VL Deep Learning for Natural Language Processing

3. Introduction to Natural Language Processing

Prof. Dr. Ralf Krestel

AG Information Profiling and Retrieval

Language and Meaning

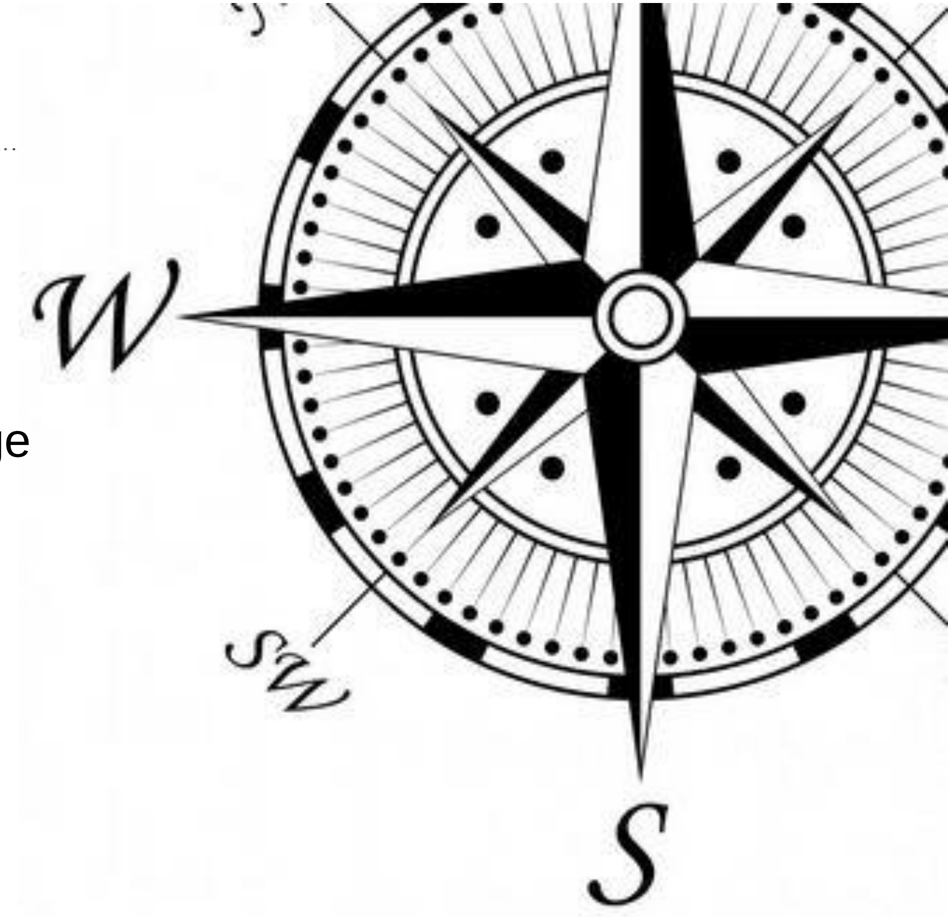
https://i.pinimg.com/originals/f8/54/6b/f8546b1135bc2549e4273c361a6a8822.jpg

Lerning Goals for this Chapter

- Understand the basic questions of Philosophy of Langauge
- Know about Zipf‘s and Heap‘s law
- Describe standard NLP pipeline
- Know common NLP tasks and be able to describe them formally
- Be able to discuss challenges and potential for deep learning regarding the standard NLP tasks

Topics Today

1. **Philosophy of Language**
2. (Statistical) Characteristics of Language
3. Natural Language Processing Pipeline



https://upload.wikimedia.org/wikipedia/en/b/b9/MagrittePipe.jpg

La trahison des images
René Magritte, 1929

Philosophy of Language

- Classical antiquity
  - Platoon
    - Theory of forms
    - Predication
  - Aristoteles
    - Sitional calculus
- Modern
  - Gottlob Frege
    - Modern Logic
  - Wilard Van Orman Quine
  - Bertrand Russell
  - Ludwig Wittgenstein
- Middle Ages
  - Abaelard
  - Duns Scotus
  - Wiliam of Ockham
    - Nominalism
- Becomes its own discipline around 1900
  - Analytic Philosophy
  - Up to then: Language as intermediary between reality and conciousness

Analysis of Language as Philosophical Method

- „Linguistic turn“
  - Richard Rorty describes this as "the view that philosophical problems can be solved or resolved either by reforming language or by better understanding the language we currently use.”
- Philosophy of ideal language
  - Natural languages are deficient (various inaccuracies)
  - Do not satisfy the strict requirements of logic
  - The goal of this approach is to revise or even replace natural languages for purposes of science with an ideal, formal language.
- Philosophy of normal language
  - Natural languages are not deficient
  - Completely useful for the purpose for which they are used, namely, for communication in the social environment
  - The task of philosophy of language is to describe or explain by pointing out conceptual or regulative relations.

https://sites.psu.edu/franceskblanchette/files/2019/11/JuryOfEnglishMajors.jpg

Consciousness – Language – Reality

- Consciousness – Language
  - Language aquisition (Chomsky, Piaget)
  - Communication (Bühler, Shannon)
  - Hermeneutic (Schleiermacher, Heidegger, Gadamer)
  - Semantic relativism (Sapir, Whorf)

- Language – Reality
  - Reference (Frege, Russell, Strawson, Kripke)
  - Meaning (Frege, Wittgenstein, Quine)

- Language – Action
  - Speech acts (Austin, Searle)
  - Implicature (Grice)

Reference (Extension & Intension)

- There are referring expressions: The name "Socrates" refers to the Greek philosopher.
- The referential theory of meaning states that the meaning of an expression consists in its reference.
- The meaning of ambiguous words (e.g. "bank") can be explained by this: The extensionality thesis states that terms are completely determined by their extensional domain. (Obviously, the set of all seats is a different set than the set of all financial institutions).
- Problem: "The evening star is the morning star".
  - The expression "evening star" and the expression "morning star" have the same reference, namely the planet Venus, but the first expression denotes the brightest star in the evening, the second the brightest star in the morning.
  - Nevertheless, it seems plausible that the one who thinks of the evening star uses a different term than the one who thinks of the morning star. The difference lies, according to Frege, not in the extension, but in the way of referring to the denoted object, i.e. the intension.

