Wikidata Linking

Jawhar M’barek, Farshad Nazifi, Michael Tebbe, Jihed Zaoueli

Nicht einmal Mediziner kennen alle Krankheiten auswendig

Lösung: Die automatische Erkennung von Krankheiten in wissenschaftlichen Publikationen und Verlinkung mit dem Nachschlagewerk ICD10.

https://www.pexels.com/photo/healthy-clinic-doctor-health-42273/

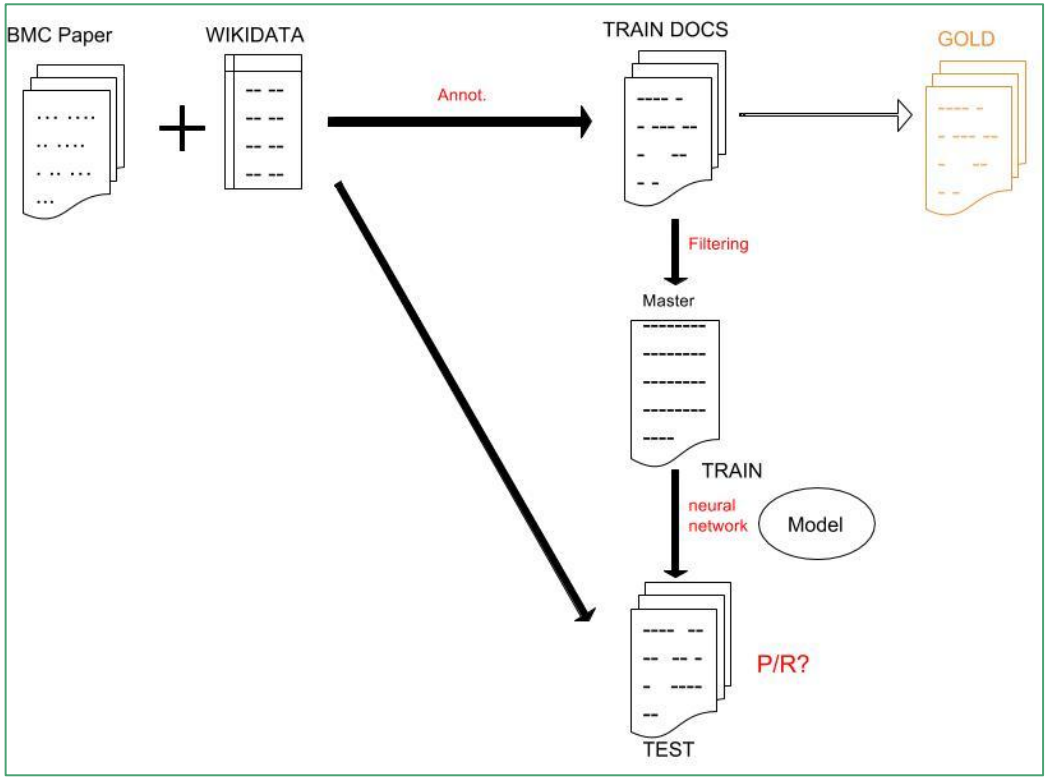
Gliederung

1. Vorgehen
2. Daten
3. Gold-Standard
4. Matching
5. Training
6. Ausblick

Vorgehen

Vorgehen

1. Extraktion
   a. Krankheiten aus Wikidata
   b. Texte aus BMC
2. Annotierung mit Matcher
3. Erstellung Goldstandard
4. Filterung für Training
   a. Traindokument
   b. Testdokument
5. Evaluation Trainingsdaten
6. Training
7. Evaluation Hyperparameter



