Introduction to Natural Language Processing

[DAT640] Information Retrieval and Text Mining

Petra Galuscakova
University of Stavanger

September 3, 2024

In this module

1. Language and Natural Language Processing

2. Levels of Language Description

3. Text Processing Tools and Linguistic Data

Recap: Information Retrieval

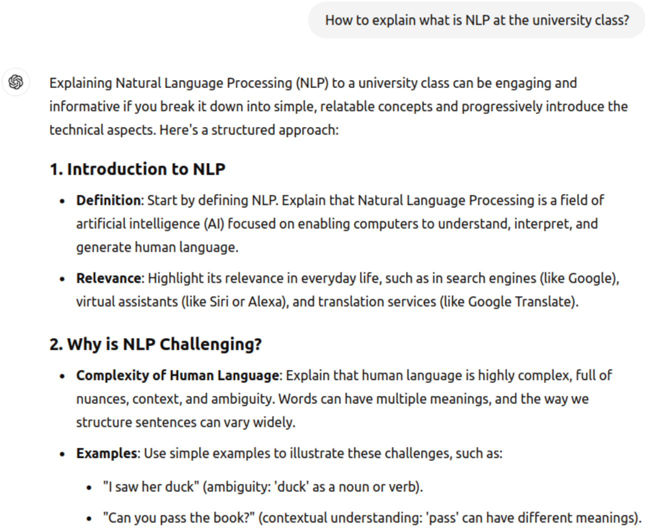
- Kahoot

Language and Natural Language Processing

What is NLP?

- Wikipedia: Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.
- Deep understanding of broad language
- Identify the structure and meaning of words, sentences, texts and conversations

Chatbots



How to explain what is NLP at the university class?

Explaining Natural Language Processing (NLP) to a university class can be engaging and informative if you break it down into simple, relatable concepts and progressively introduce the technical aspects. Here's a structured approach:

1. Introduction to NLP

- **Definition**: Start by defining NLP. Explain that Natural Language Processing is a field of artificial intelligence (AI) focused on enabling computers to understand, interpret, and generate human language.
- **Relevance**: Highlight its relevance in everyday life, such as in search engines (like Google), virtual assistants (like Siri or Alexa), and translation services (like Google Translate).

2. Why is NLP Challenging?

- **Complexity of Human Language**: Explain that human language is highly complex, full of nuances, context, and ambiguity. Words can have multiple meanings, and the way we structure sentences can vary widely.
- **Examples**: Use simple examples to illustrate these challenges, such as:
  - "I saw her duck" (ambiguity: 'duck' as a noun or verb).
  - "Can you pass the book?" (contextual understanding: 'pass' can have different meanings).

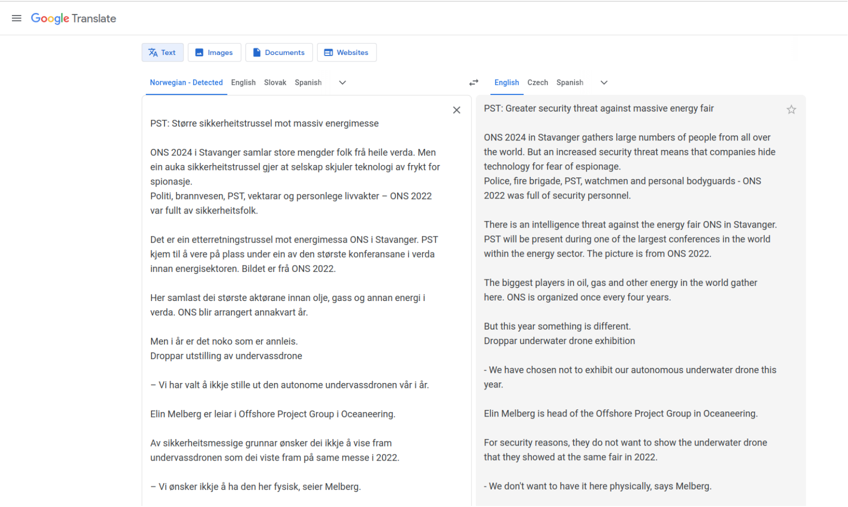
Speech Assistants

"Hey Alexa"
"Hey Siri"
"Hey Google"

