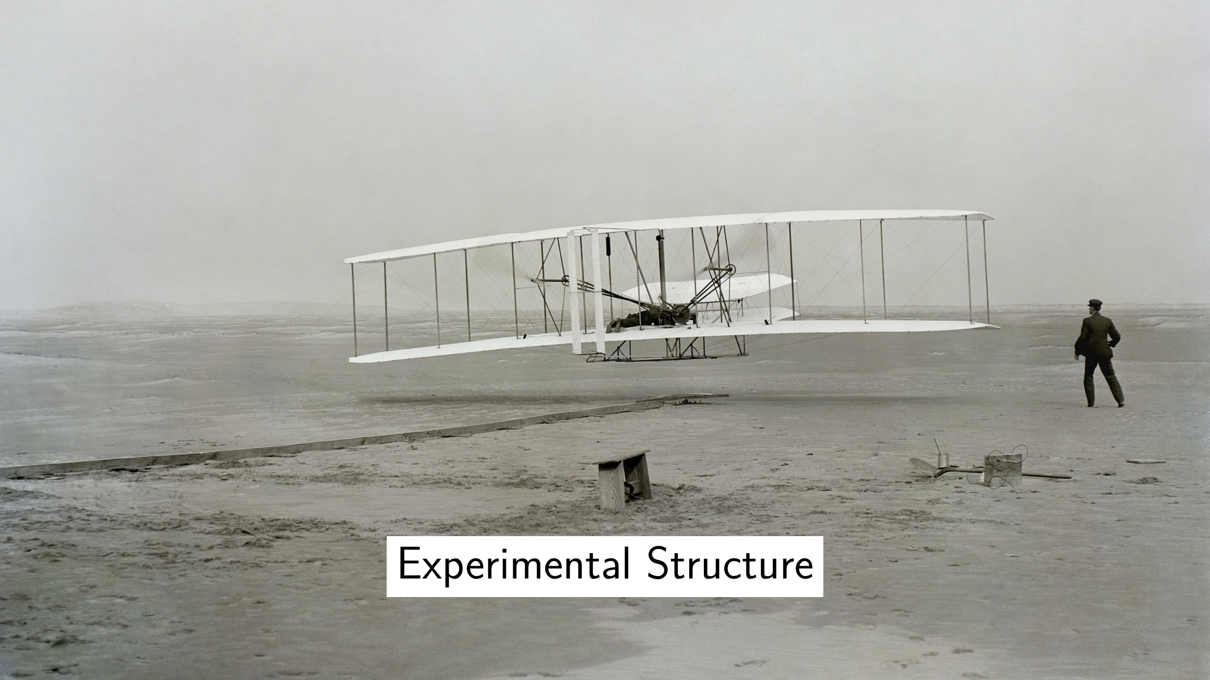
Natural Language Processing 1: Big Picture

HS Sprachtechnologie für eine bessere Welt (Winter term 2022/23)

