VL Deep Learning for Natural Language Processing

3. Introduction to Natural Language Processing and Text Mining

Prof. Dr. Ralf Krestel

AG Information Profiling and Retrieval

Summary NN

Learning Goals for this Chapter

- Describe a 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 standard NLP tasks
- Explain text mining and identify TM tasks
- List various text mining tasks, naming
  - Difficulties/challenges
  - Traditional approaches
- Know the limitations of traditional TM applications

https://www.kdnuggets.com/images/cartoon-turkey-data-science.jpg

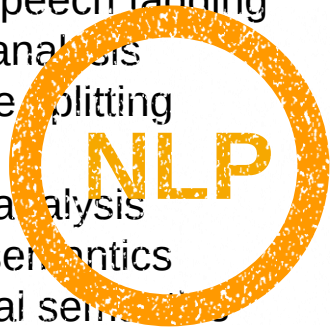
Topics Today

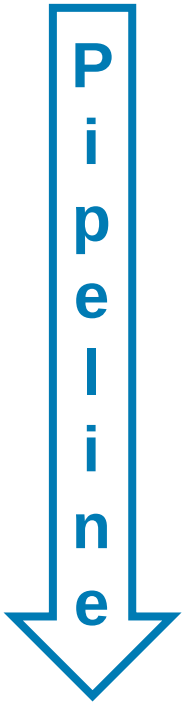
1. **Natural Language Processing Pipeline**
2. Text Mining Applications
3. Summary

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

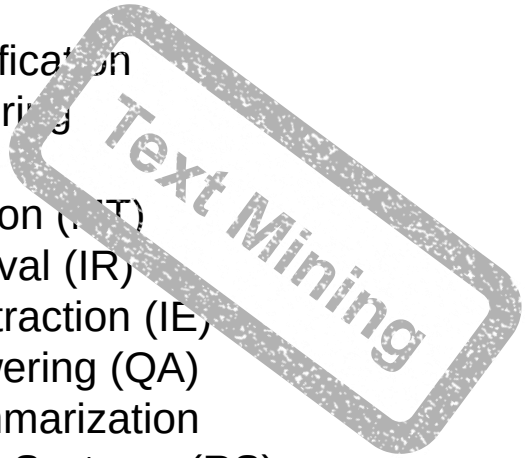
Common NLP Tasks & Text Mining Applications

- 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





- Applications
  - Document Classification
  - Document Clustering
  - Topic Modeling
  - Machine translation (MT)
  - Information retrieval (IR)
    - Information extraction (IE)
    - Question answering (QA)
    - Automatic summarization
    - Recommender Systems (RS)
  - Knowledge Graphs (KG)
  - 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

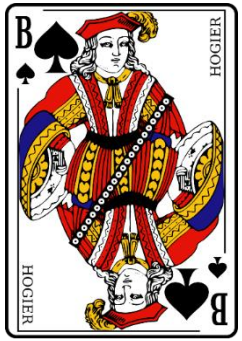
- **OCR, speech recognition**
  - Generate/Extract text from image or audio files
- **Tokenization**
  - Aka word segmentation
  - Forming words from sequence of characters
  - Surprisingly complex in English, can be harder in other languages
  - Basic assumption: any sequence of alphanumeric characters 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

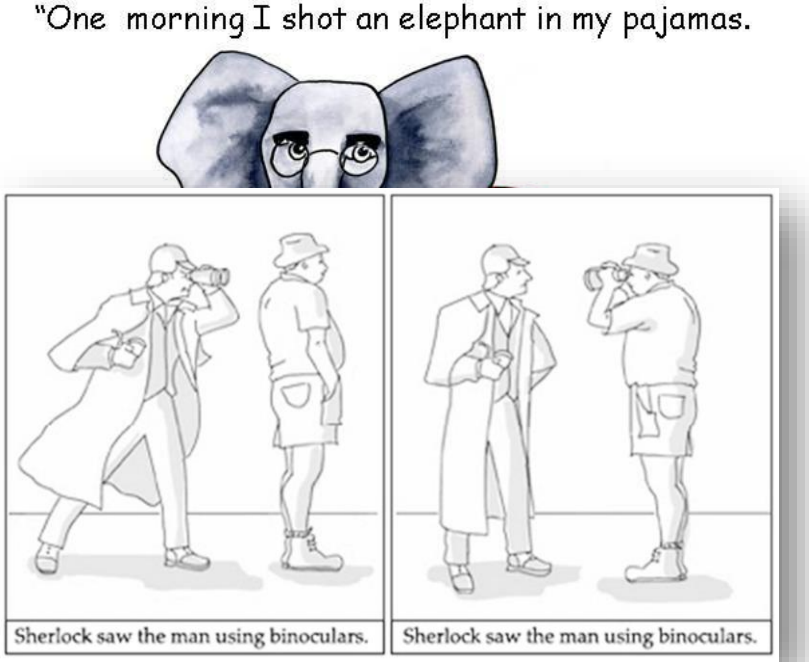
In German?

- 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.

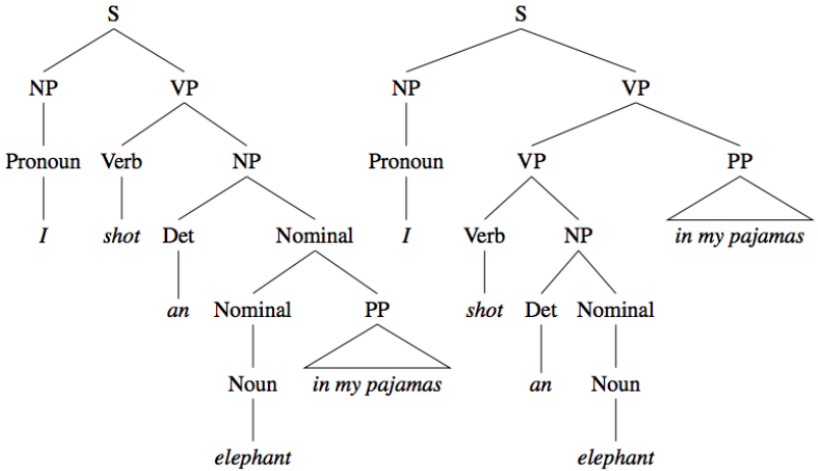
Parsing

- Shallow parsing („chunking“)

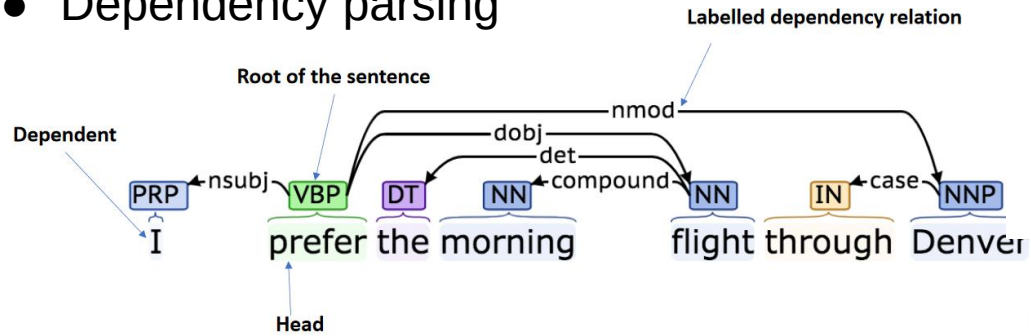
The morning flight from Denver has arrived.

