Introduction to Text Mining

Part I: Overview

Henning Wachsmuth

https://cs.upb.de/css

Outline of the course

I. Overview
   - What Is Text Mining?
   - Challenges
   - Techniques and Approaches

II. Basics of Linguistics

III. Text Mining using Rules

IV. Basics of Empirical Research

V. Text Mining using Grammars

VI. Basics of Machine Learning

VII. Text Mining using Clustering

VIII. Text Mining using Classification and Regression

IX. Practical Issues

X. Text Mining using Sequence Labeling

What Is Text Mining?

Motivation

Text Mining in a Nutshell

- Automatic discovery of previously unknown information of high quality in large amounts of mostly unstructured natural language text.

Major text mining steps

1. Information retrieval (IR). Gather potentially relevant input texts.
2. Natural language processing (NLP). Analyze the texts to identify and structure relevant information.
3. Data mining (DM). Discover patterns in the structured information.

Observations

- The types of information to be discovered are specified beforehand, i.e., text mining tackles given tasks.
- NLP is the primary focus in general (and in this course).

  A simplified view of text mining is to see it just as NLP on text.

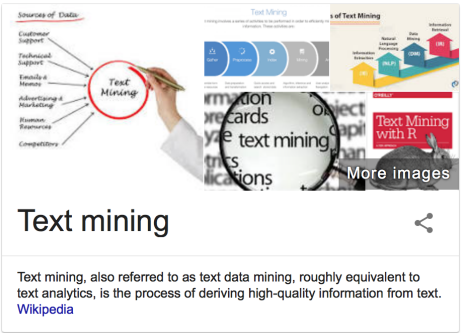
Motivation

Why Text Mining?

Applications

- Web search engines shift to directly returning relevant information from web pages in response to queries.
- Intelligent personal assistants analyze textual representations of questions to answer them.
- Big data analytics finds relations, patterns, and facts in vast amounts of often textual data.

... and many others



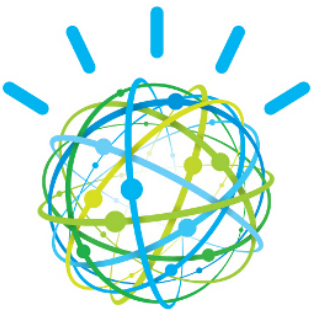
http://www.google.com/#q=text+mining

Characteristics

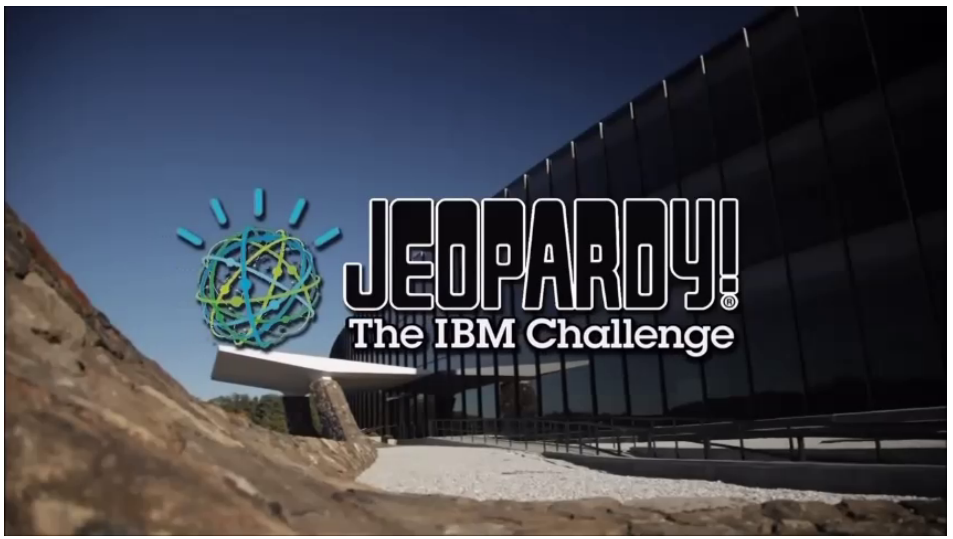
- Input data may be huge
- Output information is “hidden”
- Mining must be automatic, fast, and accurate

Text Mining Applications: Watson

IBM Watson



- State-of-the-art question answering system
- Nowadays in many decision support applications
- Initially developed to tackle the “Jeopardy!” task

