Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Hypothesis Testing
- Text Corpora
- Data Acquisition
- Data Annotation

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research questions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research questions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

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

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research questions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

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

Descriptive versus inferential statistics:

- **Descriptive.** Procedures for summarizing and comprehending a sample or distribution of values. Used to describe phenomena.

  1 2 2 2 $\rightarrow$ mean M = 1.75

- **Inferential.** Procedures that help draw conclusions based on values. Used to generalize inferences beyond a given sample.

  The average number is significantly greater than 1.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

> *"What is the effectiveness of parent education when given problem children?"*

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

> *"What is the effectiveness of parent education when given problem children?"*

Example of a well-formulated question:

> *"What is the effect of the STEP parenting program on the ability of parents to use natural, logical consequences (as opposed to punishment) with their child who has been diagnosed with bipolar disorder?"*

Empirical Research

Empirical Research in NLP

- Corpus linguistics.
  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.
- Evaluation measures.
  An evaluation of the quality of an approach is important, especially of its effectiveness.
- Experiments.
  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.
- Hypothesis testing.
  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Empirical Research

Empirical Research in NLP

- Corpus linguistics.
  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.

- Evaluation measures.
  An evaluation of the quality of an approach is important, especially of its effectiveness.

- Experiments.
  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.

- Hypothesis testing.
  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Empirical Research

Evaluation Measures

- An evaluation measure quantifies the quality of an approach on a specific task and text corpus.
- Approaches 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.
- Measures: accuracy, precision, recall, F_1-score, ... (later in this unit).
- High effectiveness is the primary goal of any NLP approach.

Efficiency

- The costs of an approach in terms of the consumption of time or space.
- Measures: overall runtime, training time, memory consumption, ...

  Efficiency is not the scope of this course.

Empirical Research

Effectiveness

- The effectiveness of an NLP approach is the extent to which the output information of the approach is correct.

Evaluation of classification effectiveness

- All NLP tasks, where instances of some output information type C are to be inferred, can be evaluated as a binary classification task.
- Check for each possible candidate instance of C whether the decision of an approach to infer the instance matches the ground truth.

Evaluation of regression effectiveness

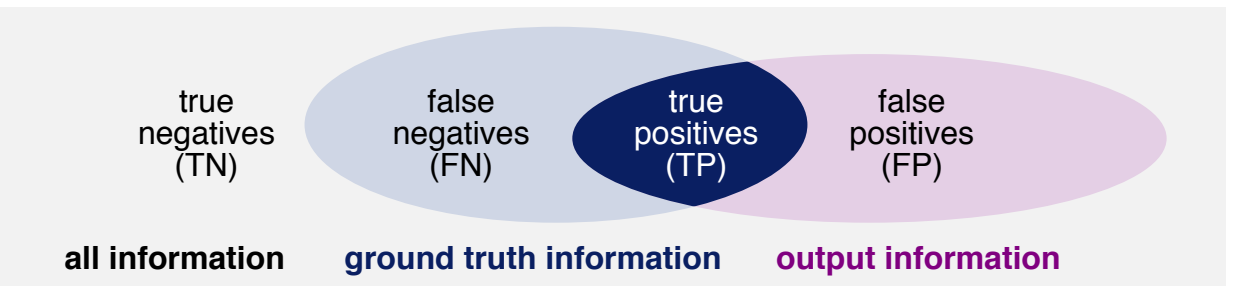
- In tasks where numeric values have to be predicted, the regression error is usually evaluated.
- Check for each value predicted for an instance by an approach how different the value is from the ground-truth value of the instance.

Empirical Research

Classification Effectiveness: Instance Types

Instance types of an NLP approach on a task

- Positives. The output information instances the approach has inferred.
- Negatives. All other possible instances.



Instance types in the evaluation of the task

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

Empirical Research

Classification Effectiveness: Evaluation based on the Instance Types

Example: Sentiment analysis

- Assume the sentiment of comments to videos is labeled as “positive”, “negative”, or “neutral”.
  Don’t confuse these labels with the instance types above!

Which of the following approaches is better?

- Approach 1. Classifies the first 70 of 100 comments correctly.
- Approach 2. Classifies the last 80 of the same 100 comments correctly.

Which dataset appears to be “easier”?

- Dataset 1. 800 out of 900 comments classified correctly.
- Dataset 2. 500 out of 600 comments classified correctly.

True vs. false instances

- A straightforward way to answer these questions is to compare the proportion of true instances under all instances.

