#DLUPC

Introduction to NLP

Magdalena Biesialska

magdalena.biesialska@upc.edu

PhD Candidate

Universitat Politecnica de Catalunya

Technical University of Catalonia

Outline

Introduction to NLP

- Definitions and key concepts

Linguistic levels

- Morphology
- Syntax
- Semantics
- Pragmatics

Tasks

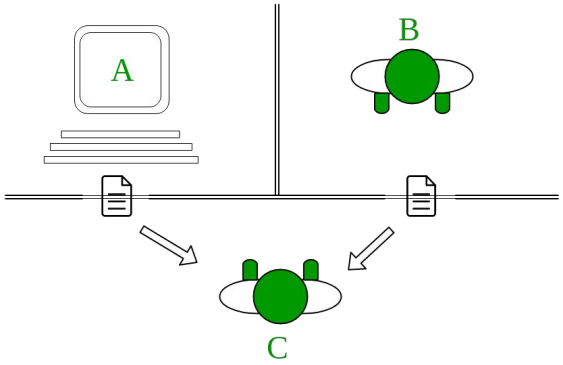
Introduction to NLP

AI, Turing Test and Natural Language Processing

- **Can machines think?** (Turing, 1950)
- How can we determine **if a machine has achieved the ability to behave intelligently**?

Would a human find out that she speaks with a computer?

Using language as humans do is sufficient to test for intelligence



Results of the 2018 Finals

None of the chatbots competing in the finals managed to fool a judge into believing it was human. The judges ranked the chatbots according to how human-like they were. Scores out of 100% were:

1. Mitsuku 33%
2. Tutor 30%
3. Colombina 25%

image source: https://www.geeksforgeeks.org/turing-test-artificial-intelligence/

Definitions

- What is **language**?

Perspective	Definition	Focus
Structural (Saussure)	Language is a system of structurally related elements to encode meaning	Elements of the system: phonological, morphological, grammatical, lexical
Functional (the Prague school)	Language is a function of communication	Pragmatics, discourse, semantic and communicative nature
Behavioural (Skinner)	Language is the result of complex human behaviour	Setting and consequences of verbal behaviour, pragmatics
Generative (Chomsky)	Language is innate and learned skill for humans	Syntax, morphology, phonology

Natural/human language is a system intended to communicate ideas from a speaker to a hearer [1]

[1] Winograd, T. (1972) "Understanding natural language". *Cognitive psychology*, 3(1):1– 191.

Definitions

- What is **Natural Language Processing (NLP)**?

  It is an interdisciplinary field that encompasses:

  - Computer Science
  - Linguistics

  sometimes also referred to as *Computational Linguistics*

- What are the main **goals of NLP**?

  - Represent and ultimately “understand” the meaning of language
  - Process natural language to perform tasks, e.g.: machine translation (Google Translate), question answering (Alexa, Siri, Cortana, Google Home)

Definitions

- What is **language**?

  Natural/human language is a system intended to communicate ideas from a speaker to a hearer

- What is **Natural Language Processing (NLP)**?

  Natural Language Processing (Computational Linguistics) is the field of designing and applying methodology of computer science and linguistics to the processing of natural language.

- What does distinguish **NLP systems** from other data processing systems?

  Natural Language Processing systems use their “knowledge” of language to process data.

Definitions

- What is **language**?

  Natural/human language is a system intended to communicate ideas from a speaker to a hearer

- What is **Natural Language Processing (NLP)**?

  Natural Language Processing (Computational Linguistics) is the field of designing and applying methodology of computer science and linguistics to the processing of natural language.

- What does distinguish **NLP systems** from other data processing systems?

  Natural Language Processing systems use their “knowledge” of language to process data.

Key concepts

How do NLP systems “know” language?

Discreteness, compositionality and sparseness

Properties of natural language:

- **Discreteness**

  Words and characters are discrete symbols

- **Compositionality**

  Sentences and phrases are built from words, and words are built from characters

- **Sparseness = Discreteness + Compositionality**

  In practice there is infinite number of ways words can be combined to create meanings

Zipf's law

- **Zipf's law (1949)**
  - Cross-linguistic, quantitative measure of the lexical diversity of languages
  - Determines how words (tokens) are distributed across documents

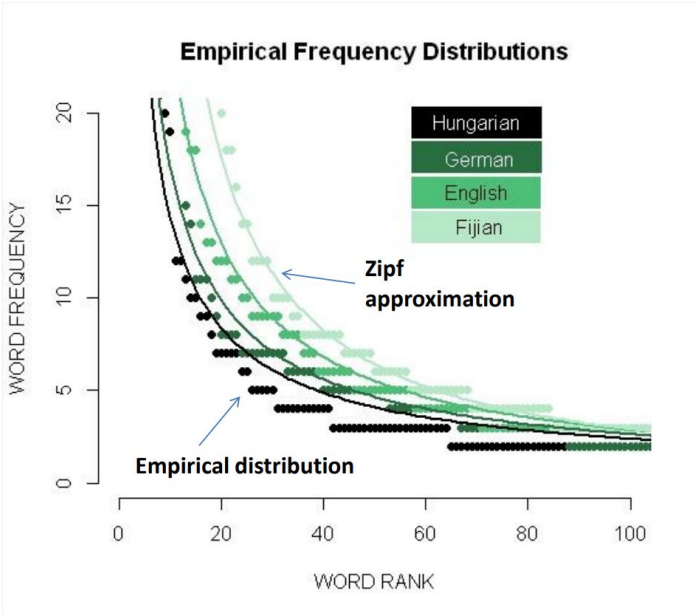
The relative frequency of a word *f(ω)* is inversely proportional to its rank *z*