Meaning

- Traditional theories of meaning assume that meaning is used to denote an object.
- Problem:
  - Sentences containing expressions that do not refer to anything.
    - E.g.: "Pegasus is a winged horse" -> would not have any meaning
  - In addition, there are many expressions, such as conjunctions and prepositions, which do not seem to refer to anything.
- Modern theories of meaning in the spirit of the philosophy of ordinary language ask how it comes about that a sign has meaning at all.
  - Meaning of an expression is not an object, but determined by the use of the sign
    - Merely a description of language, not an explanation (Wittgenstein)
  - Replacement of the concept of meaning by the term verification:
    What a proposition means is determined by how it is checked (verified) as to its truth.

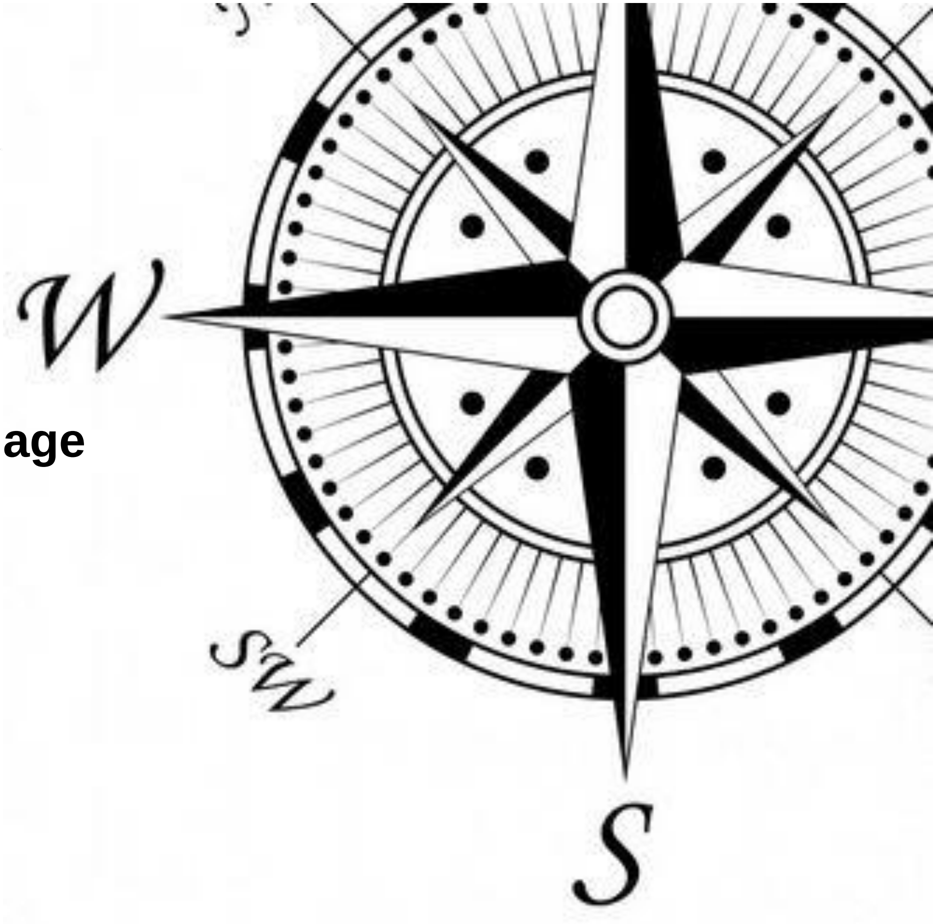
Exercise

- What is the meaning of the following sentences? Which of the sentences are "true"? Which of them are false? Which ones are "neither true nor false"? Which ones are meaningless? Which ones are incomprehensible? Why? Which propositions are analytical definitions and say nothing about the empirical world? Which propositions say something about the empirical world? ...

- Trees are plants with roots, a trunk, branches and leaves or needles.
- There is a cherry tree in our garden.
- Trees have a soul.
- We have not seen the forest for the trees.
- Trees treeing the world.
- Climb the tree!
- The tree in love dances.
- Conifers are ugly.
- We must not cut down these trees!
- Trees convert CO2 into cellulose with the help of sunlight and water. Oxygen remains as "waste".
- Trees are symbols of life.
- A lone tree stood by the side of the road.
- "Tree" is a subordinate term to "plant". And plant is a subordinate term to "living being". So a tree is a living thing.
- "Tree" is a noun.

Start 5 4 3 2 1 End

Topics Today



1. Philosophy of Language
2. **(Statistical) Characteristics of Language**
3. Natural Language Processing Pipeline

Modern Languages

- Very large vocabulary
  - Duden contains ~150k German words
    - + situational creations + compounds
- English has the most words
  - Multiple words for the same thing

- Active vs. passive
  - Speaking vs. understanding

- Some words are "more important" than others
  - Frequency of occurrence
  - Basic idea in information retrieval (tf*idf)

https://img.welt.de/img/kultur/mobile167820245/0681627457-ci23x11-w1136/Duden-Das-Wort-Arschrunzeln.jpg

