Introduction to Natural Language Processing

Part V: Basics of Empirical Methods

Henning Wachsmuth

https://ai.uni-hannover.de

Learning Objectives

Concepts

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

Methods

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

Outline of the Course

Introduction

Empirical Methods

Empirical method

- A quantitative method that analyzes numbers and/or statistics to study a *research question* on behaviors or phenomena
- Derives knowledge from experience (rather than from theory or belief)

Quantitative vs. qualitative methods

- Quantitative. Characterized by objective measurements
- Qualitative. Emphasizes the understanding of human experience

Descriptive and inferential statistics

- Descriptive. Methods for summarizing a sample or a distribution of values; used to *describe phenomena*

  4.5, 5, 6, 6.5, 6.5, 7, 7, 7, 7.5, 8 $\rightarrow$ mean M = 6.5

- Inferential. Methods for drawing conclusions based on values; used to *generalize inferences* beyond a given sample

  The average number is significantly higher than 5.

Empirical Methods

NLP and Empirical Methods

NLP (recap)

- Understanding and generating human-readable text (or speech) using computational methods
- Use of rule-based or statistical approaches for this purpose
- The output information produced is not always correct.

Elements of empirical methods in NLP

- Evaluation measures. Quantification of the quality of a method, especially its *effectiveness*
- Empirical experiments. Evaluation of the quality on *text corpora* and comparison to alternative methods
- Hypothesis testing. Use of statistical methods to “proof” the quality of a method in comparison to others

Evaluation Measures

Evaluation Measures

Evaluation measures in NLP

- In NLP, an evaluation measure quantifies the quality of a method on some task and data.
- Methods can be ranked with respect to an evaluation measure.
- Quality is assessed in terms of *effectiveness* or *efficiency*.

Effectiveness

- The extent to which the output information of an approach is correct
- High effectiveness is the primary goal of any NLP method.
- Classification measures. Accuracy, precision, recall, F_1-score, ...
- Regression measures. Mean absolute/squared error, ...

Efficiency

- The costs of a method in terms of the consumption of time or space
- Measures. Run-time (per unit), training time, memory consumption, ...

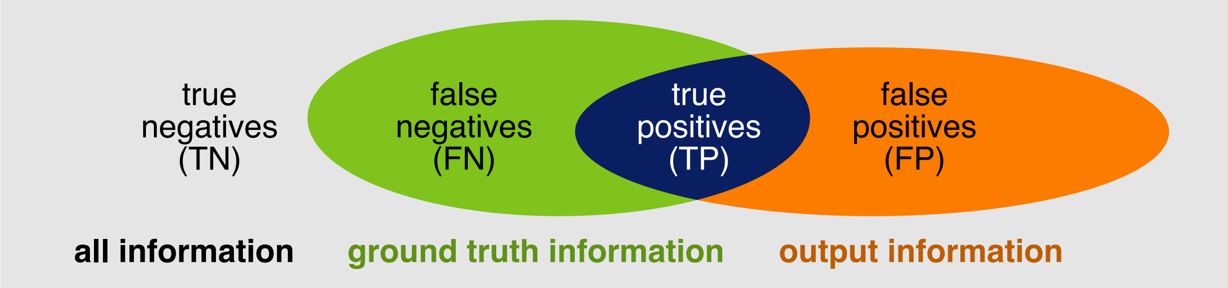
  Efficiency will be discussed at the end of this course.

Classification Effectiveness

Classification tasks

- The instances of each class can be evaluated in a binary manner.
- For each instance, check whether its class matches the ground truth.

- Positives. The class instances a given approach has inferred
- Negatives. All other possible instances



Instance types in the evaluation

- True positive (TP). A positive that belongs to the ground truth
- False positive (FP). A positive that does not belong to the ground truth
- False negative (FN). A negative that belongs to the ground truth
- True negative (TN). A negative that does not belong to the ground truth

Classification Effectiveness

Accuracy

Accuracy

- The accuracy A is a measure of the correctness of an approach.
- A answers: How many classification decisions are correct?
- For $k = 2$ classes, accuracy is the ratio of true 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 adequate when all classes are of similar importance.
- Examples: Sentiment analysis, part-of-speech tagging, ...

“The”/DT “man”/NN “sighed”/VBD “.”/. “It”/PRP “‘s”/VBZ “raining”/VBG ...

Classification Effectiveness

Limitations of Accuracy

Example: Spam detection

- Assume 4% of the mails that your mail server lets through are spam.
- What is the accuracy of a spam detector that always predicts "no spam" for mails?

When *not* to use accuracy?

- In tasks where one class is rare, high accuracy can be achieved by never predicting the class.

