VL Deep Learning for Natural Language Processing

5. Text Mining

Prof. Dr. Ralf Krestel

AG Information Profiling and Retrieval

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."

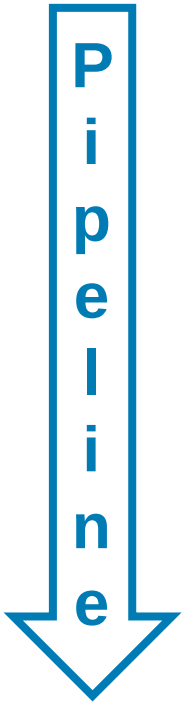
https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

- 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

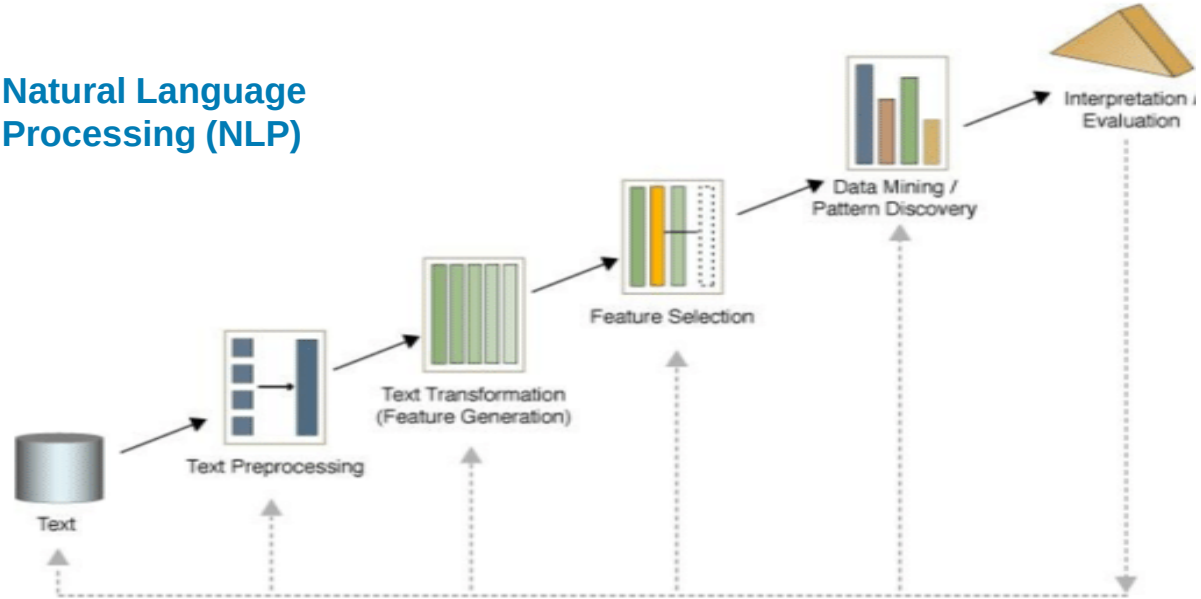


- (Text Mining) Applications
  - Document Classification
  - Document Clustering
  - Machine translation (MT)
  - Topic Modeling
  - Information retrieval (IR)
    - Information extraction (IE)
    - Question answering (QA)
    - Automatic summarization
    - Recommender Systems (RS)
  - Knowledge Graphs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

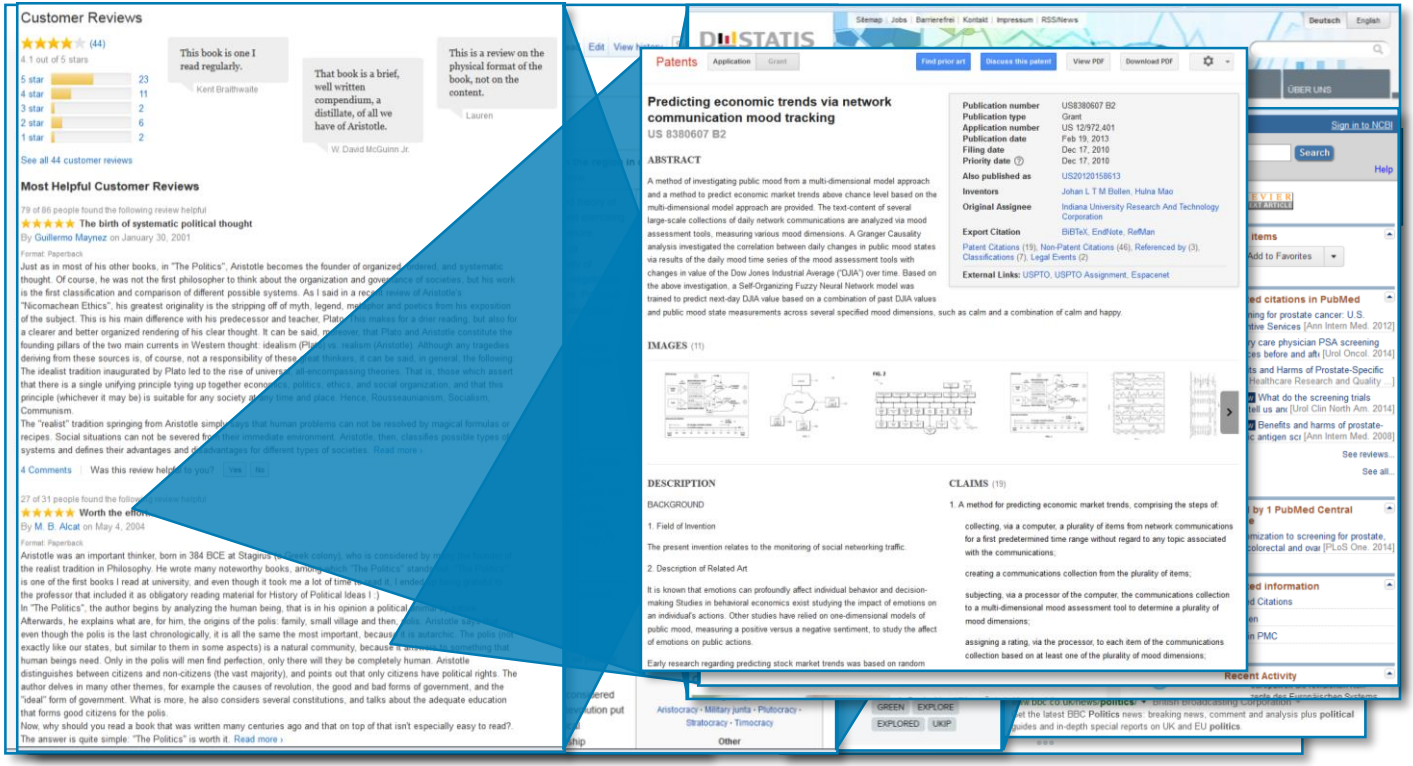
Text Mining ~ Data Mining with Text

- 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)



Different Types of Text



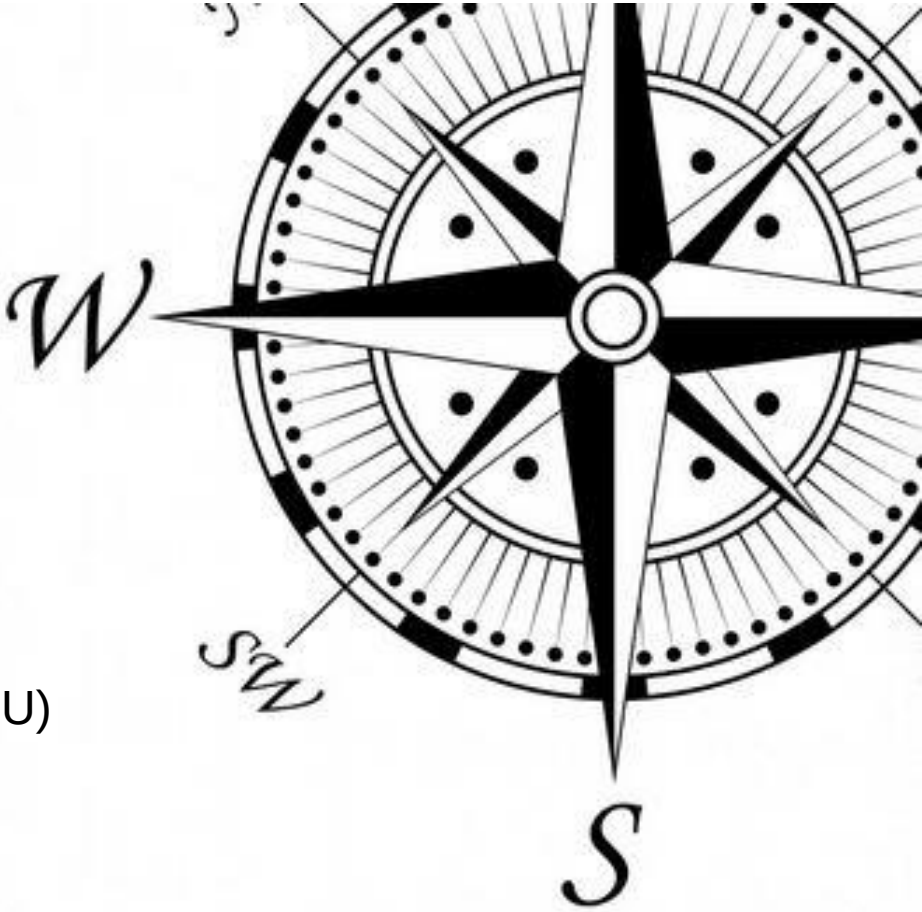
Lerning Goals for this Chapter

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

- Be able to explain text mining
- Identify text mining tasks
- Being able to list various text mining tasks, naming
  - Difficulties/challenges
  - Traditional approaches
- Know the limitations of traditional text mining applications

Topics Today

1. **Document Classification**
2. Document Clustering
3. Topic Modeling
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)

