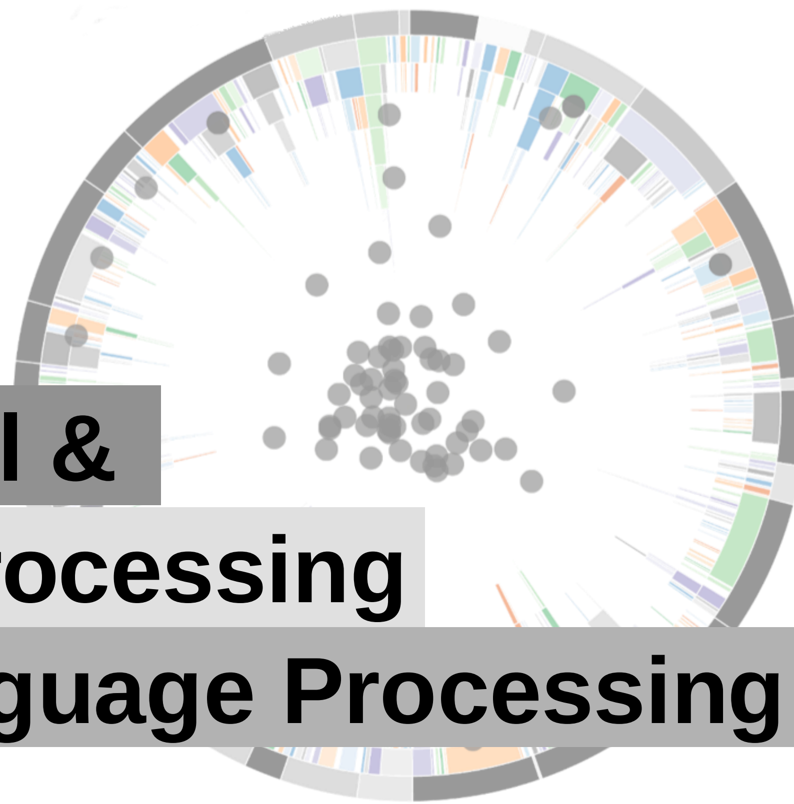
Cluster of Excellence
The Politics of Inequality

Steinbeis Transfer Center
Linguistic Data Analysis

UNIVERSITÄT
PASSAU



Information Retrieval & Natural Language Processing

Week 8: Natural Language Processing

Annette Hautli-Janisz, Prof. Dr.

19 December 2024

What is Natural Language Processing?

NLP is a field at the intersection of

- computer science
- artificial intelligence
- and linguistics.

Goal: for computers to process or “understand” natural language in order to perform tasks that are useful, e.g.,

- making appointments, buying things
- question answering (Siri, Google Assistant, Facebook, …)

Fully understanding and representing the meaning of language (or even defining it) is a difficult goal.

The field of Natural Language Processing

One of the core areas of Artificial Intelligence, because language is a key factor for communicating between humans and machines.

Field has seen an exponential growth in the last 10 years, main conferences are ACL, NAACL, COLING, EMNLP etc.

For papers, see the ACL anthology.

ACL Anthology

(A tiny sample of) NLP applications

Applications range from simple to complex:

- Spell checking, keyword search, finding synonyms
- Extract information from websites such as product price, dates, location etc.
- Classification: reading level of school texts, positive/negative sentiment of longer documents
- Machine translation
- Spoken dialogue systems
- Complex question answering

NLP in industry… is taking off

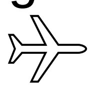
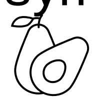
- Search (written and spoken)
- Online advertisement matching
- Automated/assisted translation
- Speech recognition
- Chatbots/dialogue agents
    - Automating customer support
    - Controlling devices
    - Ordering goods

What's 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
- using an encoding which little kids can quickly learn (amazingly!)

A human language is a discrete/symbolic/categorical signaling system

- airplane =  ; avocado = 
- with very minor exceptions for expressive signaling ("I loooooove it." or "Whooompppaaaa")
- presumably because of greater signal reliability.
- symbols are not just an invention of logic or classical AI.

What's special about human language?

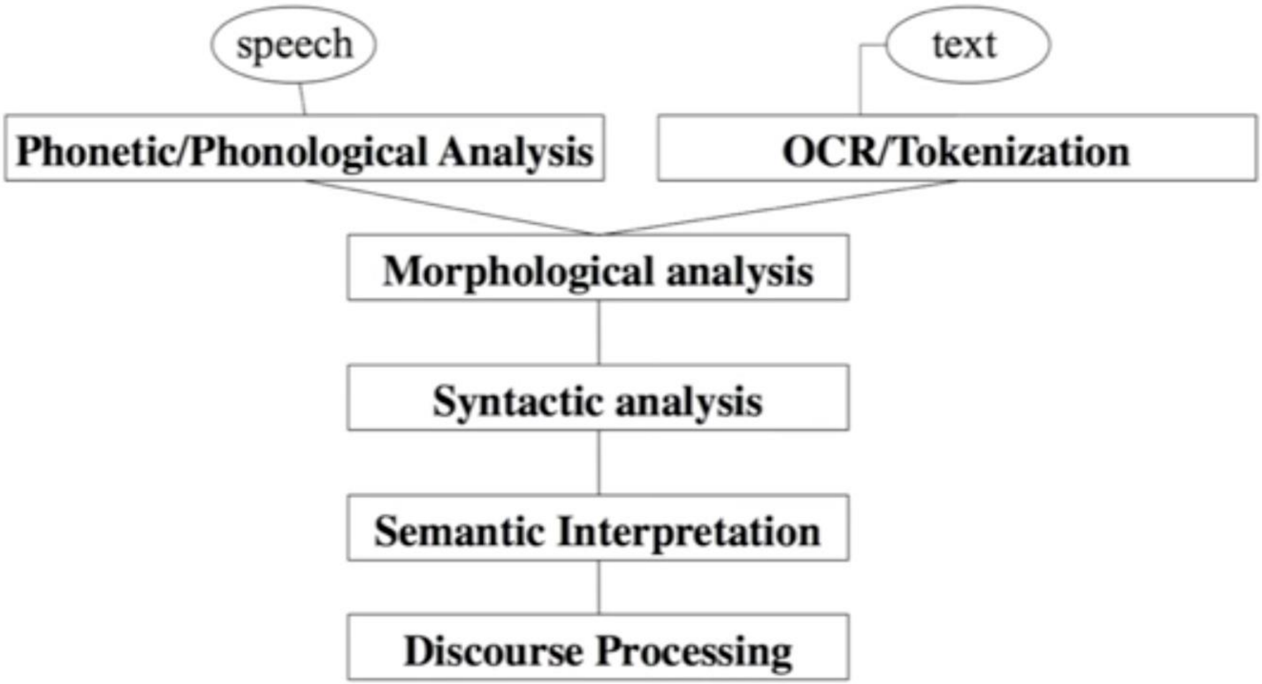
The categorical symbols of a language can be encoded as a signal for communication in several ways:

- sound
- gesture
- images (writing)

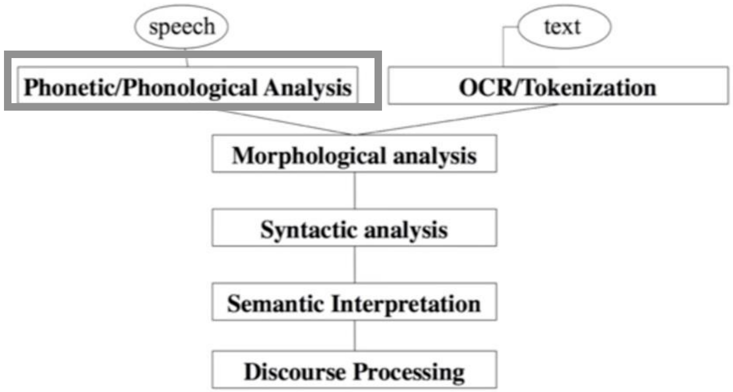
The symbol is invariant across different encodings!

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.
Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.

Different linguistic levels in Natural Language Processing



Different linguistic levels in Natural Language Processing



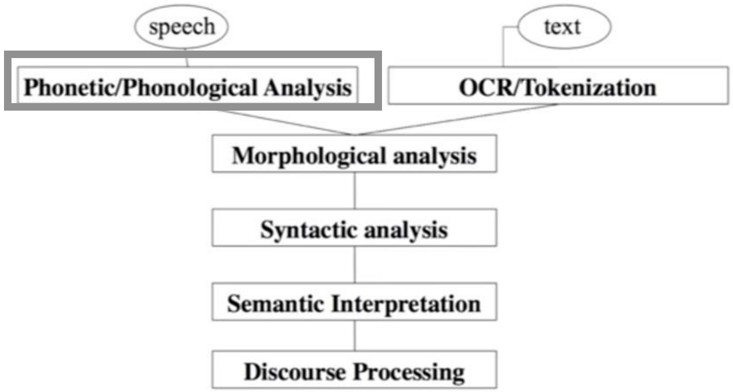
Phonetics is the study of human sound:

- production (articulatory)
- transmission (acoustic)
- perception (auditive)

Sounds can be divided into

- consonants
    - place
    - manner of articulation
    - voice
- vowels

Different linguistic levels in Natural Language Processing

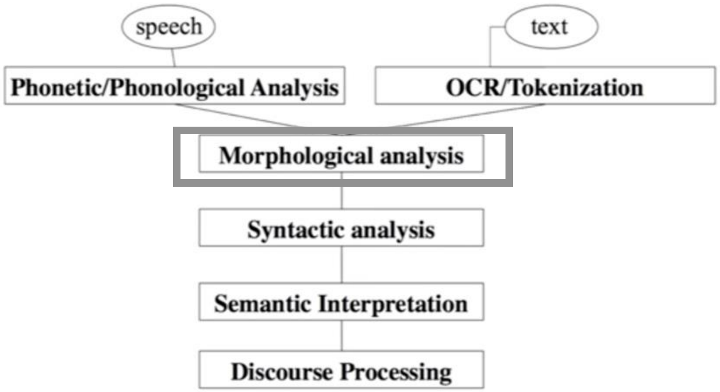


Phonology is the study of the sound system within a language and across languages.