  4% spam $\rightarrow$ 96% accuracy by always predicting "no spam"

- This includes tasks where the correct output information covers only portions of text, such as in entity recognition.

  "Apple rocks." $\rightarrow$ Negatives: "A", "Ap", "App", "Appl", "Apple ", "Apple r", ...

- Accuracy is inadequate when true negatives are of low importance.

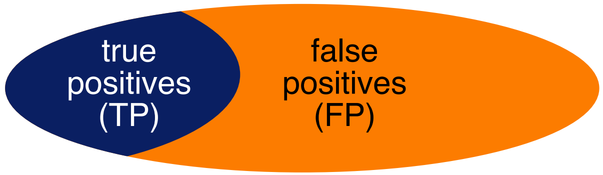
Classification Effectiveness

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?

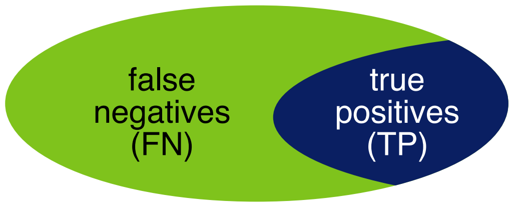
$$P = \frac{|TP|}{|TP| + |FP|}$$



Recall

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

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



Observation

- True negatives not included in formulas

Classification Effectiveness

Idea of Precision and Recall

Example: Spam detection (revisited)

- Assume 4% of the mails that your mail server lets through are spam.
- What precision and recall does the “always no spam” detector have for detecting spam?

https://datenschutz.org

Idea of precision and recall

- Put the focus on a specific class (here: “spam”).
- The typical case is that the true negatives are irrelevant.
- If multiple classes are important, precision and recall can be computed for each class (see below).

Example: Spam detection (a last time)

- What precision and recall does an “always spam” detector have?

$P = 0.04 \quad R = 1.0$

Classification Effectiveness

F_1-Score

What is better?

- A precision of 0.52 and a recall of 0.52 (option a).
- A precision of 0.04 and a recall of 1.00 (option b).
- Often, a single effectiveness value is desired.

Problem with the mean

- In the above example, the mean would be the same for both options.
- But 96% of the found instances are wrong for option b.

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.

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

Option a: $F_1 = 0.52$, option b: $F_1 = 0.08$

Classification Effectiveness

F_1-Score in Boundary Detection Tasks

Boundary errors

- A common error in tasks where text spans need to be annotated is to choose a (slightly) wrong boundary of the span.

Entities:	"First Bank of Chicago stated..." vs. "First Bank of Chicago stated..."
Sentences:	"Max asked: 'What's up?'" vs. "Max asked: 'What's up?'"

Issue with boundary errors

- Boundary errors leads to both an FP and an FN.
- Identifying nothing as a positive would increase the F_1-score.

How to deal with boundary errors?

- Different accounts for the issue have been proposed, but the standard F_1 is still used in most evaluations.
- A relaxed evaluation is to consider some character overlap (e.g., >50%) instead of exact boundaries.

Classification Effectiveness

Micro-Averaging and Macro-Averaging

Evaluation of multi-class tasks

- In general, each class in a multi-class task can be evaluated binarily.
- Accuracy can be computed for any number k of classes (as seen).
- The other measures must be combined with micro- or macro-averaging.

Micro-averaged precision (recall and F_1-score analog)

- Micro-averaging takes into account the number of instances per class, so larger classes get more importance.

$$Micro\text{-}P = \frac{|TP_1| + \ldots + |TP_k|}{|TP_1| + \ldots + |TP_k| + |FP_1| + \ldots + |FP_k|}$$

Macro-averaged precision (recall and F_1-score analog)

- Macro-averaging computes the mean result over all classes, so each class gets the same importance.

$$Macro\text{-}P = \frac{P_1 + \ldots + P_k}{k}$$

Classification Effectiveness

Confusion Matrix

Confusion matrix

- Each row refers to the ground-truth instances of one of k classes.
- Each column refers to the classified instances of one class.
- The cells contain the numbers of correct and incorrect classifications of a given approach.

Ground truth	Classified as			
	Class a	Class b	...	Class k
Class a	$\|TP_a\|$	$\|FP_b \cap FN_a\|$	...	$\|FP_k \cap FN_a\|$
Class b	$\|FP_a \cap FN_b\|$	$\|TP_b\|$	...	$\|FP_k \cap FN_b\|$
...	...	...	...	...
Class k	$\|FP_a \cap FN_k\|$	$\|FP_b \cap FN_k\|$	...	$\|TP_k\|$

Why confusion matrices?

- Used to analyze errors, to see which classes are confused
- Contains all values for computing micro- and macro-averaged results

Classification Effectiveness

Computation of Micro- and Macro-Averaged Values

Example: Evidence classification

- Assume an approach that classifies candidate evidence statements as being an “anecdote”, “statistics”, “testimony”, or “none” is given.