[1]

[1] https://cyfuture.com/blog/domineering-power-of-voice-technology-the-dawn-of-voice-assistants-is-here/

Machine Translation



PST: Større sikkerheitstrussel mot massiv energimesse

ONS 2024 i Stavanger samlar store mengder folk frå heile verda. Men ein auka sikkerheitstrussel gjer at selskap skjuler teknologi av frykt for spionasje.
Politi, brannvesen, PST, vektarar og personlege livvakter – ONS 2022 var fullt av sikkerheitsfolk.

Det er ein etterretningstrussel mot energimessa ONS i Stavanger. PST kjem til å vere på plass under ein av den største konferansane i verda innan energisektoren. Bildet er frå ONS 2022.

Her samlast dei største aktørane innan olje, gass og annan energi i verda. ONS blir arrangert annakvart år.

Men i år er det noko som er annleis.
Droppar utstilling av undervassdrone

– Vi har valt å ikkje stille ut den autonome undervassdronen vår i år.

Elin Melberg er leiar i Offshore Project Group i Oceaneering.

Av sikkerheitsmessige grunnar ønsker dei ikkje å vise fram undervassdronen som dei viste fram på same messe i 2022.

– Vi ønsker ikkje å ha den her fysisk, seier Melberg.

PST: Greater security threat against massive energy fair

ONS 2024 in Stavanger gathers large numbers of people from all over the world. But an increased security threat means that companies hide technology for fear of espionage.
Police, fire brigade, PST, watchmen and personal bodyguards - ONS 2022 was full of security personnel.

There is an intelligence threat against the energy fair ONS in Stavanger. PST will be present during one of the largest conferences in the world within the energy sector. The picture is from ONS 2022.

The biggest players in oil, gas and other energy in the world gather here. ONS is organized once every four years.

But this year something is different.
Droppar underwater drone exhibition

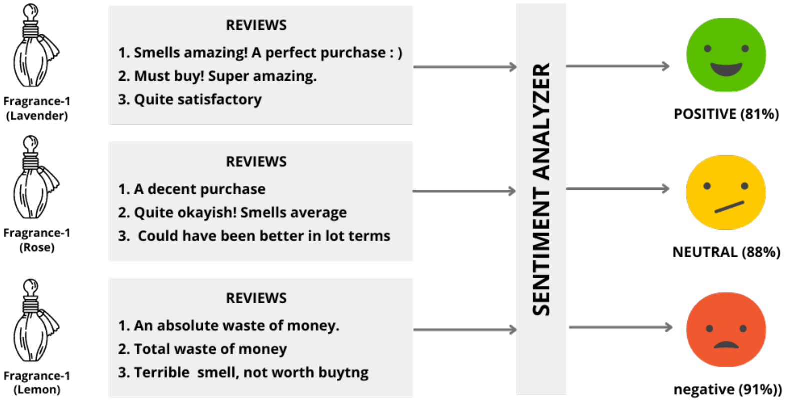
- We have chosen not to exhibit our autonomous underwater drone this year.

Elin Melberg is head of the Offshore Project Group in Oceaneering.

For security reasons, they do not want to show the underwater drone that they showed at the same fair in 2022.

- We don't want to have it here physically, says Melberg.

