School of CIT –
Social Computing
Research Group

Technical
University
of Munich

Advanced Natural Language Processing
CIT4230002

Prof. Dr. Georg Groh
Tobias Eder, M.A. M.Sc.

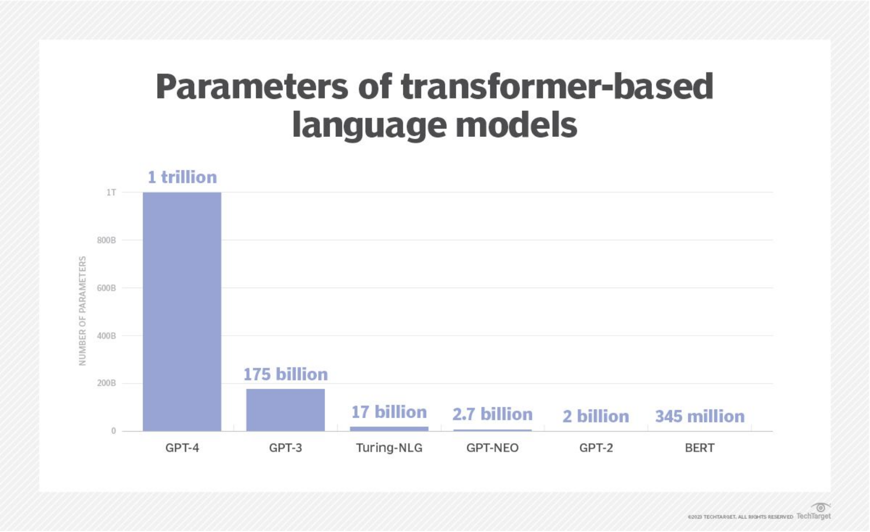
Data for NLP

Outline

- **Why NLP needs (a lot of) data**
- Datasets and data sources
- Unlabeled data and pre-training
- Annotation and labeling
- Augmentation and evaluation

NLP and Data

- NLP has become increasingly data hungry over the last 6 years



- Rising number of parameters in architectures
- Increasing amount of required training
- Broad training objectives

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

- What is the structure of language?

Saussure

Sign
The object / thing

Signifier
The physical existence
(sound, word, image)