Confusion matrix of the results

Ground-truth	Classified as			
	Anecdote	Statistics	Testimony	None
Anecdote	199	5	35	183
Statistics	17	29	0	27
Testimony	30	1	123	71
None	118	7	36	1455

Total		Precision
TP	FP	per class
199	165	0.55
29	13	0.69
123	71	0.63
1455	281	0.84

Micro- vs. macro-averaged precision (recall and F_1-score analog)

$$Micro\text{–}P = \frac{199+29+123+1455}{199+29+123+1455+165+13+71+281} = 0.77$$

$$Macro\text{–}P = \frac{0.55+0.69+0.63+0.84}{4} = 0.68$$

Regression Effectiveness

Regression task

- In a regression task, numeric values are predicted for instances from a continuous scale.
- The scale is usually, but not necessarily, predefined.

Example: Automatic essay grading

- Given a set of n student essays, automatically assign each essay i a score $1 \leq y_i \leq 4$.
  The 4-point scale is the default in today's grading systems.

https://okpolicy.org

Prediction errors

- In many regression tasks, it is unlikely to predict the exact values of instances. Therefore, accuracy is not the primary measure.
- Instead, the *error* of the predicted values $Y = (y_1, \ldots, y_n)$ compared to the ground-truth values $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)$ is in the focus.

Regression Effectiveness

Types of Prediction Errors

Mean absolute error (MAE)

- The mean difference of predicted to ground-truth values
- The MAE is robust to outliers, i.e., it does not treat them specially.

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

Mean squared error (MSE)

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

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

- Sometimes, also the root mean squared error (RMSE) is computed, defined as $RMSE = \sqrt{MSE}$.

Regression Effectiveness

Computation

Example: Automatic essay grading (revisited)

- Assume we have three automatic essay grading approaches applied to 10 essays resulting in the following scores:

https://okpolicy.org

	Essay										Prediction error	
Approach	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**MAE**	**MSE**
Approach 1	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	0.88	1.04
Approach 2	1.0	3.2	2.0	2.1	3.0	3.1	2.8	3.1	1.2	4.0	0.55	1.28
Approach 3	1.5	2.0	1.5	2.5	2.0	2.7	3.3	3.5	3.2	3.6	0.58	0.40
Ground truth	1.0	1.0	2.0	2.0	3.0	3.0	3.0	3.0	4.0	4.0	0.00	0.00

Which approach is best?

- Approach 1 trivially always predicts the mean $\rightarrow$ useless in practice.
- Approach 2 has a better MAE than Approach 3, but fails with its MSE.
- Whether MAE or MSE is more important, depends on the application.

  In essay grading, outliers are particularly problematic.

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 effectiveness 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'*?"

Corpus-based experiments vs. user studies

- We consider the empirical evaluation of approaches on *corpora* here.
- A whole different branch of experiments is related to *user studies*.

  Not covered in this course

Text Corpora

Text corpus (plural: 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

Need for text corpora

- NLP approaches are developed and evaluated on text corpora.
- Without a corpus, it's hard to develop a strong approach — and impossible to reliably evaluate it.

*"It's not the one who has the best algorithm that wins.
It's who has the most data."*

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.
- The types are specified by an annotation scheme.

Time entity **Organization entity**

" 2014 ad revenues of Google are going to reach

Reference **Time entity**

$20B. The search company was founded in '98.

Reference **Time entity** **Founded relation**

Its IPO followed in 2004. [...] "

Topic: "Google revenues" **Genre:** "News article"

Annotated corpora in NLP

- Usually, a corpus contains annotations of information types of interest in a task or domain.

Text Corpora

Ground Truth vs. Automatic Annotation

Manual annotations

- The annotations of a text corpus are usually created manually.
- To assess the quality of manual annotations, inter-annotator agreement is computed based on texts annotated multiple times.

  Standard chance-corrected measures: Cohen's κ, Fleiss' κ, Krippendorff's α, ...

Ground-truth annotations

- Manual annotations assumed to be correct are called the *ground truth*.
- NLP usually learns from ground-truth annotations.

Automatic annotation

- Technically, NLP algorithms can be seen as just adding annotations of certain types to a processed text.
- The automatic process usually aims to mimic the manual process.