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 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)

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

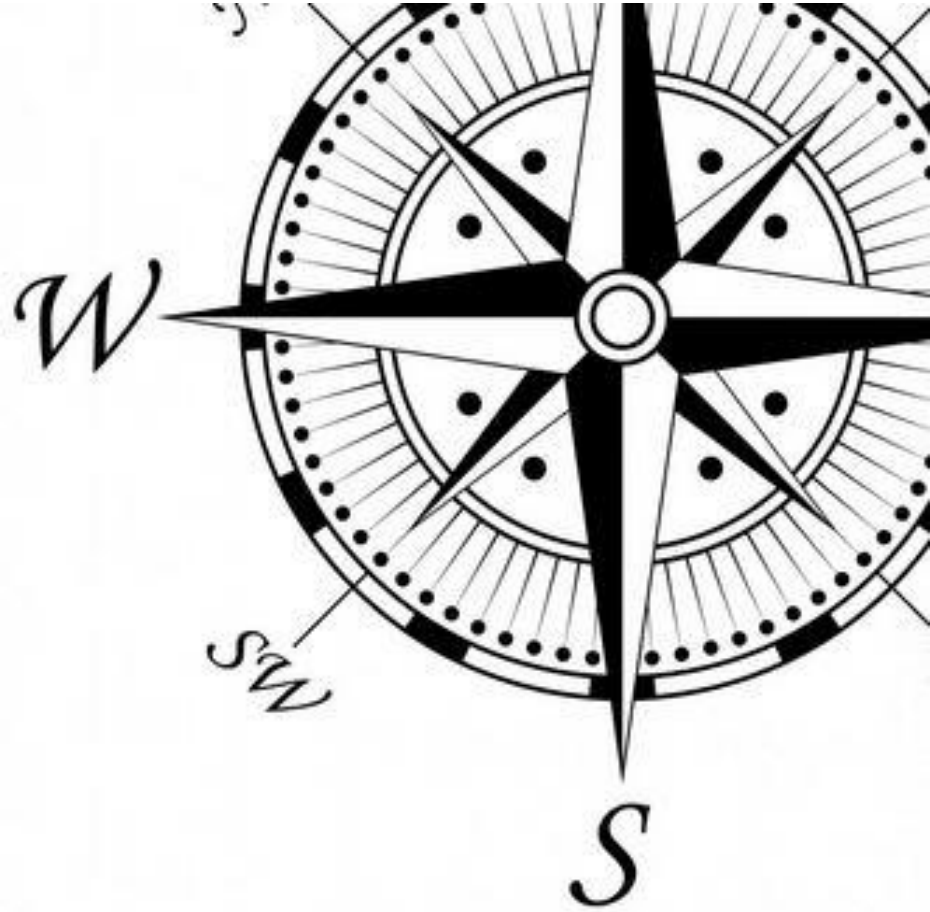
Exercise

- 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?

- 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

Topics Today

1. Natural Language Processing Pipeline
2. **Text Mining Applications**
3. Summary



What is Text Mining?

"**Text mining** (also referred to as *text analytics*) is an artificial intelligence (AI) technology that uses **natural language processing** (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms."

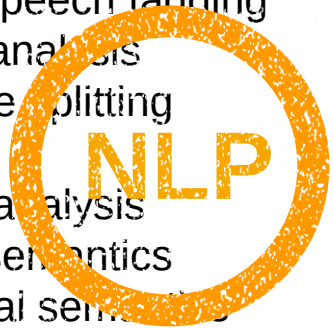
https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

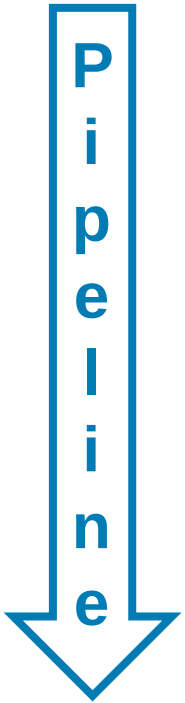
- Data mining in texts: Finding useful and interesting patterns in a corpus

- Text mining vs. Information retrieval
- Data mining vs. Database query

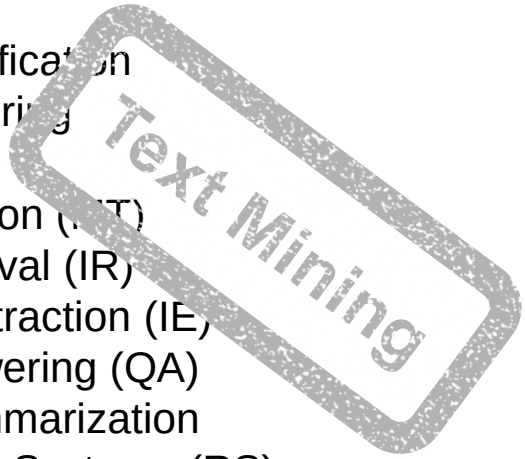
Text Mining ~ Higher Level NLP

- 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





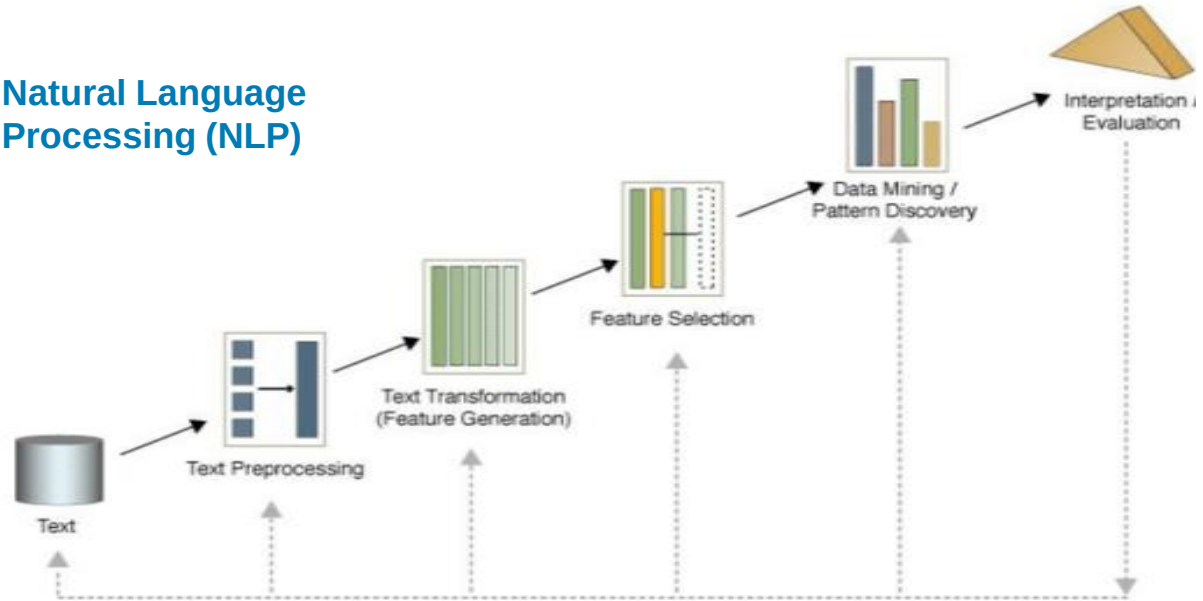
- Applications
  - Document Classification
  - Document Clustering
  - Topic Modeling
  - Machine translation (MT)
  - Information retrieval (IR)
    - Information extraction (IE)
    - Question answering (QA)
    - Automatic summarization
    - Recommender Systems (RS)
  - Knowledge Graphs (KG)
  -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



Text Mining ~ Data Mining with Text

- Text Preprocessing
  - Syntactic/semantic
- Feature generation
  - Bag-of-Words
- Feature selection
  - Statistics
- Text/Data mining
  - Classification
  - Clustering
  - Prediction
- Analysis of results
  - Visualization
  - Aggregation

Natural Language Processing (NLP)

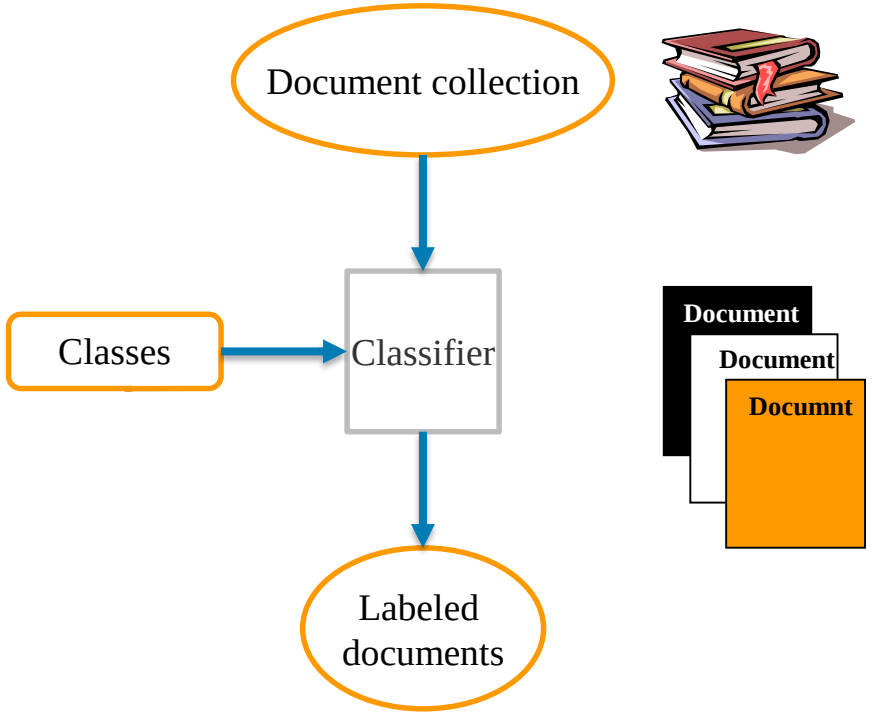


Document Classification

- Given:
  - Collection of documents
  - Different classes

- Goal:
  - Assign classes to documents
  - Multi-class vs. binary classification