Daten

Daten

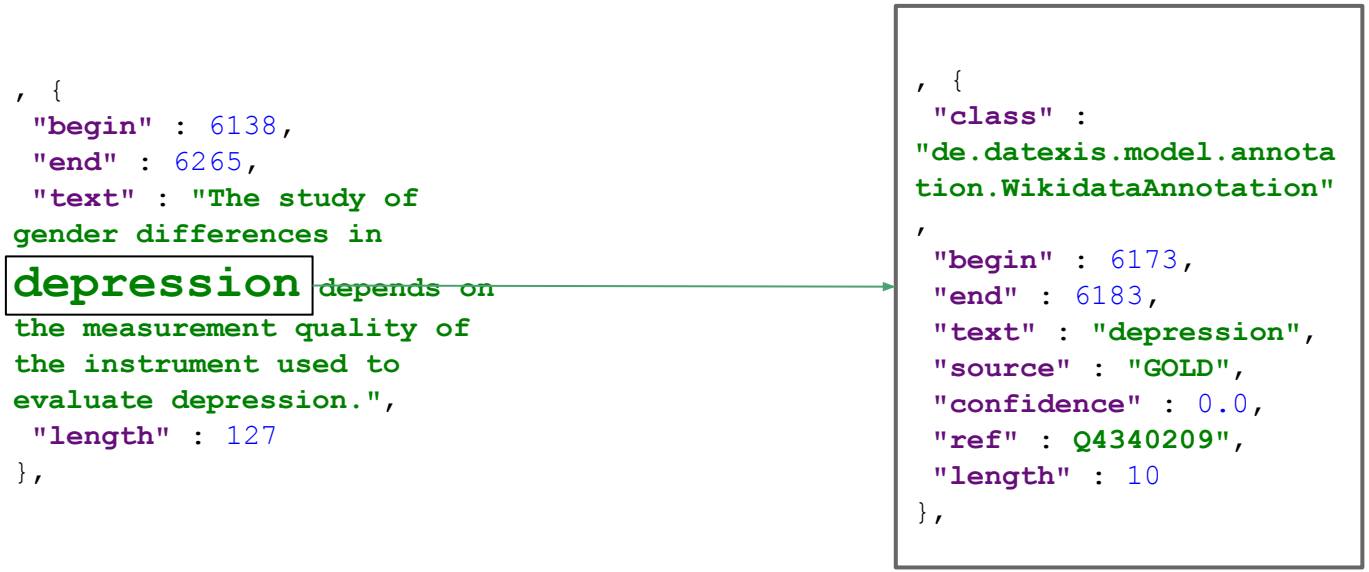
- Extrahierte englischsprachige Artikel aus dem BMC-Korpus
  - Format: Plain text
  - Enthalten: Titel, Textabschnitte, Bildunterschriften und Quellen
  - Nicht enthalten: Tabelleninhalte

- Liste von Krankheiten in Englisch aus Wikidata
  - Format: JSON
  - Enthalten: Synonyme, ICD10- und UMLS-Referenzen

Daten - Analyse

	Anzahl Dokumente	Ø Länge in Sätzen	Ø Matches pro Dokument	unique Matches	Ø Matches pro Satz
Ausgangsdokumente	1561	76,0	15,0	392	0,19
Trainingsdokumente	2	857	1256	330	1,46
Testdokumente	1	262	378	62	1,44

Daten


, {
"begin" : 6138,
"end" : 6265,
"text" : "The study of gender differences in depression depends on the measurement quality of the instrument used to evaluate depression.",
"length" : 127
},
, {
"class" : "de.datexis.model.annotation.WikidataAnnotation",
"begin" : 6173,
"end" : 6183,
"text" : "depression",
"source" : "GOLD",
"confidence" : 0.0,
"ref" : Q4340209",
"length" : 10
},
Satz
Annotation

Goldstandard

Gold-Standard

- 40 Dokumente (8125 Sätze)
  - Maschinenlesbar
  - eigener Importer
- Kappa-Wert nicht gemessen
- Aber: Annotationsregeln festgelegt
  - z. B. Adjektivformen (obesity -> obese): Nein
    Teilwörter (depression subscale): Ja