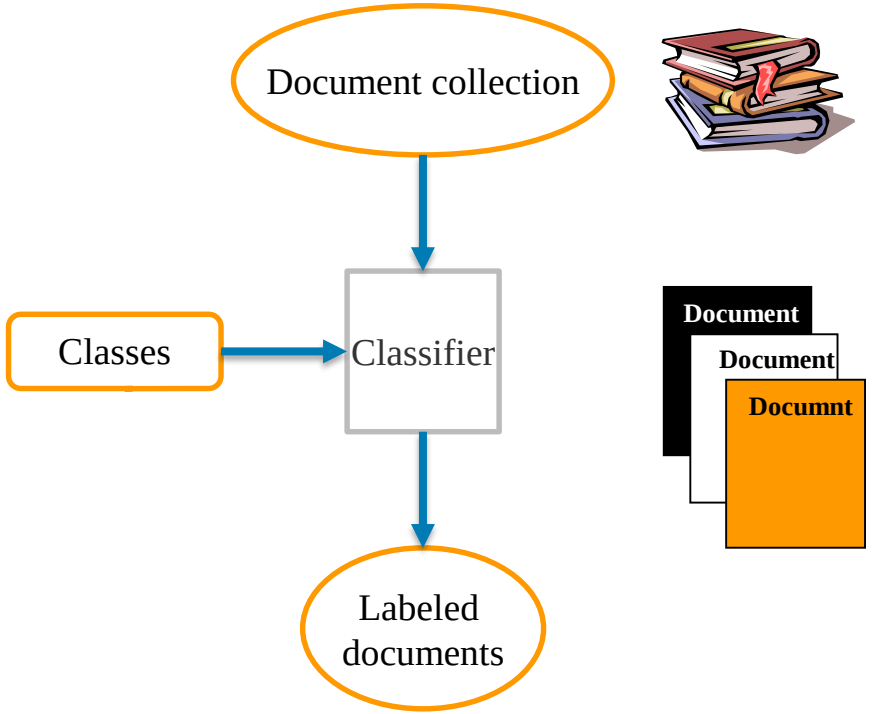


Document Classification

- Given:
  - Collection of documents
  - Different classes

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

  - Single- vs. multi-label classification

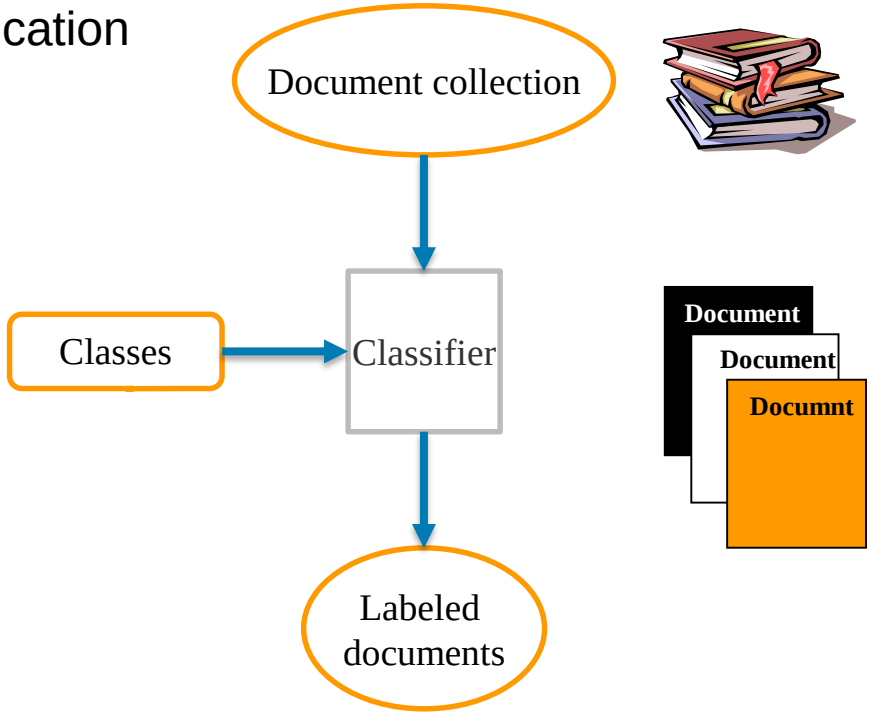


Sentiment Analysis

- 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 $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1|c) \cdot P(w_2|c) \cdot \ldots \cdot P(w_n|c) \cdot P(c)$
  - ⟶ Reformalating $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in \{1 \ldots n\}} P(w_i|c)$
  - ⟶ 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

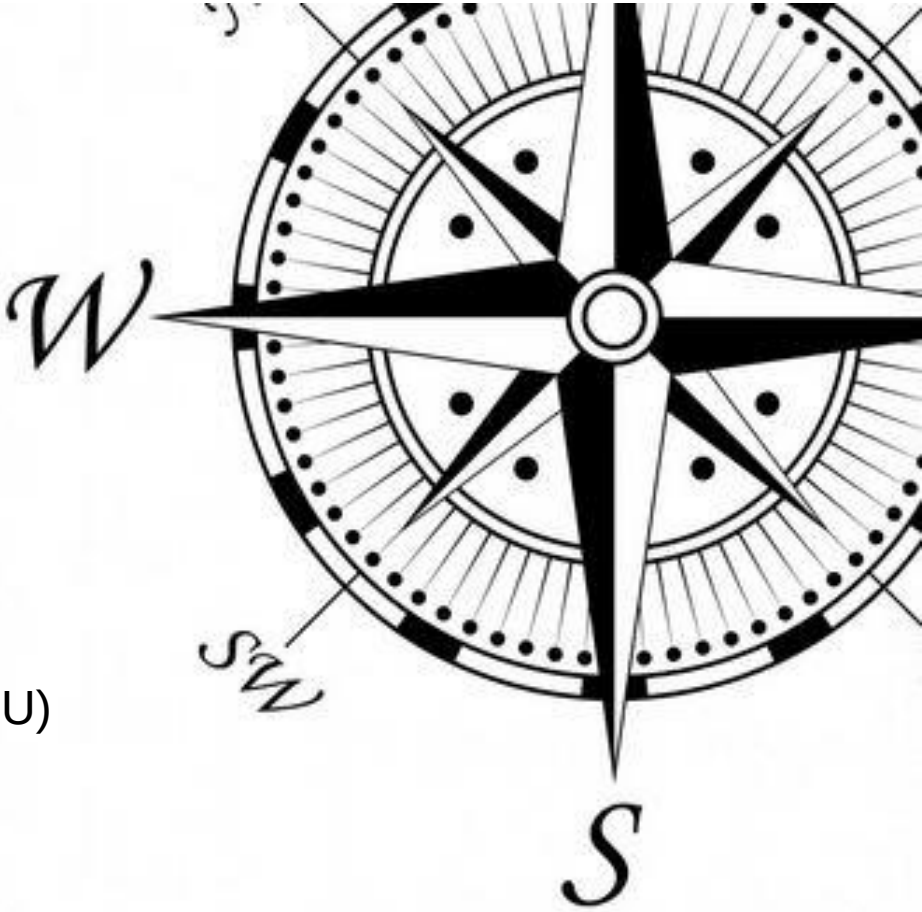
How to estimate these probabilities?

Exercise

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

Topics Today

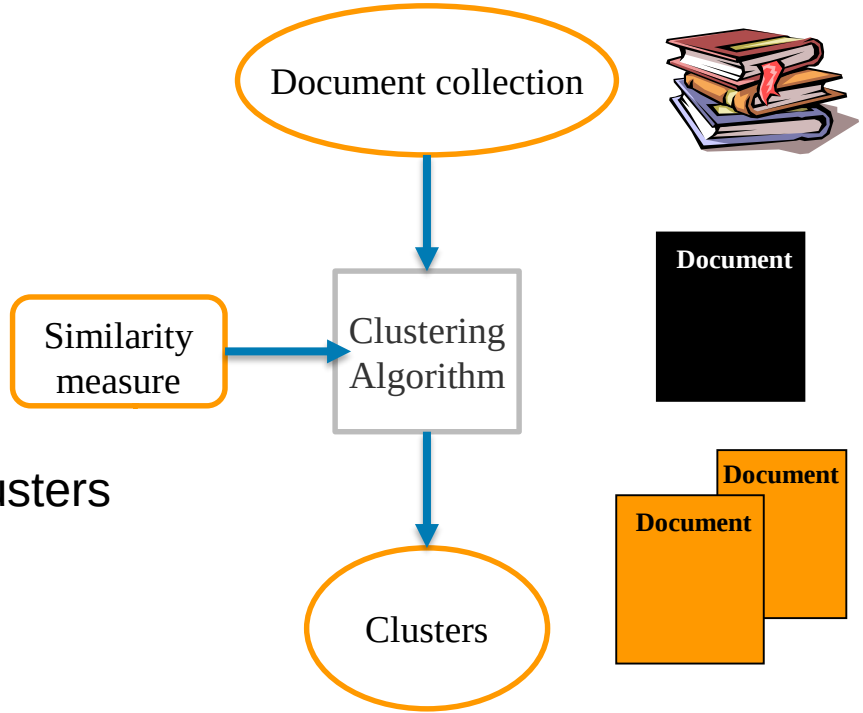
1. Document Classification
2. **Document Clustering**
3. Topic Modeling
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)



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



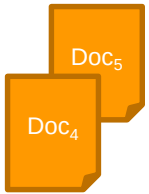


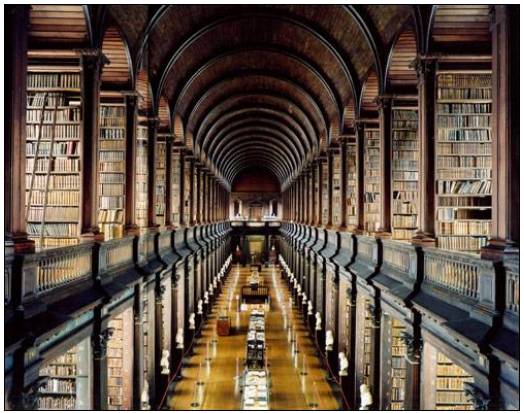



Representation

- 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



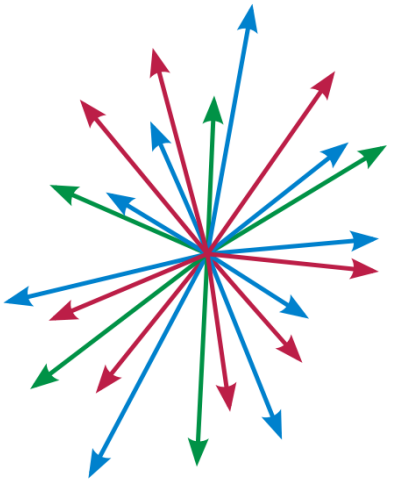


Vector Space Model I

- 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

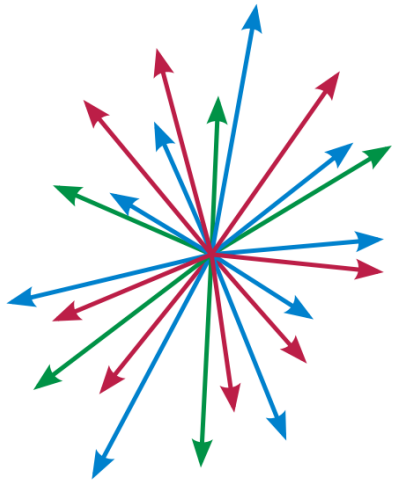


https://commons.wikimedia.org/wiki/File:Vector_space_illust_rotate.svg

Vector Space Model I

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

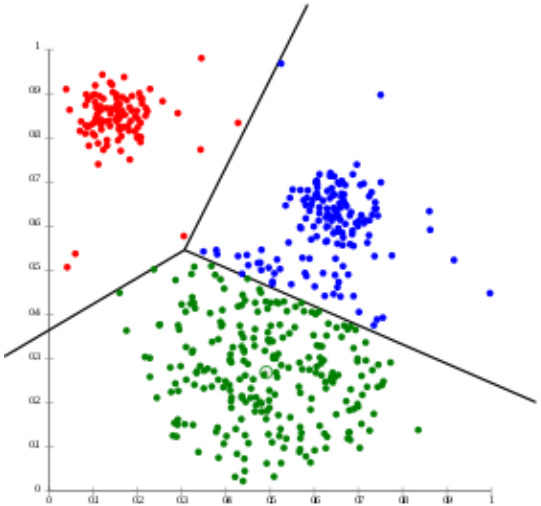
	$Term_1$	$Term_2$	$\ldots$	$Term_t$
Doc_1	d_{11}	d_{12}	$\ldots$	d_{1t}
Doc_2	d_{21}	d_{22}	$\ldots$	d_{2t}
$\vdots$	$\vdots$			
Doc_n	d_{n1}	d_{n2}	$\ldots$	d_{nt}