  - Single- vs. multi-label classification

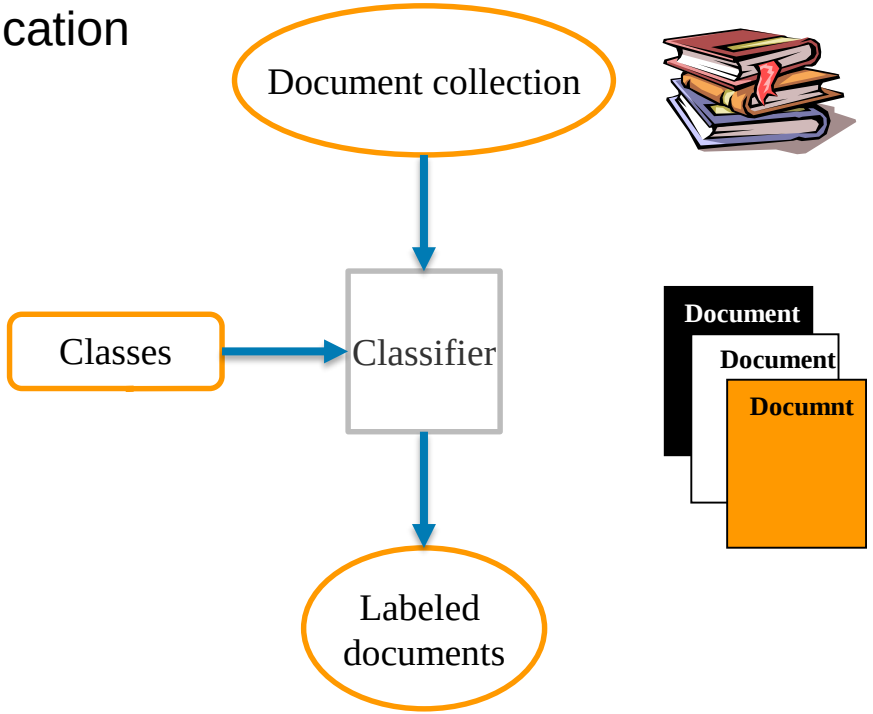


Sentiment Analysis

- One specific kind of document classification

- Given:
  - Collection of documents
    - Product reviews, tweets, …
  - Three classes
    - Pos, neu, neg

- Goal:
  - Assign classes to documents, sentences, or aspects
  - Usually with score or probability



Sentiment Analysis as ML Problem

- Naive Bayes Classification:
  - Which class c has the highest probability? $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$
  - ⟶ Bayes rule $\hat{c} = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)}$
  - ⟶ Simplification $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c)$
  - ⟶ Bag-of-Words repräsentation $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1, w_2, \ldots, w_n|c)P(c)$
  - ⟶ Naive Bayes assumption $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1|c) \cdot P(w_2|c) \cdot \ldots \cdot P(w_n|c) \cdot P(c)$
  - ⟶ Reformulating $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in \{1 \ldots n\}} P(w_i|c)$
  - ⟶ Logarithm $\hat{c} = \underset{c \in C}{\operatorname{argmax}} \log P(c) + \sum_{i \in \{1 \ldots n\}} \log P(w_i|c)$
- Other classification approaches
  - Rule-based (sentiment lexikon)
  - Support vector machine

How to estimate these probabilities?

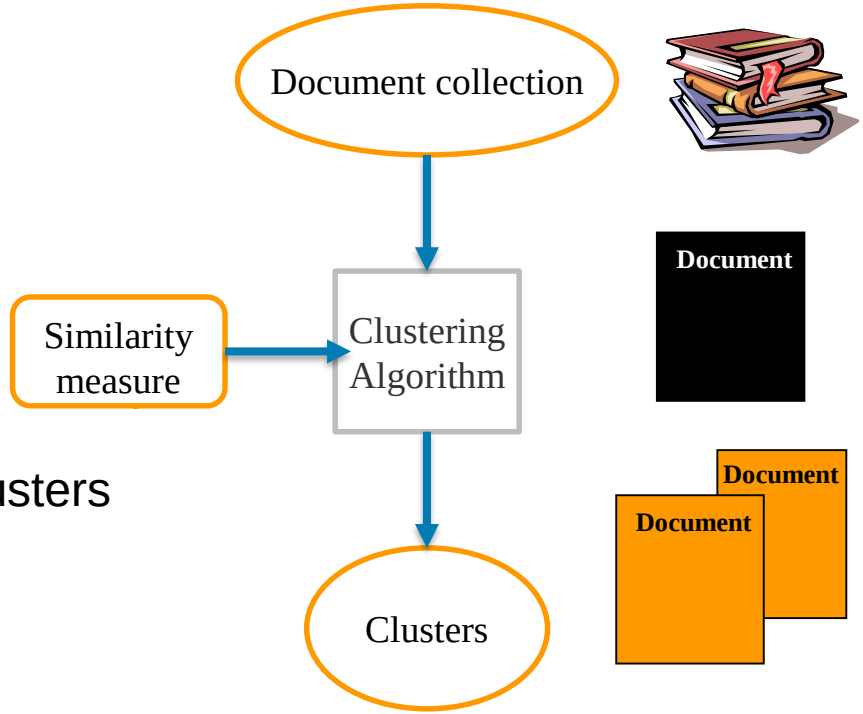
Exercise

- What are the four ways to transform a multi-label classification problem into a single-label classification problem?

Document Clustering

- Given:
  - Collection of documents
  - Similarity measure
    - Euclidean, cosine, etc.

- Goal:
  - Grouping of documents
  - Similar documents in same cluster
    Dissimilar documents in different clusters
  - Usually non-overlapping



Similarity



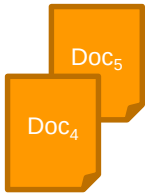


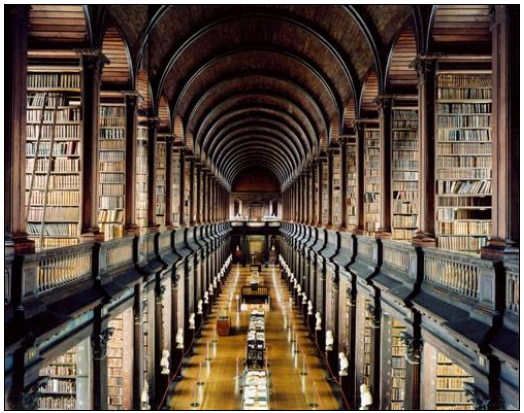



Representation

- In an "information space"
- E.g. spatial distribution in libraries
  - (Thematically) Similar books are close to each other
- Can this principle be transferred to a virtual space?
- Idea: represent documents as points in an abstract semantic space
- Similarity is then measured by distance

- Traditional representation of text data:
  - Bag-of-Words model
  - TF-IDF
  - Vector space model



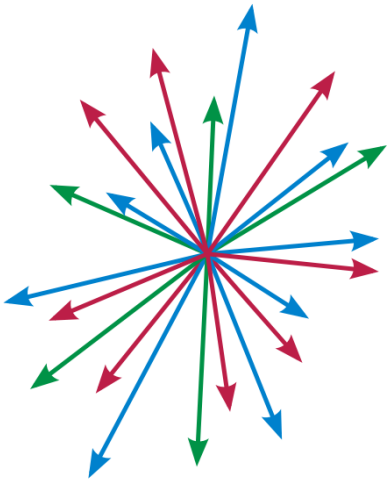


Vector Space Model I

- Proposed by Gerard Salton (Salton, 1975)
- Still very important
- Information retrieval based on it
- Simple, intuitive
- Can be easily weighted (TF-IDF)

- Documents are represented as n-dimensional, reel-valued points in a vector space $\mathbb{R}^n$, with n size of vocabulary
- Normally, n is very large: >100,000 terms

- Each term spans its own dimension
- Documents can then be represented as incidence vectors

