Statistical Natural Language Processing

Part II: Basics of Data Science

Henning Wachsmuth

https://ai.uni-hannover.de

Learning Objectives

Concepts

- The need for annotated text corpora
- Standard evaluation measures in NLP
- The most relevant basics from statistics

Methods

- Development and evaluation of methods on text corpora
- Selection of the right evaluation measure for a task
- Measuring of effectiveness in NLP
- The study of hypotheses with significance tests

Notice

- Some concepts and methods are reviewed here briefly only.
- For more details, see for example my bachelor's lecture on empirical methods in NLP: part05-empirical-methods.pdf

Outline of the Course

I. Overview

II. Basics of Data Science
   - Text Corpora
   - Evaluation Measures
   - Empirical Experiments
   - Conclusion

III. Basics of Natural Language Processing

IV. Representation Learning

V. NLP using Clustering

VI. NLP using Classification and Regression

VII. NLP using Sequence Labeling

VIII. NLP using Neural Networks

IX. NLP using Transformers

X. Practical Issues

Text Corpora

Text Corpora

Text corpus (plural text *corpora*)

- A principled collection of (mostly real-world) natural language texts with known properties, compiled to study a language problem

  Examples: 200,000 product reviews for sentiment analysis,
  1000 news articles for part-of-speech tagging, ...

- The texts in a corpus are often annotated, at least for the problem to be studied.

  Examples: Sentiment polarity of a full text,
  part-of-speech tags of each token, ...

https://pixabay.com

Dataset

- A subset of a corpus used for development or evaluation
- NLP methods are trained and tested on the datasets of a corpus.
- Without a corpus, it is hard to develop a strong method — and even harder to reliably evaluate it.

Text Corpora

Annotations

Annotation

- An annotation marks a text or a span of text as representing meta-information of a specific type.
- It may also be used to specify relations between other annotations.

Topic: ”Google revenues“ **Genre:** ”News article“

Manual vs. automatic annotation

- Manual. Most corpora are annotated by human experts or lay persons.
  NLP methods are developed based on such *ground-truth* annotations.
- Automatic. Technically, many NLP methods add annotations to texts.

Corpus Creation in 10 Steps

Input

1. Text compilation. Choose the texts to be included.
2. Annotation scheme. Define for what variables to annotate the texts.
3. Text preprocessing. Prepare texts for annotation.

Annotation process

4. Annotation sources. Decide who provides annotations.
5. Annotation guidelines. Define how to annotate.
6. Pilot annotation. Test the annotation process.
7. Inter-annotator agreement. Compute how reliable the annotations are.

Output

8. Postprocessing. Fix errors and filter annotations.
9. File representation. Store the annotated texts adequately.
10. Dataset splitting. Create subsets for training and testing.

Corpus Creation in 10 Steps

1. Text Compilation

Text compilation

- The first step is to collect the texts to be included.
- They should represent the application scenario of the task studied.
- Several types of potential data bias may need to be accounted for.
- Also, copyrights have to be considered.

Main design decisions

- Size. Usually, the more the better, but annotation must remain doable
- Domains. Topics, genres, languages, etc. (or combinations) to consider
- Confounders. Variables to control for (via balancing, defined ranges, ...)

  Examples: Publication time, length, or author

Example: ArguAna TripAdvisor corpus (Wachsmuth et al., 2014)

- 2100 English hotel reviews to annotate (+ 196k extra)

  All reviews were selected from an existing corpus (Wang et al., 2010).
- 300 reviews each for 7 locations, 420 each with rating 1–5