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

Document Clustering as ML Problem

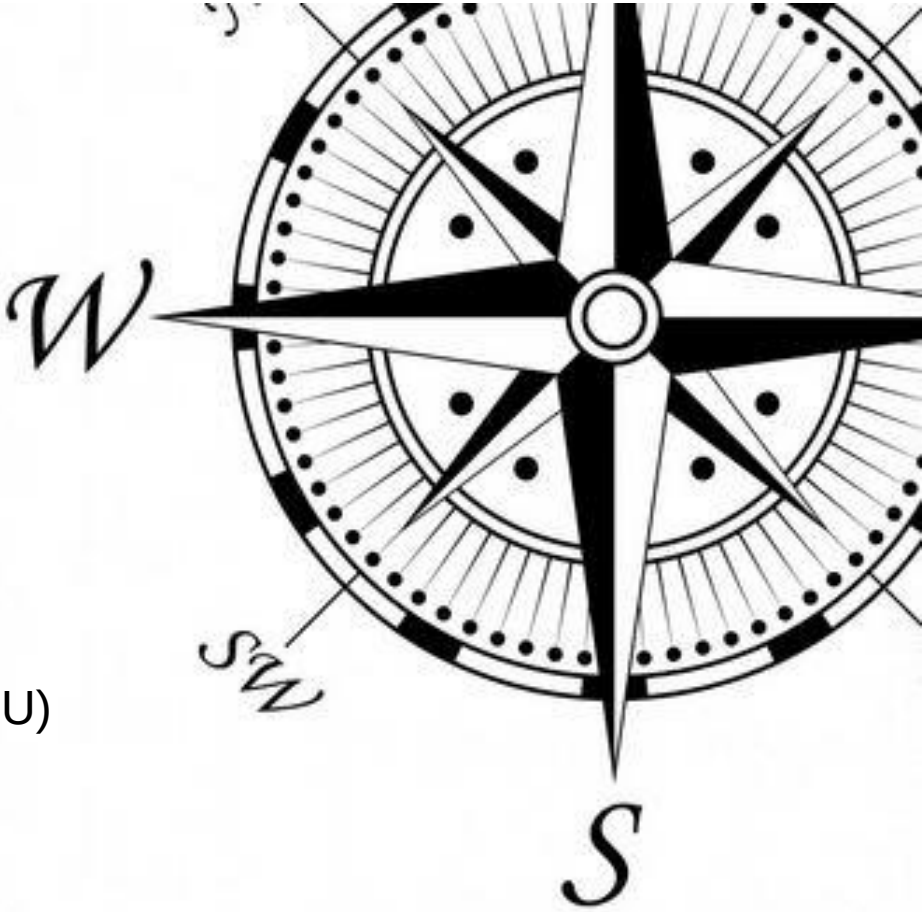
- Given: documents in a vector space
- Goal: Group similar documents together
  - Similarity ≅ Close distance in vector space
- Necessary: distance function (similarity measure)
  - Euclidian distance
  - Manhattan distance
  - …
- Many clustering algorithms
  - „Hard“ (non-overlapping)
    - E.g. k-means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



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

Topics Today

1. Document Classification
2. Document Clustering
3. **Topic Modeling**
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)



Probabilistic Topic Models

Document	Topics
424	6,8
143	1,2,7
449	3,5
756	2,7
…	…

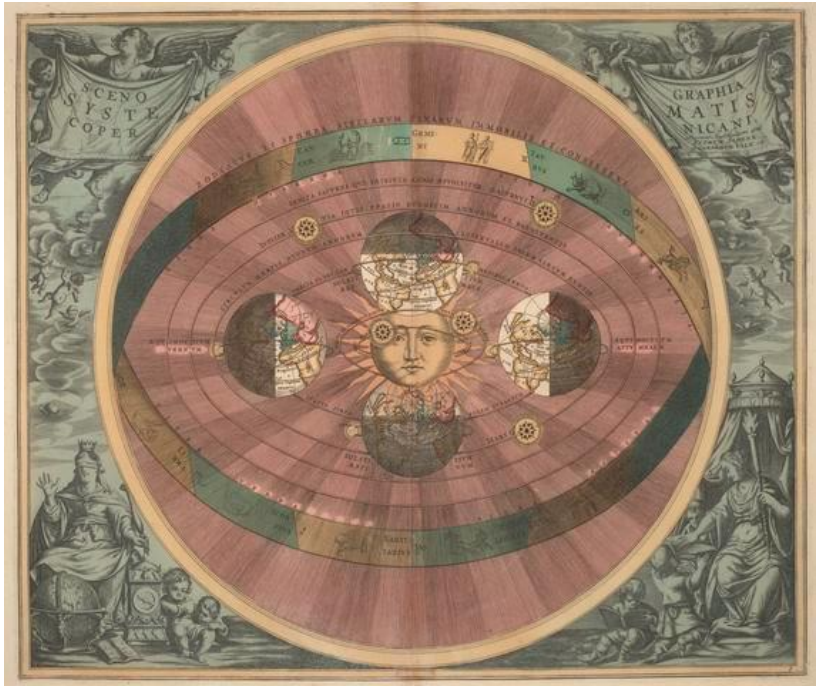
- Summarize
- Visualize
- Predict
- Personalize
- Analyze

Topic	Most Probable Terms
1	algorithm problem point cluster code vector graph datum
2	cell model orientation cortex neuron cortical input map
3	circuit chip analog input output current figure voltage
4	control model system learn motor movement trajectory
5	function theorem bound number result algorithm bind
6	image object feature recognition pixel face view figure
7	learn error weight training generalization noise rate result
8	network unit input output learn weight layer hidden neural
9	rule representation structure sequence tree language learn
…	

What are Topic Models?

- Astronomical model
  - E.g. heliocentric, explains day and night and some other observations

- Topic model
  - Explains why words occur together in a document



The Copernican system by Andreas Cellarius from the *Harmonia Macrocosmica* (1708)

Generative View

1.700 Articles:

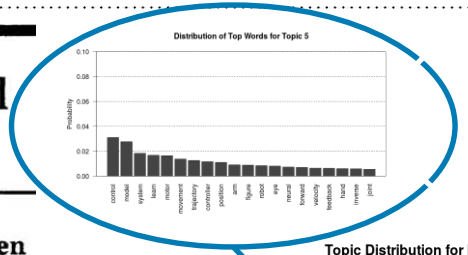
Neural Information Processing Systems

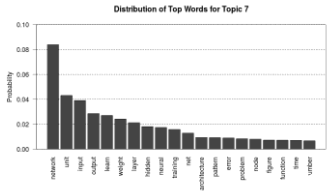
Neural Networks Structured for Control Application to Aircraft Landing

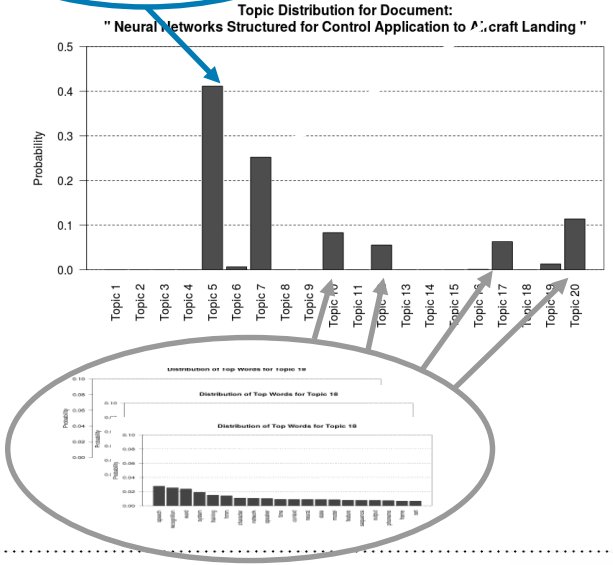
Charles Schley, Yves Chauvin, Van Henkle, Richard Golden
Thomson-CSF, Inc., Palo Alto Research Operations
630 Hansen Way, Suite 250
Palo Alto, CA 94306

Abstract

We present a generic neural network architecture capable of controlling non-linear plants. The network is composed of dynamic, parallel, linear maps gated by non-linear switches. Using a recurrent form of the back-propagation algorithm, control is achieved by optimizing the control gains and task-adapted switch parameters. A mean quadratic cost function computed across a nominal plant trajectory is minimized along with performance constraint penalties. The approach is demonstrated for a control task consisting of landing a commercial aircraft in difficult wind conditions. We show that the network yields excellent performance while remaining within acceptable damping response constraints.

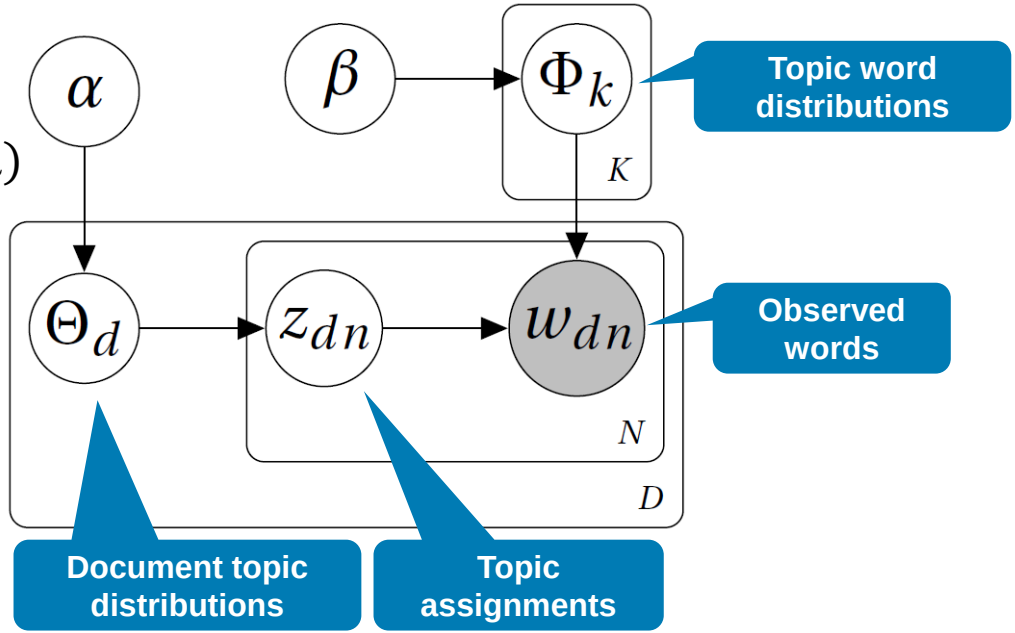






Latent Dirichlet Allocation

- Draw K topics $\Phi_k \sim Dir(\beta)$
- For each of D documents
  - Draw topic proportions $\Theta_d \sim Dir(\alpha)$
  - For each of N_d words
    - Draw topic $z \sim Mult(\Theta_d)$
    - Draw word $w \sim Mult(\Phi_z)$
- Learning the parameters/ Fitting the model via
  - Gibbs sampling [GS04]
  - Variational Bayes
  - …