Red / Leaf / Round / Apple

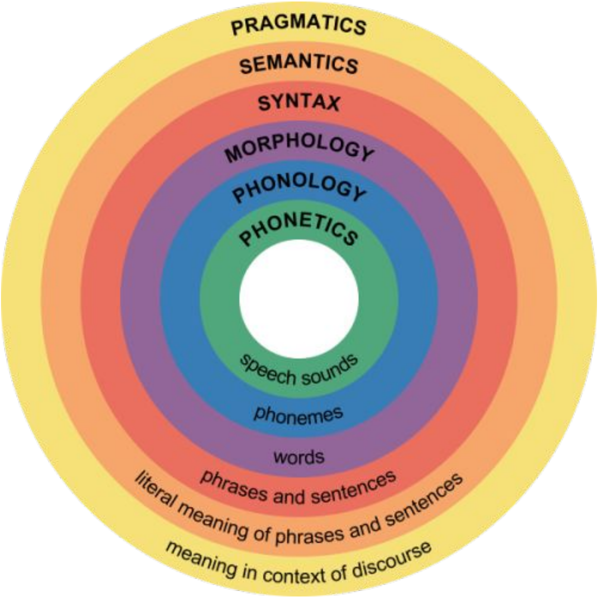
Signified
The mental concept

Fruit / Apple / Freshness / Healthy / Temptation / Teacher's pet / Computer

- Language = Systems of symbolic representation

NLP and Data | Semantics vs Pragmatics

- Additionally we are dealing with the layer of Pragmatics

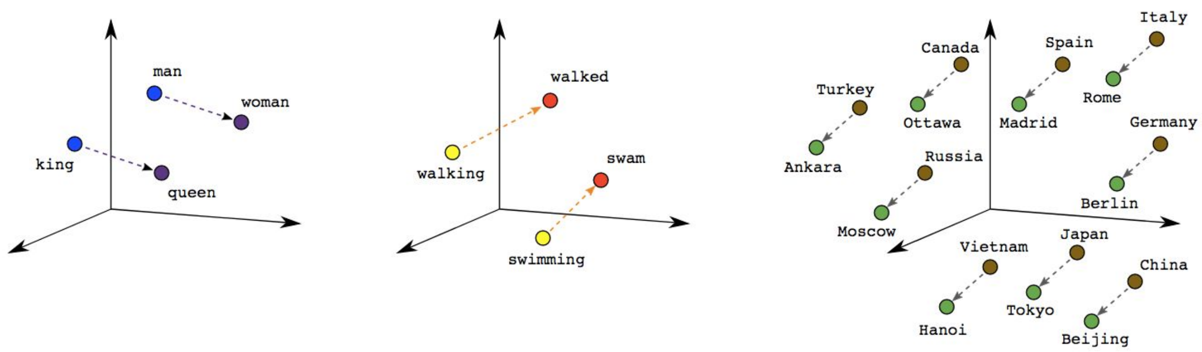


- What kind of information is left unsaid within a sentence?
- Meaning left up to specific contexts

- Capturing semantics:
  - Past: Formal logic and meaning representation languages

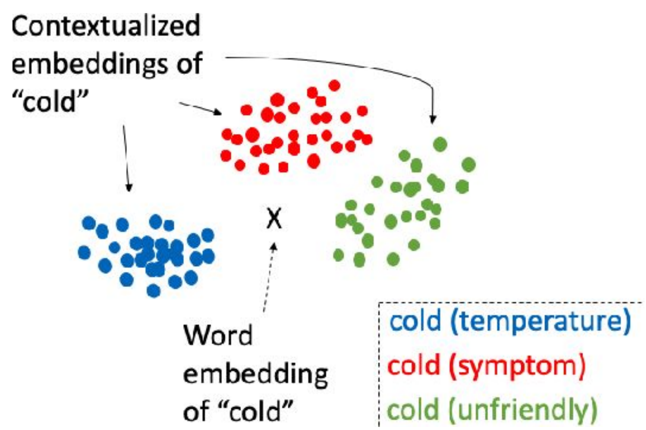
    - *A restaurant near CMU serves Indian food* ∃x Restaurant(x) ∧ Near(x, CMU) ∧ Serves(x, Indian)
    - *All expensive restaurants are far from campus* ∀x Restaurant(x) ∧ Expensive(x) ⇒ ¬Near(x, CMU)

  - Modern: Word representation through embeddings
  - Popular architectures 10 years ago: Word2Vec or GloVe



Male-Female Verb Tense Country-Capital

- But it gets more complicated…
- Words don't have the same meaning everywhere
  □ We need contextual embeddings (ELMo)



- Additionally we might want to look at the meaning of whole sentences
  □ Sentence embeddings (BERT)

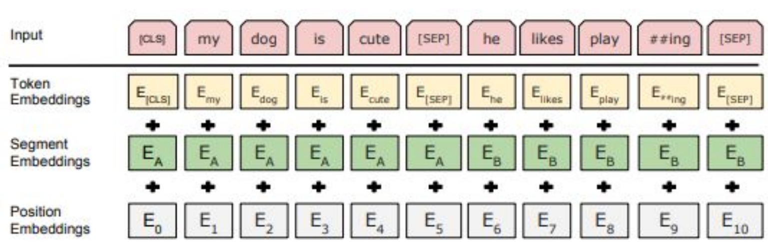


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

NLP and Data | Problems with Representation Learning

- When learning representations from text we need to see words in all sorts of contexts
- We additionally want these contexts to be relevant to our tasks
- LLMs are good because they offer high representational power but their output is not trivial to influence Prompting

- Language and its function is already complex, additionally there are problems such as:
  - Multilinguality
  - Pragmatics of language
  - Changing contexts

Outline

- Why NLP needs (a lot of) data
- **Datasets and data sources**
- Unlabeled data and pre-training
- Annotation and labeling
- Augmentation and evaluation

Datasets and data sources

- In NLP there are three types of data we mostly encounter:
  - Unlabeled text data
  - Naturally labeled data
  - Annotated data
- The amount of available data decreases significantly at each step
- When confronted with a NLP task first ask:
  - What kind of data might be beneficial to solve this task?
  - Are there ways with which we can leverage existing data for our task?