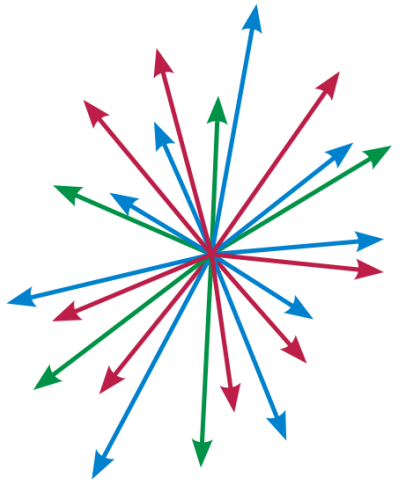


https://commons.wikimedia.org/wiki/File:Vector_space_illust_rotate.svg

Vector Space Model I

- Document collection can then be represented as a matrix of term weights
  - $V = \{Term_1, Term_2, \ldots, Term_t\}$
  - $Doc_i = (d_{i1}, d_{i2}, \ldots, d_{it})$

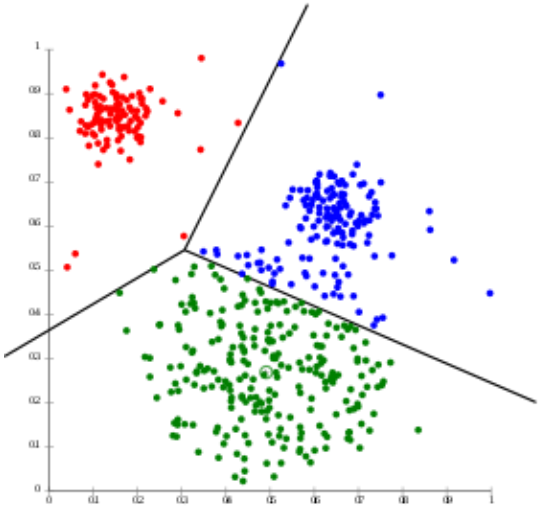
	$Term_1$	$Term_2$	$\ldots$	$Term_t$
Doc_1	d_{11}	d_{12}	$\ldots$	d_{1t}
Doc_2	d_{21}	d_{22}	$\ldots$	d_{2t}
$\vdots$	$\vdots$			
Doc_n	d_{n1}	d_{n2}	$\ldots$	d_{nt}



https://nl.wikipedia.org/wiki/Joan_Blaeu#/media/File:Theatrum_II_1640_Index_GMG430_220_%26_221.jpg

Document Clustering as ML Problem

- Given: documents in a vector space
- Goal: Group similar documents together
  - Similarity ≅ Close distance in vector space
- Necessary: distance function (similarity measure)
  - Euclidian distance
  - Manhattan distance
  - …
- Many clustering algorithms
  - „Hard“ (non-overlapping)
    - E.g. k-means
  - „Soft“ (overlapping or fuzzy)
    - E.g. EM clustering

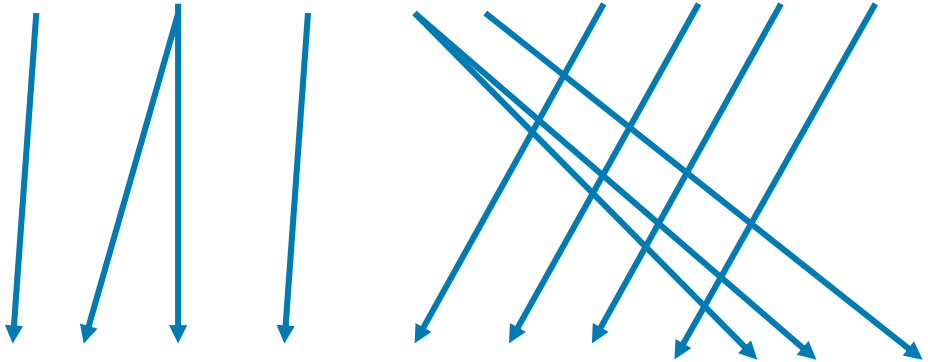


https://de.wikipedia.org/wiki/Clusteranalyse

Machine Translation

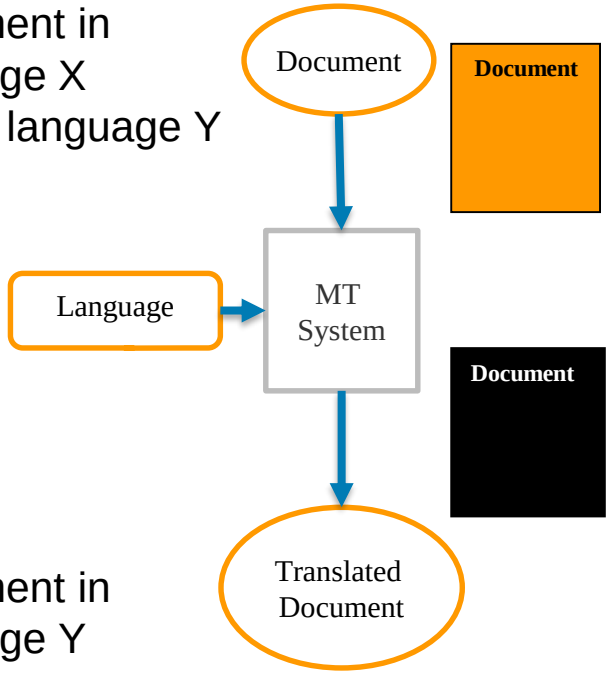
- The task of machine translation is to translate a sentence x in one language (**source**) into a sentence in another language (**target**).

x: Ein Männlein steht im Walde ganz still und stumm.



y: A little man stands quite still and silent in the forest.

- Given:
  - Document in language X
  - Target language Y



- Goal:
  - Document in language Y

1990–2010: Statistical Machine Translation

- SMT Idea: Learn a probabilistic model from data
- E.g. we want to find the best German sentences y, given the English sentence x

$$\underset{y}{argmax}\, P(y|x)$$

- Using Bayes rule:

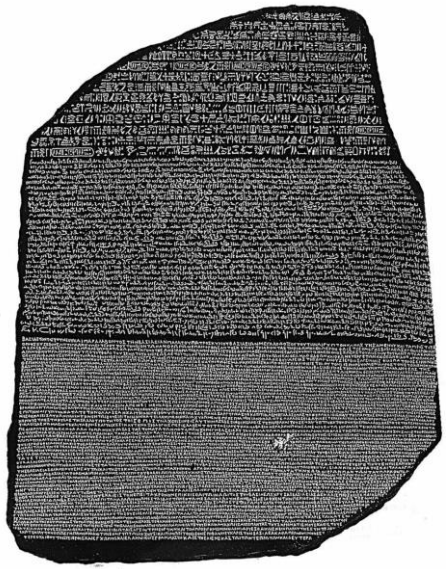
$$= \underset{y}{argmax}\, P(x|y)P(y)$$

Translation model:
- **Describes how to translate words and phrases**
- **Learnt using parallel corpora**

Language model:
- **Describes how good German looks like**
- **Learnt using a monolingual corpus**

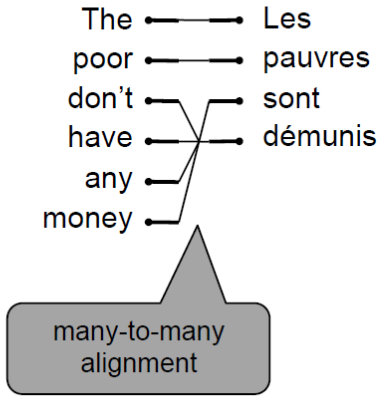
Machine Translation as ML Problem I

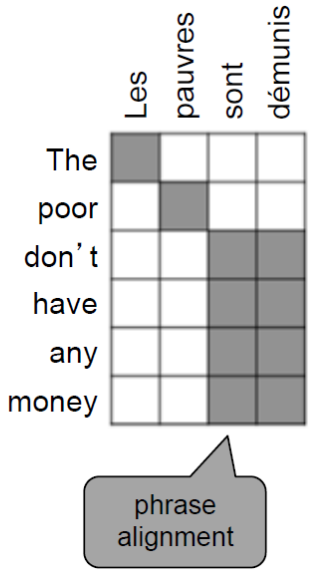
- Statistical Machine Translation (SMT)
- How to learn the translation model $P(x|y)$?
  - With a very large amount of parallel data!
- More closely, we do not want to learn $P(x|y)$, but $P(x, a|y)$, where a is an alignment.
- **Alignment** is the mapping of English words to German words within our sentences x and y.
- A number of factors influence the learning of $P(x, a|y)$:
  - Probabilities of certain assignments
    - Also depends on the position in the sentence
  - Probabilities of fertility of certain words …