Empirical Research

Classification Effectiveness: Accuracy

- The accuracy A is a measure of the correctness of an approach.
- How many classification decisions are correct?

$$A = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

When to use accuracy?

- Accuracy is adequate when all classes are of similar importance.
- This usually holds for text classification tasks, such as sentiment analysis, part-of-speech tagging, and similar.

"The"/DT "man"/NN "sighed"/VBD "."/. "It"/PRP "'s"/VBZ "raining"/VBG ...

- Also, accuracy may make sense where virtually every span of a text needs to be annotated, such as in sentence splitting.

"The man sighed. _ It's raining cats and dogs, he felt."

Empirical Research

Classification Effectiveness: Limitations of Accuracy

Example: Spam detection

- Assume 5% of the mails that your mail server lets through are spam.
- What accuracy does a spam detector have that always predicts “no spam” for these mails?

When *not* to use accuracy?

- In tasks where the positive class is rare, high accuracy can be achieved by simply inferring no information.

  5% spam $\rightarrow$ 95% 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.

Empirical Research

Classification Effectiveness: Precision and Recall

Precision

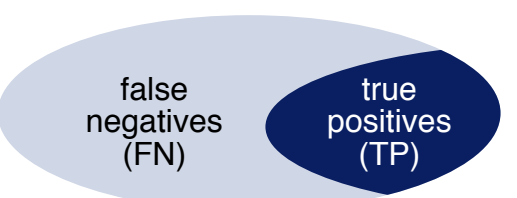
- The precision $prec$ is a measure of the exactness of an approach.
- Ratio of the found instances that are correct

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

Recall

- The recall rec is a measure of the completeness of an approach.
- Ratio of the correct instances that have been found

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



Observation

- True negatives are ignored by precision and recall

Empirical Research

Classification Effectiveness: Precision and Recall Implications

Example: Spam detection (revisited)

- Assume 5% of the mails that your mail server lets through are spam.
- What precision and recall does the “always no spam” detector have?

Idea of precision and recall

- Put the focus on a specific class (here: “spam”).
- The typical case is that the true negatives are not in the focus.
- If multiple classes are important, precision and recall can be computed for each class.

Example: Spam detection (a last time)

- Precision and recall of “always (no) spam” detectors on the spam class

no spam:	$prec$ = 1.0	no (wrong) prediction	rec = 0.0
spam:	$prec$ = 0.05		rec = 1.0

Empirical Research

Classification Effectiveness: Interplay between Precision and Recall

Perfect precision and recall

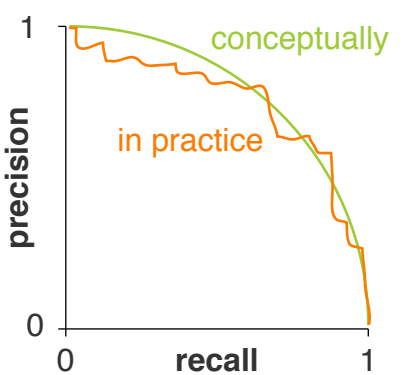
- A recall of 1.0 is mostly trivial; just assume every instance to be a TP.
  Only hard if there are too many instances, or if finding them is already a challenge.
- A precision of 1.0 is a definition problem: returning nothing might just do it.
  Otherwise, a precision of 1.0 is a bit more complicated: need to find at least one TP.

Precision vs. recall

- What is more important depends on the application.
- Usually, both precision and recall are to be maximized.

Trade-off between precision and recall

- The more true positives should be found, the more likely it is to choose also false instances.
- This leads to a typical precision-recall curve.
  Can be used to choose a method's operation point.



Empirical Research

Classification Effectiveness: F_1-Score

What is better?

- A precision of 0.51 and a recall of 0.51 (Option A).
- A precision of 0.07 and a recall of 0.95 (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 93% of the found instances are wrong with Option B.

F_1-score (aka F_1-measure)

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

$$F_1 = \frac{2 \cdot prec \cdot rec}{prec + rec}$$

Option A: $F_1 = 0.51$, Option B: $F_1 = 0.13$.

Empirical Research

Classification Effectiveness: F_1-Score Generalization

F_β-Score

- The 1 in the F_1-score in fact denotes a weighting factor.
- The general weighted harmonic mean is the F_β-score:

$$F_\beta = \frac{(1+\beta^2) \cdot prec \cdot rec}{(\beta^2 \cdot prec) + rec}$$

Problem with the weighting

- $\beta > 1$ gives more weight to precision, $\beta < 1$ gives more weight to recall.
- It is unclear how to interpret a particular choice of β.
- Therefore, nearly always $\beta = 1$ is used in practice.