$$f(\omega) = \frac{c}{z^{\alpha}} = c \cdot z^{-\alpha}$$

where $c = 1$ and $\alpha = 1$

- **Zipf's law of abbreviation**

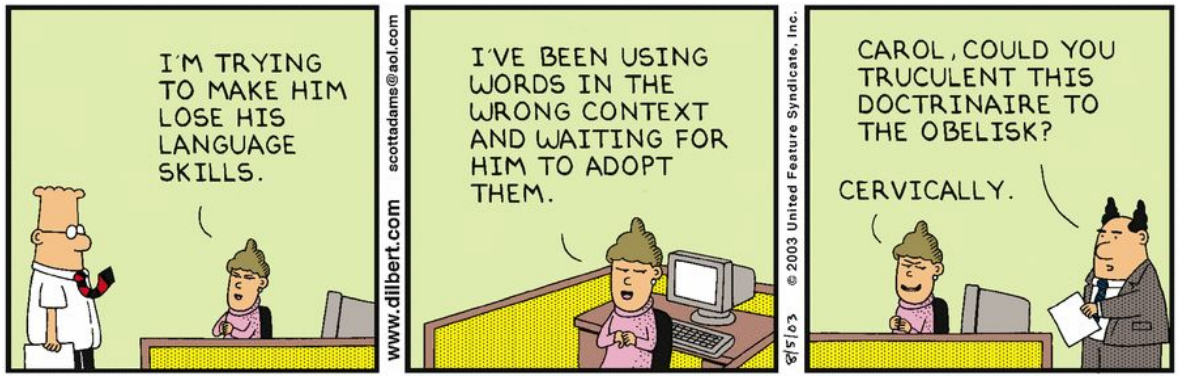
  Words that are frequent tend to be short



based on: Christian Bentz (2013) "Zipf's law and the grammar of languages. Synthetic and analytic encoding strategies across languages of the world. Language Sciences in the 21st century". Cambridge

Distributional hypothesis and context

- The **meaning of a word** can be **inferred from** the **contexts** in which it occurs
- Over the years many NLP systems learned word generalization based on this principle



"You shall know a word by the company it keeps"

– John R. Firth (1957)

Challenges in NLP

- Ambiguity of natural language (unlike programming or any other formal languages)
- Rules exist, but have many exceptions
- Language is infinite ⇒ out-of-vocabulary words, out-of-domain
- Annotation of corpora is expensive (time, money)
- Language phenomena:
  - neologisms, non-standard orthography, use of emojis 
  - idioms
  - humor, sarcasm, irony

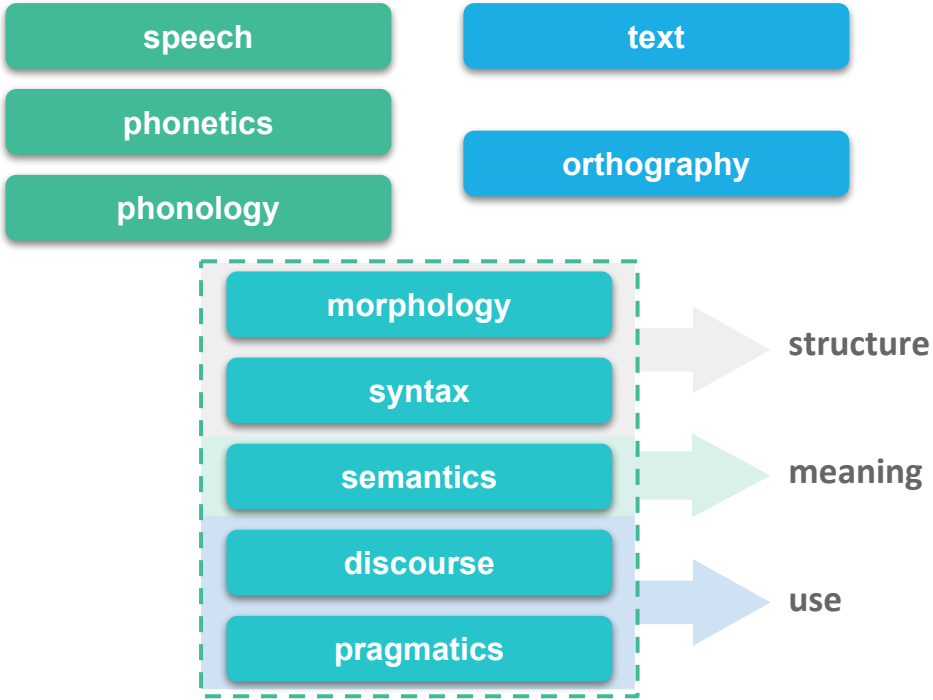
Approaches in NLP

- 1950s - 1990s Rule-based
- 1990s - 2000s Corpus-based statistics
- 2000s - ~2014 Supervised machine learning
- ~2014 - now Deep learning
- Future … Deep learning used to write rules?