Development and Evaluation

Dataset

- A sub-corpus of a corpus that is compiled and used for developing and/or evaluating approaches to specific tasks

Development and evaluation based on datasets

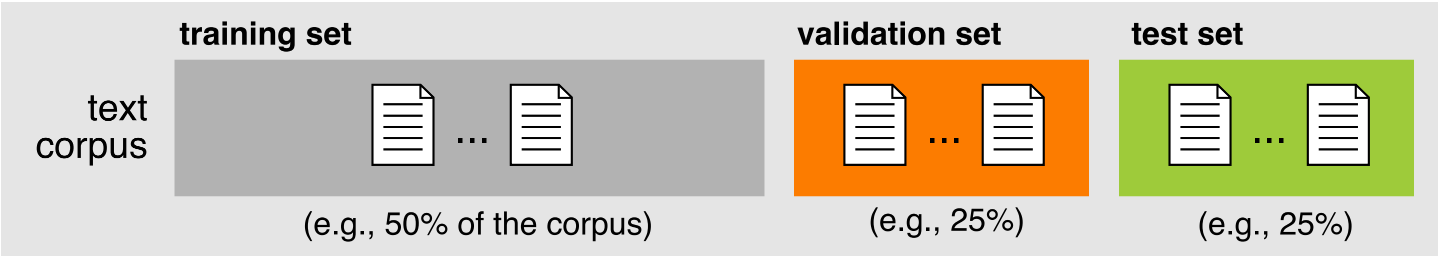
1. An approach is developed based on a set of training instances.
2. The approach is applied to a set of test instances.
3. The output of the approach is compared to the ground truth of the test instances using evaluation measures.
4. Steps 1–3 may be iteratively repeated to improve the approach.

Corpus splitting

- The split of a corpus into datasets should represent the task well.
  Out of scope here. Example: No overlap of instances from one text in different sets.
- The way a corpus is split implies how to evaluate.
- Main evaluation types. *Training, validation, and test* vs. *cross-validation*

Development and Evaluation

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.

Development and Evaluation

Cross-Validation

(Stratified) n-fold cross-validation

- A corpus is split into n dataset folds of equal size, usually $n = 10$.
  Stratification: The target variable distribution is kept stable across folds.
- n runs. The evaluation results are averaged over n runs.
- i-th run. The i-th fold is used for evaluation (validation). All other folds are used for development (training).

Pros and cons of cross-validation

- Often preferred when data is small, as more data is given for training
- Cross-validation avoids potential bias in a corpus split.
- Random splitting often makes the task easier, due to corpus bias.

Development and Evaluation

Variations

Repeated cross-validation

- Often, cross-validation is repeated multiple times with different folds.
- This way, coincidental effects of random splitting are accounted for.

Leave-one-out validation

- Cross-validation where n equals the number of instances
- This way, any potential bias in the splitting is avoided.
- But even more data is given for training, which makes a task easier.

Cross-validation + test set

- When doing cross-validation, a held-out test set is still important.
- Otherwise, repeated development will overfit to the splitting.

Comparison

Example: Evidence classification (revisited)

- Assume an evidence classification approach obtains an accuracy of 60% on a given test set, how good is this?

Selected factors that influence effectiveness

- The number of classes and their distribution in the training set
- The class distribution in the test set
- The heterogeneity of the test set
- The similarity between training, validation, and test set
- The representativeness of the test set
- The complexity of the task

Observation

- Some factors can be controlled or quantified, but not all.
- To assess the quality of an approach, we need *comparison*.

Comparison

Upper Bounds and Lower Bounds

Why comparing?

- A new approach is seen as useful if it is better than other approaches, usually measured in terms of effectiveness.
- Approaches may be compared to a *gold standard* and to *baselines*.

Gold standard (upper bound)

- The gold standard represents the best possible result on a given task.
- For many tasks, the effectiveness that humans achieve is seen as best.
- If not available, the gold standard is often equated with the ground truth in a corpus. This means: perfect effectiveness.

Baseline (lower bound)

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

Comparison

Types of Baselines

Trivial baselines

- Methods that can easily be derived from a given task or dataset
- Used to evaluate whether a new approach achieves anything

Standard baselines

- Methods that are often used for related tasks
- Used to evaluate how hard a task is

Sub-approaches

- Sub-approaches of a new approach
- Used to analyze the impact of the different parts of an approach

State of the art

- The best published methods for the addressed task (if available)
- Used to verify whether a new approach is best

Comparison

Exemplary Baselines

Example: Evidence classification (revisited)

- Assume an evidence classification approach obtains an accuracy of 60% on a given test set, how good is this?

https://pixabay.com

Exemplary data distribution (Al Khatib et al., 2016)

• Four classes.	"anecdote",	"statistics",	"testimony",	"none" (majority)
• Test distribution.	18%	3%	10%	69%

Potential baselines

- Trivial. Random guessing achieves an (expected) accuracy of 25%.
- Trivial. Always predicting the majority achieves 69%.
- Standard. Using the distribution of word {1, 2, 3}-grams achieves 76%.
- State of the art. The best published value is 78%. (Al Khatib et al., 2017)

Comparison

Implications

When does comparison work?