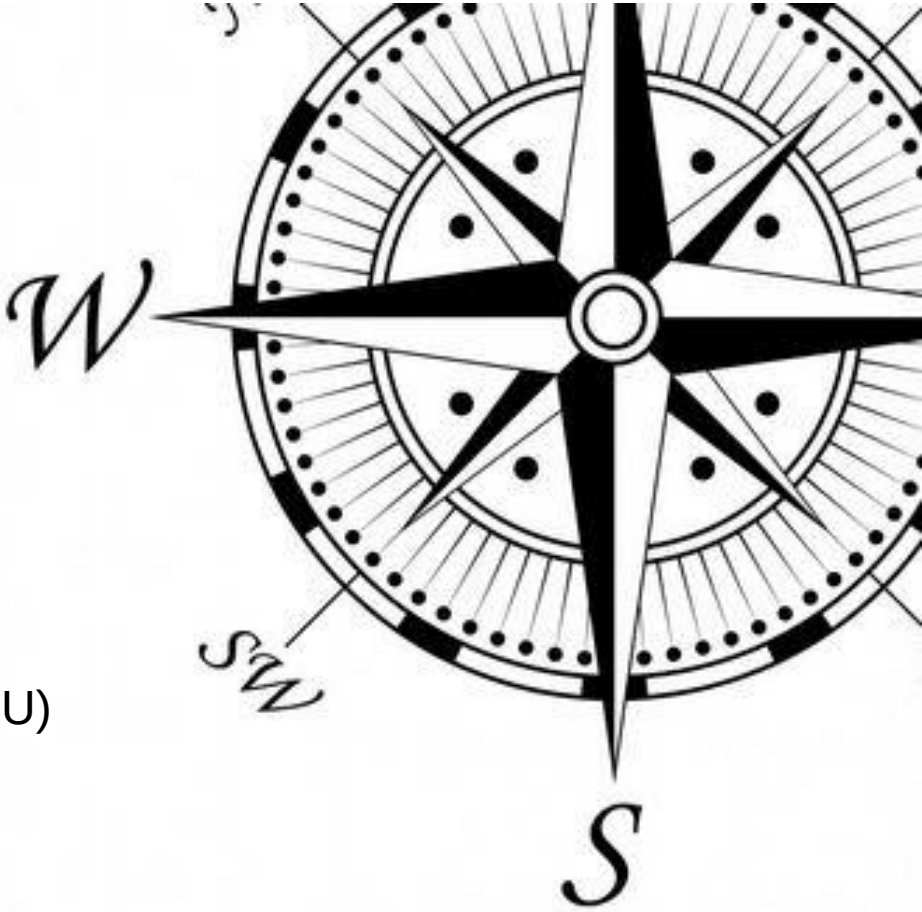


Blei, D. et al. Latent Dirichlet allocation. JMLR, 3, 2003.

Griffiths, T. and Steyvers, M. Finding scientific topics. *PNAS*, 101.suppl 1, 2004.

Topics Today

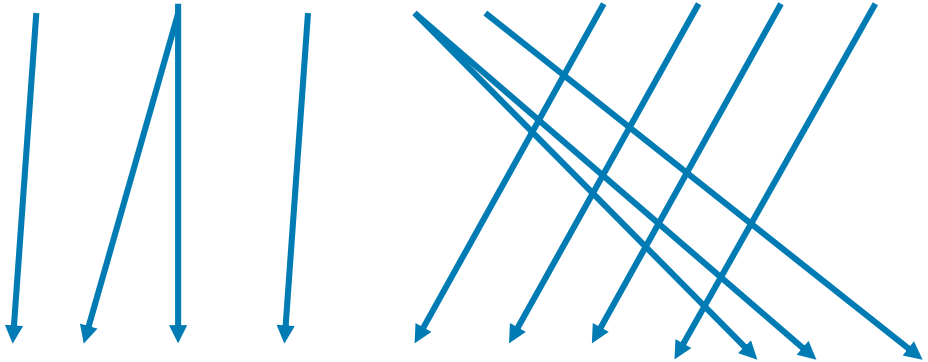
1. Document Classification
2. Document Clustering
3. Topic Modeling
4. **Machine Translation (MT)**
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)



Machine Translation

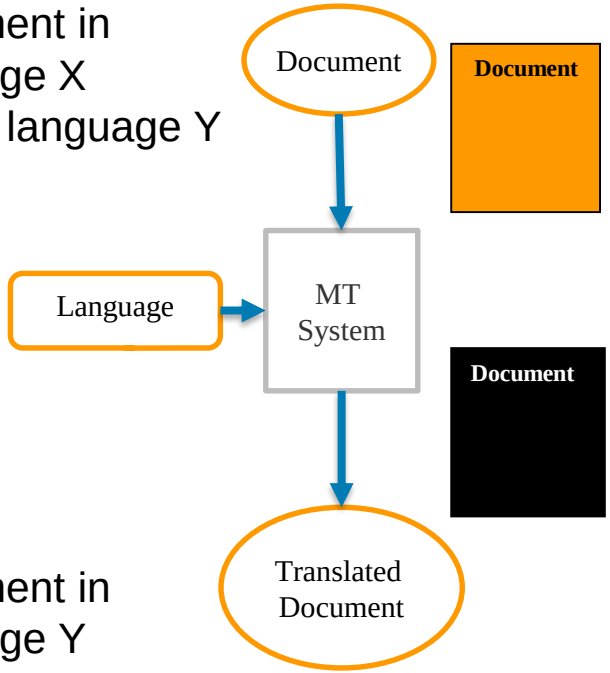
- The task of machine translation is to translate a sentence x in one language (**source**) into a sentence in another language (**target**).

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:

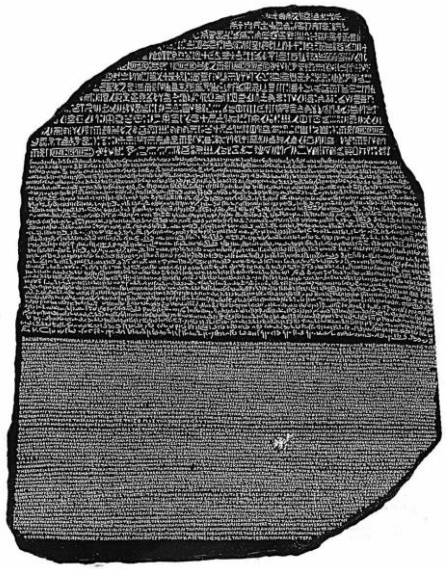
$$= \underset{y}{argmax}\, P(x|y)P(y)$$

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

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

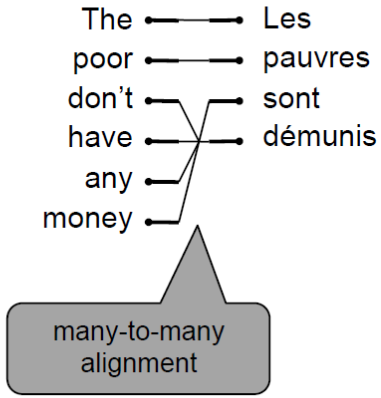
Machine Translation as ML Problem I

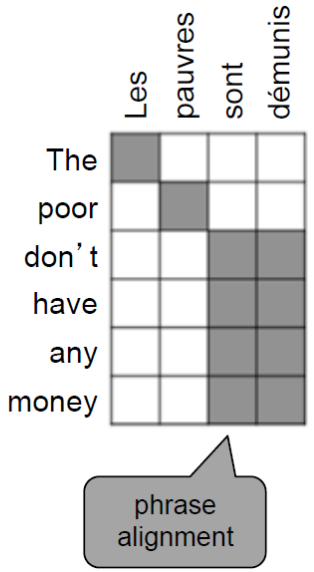
- Statistical Machine Translation (SMT)
- How to learn the translation model $P(x|y)$?
  - With a very large amount of parallel data!
- More closely, we do not want to learn $P(x|y)$, but $P(x, a|y)$, where a is an alignment.
- **Alignment** is the mapping of English words to German words within our 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



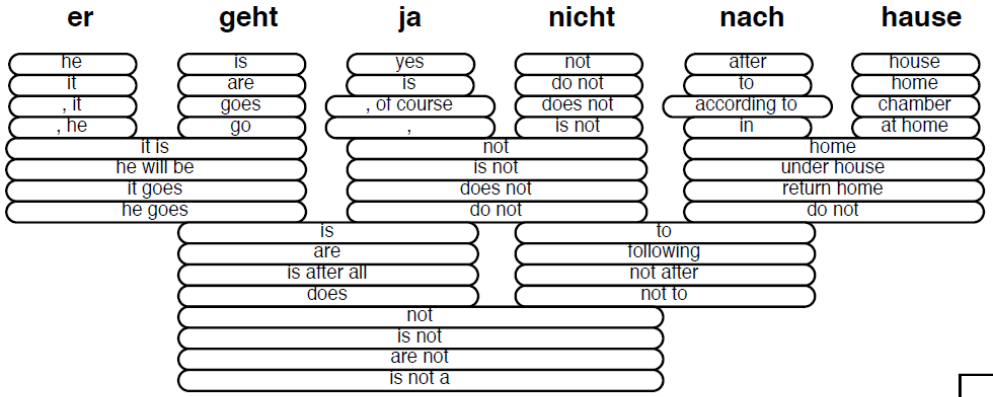


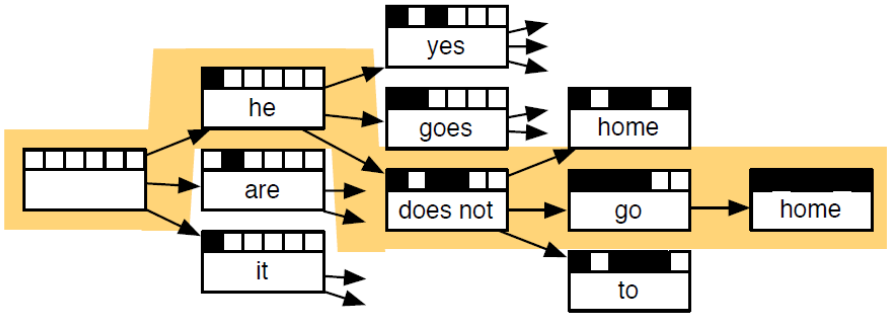
Machine Translation as ML Problem II

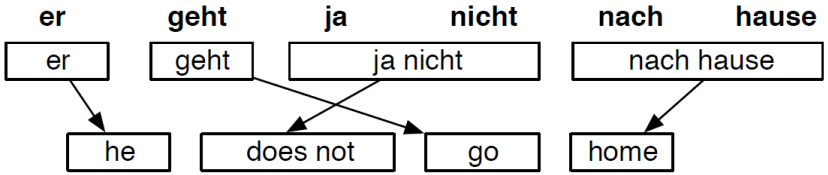
$$\underset{y}{arg\max}\, P(x, a|y)P(y)$$