Nils Reiter,
nils.reiter@uni-koeln.de

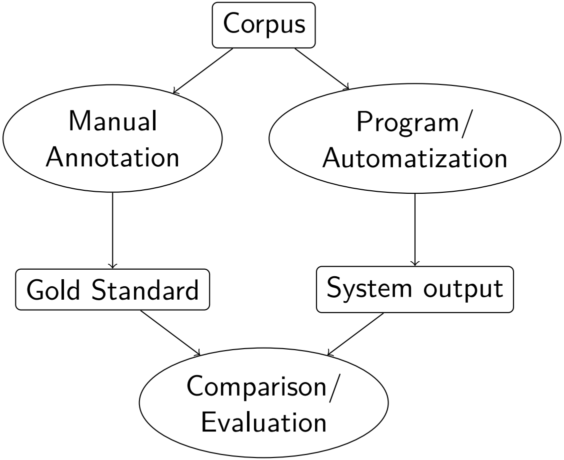
October 25, 2022

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

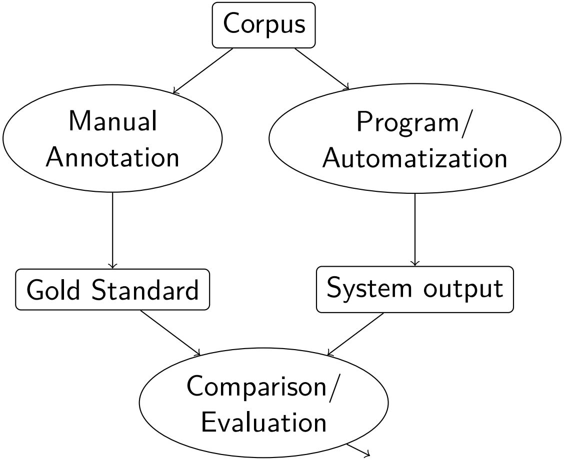


Experimental Structure

Experiments

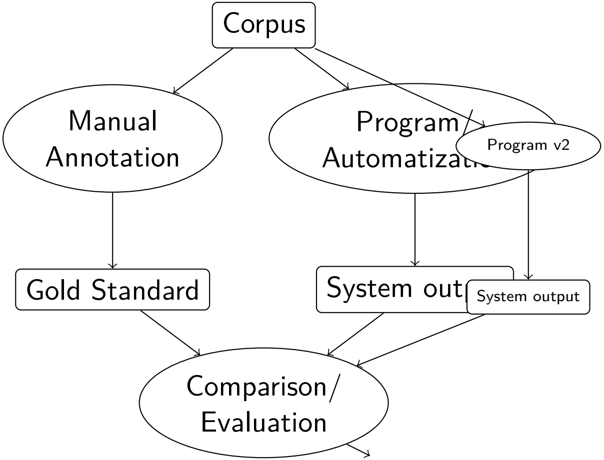


Experiments



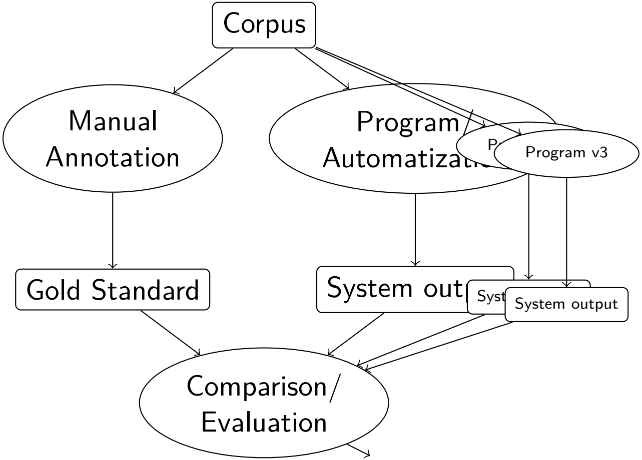
	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...
v3	15.3	21.8	...

Experiments



	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...
v3	15.3	21.8	...

Experiments



	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...
v3	15.3	21.8	...