Zipf's Law I

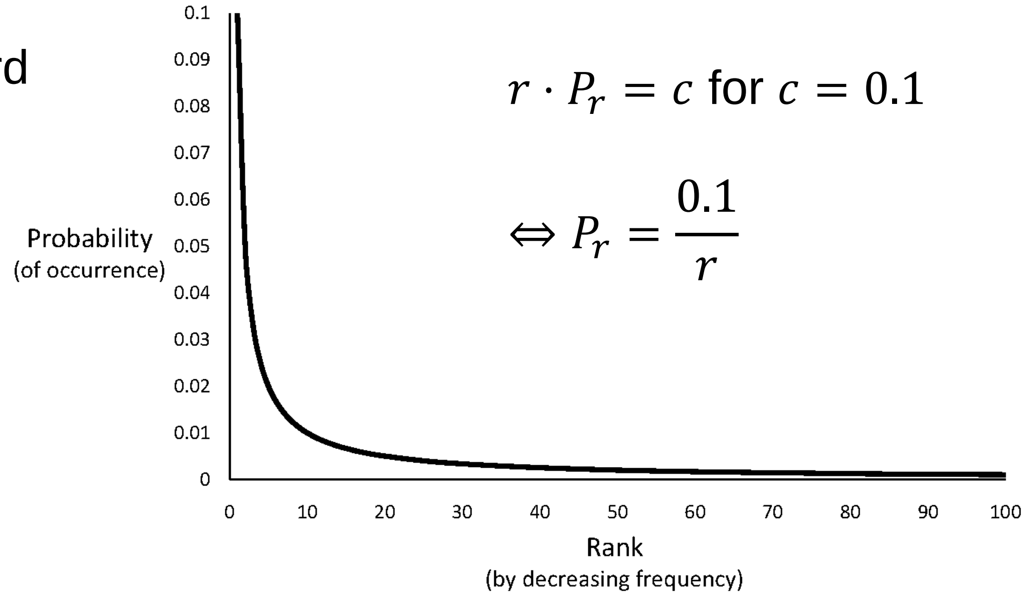
- Distribution of word frequencies is very skewed
  - A few words occur very often, many words hardly ever occur
  - Two most common words ("the", "of") make up about 10% of all word occurrences in text documents
  - Top 6 words account for 20% of text.
  - Top 50 words account for 40% of text.
  - And: 50% of all words in a large sample occur only once.

George Kingsley Zipf
(1902–1950)

http://www.lib.jgypk.u-szeged.hu/alknyelv/idegenek/klasszikusok/Zipf/nyelvesz.htm

Zipf's Law II

- Zipf's "law":
  - Observation that **rank *r*** of a word times its **frequency *f*** is approximately a **constant *k***
  - Assuming words are ranked in order of decreasing frequency
  - $r \cdot f \approx k$ or $r \cdot P(w_r) \approx c$
  - where $P(w_r)$ is occurrence probability of word *w* with rank *r*
  - and $c \approx 0.1$ for English



$r \cdot P_r = c$ for $c = 0.1$

$$\Leftrightarrow P_r = \frac{0.1}{r}$$

http://www.wikimedia.org

Example News Collection (TREC AP89)

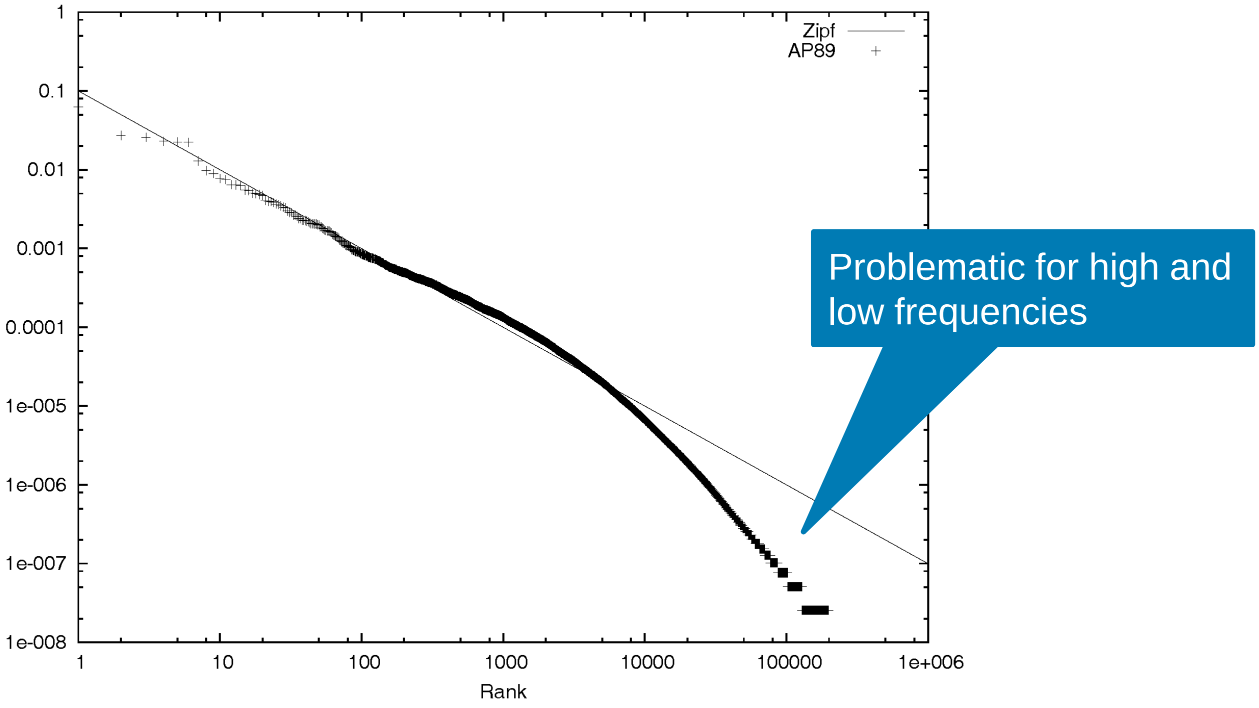
- Collection of Associated Press articles from 1989

Number of Documents	84,678
Number of Words	39,749,179
Vocabulary Size	198,763
Words Appearing More Than 1000 Times	4,169
Words Appearing Exactly Once	70,064

Word	*Freq.*	r	$P_r(\%)$	$r.P_r$	*Word*	*Freq*	r	$P_r(\%)$	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Zipf‘s Law: TREC AP89