Sentiment Analysis



Fragrance-1 (Lavender)

REVIEWS

1. Smells amazing! A perfect purchase :)
2. Must buy! Super amazing.
3. Quite satisfactory

Fragrance-1 (Rose)

REVIEWS

1. A decent purchase
2. Quite okayish! Smells average
3. Could have been better in lot terms

Fragrance-1 (Lemon)

REVIEWS

1. An absolute waste of money.
2. Total waste of money
3. Terrible smell, not worth buytng

SENTIMENT ANALYZER

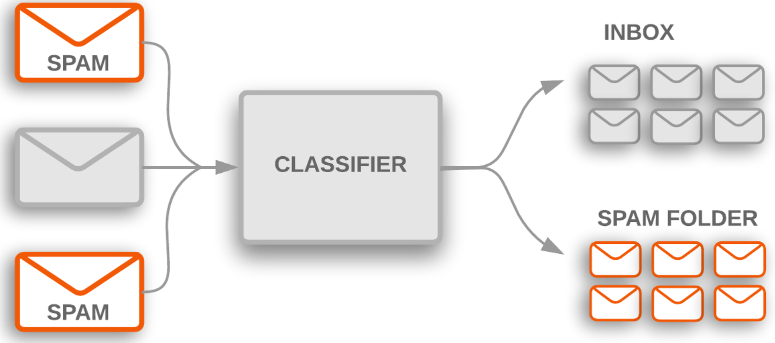
POSITIVE (81%)

NEUTRAL (88%)

negative (91%)) [2]

[2] https://www.linkedin.com/pulse/decoding-emotions-using-text-data-natural-language-roy-rachman-sedik/

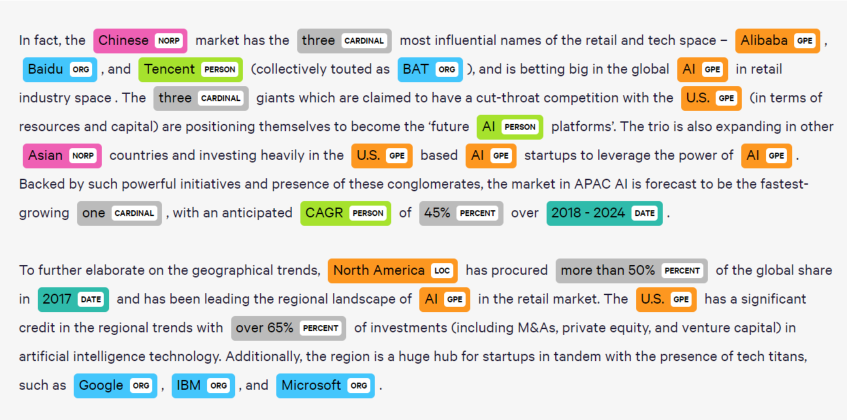
Text Classification



[3]

[3] https://developers.google.com/machine-learning/guides/text-classification

Named Entity Recognition



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

[4]

[4] https://medium.com/@alessandropaticchio/named-entity-recognition-from-scratch-e76b9b3affad

What is Language?

- Language is a system of conventional spoken, manual (signed), or written symbols by means of which human beings, as members of a social group and participants in its culture, express themselves. The functions of language include communication, the expression of identity, play, imaginative expression, and emotional release.[5]
- What is hard about language processing?
  - Language is multimodal
  - There are many languages
  - In the same language, there can be many dialects
  - Ambiguity (lexical ambiguity, syntactic ambiguity, ...)
  - Sarcasm, mood, jokes, ...
  - Spelling errors
  - Colloquialisms and slang

[5] https://www.britannica.com/topic/language

Levels of Language Description

Levels of Language Description

- Phonetics [Sounds; (nearly) language independent]
- Phonology [Sound patterns, language dependent abstraction over sound]
- Morphology [Word structure]
- Syntax [Sentence structure]
- Semantics [Literal meaning]
- Pragmatics [Meaning in context]

Written vs. Spoken Language [6]

- Written texts historically tended to be more carefully worded and better organized than spoken texts, they contain fewer errors, hesitations, and incomplete sentences. .. but Twitter, Slack, ...
- Writing is usually planned in advance, is often proofread
- Writing contains information not available in speech (sections, author name, ...)
- Spelling is more uniform across different individuals, places and times using the same language than is pronunciation.
- Writing styles change much more slowly than speech styles, and so writing seems more 'permanent' and 'authoritative'.