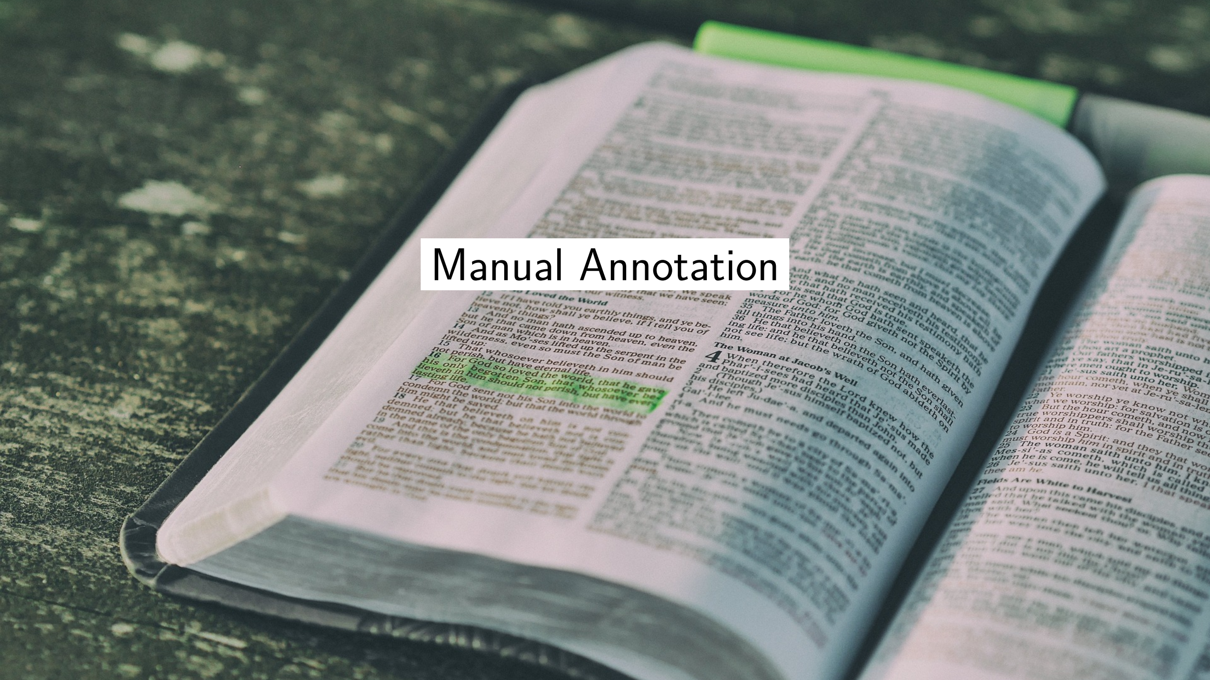
Experiments

- Reproducibility
- Hypotheses about the operationalisation of language/text phenomena

Example

- Position within a sentence is indicative for the part of speech
- Meaning of a word depends on its context
- The protagonist of a play is the character who talks the most

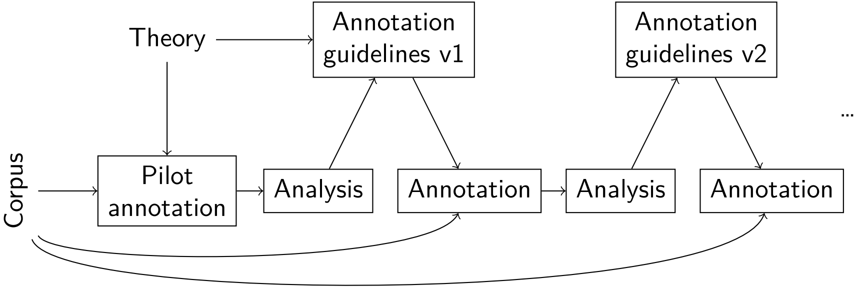
Manual Annotation



Annotation

- Interdisciplinary ‘false friend’
- Different meanings in different disciplines
  - Adding TEI/XML markup: DH community
  - Adding comments to page margins: Hermeneutic traditions
    - Literary studies, bible studies
  - Assigning categories to textual material: (computational) linguistics

Annotation Workflow



Hovy/Lavid (2010); Pagel et al. (2018)

Annotation guidelines

- Describe the way to create the machine-readable truth
- What is to be annotated (which words)
- Working definitions or tests for categories
- Living documents: Need to be iteratively improved
- Community-wide accepted standards are needed

Annotation Analysis

- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified ('Inter-Annotator-Agreement', IAA)

  Cohen, 1960; Fleiss, 1971; Fournier, 2013; Mathet et al., 2015
- Inter- und Intra-AA
- … it's also a good idea to talk to the annotators