Example: the ‘s’ in ‘brakes’ versus ‘waves’

→ Understand how speech sounds transform depending on situations or their position in syllables, words, and sentences.

Different linguistic levels in Natural Language Processing



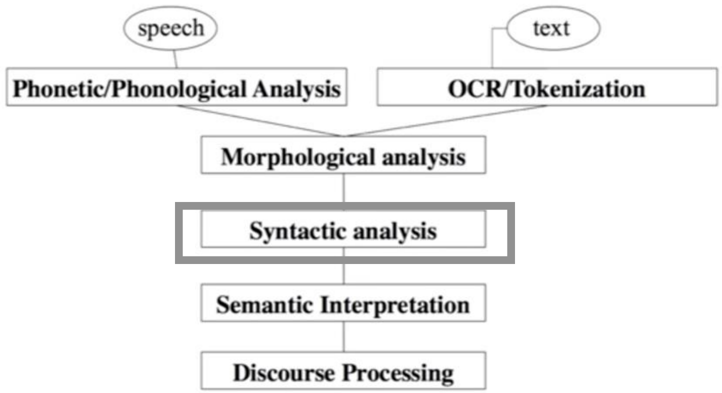
Morphology is the study of words.

Morphemes are the minimal units of words, they have a meaning and cannot be subdivided further.

Morphological parsing is the process of determining the morphemes from which a given word is constructed, for instance with Finite State Transducers.

Example: foxes:fox+Noun+Pl

Different linguistic levels in Natural Language Processing

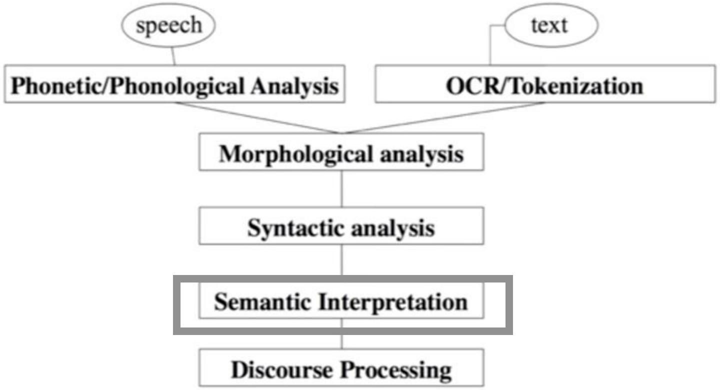


Syntax is the grammatical structure of words and phrases to create coherent sentences.

Syntactic parsing has attracted huge attention in the last 20 years of NLP research.

More on this later.

Different linguistic levels in Natural Language Processing



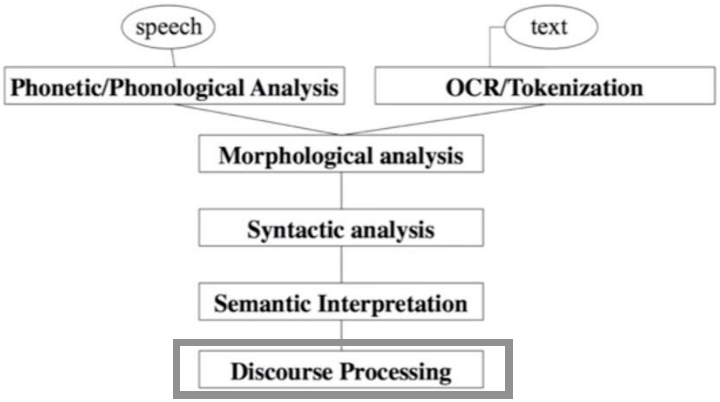
Semantics is the study of the meaning of individual words and phrases in a language or in a particular context.

Semantic representation: expressing the meaning of language.

Semantic interpretation: deriving meaning from a word stream, e.g., mapping syntactic structure to semantic structure.

Latest developments in NLP: use distributed word representations to approximate word meaning (see last lectures).

Different linguistic levels in Natural Language Processing



Pragmatics deals with the meanings and effects which come from the use of language in particular situations.

One of the hard areas of NLP: looking beyond the sentence, determine the meaning and the function of a sentence within a context.

E.g., argument mining. More on this later.

What makes NLP hard? Ambiguity!

Phonetic: “wreck a nice beach”

Word sense: “I went to the bank”

Part of speech: “I made her duck.”

Attachment: “I saw a man with a telescope.”

Coordination: “If you love money problems show up.”

Speech act: “Have you emptied the dishwasher?”

What makes NLP hard? Ambiguity!

7 May 2021

Crash blossoms – headlines with double meaning.

“Teacher Strikes Idle Kids”

“Hospitals Are Sued by 7 Foot Doctors”

“Local High School Dropouts Cut in Half”


BBC Home News Sport Reel More

NEWS Menu

Newsbeat

Knife crime: St John Ambulance to teach teens to help stab victims

What makes NLP hard? Text-only.

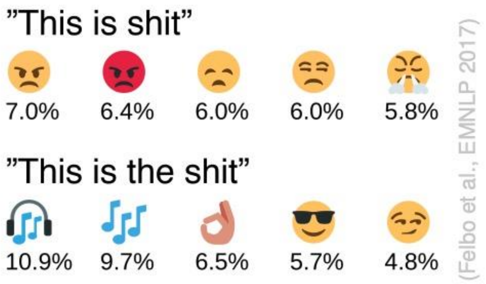
Depending on the focus in the sentence, different meanings result:

I never said she stole my money.	Someone else said it, but I didn't.
I *never* said she stole my money.	I simply didn't ever say it.
I never *said* she stole my money.	I might have implied it in some way. But I never explicitly said it.
I never said *she* stole my money.	I said someone took it. But I didn't say it was her.
I never said she *stole* my money.	I just said she probably borrowed it.
I never said she stole *my* money.	I said she stole someone else's money.
I never said she stole my *money*.	I said she stole something of mine. But not my money.

What makes NLP hard? Non-standard language

Colloquial language

- Non-standard writing. “@justinbieber Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever”
- Informal use. “This is sh*t” vs. “This is the sh*t”



Special phrases

- Tricky entities. “Let it Be was recorded”, “mutation of the for gene”, ...
- Idioms. “get cold feet”, “lose face”, ...
- Neologisms. “unfriend”, “retweet”, “hangry”, ...

Tricky segmentation

- Hyphens. “the New York-New Haven Railroad”
- Punctuation. “She was a Dr. I was not.”
- Whitespaces. “ 本を読む ”, “Just.Do.It.”

What makes NLP hard? Practical issues

Effectiveness issues

- Effectiveness. The extent to which the output of a method is correct
- Methods may not be effective enough for use in real-life applications.