- Variations of a task may affect its complexity.
- The same task may have different complexity on different datasets.
- Only in *exactly* the same experiment setting, two approaches can be compared reasonably.

Example: Evidence classification (a last time)

- Assume evidence classification approach A obtains an accuracy of 79%, and approach B 78% in exactly the same setting.
- Is A better than B?

https://pixabay.com

How to know that some effectiveness is better?

- Effectiveness differences may be coincidence.
- The significance of differences can be “proven” statistically.

Hypothesis Testing

Statistics

Variable

- An entity that can take on different quantitative or qualitative values

  A variable thereby represents a distribution of values.
- Independent. A variable X that is expected to affect another variable
- Dependent. A variable Y that is expected to be affected by others

  Other types not in the focus here: Confounders, mediators, moderators, ...

Possible causes $X_1, \ldots, X_k \quad \rightarrow \quad$ Effect Y

Scales of variables

- Nominal. Values that represent discrete, separate categories
- Ordinal. Values that can be ranked by what is better
- Interval. Values whose distance can be measured
- Ratio. Interval values that have a "true zero"

  A true zero indicates the absence of what is represented by a variable.

Interval vs. ratio scale test

- Only for ratios, it is right to say that a value is twice as high as another.

Statistics

Variables and Scales

What is independent, what is dependent?

"Does our sentiment analysis approach achieve higher accuracy with features based on part-of-speech tags than without them?"

Independent: features based on part-of-speech tags
Dependent: accuracy

What type of scale?

1. Celsius temperature
2. Exam grades
3. Phone prices
4. Colors
5. Text length

1. Interval   2. Ordinal   3. Ratio   4. Nominal   5. Ratio

Descriptive Statistics

Descriptive statistics

- Measures for summarizing (samples $\tilde{X}$ of) distributions of values X
- Used to describe phenomena

Measures of central tendency

- 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.

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

- Median. The middle value Mdn of the ordered values in a sample

  Even size: The value halfway between the two middle values

$$Mdn = (\tilde{X}_{\lfloor \frac{n+1}{2} \rfloor} + \tilde{X}_{\lceil \frac{n+1}{2} \rceil}) \;/\; 2$$

- Mode. The value Md with the greatest frequency in a sample

Descriptive Statistics
Disperson

Measures of dispersion

- Range. The distance r between minimum and maximum

$$r = \tilde{X}_{max} - \tilde{X}_{min}$$

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

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

$$\text{biased}: \; s^2 = \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i - M)^2 \qquad \text{unbiased}: \; s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{X}_i - M)^2$$

- Standard deviation. The square root s of the variance

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

Biased vs. unbiased variance

- The biased variance formula tends to underestimate the real variance of the distribution.
- For samples, the unbiased variance formula is used in statistics.

  The division by $n-1$ instead of n corrects for the small sample size.