Indirect Annotations

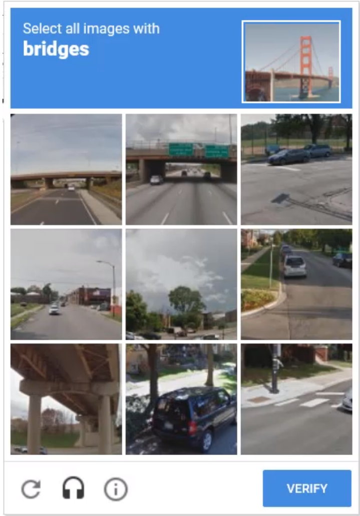
- Annotations as a by-product of games
  - https://www.artigo.org — Kohle (2010)
  - https://anawiki.essex.ac.uk/phrasedetectives/ — Chamberlain et al. (2008)

Indirect Annotations

- Annotations as a by-product of games
  - https://www.artigo.org — Kohle (2010)
  - https://anawiki.essex.ac.uk/phrasedetectives/ — Chamberlain et al. (2008)
- Captchas for OCR correction

Indirect Annotations

- Annotations as a by-product of games
  - https://www.artigo.org 2010)
  - https://anawiki.essex.ac.uk/phrasedetectives 08)
- Captchas for OCR correction

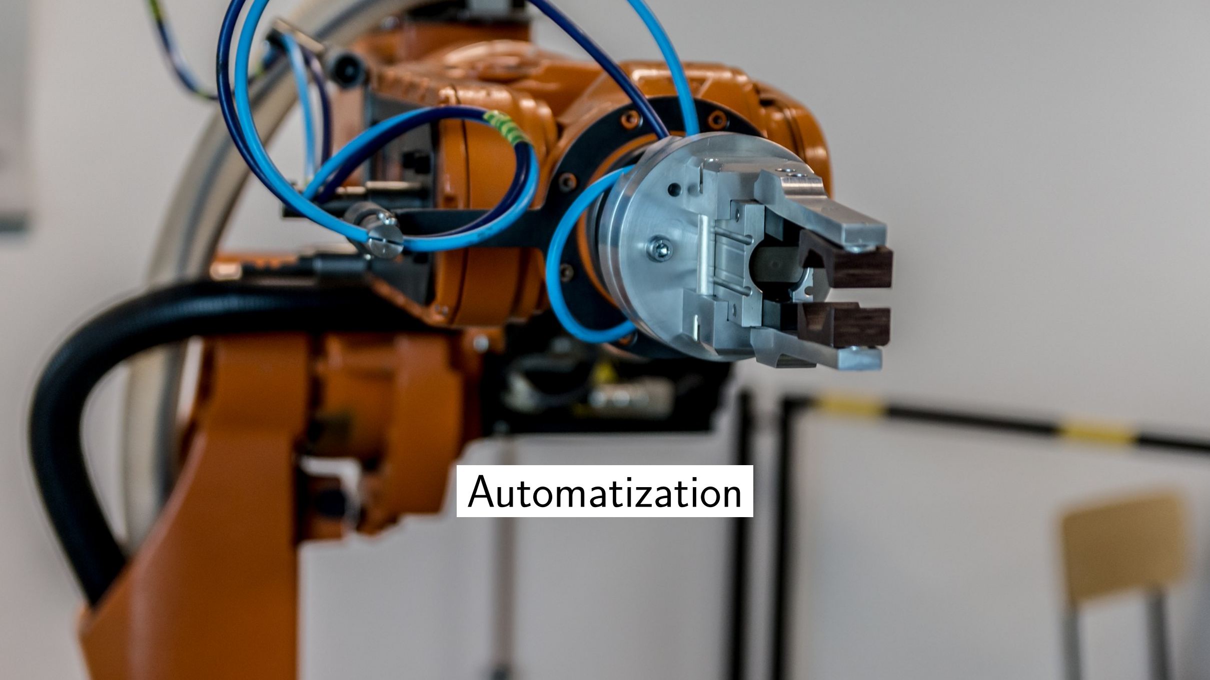


Learning from Raw Data

- Train on things that are already there
- word2vec: Is 'dog' a context word of 'lazy'? Mikolov et al. (2013)
- BERT Devlin et al. (2019)
  - Can you fill in this blanked word? ("masked language modeling", MLM)
  - Are these two sentences natural neighbours? ("next sentence prediction", NSP)