based on: Yoav Goldberg (2018) http://u.cs.biu.ac.il/~89-680/nlp-intro-01.pdf

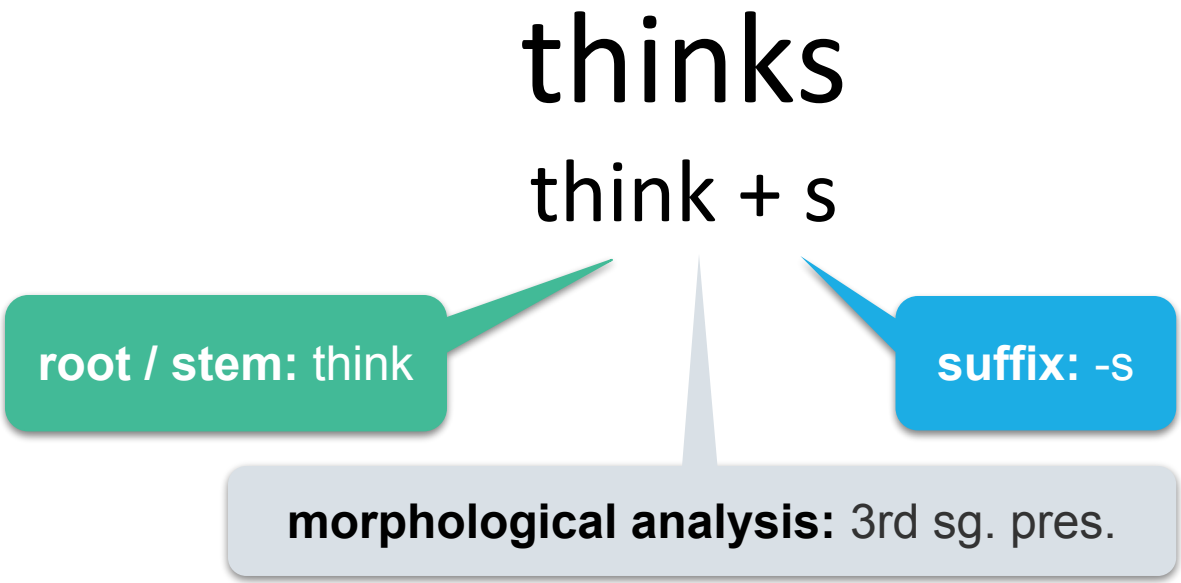
Linguistic levels

Levels of Language Analysis



Morphology

Morpheme - the smallest meaningful unit of language



Morphology

Morphology - the study of the structure of words

Morphological processes:

- **inflection** - systematic modifications of a *root* form, by means of *prefixes* and *suffixes,* to express grammatical distinctions (e.g. tense, number, plurality), e.g think, think***s***, think***ing***, think***ed***
- **derivation** - new words are derived from existing ones, e.g. adjective→adverb (nice→nice***ly***), adjective→verb (wide→wid***en***), verb→adjective (use→use***ful***), verb→noun (manage→manag***er***)
- **compounding** - concatenating at least two words into a new word, e.g. *state-of-the-art*