Alignments

1. Besides 1-to-1 alignments there are other possibilities:
2. 1-to-0 or 0-to-1
   - Some words do not have counterparts in other languages
3. 1-to-many
   - These are fertile words
4. Many-to-1
5. Many-to_many
   - Phrases



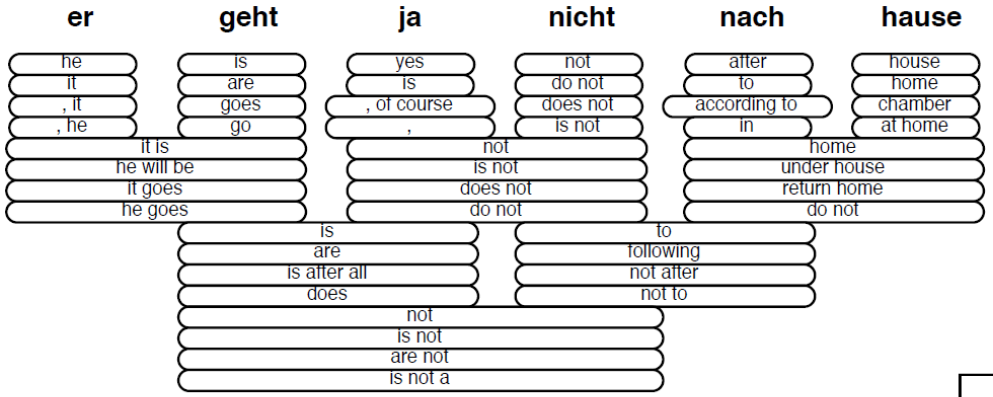


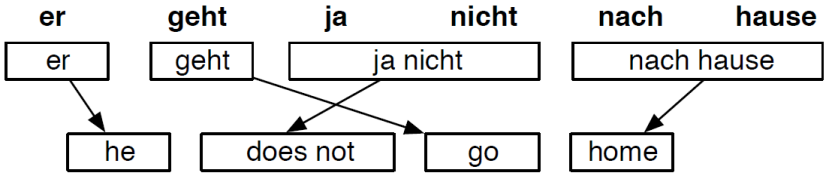
Machine Translation as ML Problem II

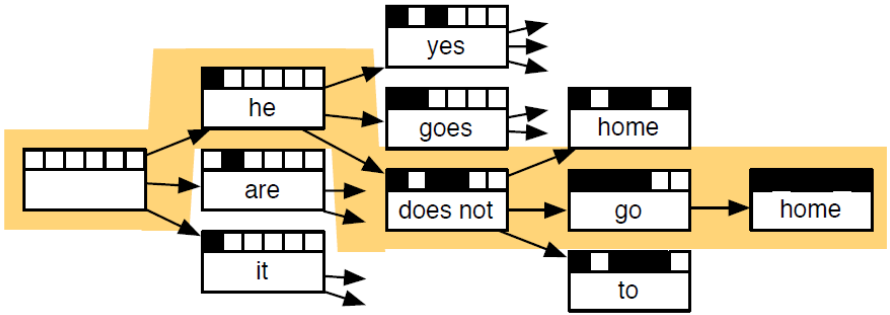
$$\underset{y}{arg\max}\, P(x, a|y)P(y)$$

- Methods to compute argmax:
  - Iterate through all possible y to compute the probabilities
    - Way too expensive!
    - Waaaaaaay too expensive!!!!!!!!!!!
  - Heuristic search algorithm that slowly, step-by-step builds up a translation and ignores unlikely translation paths

Heuristic Search







Machine Translation as ML Problem III

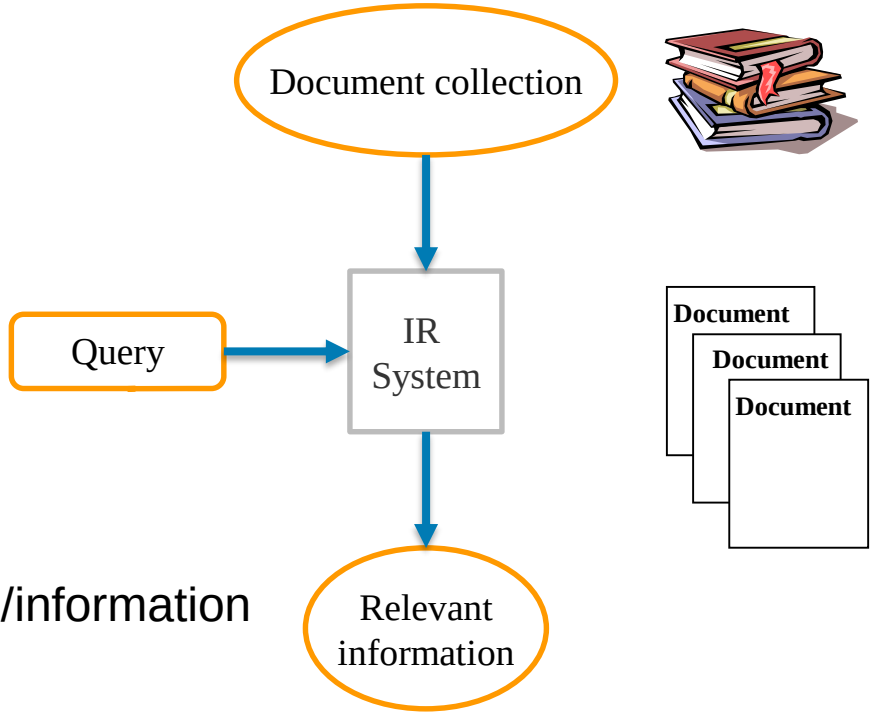
- SMT is a huge research field
  - Own, specialized conferences, challenges, …
- Best SMT-systems are very complex
  - Easy to fill a whole semester!
  - Typically many independent components
  - A lot of feature engineering
    - Depending on involved languages
  - Additional resources needed
    - Equivalent phrases, dictionaries, synonyms, …
    - Need to be created and maintained
  - A lot of manual effort
    - Development and maintainance of whole system
    - For each pair of languages seperately!

Information Retrieval Tasks

- Adhoc Retrieval
- Information Extraction (IE)
- Question Answering (QA)
- Automatic Summarization
- Recommender Systems (RS)

Information Retrieval (IR)

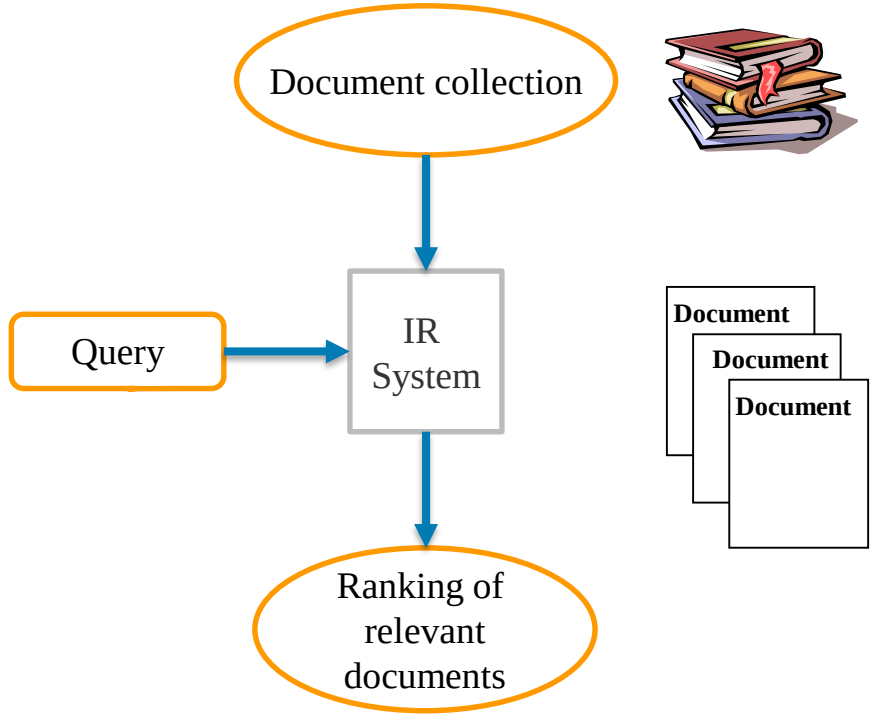
- Given:
  - Document collection
  - Query
    - Implicit
    - Explicit



- Goal:
  - Ranking of facts/documents/sentences/information
  - Sorted by relevance wrt. query
  - Optional with a score