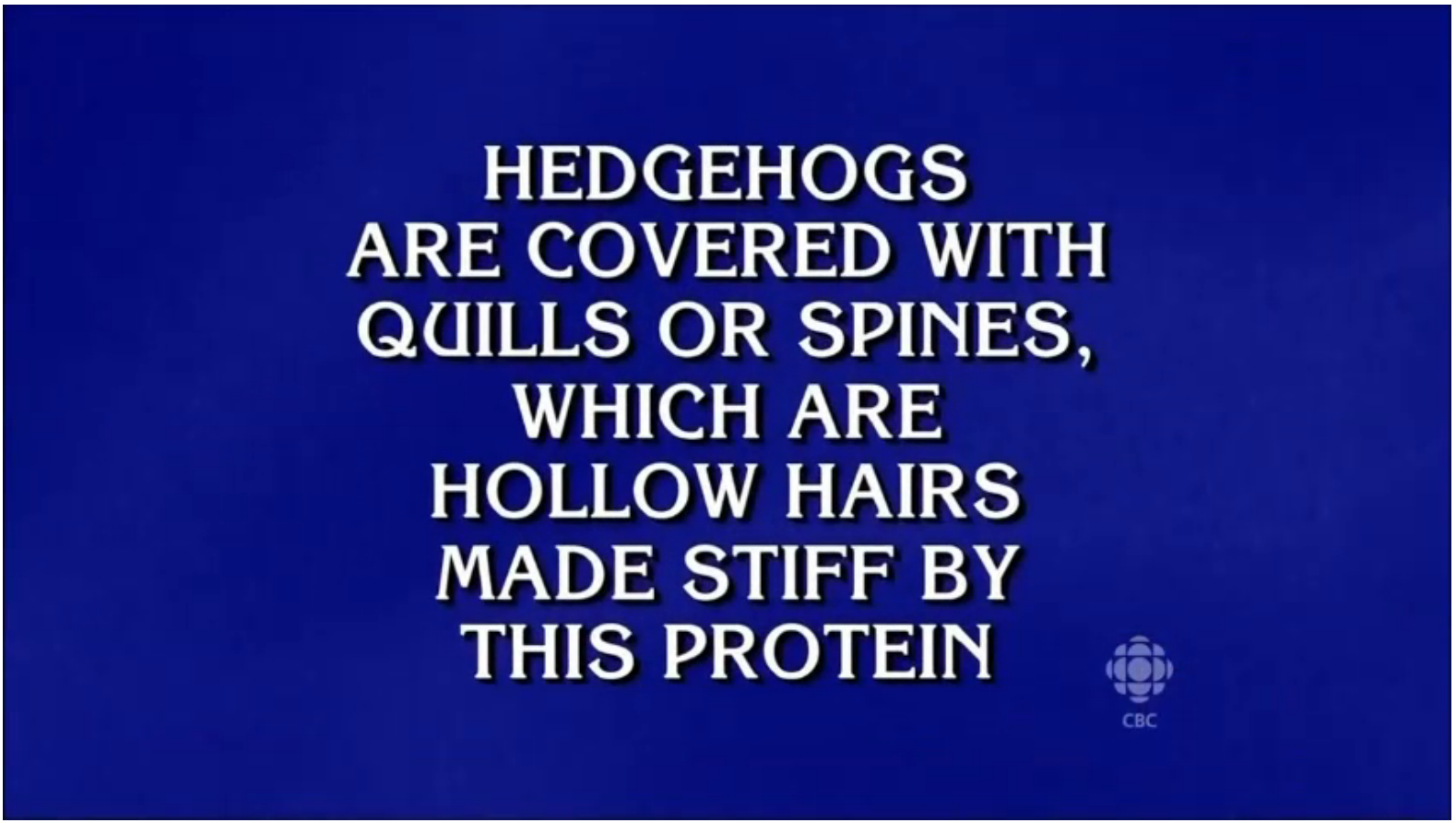


The IBM Challenge in 2011

- Watson plays against the best Jeopardy! champions
  `https://www.youtube.com/watch?v=P18EdAKuC1U`