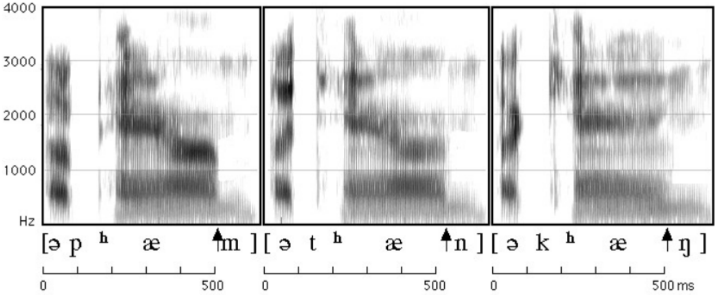
[6] Source: Jiri Hana

Written vs. Spoken Language

- Spoken language existed much earlier than writing. Writing was most likely invented in Sumer (Mesopotamia, current Iraq) about 5500 years ago. Language probably exists for 40,000 years or more.
- Societies which only speak their language and do not write it. No society uses only a written language (with no spoken form).
- We learn to speak before we learn to write.
- Most people say more during one month than they write during their entire lives.
- Writing must be taught, whereas spoken language is ‘acquired automatically’.
- Speech contains information that writing lacks (intonation, stress, voice quality ...)

Phonetics

- Technical word for a speech sound is phone (hence, phonetics)
- Phonetics is the study of speech sounds: how they are produced, how they are perceived, what their physical properties are.



Phonetics vs. Phonology

- Phonetics studies how sounds really sound, while phonology studies how they sound to speakers of some language.
- A phoneme is any set of similar speech sounds that is perceptually regarded by the speakers of a language as a single distinct unit, a single basic sound, that helps distinguish one word from another.
- It is sometimes difficult for native speakers of a language to tell the difference between sounds which may be completely distinct for speakers of another language.
- a. English: pit [phIt] vs. spit [spIt][7]
  b. Hindi: [phu:l](fruit) vs. [pu:l] (moment)
- English speakers consider [p] and the [ph] to be the same sound, despite some irrelevant articulatory details. For Hindi speakers, the same details are enough to completely differentiate the two sounds, making them as different as [p] and [b] for English speakers.

[7] (https://dictionary.cambridge.org/pronunciation/)

Phonetics and Phonology Problems

- Help to solve following problems:
  - Recognize words
  - Find 'sentences' in speech
  - Recognize a question
  - Recognize a mood of the speaker
  - Find region of origin of the speaker
  - Even if the speaker talks English, find the native language

Morphology

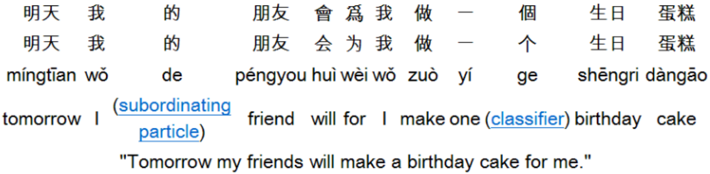
- Morphology is the study of the internal structure of words
- Morphemes are the smallest linguistic units which have a meaning or grammatical function
- Words are composed of morphemes (one or more)
- sing-er-s, home-work, un-kind-ly, flipp-ed, de-nation-al-iz-ation

Morphemes

- Content morphemes: carry some semantic content (car, -able, un-)
- Functional morphemes: provide grammatical information [the, and, -s (plural), -s (3rd singular)]
- Root: nucleus of the word that affixes attach too
- Affix: a morpheme that is not a root; it is attached to a root
- Suffix (after the root), prefix (before the root), infix (inside the root)
- Lemma: A form chosen by convention (e.g., nom. sg. for nouns, infinitive for verbs) to represent a set of word's morphological variants (lexemes)
- Also called the canonical/base/dictionary/citation form: e.g.: break, breaks, broke, broken, breaking have the same lemma to break