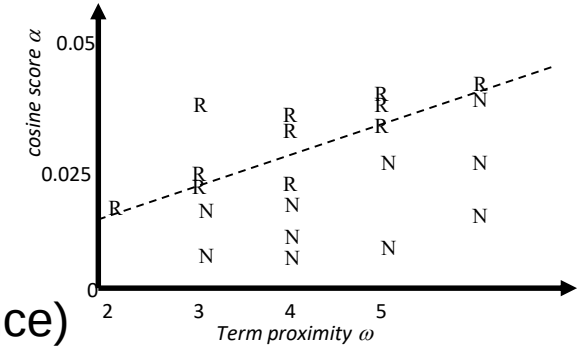
Ad-hoc Retrieval

- Given:
  - Document collection
  - User query
    - Keywords
    - Free text
- Special case: **web search**
- Goal:
  - Ranking of documents
  - Sorted by relevance wrt. query
  - Optional with a score



Ad-hoc Retrieval as ML Problem I

- Traditionally used for ad hoc retrieval:
  - Calculation of similarities (query - document)
  - Vector space model
  - Additional features
  - Weighting of features and parameter adjustment of similarity measure by hand (trial and error + experience)
  - A ranking problem
- (Supervised) machine learning
  - Classification of documents into relevant vs. non-relevant.
  - Issues:
    - Dependent on query: features must be independent
    - No ranking
  - But: Weighting and parameters can be learned with this method



Nallapati, R. Discriminative models for information retrieval. In Proceedings of *SIGIR, 2004.*

Ad-hoc Retrieval as ML Problem II

- Document classification not quite right for Ad-hoc IR:
  - Classification: assigning a document to an unordered set of a class.
  - Regression: mapping to a real number
  - **Ordinal regression**: mapping to an ordered set of classes
- Advantage of this problem formulation:
  - The relationships between relevance levels can be modeled
  - Documents are not absolutely relevant, but only relative to other documents and for a specific query.
- Training data from query log data consisting of features of query-document pairs and relevance estimation
- "Learning to Rank" (LtR, L2R, LetoR)
  - Pointwise, pairwise, listwise

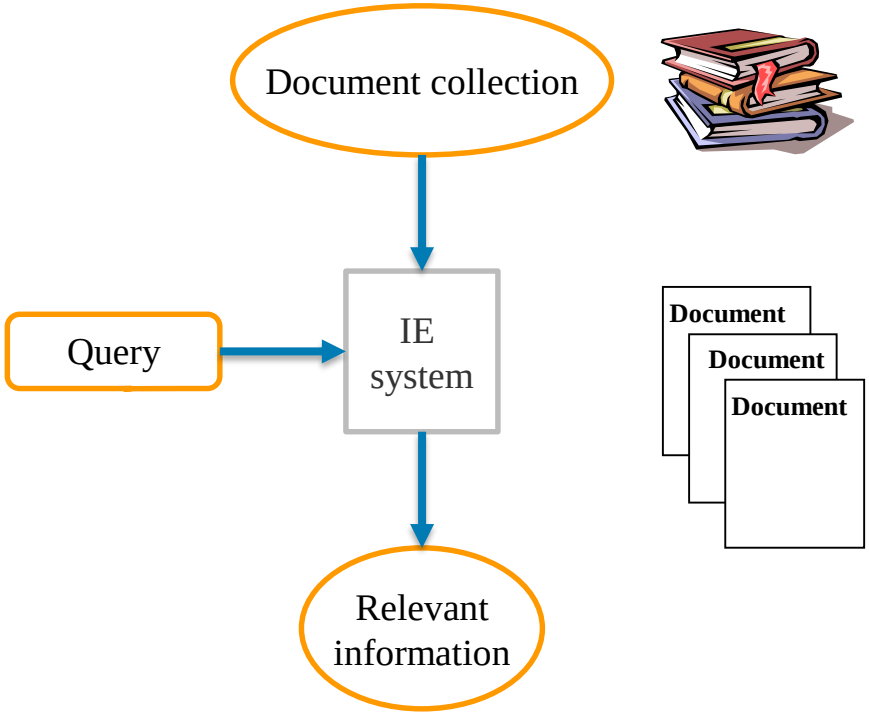
A rather obscure subfield of statistics, but just what we need

Information Extraction (IE)

- Given:
  - Document collection
  - User query
    - Clearly defined
    - Limited

- Goal
  - Set of information
  - In a given, structured output format
  - Optional: probabilities/score

Examples?



Information Extraction

- Introduced/Formalized by Message Understanding Conferences (MUC)
  - Challenges organized by DARPA 1987 – 1997
  - Various extraction tasks based on news articles
    - Nautical operations
    - Terrorist activities
    - Microelectronics
    - Persons in companies
    - Space travel
  - Benchmark evaluation
  - Structured output

<doc>
<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO> Copyright (c) 1989 Jiji Press Ltd.; </SO>
<TXT> BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO JAPAN. THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO., CAPITALIZED AT 20 MILLION NEW TAIWAN DOLLARS, WILL START PRODUCTION IN JANUARY 1990 WITH PRODUCTION OF 20,000 IRON AND "METAL WOOD" CLUBS A MONTH. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 50,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID. THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID. BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUB PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES. WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</doc>

<ORGANIZATION-0592-3> :=
ORG_NAME: "TAGA CO."
ORG_DESCRIPTOR: "A JAPANESE TRADING HOUSE“
/"A COMPANY ACTIVE IN TRADING WITH TAIWAN“
ORG_TYPE: COMPANY
ORG_NATIONALITY: JAPAN

<ORGANIZATION-0592-4> :=
ORG_NAME: "BRIDGESTONE SPORTS TAIWAN CO.“
ORG_TYPE: COMPANY
ORG_DESCRIPTOR: "A JOINT VENTURE“
ORG_LOCALE: KAOHSIUNG CITY /KAOHSIUNG PROVINCE
ORG_COUNTRY: TAIWAN

Information Extraction as ML Problem

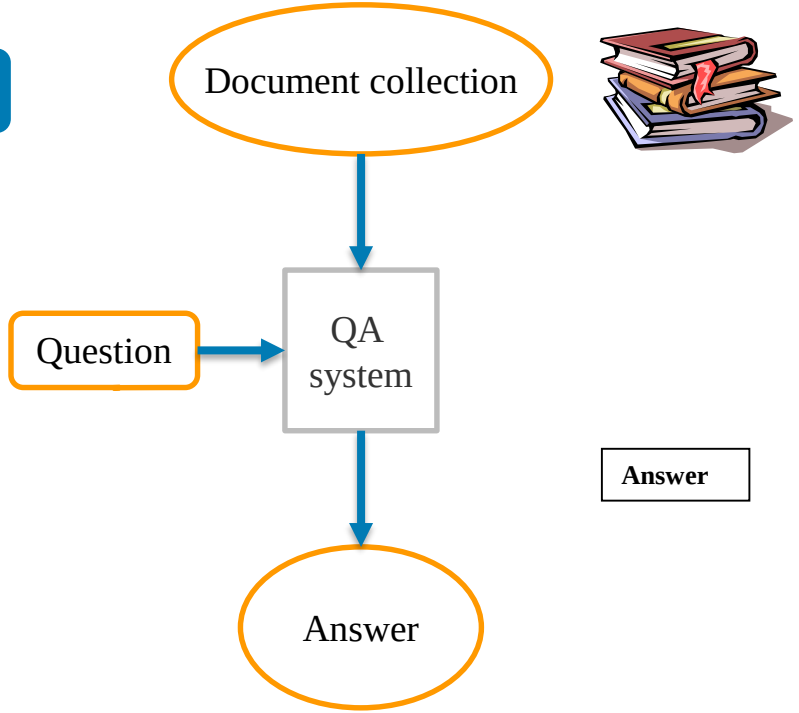
- Does a word belong to a certain piece of information?
  - Binary classification (Yes/No)
  - Typically multi-class classification
  - E.g. Named Entity Recognition (NER)
    - Beginning of a NE; Part of a NE; No NE
  - Can also be more complex: Relation extraction, Event extraction.
  - Traditional: Simple manually created rules based on
    - POS tags, case sensitivity, dictionaries, ...
- With ML:
  - Learning these rules using training data
  - Hidden Markov models (probability for certain classes)
  - Conditional Random Fields (inclusion of multiple features)

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282.289.