Morphology

The morpheme per word ratio differs by language

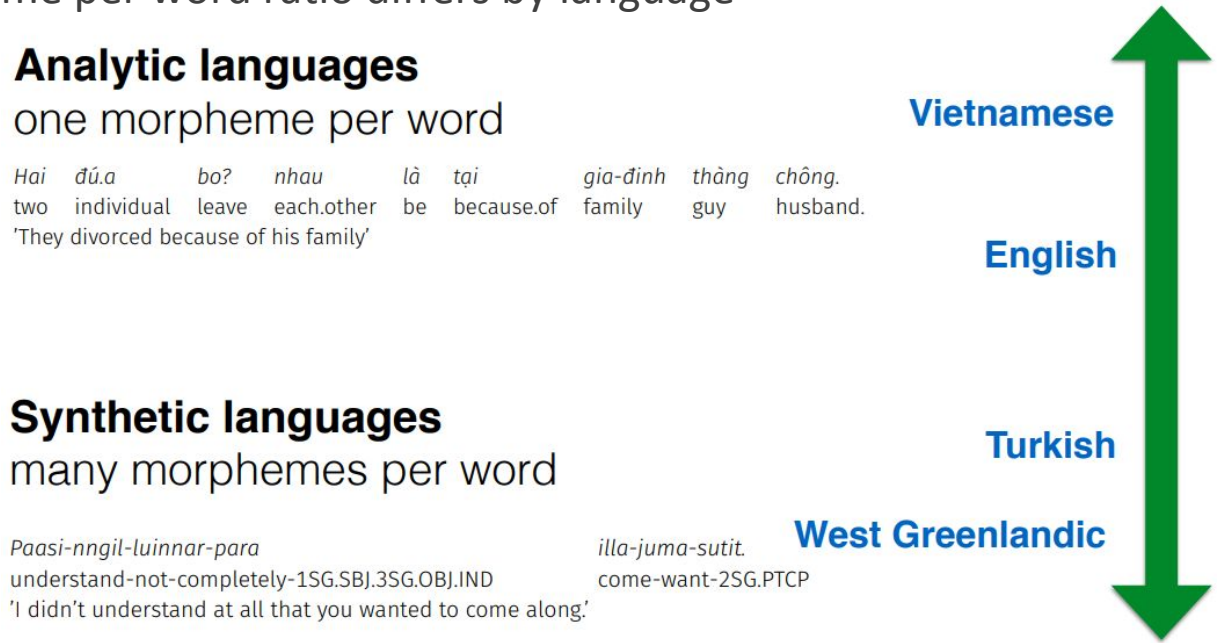


image source: Adam Lopez (2018) https://www.inf.ed.ac.uk/teaching/courses/nlu/assets/slides/2018/l06.pdf

Morphology

The degree to which morphemes are segmentable differs by language

Agglutinative
morphemes have clear boundaries

Swahili
Turkish
Basque

Fusional
morphemes are not easily distinguishable from the root or other morphemes

Spanish
Latin

Syntax

Syntax - the study of the rules for constructing phrases and sentences

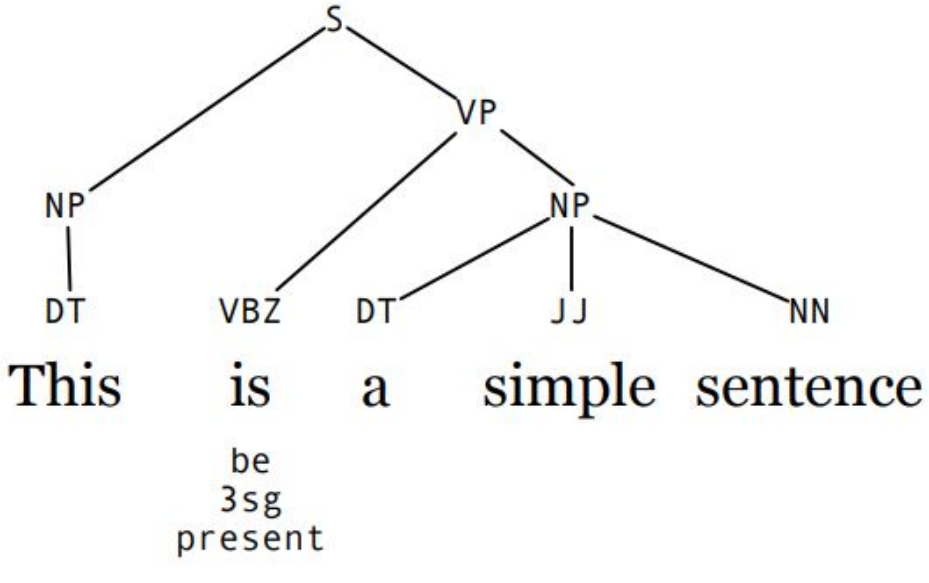


image source: Sharon Goldwater(2018) https://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp01.pdf

Syntax

Major open classes (English):

Class	Subclasses	
Nouns	**Proper nouns** - names of specific entities or persons (e.g. Google, Barcelona, Alan Turing)	
	Common nouns	**Count nouns** - nouns which can occur in both singular and plural forms (e.g dog / dogs), can be enumerated (one dog, two dogs, …)
		Mass nouns - uncountable nouns, treated as a group, volume, mass or quantity (e.g. snow, wine, knowledge)
Verbs	**Auxiliaries** (closed subclass)	**Copula** - indicates present/past/future (tense), if action or event is completed (aspect) (e.g. *be*, *do*, *have*)
		Modal verbs - the mood associated with the action or event: possibility (e.g. *can*, may), necessity (e.g. *must*)

Syntax

Major open classes (English):

Class	Subclasses
Adjectives	Properties or qualities, for concepts such as color (e.g. green, white), value (e.g. good, bad) among others
Adverbs	**Directional (locative) adverbs** - the direction or location of some action (e.g. there, school, downhill, parallel)
	Degree adverbs - the extent of some property, action, process (e.g. many, much, very)
	Manner adverbs - the manner of some action, process (e.g. rapidly, professionally)
	Temporal adverbs - the time that some action or event happened (e.g. today, Wednesday)

Syntax

Closed classes (English):

Class	
Prepositions	on, at, from, to, under, over, near, with, by
Determiners	a, an, the
Pronouns	I, she, others, who
Conjunctions	and, but, or, as, if, when
Particles	up, down, on, off, in, out, at, by
Numerals	one, two, three, first, second, third