Efficiency issues

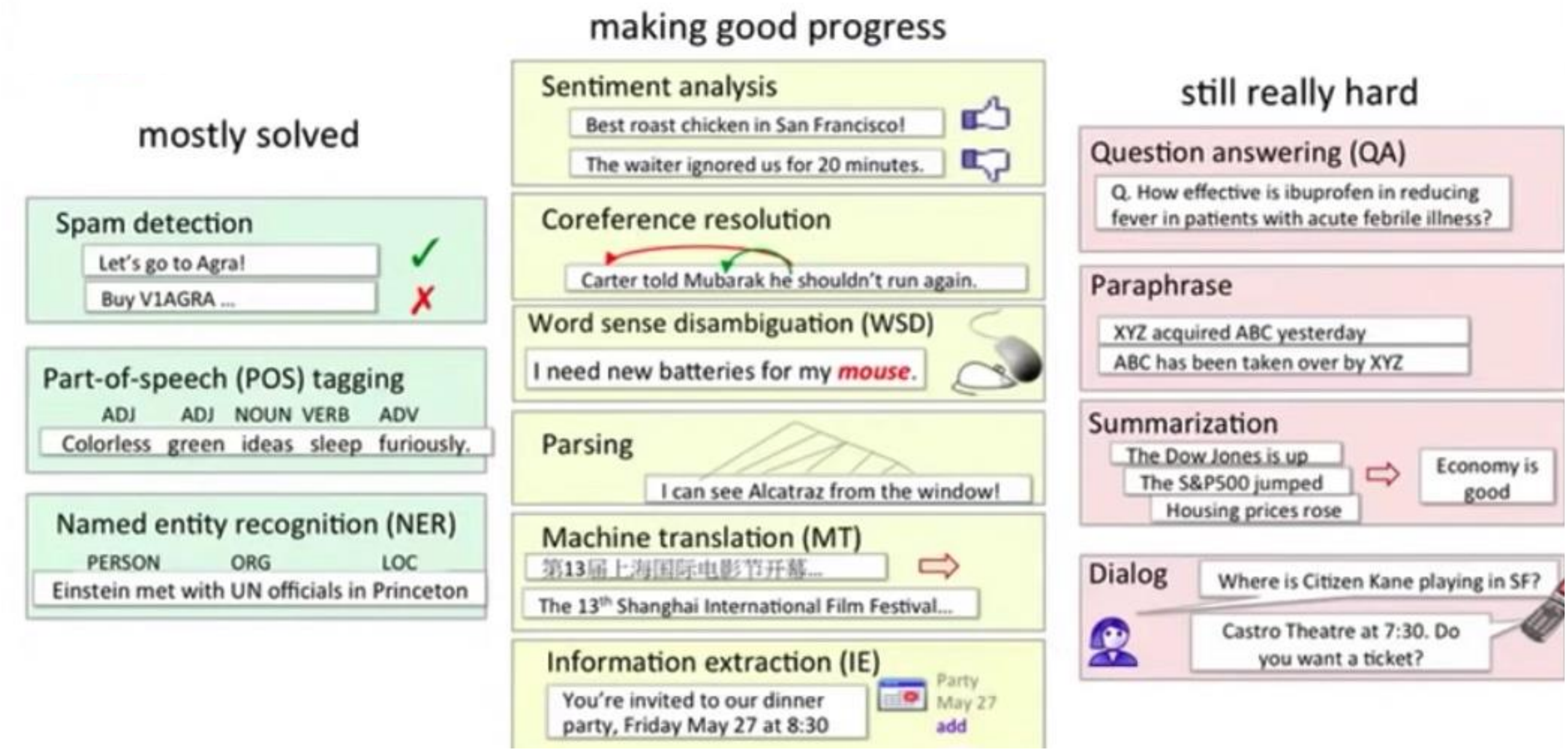
- Efficiency. The run-time, space, or energy consumption of a method
- Methods may not be efficient enough when applied to big text amounts.

Robustness issues

- Robustness. The effectiveness of a method across domains of text
- Methods may not be robust enough on data different from training data.

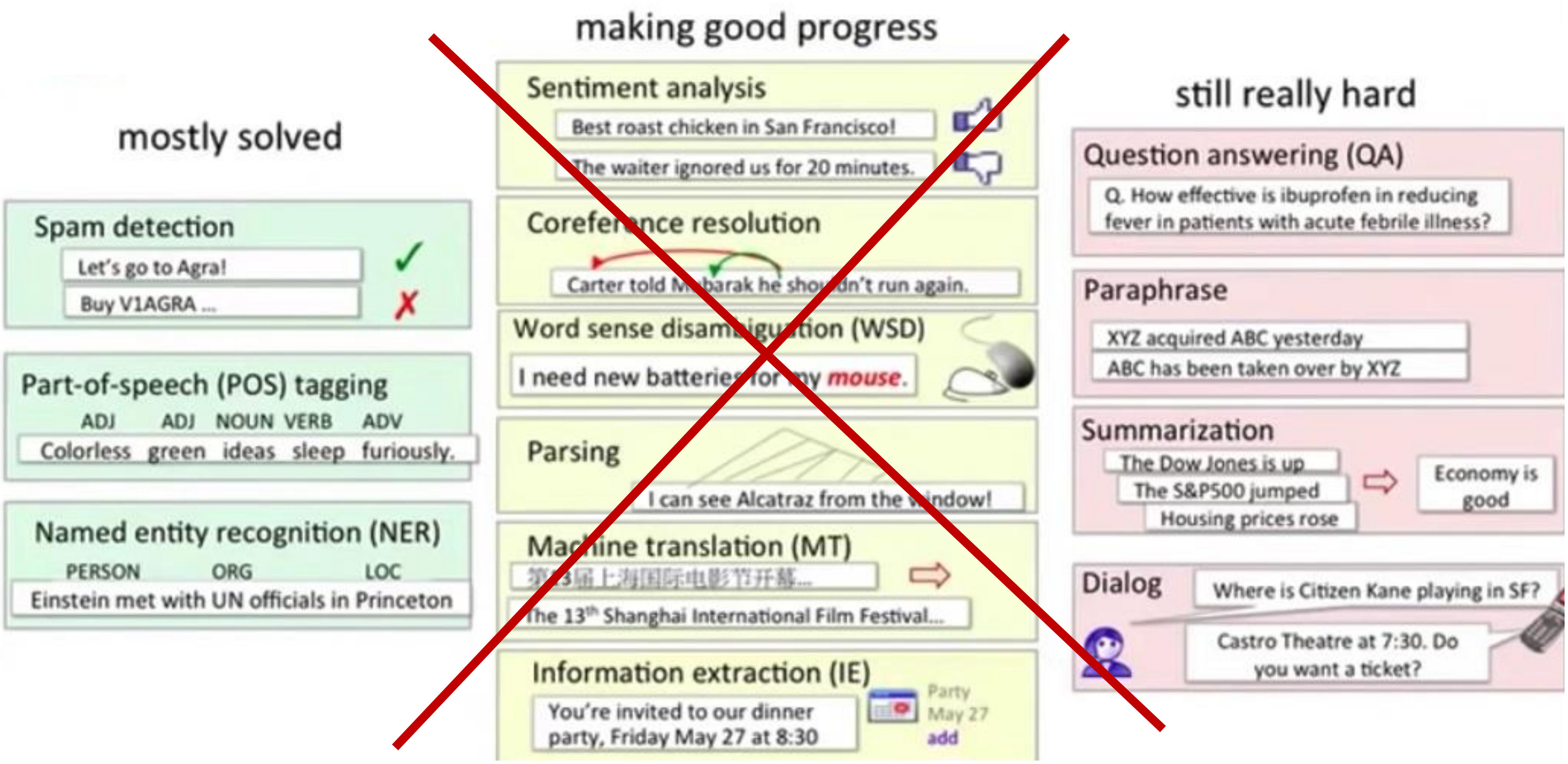
Language technology

From Dan Jurafsky, Stanford University, about 10 years ago:



Neural models & LLMs

From Dan Jurafsky, Stanford University:



Types of NLP tasks

Analysis and synthesis:

- Analysis: The decoding of structured information from text
- Synthesis: The encoding of (structured) information into text

Also: Natural Language Understanding (NLU) versus Natural Language Generation (NLG)

Types of NLP tasks

Analysis tasks, selected:

- Token and sentence splitting
- Stemming and lemmatization
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition

Synthesis tasks, selected:

- Lexicon generation
- Spelling correction
- Summarization
- Text style transfer

Example: Information extraction

Task: Identify entities, their attributes and their relations in a given text.

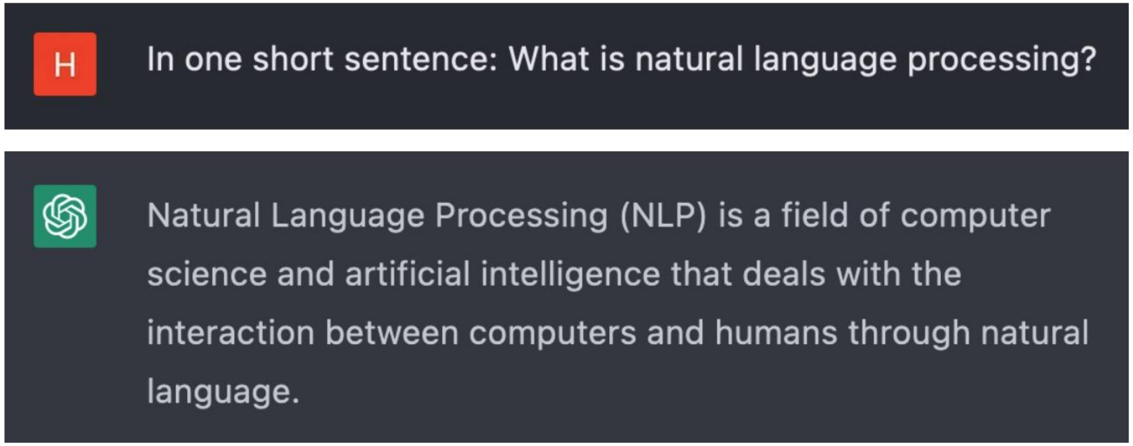
Example: Extract a company's founding dates from a newspaper article.

Possible approach?

Example: Language modeling

Task: Extend a text word by word until an appropriate ending is reached.

Example: Answer a user's question to a chatbot.



Possible approach:

Train general language model on huge amounts of text examples

Fine-tune model on question-answer training pairs

Terms in NLP

Task. A specific problem with a defined input and desired output (e.g., constituency parsing, summarization, ...)

Technique. A general way of analyzing/synthesizing a text (e.g., language model, ...)