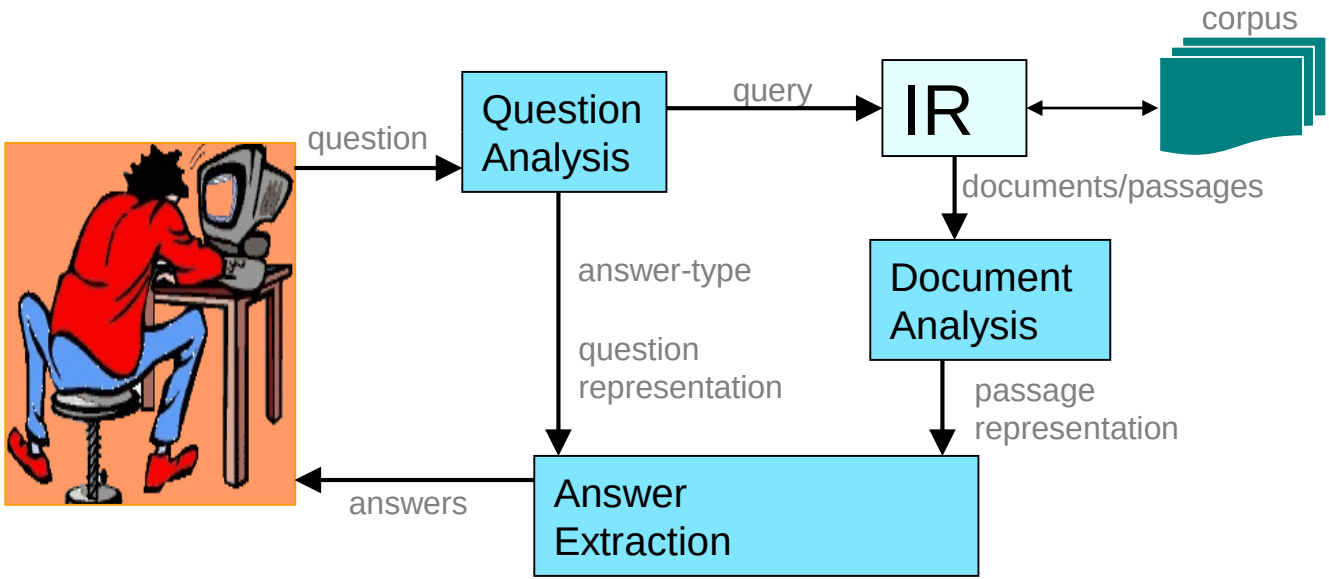
Question Answering (QA)

- Given:
  - Document collection
  - User query
    - Grammatically correct Question
    - Whole sentence
- Special case: document collection = web
- Goal:
  - Grammatically correct answer
  - Whole sentence
  - Optionally: Top-k most likely answers

Examples?



QA: System Architecture



QA: Systems

- IR-based systems
  - Rely on large amounts of information on the web or in ontologies
  - Typically two main components:
    1. Finding relevant documents (classical IR)
    2. Finding the answer in a document
- Knowledge-based systems
  - A knowledge base (KB) contains triple;
    - e.g., extracted from Wikipedia infoboxes
  - **Mapping a question to a query via the KB**
- Hybrid systems
  - Use multiple sources: Text and KBs
  - DeepQA, IBM

Better suited for open-domain QA

Better suited for closed-domain QA

Question Answering as ML Problem

- Traditionally a system with individual components
  - Determination of the question type
    →**Classification** (categorization)
  - Finding relevant documents
    →**IR**
  - Analysis of documents
    →**NLP**
    - KBC (knowledge base completion)
  - Extracting answers
    →**IE**
    - Reading comprehension

Passage: One day, I was studying at home. Suddenly, there was a loud noise...A building in my neighborhood was on fire...A few people jumped out of the window... Those who were still on the second floor were just crying for help...Firefighters arrived at last. They fought the fire bravely. Water pipes were used and a ladder was put near the second-floor window. Then the people inside were taken out by the firefighters...Thanks to the firefighters, the people inside were saved and the fire was put out in the end, but many things, such as desk, pictures and clothes, were damaged.

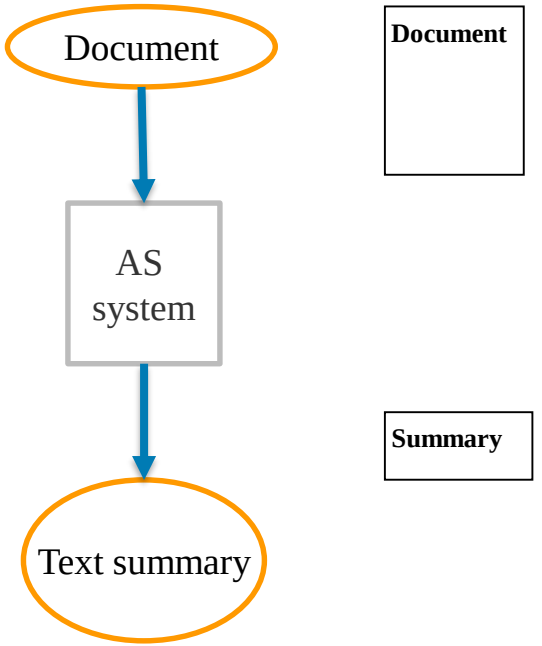
Question: *How did the people who didn't jump out of the window get out of the building?*

Option A: They were taken out by the firefighters.
Option B: They climbed down a ladder by themselves.
Option C: They walked out after the fire was put out.
Option D: They were taken out by doctors
Correct Option: A

Automatic Summarization

- Given:
  - One document
    - Or multiple documents (multi-doc summarization)
  - Optionally: a question/topic (focused summary)
  - Max length

- Goal:
  - Summary of content
  - Grammatically correct sentences



Automatic Summarization as ML Problem

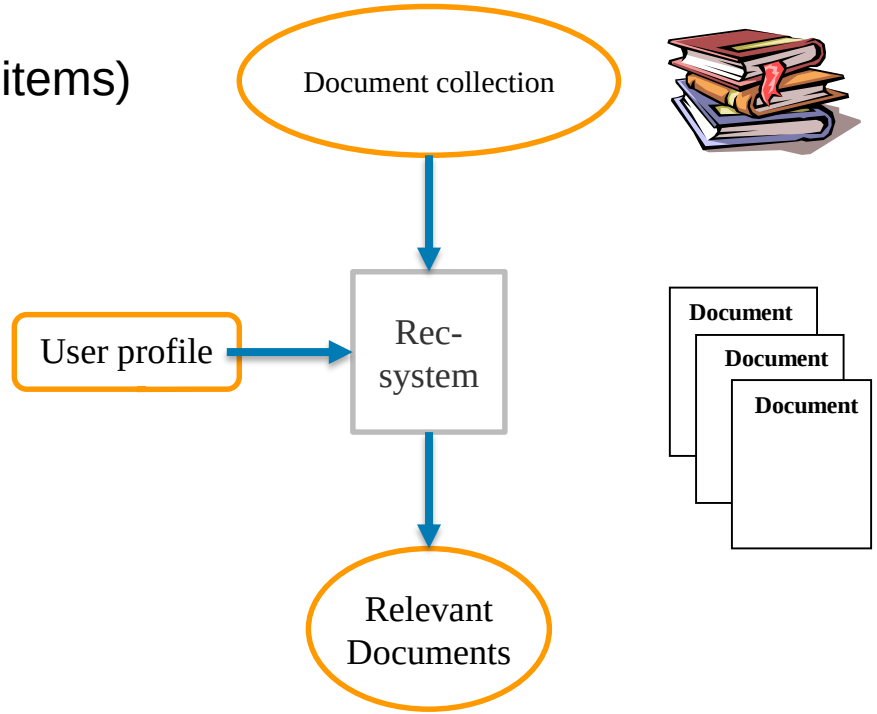
- Extractive
  - Find salient sentences (or parts)
  - Put them together to form a coherent text
  - Based on heuristics
    - Position in text
    - Entities
    - Tf*idf
    - …

- Generative
  - Understand the text
  - Generate a condensed version of the text
  - Based on template filling

Recommender Systems (RS)

- Given:
  - Document collection (more generally: items)
  - Implicit user query
    - Context
    - History
    - …

  } User profil



- Goal:
  - Ranking of documents (items)
  - Sorted by relevance to user
  - Optionally with score

RS as ML Problem

- Find similar items
  - Knowledge-based
  - Content-based
    - Clustering
  - Collaborative/community-based
    - Collaborative filtering (CF)
  - Hybrid