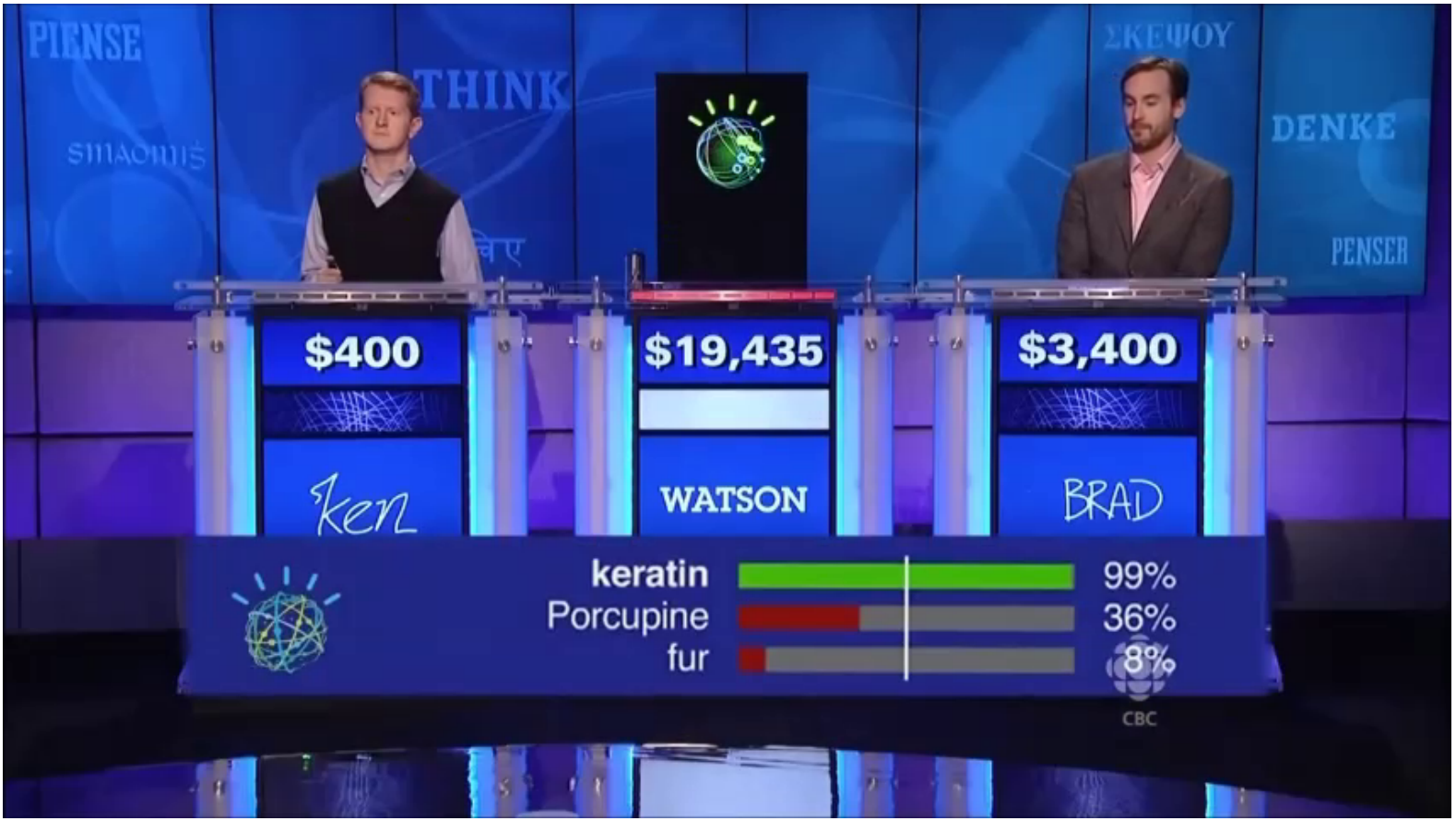
Text Mining Applications: Watson

Example “Question”