Morphology and Languages

- Two basic morphological types of languages:
- Analytic (isolating) languages: Sentences are sequences of single-morpheme words. e.g. Vietnamese and Classical Chinese



明天	我	的	朋友	會	爲	我	做	一	個	生日	蛋糕
明天	我	的	朋友	会	为	我	做	一	个	生日	蛋糕
míngtīan	wǒ	de	péngyou	huì	wèi	wǒ	zuò	yí	ge	shēngri	dàngāo
tomorrow	I	(subordinating particle)	friend	will	for	I	make	one	(classifier)	birthday	cake

"Tomorrow my friends will make a birthday cake for me."

- Synthetic languages: Affixes are added to roots.

Synthetic Languages Subtypes

- Agglutinating:
  - Each morpheme has a single function, it is easy to separate them
  - E.g., Uralic languages (Estonian, Finnish, Hungarian), Turkish, Basque, ...
  - Japanese:
    - taberu (I'll eat it)
    - tabetai (I want to eat it)
    - tabetakunai (I don't want to eat it)
    - tabetakunakatta (I didn't want to eat it)
- Fusional:
  - Like agglutinating, but affixes tend to 'fuse together', One affix has more than one function.
  - E.g., Slavic, Romance languages, Greek, ...
  - Czech matk-a 'mother' – 'a' means the word is a noun, feminine, singular, nominative
- Polysynthetic:
  - Extremely complex, many roots and affixes combined together, often one word corresponds to a whole sentence in other languages.
  - Angyaghllangyugtuq – 'he wants to acquire a big boat' (Eskimo)

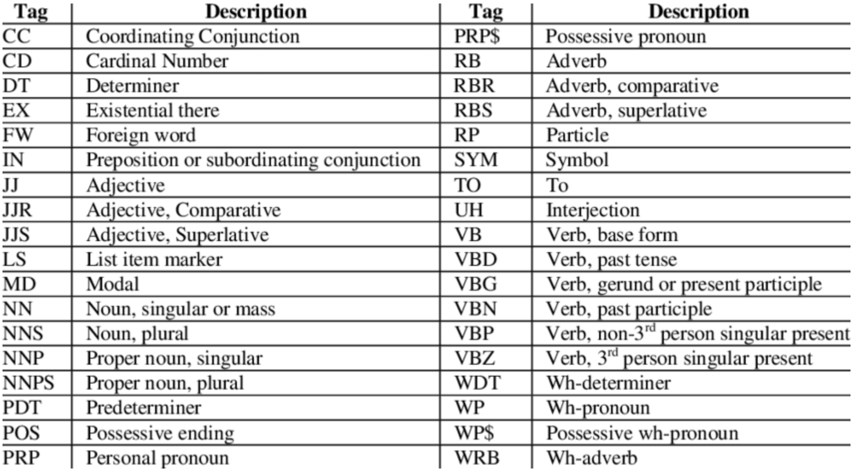
Syntax

- The part of linguistics that studies sentence structure.
- Word order:
    - I want these books.
    - *want these I books.
- Agreement – subject and verb, determiner and noun, ... often must agree:
    - He wants this book.
    - *He want this book.
- Hierarchical structure – what modifies what:
    - We need more (intelligent leaders). (more of intelligent leaders)
    - We need (more intelligent) leaders. (leaders that are more intelligent)
- Syntax is not about meaning. Sentences can have no sense and still be grammatically correct:
- Colorless green ideas sleep furiously. – nonsense, but grammatically correct

Part of Speech (POS) Tags

- Words in a language behave differently from each other.
- But not each word is entirely different from all other words in that language.
- Words can be categorized into parts of speech (lexical categories, word classes) based on their morphological, syntactic and semantic properties.
- Open categories (open to additions):
  - Verb, noun, pronoun, adjective, numeral, adverb
  - Subject to inflection
  - Potentially unlimited number of words
- Closed categories:
  - Preposition, conjunction, article, interjection, particle
  - Finite and small number of words
- Ambigous: Time [V,N] flies [V,N] like [V,Prep] an [DET] arrow [N].