- Log-Log-Graph



Heap‘s Law

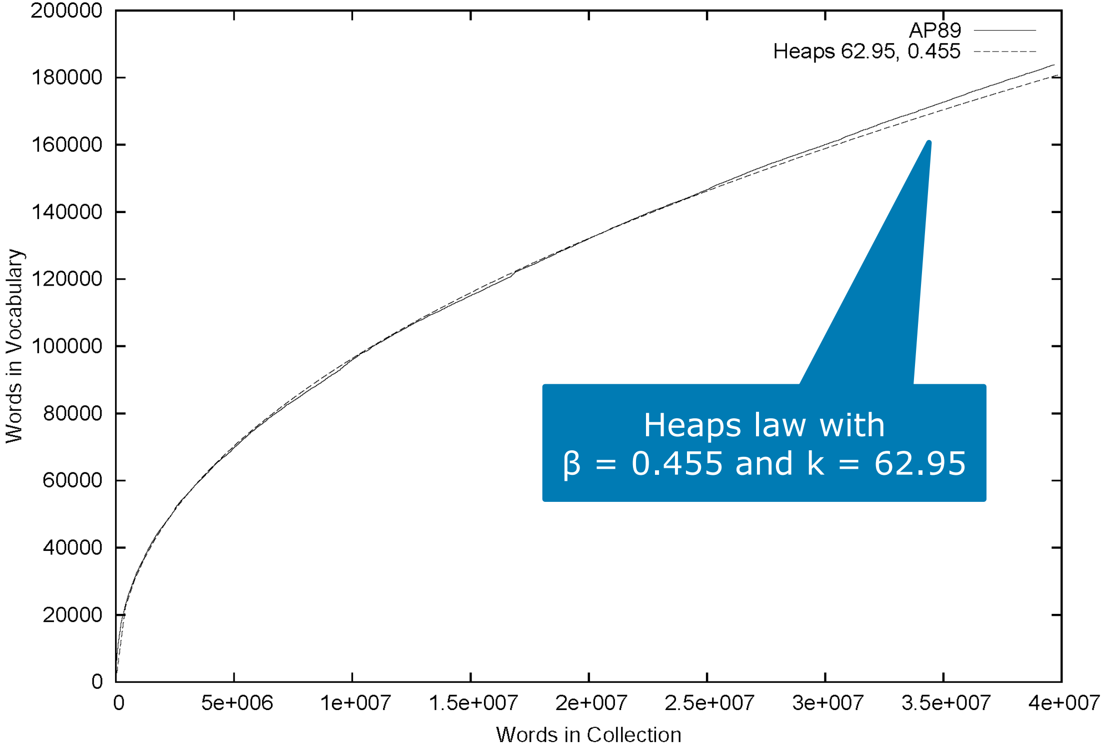
- As **corpus grows**, so does **vocabulary size**
  - But: Fewer new words when corpus is already large
- Observed relationship (*Heaps’ Law, found empirically*):

$$V = k \cdot N^{\beta}$$

$$\log V = b \cdot \log N + \log k$$

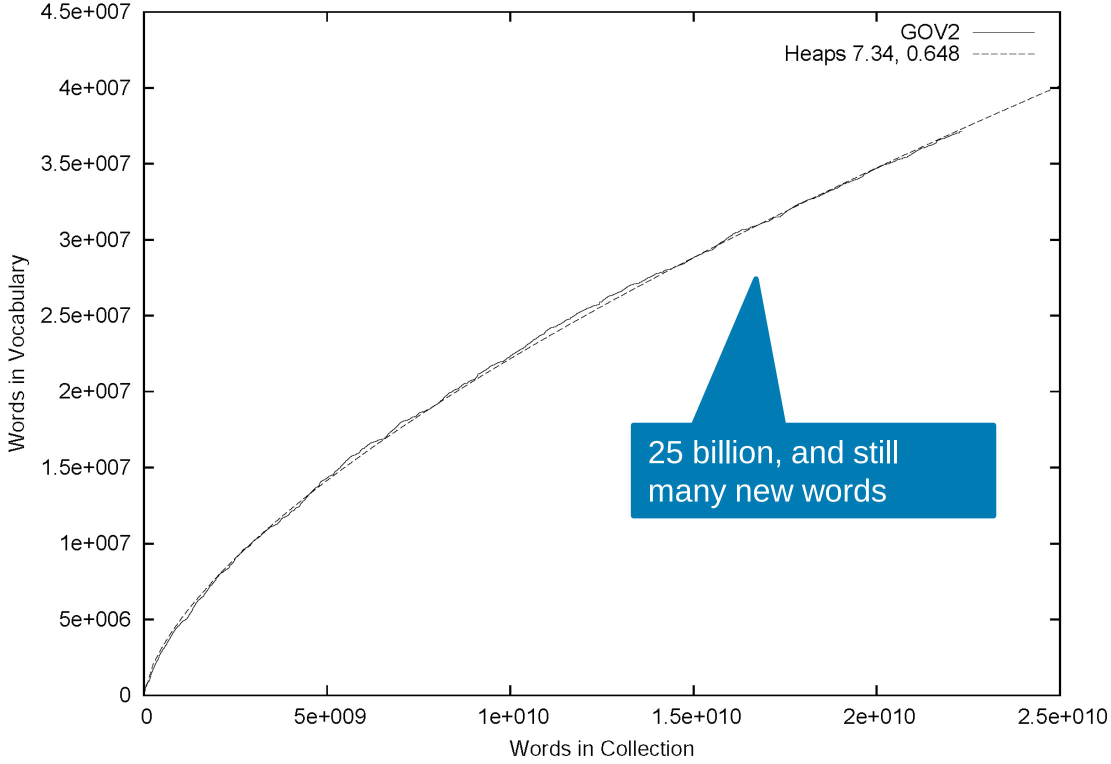
  - where V is vocabulary size (number of unique tokens)
  - N is the number of tokens in corpus (non-unique)
  - k, β are parameters that vary for each corpus
  - β: how fast the vocabulary size increases as the corpus grows
  - typical values given are $10 \leq k \leq 100$ and $\beta \approx 0.5$
- Example
  - $n = 1{,}000{,}000$ $\quad k = 50$ $\quad \beta = 0.5$ $\qquad v = 50 \cdot 1{,}000{,}000^{0.5} = 50{,}000$