Empirical Research

Classification Effectiveness: F_1-Score Issue in Tasks with Boundary Detection

Boundary errors

- A common error in tasks where text spans need to be identified and classified 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 lead to both an FP and an FN.
- Identifying nothing would increase the F_1-score compared to an almost match.

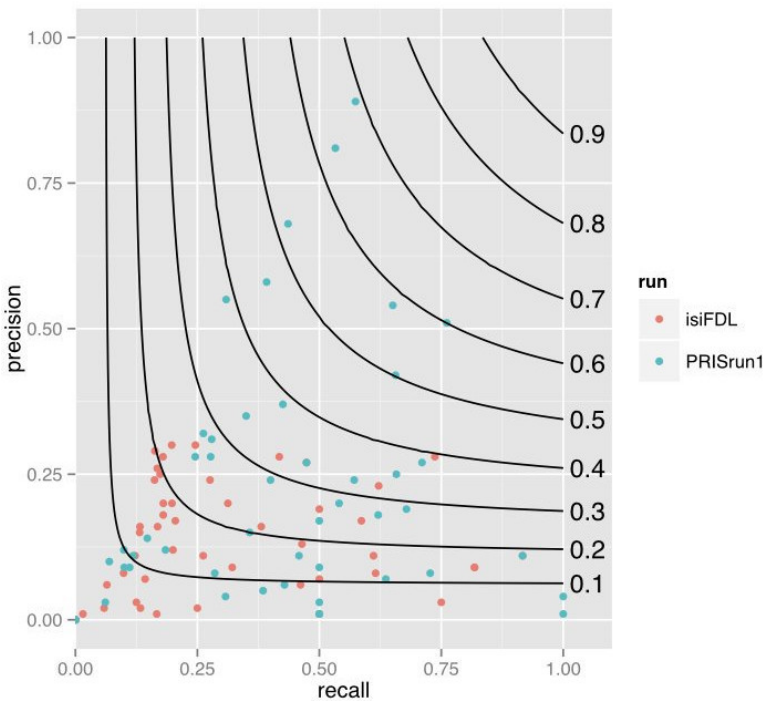
How to deal with boundary errors

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

Empirical Research

Classification Effectiveness: Other F_1-Score Issues

- Is $F_1 = 0.87$ or $F_1 = 0.89$ better?
  - Natural question: What are precision and recall?
  - Then why average them at all?
- F_1-score averages two things; averaging over lots of classes / observations yields some value that can hardly be interpreted anymore (but people still try, see next slides).
- Usually, precision drops when recall increases and vice versa.
- Instead of hiding that in an average, plotting the F "isocurves" (precision / recall values yielding the same F_1-scores) helps to differentiate



[Ian Soboroff via Twitter, https://twitter.com/ian_soboroff/status/1110903162171465728?lang=en]

Empirical Research

Classification Effectiveness: Micro- 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.
- Other results need to be combined with micro- or macro-averaging.

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

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

$$Micro{-}prec = \frac{|TP_1| + \ldots + |TP_k|}{|TP_1| + \ldots + |TP_k| \quad + \quad |FP_1| + \ldots + |FP_k|}$$

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

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

$$Macro{-}prec = \frac{prec_1 + \ldots + prec_k}{k}$$

Empirical Research

Classification Effectiveness: Confusion Matrix for Micro- and Macro-Averaging

- Each row refers to the ground-truth instances of one of k classes.
- Each column refers to the classified instances of one of the classes.
- The cell values illustrate the 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	$\|FP_a \cap FN_k\|$	$\|FP_b \cap FN_k\|$	...	$\|TP_k\|$

Confusion matrices for what?

- Used to analyze errors, to see which classes are confused with which.
- Also shows the basis of computing micro- and macro-averaged results.