- 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





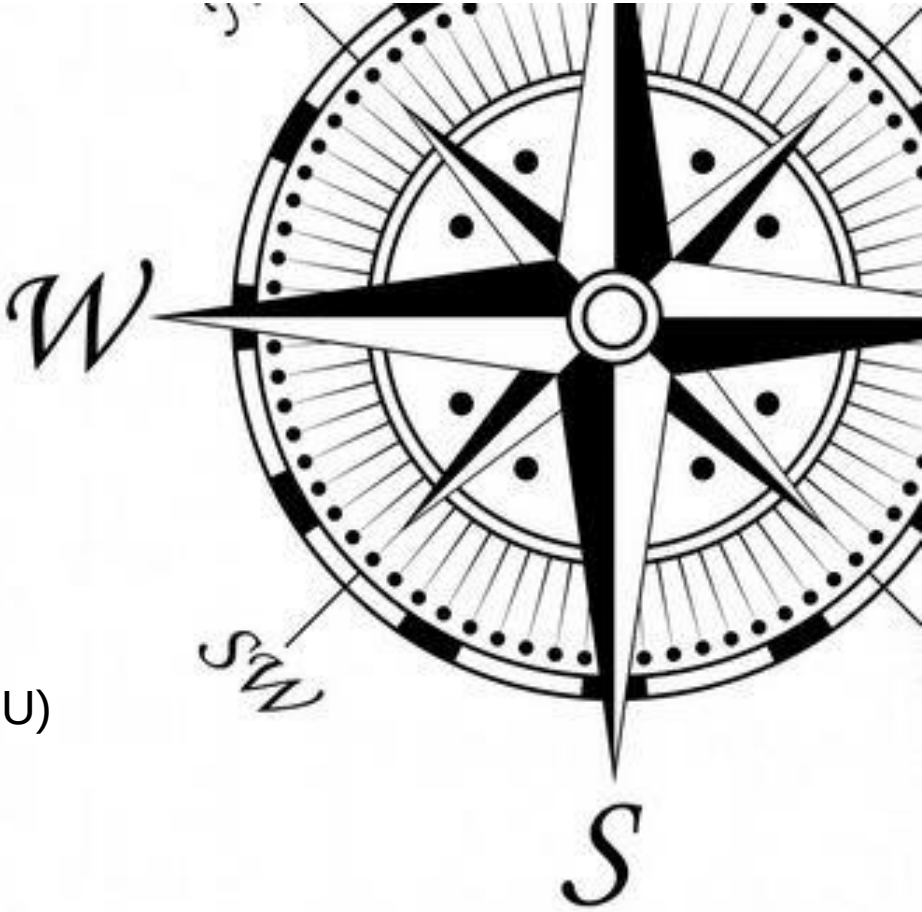


Machine Translation as ML Problem III

- 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!

Topics Today

1. Document Classification
2. Document Clustering
3. Machine Translation (MT)
4. Topic Modeling (TM)
5. **Information Retrieval (IR)**
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)

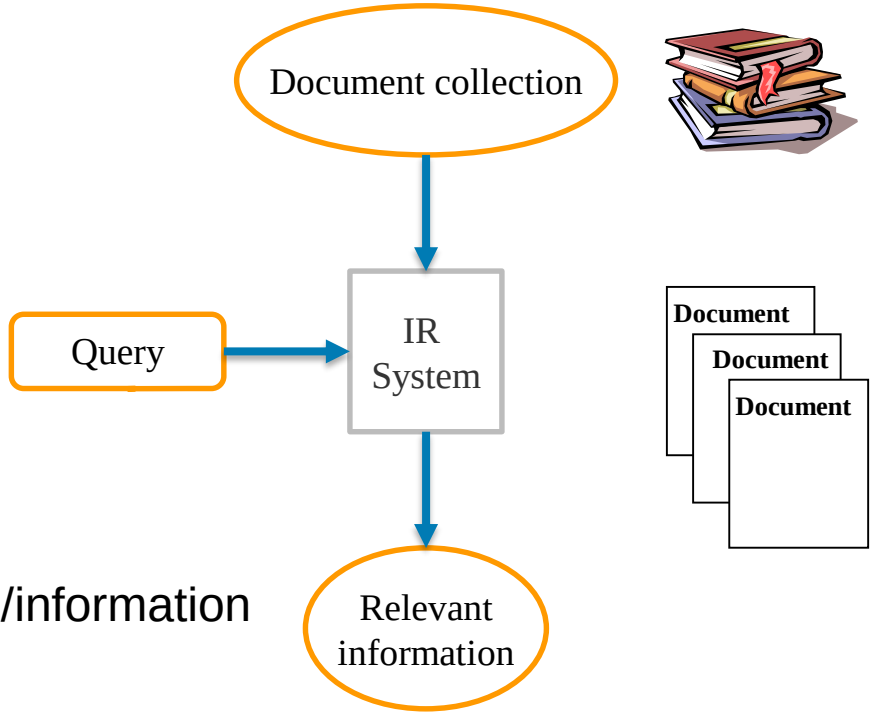


Information Retrieval Tasks

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

Information Retrieval (IR)

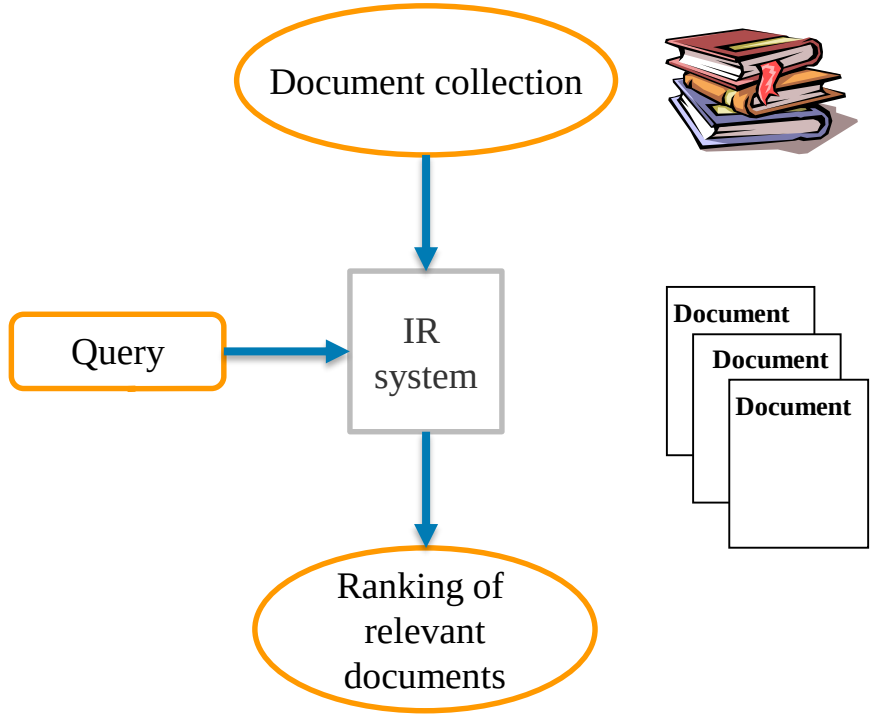
- Given:
  - Document collection
  - Query
    - Implicit
    - Explicit



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

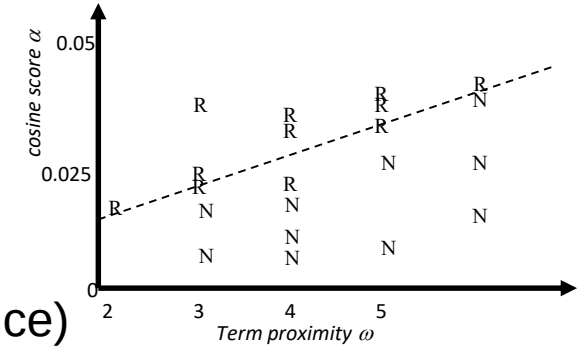
Ad-hoc Retrieval

- 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

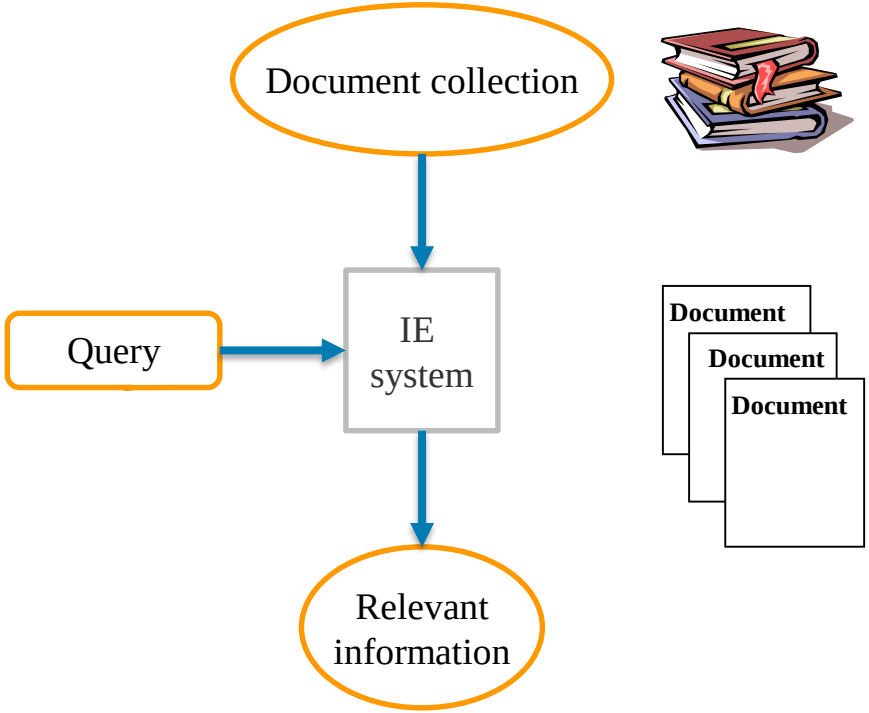
A rather obscure subfield of statistics, but just what we need

