Preposition Sense Disambiguation

TTLab - Text2Scene Praktikum

Dirk Neuhäuser Tim Rosenkranz Tobias Marzell

Prof. Dr. Alexander Mehler, Alexander Henlein

11. September 2020

Aufgabe

Motivation

Preposition-Sinn wichtig für Verständnis:

> **Auf** *der Wache* bedeutet eigentlich, dass man **in** dem Gebäude ist
>
> **Auf** *dem Tisch* bedeutet, dass etwas wirklich **auf** dem Tisch ist

Zur Visualisierung braucht man diese Informationen

Nice to Know

SemEval Benchmark:

- **SemEval-Winner 2007** - Max-Entropy-Ansatz: Accuracy bis 75 %
- **Litkowski 2013** - Lemmatizer und Dependency-Parser: Accuracy: 86 %
- **Semi-Supervised(2016)** - Ausnutzen von mehreren Sprachen: Accuracy bis 80 %
- **Hongyu Gong(2018)** - Geographische Context-Vektoren: Accuracy bis 80 %

Unsere Aufgabe

- Datenbeschaffung und -bereinigung
- Trainieren von 2 Taggern (hyperparamter-optimiert)
- Einbinden in den Text-Imager
- Semi-Supervised Ansatz Wenn möglich einbinden

Datenbeschaffung und -bereinigung (Tim und Dirk)

Daten

- SemEval 2007 oft als Benchmark Datensatz
- schwer zu bekommen online
- 34 englische Präpositionen mit insgesamt 224 verschiedene Sinnen
- Je preposition 2 xml (Trainingsdata und label-defintionen)

Datenbeispiel

Trainingsdata:

```
<instance id="with.p.fn.340887" docsrc="FN">
  <answer instance="with.p.fn.340887" senseid="7(5)"/>
  <context>
    Scott threw himself <head>with</head> enthusiasm into this exacting assignment .
  </context>
</instance>


<instance id="with.p.fn.342218" docsrc="FN">
  <answer instance="with.p.fn.342218" senseid="11(7b) 7(5)"/>
  <context>
    He took another sniff , wrinkling his nose <head>with</head> distaste .
  </context>
</instance>
```

Definitionen:

```
<sense id="5">
        <definition>because of / due to (the physical/mental presence of) (e.g., boiling with anger, shining with dew)</definition>
        <majorcluster>CAUSE</majorcluster>
        <pprojmap type="equivalent" targetid="11(7b)"/>
        <notes></notes>
</sense>
<sense id="6">
        <definition>indicating the manner or circumstances (but not cause or motivation) of something (e.g., fix with precision)</d
        <majorcluster>MANNER</majorcluster>
        <pprojmap type="equivalent" targetid="7(5)"/>
        <notes></notes>
</sense>
```

Aufbereitete Daten

sentence_id	sentence	label_id	definition
0	The USSR 's fifteen union republics , united <head>on</head> a supposedly voluntary basis , formed an ` integral , federal , multinati	22	having (the thing mentioned) as criteria used in judgment or evaluation or as a/the justi
1	Taking usage rate as a variable essentially means segmenting <head>on</head> the basis of volume purchased .	22	having (the thing mentioned) as criteria used in judgment or evaluation or as a/the justi
2	She pinched bruises <head>on</head> her daughter 's inner arm , and had poured hot tea on both daughters .	1	indicating a surface that, due to contact with, serves as the cause or means of causing
3	Gingerly I squeezed a bit <head>on</head> my fingertip .	1	indicating a surface that, due to contact with, serves as the cause or means of causing
4	Tom squeezed <head>on</head> the reins and they came to a halt .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
5	I pointed my piece at Johnny and squeezed <head>on</head> the trigger .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
6	Madeleine and Victorine stood behind his wheelchair , like nurses , while the two girls huddled <head>on</head> the windowseat .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
7	She would kneel <head>on</head> the floor behind the sofa and pull out the tall books of art reproductions .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
8	She knelt <head>on</head> the cold stone floor and carefully placed some coals on the dying embers in the grate .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
9	Sung was kneeling <head>on</head> the top of the dyke , staring across at the House as the dawn broke .	13	physically in contact with and supported by (a surface) (e.g., the book on the table)
10	` The premises are freehold , " he continued , not responding to my comment , but at least no longer leaning <head>on</head> the co	13	physically in contact with and supported by (a surface) (e.g., the book on the table)

Huggingface-Bert (Dirk)