https://pixabay.com

Corpus Creation in 10 Steps

1. Text Compilation: Representativeness and Balance

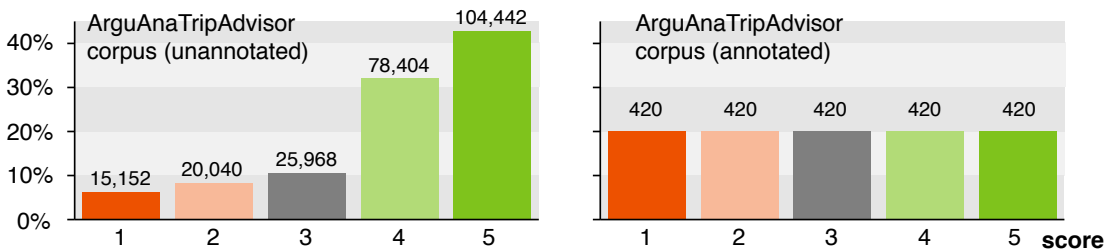
Representativeness

- A text compilation is representative for some variable X, if it includes the full range of variability of texts with respect to X.
- This is important for generalization, since a corpus governs what can be learned about a given task or domain.
- For evaluation, a representative distribution of texts is usually favorable.

Balance

- A text compilation is balanced if all values of X are represented evenly.
- For development, a balanced distribution may be favorable.

Example: ArguAna TripAdvisor corpus



Corpus Creation in 10 Steps

2. Annotation Scheme

Annotation scheme

- The definition of all annotation types to be considered in a task
- Clarifies syntax, semantics, and/or pragmatics behind each type
- Models what can be studied about the task on a corpus explicitly

Example: ArguAna TripAdvisor corpus

- Sentiment. Each statement classified as positive, negative, or neutral

  A statement was defined to be $\geq$ 1 clause, $\leq$ 1 sentence, and meaningful on its own.
- Aspects. Each aspect of a hotel marked
- Ratings. Each review scored for several quality dimensions

title: *great* **location**, *bad* **service** **sentiment score:** 2 of 5

body: *stayed at the darling harbour holiday inn. The* **location** *was great, right there at China town, restaurants everywhere, the monorail station is also nearby. Paddy's market is like 2 mins walk.* **Rooms** *were however very small. We were given the 1st floor* **rooms**, *and we were right under the monorail track, however* **noise** *was not a problem.* **Service** *is terrible.* **Staffs** *at the* **front desk** *were impatient, I made an enquiry about* **internet access** *from the* **room** *and the person on the phone was rude and unhelpful. Very shocking and unpleasant encounter.*

Corpus Creation in 10 Steps

3. Text Preprocessing

Text preprocessing

- The preparation of corpus texts for their manual annotation

Usual preprocessing steps

- Input files are converted into a common, usually simple format.
- Metadata is stored, in case it is considered relevant.
- The texts are analyzed, usually automatically, in order to create the instances to be annotated.

Example: ArguAna TripAdvisor corpus

- Originally, the input reviews were crawled HTML pages.
- The review contents were converted to plain text.
- The review ratings and other metadata were stored in annotations.
- Each text was automatically pre-segmented into statements.

  The rule-based segmentation algorithm used is provided with the corpus.

Corpus Creation in 10 Steps

4. Annotation Sources

Expert annotation

- Experts for a task or domain manually annotate each corpus text.
- Usually best results, but often time and cost intensive

Crowd-based annotation

- A crowdsourcing platform is used for manual annotation (e.g., *Upwork*)
- Access to many lay annotators (cheap) or semi-experts (not that cheap)
- Distant coordination overhead; results for complex tasks unreliable

Distant supervision

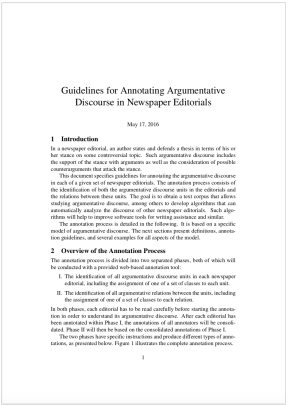
- Annotations are (semi-)automatically derived from existing metadata.
- Enables large corpora, but annotations may be noisy

Example: ArguAna TripAdvisor corpus

- Sentiment. Crowd-based annotation, three annotators each
- Aspects. Expert annotations, one expert per review (two for a sample)
- Ratings. Distant supervision; ratings obtained from review metadata

Corpus Creation in 10 Steps

5. Annotation Guidelines



Guidelines for Annotating Argumentative Discourse in Newspaper Editorials

May 17, 2016

1 Introduction

In a newspaper editorial, an author states and defends a thesis in terms of his or her stance on some controversial topic. Such argumentative discourse includes the support of the stance with arguments as well as the consideration of possible counterarguments that attack the stance.