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

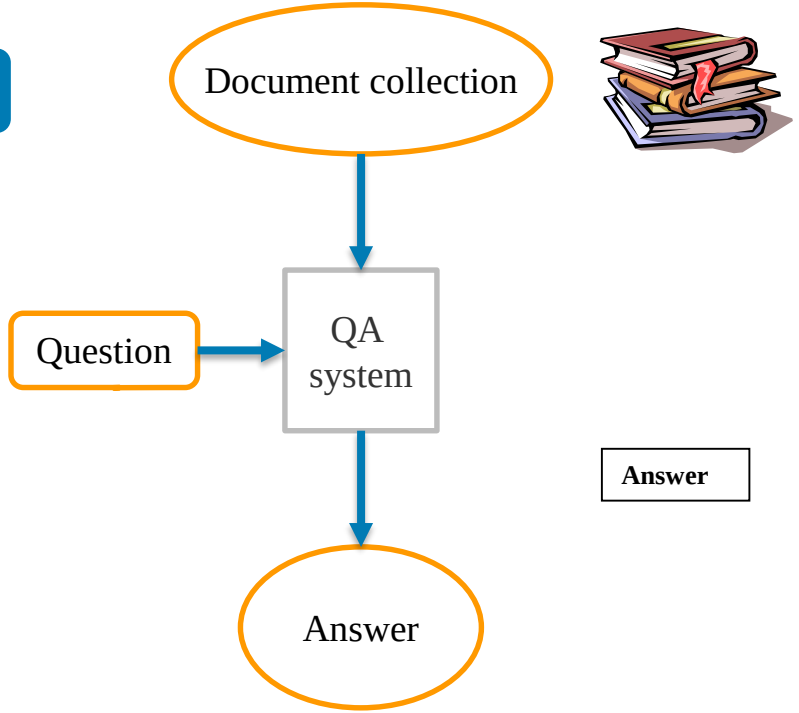
- Does a word belong to a certain piece of information?
  - Binary classification (Yes/No)
  - Typically multi-class classification
  - E.g. Named Entity Recognition (NER)
    - Beginning of a NE; Part of a NE; No NE
  - Can also be more complex: Relation extraction, Even extraction.
  - Traditional: 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.

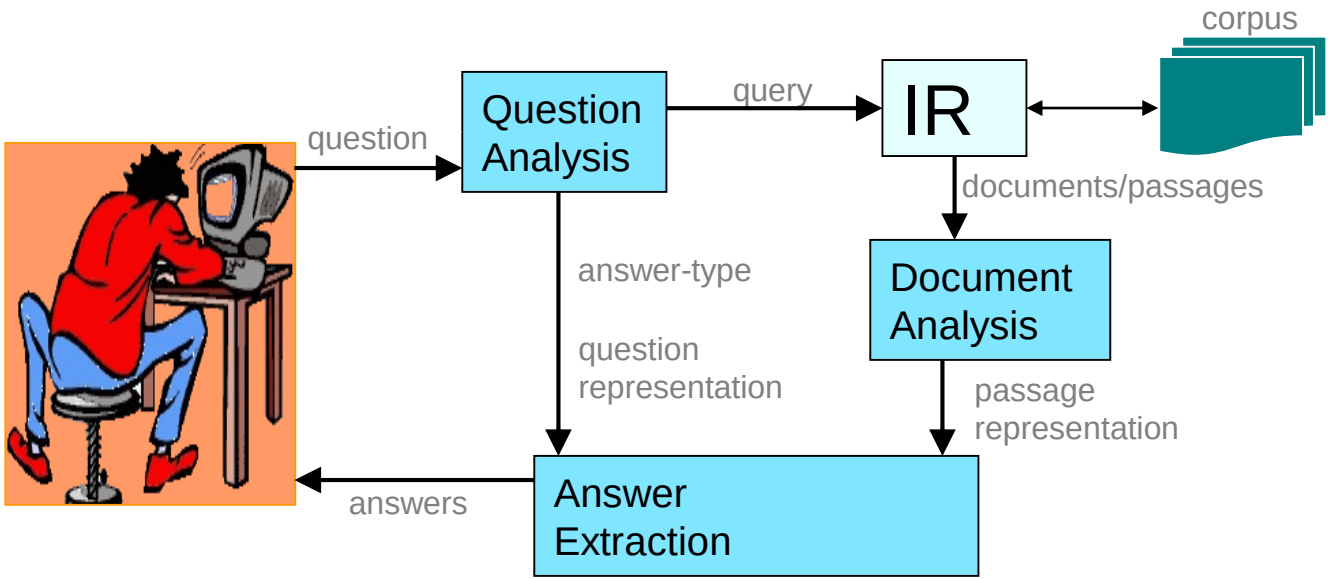
Question Answering (QA)

- 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.

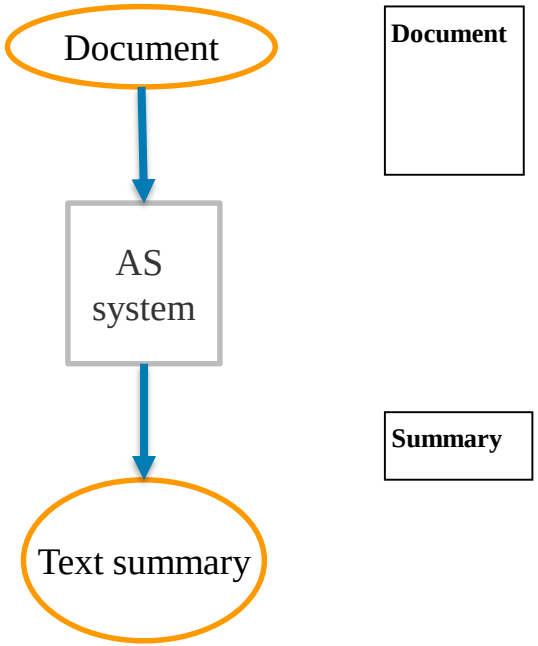
Question: *How did the people who didn't jump out of the window get out of the building?*

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

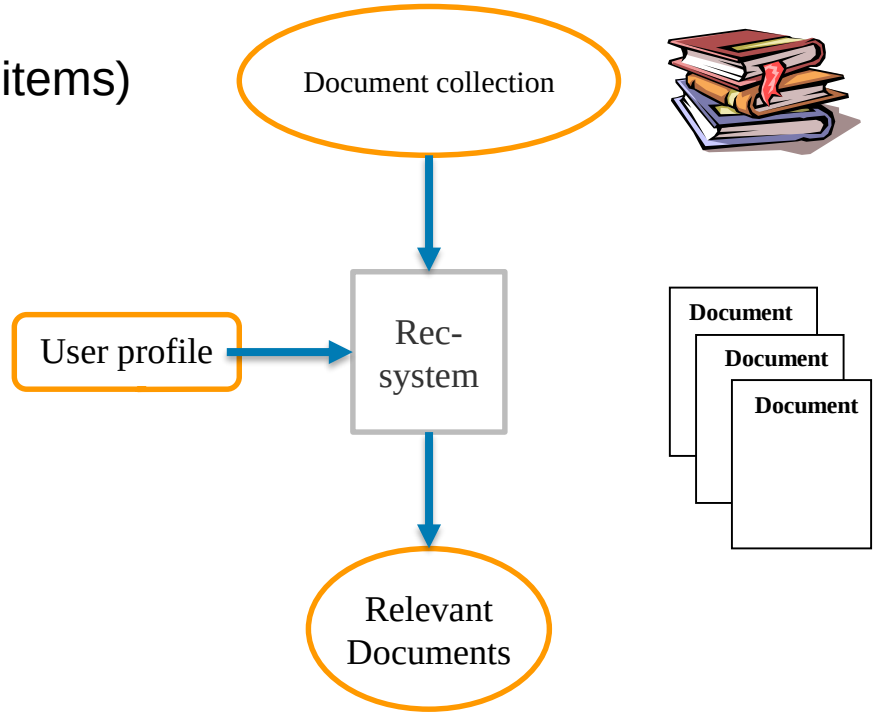
- Extractive
  - Find salient sentences (or parts)
  - Put them together to form a coherent text
  - Based on heuristics
    - Position in text
    - Entities
    - Tf*idf
    - …

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

Recommender Systeme (RS)

- Given:
  - Document colleciton (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

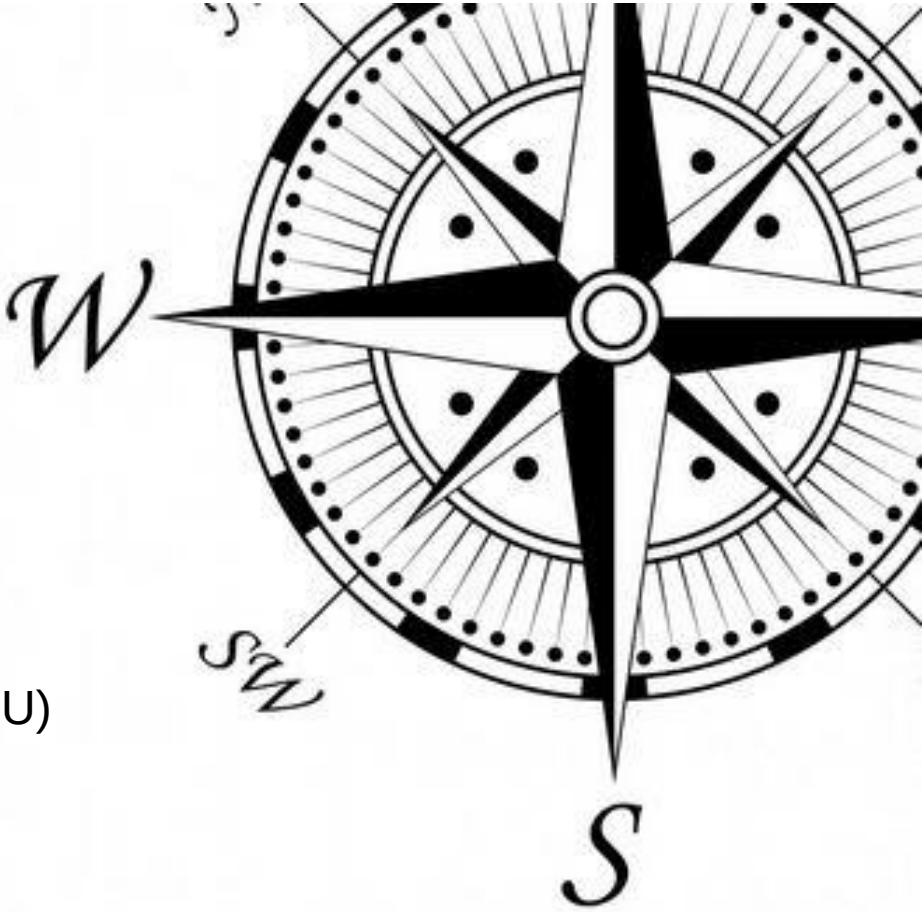
Exercise

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

Start 5 4 3 2 1 End

Topics Today

1. Document Classification
2. Document Clustering
3. Topic Modeling (TM)
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. **Knowlege Graphs (KG)**
7. Natural Language Generation (NLG)
8. Natural Language Understanding (NLU)



How to Store Knowledge?