Learning from Raw Data

- Train on things that are already there
- word2vec: Is 'dog' a context word of 'lazy'? Mikolov et al. (2013)
- BERT Devlin et al. (2019)
  - Can you fill in this blanked word? ("masked language modeling", MLM)
  - Are these two sentences natural neighbours? ("next sentence prediction", NSP)
- Training data available in abundance
  - As long as there is digital data for a language
  - ⚠ Difficult to control what exactly is in there
    - More obvious for text-image data sets Birhane et al. (2021) haveibeentrained.com


Automatization

Welche Methoden kennen Sie?

Systems

- Predict annotations
- Ideally: The same annotations as a human (the correct ones)
- Parameters
  - On what exactly does the program make predictions?
  - What information, criteria and features does it need?

Systems

- Predict annotations
- Ideally: The same annotations as a human (the correct ones)
- Parameters
  - On what exactly does the program make predictions?
  - What information, criteria and features does it need?

System types

- Rule-based (not so popular anymore)
- Supervised machine learning
  - Deep learning

Supervised Systems

- Classification: Assign items into previously known categories
  - Sequence labeling: Special case. Class for item n depends on item $n-1$
- Learn patterns from annotated data
- Relations between input (X) and output (Y)
  - Can be an n-to-m relation, but mostly n-to-1 (i.e., we predict a single target category)

Features

- The properties of a item that is to be classified
- Classical machine learning
  - Manual coding of explicit, scientifically validated features: Feature extraction
  - "Translation" of the corpus into feature vectors
  - Feature engineering
  - Design and implementation of feature extractors
  - Linguistic features need to be determined somehow
    $\rightarrow$ Dependencies, modularization

Features

- The properties of a item that is to be classified
- Classical machine learning
  - Manual coding of explicit, scientifically validated features: Feature extraction
  - "Translation" of the corpus into feature vectors
  - Feature engineering
  - Design and implementation of feature extractors
  - Linguistic features need to be determined somehow
    $\rightarrow$ Dependencies, modularization
- Deep learning
  - Embeddings used as features
    - A word is mapped onto an n-dimensional vector, which is then put into the ML system
    - Vector dimensions = features
    - But not interpretable anymore

Parameters and Hyper Parameters

Parameters

- What is learned by the algorithm during training
  - E.g., probability/frequency of feature F and class C ($=$ weights)
- Parameters are stored in the model

Parameters and Hyper Parameters

Parameters

- What is learned by the algorithm during training
  - E.g., probability/frequency of feature F and class C ($=$ weights)
- Parameters are stored in the model

Hyper Parameters

- Set during the training process by us
  - E.g., number of training epochs in a neural network, data set size, ...
- Not automatically optimised, but important for performance

Parameters and Hyper Parameters

Parameters

- What is learned by the algorithm during training
  - E.g., probability/frequency of feature F and class C ($=$ weights)
- Parameters are stored in the model

Hyper Parameters

- Set during the training process by us
  - E.g., number of training epochs in a neural network, data set size, ...
- Not automatically optimised, but important for performance
- Development set: Find optimal hyper parameters