Penn Treebank Tagset



Tag	Description	Tag	Description
CC	Coordinating Conjunction	PRP$	Possessive pronoun
CD	Cardinal Number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	To
JJR	Adjective, Comparative	UH	Interjection
JJS	Adjective, Superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

POS Tags Demo

http://lindat.mff.cuni.cz/services/udpipe/

Nouns of Special Interest

- Synonyms: words with almost the same meaning (big - large)
- Antonyms: words with opposite meanings (big - small)
- Hyponyms and Hypernyms: words that refer to members of a larger category: pigeon, crow, and hen are all hyponyms of bird and animal; bird and animal are both hypernyms of pigeon, crow, and hen.
- Meronyms - a word that denotes a constituent part or a member of something. For example, apple is a meronym of apple tree.
- Homonyms - words that sound the same (tail – tale)

Nouns and Noun Phrases

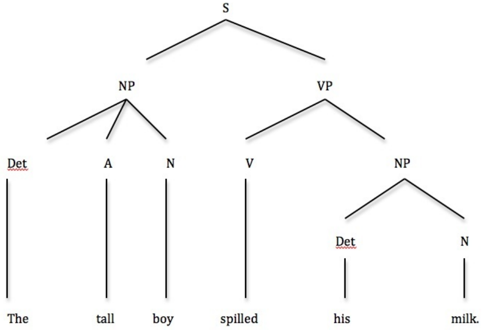
- In English: a determiner followed by a noun, or determiner followed by an adjective followed by a noun, or a single noun, ...
- NP -> Det N [the cat]
  NP -> Det A N [those noisy cats]
  NP -> N [cats]
  NP -> A N [noisy cats]
  NP -> (Det) (A) N [cats, noisy cats, the cat, those noisy cats]
- Prepositional phrases
  PP -> P NP [about those noisy cats]

Noun Phrases and Sentences

- In English: a sentence consists of a subject (usually a noun phrase) followed by a verb which is sometimes followed by an object (another noun phrase), prepositional phrases etc.
- Alphons slept. – Subject + V [S → NP V]
  Alphons saw his dog. – Subject + V + Object [S → NP V NP]
  Alphons asked for a cake. [S → NP V PP]
  Alphons begged a cake from his dog. [S → NP V NP PP]

Phrase Trees

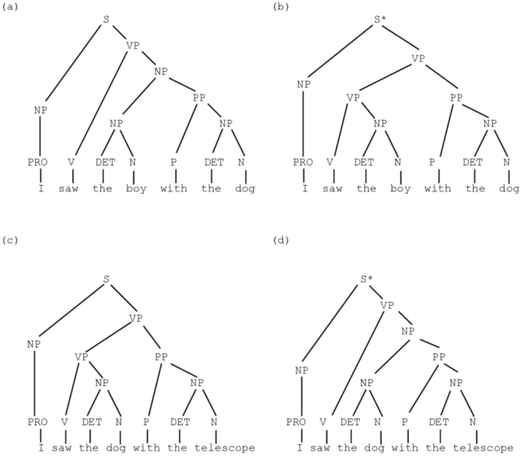
- Phrases are created from other phrases or words. Sentence is the biggest phrase.
- We can depict the fact that a sentence is built from smaller parts by a diagram.



8

[8] https://web.mnstate.edu/houtsli/tesl551/Syntax/page3.htm

Syntactic Ambiguity

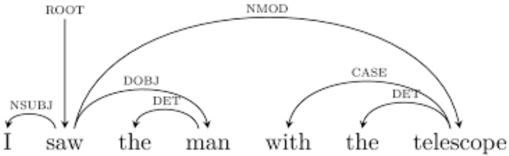


9

[9] Sagae, Kenji, Macwhinney, Brian and Lavie, Alon: Automatic Parsing of Parent-child Interactions, 2004

Dependency Trees

- Dependency grammar is a description of a dependency structure of a sentence, i.e. the structure of dependency relations between the elements of a sentence.
- Dependency is an asymmetric binary relation between language units: governing head and dependent modifying unit.
- The verb is always the head of the sentence tree.