This document specifies guidelines for annotating the argumentative discourse in each of a given set of newspaper editorials. The annotation process consists of the identification of both the argumentative discourse units in the editorials and the relations between these units. The goal is to obtain a text corpus that allows studying argumentative discourse, among others to develop algorithms that can automatically analyze the discourse of other newspaper editorials. Such algorithms will help to improve software tools for writing assistance and similar.

The annotation process is detailed in the following. It is based on a specific model of argumentative discourse. The next sections present definitions, annotation guidelines, and several examples for all aspects of the model.

2 Overview of the Annotation Process

The annotation process is divided into two separated phases, both of which will be conducted with a provided web-based annotation tool:

I. The identification of all argumentative discourse units in each newspaper editorial, including the assignment of one of a set of classes to each unit.

II. The identification of all argumentative relations between the units, including the assignment of one of a set of classes to each relation.

In both phases, each editorial has to be read carefully before starting the annotation in order to understand its argumentative discourse. After each editorial has been annotated within Phase I, the annotations of all annotators will be consolidated. Phase II will then be based on the consolidated annotations of Phase I.

The two phases have specific instructions and produce different types of annotations, as presented below. Figure 1 illustrates the complete annotation process.

1

Annotation guidelines

- To obtain reliable annotations, the annotators get guidelines on what and how to annotate.
- Guidelines include concepts, the annotation scheme, the annotation process, and often examples.

Length as a design decision

- Guidelines may span dozens of pages, but may also be very short.
- The more complete, the more annotations will reflect the authors' view.
- The more concise, the more decisions are left to the annotators' view.

Example: ArguAna TripAdvisor corpus

- For crowd-based sentiment, we had very short guidelines (+ examples):

> *"When visiting a hotel, are the following statements positive, negative, or neither?"*
>
> Notes. (1) Pick *neither* only for facts, not for unclear cases. (2) Pay attention to subtle statements where sentiment is expressed implicitly or ironically. (3) Pick the most appropriate answer in controversial cases.

Corpus Creation in 10 Steps

6. Pilot Annotation

Pilot annotation

https://pixabay.com

- Before a complete corpus is annotated, guidelines are usually tested on a small sample.
- The goal is to identify unclear parts, overseen and hard cases, and general annotation problems.
- Guidelines are often written incrementally based on pilot studies.

Annotators in pilot study

- Experts may discuss and align their annotation based on pilot results.
- Sometimes, the set of annotators is chosen based on pilot results.
- Rule of thumb. If authors don't agree, annotators won't either.

  Al Khatib et al. (2016) ommitted to annotate argumentative relations for this reason.

Example: ArguAna TripAdvisor corpus

- Sentiment. The guideline above was best among multiple variations.
- Aspects. The decision to use experts was based on pilot crowd tests.

Corpus Creation in 10 Steps

7. Inter-Annotator Agreement

Inter-annotator agreement (aka inter-rater reliability, inter-coder agreement)

- Quantification of the similarity of annotations by multiple annotators
  Commonly 2–5 annotators, sometimes more (especially in crowdsourcing)
- Between 1.0 (perfect agreement) and –1.0 (systematic disagreement)
  0.0 means random agreement.

Why inter-annotator agreement?

- Captures the reliability (or homogeneity) of the annotations of a corpus
- Gives a rough idea of how effective an algorithm may become
- Dilemma. Low agreement may indicate bad guidelines or insufficient training—but also just a subjective task.

Basis for computing agreement

1. Either, each corpus instance is annotated by multiple annotators.
2. Or, a sample is annotated multiple times, and the rest once each.
   1. is statistically more reliable and allows filtering, majority voting, etc.; 2. is cheaper.