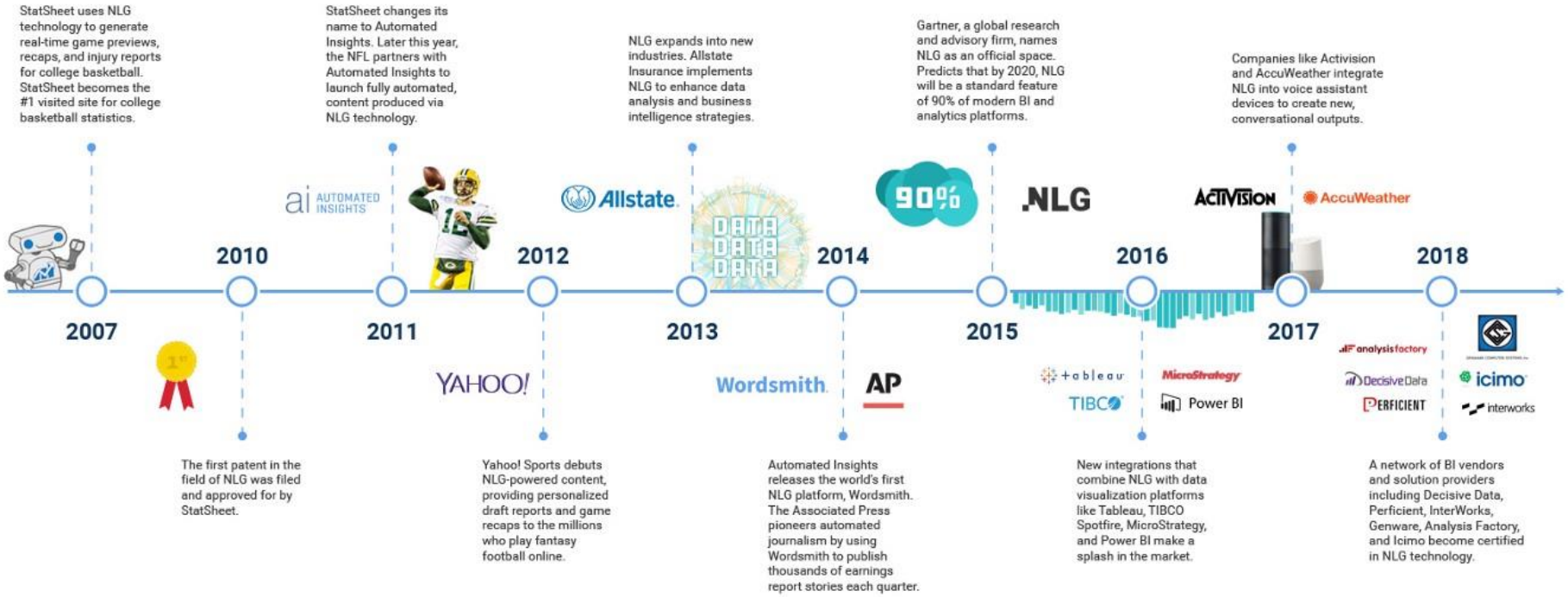



https://miro.medium.com/max/623/1*hQAQ8s0-mHefYH83uDanGA.gif

Exercise

- How can IR be formulated as deep learning problem?
  - Input?
  - Output?
  - What measure should be optimized?

History of Natural Language Generation (NLG)



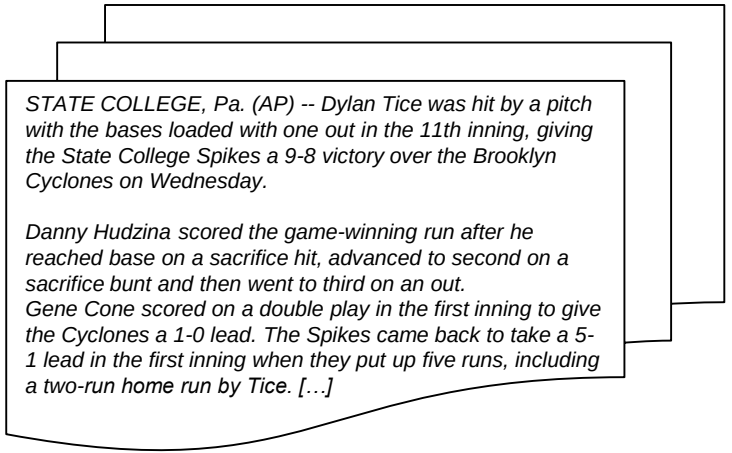
https://medium.com/@AutomatedInsights/the-history-of-natural-language-generation-5b4c3fa2f9f9

NLG Systems

- NLG as an author
  - Story telling, jokes, art
  - Automated journalism
    - Weather reports
    - Stock market descriptions
    - Sports game summarization
  - Museum artifacts descriptions
  - Customer relationship management
    - Personal letters to customers
    - Chatbots
  - Summarization
    - Extractive and generative

https://www.ap.org/press-releases/2016/ap-expands-minor-league-baseball-coverage

- NLG as an author aid
  - NLG in augmentative and alternative communication
  - Machine translation (generation from interlingua)



What is Natural Language Understanding?

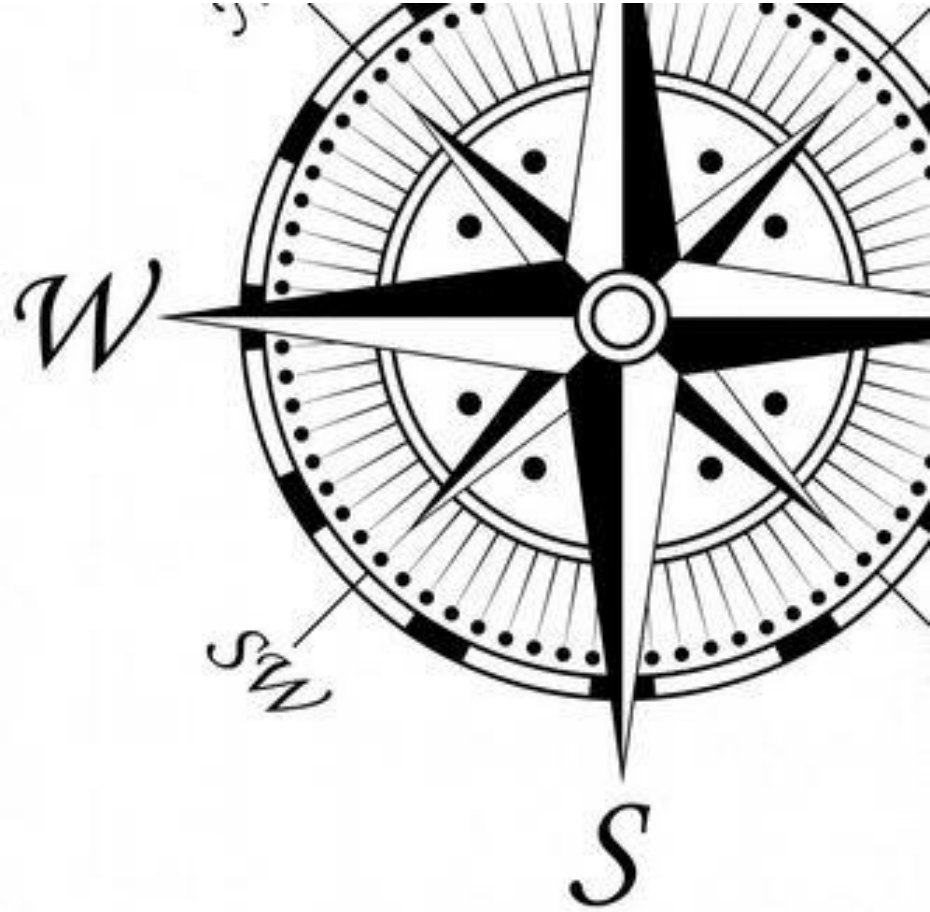
- “Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural language concepts.”

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

Other examples:
- Numeric database
- Ontology
- Knowledge graph

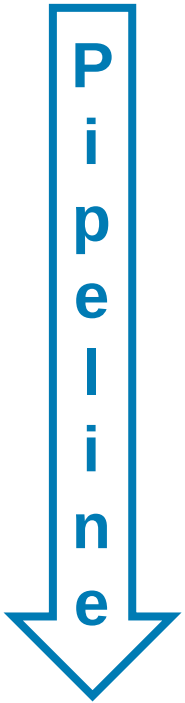
Topics Today

1. Natural Language Processing Pipeline
2. Text Mining Applications
3. **Summary**



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



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

Learning Goals for this Chapter

- Describe a 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 standard NLP tasks
- Explain text mining and identify TM tasks
- List various text mining tasks, naming
  - Difficulties/challenges
  - Traditional approaches
- Know the limitations of traditional TM applications

https://www.kdnuggets.com/images/cartoon-turkey-data-science.jpg

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/