University of Stuttgart
Germany

ML-based Emotion Role Labelling

Emotion Analysis
Assignment 4

Felix Bühler
Carlotta Quensel
Max Wegge

Motivation
Research Question #1: Data Choice

- **Research Question #1**

- emotion roles are determined semantically
  »» this information is partially included in the syntactic structure
- label the emotion target which is often an NP (e.g., person/institution the emotion is directed at)

» **(How) does the choice of training data influence the result of the classifier?**

- **Data Choice**

- corpora from different domains:
  »» different syntactic structures
  - GoodNewsEveryone (GNE)
    »» news headlines
    »» include 'ungrammatical' telegram style sentences
  - Reman
    »» complex sentences with three segments from literature
  - Electoral Tweets
    »» everyday language usage from twitter users

» **train and evaluate our models on all three of these very different corpora**

Motivation

Research Question #2: Method Choice

- **Research Question #2**

- sequence labelling is more complex than nominal classification

  »» needs context information

» **How do a naïve and a complex algorithm differ in performance?**

- **Method Choice**

- **compare a naïve approach without much context to a complex method**
  - Hidden Markov model (HMM)

    »» takes into account the context of prior labels (but not of tokens)
  - Transformer

    »» DL-method

    »» pretrained on external data (RoBERTa model)

    »» whole sequence as context for each individual token

Method
Hidden Markov | Viterbi

- trained on observations in the training data
- easily trained only with frequencies:

»» emission probabilities:

compute for all tokens: $\frac{\text{frequency of token,tag}-\text{pair}}{\text{overall token frequency}}$

»» transition probabilities

compute for every tag O, B and I: $\frac{\text{frequency of }\text{tag}_1\text{,}\text{tag}_2\text{ bigram}}{\text{frequency of }\text{tag}_2}$

»» prior probabilities:

compute the relative frequency of each tag as the first tag

- the best labels for a token sequence are the ones with the highest product of probabilities
- Viterbi is used to determine the labels with the maximum sequence probability

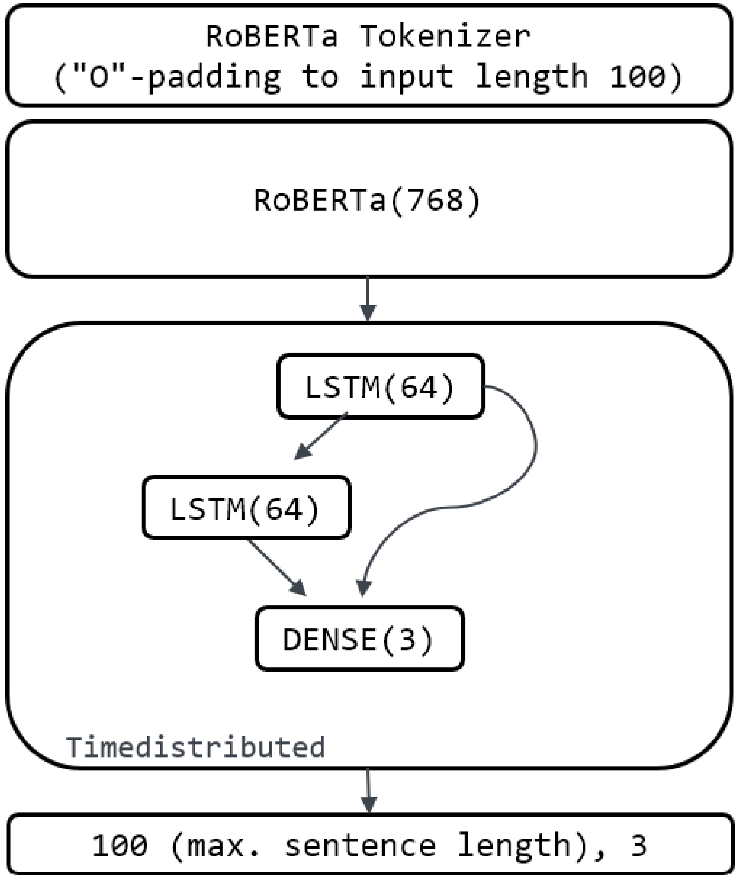
Method
RoBERTa

- maximum input length 100 words

  »» most sequences are shorter
- 1st layer: pre-trained RoBERTa model
- 2nd & 3rd layer: bidirectional LSTM (64 units)
- 4th layer: Dense-Layer to combine all features