Heap‘s Law: TREC AP89



Heaps law with
β = 0.455 and k = 62.95

Heap's Law: Web Corpus GOV2



http://www.wikimedia.org

Exercise

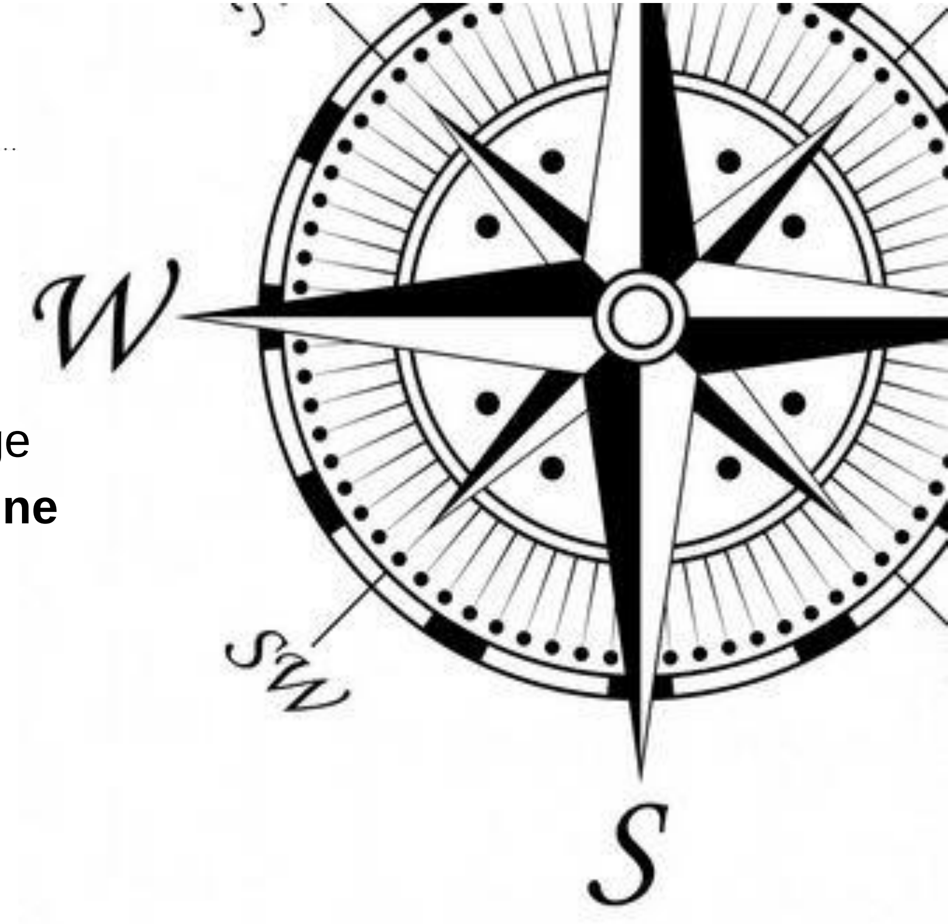
- What would be the population of Stuttgart according to Zipf‘s law?

Rank r	City	Population f
1	Berlin	3 669 000
2	Hamburg	1 847 000
3	München	1 484 000
4	Köln	1 088 000
5	Frankfurt	763 000
6	Stuttgart	636 000

Start 5 4 3 2 1 End

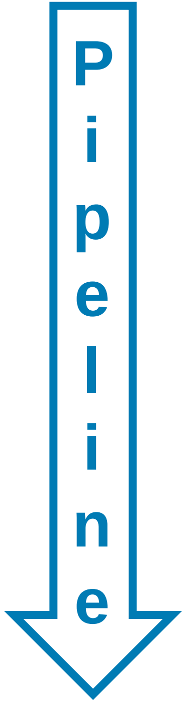
Topics Today

1. Philosophy of Language
2. (Statistical) Characteristics of Language
3. **Natural Language Processing Pipeline**



Common NLP Tasks & (Text Mining) Applications

- Preprocessing
  - OCR, speech recognition
  - Tokenization
  - Normalization
- Morphological analysis
  - Stemming, lemmatization
  - Part-of-speech tagging
- Syntactic analysis
  - Sentence splitting
  - Parsing
- Semantic analysis
  - Lexical semantics
  - Relational semantics
  - Discourse



- (Text Mining) Applications
  - Document Classification
  - Document Clustering
  - Machine translation (MT)
  - Information retrieval (IR)
    - Information extraction (IE)
    - Question answering (QA)
    - Automatic summarization
    - Recommender Systems (RS)
  - Natural language generation (NLG)
  - Natural language understanding (NLU)

NLP History

- **Symbolic NLP (1950s–early 1990s)**
  - John Searle's Chinese room experiment: Given a collection of rules, the computer emulates natural language understanding (or other NLP tasks) by transforming the input into output applying those rules.
  - Requires complex sets of **hand-written rules**
- **Statistical NLP (1990s–2010s)**
  - "statistical revolution“
  - Introduction of machine learning (supervised, semi-supervised, and unsupervised)
  - Heavy **feature engineering** necessary
- **Neural NLP (2010s–present)**
  - representation learning
  - deep **neural networks**

Preprocessing

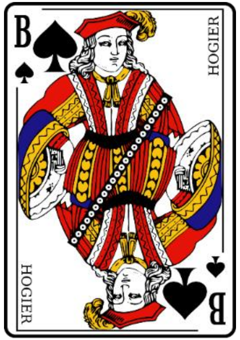
- **OCR, speech recognition**
  - Generate/Extract text from image or audio files
- **Tokenization**
  - Aka word segmentation
  - Forming words from sequence of characters
  - Surprisingly complex in English, can be harder in other languages
  - Basic assumption: any sequence of alphanumeric characters of length > 3
- **Normalization**
  - Changing any upper-case letter to lower-case
    - aka. case-folding, lower casing, or downcasing
- Example:
  - “Bigcorp’s 2007 bi-annual report showed profits rose 10%.”
  - becomes “bigcorp 2007 annual report showed profits rose”