Empirical Research

Classification Effectiveness: Computing Micro- and Macro-Averages

Example: Evidence classification

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

Confusion matrix of the results

Ground-truth	Classified as				Total		Precision
	Anecdote	Statistics	Testimony	None	TP	FP	per class
Anecdote	199	5	35	183	199	165	0.55
Statistics	17	29	0	27	29	13	0.69
Testimony	30	1	123	71	123	71	0.63
None	118	7	36	1455	1455	281	0.84

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

- $Micro{-}prec = \frac{199+29+123+1455}{199+29+123+1455+165+13+71+281} = 0.77$
- $Macro{-}prec = \frac{0.55+0.69+0.63+0.84}{4} = 0.68$

Empirical Research

Regression Effectiveness

Regression task

- A regression task requires to predict numeric values for instances from some usually but not necessarily predefined scale.
- In NLP, typical regression tasks are automatic essay grading, review rating prediction, etc.

Example: Automatic essay grading

- Given a set of n student essays, automatically assign each essay i a score $y_i \in \{1, \ldots, 4\}$.

  The 4-point scale is the default in today's grading systems.

Regression errors

- In many regression tasks, it is unlikely to perfectly predict the value of instances, which is why accuracy is often not the primary measure.
- The focus is rather on the mean regression error of the predicted values $Y = (y_1, \ldots, y_n)$ compared to the ground-truth values $\hat{Y} = (\hat{y_1}, \ldots, \hat{y_n})$.

Empirical Research

Regression Effectiveness: Types of Regression Errors

Mean absolute error (MAE)

- The MAE is the mean difference of predicted to ground-truth values.
- It is robust to outliers.

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

Mean squared error (MSE)

- The MSE is the mean squared difference of predicted to ground-truth values.
- It 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 $RMSE = \sqrt{MSE}$.

Empirical Research

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.

	Essay										Regression 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	1	2	2	3	3	3	3	4	4	0.00	0.00

Which approach is best?

- Approach 1 trivially always predicts the mean → 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.

Empirical Research

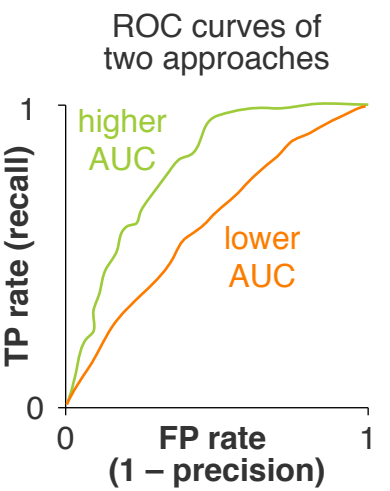
Other Measures

Notice

- Accuracy, precision, recall, F_1-score, and mean absolute/squared error are standard effectiveness measures.
- They are still sometimes not suited for specific settings in which other measures are more useful (cf. the F_1-score issues).

Selection of other measures

- Error rate. Simply 1 – accuracy.
- Labeled attachment score. Proportion of fully correctly classified tokens in syntactic parsing.
- Precision@k. Precision within the top k results of a ranking problem (also recall@k is used where it makes sense).
- Area under curve (AUC). Expected proportion of positives ranked before a negative, based on the receiver-operating characteristic (ROC) curve.



Empirical Research

Experiments

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

Intrinsic vs. extrinsic effectiveness evaluation

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

  "What accuracy has a part-of-speech tagger X on the dataset Y?"

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

  "Does the ouput of a part-of-speech tagger X improve sentiment analysis on Y?"

Corpus-based experiments vs. user studies

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

  Not covered in this course.

Empirical Research

Datasets

- A dataset is 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 then applied to a set of unseen test instances.
3. The output information of the approach is compared to the ground-truth instances in terms of certain evaluation measures.
4. Steps 1–3 may be iteratively repeated to further improve the approach.

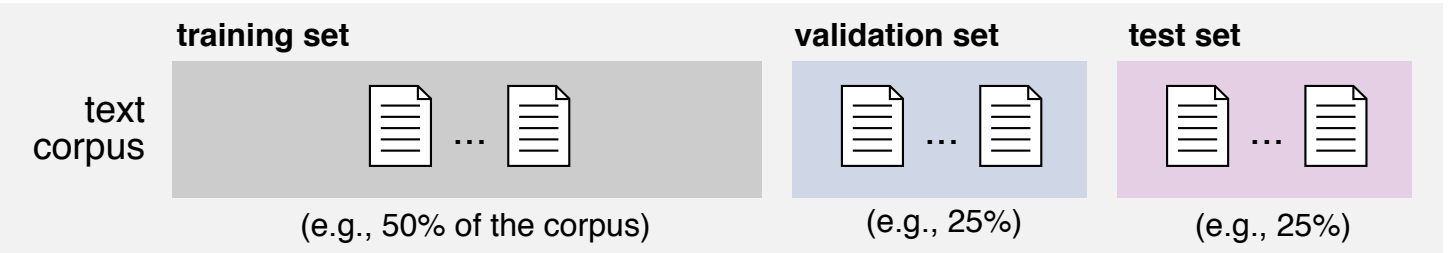
Types of evaluation

- The way a corpus is split implies how to use the datasets.
- Main alternatives. Training, validation, and test set vs. cross-validation.
- The splitting may have certain constraints, but this is out-of-scope here.