Algorithm. A specific implementation of a technique (e.g., GPT-3, ...)

Model. The configuration of an algorithm resulting from training (e.g., CKY parsing on Penn Treebank, GPT-3 fine-tuned on a set of Q&A pairs, ...)

Method. May refer to an algorithm, model

Applications

Software that employs NLP to solve real-world problems. This includes tools, web services etc.

Why applications?
Automate human tasks and/or improve human performance.
Use cases:

Examples?

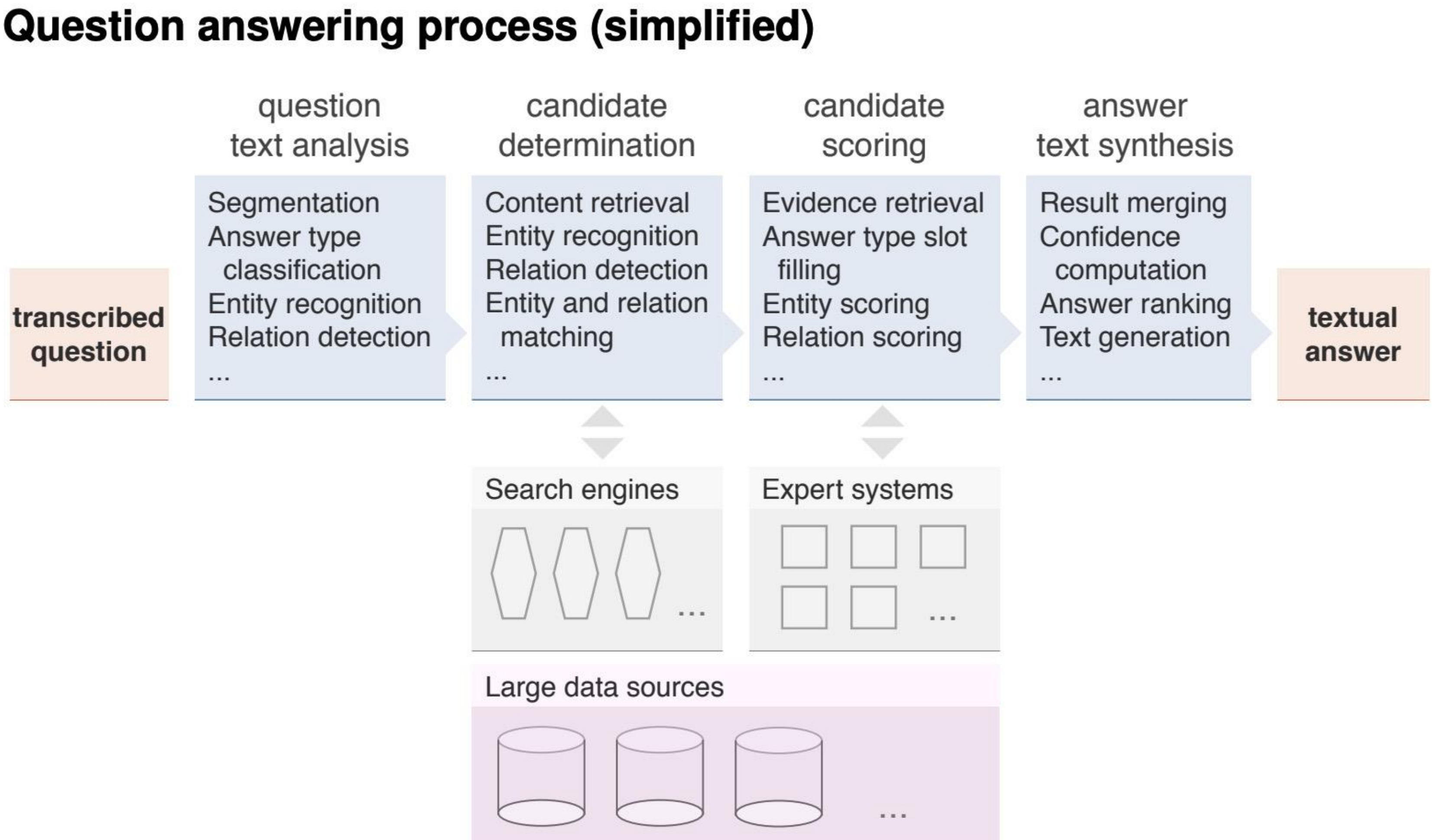
Today: Focus on computational methods rather than applications.
Applications motivate why we deal with specific methods.

One application: IBM's Watson

IBM's Supercomputer, became famous in 2011.

https://www.youtube.com/watch?v=P18EdAKuC1U

One application: IBM's Watson



Evolution of NLP applications

Selected milestones:

February 2011: Watson wins Jeopardy.

October 2011: October 2011. Siri starts on the iPhone

August 2014. Skype translates conversations in real time

May 2018. Google Duplex makes phone call appointments

February 2019. Project Debater competes in entire debates

November 2022. ChatGPT leads conversations on any topic

Different ways to do NLP

NLP using rules.

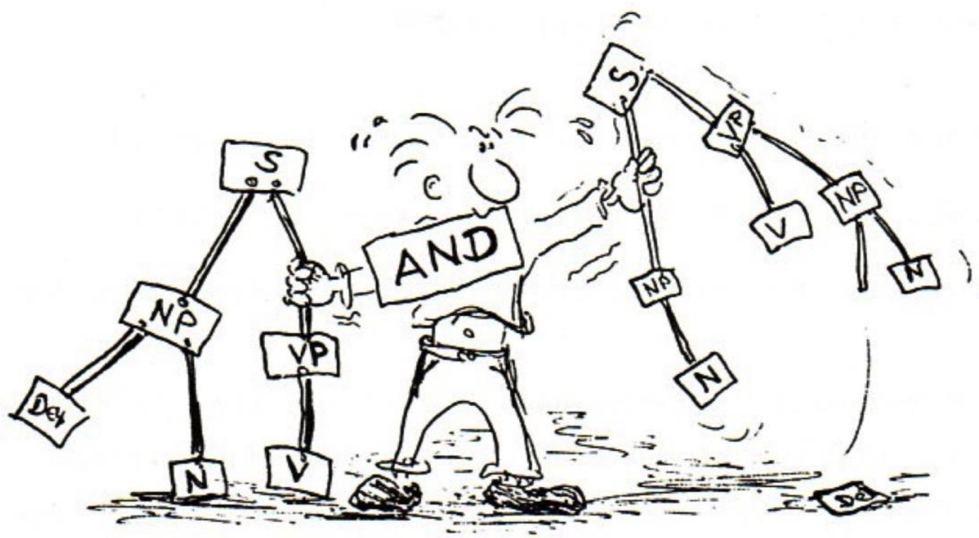
NLP using lexicons.

NLP using context-free grammars.

NLP using language models.

Rule-based NLP

The quality of any rule-based NLP rises and falls with the knowledge of the human expert.



Encoding of knowledge: decision/rewrite rules, string matching, lexicons, grammars, etc.

Rule-based NLP

Hand-crafted decision trees

- Series of decision rules that classify or segment input text spans

```
if char ∈ {'.', '?', '!'} then return true
                          else return false
```

Finite-state transducers

- Series of rules that rewrite matching input spans into output spans

```
(*vowel*) y → i
```

If a span contains vowel and ends with ‘y’, replace ‘y’ with ‘i’.

Template-based generation

- Predefined string templates filled with information to create new text

“I am `<stance>` `<issue>`, because `<reason>`.”

Example: “I am *con death penalty* because *the death penalty kills innocent people.*”

Rule-based NLP

Alternative to hand-crafted rules?

- (Semi-) Automatic definition of implicit or explicit rules using statistics derived from a given dataset
- Done with probabilistic techniques or machine learning
- Aka: Statistical inference or the data-driven approach

Rule-based vs. statistical methods

- For most analysis and synthesis tasks, the best results are nowadays achieved with statistical/neural techniques.
- Particularly in industry, rule-based techniques are still used, because they are often well-controllable and explainable.
- All rule-based methods have a statistical counterpart in some way.

Different ways to do NLP

NLP using rules.

NLP using lexicons.

NLP using context-free grammars.

NLP using language models.

Lexicon-based NLP

Lexicon: A repository of terms (in terms of words or phrases) that represents a language, a vocabulary, or similar

Observations:

- Lexicons often store additional information along with a term.
- Lexicons often have an explicit ordering, for example, alphabetically.