Descriptive Statistics

Example

Measures for an ordered sample of 10 values

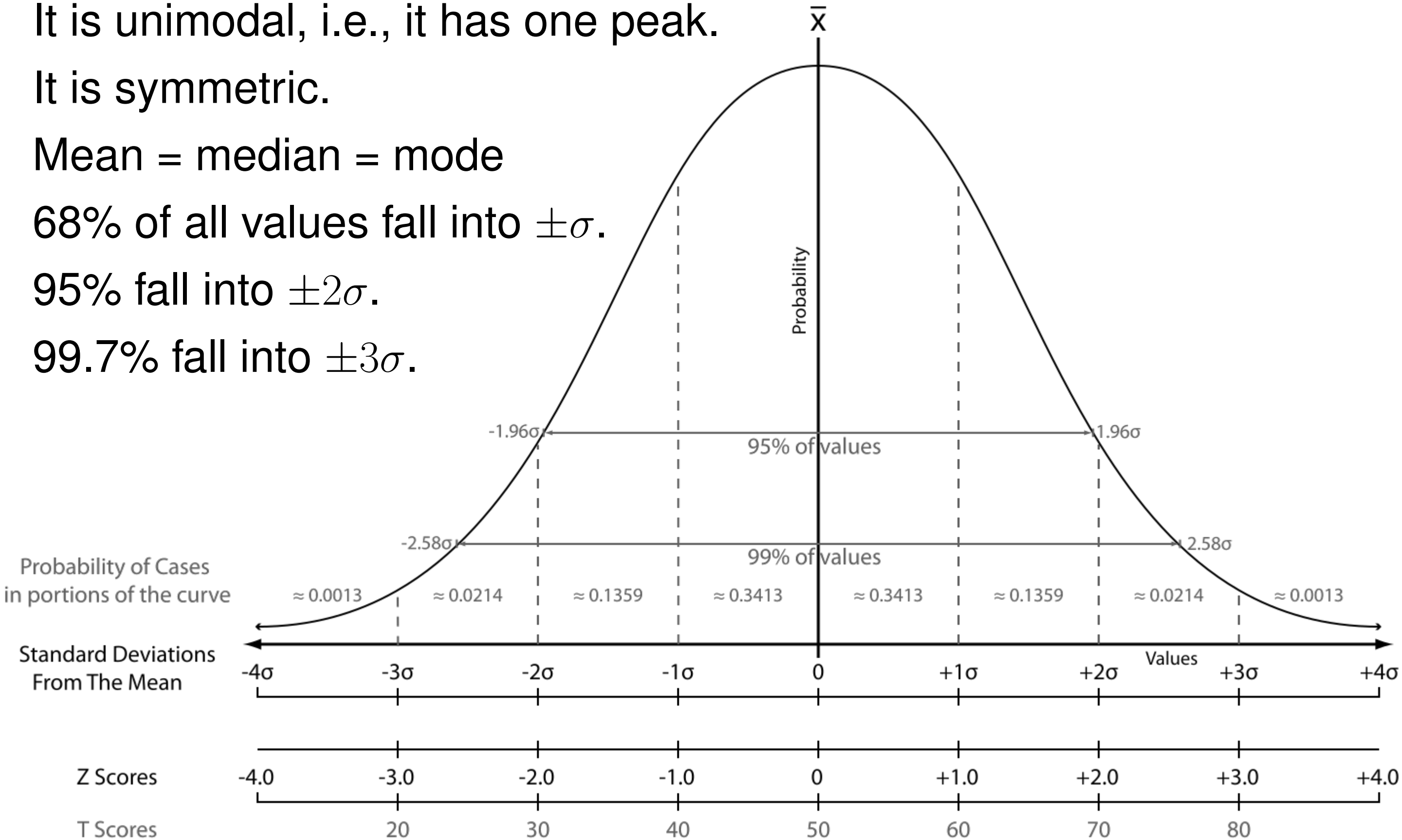
$$
\begin{aligned}
\tilde{X} &= (1, 3, 3, 3, 5, 6, 6, 7, 10, 15) \\
M &= \tfrac{1}{10}\textstyle\sum_{i=1}^{10} \tilde{X}_i \quad = \quad 5.9 \\
Mdn &= (\tilde{X}_4 + \tilde{X}_5)\ /\ 2 \quad = \quad 5.5 \\
Md &= 3 \\
r &= \tilde{X}_{10} - \tilde{X}_1 \quad = \quad 14 \\
s^2 &= \tfrac{1}{9}\textstyle\sum_{i=1}^{10}(\tilde{X}_i - M)^2 \quad \approx \quad 15.97 \\
s &= \sqrt{s^2} \quad \approx \quad 4.00
\end{aligned}
$$

Descriptive Statistics

Normal Distribution

Normal distribution (aka Gaussian distribution)

- The frequency distribution that follows a normal curve
- It is unimodal, i.e., it has one peak.
- It is symmetric.
- Mean = median = mode
- 68% of all values fall into $\pm\sigma$.
- 95% fall into $\pm 2\sigma$.
- 99.7% fall into $\pm 3\sigma$.



https://commons.wikimedia.org

Descriptive Statistics

Standard Scores

Standard score

- Indicates how many standard deviations a value is away from the mean of a distribution X

z-score

- Indicates the precise location of a value X_i within a distibution X

  Positive if above the mean, negative otherwise

$$z = \frac{X_i - \mu}{\sigma} \quad \text{approximated as} \quad z = \frac{\tilde{X}_i - M}{s}$$

t-score

- Transforms a value $\tilde{X}_i$ from a sample of size n into a standardized comparable form

  Usually used for small samples (with less than ∼30 values)

$$t = \frac{\tilde{X}_i - M}{s/\sqrt{n}}$$

Inferential Statistics

Inferential statistics

- Procedures that help study *hypotheses* based on values
- Used to make inferences about a distribution beyond a given sample

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.

Inferential Statistics

Hypotheses

Two types of hypotheses

- Directional. Specifies the direction of an expected difference

> "The RMSE of our approach is statistically significantly lower than the RMSE reported by Persing et al. (2015)."

