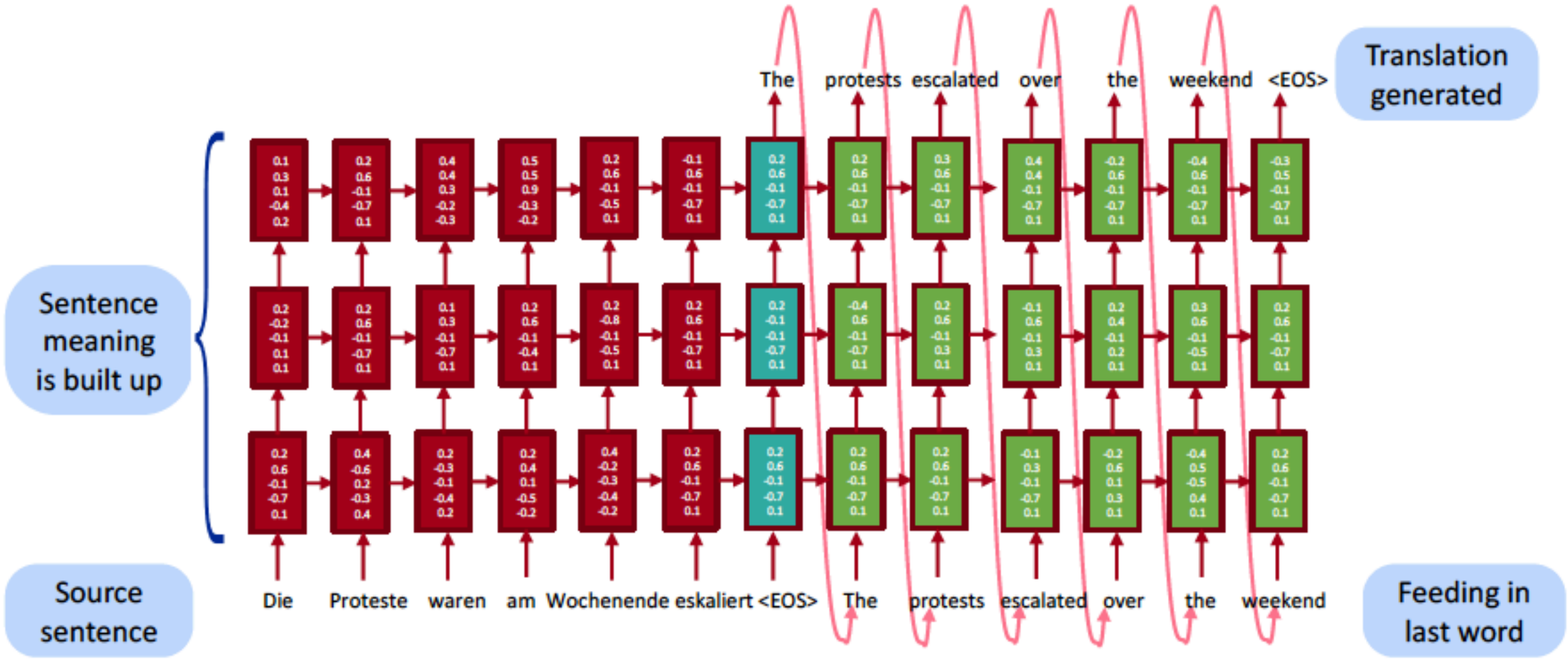
- Instructable agents (Azaria et al AAAI 2016)
  - Allow user to define a new command and a sequence of actions to perform it

- I'm stuck in traffic and will be late." The assistant **may not know what to do**, so the user may then explain **"first, use GPS to estimate my time of arrival, then see who I am meeting and send them an email indicating that I'll be late**." The assistant now understands how to handle similar situations in the future.

- Teaching the system that "drop a note to Bill" has the same meaning as "send an email to Bill".