Syntax

Syntactic analysis - uses grammar to determine what sentences are valid

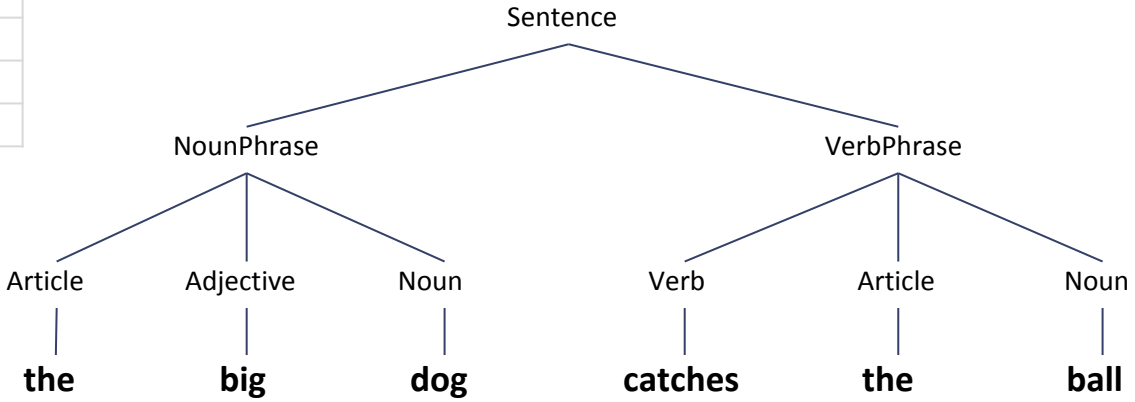
- **Constituent** - groups of words may behave as a single unit or phrase (e.g. *noun phrase*, *verb phrase*)
- **Context-Free Grammar (CFG)** - a set of production (recursive rewriting) rules used to generate string patterns. Contains the following components:
  - **terminals** - a finite set ***T*** of terminal symbols (characters of the alphabet)
  - **non-terminals** - a finite set ***N*** of non-terminal symbols
  - **productions rules** - a finite set ***P*** of production rules: $P \subset N \times (N \cup T)^*$, where $A \in N$ and $\alpha \in (N \cup T)^*$ and $A \rightarrow \alpha$
  - **start symbol** - non-terminal symbol ***S*** that appears in the first string generated by the grammar
- **Context-free languages (CFL)** - languages that can be constructed from individual strings by concatenation, union, and recursion

Syntax

Parse tree: *The big dog catches the ball*

Grammar	
Sentence	→ NounPhrase, VerbPhrase
VerbPhrase	→ Verb, NounPhrase
NounPhrase	→ Article, Noun
NounPhrase	→ Article, Adjective, Noun

Lexicon	
Word	**Class**
dog	Noun
catches	Verb
big	Adjective
ball	Noun
the	Article



Syntax

Word order:

- **SVO** - the *verb* is between the *subject* and *object*, e.g. English, German, French, Mandarin
- **SOV** - the *verb* is at the end of basic clauses, e.g. Hindi, Japanese
- **VSO** - the *verb* is at the beginning of basic clause, e.g. Irish, Classical Arabic, Biblical Hebrew

Semantics

Semantics - the study of the meaning of words, phrases and sentences

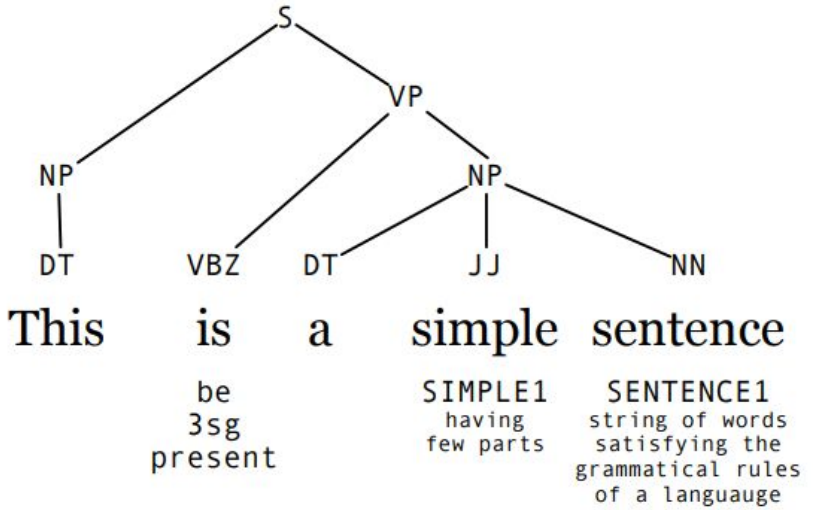


image source: Sharon Goldwater(2018) https://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp01.pdf

Semantics

- **lexical semantics** - the study of the meaning of individual words
  - **hypernymy** and **hyponymy** - *animal* is a **hypernym** of *dog* whereas *dog* is a **hyponym** of *animal*
  - **synonymy** and **antonymy** - *fashionable* is a **synonym** of *trendy* whereas *short* is an **antonym** of *long*
  - **holonymy** and **meronymy** - *body* is a **holonym** of *hand* whereas *engine* is a **meronym** of *car*
  - **homonymy**, **homophony** and **polysemy** - **homonym:** *bat*, **homophony:** *write/right*, **polysemy:** *firm*