  Example: No overlap of instances from the same text in different datasets.

Empirical Research

Types of Evaluation: Training, Validation, and Test Set



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 towards 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 Research

Types of Evaluation: Cross-Validation

(Stratified) n-fold cross-validation

- Randomly split a corpus into n datasets of equal size, usually $n = 10$.
- The development and evaluation consist of n runs. The evaluation results are averaged over all n runs.
- In the i-th run, the i-th fold is used for evaluation (testing). 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.

Empirical Research

Types of Evaluation: Variations

Repeated cross-validation

- Often, cross-validation is repeated multiple times with different folds.
- This way, coincidental effects of the 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.

Empirical Research

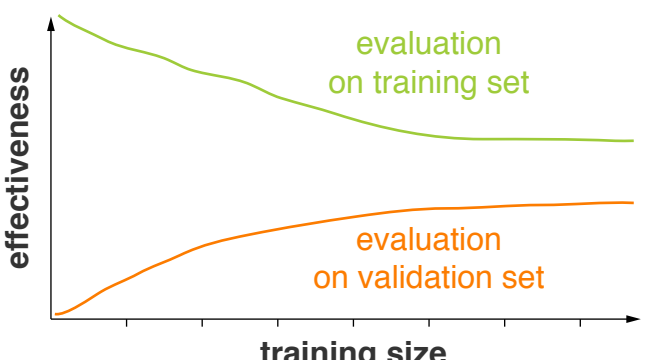
Training Data

How much training data is needed?

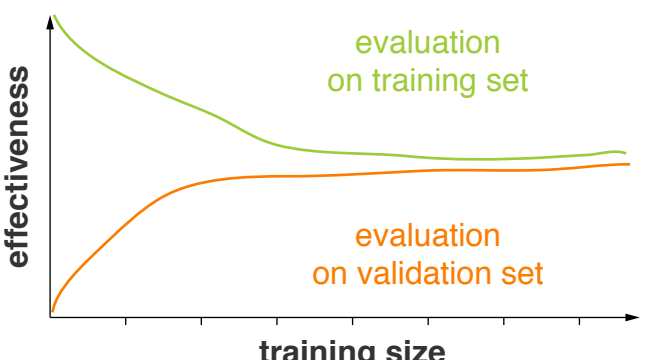
- In general, hard to say.
- Depends on the task, the heterogeneity of the data, . . .

One way to find out

- Test different training sizes.
- Evaluate effectiveness on training set and on validation set.



(more data is likely to help)



(more data likely won't help)

- Validation effectiveness is unlikely to ever exceed training effectiveness.

Empirical Research

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 and distribution of classes.
- The class distribution in the test set.
- The heterogeneity of the test set.
- The similarity between training 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.

Empirical Research

Comparison: Upper and Lower Bounds

Why comparing?

- A new approach is useful if it is better than previous approaches (usually measured in terms of effectiveness) or if it implements a completely new idea and is not too much worse than previous best approaches.
- Usually, approaches are 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.
- For simplicity, the gold standard is often equated with the ground-truth annotations in a corpus, so 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 aims to be better than all reasonable baselines.

Empirical Research

Comparison: Types of Baselines

Trivial baselines

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

Standard baselines

- Approaches 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 approaches for the addressed task (if available).
- Used to verify whether a new approach is best.

Empirical Research

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?

Exemplary dataset and task parameters [Al-Khatib et al., 2016]

- Four classes. “anecdote”, “statistics”, “testimony”, “none” (majority)
- Test distribution. 18% 3% 10% 69%

Potential baselines

- Trivial. Uniformly random guessing achieves an accuracy of 25%.
- Trivial. Always predicting the majority achieves 69%.
- Standard. Using word {1, 2, 3}-grams frequencies achieves 76%.
- State of the art. The best published result is 78%. [Al-Khatib et al., 2017]

Empirical Research

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 reasonably be compared.