Huggingface-Bert

Huggingface library:

- Leichte Umsetzung von State-of-the-Art Modellen
- Unterstützt bert

Bert:

- Released: November 2019
- Sehr(!) gut vortrainiert
- Knackt gleich in mehreren Disziplinen die State-of-the-Art (GLUE, SQuAD, SWAG)

Huggingface-Bert-Trainer

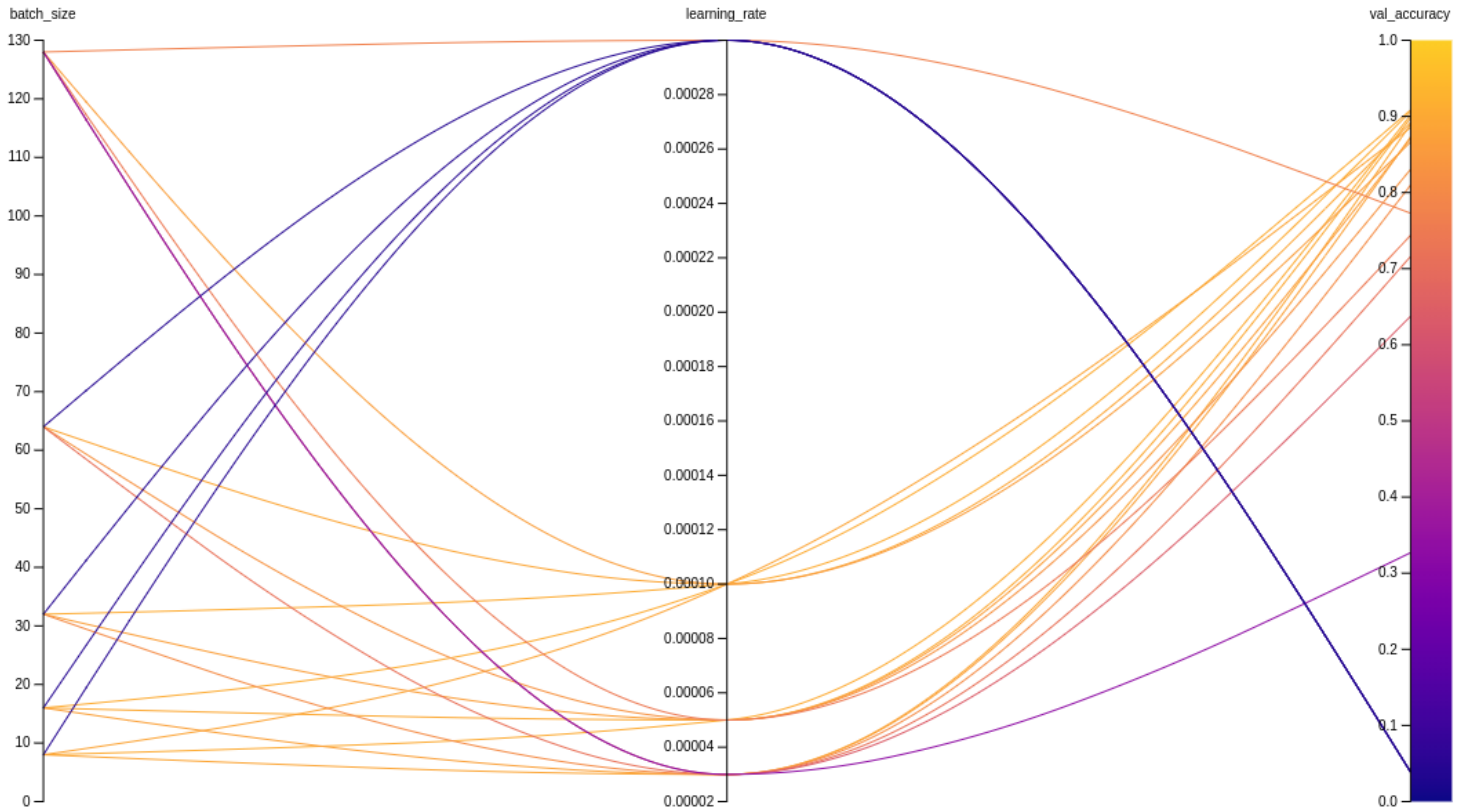
1. Daten einlesen (90:10 train:val Split)
2. Daten tokenisieren (pretrained tokenizer von hf für Bert): Sätze werden zu **Input-Ids** und **Attention-Masks**
3. bert Modell initialisieren als **Klassifizierungsmodell** (mit 224 verschiedenen Outputs)
4. Trainingsschleife