Example: Parts of Speech

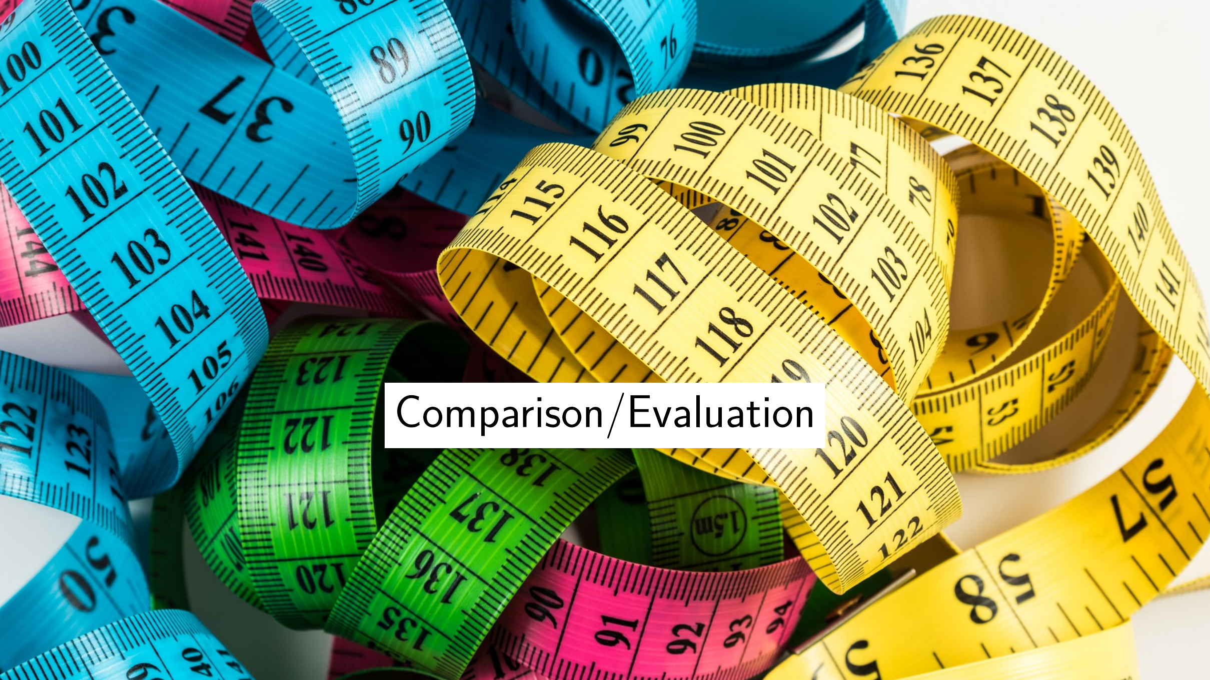
Feature	Data type
Case	Binary
Length	> 0
Sentence initial	Binary

Table: Features

Token	Case	L.	S. initial
Der	u	3	Y
Hund	u	4	N
bellt	l	5	N
.	?	1	N
Die	u	3	Y
Katze	u	5	N
schnurrt	l	8	N
.	?	1	N

Table: Feature extraction

Comparison / Evaluation



Evaluation

Intrinsic

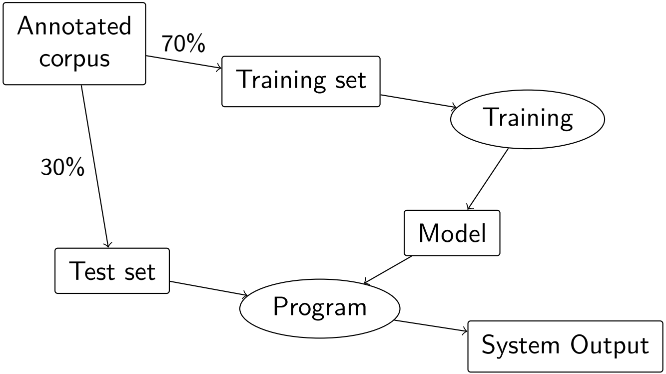
- Compare the automatically produced annotations with the gold standard
- Can be quantified (similar to IAA)
  - *precision*, *recall*, *f-score*
- System treated as a black box

Extrinsic

- Use of the program in another program that can be evaluated
  - *downstream tasks*
  - e.g., use of a PoS tagger in a machine translation system

Intrinsic Evaluation

- Goal: Predict the quality on new data
- The program cannot have seen the data, so that it's a realistic test



Classification Evaluation Metrics

(MS99, 267 ff.)