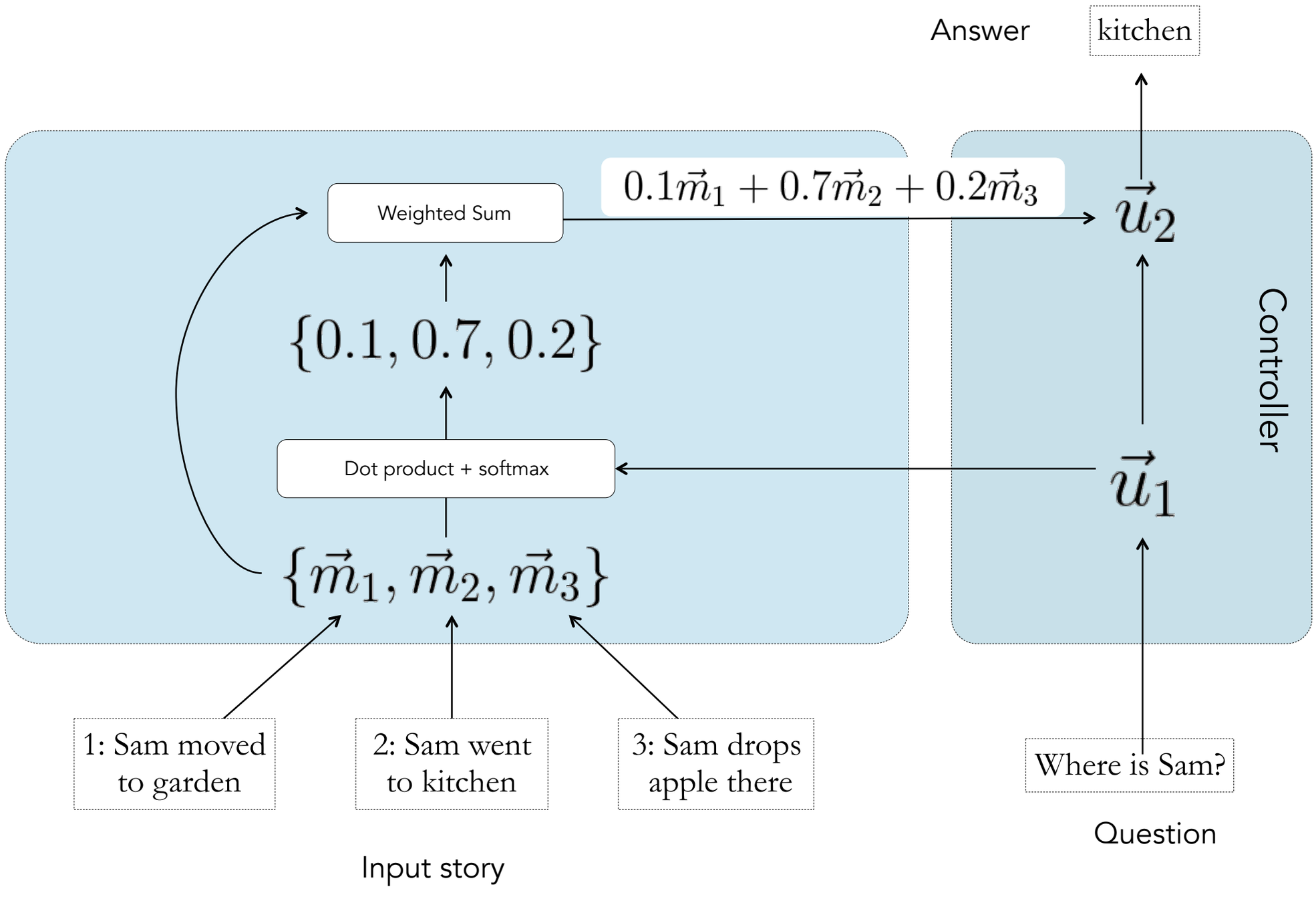
Neural Machine Translation

- Source sentence is mapped to vector, then output sentence generated



Now live for some languages in Google Translate (etc.), with big error reductions!

Neural Question Answering



5. Course Logistics

What I hope to teach:

- Fundamental ideas used in key NLP components
- Some big picture understanding of human languages and the difficulties in understanding and producing them
- An understanding of and ability to build systems for some of the major problems in NLP:
  - text classification, parsing, machine translation, …

Course outline

1. Probabilistic language models
   - *Define probability distributions over text sequences*
2. Text classifiers
   - *Infer attributes of a piece of text by "reading" it*
3. Semantics
   - *Represent meaning*
4. Sequence models
   - *Convert sequences into other sequences*
5. Parsing
   - *Parse sentences into syntactic representations*
6. Machine translation
   - *Map text in one language to text in another*

Grading policy

- Grade
  - 80% (5) programming assignments
  - 20% final project

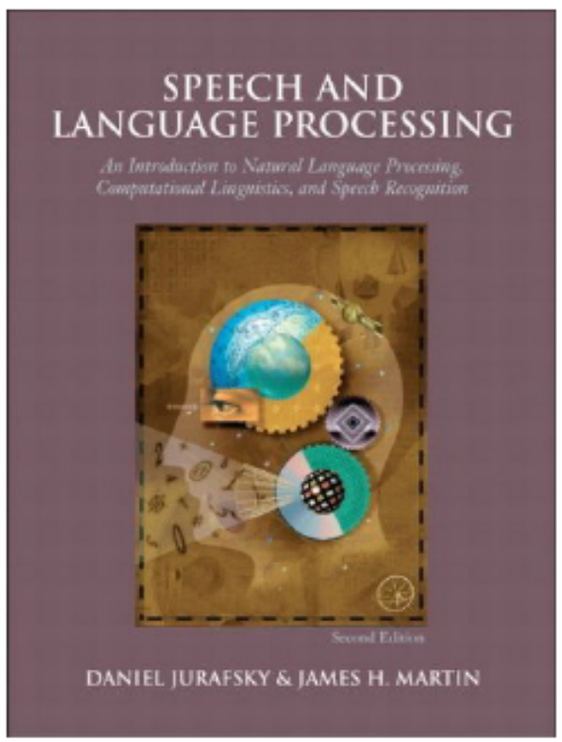
- Participation
  - Attend classes
  - Ask questions; answer questions in class and on Piazza

Readings

- Reference texts
  - Jacob Eisenstein Book
  - Michael Collins notes

- Research articles

- Other texts: Jurafsky and Martin



- Instructor: Ndapa Nakashole (ndapa@eng.ucsd, CSE 4108)

- TAs
  - Sharathabhinav Dharmaji sdharmaj@ucsd.edu
  - Tayal, Piyush ptayal@ucsd.edu
  - 3rd coming soon …

- Course website
  - http://cseweb.ucsd.edu/~nnakashole/teaching/256_sp19.html

Acknowledgements

- Includes content from
  - Noah Smith
  - Chris Manning
  - and indirectly many others …