10

[10] https://wstyler.ucsd.edu/talks/l6_23_parsing_handout.html

Syntactic Parsing Demo

http://lindat.mff.cuni.cz/services/udpipe/

Semantics

- The part of linguistics that studies meaning in language:
    - The meanings of words
    - How word meanings are combined to give the meaning of a sentence
- Semantics deals with literal meaning.
- Pragmatics deals with the intended meaning, with the usage of language, with language in context, etc.

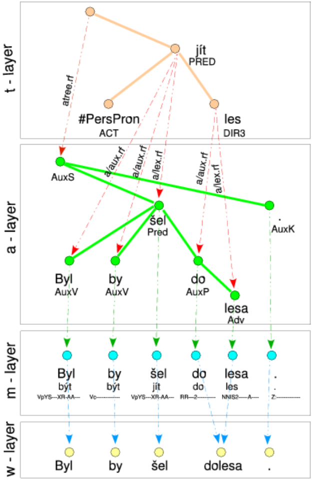
Semantic Roles

- Semantic roles are used to indicate the role played by each entity in a sentence.
- Agent – one who deliberately does the action
- Cause – mindlessly performs the action
- Experiencer – has sensory or mental experience
- Patient – thing that the action happens to
- Theme – thing or being that is in a state/location
- Source – origin of a change in location/possesion
- Goal/recipient – endpoint of a change in location/possesion
- Instrument – the means of accomplishing the action

Semantic Roles

- a. The janitor opens the door with a key. (key – instrument)
- b. The key opens the door. (key – instrument)

Layers of Sentence Description



11

[11] https://ufal.mff.cuni.cz/pdt2.0/browse/doc/pdt-guide/en/html/ch02.html

Discourse and Coreference

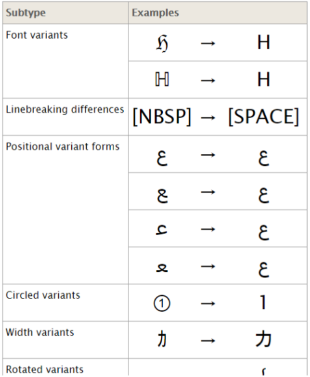
- Discourse
  - Language consists of collocated, related groups of sentences. We refer to such a group of sentences as a discourse.
  - Cooperative, one-way conversation.
  - Deliver information from the speaker/writer to the listeners/readers
- Dialogue
  - Cooperative, two-way conversation.
  - Goal is for participants to exchange information and build relationships with one another.
- Coreference: referring expressions that are used to refer to the same entity.
- Anaphora: reference to a previously introduced entity.

Text Processing Tools and Linguistic Data

Unicode

- ASCII (128 chars) → 8bit codepages (256 chars) → Unicode (65,536 chars)
- UTF-8 is a way of encoding Unicode, is capabale of encoding 1M+ characters
- UTF-8 size of encoded character varies, the more frequent should take less space
- UTF-8-, -16 and -32 differ in the coding approach. UTF-8 requires 8, 16, 24 or 32 bits (one to four bytes) to encode a Unicode character, UTF-16 requires either 16 or 32 bits to encode a character, and UTF-32 always requires 32 bits to encode a character.
- The first 128 Unicode code points, U+0000 to U+007F, which are used for the C0 Controls and Basic Latin characters and which correspond to ASCII
- If possible, make sure that all your data is UTF-8 and all your software assumes UTF-8 everywhere

Unicode Normalization



Subtype	Examples
Font variants	ℌ → H
	ℍ → H
Linebreaking differences	[NBSP] → [SPACE]
Positional variant forms	ﻉ → ع
	ﻊ → ع
	ﻋ → ع
	ﻌ → ع
Circled variants	① → 1
Width variants	ｶ → カ
Rotated variants	