- Accuracy: How many items were correctly classified over all classes?
  (one value for everything)
- Precision: How many of the items classified as category C actually belong to category C?
  (one value per category)
- Recall: How many of the items in category C have been classified as C
  (one value per category)
- F-Score: Harmonic mean between precision and recall

Baseline

- What does an evaluation score tell us?
  - Nothing, if not compared to anything

Baseline

- What does an evaluation score tell us?
  - Nothing, if not compared to anything
- Artificial baselines
  - Majority baseline: Classify everything into the most frequent class
  - Random baseline: Classify everything at random

Baseline

- What does an evaluation score tell us?
  - Nothing, if not compared to anything
- Artificial baselines
  - Majority baseline: Classify everything into the most frequent class
  - Random baseline: Classify everything at random
- Self baselines
  - Take a single feature (classical machine learning)
  - Pre-trained embeddings
  - BERT without fine-tuning

Baseline

- What does an evaluation score tell us?
  - Nothing, if not compared to anything
- Artificial baselines
  - Majority baseline: Classify everything into the most frequent class
  - Random baseline: Classify everything at random
- Self baselines
  - Take a single feature (classical machine learning)
  - Pre-trained embeddings
  - BERT without fine-tuning
- Foreign baselines
  - Last year's system
  - Competition system
  - Shared task winner

If baseline has hyper parameters, they need to be optimized as well (for a fair comparison)

Results

	P	R	F
Baseline 1	...	...	...
Baseline 2	...	...	...
Variant 1	...	...	...
Variant 2	...	...	...
Variant 3	...	...	...

Table: A typical results table

Error Analysis

- Systems do not deliver perfect results (i.e., scores are below $100\,\%$)
- What can we say about the remaining errors?

Error Analysis

- Systems do not deliver perfect results (i.e., scores are below $100\,\%$)
- What can we say about the remaining errors?
- Workflow
  - Extract n errors, inspect them manually
  - Can we detect regularities/patterns in them? E.g., why they were misclassified?
  - Ideally, error analysis makes quantitative statements about error sources

Error Analysis

- Systems do not deliver perfect results (i.e., scores are below $100\,\%$)
- What can we say about the remaining errors?
- Workflow
  - Extract n errors, inspect them manually
  - Can we detect regularities/patterns in them? E.g., why they were misclassified?
  - Ideally, error analysis makes quantitative statements about error sources
- Directions for further improvements of the system

Analysis != Generation

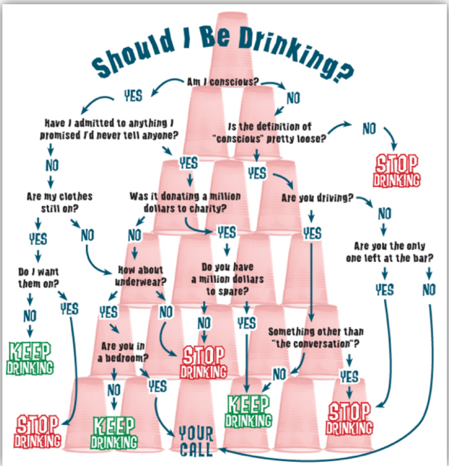
- Analysis: Text as input, annotations as output
- Generation: Some data as input, text as output
  - Machine translation, digital assistants, summarization, ...
- Different kinds of systems (not classification)
- Different evaluation metrics
  - Machine translation: BiLingual Evaluation Understudy (BLEU) Papineni et al. (2001)
    - Weighted overlap between reference and system

Decision Trees


Should I Be Drinking?
Am I conscious?
YES
NO
Have I admitted to anything I promised I'd never tell anyone?
Is the definition of "conscious" pretty loose?
NO
STOP DRINKING
NO
YES
YES
Are my clothes still on?
Was it donating a million dollars to charity?
Are you driving?
NO
YES
NO
NO
YES
YES
Are you the only one left at the bar?
Do I want them on?
How about underwear?
Do you have a million dollars to spare?
YES
NO
NO
YES
YES
NO
NO
YES
KEEP DRINKING
Something other than "the conversation"?
Are you in a bedroom?
STOP DRINKING
NO
YES
NO
YES
KEEP DRINKING
STOP DRINKING
KEEP DRINKING
YOUR CALL
STOP DRINKING

Decision Trees

Prediction Model – Toy Example



- What are the instances?
  - Situations we are in (this is not really automatisable)
- What are the features?
  - Consciousness
  - Clothing situation
  - Promises made
  - Whether we are driving
  - ...