Huggingface-Bert-Hyperparameter-Optimierung

Google empfielt

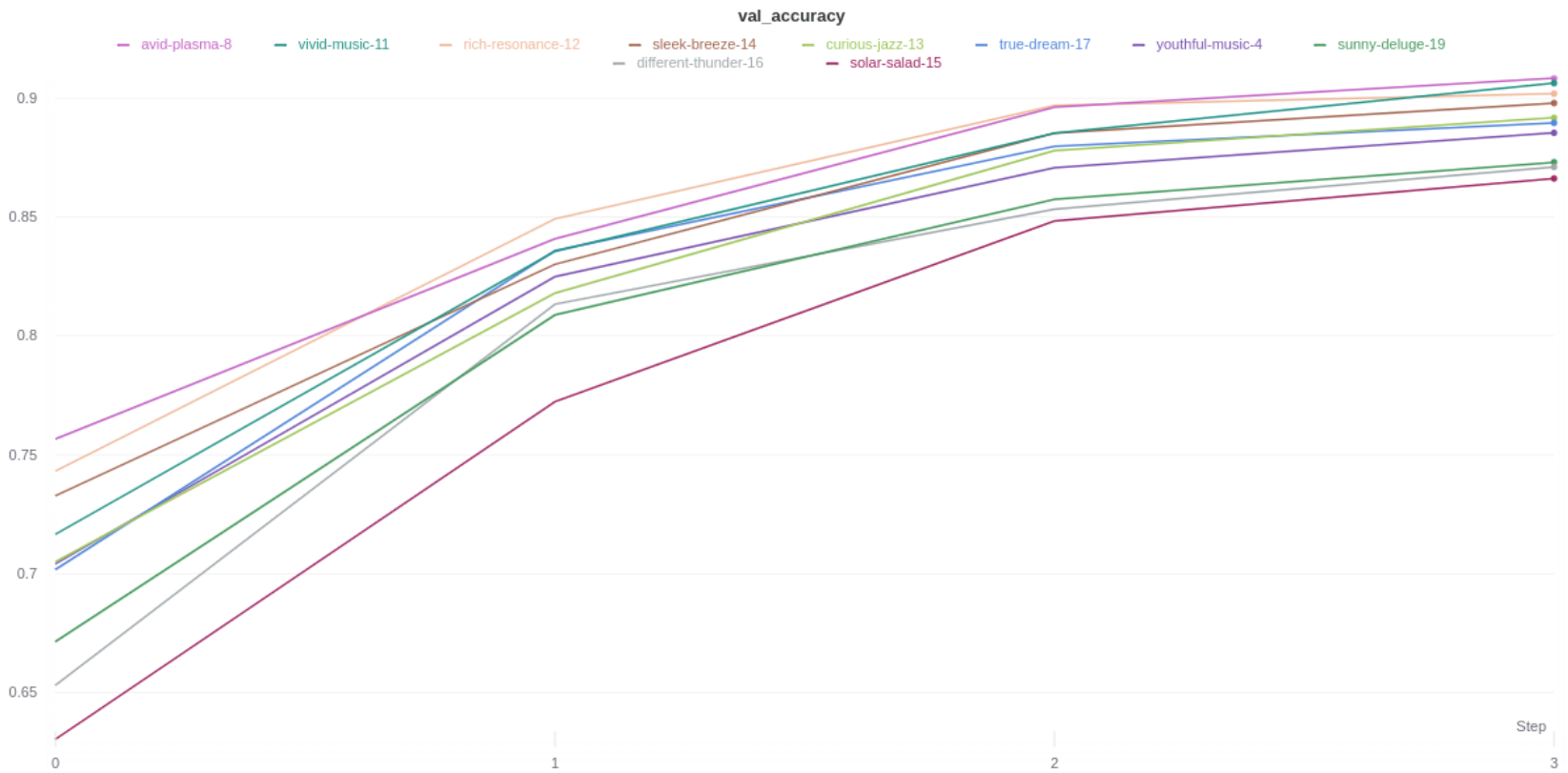
- Epochen: 4
- Optimizer: Adam
- Learning-Rate aus [3e-4, 1e-4, 5e-5, 3e-5]
- Batch-Size aus [8, 16, 32, 64, 128]
- Rest: schon vorgegeben