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?

How to know that some effectiveness is better?

- Effectiveness differences may be coincidence.
- The significance of observed differences can be tested statistically.

Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Hypothesis Testing
- Text Corpora
- Data Acquisition
- Data Annotation

Hypothesis Testing

Statistics

Variable

- An entity that can take on different numeric (quantitative) or nonnumeric (qualitative) values.
- Independent. A variable X that is expected to affect another variable.
- Dependent. A variable Y that is expected to be effected 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 / categorical. Values that represent discrete, separate categories.
- Ordinal. Values that can be ordered / ranked by what is better.
- Interval. Values whose difference can be measured.
- Ratio. Interval values that have an absolute zero.

Interval vs. ratio scale test

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

Hypothesis Testing

Statistics: Variables and Scales

What is independent, what is dependent?

> *Do at-risk high school seniors who participate in a study skills program have a higher graduation rate than at-risk high school seniors who do not participate in a study skills program?*

Independent: participation in study skills program

Dependent: graduation rate

What type of scale?

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

1. Interval (Ratio in Kelvin) 2. Ordinal 3. Ratio 4. Nominal 5. Ratio

Hypothesis Testing

Descriptive Statistics

- Measures for summarizing and comprehending distributions of values.
- Used to describe phenomena.

Measures of central tendency

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

  M is used for a sample of values, μ for a population.

  $\tilde{X}$ = (69, 77, 77, 77, 84, 85, 85, 87, 92, 98) $\quad\rightarrow\quad$ $M = \frac{1}{10}\sum_{i=1}^{10}\tilde{X}_i$ = 83.1

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

  For an even number of values, the value halfway between the two middle values (or sometimes the one left / right from the middle).

  $\tilde{X}$ = (69, 77, 77, 77, 84, 85, 85, 87, 92, 98) $\quad\rightarrow\quad$ $median$ = 84.5

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

  $\tilde{X}$ = (69, 77, 77, 77, 84, 85, 85, 87, 92, 98) $\quad\rightarrow\quad$ $mode$ = 77

Hypothesis Testing

Descriptive Statistics: Central Tendency and its Disperson

When to use what tendency measure?

- Mean. For (rather) symmetrical distributions of interval / ratio values.
- Median. For ordinal values and skewed interval / ratio distributions.
- Mode. For nominal values.

Measures of dispersion

- Range. The distance r between minimum and maximum.

  $\tilde{X}$ = (69, 77, 77, 77, 84, 85, 85, 87, 92, 98) $\quad\rightarrow\quad$ $r = \tilde{X}_{max} - \tilde{X}_{min}$ = 29

- Variance. The arithmetic mean s^2 of the squared differences between each value and the mean.

  s is used for a sample of values, σ for a population.

  $\tilde{X}$ = (69, 77, 77, 77, 84, 85, 85, 87, 92, 98) $\quad\rightarrow\quad$ $s^2 = \frac{1}{10}\sum_{i=1}^{10}(\tilde{X}_i - M)^2$ = 70.54

- Standard deviation. The square root s of the variance.