- If your task is building on previous work
  □ Safe to start with the same data

Datasets and data sources | Finding Existing Data

- There are plenty of resources for available NLP data

Datasets

- https://huggingface.co/docs/datasets/index

- http://www.elra.info/en/lrec/shared-lrs/

Papers With Code

- https://paperswithcode.com/area/natural-language-processing

Datasets and data sources | Unlabeled

- In the NLP context: Raw text data, mostly without added metadata
- Vast majority of the sources: Web
- What is it good for?
  - Autoregressive training
  - Masking tasks
  - Pre-training (more on this later)

- Example datasets:
  - Common Crawl https://commoncrawl.org/
  - Wikipedia Dump https://dumps.wikimedia.org/

Datasets and data sources | Naturally Labeled

- Some tasks might afford their own implicit labels
- This can occur in parallel data i.e. for machine translation or in co-occurring pieces of text such as television show scripts and an episode synopsis
- Another possibility is meta information that might be used as an indicator for a task, e.g. written reviews and a star rating for general sentiment prediction

- Naturally labeled data can be a huge boon for your tasks
- Think about ways in which the text you might want to use can already come with a label of sorts.
- Example datasets:
  - DCEP https://joint-research-centre.ec.europa.eu/language-technology-resources/dcep-digital-corpus-european-parliament_en

Datasets and data sources | Annotated

- This includes all data where an explicit intentional label was given
- Labels can vary in complexity and format and are usually very task-dependent
- Use-cases range form classification to complex tasks such as question answering and summarization
- Quality of existing annotated data is often hard to evaluate
- Many tasks might require to mix data sources

- Example datasets:
  - Large-Scale Hate Speech LREC https://github.com/avaapm/hatespeech
  - WCEP Summarization Data https://github.com/complementizer/wcep-mds-dataset

Datasets and data sources | Data collection

- If the data you need is not already released in dataset form?

  □ PANIC

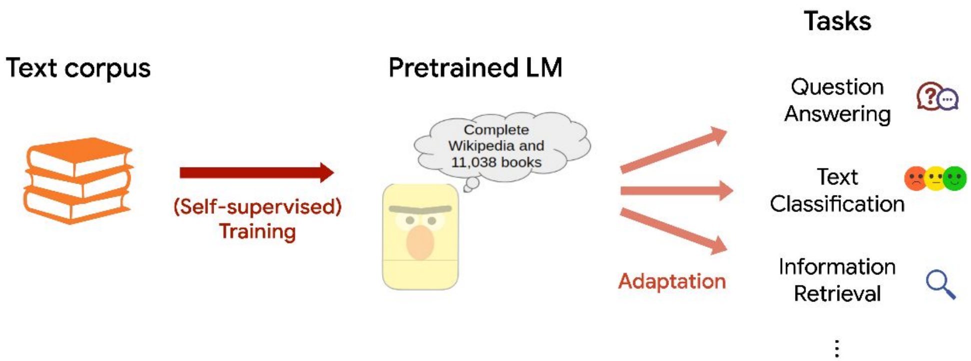
- Or: Build your own dataset
- The nice way: Get the data from some platform offering a permissive API
- If not..
  - Use web scraping tools such as..
  - Scrapy
  - Selenium

Outline

- Why NLP needs (a lot of) data
- Datasets and data sources
- **Unlabeled data and pre-training**
- Annotation and labeling
- Augmentation and evaluation

Unlabeled data and pre-training

- Recap: Pre-training is a general term describing a type of transfer learning where training is first done on an unrelated objective before later training the same architecture on the actual objective
- Other terminology related to pre-training:
  - Multitask learning – general term for training on multiple tasks
  - Transfer learning – applying learned knowledge on another task
  - Few-shot / Zero-shot learning – learning to perform a task with very few (zero) labeled examples