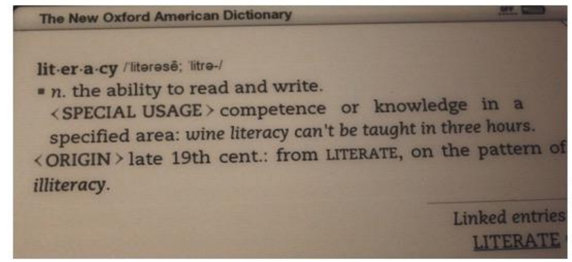
Types of lexicons:

- Terms only. Term lists, language lexicons, vocabularies
- Terms with definitions. Dictionaries, glossaries, thesauri
- Terms with information. Gazetteers, frequency lists, confidence lexicons

Lexicon-based NLP

Dictionary

- A list of terms along with their definitions, grammatical information, and more
- Could be used to compare term meaning



The New Oxford American Dictionary

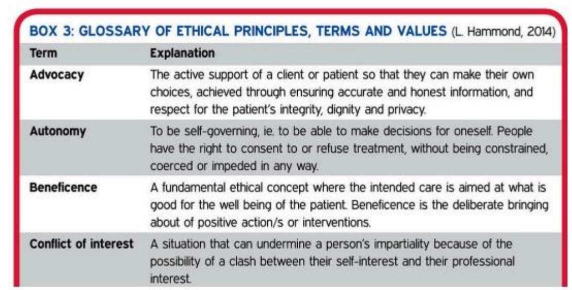
lit·er·a·cy /ˈlitərəsē; ˈlitrə-/
▪ *n.* the ability to read and write.
‹SPECIAL USAGE› competence or knowledge in a specified area: *wine literacy can't be taught in three hours.*
‹ORIGIN› late 19th cent.: from LITERATE, on the pattern of *illiteracy*.

Linked entries
LITERATE

http://flickr.com

Glossary

- A vocabulary with term definitions
- Could be used to compare term meaning



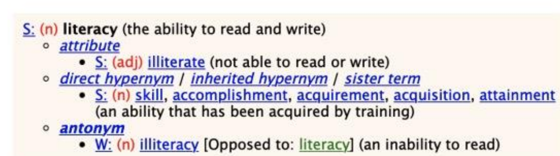
BOX 3: GLOSSARY OF ETHICAL PRINCIPLES, TERMS AND VALUES (L. Hammond, 2014)

Term	Explanation
Advocacy	The active support of a client or patient so that they can make their own choices, achieved through ensuring accurate and honest information, and respect for the patient's integrity, dignity and privacy.
Autonomy	To be self-governing, ie. to be able to make decisions for oneself. People have the right to consent to or refuse treatment, without being constrained, coerced or impeded in any way.
Beneficence	A fundamental ethical concept where the intended care is aimed at what is good for the well being of the patient. Beneficence is the deliberate bringing about of positive action/s or interventions.
Conflict of interest	A situation that can undermine a person's impartiality because of the possibility of a clash between their self-interest and their professional interest.

http://flickr.com

Thesaurus

- A dictionary of synonyms, with (possibly hierarchical) information on related terms
- Used e.g. to find similar terms



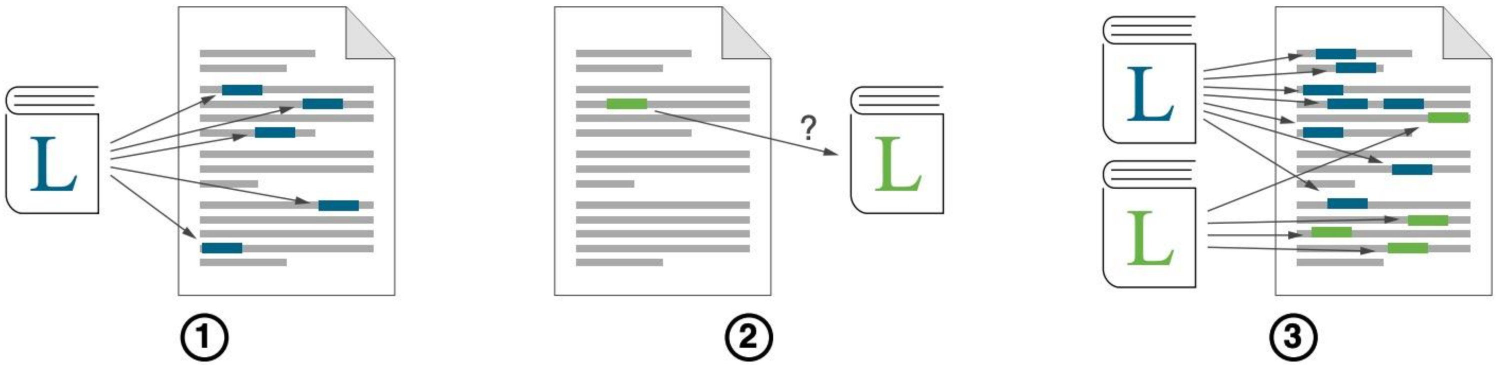
S: (n) literacy (the ability to read and write)
- attribute
  - S: (adj) illiterate (not able to read or write)
- direct hypernym / inherited hypernym / sister term
  - S: (n) skill, accomplishment, acquirement, acquisition, attainment (an ability that has been acquired by training)
- antonym
  - W: (n) illiteracy [Opposed to: literacy] (an inability to read)

http://princeton.edu

Lexicon-based NLP

Lexicon matching

- The identification of concepts in natural language texts, each being represented by a lexicon
- This requires to decide when a matching term refers to a concept.
- Main goals include to extract concept instances or to assess texts.



When to use lexicon matching?

1. A given lexicon can be used to find all term occurrences in a text.
2. The existence of a given term in a lexicon can be checked.
3. The density or distribution of vocabularies in a text can be measured.

Lexicon-based NLP

Benefits:

Lexicon matching is particularly reliable for unambiguous terms.
For entity types such as location names, huge gazetteer lists exist.
The idea of matching a lexicon is well-explainable.

Limitations:

Information that is not in the employed lexicons can never be found.
Ambiguous terms require other methods for disambiguation.
Composition of related information is hard to model with lexicons.

Different ways to do NLP

NLP using rules.

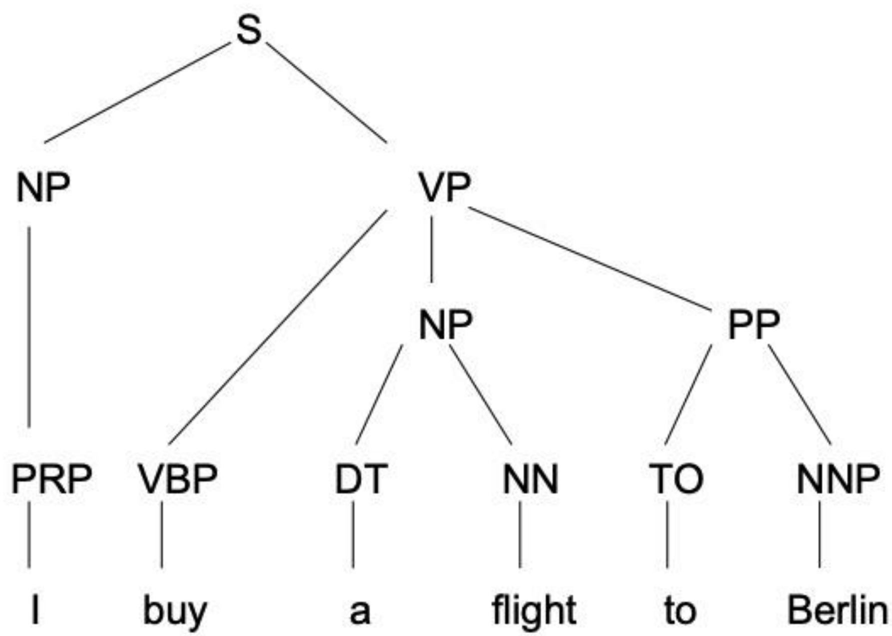
NLP using lexicons.

NLP using context-free grammars.

NLP using language models.

Context-free grammars

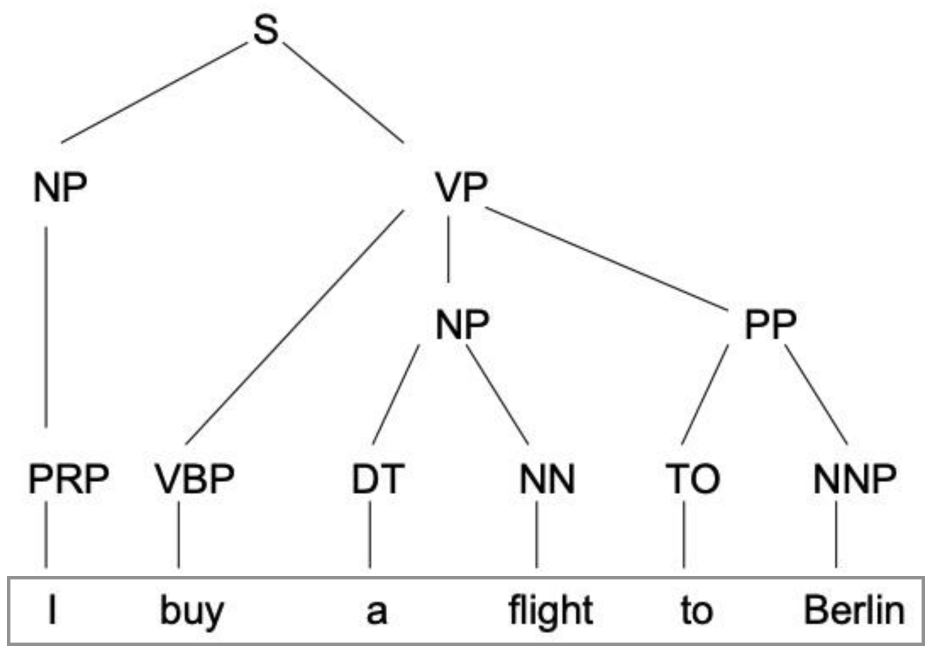
A grammar *G* consists of terminal nodes *T*, non-terminals *N*, a start symbol *S* and rules *R.*