Matching

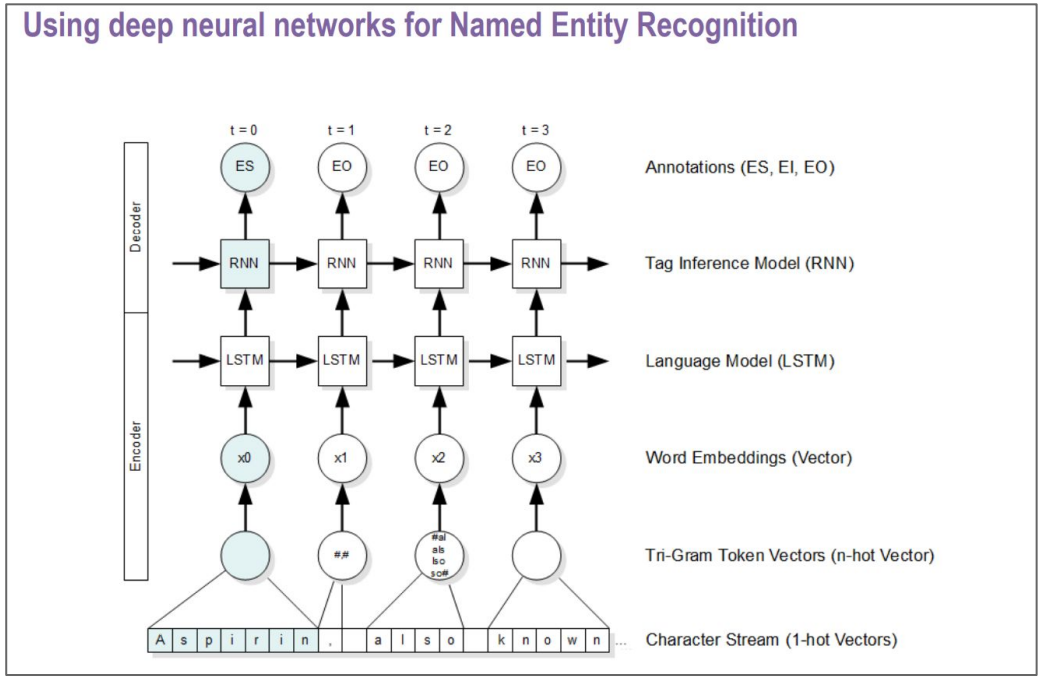
Matching

Ansatz	Precision	Recall	F1-Score
Containment	56%	54%	0,45
Boundaries	70%	56%	0,56
Stemming	66%	64%	0,64
Fuzzy	61%	67%	0,65