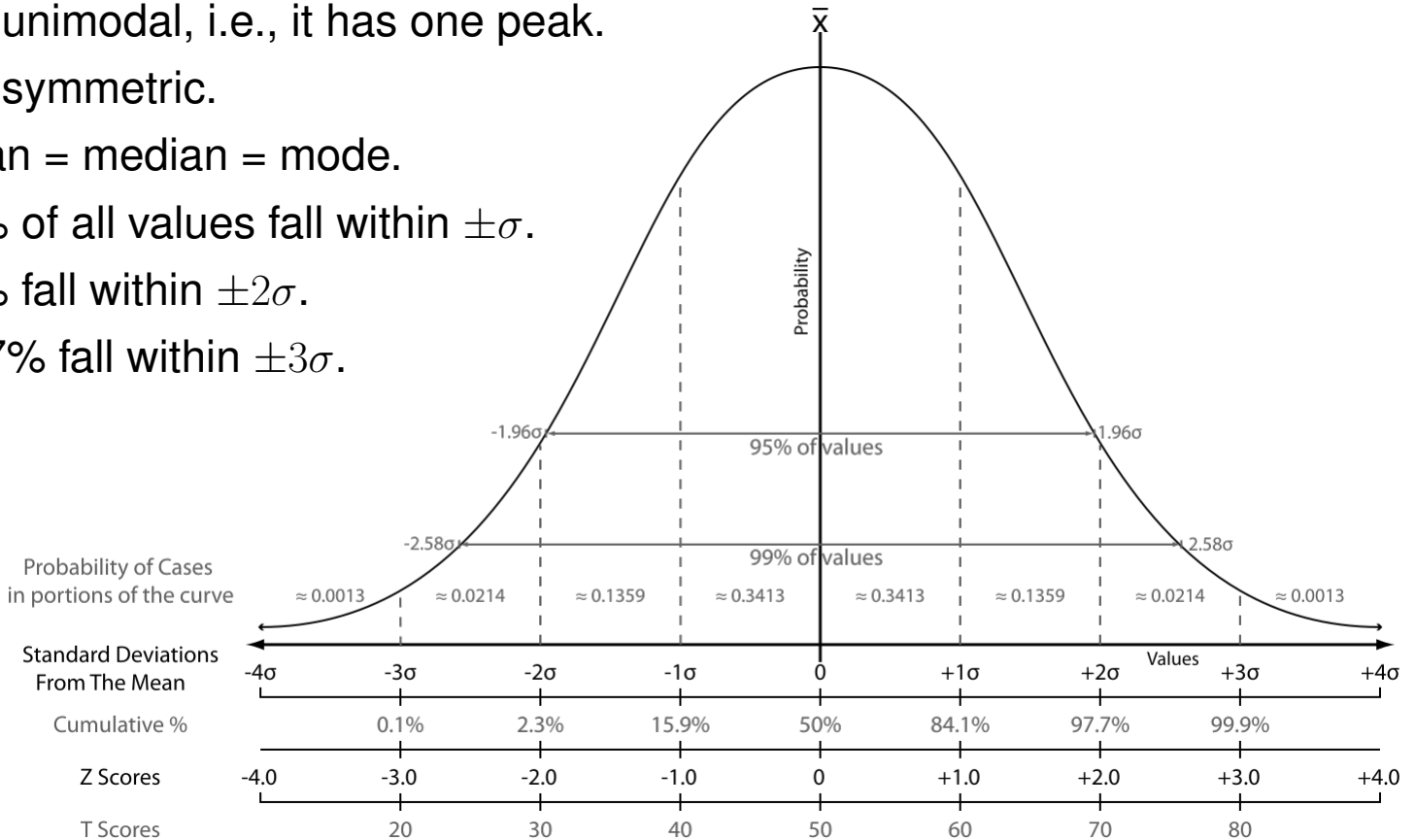
  69, 77, 77, 77, 84, 85, 85, 87, 92, 98 $\quad\rightarrow\quad$ $s = \sqrt{s^2}$ = 8.39

Hypothesis Testing

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 within $\pm\sigma$.
- 95% fall within $\pm 2\sigma$.
- 99.7% fall within $\pm 3\sigma$.



Hypothesis Testing

Descriptive Statistics: Standard Scores

Standard score

- Indicates how many standard deviations a value is 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.

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

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.

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

Hypothesis Testing

Inferential Statistics

- Procedures that help draw conclusions based on values.
- Used to make inferences about a population beyond a given sample.

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 due to sampling error).
- The likelihood is specified in terms of a probability, called the p-value.
- Significance is given if p is $\leq$ a specified significance level α.

  Usually, $\alpha = 0.05$ or $\alpha = 0.01$. But a lot of discussions around "p-hacking".

What is a good hypothesis? (Bartos, 1992)

- Is founded in a problem statement and supported by research.
- Is testable (i.e., it is possible to collect data to answer the question).
- States an expected relationship between variables.
- Is stated as simply and concisely as possible.

Hypothesis Testing

Inferential Statistics: Hypotheses

Two competing hypotheses

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

  "There will be a statistically significant difference in dropout rates of students who use drugs and students who do not use drugs."

- Null hypothesis (H_0). Antithesis to H.

  Results supporting H, if any, are due to chance or sampling error.

  "There will be no statistically significant difference in dropout rates of students who use drugs and students who do not use drugs."

Two types of hypotheses

- Directional. Specify the direction of an expected difference. Indicates that a *one-tailed test* needs to be conducted.
- Non-directional. Specifies only that a difference is expected, without a direction. Indicates that a *two-tailed test* needs to be conducted.

