

# Transformer Models and Related Work

## HS Experimentelles Arbeiten in der Sprachverarbeitung

Nils Reiter

`nils.reiter@uni-koeln.de`

November 10, 2022

## Section 1

### Transformer Models

# Introduction

- ▶ Transformer: A breakthrough in natural language processing
- ▶ BERT (Google): First successful transformer model
- ▶ RoBERTa (Facebook): Gradual improvement over BERT

Devlin et al. (2019)

Liu et al. (2019)

# Introduction

- ▶ Transformer: A breakthrough in natural language processing
- ▶ BERT (Google): First successful transformer model
- ▶ RoBERTa (Facebook): Gradual improvement over BERT
- ▶ General idea
  - ▶ Transformer architecture
  - ▶ Process whole input at once (max. 512 tokens, = bidirectional)
  - ▶ Pre-training and fine-tuning on different tasks

Devlin et al. (2019)

Liu et al. (2019)

# Process Whole Input

- ▶ Classical language modeling
  - ▶  $n$ -gram models, with  $n \leq 4$
  - ▶ Context: Only previous tokens
- ▶ Bidirectional models
  - ▶ Context: Both directions
  - ▶ Much longer distances

## Pre-Training and Fine-Tuning

- ▶ Transformer models are trained on huge data sets
- ▶ Training one from scratch requires significant resources (time/money)
- ▶ Pre-trained models are shared freely
- ▶ Recipe: Take a pre-trained model and fine-tune it on your task
  - ▶ Pre-trained model contains an abstract language representation

## Pre-Training and Fine-Tuning

- ▶ Transformer models are trained on huge data sets
- ▶ Training one from scratch requires significant resources (time/money)
- ▶ Pre-trained models are shared freely
- ▶ Recipe: Take a pre-trained model and fine-tune it on your task
  - ▶ Pre-trained model contains an abstract language representation
- ▶ Fine-tuning
  - ▶ Any language-related task!

## Pre-Training and Fine-Tuning

- ▶ Transformer models are trained on huge data sets
- ▶ Training one from scratch requires significant resources (time/money)
- ▶ Pre-trained models are shared freely
- ▶ Recipe: Take a pre-trained model and fine-tune it on your task
  - ▶ Pre-trained model contains an abstract language representation
- ▶ Fine-tuning
  - ▶ Any language-related task!
- ▶ In contrast: Classical machine learning
  - ▶ Train everything at once
  - ▶ All that the model learns is taken from one data set



# Transformer Training Tasks

## Masked Language Modeling (MLM)

- ▶ Sentence-wise
- ▶ 15% of the tokens are “masked” by a special token
- ▶ Model predicts these, having access to all other tokens

# Transformer Training Tasks

## Masked Language Modeling (MLM)

- ▶ Sentence-wise
- ▶ 15% of the tokens are “masked” by a special token
- ▶ Model predicts these, having access to all other tokens

## Next sentence prediction (NSP)

- ▶ Two (masked) sentences are concatenated
- ▶ Model has to predict whether second sentence follows on the first or not
- ▶ Not used in RoBERTa

# Parameters and Hyperparameters

- ▶ Parameters: Learned during training
  - ▶ Weights between neurons, tree structure, thresholds, ...
- ▶ Hyperparameters: Our decisions
  - ▶ Choice of ML algorithm
  - ▶ Network layout
  - ▶ Feature set
  - ▶ ...

# Multi-Task-Learning

- ▶ Old idea: Some tasks may be solved at once, because they are related
  - ▶ E.g.: pos-tagging and lemmatisation
- ▶ A single model that predicts two categories
- ▶ Easy to integrate in neural networks

## Section 2

### Unsere Experimente

# Themen

- A Detecting and Rating Humor and Offense
- B Patronizing and Condescending Language Detection
- C Named Entity Recognition and Classification (on historical newspapers)

# Themen

- A Detecting and Rating Humor and Offense
- B Patronizing and Condescending Language Detection
- C Named Entity Recognition and Classification (on historical newspapers)

Heute: Erstes brainstorming: Worauf wollt Ihr Euch fokussieren? Was müsstet Ihr dafür tun?

## Section 3

### Related Work



## Related Work I

- ▶ NLP-Papiere dokumentieren den Forschungsstand (“State of the Art”, “Related Work”)
- ▶ Beginn einer jeden Projektarbeit: Recherche zum Forschungsstand

# Warum?

- ▶ Forschungsstand stellt Kontext her: Was kann als bekannt vorausgesetzt werden?
- ▶ Wissenschaftliche Artikel leisten einen neuen Beitrag
  - ▶ Beitrag muss im Verhältnis zum Forschungsstand sichtbar werden
- ▶ Vollständigkeit / Adäquatheit der Auswahl
- ▶ Unvollständiger Forschungsstand kann zum Ablehnen einer Veröffentlichung führen

# Was sucht man?

## Etablierter Task

- ▶ Welche Ansätze lösen den gleichen Task?
- ▶ Welche (annotierten) Datensätze wurden publiziert? Unterscheiden die sich vielleicht in den Details?

# Was sucht man?

## Etablierter Task

- ▶ Welche Ansätze lösen den gleichen Task?
- ▶ Welche (annotierten) Datensätze wurden publiziert? Unterscheiden die sich vielleicht in den Details?

## Neuer Task

- ▶ Gibt es Ansätze, ein ähnliches Problem zu lösen?
  - ▶ Task-Typen (Textklassifikation, Sequence labeling, ...) helfen bei der Einschätzung was "ähnlich" ist
- ▶ Gibt es Datensätze, die die Bearbeitung des Problems erleichtern?

## Wo sucht man?

- ▶ ACL Anthology
- ▶ Survey-Artikel
  - ▶ Übersichtsartikel, die den Stand (zu einem Zeitpunkt) zusammenfassen und systematisieren
- ▶ “Related-Work”-Abschnitte von Artikeln
- ▶ Google Scholar

## Wo sucht man?

- ▶ ACL Anthology
- ▶ Survey-Artikel
  - ▶ Übersichtsartikel, die den Stand (zu einem Zeitpunkt) zusammenfassen und systematisieren
- ▶ “Related-Work”-Abschnitte von Artikeln
- ▶ Google Scholar

### Im Regelfall **nicht** zitierfähig

- ▶ YouTube-Videos
- ▶ Blogs
- ▶ Webseiten von Firmen
- ▶ Studentische Arbeiten

## Hausaufgabe 3

- ▶ Recherchieren Sie den Forschungsstand zum Thema "Fake News Detection"
- ▶ Erstellen Sie eine kuratierte Liste von mindestens ~~10~~ wiss. Veröffentlichungen
- ▶ Die Liste enthält (in Stichpunkten): Angaben zur Methodik, verwendete Daten, Ergebnisse, besondere Beobachtungen
- ▶ Kuratiert
  - ▶ Nicht wahllos, sondern die "interessanten" Beiträge zum Thema
  - ▶ Nur wissenschaftliche Veröffentlichungen

5