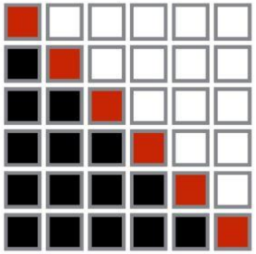


- Unlabeled data lends itself well to be used for pre-training objectives

Unlabeled data and pre-training

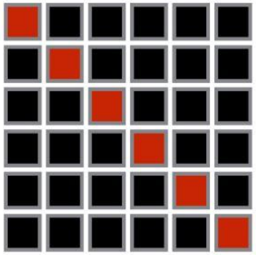
- Common pre-training strategies:
- Autoregressive language modeling
  - ☐ Used for prompting / text generation (e.g. GPT)

$$P(X) = \prod_{i=1}^{|X|} P(x_i | x_1, \ldots, x_{i-1})$$



- Masked language modeling
  - ☐ Common in pre-training + task fine-tuning (e.g. BERT)

$$P(X) \neq \prod_{i=1}^{|X|} P(x_i | x_{\neq i})$$



Unlabeled data and pre-training | Leveraging your data

- In general more training data is beneficial but not always needed
- Paradoxically larger models will learn faster than smaller models (Kaplan et al. 2020)
- However this means a smaller model can be improved further by letting it train longer

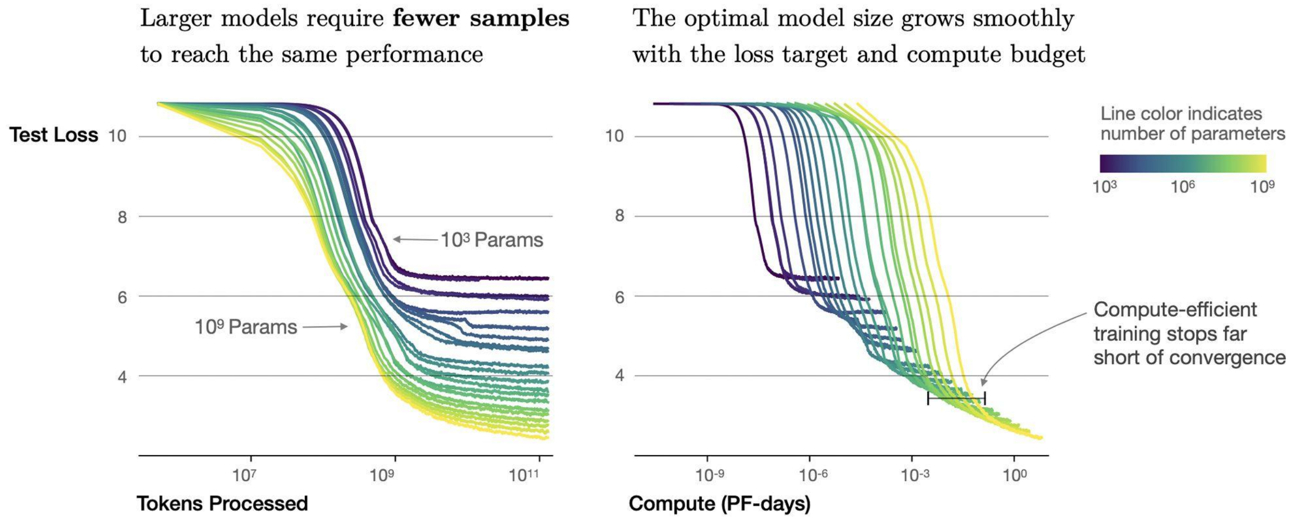
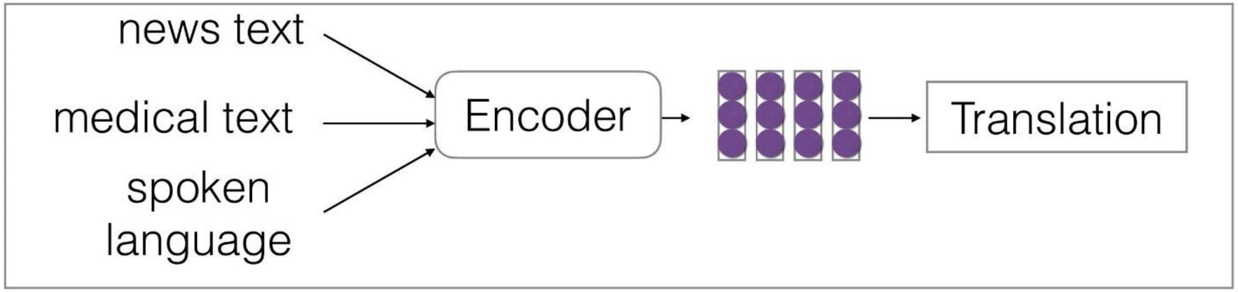