Hypothesis Testing

Four Steps of Hypothesis Testing

1. State the null hypothesis H_0 and the research hypothesis H.
2. Choose a statistical significance level α.
3. Choose and carry out an appropriate statistical test to get the p-value.
4. Make a decision regarding H_0 and H (i.e., reject or fail to reject H_0).

Statistical significance level

- Always chosen before the test.
- Quantifies the acceptable risk that H_0 is wrongly rejected.
- A choice of α = 0.05 means that one is willing to accept 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

- 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

Effect Size

Statistical significance vs. effect size

- Significance does not state how large a difference is.
- The effect size describes the magnitude of the difference.

Effect size measure Cohen's d

- The effect size is usually computed based on the standard deviations:

$$d = \frac{M_1 - M_2}{s}$$

- Small effect: $d \geq 0.2$, medium effect: $d \geq 0.5$, large effect: $d \geq 0.8$.

Notice

- The focus is largely on significance in NLP (and in this course) but effect sizes and confidence intervals should also be reported.

Hypothesis Testing

What Test to Choose

Parametric vs. non-parametric tests

- A significance test needs to be chosen that fits the data.
- Parametric tests are more powerful and precise (i.e., it is more likely to detect a significant effect when one truly exists).
- Non-parametric tests make fewer assumptions in order to be applicable.

Overview of common significance tests

Parametric test	**Non-parametric correspondent**
Independent t-test	Mann-Whitney Test
Dependent and one-sample t-test	Wilcoxon
One way, between group ANOVA	Kruskal-Wallis
One way, repeated measures ANOVA	Friedman Test
Factorial ANOVA	–
MANOVA	–
Pearson	Spearman, Kendall's τ, χ^2
Bivariate regression	–

Hypothesis Testing

Assumptions

Assumptions of all significance tests

- Sampling. A sample needs to be a random sample from the population.
- Values. The values within each variable must be independent.

Assumption of all parametric tests

- Dependent variable. Needs to have an interval or ratio scale.
- Distribution. The population distribution needs to be normal.

  Evaluated either by investigating histograms or by using normality tests, such as the Shapiro-Wilk test (< 50 values) or the Kolmogorov-Smirnov test (> 50).
- Variance. The distributions need to have the same variances.

  Evaluated 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.

Hypothesis Testing

The Student's t-Test

What is the student's t-test?

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

- A parametric statistical significance test for small samples (say, $n \leq 30$).
- Types. Independent t-test, one-sample t-test, dependent t-test.

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 whether one value is higher or lower than another one.
- Two-tailed. Test whether two values are different from each other.

One sample vs. paired samples

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

Hypothesis Testing

One-Sample t-Test

- Compares the mean M of a sample $\tilde{X}$ of size n from a distribution X to a known population mean μ .
- $n - 1$ degrees of freedom.

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

- Compute the mean M of all sample values $\tilde{X}$.
- Compute the estimated population variance: $s^2 = \sum_{i=1}^{n} \frac{(\tilde{X}_i - M)^2}{n-1}$

  Division by $n - 1$ corrects for the small sample size.
- 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".
- Compute the t-score: $t = \frac{M - \mu}{s_M}$

Hypothesis Testing

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 usually 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

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

Hypothesis Testing

Independent t-Test

- Compares two independent samples $\tilde{X}, \tilde{X}'$ of size n from the same distribution X.
- $2n - 2$ 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

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

Hypothesis Testing

The Student's t-Test: What to do with the t-Score?

t-distribution

- Distribution based on the normal distribution for small sample sizes that captures the location of the sample mean relative to the true mean.
- Dependent on the degrees of freedom (DoF).
- Statistics tools, such as *R*, 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 the table

- Compare t-score with value at given DoF and α ($= 1 -$ confidence).
- If t-score > value, then H_0 can be rejected (result is significant).
- Otherwise, H_0 cannot be rejected.

Hypothesis Testing

Example: One-Tailed One-Sample t-Test

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

1. Define hypotheses.

$$H\text{: RMSE} - 0.244 < 0 \qquad H_0\text{: RMSE} - 0.244 \geq 0$$

2. Set significance level $\alpha = 0.05$ and compute $n = 5$ RMSE values.

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

3. Compute mean of the sample values.

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

4. Compute estimated population variance and standard error.

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

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

5. Compute t-score and make decision.

4 degrees of freedom, critical t-value from table is 2.132.

$t = \frac{0.226-0.244}{0.0114} = -1.579 \quad \rightarrow \quad 1.579 < 2.132$, so H_0 cannot be rejected.