Decision Trees

Prediction Model

- Each non-leaf node in the tree represents one feature
- Each leaf node represents a class label
- Each branch at this node represents one possible feature value
  - Number of branches $= |v(f_i)|$ (number of possible values)

Decision Trees

Prediction Model

- Each non-leaf node in the tree represents one feature
- Each leaf node represents a class label
- Each branch at this node represents one possible feature value
  - Number of branches $= |v(f_i)|$ (number of possible values)
- Make a prediction for x:
  1. Start at root node
  2. If it's a leaf node
     - assign the class label
  3. Else
     - Check node which feature is to be tested (f_i)
     - Extract $f_i(x)$
     - Follow corresponding branch
     - Go to 2

Decision Trees

Learning Algorithm

- Core idea: The tree represents splits of the training data
  1. Start with the full data set D_{train} as D
  2. If D only contains members of a single class:
     - Done.
  3. Else:
     - Select a feature f_i
     - Extract feature values of all instances in D
     - Split the data set according to f_i: $D = D_v \cup D_w \cup D_u \ldots$
     - Go back to 2
- Remaining question: How to select features?

And now:

The cliffhanger !

to be continued on Nov. 8th

References I

Birhane, Abeba/Vinay Uday Prabhu/Emmanuel Kahembwe (2021). "Multimodal datasets: misogyny, pornography, and malignant stereotypes". In: *CoRR* abs/2110.01963. arXiv: 2110.01963. URL: https://arxiv.org/abs/2110.01963.

Chamberlain, Jon/Massimo Poesio/Udo Kruschwitz (2008). "Phrase detectives: a web-based collaborative annotation game". In: *I-SEMANTICS '08: International Conference on Semantic Systems*. Graz, Austria: Verlag der Technischen Universität Graz, pp. 42–49.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

References II

Devlin, Jacob/Ming-Wei Chang/Kenton Lee/Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Fleiss, Joseph L. (1971). "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5, pp. 420–428.

Fournier, Chris (2013). "Evaluating Text Segmentation using Boundary Edit Distance". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1702–1712. URL: http://aclweb.org/anthology/P13-1167.

References III

- Hovy, Eduard/Julia Lavid (2010). "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In: *International Journal of Translation Studies* 22.1.
- Kohle, Hubertus (2010). "Artigo. Social image tagging pour les oeuvres d'art". In: *L'art et la mesure. Histoire de l'art et méthodes quantitatives*. Ed. by Béatrice Joyeux-Prunel. Paris: Ed. Rue d'Ulm, pp. 153–164.
- Manning, Christopher D./Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.
- Mathet, Yann/Antoine Widlöcher/Jean-Philippe Métivier (2015). "The Unified and Holistic Method Gamma ( ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3, pp. 437–479.
- Mikolov, Tomáš/Kai Chen/Greg Corrado/Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv cs.CL*. URL: https://arxiv.org/pdf/1301.3781.pdf.

References IV

- Pagel, Janis/Nils Reiter/Ina Rösiger/Sarah Schulz (2018). "A Unified Text Annotation Workflow for Diverse Goals". In: *Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI 2018*. Ed. by Sandra Kübler/Heike Zinsmeister. Sofia, Bulgaria.
- Papineni, Kishore/Salim Roukos/Todd Ward/Wei-Jing Zhu (2001). *Bleu: a method for automatic evaluation of machine translation*. IBM Research Division.
- Shannon, Claude E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423.