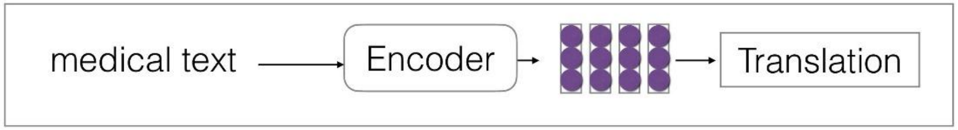
Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

Unlabeled data and pre-training | Domain Adaptation

- Training time: Same task but data from different distributions



- Test performance entirely on the low-resource domain

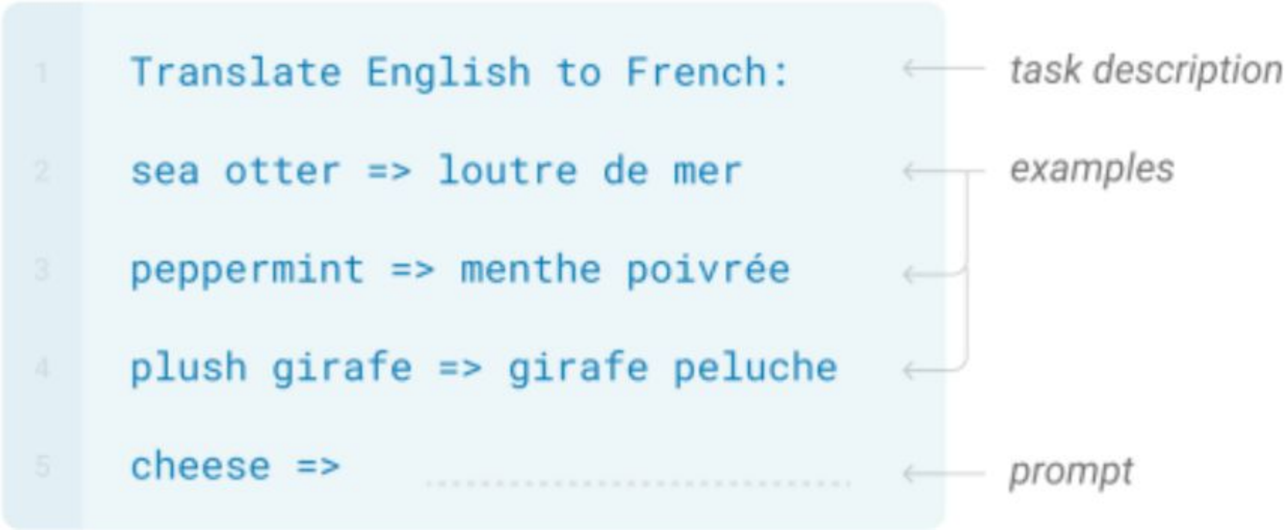


- Domains can be defined by:
  - Content
  - Style
  - Labels
- Domain shift might result in covariate or concept shift

Unlabeled data and pre-training | Prompting

- Recap: Prompting describes a form of fixed-parameter task execution
- The task is reformulated as a language modeling task
- It can be done zero-shot…

- Or few-shot style, depending on input length



Unlabeled data and pre-training | Pitfalls

- Training on unlabeled data usually means training on a lot of unseen data
- There might be problems with
  - Distribution or domain of the data
  - Reproducing unwanted biases
  - Poor data quality
  - Unverifiable and false information
- In some cases more is not always better
- Some tasks might benefit from pre-training on less, but better task-specific data

Outline

- Why NLP needs (a lot of) data
- Datasets and data sources
- Unlabeled data and pre-training
- **Annotation and labeling**
- Augmentation and evaluation

Annotation and labeling

- When do we need to annotate ourselves?
  - Insufficient labeled data
  - Unique task requiring specialized knowledge
  - Unifying labels of different datasets
  - Need to evaluate performance on an otherwise unsupervised task