- Non-directional. Specifies only that any difference is expected

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

Good hypotheses (Rockinson-Szapkiw, 2013)

- Founded in a problem statement and supported by research
- Testable, i.e., it is possible to collect data to study the hypothesis
- States an expected relationship between variables
- Phrased as simply and concisely as possible

Hypothesis Testing

Hypothesis test (aka statistical significance test)

- A statistical procedure that determines how likely it is that the results of an experiment are due to chance or sampling error
- 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
  Usually, α is set to 0.05 (default) or to 0.01.
- A choice of α = 0.05 means that there is no more than 5% chance that a potential rejection of H_0 is wrong.
  In other words, with $\geq$ 95% confidence a potential rejection is correct.

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

Testing a Hypothesis

Four steps of hypothesis testing

1. Hypothesis. State H and H_0.
2. Significance level. Choose α (always *before* the test).
3. Testing. Carry out an appropriate hypothesis test to get the p-value.
4. Decision. Depending on α and p, reject H_0 or fail to reject it.

https://xkcd.com/892/

Hypothesis Testing

What Test to Choose

Hypothesis tests

- A significance test needs to be chosen that fits the data.
- Different tests exist that make different assumptions about the data

  More on assumptions on the next slide

Parametric vs. non-parametric tests

- Parametric. More powerful and precise, i.e., it is more likely to detect a significant effect when one truly exists
- Non-parametric. Fewer assumptions and, thus, more often applicable
- Each parametric test has a non-parametric correspondent.

Parametric test	Non-parametric correspondent
One-sample and dependent t-test	Wilcoxon Signed-Rank Test
Independent t-test	Mann-Whitney Test
One way, between group ANOVA	Kruskal-Wallis
One way, repeated measures ANOVA	Friedman Test
...	...

Hypothesis Testing

Assumptions

Assumptions of all hypothesis tests

- Sampling. The sample is a random sample from the distribution.
  Notice: In NLP, each "instance" of a sample usually consists of multiple texts.
- Values. The values within each variable are independent.

Assumption of all parametric tests

- Scale. The dependent variable has an interval or ratio scale.
- Distribution. The given distributions are normally distributed.
  Tested by checking histograms or by using normality tests, e.g., the Shapiro-Wilk test.
- Variance. Distributions that are compared have similar variances.
  Tested using Levene's Test, Bartlett's test, or scatterplots and Box's M.

Test-specific assumptions

- In addition, specific tests may have specific assumptions.
- Depending on which are met, an appropriate test is chosen.

The Student's t-Test

Student's t-test

- A parametric hypothesis test for small samples ($\sim n \leq 30$)
- Computes a t-*score* from which significance can be derived
- Types. One-sample t-test, dependent t-test, independent t-test

  The term *student* was simply used as a pseudonym by the inventor.

Test-specific assumptions

- The independent variable has a nominal scale.
- t-tests are robust over moderate violations of the normality assumption.

One-tailed vs. two-tailed

- One-tailed. Test a directional hypothesis
- Two-tailed. Tests a non-directional hypothesis

One sample vs. paired samples

- One sample. A sample mean is compared to a known value.
- Paired samples. Two sample means are compared to each other.

The Student's t-Test

t-Score

t-distribution

- Variation of the normal distribution for small sample sizes
- Dependent on the *degrees of freedom (DoF)* in an experiment

  Put simply, DoF is the number of potential variations in the computation of a value.

- Statistics libraries (e.g., in Python) can compute t-distributions.
- Otherwise, tables exist with the significance confidences of t-values.

  https://en.wikipedia.org/wiki/Student%27s_t-distribution

	95%	**97.5%**	**99%**	**99.5%**	**99.9%**	**99.95%**	*One-tailed*
DoF	**90%**	**95%**	**98%**	**99%**	**99.8%**	**99.9%**	*Two-tailed*
3	2.353	3.182	4.541	5.841	10.21	12.92	
4	2.132	2.776	3.747	4.604	7.173	8.610	
...	...	...	...	...	...	...	

How to use this table?

- Compare t-score with value at given DoF and α ($=1-$ confidence).
- If t-score > value, then H_0 can be rejected; otherwise not.

The Student's t-Test

One-Sample t-Test

One-sample t-test

- Compares the mean M of a sample $\tilde{X}$ of size n from a distribution X to a known distribution mean μ
- $n-1$ degrees of freedom (since the n-th value is implied by M)

Example research question

- "Does our essay grader improve over the best result reported so far?"

> H_0. "The RMSE of our approach is *not* statistically significantly lower than the RMSE reported by Persing et al. (2015)."