  »» layers 2-4 used in a Time-Distributed-Layer to produce a prediction for each token

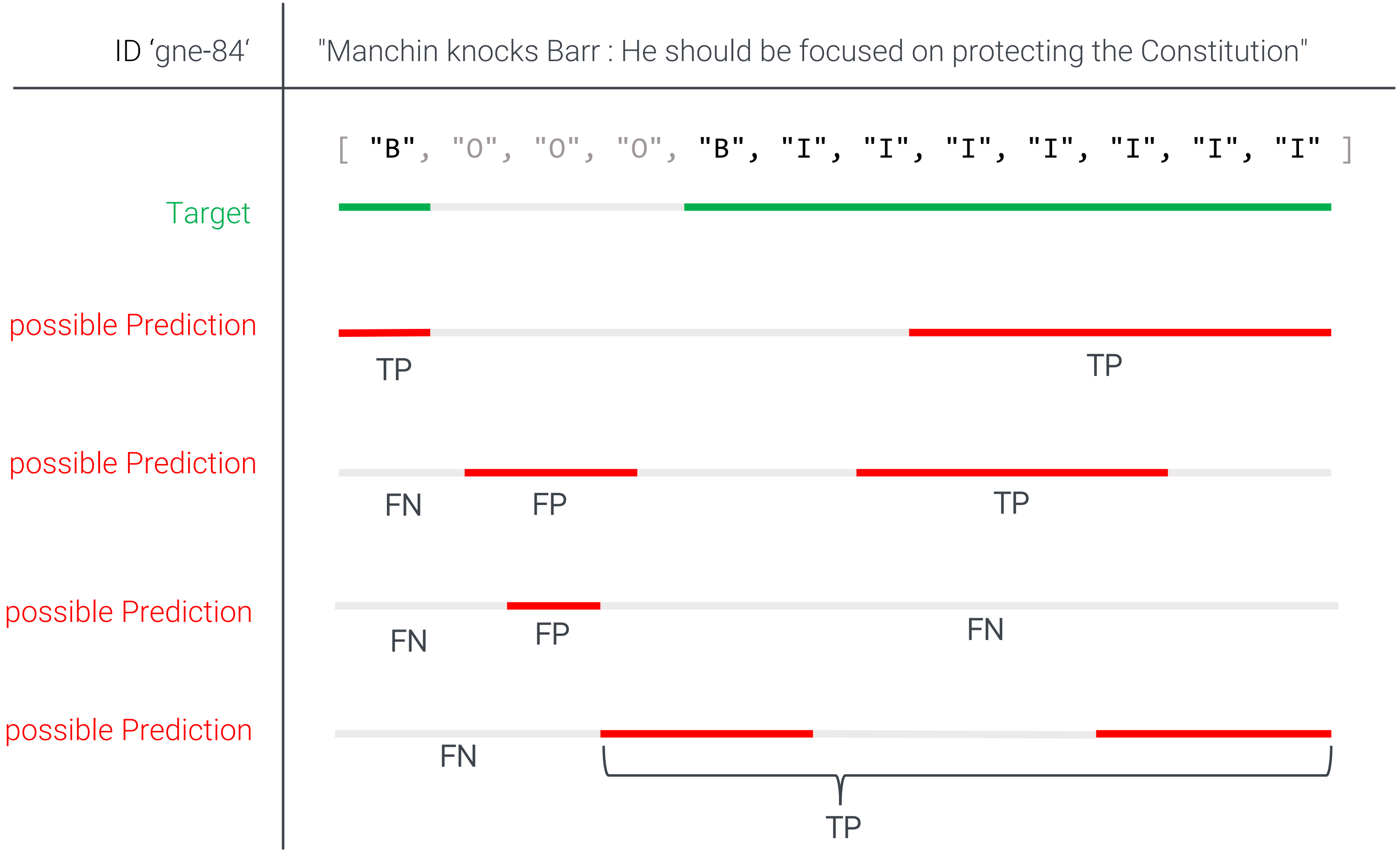
  »» residual connection between the first LSTM and the Dense-Layer to improve accuracy
- added *ReduceLROnPlateau* to reduce the learning-rate for internal metrics when learning stagnates