Huggingface-Bert-Ergebnisse

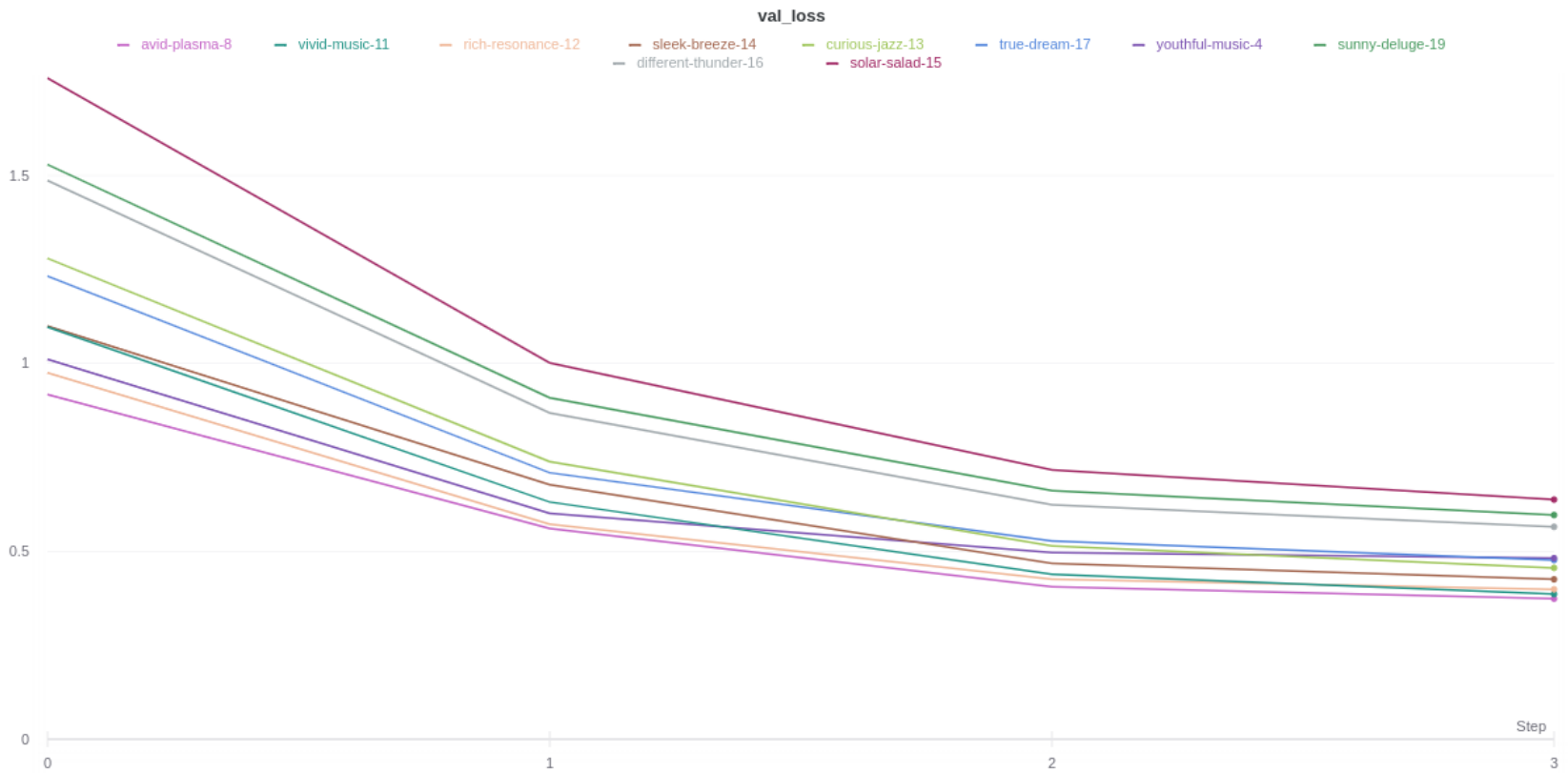


Beste Acc: 0.9084 mit Learning-Rate = 1e-4 und Batch-Size = 16

Huggingface-Bert-Ergebnisse



Huggingface-Bert-Ergebnisse



Huggingface-Bert-Evaluation

```
sentence : I am <head>in</head> big trouble
prediction : indicating a state/condition/form, often a mental/emotional one that is being experienc


sentence : I am <head>in</head> a big airplane
prediction : indicating a LOCATION that surrounds (or can be viewed as surrounding) something else (


sentence : I am <head>in</head> New York
prediction : indicating a LOCATION that surrounds (or can be viewed as surrounding) something else (


sentence : <head>In</head> 2020 Donald Trump will be re-elected
prediction : indicating a TIME_PERIOD during which something happens or continues (e.g., in the morn


sentence : I always see you <head>in</head> my dreams.
prediction : indicating a LOCATION that surrounds (or can be viewed as surrounding) something else (


sentence : I am speaking <head>in</head> portuguese.
prediction : indicating the language, medium, or means of encoding (e.g., spoke in German)


sentence : Donald Trump appears <head>in</head> a weird manner.
prediction : indicating a manner that something happens or is done, often somewhat idiomatic (e.g.,
```