Text Mining Applications: Watson

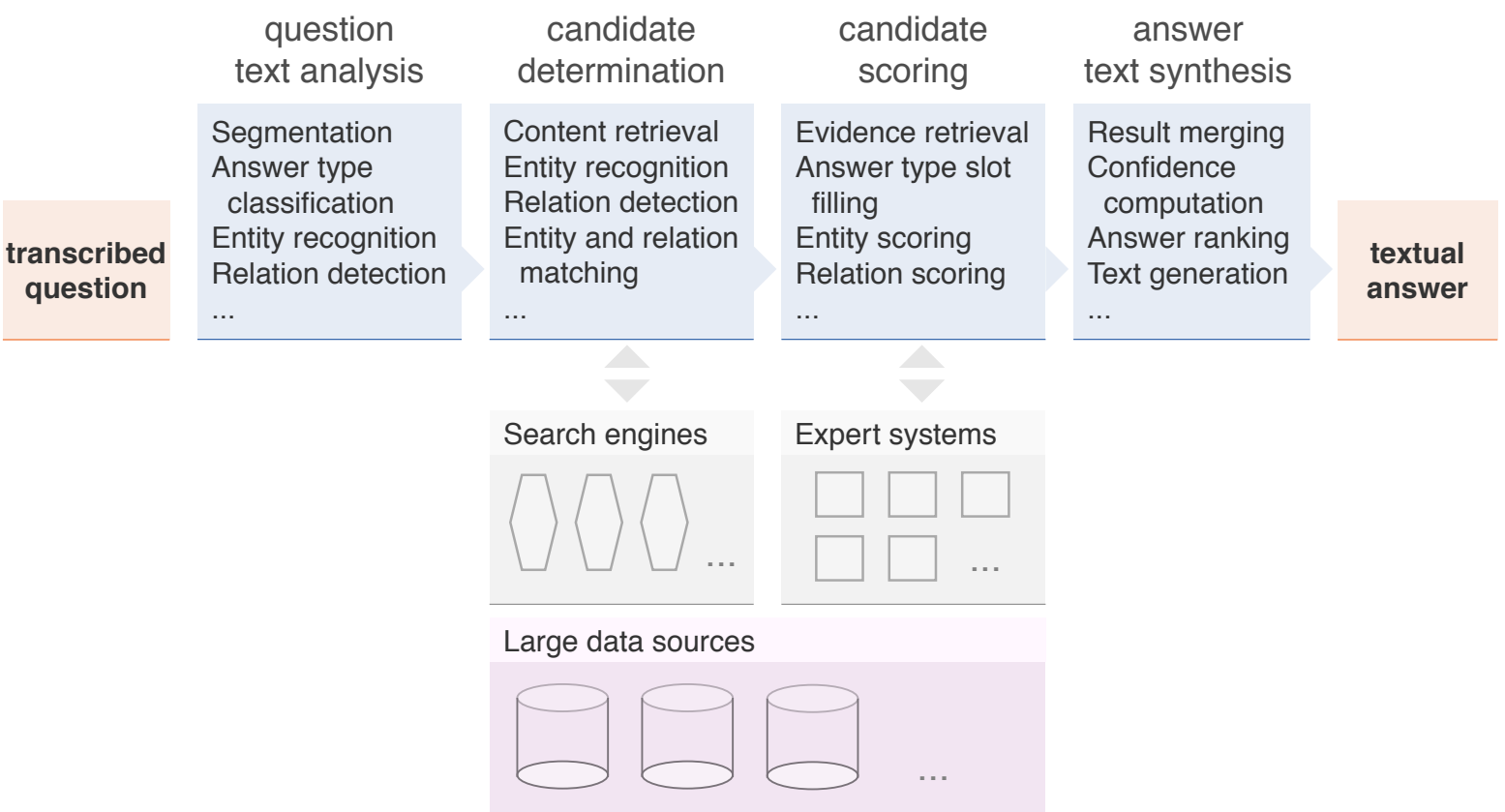
Watson's "Answer"



Text Mining Applications: Watson

Where Is Text Mining Involved?

Watson's question answering process (simplified)



Evolution of Natural Language Processing Applications

Selected milestones

- February 2011. IBM's Watson wins Jeopardy
  `https://www.youtube.com/watch?v=P18EdAKuC1U`

- October 2011. Siri starts on the iPhone
  `https://www.youtube.com/watch?v=gUdVie_bRQo`

- August 2014. Skype translates conversations in real time
  `https://www.youtube.com/watch?v=RuAp92wW9bg`

- May 2018. Google does phone call appointments
  `https://www.youtube.com/watch?v=pKVppdt_-B4`

- June 2018. IBM Debater competes in classical debates
  `https://www.youtube.com/watch?v=UeF_N1r91RQ`

Observations

- Text mining inside, i.e., all main processing tasks are done on text.
- None of these applications works perfectly.