12

- unicodedata Python library: https://docs.python.org/3/library/unicodedata.html

[12] Zdenek Zabokrtsky

Tokenization, Stopwords, Stemming

- Covered in M1
- Segmentation: Parsing a string into individual sentences
- Tokenization: Parsing a string into individual words (tokens)
- Stemming: Reduce the different forms of a word that occur to a common stem
- Stopwords removal: remove function words that have little meaning apart from other words: the, a, an, that, those

Lemmatization

- The goal of both stemming and lemmatization is to reduce inflectional forms
- Stemming usually refers to a heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
- Stemming often works fast and well in English and performs better in IR
- Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
- am, are, is, was, being, will be -> to be

Corpora

- Tony McEnery: 'Corpus data are, for many applications, the raw fuel of NLP, and/or the testbed on which an NLP application is evaluated.'
- Corpus (plural corpora) is a collection of linguistic data, either written texts or a transcription of recorded speech
- Might be text or multimodal
- Based on language included: Monolingual, Multilingual, and Parallel (aligned pairs of texts in two or more languages)
- Might be focused on different data sources (e.g. books, news text, Web, law texts, ...)
- Might be annotated with different information (named entities, coreference, sentiment, ...)
- Treebanks are annotated databases with syntactic parse trees

Shared Corpora

- LRE Map https://lremap.elra.info/
- Linguistic Data Consortium: https://www.ldc.upenn.edu/
- LINDAT/CLARIN Repository: https://lindat.mff.cuni.cz/repository/
- Norwegian: CLARINO: https://tekstlab.uio.no/clarino/ and https://repo.clarino.uib.no/xmlui/
- Norwegian: National Library of Norway: https://www.nb.no/sprakbanken/en/sprakbanken/

WordNet

- A large lexical database, or “electronic dictionary,” developed and maintained at Princeton
- Includes most English nouns, verbs, adjectives, adverbs
- Organized by meaning: words in close proximity are semantically similar
- WordNet groups (roughly) synonymous, denotationally equivalent, words into unordered sets of synonyms (’synsets’):
  - hit, beat, strike
  - big, large
  - queue, line
- https://wordnet.princeton.edu/

Natural Language Toolkit (NLTK)

- Provides an interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
- https://www.nltk.org/

spaCy

- spaCy is a library for advanced Natural Language Processing in Python and Cython. It's built on the very latest research, and was designed from day one to be used in real products.
- spaCy comes with pretrained pipelines and currently supports tokenization and training for 70+ languages. It features state-of-the-art speed and neural network models for tagging, parsing, named entity recognition, text classification and more, multi-task learning with pretrained transformers like BERT, as well as a production-ready training system and easy model packaging, deployment and workflow management.
- https://pypi.org/project/spacy/

Exercise

Exercise E5-1 SpaCy Library

Summary

- Introduction to NLP
- Levels of language description
- Tools and data for NLP

Reading and References

- *Lecture Notes on Natural Language Processing*, Jordan Boyd-Graber[13]
- *Natural Language Processing*, Institute of Formal and Applied Linguistics [14]
- *Introduction to Linguistics* , Jiri Hana [15]
- *Speech and Language Processing*, Dan Jurafsky and James H. Martin [16]

[13] http://users.umiacs.umd.edu/~jbg/teaching/CMSC_470/

[14] https://ufal.mff.cuni.cz/courses/npfl124

[15] https://ufal.mff.cuni.cz/courses/npfl063

[16] https://web.stanford.edu/~jurafsky/slp3/

Early Course Dialogue

- Students elect the student representative
- Students fill out the anonymous feedback form
- Raw feedback form results are shared with the student representative
- The instructor will summarize the feedback and share it with the student representative
- If approved, the summarized feedback form will be shared with the university

Early Course Dialogue

https://tinyurl.com/yhc5za8a