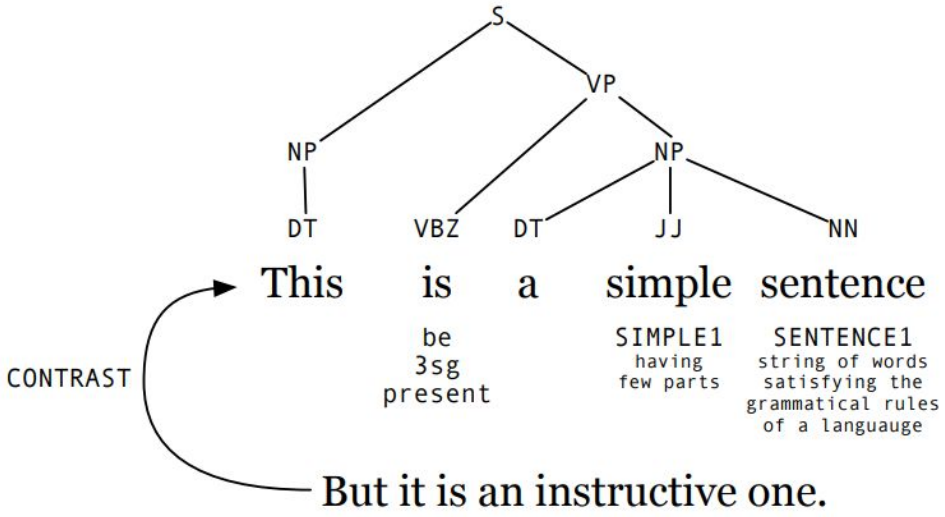
Semantics

- the study how word meanings are related to each other
  - **collocations** - *strong tea*, *powerful computer*, *high mountains*, *tall trees*
  - **idioms** - *beat around the bush*, *once in a blue moon*
  - **scope** - quantifiers and operators have scope, e.g. *Every student likes some book.*

The meaning of a sentence is usually a combination of the meaning of its words - it is important for interpretation to take into account syntax information

Pragmatics

Pragmatics - the study of how language is used in context



Pragmatics

Context refers to both the linguistic and situational context

- **Discourse**
  - discourse connectives: *however*, *furthermore*, *if … then*
  - coreference: anaphora - reference to an entity previously introduced into the discourse

- TEMPORAL
  - Asynchronous
  - Synchronous: precedence, succession
- CONTINGENCY
  - Cause: result, reason
  - Pragmatic cause: justification
  - Condition: hypothetical, general, unreal present, unreal past, real present, real past
  - Pragmatic condition: relevance, implicit assertion
- COMPARISON
  - Contrast: juxtaposition, opposition
  - Pragmatic contrast
  - Concession: expectation, contra-expectation
  - Pragmatic concession
- EXPANSION
  - Conjunction
  - Instantiation
  - Restatement: specification, equivalence, generalization
  - Alternative: conjunctive, disjunctive, chosen alternative
  - Exception
  - List

image source: Prasad et al., 2008 “The hierarchy of discourse relation in the Penn Discourse Treebank annotations”

Tasks


YOU GOT THIS EMAIL SIX DAYS AGO. DO YOU WANT TO FOLLOW UP?
OH, RIGHT, I SHOULD DO THAT.
TYPE TYPE.
DID YOU FORGET THE ATTACHMENT?
OOPS, YES.
TODAY IS THE RECIPIENT'S BIRTHDAY. DID YOU WANT TO MENTION THAT?
WAIT, IT IS?

Text processing

Every downstream NLP task needs previously processed data.

The following text processing steps are usually performed:

- Normalization
- Tokenization / Segmentation
- Cleaning
- Truecasing

Classification of tasks with respect to linguistic levels

- Morphology
  - Lemmatization / Stemming
- Syntax
  - Part-of-Speech (POS) tagging
  - Syntactic Parsing
- Semantics
  - Named-Entity Recognition (NER)
  - Semantic Role Labelling / Semantic Parsing
  - Word Sense Disambiguation
- Pragmatics
  - Coreference resolution and Anaphora resolution
  - Sentiment analysis
  - Question answering (QA)
  - Machine Translation (MT)

Part of Speech (POS) tagging

POS tagging is the process of assigning a part-of-speech or other lexical class to each word in a corpus.

VB DT NN .
Book that flight .

VBZ DT NN VB NN ?
Does that flight serve dinner ?