Text Mining Tasks

Information Extraction

Example: Extract company's founding dates from news articles

Output: Founded("Google", 1998)

Typical text analysis steps

1. Lexical and syntactic preprocessing
2. Named and numerical entity recognition
3. Reference resolution
4. Entity relation extraction

Text Mining Tasks

Text Classification

Example: Classify topic and sentiment of customer reviews

> *“ This was truly a lovely hotel to stay in . The staff were all friendly and very helpful . The location was excellent . The atmosphere is great and the decor is beautiful. “*

Output: Topic("hotel"), Sentiment("positive")

Typical text analysis steps

1. Lexical and syntactic preprocessing
2. Feature computation
3. Supervised classification

Notice

- Applications often combine information extraction and text classification.

Analysis Levels

Levels of language analysis

- Phonetics. The physical aspects of speech sounds.
- Phonology. The linguistic sounds of a particular language.
- Morphology. The senseful components of words (and other *tokens*).
- Syntax. The structural relationships between words, usually within a sentence (or a similar utterance).
- Semantics. The meaning of single words and compositions of words.
- Discourse. Linguistic units larger than a single sentence, such as paragraphs and complete documents.
- Pragmatics. How language is used to accomplish goals.

Notice

- Phonetics and phonology are usually disregarded in text mining.
- Some language phenomena cannot clearly be assigned to one level, such as sentiment.

Text Analyses

Lexical and syntactic

- Tokenization
- Sentence splitting
- Paragraph detection
- Stemming
- Lemmatization
- Part-of-speech tagging
- Similarity computation
- Spelling correction
- Phrase chunking
- Dependency parsing
- Constituency parsing

... and some more

Semantic and pragmatic

- Term extraction
- Numerical entity recognition
- Named entity recognition
- Reference resolution
- Entity relation extraction
- Temporal relation extraction
- Topic detection
- Authorship attribution
- Sentiment analysis
- Discourse parsing
- Spam detection
- Argument mining

... and many many more

Notice

- The tasks in green are going to be detailed in this course.

Terminology

Term usage in this course

- Text mining application. A technology that tackles a real-world problem using text mining.
- Text mining task. A specific problem of inferring meta-information about a text. Requires one or more text analyses.
- Text mining approach. A specific computational method to tackle some text mining task.
- Text analysis. A processing of text that infers specific meta-information.
- Text analysis algorithm. A specific computational method to address some text analysis.
- Text analysis technique. A general method of how to analyze a text.

Notice

- Meta-information is usually represented as annotations of a text.
- Term usage has slight variations and inconsistencies in the literature.

Challenges

Ambiguity in Natural Language

Ambiguity

- Fundamental challenge of processing natural language
- Pervasive

Several types of ambiguity

- Phonetic. “wreck a nice beach”
- Word sense. “I went to the bank”.
- Part of speech. “I watch my watch.”
- Attachment: “I saw a kid with a telescope.”
- Scope of quantifiers. “I didn’t buy a car.”
- Presuppositions. “I believe Max’ car is red, even though Max never bought nor received a car.”
- Speech act. “Have you emptied the dishwasher?”

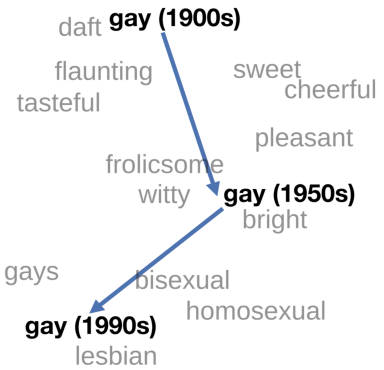
... and many more

Ambiguity in Natural Language

Word Senses

Word sense ∼ Meaning of a word

- "daffodil" has only one sense
- "make" has 49 senses
- "gay" changed its (main) sense over time



(Hamilton et al., ACL 2016)

Example: "ride" has 16 senses, here is a selection:

- ride over, along, or through
- sit and travel on the back of animal, usually while controlling its motions
- be carried or travel on or in a vehicle
- be contingent on
- harass with persistent criticism or carping
- keep partially engaged by slightly depressing a pedal with the foot
- continue undisturbed and without interference
- move like a floating object

Ambiguity in Natural Language

Ambiguity Types Interfere

Interpretations of “I made her duck”

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the duck she owns.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into undifferentiated waterfowl.