Context-free grammars

A grammar *G* consists of terminal nodes *T*, non-terminals *N*, a start symbol *S* and rules *R.*

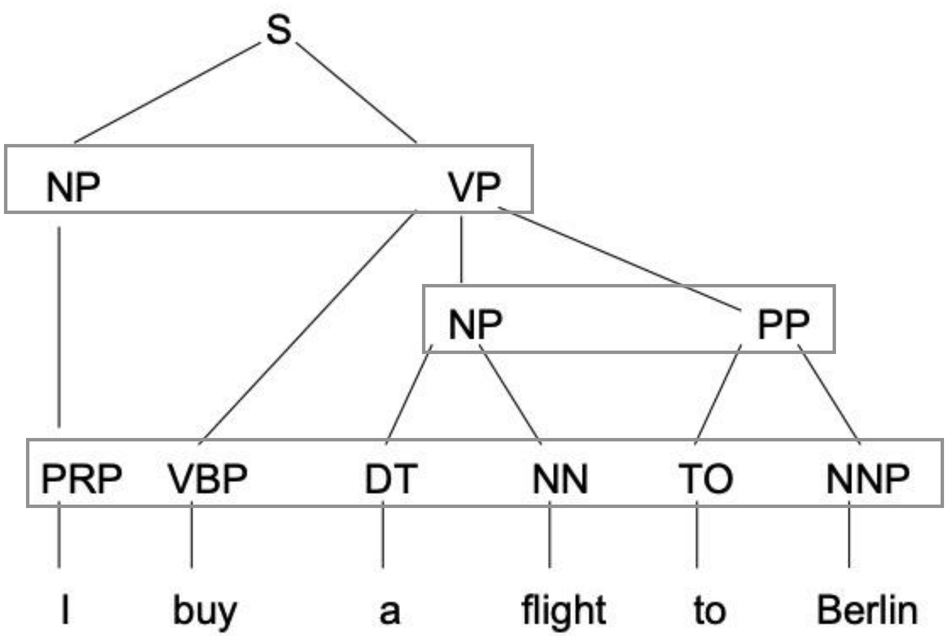
Terminals are the set of words in a sentence.



Context-free grammars

A grammar *G* consists of terminal nodes *T*, non-terminals *N*, a start symbol *S* and rules *R.*

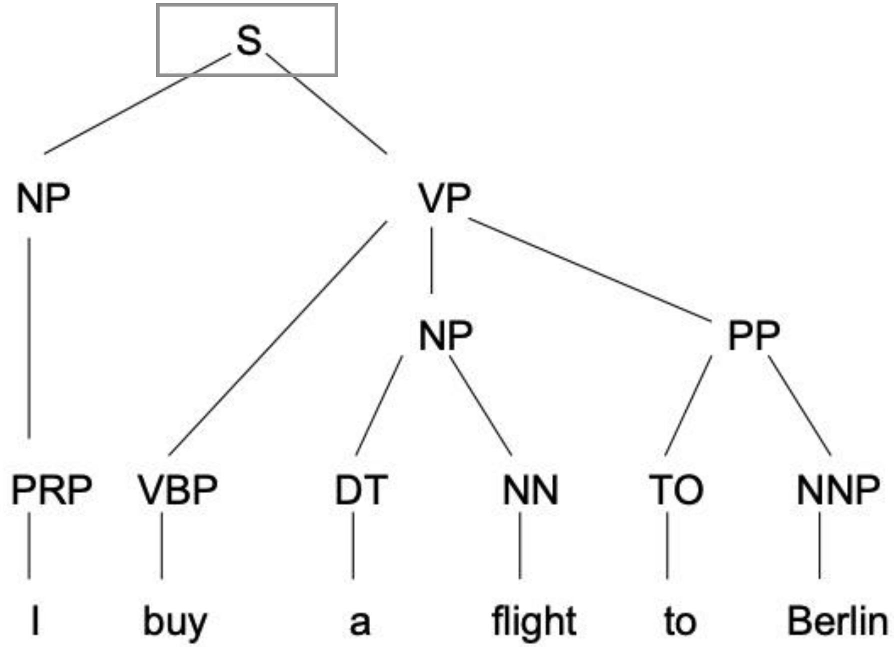
Non-terminals are the constituents in a sentence.



Context-free grammars

A grammar *G* consists of terminal nodes *T*, non-terminals *N*, a start symbol *S* and rules *R.*

The start symbol is the main constituent in a sentence.

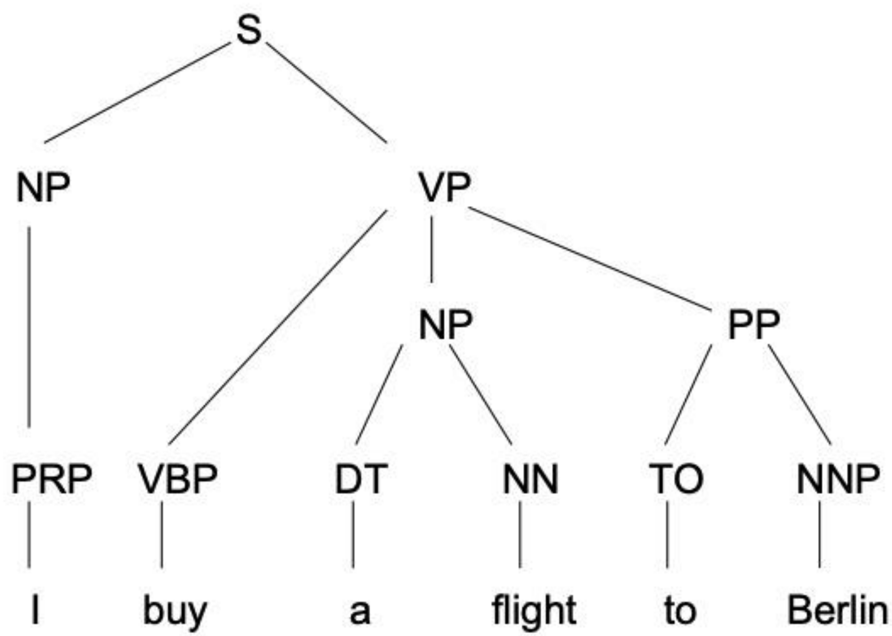


Context-free grammars

A grammar *G* consists of terminal nodes *T*, non-terminals *N*, a start symbol *S* and rules *R*.

Equations of the form $A \rightarrow a$,

e.g.,
S → NP VP
VP → VBP NP PP



Context-free grammars

A context-free grammar (CFG) is a formal grammar whose production rules are of the form $A \rightarrow a$, with A being a single nonterminal symbol and a a string of terminals and nonterminals.

The grammar is context-free if the production rules can be applied regardless of the context of a nonterminal.

Parses result from wide range of grammatical formalisms, e.g.

- phrase structures: ordered, labelled trees that express hierarchical relations among (groups of) words → constituency parsing
- dependency structures: binary grammatical relations between words in a sentence → dependency parsing

Parsing in NLP

Parsing is a central part in natural language processing systems

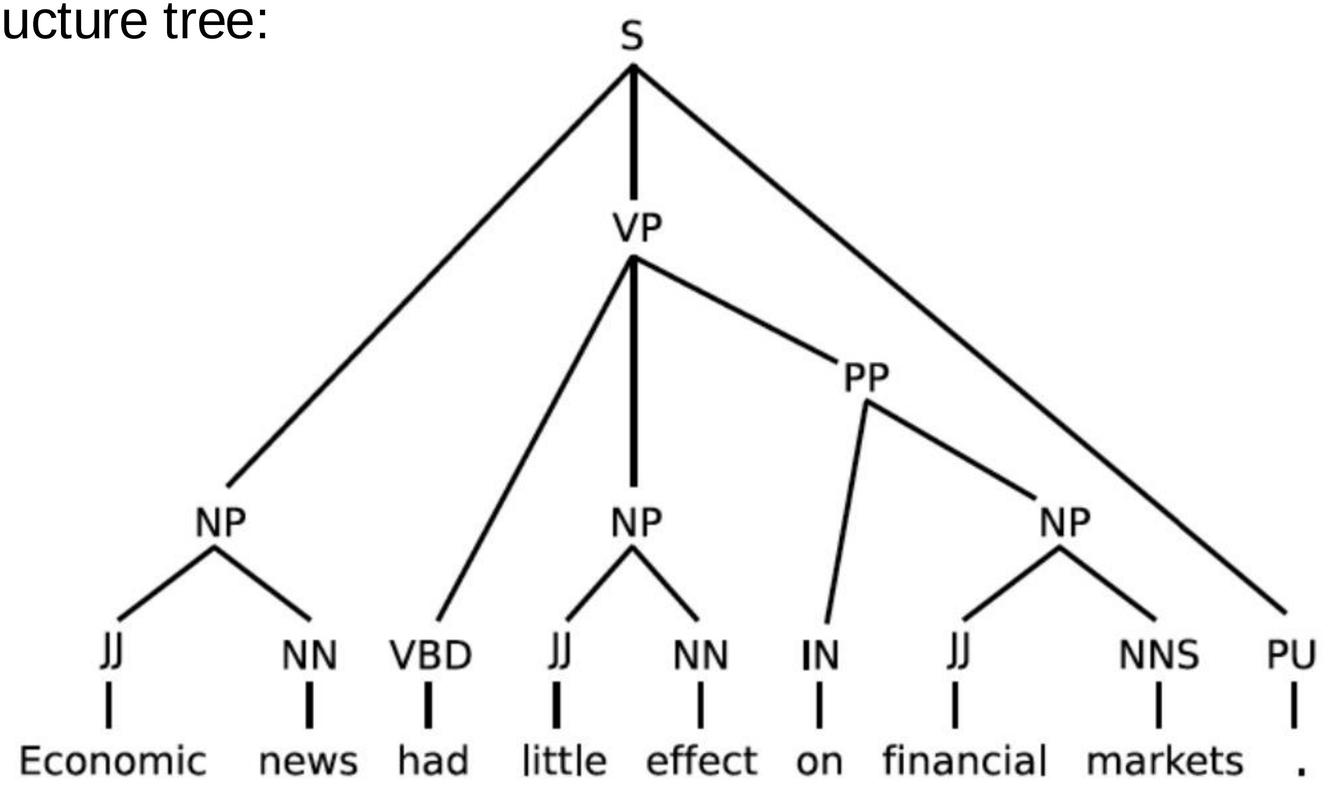
- The parsing accuracy impacts the success of an application as a whole
- The set of parses for a given input sentence is typically very large
  - → syntactic disambiguation
  - → effective storage