- In a **Database**
  - Relational / Graph / NoSQL DB
- In an **Ontology**
  - Metadata and how they relate (triples)
- In a **Knowledge Base**
  - Data and how they relate (fact triples)
- In a **Knowledge Graph**
  - Connecting the fact triples

Example Database

Books

Title	Author	Publisher	Year Published	Followed By
To Kill a Mockingbird	Harper Lee	J. B. Lippincott Company	1960	Go Set a Watchman
Go Set a Watchman	Harper Lee	HarperCollins, LLC; Heinemann	2015	
The Picture of Dorian Gray	Oscar Wilde	J. B. Lippincott & Co.	1890	
2001: A Space Odyssey	Arthur C. Clarke	New American Library, Hutchinson	1968	

Publishers

Name	City	Country
J. B. Lippincott & Company	Philadelphia	United States
HarperCollins, LLC	New York City	United States
Heinemann	Portsmouth	United States
New American Library	New York City	United States
Hutchinson	London	United Kingdom

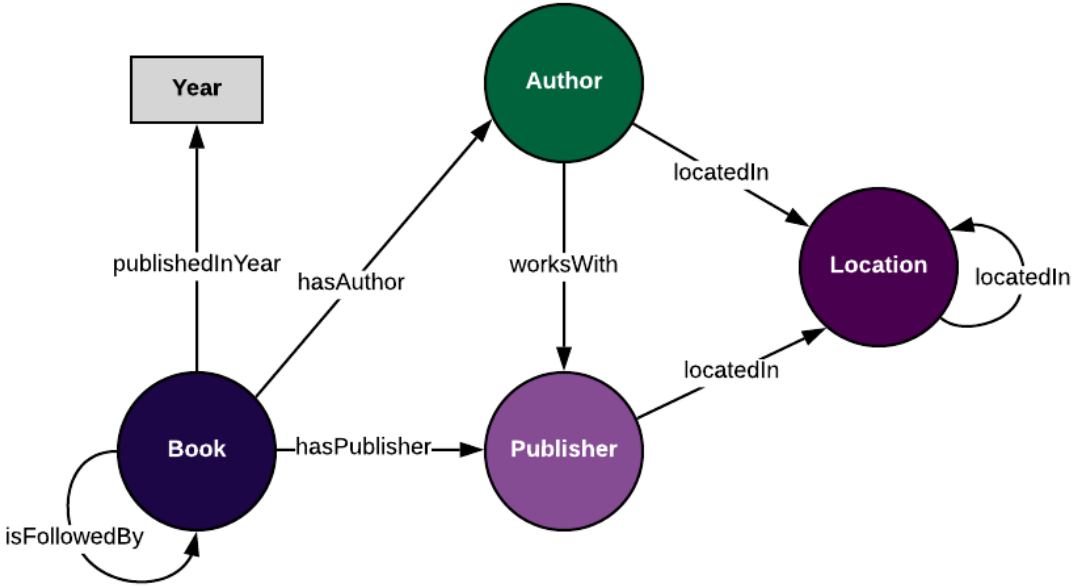
Authors

Name	Country of Birth
Harper Lee	United States
Oscar Wilde	Ireland
Arthur C. Clarke	United Kingdom

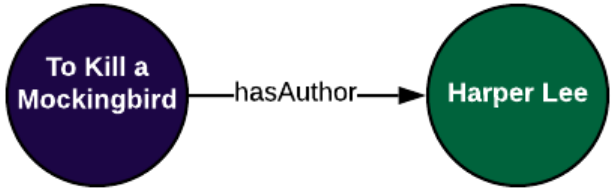
https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/

Example Ontology

- Classes
  - Books
  - Authors
  - Publishers
  - Locations
- Attributes
  - Books are published on a date
  - …
- Relations
  - Books have authors
  - Books have publishers
  - Books are followed by sequels (other books)
  - …



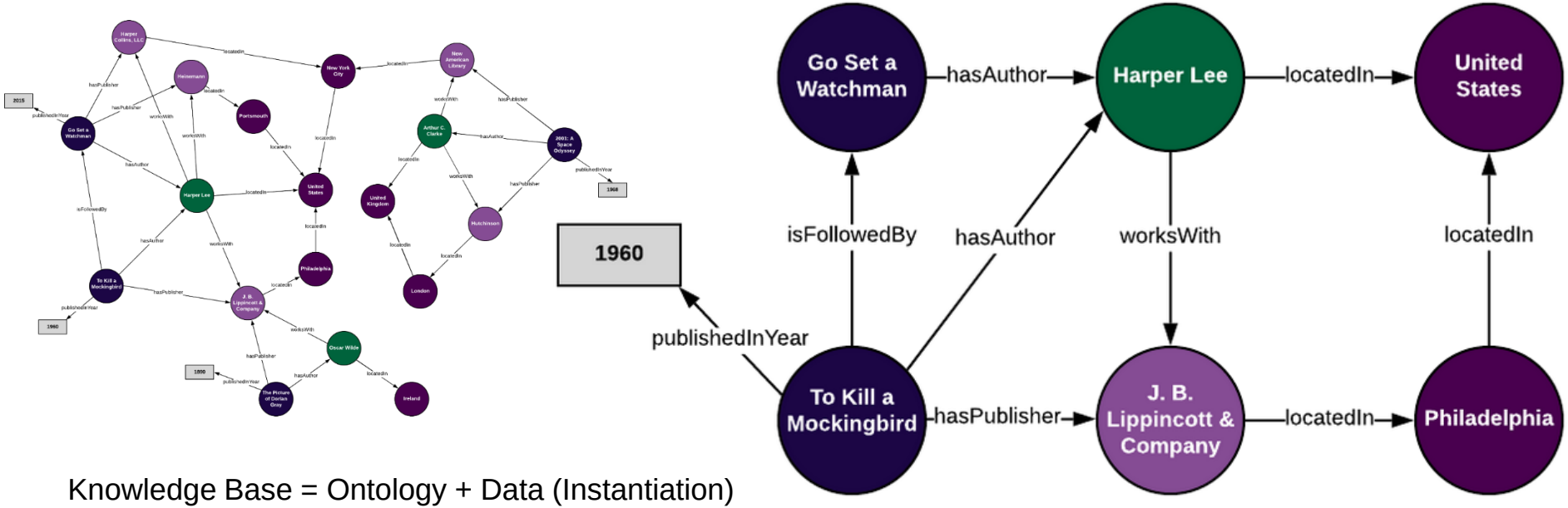
Example Knowledge Base



- Collection of facts:
  - To Kill a Mockingbird → has author → Harper Lee
  - To Kill a Mockingbird → has publisher→ JBL&C
  - To Kill a Mockingbird → published in → 1960
  - To Kill a Mockingbird → is followed by → Go Set a Watchman
  - Harper Lee → works with → JBL&C
  - JBL&C → located in → Philadelphia
  - Philadelphia → located in → United States of America
  - …

Example Knowledge Graph

- A graph containing all the triples of a knowledge base



Knowledge Base = Ontology + Data (Instantiation)

Knowledge Graph = KB + Graph

Knowledge Graph: Definition

- A Knowledge Graph is a data set that is:
  - structured (in the form of a specific data structure)
  - normalized (consisting of small units, such as vertices and edges)
  - connected (defined by the – possibly distant – connections between objects)

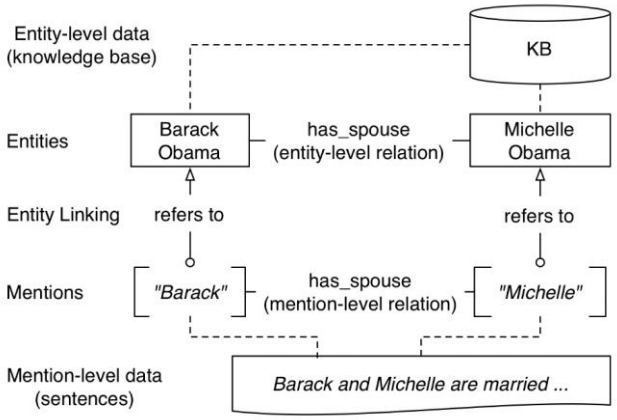
- Moreover, knowledge graphs are typically:
  - explicit (created purposefully with an intended meaning)
  - declarative (meaningful in itself, independent of a particular implementation or algorithm)
  - annotated (enriched with contextual information to record additional details and meta-data)
  - non-hierarchical (more than just a tree-structure)
  - large (millions rather than hundreds of elements)

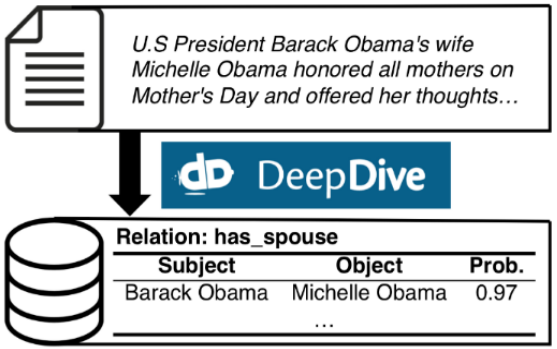
Knowledge Graph: (Counter-) Examples

- **Typical** knowledge graphs:
  - Wikidata, Yago, Freebase, DBpedia (though hardly annotated), OpenStreetMap
  - Google Knowledge Graph, Microsoft Bing Satori (presumably; we can't really know)
- **Debatable** cases:
  - Facebook's social graph: structured, normalized, connected, but not explicit (emerging from user interactions, without intended meaning beyond local relations)
  - WordNet: structured dictionary and thesaurus, but with important unstructured content (descriptions); explicit, declarative model
  - Global data from schema.org: maybe not very connected
  - Document stores (Lucene, MongoDB, …): structured, but not normalized; connections 2nd
- Primarily **not** knowledge graphs:
  - Wikipedia: mostly unstructured text; not normalized; connections (links) important but secondary (similar: the Web)
  - Relational database of company X: structured and possibly normalized, but no focus on connections (traditional RDBMS support connectivity queries only poorly)

Knowledge Graph Tasks

- Knowledge **Representation** Learning
- Knowledge Aquisition
  - Construction from scratch
  - **Completion** of existing KG
    - **Named Entity** Recognition (NER)
    - Named Entity Disambiguation (NED)
    - Named Entity Linking (NEL)
    - **Relation** Extraction (RE)
- **Reasoning** / Inference
- Today „all-in-one“ solutions, e.g.
  - Acquisition: [Han18]
  - Reasoning: [Bauer18]