Reasons for ambiguity

- “duck” and “her” are ambiguous in their part of speech.
- “make” has different meanings.
- “make” can take one object or two.

Ambiguity in Natural Language

Pragmatic Ambiguity May even Be Worse

Purpose of "I never said she stole my money."

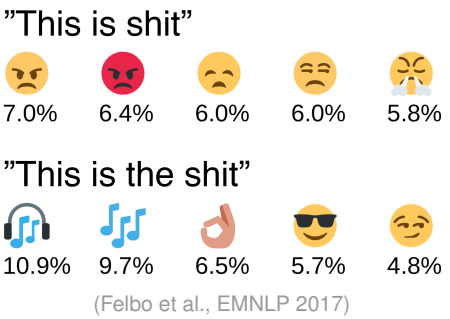
I never said she stole my money.	Someone else said it, but I didn't.
I *never* said she stole my money.	I simply didn't ever say it.
I never *said* she stole my money.	I might have implied it in some way. But I never explicitly said it.
I never said *she* stole my money.	I said someone took it. But I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

Other Challenges of Processing Natural Language

Selected Challenges in a Text

Tricky language

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



(Felbo et al., EMNLP 2017)

Tricky phrases

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

Tricky segmentation

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

Other Challenges of Processing Natural Language

Selected Challenges of the Context of a Text

World knowledge

- "Max and Tim are brothers." vs. "Max and Tim are fathers."
  Do Max and Tim belong to the same family?
- "I hope Trump will rethink capital punishment."
  Stance on death penalty? What location is it about? Is death penalty legal there?

Domain dependency

- "Read the book!"
  Positive sentiment in a book review. Positive in a movie review?

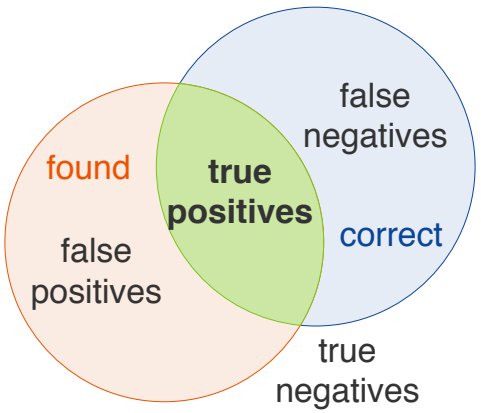
Language dependency

- "Bad"
  Sentiment in English? In German (means "bath")? In French (does not exist)?
  In Japanese (not even the characters exist)? ...

Performance Challenges of Text Mining

Effectiveness

- Text mining is rarely free of errors, due to the raised challenges.
- Effectiveness, in terms of the extent to which the found information is correct, is hence the primary goal of any text mining approach.



Efficiency

- Text mining often needs to operate on large amounts of text in few time.
- Efficiency, in terms of run-time or sometimes also space, is hence important, particularly in practical applications.

Robustness

- Text mining often needs to be applied to texts with unknown properties.
- Robustness, in term of being effective across topics, genres, or other domains of texts, is hence important, again particularly in practice.

Techniques and Approaches

Text Mining as Corpus Linguistics

Text corpus

- Text mining is usually studied in a corpus linguistics manner, i.e., approaches are developed on text corpora.
- Text corpus. A collection of real-world texts with known properties, compiled to study a language problem.

Annotation

- Annotation. Marks a text or a span of text that represents an instance of a particular type of information. Annotations represent meta-information about the marked parts.
- The texts in a corpus are often annotated for the studied problem.

Datasets

- Dataset. Sub-corpus used for developing and evaluating approaches.
- Typical: a *training set* for development, a *validation set* for evaluation during development, and a *test set* for the final evaluation.

Text Analysis Techniques

Task Dimensions

Task goal

- Clustering. A set of instances is grouped into not-predefined classes.
- Classification. Each instance is assigned a predefined class label.
- Regression. Each instance is assigned a numeric value.

... and some others

Task knowledge

- Supervised. Correct labels/values of training instances available.
- Unsupervised. No class labels/values used in development.

... and some others

Task interdependency

- Independent. Each instance is treated in isolation.
- Sequential. Instances are treated based on others in a sequence.
- Hierarchical. Instances are recursively decomposed into sub-instances.

Text Analysis Techniques

Inference Process

Knowledge-based inference

- Analysis done based on manually encoded expert knowledge.
- Knowledge represented by rules, lexicons, grammars, etc.