Process

1. Compute the mean M of all sample values $\tilde{X}$.
2. Compute the variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{X}_i - M)^2$
3. Compute the standard deviation of the distribution of means: $s_M = \sqrt{\frac{s^2}{n}}$

   Also called *standard error*; division by n normalizes into the t-distribution
4. Compute the t-score: $t = \frac{M - \mu}{s_M}$

The Student's t-Test

Example: One-Tailed One-Sample t-Test

"The essay grading approach achieves a lower RMSE than 0.244"

1. State hypotheses and define significance level.

   H: RMSE $- 0.244 < 0$ $\quad$ H_0: RMSE $- 0.244 \geq 0$ $\quad$ $\alpha = 0.05$

2. Given a sample (say, $n = 5$), compute RMSE values.

   $\tilde{X} = (0.226, 0.213, 0.200, 0.268, 0.225)$

3. Compute sample mean, variance, and standard error.

   $M = \frac{1}{5} \cdot (0.226 + 0.213 + 0.200 + 0.268 + 0.225) = 0.226$

   $s^2 = \frac{(0.226-0.226)^2+(0.213-0.226)^2+(0.200-0.226)^2+(0.268-0.226)^2+(0.225-0.226)^2}{4} = 0.00065$

   $s_M = \sqrt{\frac{0.00065}{5}} = 0.0114$

4. Compute t-score and make decision.

   $t = \frac{0.244-0.226}{0.0114} = 1.579$ $\quad$ 4 DoFs $\quad$ critical t-value from table is 2.132.

   $\rightarrow$ $\quad$ $1.579 < 2.132$, so H_0 cannot be rejected.

The Student's t-Test

Dependent t-Test

Dependent t-test (aka paired-sample test)

- Compares two samples $\tilde{X}, \tilde{X}'$ of size n from the same distribution X, taken at different *times* (i.e., they may have changed in between)
- $n - 1$ degrees of freedom

Example research question

- "Does adding POS tags improve our sentiment analysis approach?"

> H_0. "The accuracy of our approach is not statistically significantly higher with POS tags than without POS tags."

Process

1. Compute each difference $\Delta_i = \tilde{X}_i - \tilde{X}'_i$ between the paired samples.
2. Compute the mean M of all differences Δ.
3. Compute the variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\Delta_i - M)^2$
4. Compute the standard error: $s_M = \sqrt{\frac{s^2}{n}}$
5. Compute the t-score: $t = \frac{M-0}{S_M} = \frac{M}{S_M}$

The Student's t-Test

Independent t-Test

Independent t-test

- Compares two independent samples $\tilde{X}, \tilde{X}'$ of size n from the same distribution X
- $2 \cdot (n-1) = 2n-2$ degrees of freedom

Example research question

- "Are the predicted essay grades different from the gold standard?"

> H_0. "There is no statistically significant difference between the gold standard scores and the scores predicted by the approach."

Process

1. Compute the means M, M' of all sample values of $\tilde{X}, \tilde{X}'$.
2. Compute the variances: $s_1^2 = \sum_{i=1}^{n} \frac{(\tilde{X}_i - M)^2}{n-1}, \quad s_2^2 = \sum_{i=1}^{n} \frac{(\tilde{X}_i' - M')^2}{n-1}$
3. Compute the standard error: $S_M = \sqrt{\frac{s_1^2 + s_2^2}{2}} \cdot \sqrt{\frac{2}{n}}$
4. Compute the t-score: $t = \frac{M - M'}{S_M}$

Hypothesis Testing

Alternatives

What if the t-test assumptions are not met?

- Test-specific assumption. Find other parametric test that is applicable.
- Assumptions of parametric tests. Find applicable non-parametric test.

  A common case is that the given values are not normally distributed.
- Assumptions of all significance tests. Hypotheses cannot be tested.

Example: Wilcoxon Signed-Rank Test

- Non-parametric alternative to dependent t-test, for small sample sizes
- Requires randomly chosen, independent paired samples, dependent variable with interval or ratio scale
- Does not require a normal distribution
- Computes a z-score based on a ranking of the differences of the pairs

  The value can also be checked against a reference table.

Conclusion

Conclusion

Empirical methods

- NLP uses empirical methods for linguistic tasks.
- An annotated text corpus represents the data of a task.
- Approaches are developed and evaluated on corpora.

https://pixabay.com

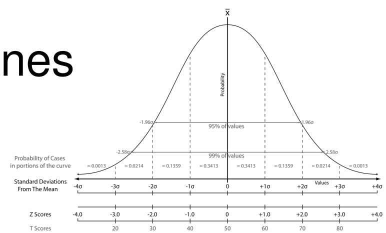
Evaluation measures

- NLP is usually evaluated for its effectiveness.
- Measures: Accuracy, F_1-score, mean squared error, ...
- Effectiveness is measured in experiments on datasets.

https://okpolicy.org

Comparison

- Need to compare approaches to reasonable baselines
- Descriptive and inferential statistics play a role.
- Hypothesis tests check whether a result is better.



https://commons.wikimedia.org

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.

- **Al Khatib et al. (2017).** Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of Argumentation Strategies across Topics. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1362–1368, 2017.