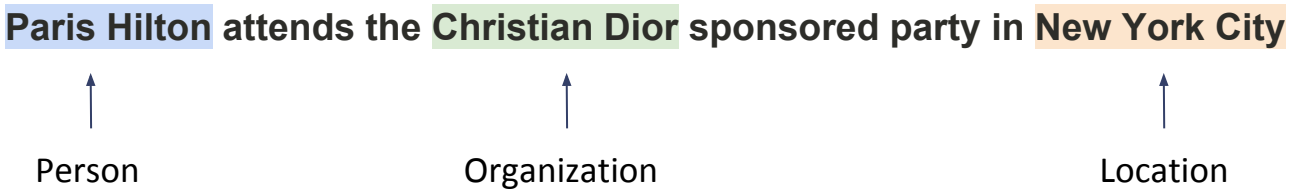
- Rule-based taggers
- Stochastic taggers (based on Hidden Markov Model)

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Named-Entity Recognition (NER)

Named-Entity Recognition deals with identifying names in a text and classifying them into predefined set of categories

Categories
Person
Location
Organization
Other



Coreference and Anaphora resolution

- **Anaphora resolution** determines an antecedent for each noun phrase that depends on other noun phrases in a discourse

Every student has **its** book

- **Coreference resolution** identifies the noun phrases and pronouns in a discourse that refer to the same entity

NLP is very exciting.

I want to learn more about Deep Learning methods used in **NLP**.

Sentiment analysis

Sentiment analysis predicts the opinion expressed in a text

★ **1**/10

Boring, predictive, repetitive...

rquku 10 October 2018

Lady Gaga can sing, but i can not say therefore that was a good movie. I was waiting the whole time when the heck this movie would end. I do not recommend it.

★ **9**/10

2018's Best

hoffmanben-327622 November 2018

This film takes you through every human emotion possible, which makes it such a great story. Anyone who has gone through personally or closely related to someone who battled addiction brings a sense of peace. It's a disease and very real. Surely to be nominated for Best Picture.

source: https://www.imdb.com/title/tt1517451/reviews

Question answering (QA)

Involves two tasks:

- finding documents that may contain an answer to a given question
- finding an answer in a document or paragraph ⇒ Reading Comprehension task

Used to develop:

- dialog systems
- chatbots

In 2011 IBM Watson won in Jeopardy! with two all-time champions

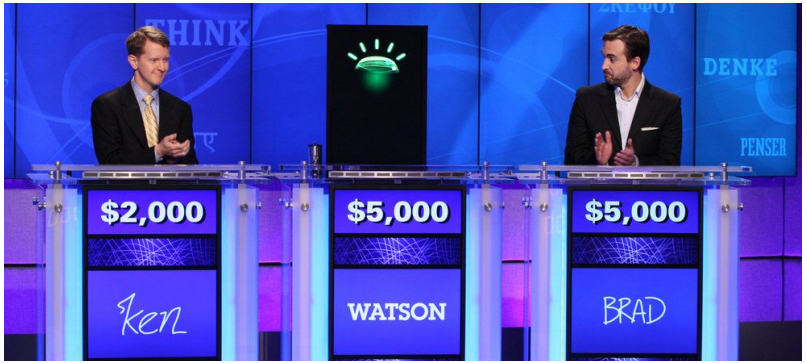


image source: New York Times https://goo.gl/images/4iMspt

Progress in NLP

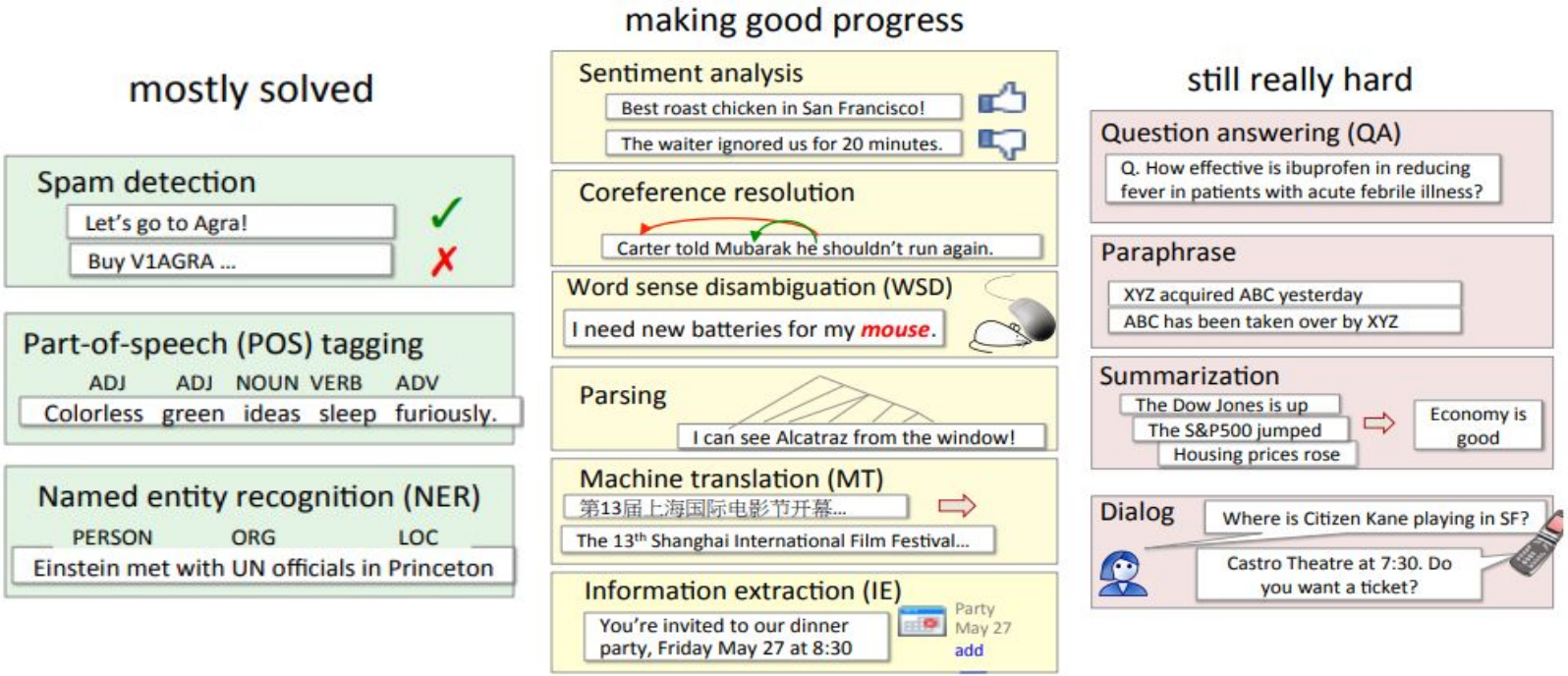
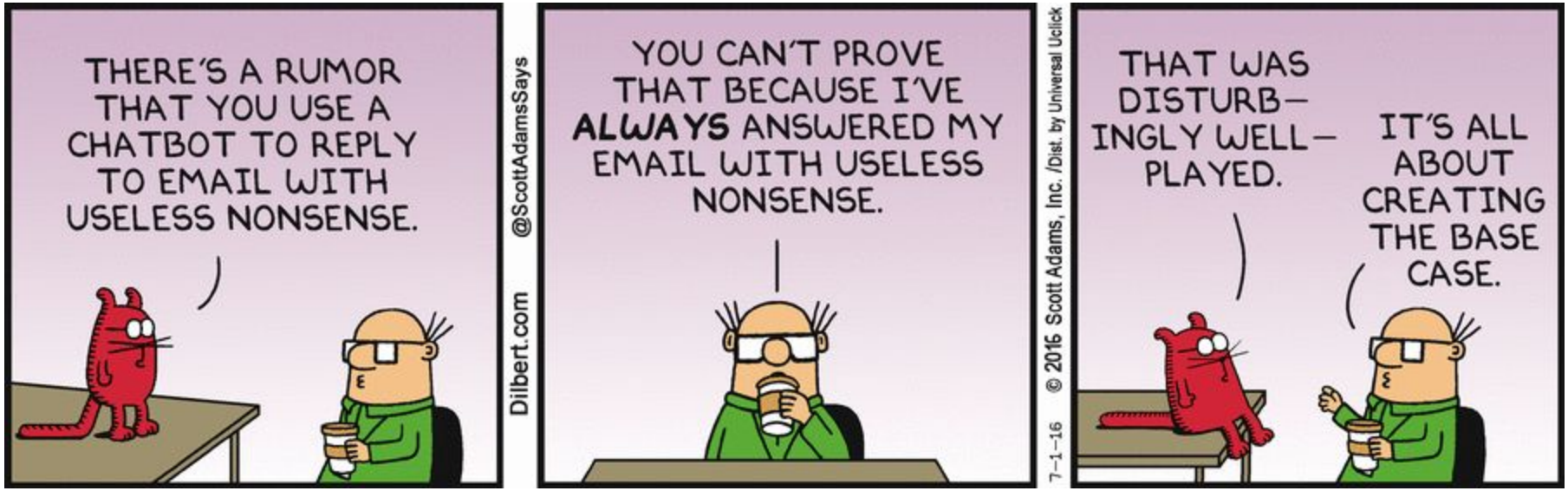


image source: Dan Jurafsky & Chris Manning http://spark-public.s3.amazonaws.com/nlp/slides/intro.pdf



We are not there yet! Current NLP systems are close to but don't achieve human parity just yet.

There's still work to be done!

Quiz

To play go to: **kahoot.it** and enter: **<**Game PIN>

Further reading

Textbooks

- Jurafsky, D. & Martin, J.H. (2009). “Speech and language processing: an introduction to natural language processing”. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (3rd ed. draft available for free) http://www.cs.colorado.edu/~martin/slp.html (2nd ed.)
- Goldberg, Y. & Hirst, G. (2017). “Neural Network Methods in Natural Language Processing”. Morgan & Claypool Publishers.
- Manning, C.D., & Schütze, H. (1999). “Foundations of Statistical Natural Language Processing”.

Articles

- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.P. (2011). “Natural Language Processing (almost) from Scratch”. *Journal of Machine Learning Research, 12*, 2493-2537.
- Goldberg, Y. (2016). “A Primer on Neural Network Models for Natural Language Processing”. *J. Artif. Intell. Res., 57*, 345-420.
- Cho, K. (2017). “Natural Language Understanding with Distributed Representation” https://github.com/nyu-dl/NLP_DL_Lecture_Note

Thank you! Q&A?