Tokenization: Issues I

- **Small words** can be important in some queries, usually in combinations
  - xp, ma, pm, ben e king, el paso, system r, master p, gm, j lo, world war II
- Both **hyphenated** and non-hyphenated forms of many words are common
  - Sometimes hyphen is not needed
  - e-bay, wal-mart, active-x, cd-rom, t-shirts
- Sometimes **hyphens** should be considered either as part of the word or a word separator
  - winston-salem, mazda rx-7, e-cards, pre-diabetes, t-mobile, spanish-speaking
- **Numbers** can be important, including decimals
  - MH 370, nokia 3250, top 10 courses, quicktime 6.5 pro, 92.3 the beat, 24103
- **Periods** can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., cs.umass.edu, F.E.A.R.

Tokenization: Issues II

- **Special characters** are an important part of tags, URLs, code in documents, …
- **Capitalized** words can have different meaning from lower case words



Why are there lower and UPPER case letters?

https://www.youtube.com/watch?v=9-clrKOp5Co

- **Apostrophes** can be a part of a word, a part of a possessive, or just a mistake
  – rosie o’donnell, can’t, don’t, 80’s, 1890’s, men’s straw hats, master’s degree, england’s ten largest cities, shriner’s

http://www.wikimedia.org
https://www.welt.de/img/newsticker/news1/mobile144186224/3291624877-ci23x11-w1280/Donald-Trump-nimmt-kein-Blatt-vor-den-Mund.jpg

Tokenization: N-Grams

- Instead of single tokens, sequences of n words, so-called n-grams
  - **bigram**: 2 word sequence, **trigram**: 3 word sequence, **unigram**: single words
  - N-grams also used at character level for applications such as OCR
- N-grams typically formed from **overlapping** sequences of words
  - i.e., move n-word “window” one word at a time in document
- Frequent n-grams are more likely to be meaningful phrases
  - „President of the USA“, „Holstein Kiel“, „Porsche 911“, „all rights reserved“
- N-grams also form a Zipf distribution (better fit than words alone)
- Google N-Grams “All Our N-gram are Belong to You”
  - Tokens: 1,024,908,267,229 sentences: 95,119,665,584
  - Unigrams: 13,588,391
  - Bigrams: 314,843,401
  - Trigrams: 977,069,902
  - Tetragrams: 1,313,818,354
  - pentagrams: 1,176,470,663

Also useful for Chinese text

https://en.wikipedia.org/wiki/All_your_base_are_belong_to_us

Morphological Analysis

- Many **morphological variations** of words
  - **inflectional** (plurals, tenses)
  - **derivational** (making verbs nouns etc.)
- In most cases, these have the same or very similar meanings
- Introduce noise when (statistically) processing words

- Solution:
  - **Stemming**
  - **Lemmatization**

- Identifying the lexical class (part-of-speech) of a word
  - **Part-of-Speech tagging**

Stemming & Lemmatization

- **Stemmers** attempt to reduce morphological variations to a **common stem**
  - Usually involves removing suffixes
  - E.g. goes, going → go but went → went?
  - Algorithmic or dictionary-based

- **Lemmatizer**: reduce words to their root forms
  - E.g. goes, went, going, gone → go
  - **More expensive** than stemming

Part-of-Speech Tagging

- Words can be categorized by their meaning (semantic), by their form (morphological), or by their use in the sentence (syntactic).
- In English there are 9 types of words
  - noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, interjection
  - Further subdivision into subclasses
- Popular tag sets
  - Penn tag set (45 tags) ⇒ Penn Treebank
  - Brown tag set (87 tags) ⇒ Brown corpus
  - STTS: Stuttgart-Tübingen tag set (55 tags) ⇒ Tiger corpus

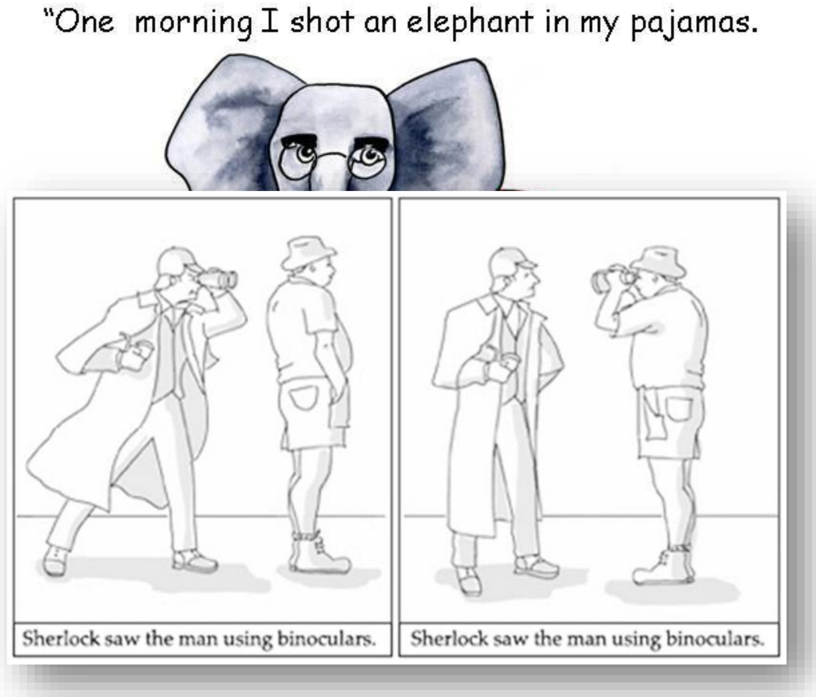
In German?