Corpus Creation in 10 Steps

7. Inter-Annotator Agreement: Overview of Measures

Joint probability measures

- Observed agreement. $\%$ of nominal instances where 2 annotators agreed
- Full agreement. $\%$ of instances where $k \geq 3$ annotators all agreed
- Majority agreement. $\%$ of instances where $> 50\%$ annotators agreed

Chance-corrected measures

- More robust, taking into account that agreement may be due to chance
- Cohen's κ. Difference of observed to chance agreement (see below)
- Fleiss' κ. "Generalization" of Cohen's κ to $k \geq 3$ annotators
- Krippendorff's α. Focus on disagreement cases, any k, any scale

Correlation measures

- Quantify the (mean) pairwise correlation of annotators for ordinal scale
- Kendall's τ. Concordance of ranks of two instance orderings (see below)
- Spearman's ρ. Monotonicity of the relation between two orderings
- Pearson's r. Linear correlation between two sets of continuous values

Corpus Creation in 10 Steps

7. Inter-annotator Agreement: Cohen's κ

Cohen's κ

- For n instances annotated by two annotators, A and B, and a set of nominal categories C:

$$\kappa := \frac{p_o - p_e}{1 - p_e} \quad \text{where} \quad p_e := \frac{1}{n^2} \sum_{c \in C} a_c \cdot b_c$$

- p_o: observed agreement on instances
- p_e: expected agreement by chance
- a_c, b_c: number of times A and B chose class c

rough interpretation

κ Range	Agreement
[-1.0, 0.0]	No
(0.0, 0.2]	Slight
(0.2, 0.4]	Fair
(0.4, 0.6]	Moderate
(0.6, 0.8]	Substantial
(0.8, 1.0]	"Perfect"

Example

- $n = 100$, $p_o = 0.75$ for two categories c and c' ($a_c = b_c = 80$, $a'_c = b'_c = 20$)

$$p_e = \frac{1}{100^2} \cdot (6400 + 400) = 0.68 \quad \text{and thus} \quad \kappa = \frac{0.75 - 0.68}{1 - 0.68} \approx 0.22$$

Example: ArguAna TripAdvisor corpus

- Sentiment. Fleiss' $\kappa = 0.67$, full 73.6%, majority 98.3%
- Hotel aspects. Cohen's $\kappa = 0.73$ (based on 546 cases)

Corpus Creation in 10 Steps

7. Inter-annotator Agreement: Kendall's τ

Kendall's τ rank correlation coefficient

- Given n instances to be ranked, let $(a_1, b_1), \ldots, (a_n, b_n)$ be their joint ranks assigned by two annotators, A and B. Then:

$$\tau := \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{n \cdot (n-1)/2}$$

- Concordant. Any $(a_i, b_i), (a_j, b_j), i < j$: $a_i < a_j$ & $b_i < b_j$ or $a_i > a_j$ & $b_i > b_j$
- Discordant. Any $(a_i, b_i), (a_j, b_j), i < j$: $a_i < a_j$ & $b_i > b_j$ or $a_i > a_j$ & $b_i < b_j$

Adjustment for ties

- τ ignores the number of ties, t_A (for $a_i = a_j$) and t_B (for $b_i = b_j$).
- A common adjustment, τ', replaces the denominator of τ by:

$$\sqrt{(\#\text{conc. p.} + \#\text{disc. p.} + t_A) \cdot (\#\text{conc. p.} + \#\text{disc. p.} + t_B)}$$

Example

- $n = 3$, rank pairs: $(1, 2)$, $(2, 3)$, $(3, 3)$
- #conc. p. $= 2$, #disc. p. $= 0$, $t_A = 0$, $t_B = 1$

$$\tau = (2-0)/3 \approx 0.67$$
$$\tau' = (2-0)/\sqrt{6} \approx 0.82$$

Corpus Creation in 10 Steps

8. Postprocessing

Postprocessing

- The consolidation of the annotated texts for the final corpus
- Includes *cleansing* of potentially wrong or inconsistent cases
- May be manual and/or automatic

Common postprocessing steps

- Resolution (or discarding) of cases where annotators disagreed
- Removal of noise in the data observed during annotation
- Merging of labels that have been assigned only rarely

Example: ArguAna TripAdvisor corpus

- Each statement was assigned its majority sentiment where available.
- The 1.7% sentiment disagreement cases were resolved manually in the context of their associated reviews.
- Wrong aspect annotation boundary errors were fixed automatically.