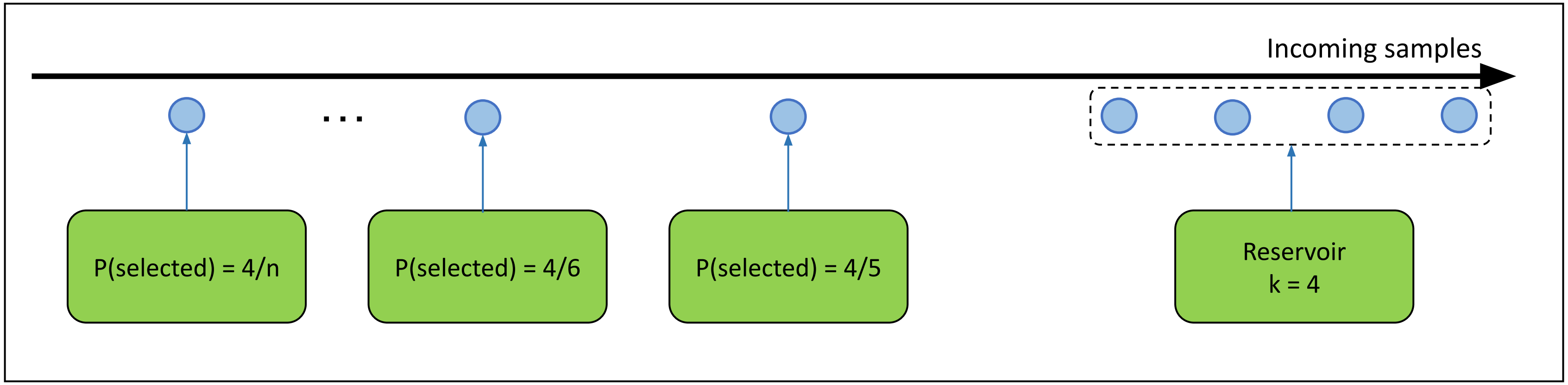
- Often the need to annotate becomes obvious after working on a project for a while. Be ready that there might be annotations needed even if you did not plan for it ahead of time

Annotation and labeling | Preparing to Annotate

- Before you start annotating…
  - Decide how much data you want to (can) annotate
  - Make sure to sample the examples to annotate from appropriate data
  - Create clear annotation guidelines (even if annotating by yourself)
  - Hire extra annotators
  - Evaluate the quality of the annotations
- How to sample data?
  - Cover appropriate domains
  - Cover language varieties and speaker demographics
  - Document the choices you made for sampling

Annotation and labeling| Sampling strategies

- Snowball sampling – For web scraping
- Stratified sampling – Drawing randomly from different subgroups (or stratas)
- Weighted sampling – For adjusting probabilities when leveraging domain knowledge
- Reservoir sampling – Randomly sampling from a stream of data



- Importance sampling – Sampling from a proposal distribution

Annotation and labeling | Guidelines and Hiring

- Loop: Try to annotate yourself, set up guidelines + iterate

What:

2 *LIST OF PARTS OF SPEECH WITH CORRESPONDING TAG* 2

Adverb—RB

This category includes most words that end in *-ly* as well as degree words like *quite*, *too* and *very*, posthead modifiers like *enough* and *indeed* (as in *good enough*, *very well indeed*), and negative markers like *not*, *n't* and *never*.

Adverb, comparative—RBR

Adverbs with the comparative ending *-er* but without a strictly comparative meaning, like *later* in *We can always come by later*, should simply be tagged as RB.

Adverb, superlative—RBS

Difficult Cases:

4 Confusing parts of speech

This section discusses parts of speech that are easily confused and gives guidelines on how to tag such cases.

CC or DT

When they are the first members of the double conjunctions *both ... and*, *either ... or* and *neither ... nor*, *both*, *either* and *neither* are tagged as coordinating conjunctions (CC), not as determiners (DT).

EXAMPLES: Either/DT child could sing.

But:

Either/CC a boy could sing or/CC a girl could dance.
Either/CC a boy or/CC a girl could sing.
Either/CC a boy or/CC girl could sing.

- After you arrived at decent guidelines:
  - Test run with a small group of annotators
  - Check problems and misunderstandings and iterate guidelines
- Lastly: Scale up annotation

Annotation and labeling | Data Needs

- How much test data?
  - Estimate based on statistical significance ($p<0.05$)
  - Determine number by power analysis (Card et al. 2020)

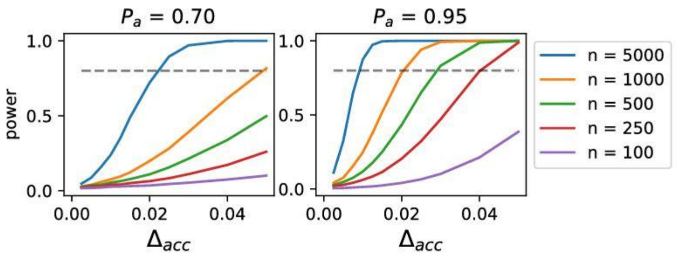


Figure 3: Power for comparing two classifiers on accuracy using paired data depends on the size of the test set (n), the expected agreement (P_a), and the expected difference in accuracy (Δ_{acc}). The dashed line shows 80% power, often taken to be a minimal requirement.

- What about training data?
  - Generally more is better (see above)
  - Can be mitigated by different strategies such as active learning

Annotation and labeling | Assessing annotation quality

- Ideally we assess the quality of our annotations
- The best way to do this is to look for inter-annotator agreement
  - Ex: Double (or triple etc) annotate data
  - Compute statistic such as Cohen's Kappa (Carletta 1996)

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - \boxed{p_o}}{1 - \boxed{p_e}}$$

Observed agreement

Expected agreement

  - How to determine expected agreement?
    □ Chance to agree randomly
- Bad results?
  - Better guidelines
  - Better annotators
  - Better task

Outline

- Why NLP needs (a lot of) data
- Datasets and data sources
- Unlabeled data and pre-training
- Annotation and labeling
- **Augmentation and evaluation**

Augmentation and evaluation

- Ideally our data has the following properties:
  - Contains enough information for modelling
  - Good coverage of the desired task
  - Reflects real inputs the model is expected to receive
  - As unbiased as possible
  - Not a result of a feedback loop
  - Has consistent labels
  - Large enough for generalization

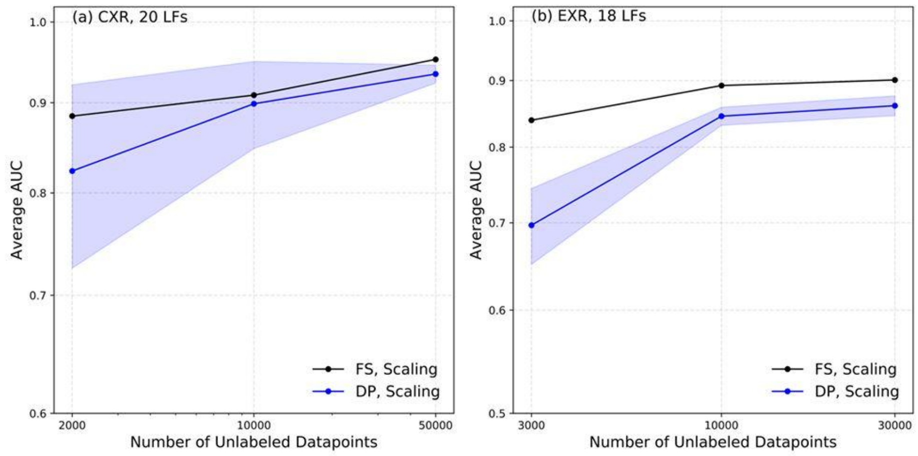
- Quite often real scenarios require compromises on some of these aspects

Augmentation and evaluation | Reporting on data

- Because of this we should document details about our used or created data alongside our research
- Popular framework: Data statements (Bender and Friedman 2018)
- Offers a checklist of things to document about a dataset, e.g.
  - Curation rationale
  - Language and variety
  - Speech situation
  - Speaker demographic
  - Annotator details
  - Etc.

Augmentation and evaluation | Weak supervision

- Weak supervision = Foregoing hand labels
- Instead rely on heuristics in the form of a labeling function (LF)
- This can be demonstrated to achieve excellent results (Dunnmon et al. 2020)



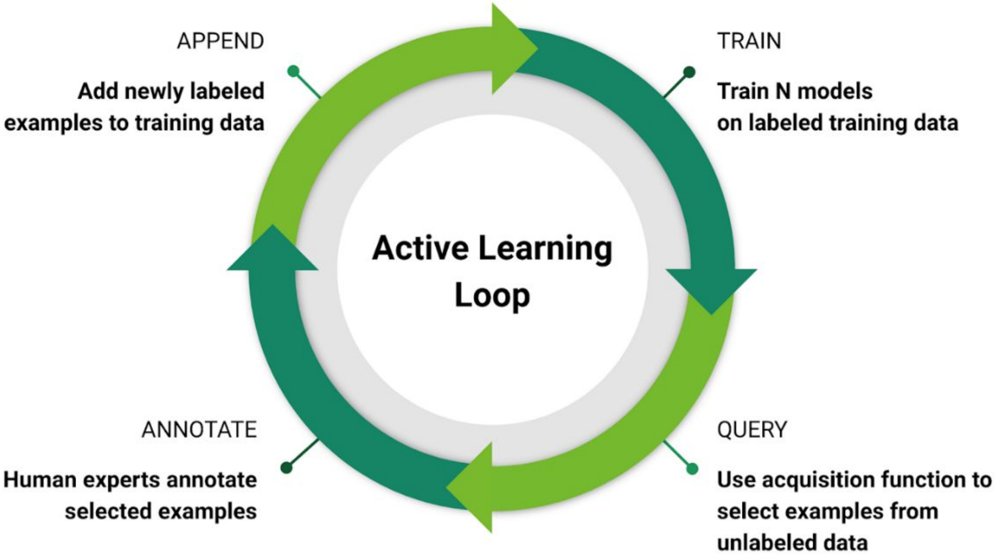
- Problems:
  - Labels can become very noisy
  - Requires extensive feature engineering for the LFs

Augmentation and evaluation | Augmentation

- Data augmentation can help at generating more data within your domain
- Popular in computer vision but generally more noisy for NLP
- In general we can perform augmentation on…
  - Character-level
  - Word-level
  - Sentence-level
- Quite often the most popular augmentation is synonym replacement on the word-level and random shuffling on the sentence-level.
- It's possible to be more fancy, e.g.
  - Back-translation (machine translation to another language, then back)
  - Rephrasing by a LLM

Augmentation and evaluation | Active Learning

- Active learning is a method to choose which data samples to label for further training
- Ex: Uncertainty measurement on unlabeled data to determine highest gain



- Other possible heuristics:
  - Disagreement by candidate models
  - Highest gradient update

Augmentation and evaluation | Evaluation

- In practice it can be very hard to judge the quality of an already released dataset
- If the dataset has its own documentation (e.g. data statements) this can be done more easily retroactively
- Often vital information about the labeling process, guidelines, sampling strategies etc. is missing
- It is therefore important to critically evaluate the data you base your models on
- At the very least adhere to proper train/test split and evaluate qualitatively as well as quantitatively

Final Takeaways

- Capturing the semantics of language is a complex task requiring large amounts of data
- There are a plethora of curated datasets available but they might not fit your task exactly
- You can leverage pre-training or pre-trained models for your own tasks
- Sometimes zero-shot methods are the way to go
- In other cases you can leverage heuristics via weak supervision or data augmentation methods
- In the case where you need to generate your own data take deliberate steps to plan and document how to acquire and (if necessary) annotate the data

Study Approach

Minimal

- Work with the slides

Standard

- Minimal approach + read Card et al. (5)

In-Depth

- = standard approach + skim through references (2) Chapter 4 and (7)

See you next time!

Resources

(1) Graham Neubig: CMU Advanced NLP 2022

(2) Chip Huyen: Designing Machine Learning Systems 2022

(3) Andriy Burkov: Machine Learning Engineering 2020

(4) Kaplan et al.: Scaling Laws for Neural Language Models 2020

(5) Card et al.: With Little Power Comes Great Responsibility 2020

(6) Jean Carletta: Assessing Agreement on Classification Tasks: The Kappa Statistic 1996

(7) Bender and Friedman: Data Statements for Natural Language Processing 2018