- Example:
  - My/PRP$ aunt/NN 's/POS can/NN opener/NN can/MD open/VB a/DT drum/NN

Syntactic Analysis

- **Sentence splitting**
  - Identifying sentence boundries
  - Very easy in English for the majority of cases
    - Simple rule: full stop followed by upper-case word
- Syntactic **parsing** (grammatical analysis)
  - Parsing: creating a parse tree from a sentence
  - Language is ambiguous
    - What is the meaning of „Fruit flies like an arrow.“?

Tricky cases?



https://www.pinterest.de/pin/the-marx-brothers--495747871456246303/

Poller, Olga. (2017). The descriptive content of names as predicate modifiers. Philosophical Studies. 174. 10.1007/s11098-016-0801-5.

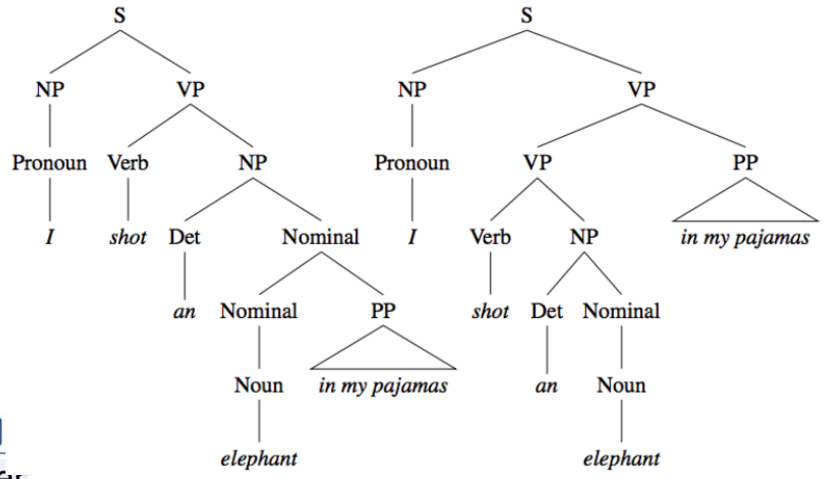
Parsing

- Shallow parsing („chunking“)

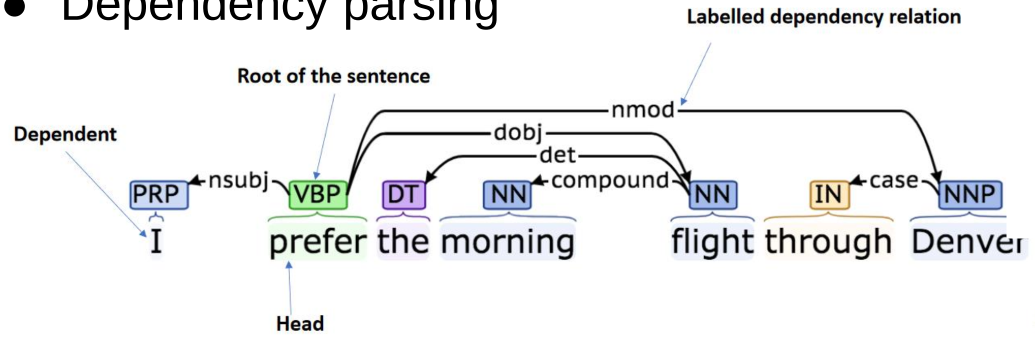
The morning flight from Denver has arrived.

NP PP NP VP

- Constituency parsing



- Dependency parsing



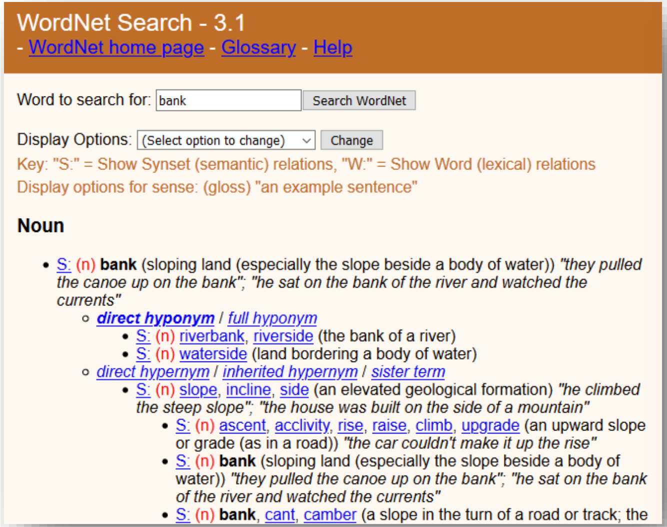
Kairit Sirts: https://courses.cs.ut.ee/LTAT.01.001/2021_spring/uploads/Main/Lecture10_2021_syntax.pdf

Semantic Analysis

- **Lexical semantics**
  - Semantics of individual words in context
  - Distributional semantics
    - How can we learn semantic representations from data?
- **Relational semantics**
  - Semantics of individual sentences
- **Discourse**
  - Semantics beyond individual sentences

Lexical Semantics

- **Word sense disambiguation (WSD)**
  - For ambiguous words, which meaning makes the most sense **in context**
    - E.g., with the help of a **lexical database** / dictionary (e.g., wordNet)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bank   Search WordNet