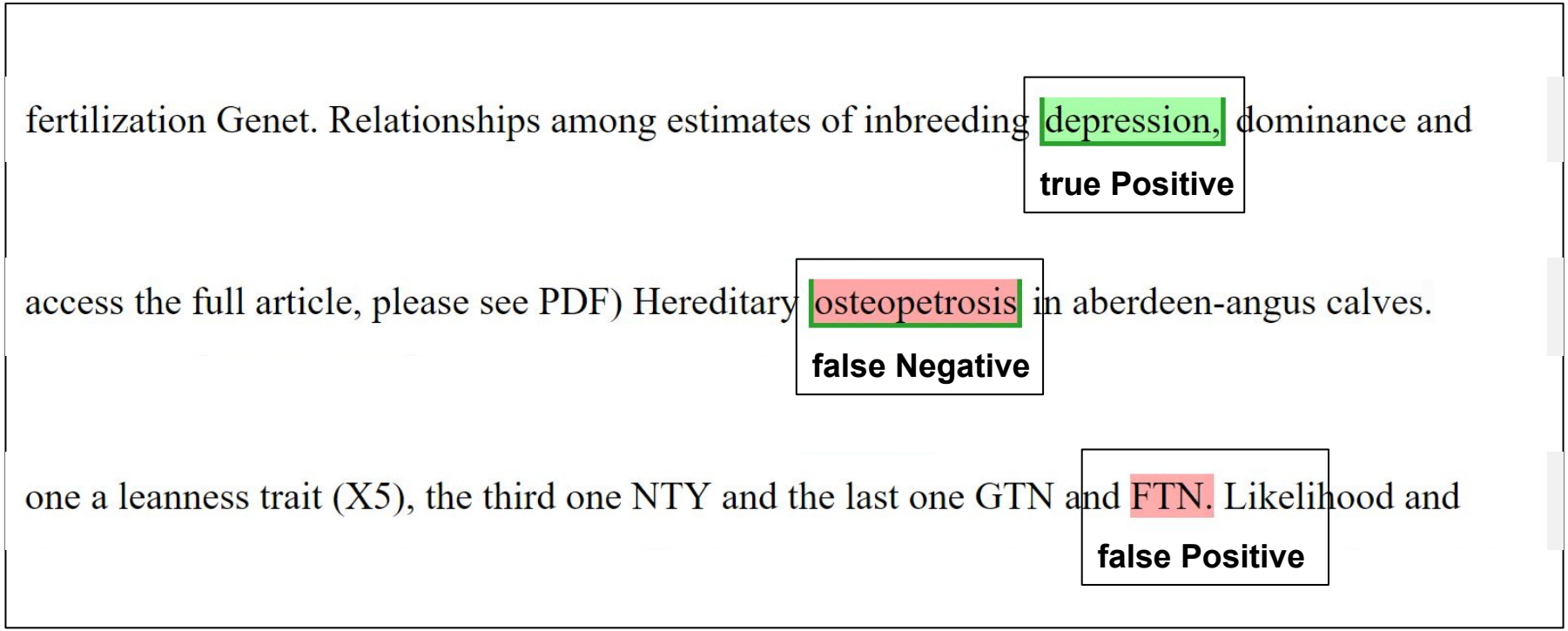
Evaluation gemessen am Gold-Standard

Training

Training - Netzarchitektur



Training - Annotierungsbeispiele



Quelle: ffwNeuro_400_lstmNeuro_200_batchSize_24_numEpochs_7_learningRate_0.01.html

Training

- Hyperparameter Evaluation mit Gridsearch
- getestete Hyperparameter:
  - Anzahl FFW-Neuronen (300-500)
  - Anzahl LSTM-Neuronen (100-300)
  - Epochen (1-10)
  - Learning Rate (0,001-0,01)

Training - Auswertung der HP

Gesamt: 33 Durchläufe

Nach Recall sortiert

FFW-Neuronen	LTSM-neuronen	Batch-size	Iterations	Epochs	Learning-Rate	Precision	Recall	F1-score
400	100	24	1	3	0,001	85.38	67.99	75.70
400	275	16	1	7	0,008	65.37	66.93	66.14
400	275	16	1	7	0.01	55.63	66.67	60.65
435	200	24	1	7	0,001	64.06	65.08	64.57
435	200	24	1	3	0,010	57.48	65.08	61.04
400	275	24	1	7	0,01	64.47	64.81	64.64

Nach Precision sortiert

FFW-Neuronen	LTSM-neuronen	Batch-size	Iterations	Epochs	Learning-Rate	Precision	Recall	F1-score
400	100	24	1	3	0,001	85.38	67.99	75.70
400	275	24	1	3	0.01	80.22	56.88	66.56
400	275	24	1	3	0,008	79.17	35.19	48.72
400	200	24	1	3	0,001	73.55	60.32	66.28
400	275	32	1	3	0,008	73.00	50.79	59.91
500	200	24	1	3	0,001	71.51	63.76	67.41

Ausblick

Erkenntnisse und Verbesserungsmöglichkeiten

- FFW-Neuronen: ca. 400
- unter 10 Epochen
- niedrige Batchsize = höherer Recall, niedrigere Precision
- LSTM-Layer & Lernrate: Daten nicht aussagekräftig

- Annotierung mit Exact Match funktioniert nur bedingt
- training mit besseren Daten nötig
- mehr Dokumente im Goldstandard
- genauere Hyperparameterevaluation

Vielen Dank für ihre Aufmerksamkeit!

Gibt es Fragen?