- intersections between target and predicted sequence are evaluated as True Positive
- empty intersections are evaluated as false classifications

  »» not predicting target sequence: False Negative

  »» predicting non-target sequence: False Positive
- multiple sequences mapped onto one are evaluated as one
- True Negatives (correctly predicting non-target sequence) are omitted

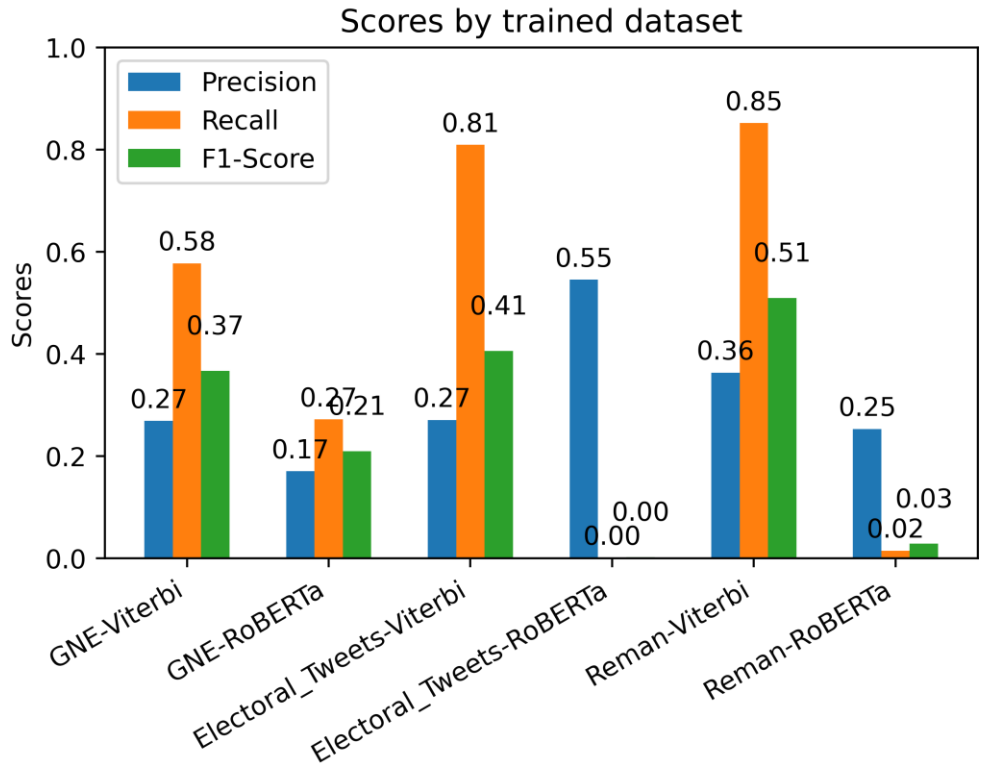


Evaluation
Method

- all models are trained on one corpus and evaluated on the other two
- Viterbi predicts too many and too long sequences regardless of training data
  »» HMM has higher probabilities for B/I → I than for B/I → O
  »» greediness biases the result with our evaluation method
- Transformer predicts too little sequences regardless of training data
  »» padded input length adds more ‘O’ to the training data
- because of the above, the recall of both algorithms is inverted
- precision is fairly low regardless of method and training data
  »» possible clue that more information is necessary (POS-tags, chunking, …)

»» complex task → complex method with limited training capacities

»» simple models do not rely as much on the optimal training time



Evaluation
Data

Viterbi

- performs best when trained on Reman
  - »» literature
    - » syntactically correct
    - » best transition probabilities
  - »» long sentences and infrequent target annotation
    - » combats greediness problem
- worst performance on GNE
  - »» news headlines are the shortest of the data sets
    - » hightens the possibilities for a target sequence
    - » amplifies greediness problem

RoBERTa-approach

- performs best when trained on GNE
  - »» possible compatibility between GNE and pre-training data
- worst performance for Reman
  - »» as the problem of sparse input matrices is amplified by the corpus' inherent sparseness.

Thank you
for listening!

Questions?