Huggingface-Bert-Evaluation

```
sentence : <head>By</head> 2021 I will have a bachelor degree.
prediction : indicating the size or amount, as of a margin (e.g., increase by 7%)


sentence : <head>By</head> the way, Joe Biden is actually better than is Opponent.
prediction : indicating the MEANS of achieving something (e.g., melt it by cooking it)


sentence : The Crisis was handled <head>by</head> the best president ever - Donald Trump.
prediction : indicating the logical subject (that is, the word that would be the subject in an act


sentence : He is leading the polls <head>by</head> far.
prediction : indicating the size or amount, as of a margin (e.g., increase by 7%)
```

Huggingface-Bert-Evaluation-Details

preposition	accuracy	occurrence in test-data	different senses
through	0.6512	43	13
inside	0.7143	7	3
round	0.7500	12	3
around	0.7632	38	5
after	0.7778	9	6
before	0.8000	5	3
above	0.8000	5	5
beneath	0.8000	5	3
off	0.8125	16	4
like	0.8462	26	7
by	0.8462	52	8
in	0.8561	139	12
behind	0.8571	14	4
on	0.8750	88	19
during	0.8750	8	2
from	0.8852	122	16
into	0.8852	61	8
over	0.8947	19	8
between	0.9130	23	6
for	0.9263	95	11
of	0.9271	288	16
to	0.9322	118	10
down	0.9394	33	3
against	0.9474	19	6
at	0.9577	71	9
across	0.9688	32	2
with	0.9832	119	14
about	0.9859	71	4
as	1.0000	17	1
along	1.0000	36	3
towards	1.0000	22	4
onto	1.0000	10	2
beside	1.0000	6	1
among	1.0000	10	3

Huggingface-Bert-Anbindung

```
<type5:Sentence begin="0" end="31" sofa="1" xmi:id="312"/>
<type5:Sentence begin="32" end="107" sofa="1" xmi:id="317"/>
<type6:WordSense begin="15" end="17" sofa="1" value="45" xmi:id="322"/>
<type6:WordSense begin="25" end="27" sofa="1" value="45" xmi:id="327"/>
<type6:WordSense begin="51" end="53" sofa="1" value="111" xmi:id="332"/>
<cas:Sofa mimeType="text" sofaID="_InitialView" sofaNum="1" sofaString="This is a test by Barack Obama. And this is another Test by the greatest president of the US - Donald Trump" xmi:id="1"/>
```

FairNLP (Tim)

FlairNLP

- Aktuell, spezialisiert für NLP-Aufgaben [Akbik, Blythe und Vollgraf 2018]
- Einfaches Framework
- Basiert auf PyTorch

Flair Trainer

- Daten im CSV-Format in Korpus laden
  - Mindestens train.csv, optional dev.csv und test.csv
  - dev und test werden ggf. von flair erstellt

Optimisierung:

- Optimisierung mit Hyperopt-Wrapper
- Mehrere Embeddings, Lernraten, etc. eingestellt
- Unbekannte Fehler
- log files im repository

Flair Optimisierung

```
Results:
- F-score (micro) 0.9835
- F-score (macro) 0.7382
- Accuracy 0.9835
```

```
Total sentences: 1640; Correct: 145; % correct: 0.08841463414634146
```

Ergebnisse des Trainings mit optimierten hyperparameter

	precision	recall	f1-score	support
__label__147	1.0000	1.0000	1.0000	18
__label__298	1.0000	1.0000	1.0000	5
__label__207	1.0000	1.0000	1.0000	9
__label__149	1.0000	1.0000	1.0000	28
__label__145	1.0000	0.8571	0.9231	7
__label__78	1.0000	1.0000	1.0000	4
__label__155	0.9815	1.0000	0.9907	53
__label__77	1.0000	1.0000	1.0000	11
__label__168	1.0000	1.0000	1.0000	1
__label__89	0.9688	1.0000	0.9841	31
__label__152	1.0000	1.0000	1.0000	17
__label__245	1.0000	1.0000	1.0000	65
__label__40	1.0000	1.0000	1.0000	7
__label__99	0.8824	1.0000	0.9375	15
__label__156	1.0000	1.0000	1.0000	67
__label__36	1.0000	1.0000	1.0000	8
__label__307	1.0000	0.9677	0.9836	31
__label__153	1.0000	1.0000	1.0000	64
__label__209	1.0000	1.0000	1.0000	1
__label__295	0.0000	0.0000	0.0000	0
__label__46	1.0000	0.8571	0.9231	7
__label__34	1.0000	0.9600	0.9796	25
__label__278	1.0000	1.0000	1.0000	14
__label__61	1.0000	1.0000	1.0000	19
__label__2	0.9167	1.0000	0.9565	11
__label__175	1.0000	1.0000	1.0000	17
__label__196	1.0000	1.0000	1.0000	29
__label__188	1.0000	1.0000	1.0000	6
__label__266	1.0000	1.0000	1.0000	7
__label__216	1.0000	1.0000	1.0000	15
__label__250	1.0000	1.0000	1.0000	3
__label__3	1.0000	1.0000	1.0000	3
__label__14	1.0000	1.0000	1.0000	10
__label__13	1.0000	1.0000	1.0000	14
__label__69	1.0000	1.0000	1.0000	4
__label__45	1.0000	1.0000	1.0000	1
__label__47	1.0000	1.0000	1.0000	2
__label__161	1.0000	1.0000	1.0000	6
__label__41	1.0000	0.8333	0.9091	6
__label__203	0.9130	0.9767	0.9438	43
__label__33	0.9600	1.0000	0.9796	24
__label__120	1.0000	1.0000	1.0000	4
__label__212	0.9583	1.0000	0.9787	46
__label__170	1.0000	1.0000	1.0000	12
__label__215	0.8000	1.0000	0.8889	8
__label__213	1.0000	0.9474	0.9730	19
__label__274	0.9032	1.0000	0.9492	28
__label__66	1.0000	0.8529	0.9206	34
__label__95	1.0000	1.0000	1.0000	29
__label__260	0.0000	0.0000	0.0000	0
__label__181	0.9524	1.0000	0.9756	40
__label__217	1.0000	1.0000	1.0000	5
__label__49	0.8750	1.0000	0.9333	7
__label__234	1.0000	1.0000	1.0000	12

Ergebnisse je Klasse

```
2020-09-11 17:03:12,480 Model: "TextClassifier(
  (document_embeddings): DocumentPoolEmbeddings(
    fine_tune_mode=none, pooling=mean
    (embeddings): StackedEmbeddings(
      (list_embedding_0): OneHotEmbeddings(
        min_freq=3
        (embedding_layer): Embedding(11805, 300)
      )
      (list_embedding_1): WordEmbeddings('glove')
    )
  )
  (decoder): Linear(in_features=400, out_features=238, bias=True)
  (loss_function): CrossEntropyLoss()
  (beta): 1.0
  (weights): None
  (weight_tensor) None
)"
```

Einstellungen der Hyperparameter für gutes Ergebnis

Flair Trainer

- Daten im CSV-Format in Korpus laden
  - Mindestens train.csv, optional dev.csv und test.csv
  - dev und test werden ggf. von flair erstellt
- Embeddings: FlairEmbeddings + DocumentRNNEmbeddings
- Mindestens rund 100 Epochen Training

Semi-Supervised (Tobias)

Datenbeschaffung

- European Parliament Proceedings Parallel Corpus 1996-2011 (http://www.statmt.org/europarl/)
- CDEC Word Aligner (https://github.com/redpony/cdec)

Daten Vorbereitung

- Tokenizen der Daten
- Alle Tokens in lower-case überführen
- Zusammenführen der beiden Corpora (special format)
- Entfernen von unvollständigen Zeilen
- Wörter alignen

Verfolgte Ansätze

- Bidirectional LSTM selbst trainieren
- Transformers Model transfer learning
- Bert Transfer Learning

Ausblick

- Verschiedene Sprachen ausprobieren
- Output in die anderen Classifier einbinden

Quellen

Quellen

Akbik, Alan, Duncan Blythe und Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *COLING 2018, 27th International Conference on Computational Linguistics*, S. 1638–1649.