Corpus Creation in 10 Steps

9. File Representation

File representation

- Usually, each corpus text is stored in a separated file.
- Large corpora may be stored in databases or indexes.
- Various file representations exist.

Common file representations

- Plain text only. One line per token, one tab per token-level annotation
- Plain text + annotation. Only text in file, extra file specifies annotations
- XMI/XML. One file for each text, one tag per annotation
- Spreadsheet. One row per text, one additional column per annotation

Example: ArguAna TripAdvisor corpus

- XMI files preformatted for the Apache UIMA framework
- Each annotation is stored as a tag with attributes and character indices.
- The annotation scheme is specified in a global descriptor XML file.

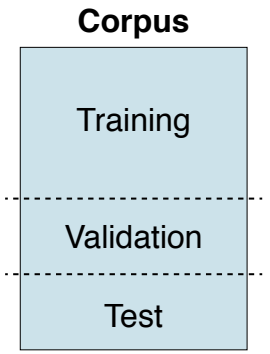
Corpus Creation in 10 Steps

10. Dataset Splitting

Dataset splitting

- How to split a corpus into training, validation, and test set (or similar), depends on the task.
- Good splits mimic the real-world situation of interest while avoiding *data leakage* that may be exploited in learning.

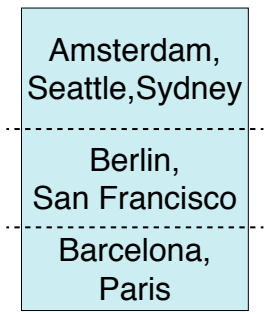
  Example: Annotations of one text should usually be in different sets.



Common splitting criteria

- Random. Split done (pseudo-) randomly
- Topic. Datasets (more or less) disjunct in terms of topic
- Time. Oldest texts for training, newest for testing

  Often, good splitting criteria are task or dataset-specific.



Example: ArguAna TripAdvisor corpus

- Location. 3 locations for training, 2 for validation, 2 for test

  This way, location-specific sentiment indicators cannot be exploited.

Evaluation Measures

Evaluation Measures

Evaluation measures in NLP

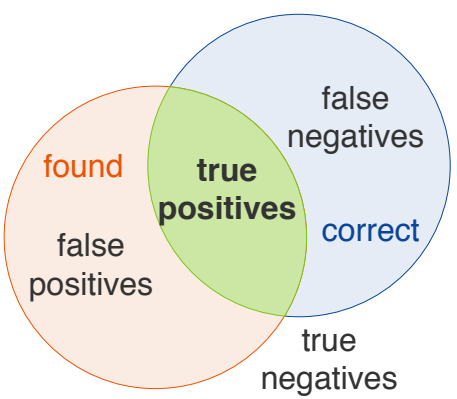
- A measure to quantify a method's quality on a given task and corpus
- Methods can be ranked with respect to an evaluation measure.
- Quality is mostly assessed in terms of *effectiveness*, that is, the extent to which the output information of a method is correct

Measuring effectiveness

- How to adequately measure effectiveness, depends on the task.
- Analysis. The output of a method is compared to the ground truth.
- Synthesis. Usually, not only one correct output exists.

Instance types in (analysis) tasks

- True positive (TP). Correctly found
- True negative (TN). Correctly not found
- False negative (FN). Mistakenly not found
- False positive (FP). Mistakenly found



Evaluation Measures

Accuracy

Accuracy

- The accuracy A is a measure of the correctness of a method.
- For $m = 2$ classes, accuracy is the ratio of positives under all instances.

$$A_{binary} := \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

- For $k > 2$ classes, accuracy is simply the ratio of true positives.

$$A_{multi} := \frac{|TP_1| + \ldots + |TP_k|}{|TP_1| + |FP_1| + \ldots + |TP_k| + |FP_k|}$$

When to use accuracy?

- Accuracy is often adequate when all classes are of similar importance.

  Examples: Sentiment analysis, part-of-speech tagging, ...