Feature-based statistical inference

- Analysis done based on statistical patterns found in training data.
- Patterns capture manually or semi-automatically encoded text features.

Neural statistical inference

- Analysis done based on statistical patterns found in training data.
- Patterns automatically encoded in neural networks.

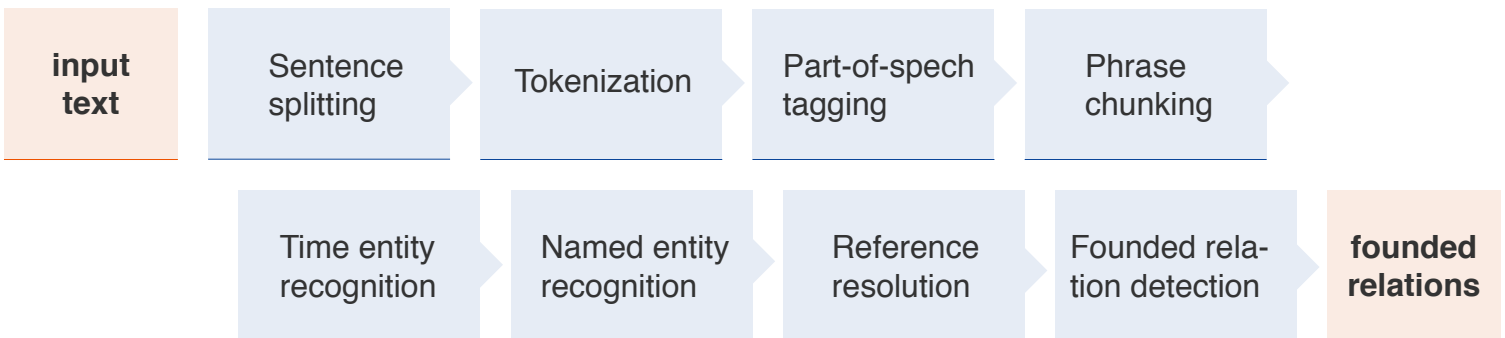
Text Analysis Pipelines

Inference Process

Pipeline approach

- The standard way to tackle a text mining task is with a pipeline.
- Text analysis pipelines. Sequentially apply a set of text analysis algorithms to the input texts.

Example extraction pipeline for company's founding dates



Alternatives

- Joint model. Realizes multiple analysis steps at the same time.
- Neural network. Often just works on the plain input text.

Development of a Text Mining Approach

Input

- Task. A text mining task to be approached.
- Text corpus. A corpus, split into development and evaluation datasets.

A typical development process

1. Analyze on training set how to best tackle the task.
2. Develop approach based on some technique that tackles the task.
3. Evaluate the performance of the approach on the validation set.
4. Repeat steps 1–3 until performance cannot be improved anymore.
5. Evaluate the performance of the final approach on the test set.

Output

- Approach. A text mining approach to tackle the given task.
- Results. Empirical performance measurements of the approach.

Conclusion

Summary

Text mining

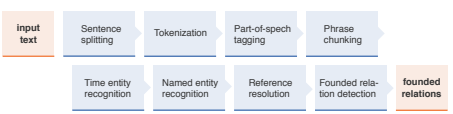
- Discovers new information in natural language text.
- Uses computational analysis at several linguistic levels.
- Has important applications in real-word technologies.

Challenges

- Natural language is ambiguous in several ways.
- Understanding requires context and world knowledge.
- Text mining aims to be effective, efficient, and robust.

This course

- Teaches how to develop text mining approaches.
- Covers several tasks and analysis techniques.
- Covers design, implementation, and evaluation.



References

Some content and examples taken from

- Emily M. Bender (2018). 100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask. Tutorial at the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), http://faculty.washington.edu/ebender/papers/Bender-ACL2018-tutorial.pdf.
- Daniel Jurafsky and Christopher D. Manning (2016). Natural Language Processing. Lecture slides from the Stanford Coursera course. https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html.
- Matthias Hagen (2018). Natural Language Processing. Slides from the lecture at Martin-Luther-Universität Halle-Wittenberg. https://studip.uni-halle.de/dispatch.php/course/details/index/8b17eba74d69784964cdefc154bb8b95.
- Daniel Jurafsky and James H. Martin (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall, 2nd edition.
- Christopher D. Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Henning Wachsmuth (2015): Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer.