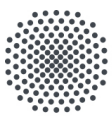
Predicting Structural Elements in German Drama

Janis Pagel*, Nidhi Sihag† and Nils Reiter*

University of Cologne*, University of Stuttgart†

Nov 18, 2021 @CHR2021

University of Cologne



University of Stuttgart
Germany

VolkswagenStiftung

Overview

- Automatic prediction of structural elements of German dramas
  - Act and scene boundaries, speaker tags, stage directions, utterances
- Experiments with different BERT models to predict the location of these elements

Example

ACT 1
Scene 1
Enter Barnardo and Francisco, two sentinels.
BARNARDO Who's there?
FRANCISCO Nay, answer me. Stand and unfold yourself.
...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

Example

ACT 1

Scene 1

Enter Barnardo and Francisco, two sentinels.

BARNARDO Who's there?

FRANCISCO Nay, answer me. Stand and unfold yourself.

...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

- Act boundaries
- Scene boundaries
- Stage directions
- Speaker tags
- Character speech

Example

ACT 1

Scene 1

Enter Barnardo and Francisco, two sentinels.

BARNARDO Who's there?

FRANCISCO Nay, answer me. Stand and unfold yourself.

...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

- Act boundaries
- Scene boundaries
- Stage directions
- Speaker tags
- Character speech

Example

ACT 1
Scene 1
Enter Barnardo and Francisco, two sentinels.
BARNARDO Who's there?
FRANCISCO Nay, answer me. Stand and unfold yourself.
...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

- Act boundaries
- Scene boundaries
- Stage directions
- Speaker tags
- Character speech

Example

ACT 1
Scene 1
Enter Barnardo and Francisco, two sentinels.
BARNARDO Who's there?
FRANCISCO Nay, answer me. Stand and unfold yourself.
...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

- Act boundaries
- Scene boundaries
- Stage directions
- Speaker tags
- Character speech

Example

ACT 1
Scene 1
Enter Barnardo and Francisco, two sentinels.
BARNARDO Who's there?
FRANCISCO Nay, answer me. Stand and unfold yourself.
...

Figure: Beginning of Shakespeare's *Hamlet*, adapted from `https://www.dracor.org/shake/hamlet`

- Act boundaries
- Scene boundaries
- Stage directions
- Speaker tags
- Character speech

Motivation

- Automatic generation of XML/TEI encoded plays from plain-text/OCRed data
- Investigation of what information helps models to distinguish between speech and stage directions

Data

- GerDraCor (`https://github.com/dracor-org/gerdracor`)
  - German Plays 1730–1940
  - 10,021,598 tokens
  - 240,794 types
  - 238,364 sentences

Data

- GerDraCor (`https://github.com/dracor-org/gerdracor`)
  - German Plays 1730–1940
  - 10,021,598 tokens
  - 240,794 types
  - 238,364 sentences

Table: Distribution of classes on the dataset.

Class	Count	Normalized Count (%)
Act	1,458	0.1
Scene	11,001	0.8
Stage	175,238	12.4
Speaker	316,451	22.4
Speech	906,635	64.2

Experiments

- Texts split into sentences using NLTK
- Prediction for each sentence: One of the five classes?
- Baseline: Conditional Random Field (CRF) model
- BERT models:
  - German bert-base-uncased
  - German bert-base-cased

Baseline

- CRF model
- Features:
  - Lowercased sentence string
  - Does sentence contain the German word ‘Akt’?
  - Does sentence contain words ‘Szene’ or ‘Scene’?
  - Does the sentence begin with uppercase letter?
  - Does the sentence only contain uppercase letters?
  - Does the sentence contain a digit?

Results

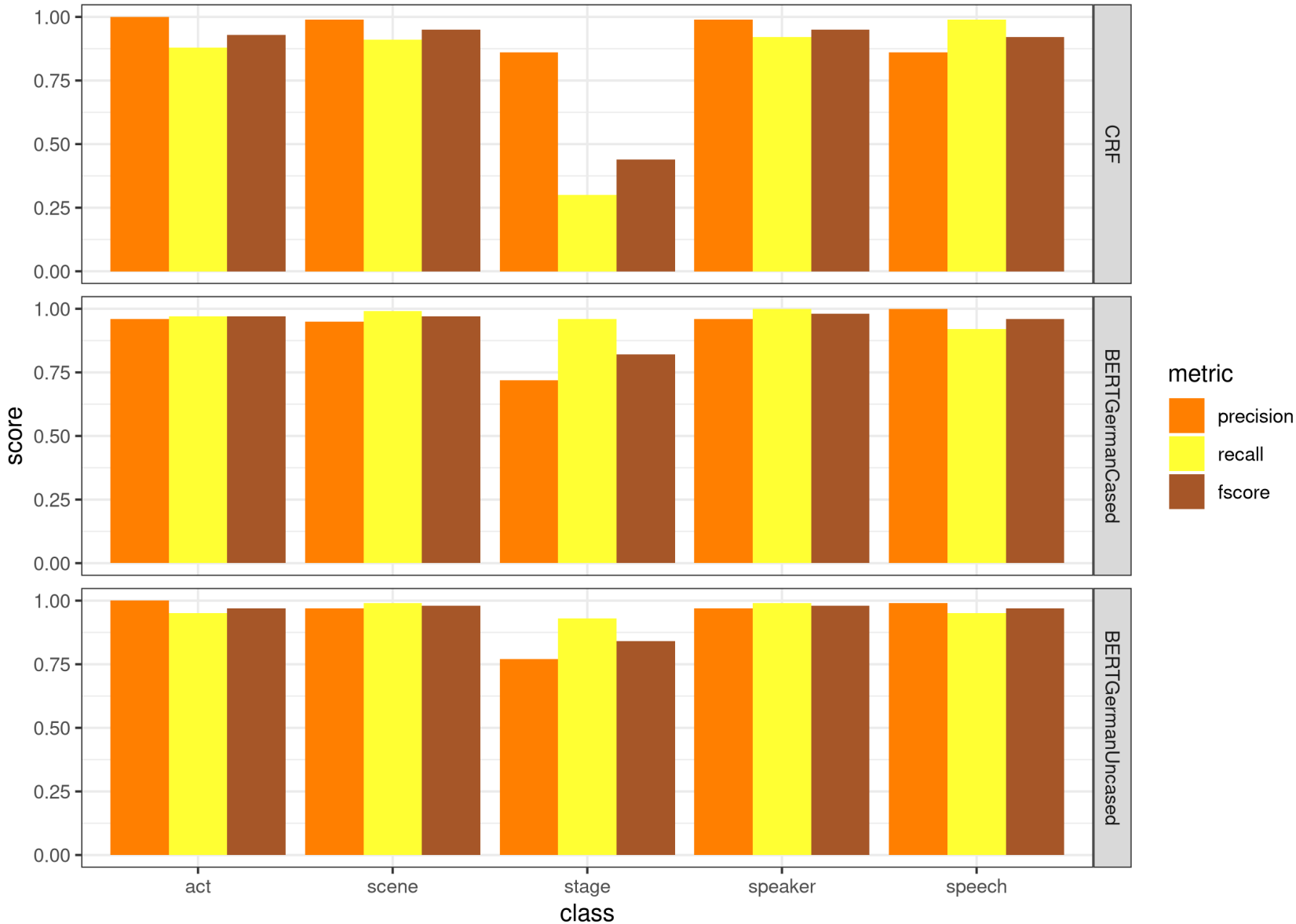


Figure: Results for running the experiments.

English Model

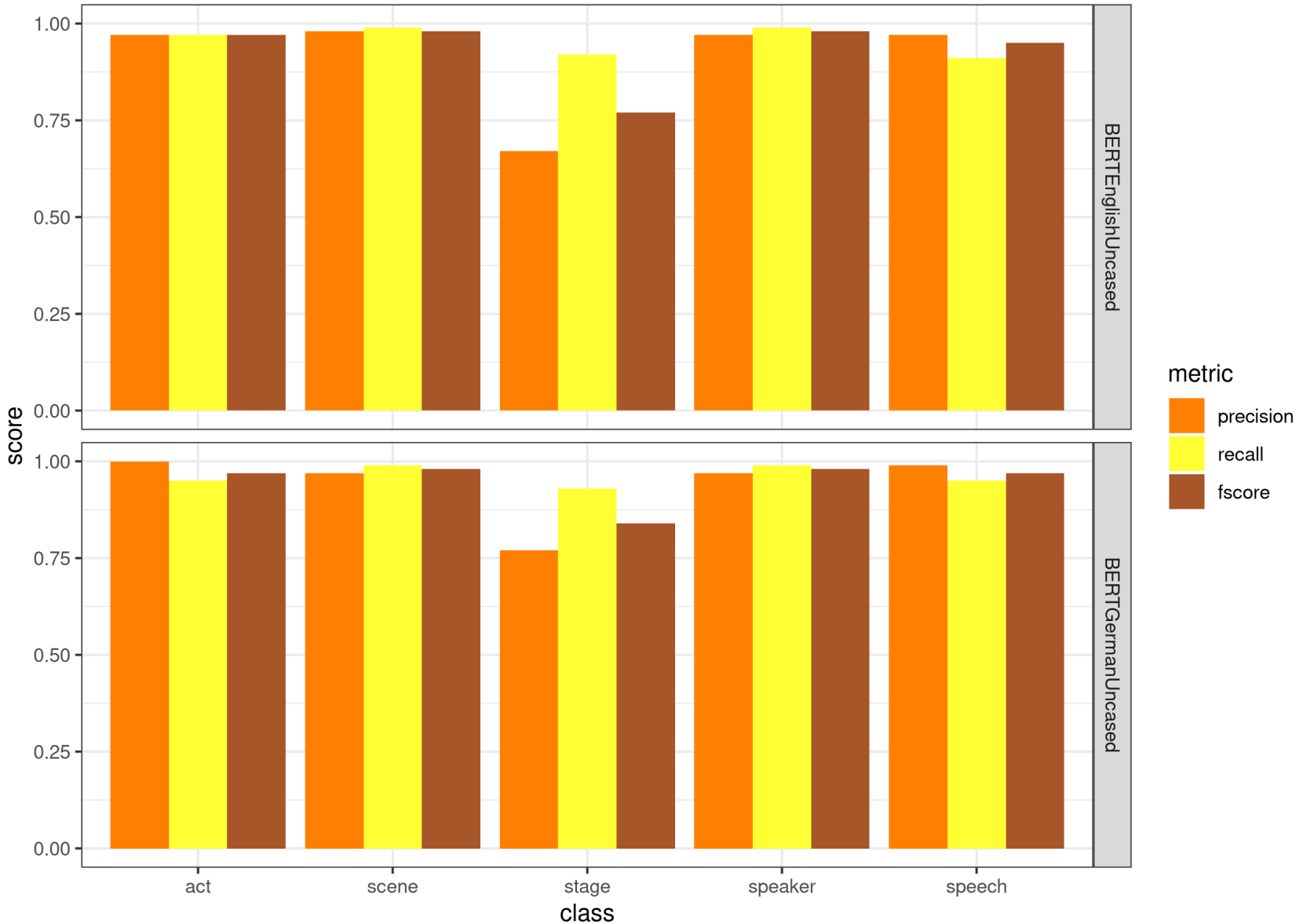


Figure: Comparing an English BERT model with the best German model.

Conclusion

- German bert-base-uncased performed overall the best
- Overt markers like act, scene and speaker are easy to predict (also by the English BERT model)
- The distinction between speech and stage direction is a more challenging task
- BERT models perform well with recall of stage directions, CRF with precision
  - Future work should combine the two models for improved results