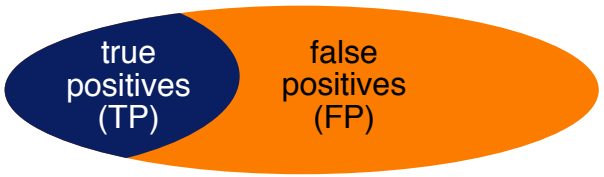
Evaluation Measures

Precision and Recall

Precision

- The precision P is a measure of the exactness of an approach.
- P answers: How many of the found instances are correct?

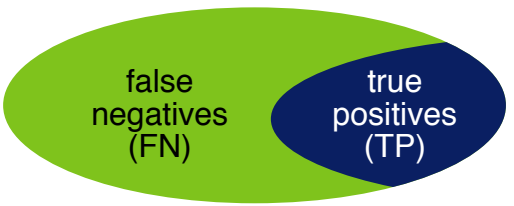
$$P := \frac{|TP|}{|TP| + |FP|}$$



Recall

- The recall R is a measure of the completeness of an approach.
- R answers: How many of the correct instances have been found?

$$R := \frac{|TP|}{|TP| + |FN|}$$



When to use precision and recall?

- Together, they are adequate if the focus is on the positive instances.

  Examples: Named entity recognition, plagiarism detection, ...

Evaluation Measures

F_1-score and Averaging

F_1-score (aka F_1-measure)

- The F_1-score is the harmonic mean of precision and recall.
- F_1 favors balanced over imbalanced precision and recall values.

$$F_1 := \frac{2 \cdot P \cdot R}{P + R}$$

Multi-class precision (recall and F_1-score analogous)

- In general, each class in a multi-class task can be evaluated binarily.
- Overall results are obtained with micro- or macro-averaging.
- Micro-averaging. Take into account the number of instances per class:

$$P_{micro} := \frac{|TP_1| + \ldots + |TP_k|}{|TP_1| + \ldots + |TP_k| \quad + \quad |FP_1| + \ldots + |FP_k|}$$

- Macro-averaging. Compute the mean result over all classes:

$$P_{macro} := \frac{P_1 + \ldots + P_k}{k}$$

Effectiveness Measures for Numerical Predictions

Evaluating real-valued predictions

- It is unlikely to predict exact real values in many tasks.
- In such cases, the focus is on the difference to the correct value.

Mean absolute error (MAE)

- The mean difference of predicted values y_i to ground-truth values $\hat{y}_i$
- MAE does not treat outliers specifically.

$$MAE := \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Mean squared error (MSE)

- The mean squared difference of predicted to ground-truth values
- MSE is sensitive to outliers.

  Sometimes, the root mean squared error (RMSE) is used: $RMSE = \sqrt{MSE}$.

$$MSE := \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Effectiveness Measures for Synthesis Tasks

Evaluating synthesis tasks

- Goal: Judge quality of generated texts.
- Problem: Multiple outputs may be correct.

Ground truth. “Abolish death penalty forever.”
Generated. “Let’s ban the death penalty.”

Two types of evaluation

- Automatic. Quantify similarity between ground truth and generated text.
- Manual. Human annotators assess the quality of generated texts.

Dilemma

- Only manual evaluation is seen as reliable, but it costs time and money.
- Automatic evaluation is needed to observe progress while developing.

Effectiveness Measures for Synthesis Tasks

Overview of Measures

Automatic evaluation measures

- BLEU. Precision of n-gram overlap with brevity penalty
- ROUGE. Recall of n-gram overlap, either for a specific n or averaged
- BERTScore. F_1 derived from similarity matching of text embeddings

  Formulas left out here for brevity

Manual evaluation measures

- Usually, quality dimensions are scored on a Likert scale (say, 1–5).
- The mean or majority judgment of annotators is used for evaluation.

  Sometimes, different candidates are also ranked relatively.

Selected quality dimensions for synthesis tasks

- Syntax. Gramaticality, fluency, naturalness, ...
- Semantics. Meaning preservation, coherence, ...
- Pragmatics. Relevance, informativeness, ...

  What dimensions to assess, depends on the task.

Empirical Experiments

Empirical Experiments

Empirical experiments in NLP

- An empirical experiment tests a hypothesis based on observations.
- The focus is here on effectiveness evaluation in NLP.