- Parses might be weighted by numerical rule values and combinations

- Most commonly: Constituency parsing and dependency parsing

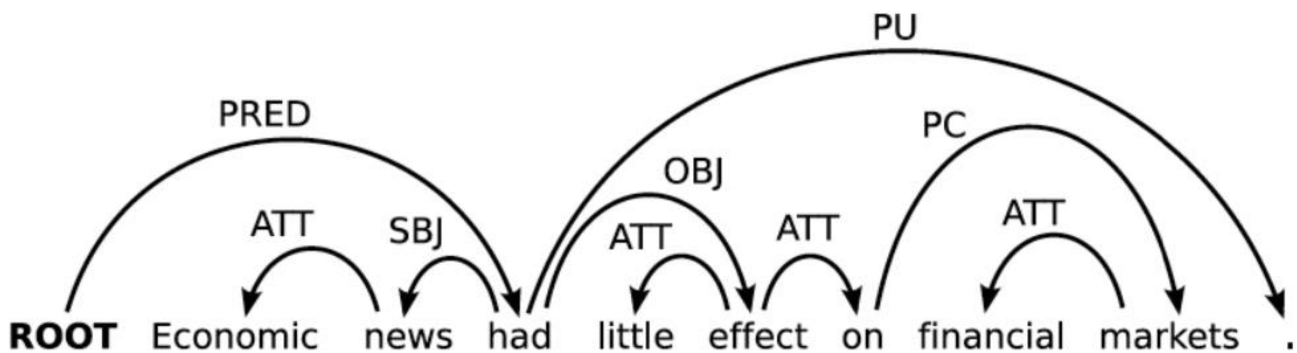
Parsing in NLP

A phrase structure tree:



Parsing in NLP

A dependency structure:



Parsing in NLP

Find structural relationships between words in a sentence.

```
(ROOT
  (S
    (NP
      (NP (NNP Steve) (NNP Paul) (NNP Jobs))
      (, ,)
      (NP
        (NP (NN co-founder))
        (PP (IN of)
          (NP (NNP Apple) (NNP Inc))))
      (, ,))
    (VP (VBD was)
      (VP (VBN born)
        (PP (IN in)
          (NP (NNP California)))))
    (. .)))
```

(http://nlp.stanford.edu:8080/parser)

Parsing in NLP

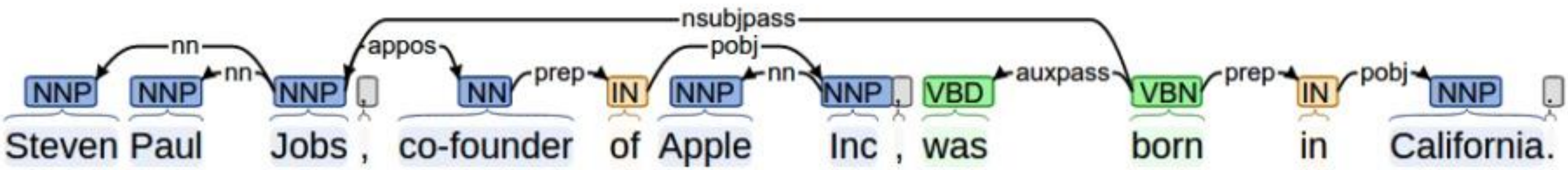
One application: Grammar checking.

```
(ROOT
  (FRAG
    (NP (PRP We))
    (PP (IN by)
      (NP (DT a) (JJ new) (NN book)))))
```

```
(ROOT
  (S
    (NP (PRP We))
    (VP (VBP buy)
      (NP (DT a) (JJ new) (NN book)))))
```

Parsing in NLP

Another application: Relation extraction.



(http://nlp.stanford.edu:8080/corenlp/)

Context-free grammars

$S \rightarrow NP\ VP$

$S \rightarrow VP$

$NP \rightarrow NN$

$NP \rightarrow PRP$

$NP \rightarrow DT\ NN$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP\ PP$

$VP \rightarrow VBP\ NP$

$VP \rightarrow VBP\ NP\ PP$

$VP \rightarrow VP\ PP$

$VP \rightarrow VP\ NP$

$PP \rightarrow TO\ NNP$

$PRP \rightarrow$ I

$NN \rightarrow$ book

$VBP \rightarrow$ buy

$DT \rightarrow$ a

$NN \rightarrow$ flight

$TO \rightarrow$ to

$NNP \rightarrow$ Berlin

Probabilistic context-free grammars

0.9	S → NP VP
0.1	S → VP
0.3	NP → NN
0.4	NP → PRP
0.1	NP → DT NN
0.2	NP → NP NP
0.1	NP → NP PP
0.4	VP → VBP NP
0.3	VP → VP PP
0.5	VP → VP NP
1.0	PP → TO NNP

1.0	PRP → I
0.6	NN → book
0.7	VBP → buy
0.8	DT → a
0.4	NN → flight
1.0	TO → to
1.0	NNP → Berlin

Probabilistic context-free grammars

Generated from a treebank, i.e., a corpus in which each sentence has been paired with a parse tree.

These are generally created by

- parsing the collection with an automatic parser
- correcting each parse by human annotators if required

The Penn Treebank is the most widely used treebank in English.

Most well-known section is the Wall Street Journal Section with 1 M words from 1987-1989.