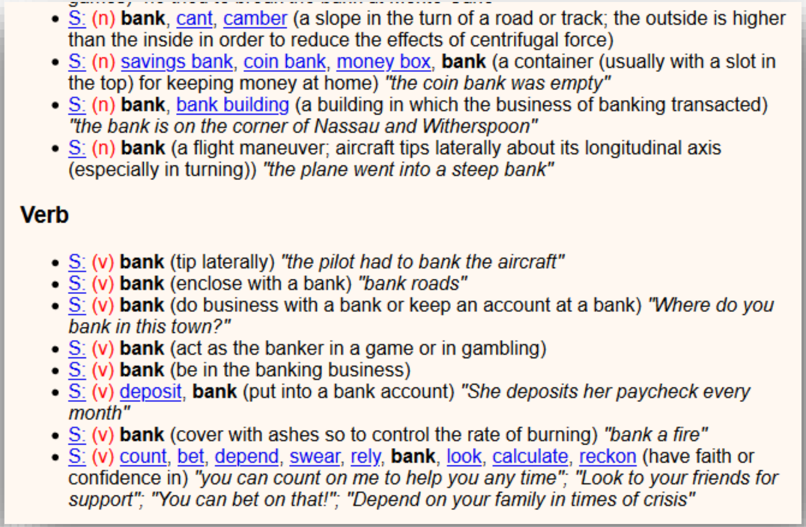
Display Options: (Select option to change)   Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
  - ***direct hyponym*** / *full hyponym*
    - S: (n) riverbank, riverside (the bank of a river)
    - S: (n) waterside (land bordering a body of water)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) slope, incline, side (an elevated geological formation) *"he climbed the steep slope"; "the house was built on the side of a mountain"*
      - S: (n) ascent, acclivity, rise, raise, climb, upgrade (an upward slope or grade (as in a road)) *"the car couldn't make it up the rise"*
      - S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
      - S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the



- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Verb

- S: (v) **bank** (tip laterally) *"the pilot had to bank the aircraft"*
- S: (v) **bank** (enclose with a bank) *"bank roads"*
- S: (v) **bank** (do business with a bank or keep an account at a bank) *"Where do you bank in this town?"*
- S: (v) **bank** (act as the banker in a game or in gambling)
- S: (v) **bank** (be in the banking business)
- S: (v) deposit, **bank** (put into a bank account) *"She deposits her paycheck every month"*
- S: (v) **bank** (cover with ashes so to control the rate of burning) *"bank a fire"*
- S: (v) count, bet, depend, swear, rely, **bank**, look, calculate, reckon (have faith or confidence in) *"you can count on me to help you any time"; "Look to your friends for support"; "You can bet on that!"; "Depend on your family in times of crisis"*

Lexical Semantics

- **Named entity recognition (NER) (includes NE typing)**
  - Which tokens map to proper names and what are their types
    - e.g., person, location, organization
- **Named entity linking (NEL) (includes NE disambiguation)**
  - Link the NE to an identifier, e.g., from a knowledge base
- **Terminology extraction**
  - Extract relevant terms from a given corpus
- **Sentiment analysis** (of words)
  - Extract subjective information based on the polarity of words
    - E.g., with the help of a sentiment lexicon (e.g., sentiWordNet)

Lexical Semantics

- **Named entity recognition (NER) (includes NE typing)**
  - Which tokens map to proper names and what are their types
    - e.g., person, location, organization
  - Example: [Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.
- **Named entity linking (NEL) (incl. disambiguation)**
  - Link the NE to an identifier, e.g., from a knowledge base (e.g., Wikipedia)

- Springfield, Alabama, unincorporated community
- Springfield, Arkansas
- Springfield, California
- Springfield, Colorado
- Springfield, Florida, a city in Bay County
- Springfield (Jacksonville), Florida, a neighborhood
- Springfield, Georgia
- Springfield, Idaho
- Springfield, Illinois, the state capital of Illinois
  - Springfield metropolitan area, Illinois
- Springfield, LaPorte County, Indiana
- Springfield, Posey County, Indiana
- Springfield, Kentucky

Relational Semantics

- **Relationship extraction**
  - Given a chunk of text, identify the relationships among named entities (e.g. who is married to whom).
- **Semantic parsing**
  - Given a piece of text (typically a sentence), produce a formal representation of its semantics
- **Semantic role labelling** (see also implicit semantic role labelling below)
  - Given a single sentence, identify and disambiguate semantic predicates (e.g., verbal frames), then identify and classify the frame elements (semantic roles).

Discourse Analysis

- **Coreference resolution**
  - Determine which words ("mentions") refer to the same objects ("entities")
  - E.g., anaphora resolution (matching pronouns with nouns or names)
- **Discourse analysis**
  - Discourse parsing, i.e., identifying the discourse structure of a text (e.g. elaboration, explanation, contrast)
  - Speech act classification (yes-no or content question, statement, assertion, etc.)
- **Recognizing textual entailment (RTE)**
  - Given two text fragments, determine if one being true entails the other
- **Topic segmentation**
  - Given a chunk of text, separate it into segments of discussed topics
- **Argument mining**
  - extraction and identification of argumentative structures

Exercise

- Which of these NLP tasks can be solved well using supervised machine learning?
  - Which with the help of unsupervised learning?
  - Which not, why not?
- For which task is deep learning very promising and why?
- What is the difference between natural and idealized language (e.g., formal logic) in terms of their processing?
- What makes evaluating the components in a processing pipeline difficult?
- Where and why is it sometimes useful to abandon strict step-by-step processing?

Start 5 4 3 2 1 End

- Tokenization
- Normalization
- Stemming
- Lemmatization
- POS tagging
- Sentence splitting
- Syntactic parsing
- WSD
- NER
- NEL
- Terminology extraction
- Sentiment analysis
- Relationship extraction
- Semantic parsing
- Semantic role labelling
- Coreference resolution
- Discourse analysis
- RTE
- Topic segmentation
- Argument mining

Lerning Goals for this Chapter

- Understand the basic questions of Philosophy of Langauge
- Know about Zipf‘s and Heap‘s law
- Describe standard NLP pipeline
- Know common NLP tasks and be able to describe them formally
- Be able to discuss challenges and potential for deep learning regarding the standard NLP tasks

Literature

- Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition
  – D Jurafsky, JH Martin. Prentice Hall, 2000.
  – https://web.stanford.edu/~jurafsky/slp3/

- Foundations of Statistical Natural Language Processing
  – CD Manning, H Schütze. MIT Press, 1999.
  – https://nlp.stanford.edu/fsnlp/

Start

3

2

1

End