Intrinsic vs. extrinsic evaluation

- Intrinsic. The effectiveness of an approach is directly evaluated on the task it is made for.

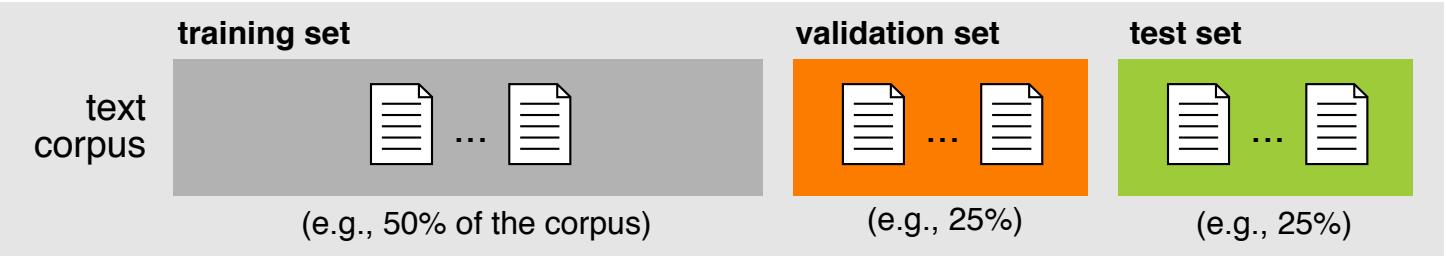
  "What accuracy does a part-speech tagger *XY* have on the dataset *D*?"

- Extrinsic. The effectiveness of an approach is evaluated by measuring how effective its output is in a downstream task.

  "Does the ouput of *XY* improve sentiment analysis on *D'*?"

Empirical Experiments

Training, Validation, and Test



Training set

- Known instances used to develop or statistically learn an approach
- The training set may be analyzed manually and automatically.

Validation set (aka development set)

- Unknown test instances used to iteratively evaluate an approach
- The approach is optimized on (and adapts to) the validation set.

Test set (aka held-out set)

- Unknown test instances used for the final evaluation of an approach
- The test set represents unseen data.

Empirical Experiments

Cross-Validation

n-fold cross-validation

- Data is split into n dataset folds of equal size, often $n = 10$.
- The evaluation results are averaged over n runs.

Training and evaluation

- In the i-th run, the i-th fold is used for evaluation (validation).
- All other folds are used for development (training).

Empirical Experiments

Comparison

Why comparing?

- A new method is seen as useful, if it is better than other methods, usually measured in terms of effectiveness.
- To test this, methods are compared to *baselines*.

Baseline

- A baseline is an alternative method that has been proposed before or that can easily be realized.
- Ideally, a new method should be better than all baselines.

Types of baselines

- Trivial. Methods that can easily be derived from a given task or dataset
- Standard. Methods that are often used for related tasks
- Ablation. Sub-methods of a newly proposed method
- State of the art. The best published method for the task (if available).

Empirical Experiments

Descriptive Statistics

Descriptive statistics

- Methods for summarizing a sample $\tilde{X}$ (or distribution X) of values in order to *describe phenomena*

Selected measures

- Mean. The arithmetic average M of a sample of values $\tilde{X}$ of size n

  M is used for a sample, μ for the whole distribution.

$$M := \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i$$

- Variance. The mean s^2 of all values' squared differences to the mean

  s is used for a sample, σ for the whole distribution.

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{X}_i - M)^2 \qquad \sigma^2 := \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i - \mu)^2$$

- Standard deviation. The square root of the variance

$$s = \sqrt{s^2} \qquad \sigma = \sqrt{\sigma^2}$$

Empirical Experiments

Inferential Statistics

Inferential statistics

- Methods for drawing conclusions based on values in order to *generalize inferences* beyond a given sample $\tilde{X}$

Two competing hypotheses

- Research hypothesis (H). Prediction about how a change in variables will cause changes in other variables

> "There is a statistically significant difference between the RMSE of our approach and the RMSE reported by Persing et al. (2015)."