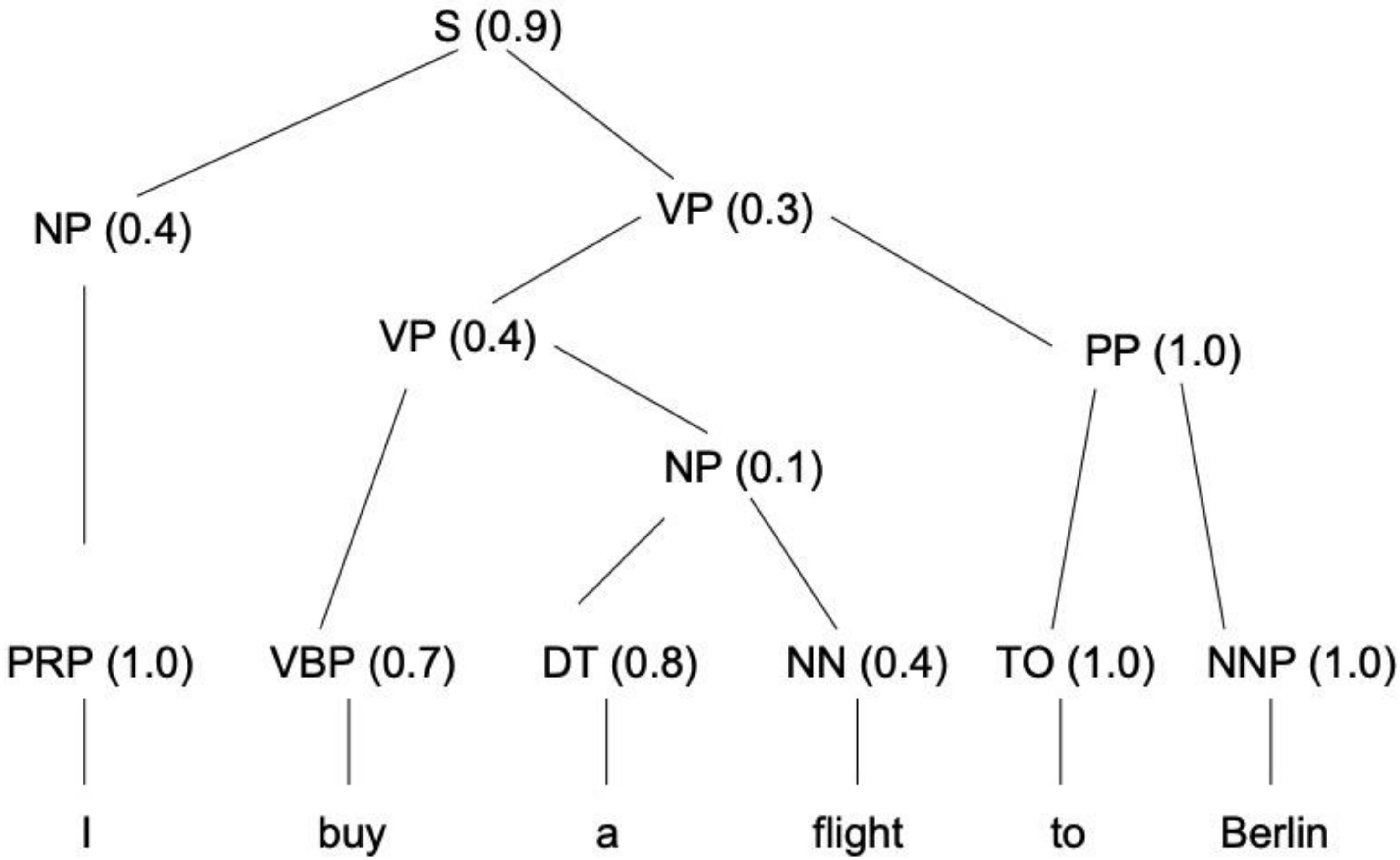
Statistical parsing

Considering the corresponding probabilities while parsing a sentence.

Selecting the parse tree which has the highest probability.

P(t): the probability of a tree t = Product of the probabilities of the rules used to generate the tree.

Statistical parsing



$$P(t) = 0.9 \times 0.4 \times 1.0 \times 0.3 \times 0.4 \times 0.7 \times 0.1 \times 0.8 \times 0.4 \times 1.0 \times 1.0 \times 1.0$$

Ambiguity is pervasive.

How does the constituency tree capture the ambiguity?

The saw the man with the telescope.

He saw the man with the telescope.

Different ways to do NLP

NLP using rules.

NLP using lexicons.

NLP using context-free grammars.

NLP using language models.

NLP using language models

Example: Next words

- Given the following sequence of words:

ChatGPT is based on a neural language __________

- Which of the following is the most likely next word?

that | model | learning | language | ...

Example: Probabilities of word sequences

- Given the following two sequences of words:

language models have become a key technique in NLP

NLP models language in key have become a technique

- Which of them seems more likely?

NLP using language models

Language model (LM)

- A language model represents a probability distribution over sequences of tokens, $s = (w_1, \ldots, w_k)$, with $k \geq 1$.
- It thus defines the probability $P(s)$ of any token sequence s.
- Also, it assigns a probability $P(w_{k+1}|s)$ to any next token w_{k+1} after s.

Where do the probabilities come from?

- $P(s)$ can be approximated by the relative frequency of s in a corpus.
- For longer s, $P(s)$ may be unreliable (or even 0) due to data sparsity.

n-gram language model

- An n-gram LM derives the probability of s from the probability of all token sequences of length n contained in s.
- $n \geq 1$ is a predefined hyperparameter of the LM.
- The larger n, the more data is needed to get reliable estimations $P(s)$.

NLP using language models: Challenges

Vanishing probabilities

- In real-world data, the probability of most token sequences s is near 0, which may lead to vanishing probabilities.
- A way to deal with this problem is to use *log probabilites*.

Unknown words and sequences

- Some tokens may never appear in a training corpus.
- Even if all tokens are known, there will always be sequences s that do not appear in a training corpus but in other data.
- A technique used to deal with these problems is called *smoothing*.

Exactness vs. generalization

- The higher n, the more exact the estimated probabilities.
- Sometimes, less context (i.e., a lower n) may aid generalization.

NLP using language models: Applications

When to use LMs?

- Probabilities of token sequences are essential in any task where tokens have to be inferred from ambiguous input.
- Ambiguity may be due to linguistic variations or due to noise.
- LMs are a key technique in generation, but are also used for analysis.

Selected applications

- Speech recognition. Disambiguate unclear words based on likelihood.

  wreck a nice beach | recognize speech

- Spelling/Grammar correction. Find likely errors and suggest alternatives.

  I booked one and Tim booked too | I booked one and Tim booked two

- Machine translation. Find likely interpretation/order in target language.

  爱国人 → love country human → country loving human

IBM’s Project Debater

Based on IBM’s Watson

Since 2018: *”When asked to discuss any topic, Watson can autonomously scan its knowledge database for relevant content, ‘understand’ the data, and argue both for and against that topic.”*

More information here: **https://www.research.ibm.com/artificial-intelligence/project-debater/**

See here for a brief overview.

What is IBM's Project Debater?

A system that brings AI to the art of debate.

Project Debater is the first AI system that can debate humans on complex topics. Project Debater digests massive texts, constructs a well-structured speech on a given topic, delivers it with clarity and purpose, and rebuts its opponent. Eventually, Project Debater will help people reason by providing compelling, evidence-based arguments and limiting the influence of emotion, bias, or ambiguity.

LEARN MORE

Some background details

Important components in the system:

- Some IBM Watson technology reused
- An in-house wikification tool:
  - Anchor lexical items to their Wikipedia page (plus any IBM wiki)
  - Use the links between Wikipedia pages to create an enormous knowledge graph (what information is related to which other bit of information, how?)
  - Only possible with mega-billion company infrastructure (technology and manpower)

Some background details

State-of-the-art machine learning techniques

- Training neural networks with strong labelled data (human annotations) and weakly labelled data (e.g., Debatepedia)
- Possible with above-mentioned infrastructure

Some background details

Apart from that:

- Lexicon-based stance classification
- New text-to-speech system to facilitate a continuous and persuasive speech for a few minutes
- The usual NLP techniques
  - Named-entity recognition
  - WordNet
  - Sentiment Analysis/Opinion Mining

Some background details

How the system was presented to the public:

- w/ two debating experts (Israelian champions in a debating competition)
  - know how to structure their speech
  - know how to argue
  - know the system and the structures it performs good at

- The debating experts contributed to developing the system
  - system was trained on their voices
  - multiple training runs to learn speaker characteristics

- System was presented for one of a relatively small set of possible topics

Read a scientific publication

The latest (and most high-profile) publication of IBM’s Project Debater is in *Nature*.

Noam Slonim et al. 2021. An autonomous debating system. Nature, Vol. 591, pp. 379-385. (also in course materials)

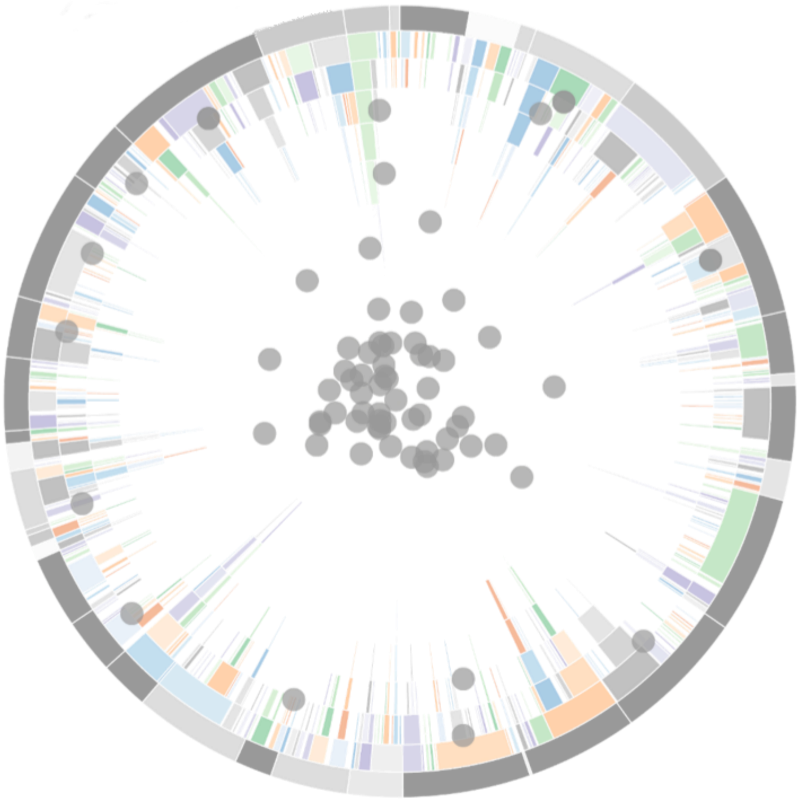
Read the paper, try to the understand it and prepare at least two questions on it, i.e., where does the paper remain unclear in your opinion, what additional information would you request, etc.? We will discuss (and answer them) in class.

Cluster of Excellence
The Politics of Inequality

Steinbeis Transfer Center
Linguistic Data Analysis

UNIVERSITÄT PASSAU

Thank you.
Questions?
Comments?



Annette Hautli-Janisz, Prof. Dr.
cornlp-teaching@uni-passau.de