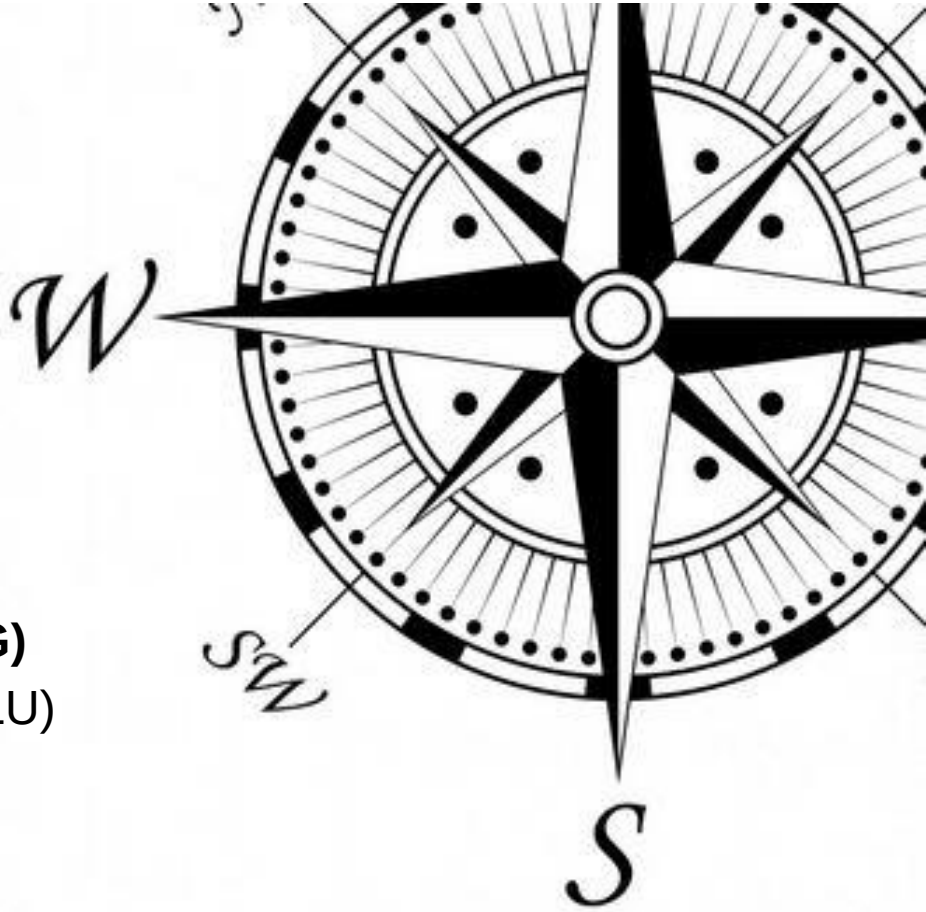




http://deepdive.stanford.edu/kbc

Topics Today

1. Document Classification
2. Document Clustering
3. Topic Modeling (TM)
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. **Natural Language Generation (NLG)**
8. Natural Language Understanding (NLU)

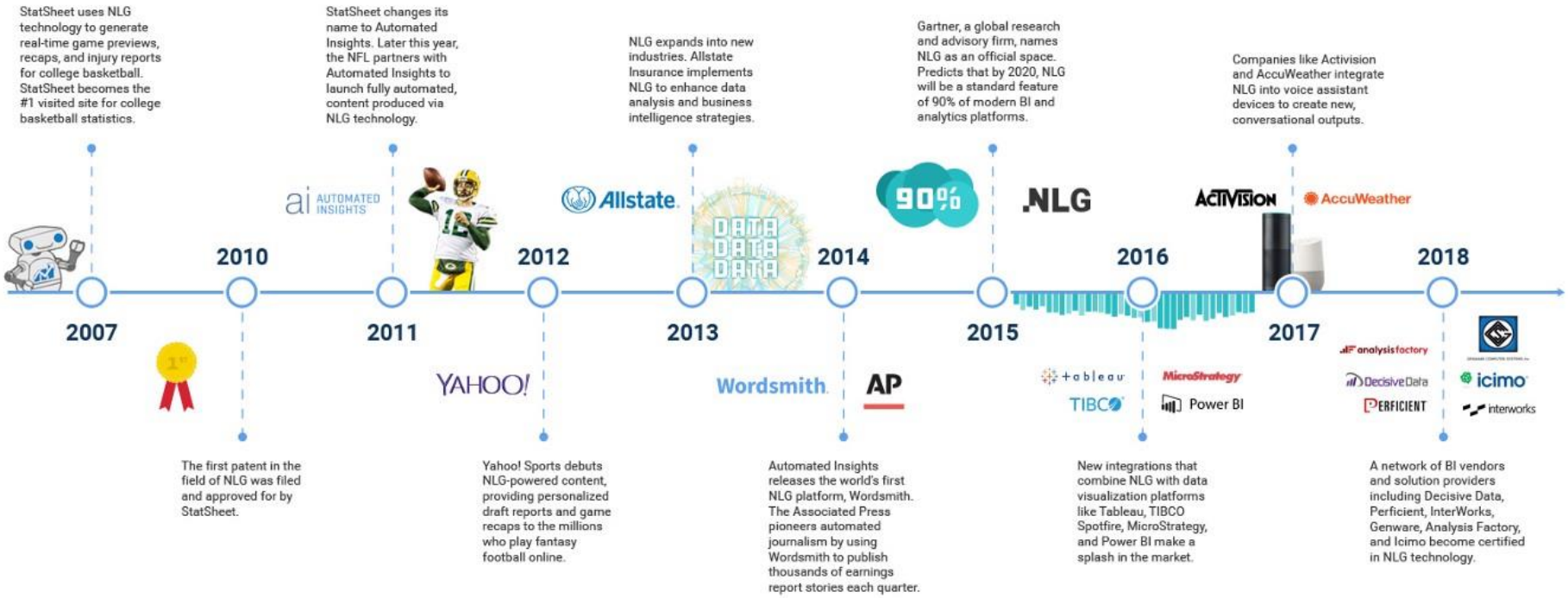


What is Natural Language Generation (NLG)

- In 2000: NLG is mapping some communication goal to some surface utterance that satisfies the goal.
  - Dale, Robert; Reiter, Ehud (2000). Building natural language generation systems. Cambridge, UK: Cambridge University Press.

- In 2017: Augmented Analytics Is the Future of Data and Analytics. **Augmented analytics**, an approach that automates insights using machine learning and natural-language generation, marks the next wave of disruption in the data and analytics market.
  - Gartner Research https://www.gartner.com/en/documents/3773164

History of Natural Language Generation (NLG)



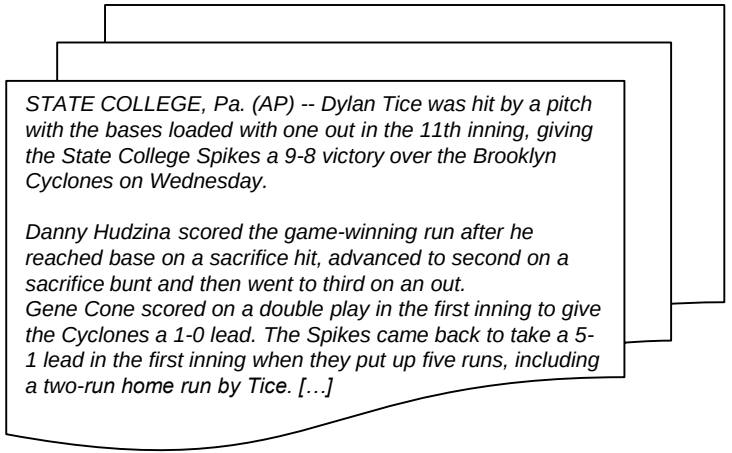
https://medium.com/@AutomatedInsights/the-history-of-natural-language-generation-5b4c3fa2f9f9

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)



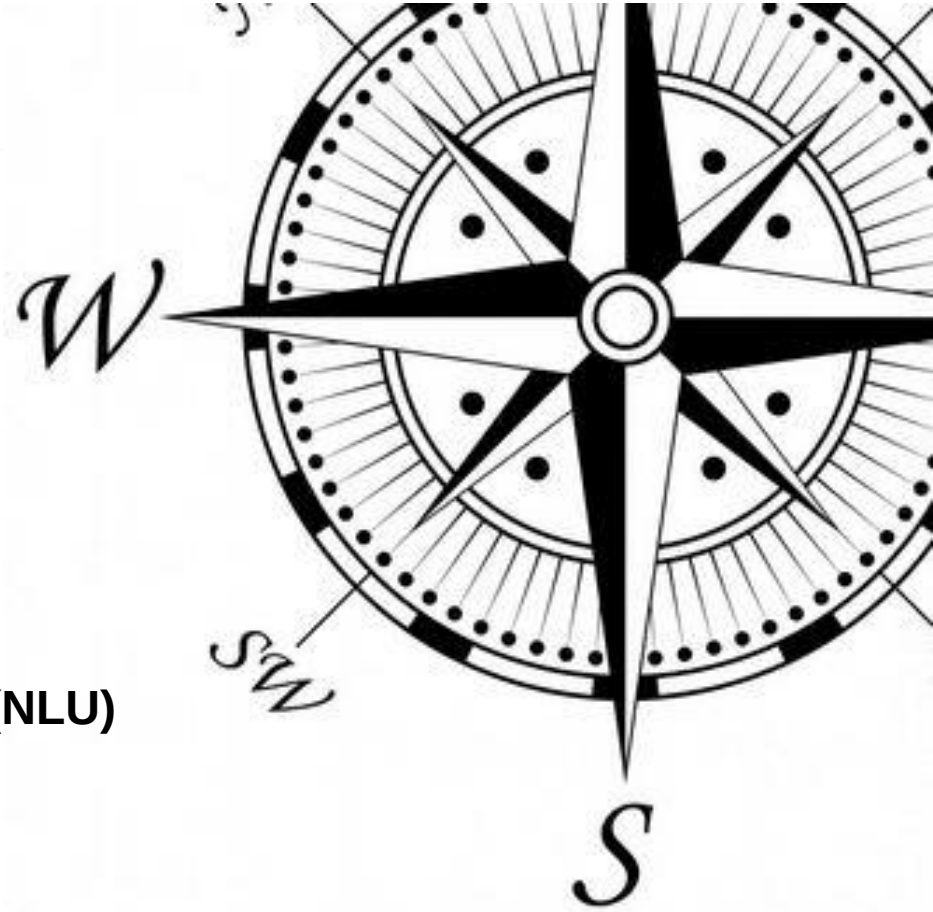
Pipeline Architecture (Pre-DL)

- Content determination
  - What information should be included in the text?
- Document structuring
  - How to organize text
- Lexicalization
  - Choosing particular words or phrases
- Aggregation
  - Composing chunks of info into sentences
- Referring expression generation
  - What properties should be used in referring to an entity
- Surface realization
  - Mapping underlying content of text to a grammatically correct sentence that expresses the desired meaning

Dale, Robert; Reiter, Ehud (2000).
Building natural language generation systems. Cambridge, UK: Cambridge University Press.

Topics Today

1. Document Classification
2. Document Clustering
3. Topic Modeling (TM)
4. Machine Translation (MT)
5. Information Retrieval (IR)
6. Knowlege Graphs (KG)
7. Natural Language Generation (NLG)
8. **Natural Language Understanding (NLU)**



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

Lerning Goals for this Chapter

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

- Be able to explain text mining
- Identify text mining tasks
- Being able to list various text mining tasks, naming
  - Difficulties/challenges
  - Traditional approaches
- Know the limitations of traditional text mining applications