- Null hypothesis (H_0). Antithesis to H

> "There is *no* statistically significant difference between the RMSE of our approach and the RMSE reported by Persing et al. (2015)."

- If H_0 is true, then any results observed in an experiment that support H are due to chance or sampling error.

Hypothesis Testing

Statistical significance test (aka hypothesis test)

- A statistical procedure that determines how likely it is that the results of an experiment are due to chance (or sampling error)
- It tests whether a null hypothesis H_0 can be rejected (and hence, H can be accepted) at some chosen *significance level*.

Significance level α

- The accepted risk (in terms of a probability) that H_0 is wrongly rejected
- A choice of $\alpha = 0.05$ means that there is at least a 95% chance that a potential rejection of H_0 is correct.

  Usually, α is set to 0.05 (default) or to 0.01.

p-value

- The likelihood (in terms of a probability) that results are due to chance.
- If $p \leq \alpha$, H_0 is rejected. The results are seen as statistically significant.
- If $p > \alpha$, H_0 cannot be rejected.

Hypothesis Testing

Carrying out a Significance Test

Main test steps

1. Hypothesis. State H and H_0.
2. Significance level. Choose α (*before* the test).
3. Testing. Carry out a significance test, which fits the data, to get the p-value.
4. Decision. Reject H_0 or fail to reject it.

https://pixabay.com, https://xkcd.com/892/

Significance tests

- Different tests exist that make different assumptions about the data.
- Parametric. More likely to detect a significant effect when one exists
- Non-parametric. Fewer assumptions and, thus, more often applicable

Parametric test	Non-parametric correspondent
Independent student's t-test	Mann-Whitney Test
Dependent and one-sample student's t-test	Wilcoxon Signed-Rank Test
...	...

Conclusion

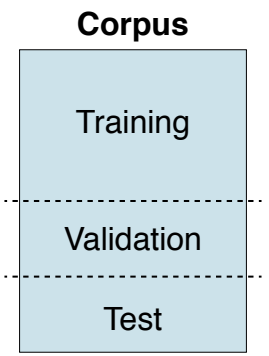
Conclusion

Data Science in NLP

- Text corpora used for development and evaluation
- Evaluation measures quantify effectiveness of methods
- Empirical experiments test methods on data

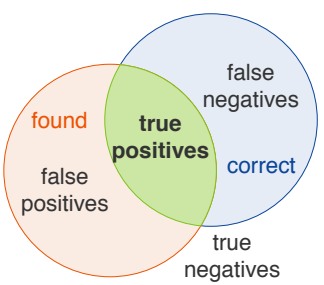
Text corpora

- Text collections compiled to study a language problem
- Often manually annotated and split into datasets
- Corpus creation is a complex, yet important process



Measures and experiments

- Various measures for analysis and synthesis tasks
- Experiments compare methods against baselines
- Statistical tests explore and “prove” quality of methods



References

Some content taken from

- **Ng (2018).** Andrew Ng. Machine Learning. Lecture slides from the Stanford Coursera course, 2018. https://www.coursera.org/learn/machine-learning
- **Jurafsky and Manning (2016).** Daniel Jurafsky and Christopher D. Manning. Natural Language Processing. Lecture slides from the Stanford Coursera course, 2016. https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html
- **Rockinson-Szapkiw (2013).** Amanda J. Rockinson-Szapkiw. Statistics Guide, 2013. http://amandaszapkiw.com/elearning/statistics-guide/downloads/Statistics-Guide.pdf
- **Wachsmuth (2015).** Henning Wachsmuth. Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer, 2015.
- **Witten and Frank (2005).** Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, CA, 2nd edition, 2005.

References

Other references

- **Al Khatib et al. (2016).** Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A News Editorial Corpus for Mining Argumentation Strategies. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3433–3443, 2016.
- **Wachsmuth et al. (2014).** Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palarkarska. A Review Corpus for Argumentation Analysis. In Proceedings of the of the 15th International Conference on Intelligent Text Processing and Computational Linguistics, pages 115–127, 2014.
- **Wang et al. (2010).** Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In: Proceedings of the 16th SIGKDD. pages 783–792, 2010.