

Named Entity Classification

Madita Huvar, Sanaz Safdel, Phillip R.-P.

January 12, 2017

- 1 Einführung
- 2 Daten & Tools
 - Tools
 - Korpus
 - Korpusklassen
- 3 Klassifizierer
 - Features für den Baseline-Klassifizierer
 - Erweitertes Featureset
 - Klassifizierertyp
 - Probleme
 - Erfahrungen mit den Korpusklassen
- 4 Evaluation
- 5 Ausblick
- 6 Referenzen

- Python 3.4+
- Scikit Learn als Klassifizierer
- liac-arff
- Weka zur Korpusanalyse

Für die Named Entity Klassifikation nutzen wir das OntoNotes Korpus 2012.

Dabei nutzen die englischen Nachrichtentexte des The Wall Street Journal. Für die Entwicklungsphase nutzten wir das im OntoNotes Korpus bereits vorgefertigte Developmenttest.

Für die Klassifikation der Named Entities werden die bereits vorgefertigten Trainings- und Testdatensets genutzt.

Table: Anzahl an atomaren Named Entities

Developmentset	Trainingset	Testset
3325	23686	2996

Für die Extraktion der Named Entities wurde ein Korpusreader erstellt. Der Reader extrahiert alle Named Entities, inklusive POS-tags der einzelnen Tokenm, Phrasenart, Kontextwörtern (ne-1, ne+1), und ordnet sie ihren Klassen zu.

Beispiel:

```
{'PERSON':[['Peter', 'NNP'], ['Mokaba', 'NNP'], 'NP', ('Says', ',')]}
```

Table: Klassen im OntoNotes Korpus

Klassen	Trainingset
ORG	5788
PERSON	3756
GPE	3601
NORP	1484
PERCENT	1061
CARDINAL	1852
MONEY	1509
DATE	4080
FAC, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, TIME, QUANTITY, ORDINAL	< 1800

Table: Balancierte Klassen

Klassen	Beschreibung
PERSON	People, including fictional
NORP_GPE	Nationalities or religious or political groups Countries, cities, states
ORGANIZATION	Companies, agencies, institutions, etc.
DATE	Absolute or relative dates or periods
PERCENT_MONEY_CARDINAL	Percentage (including “%”) Monetary values, including unit Numerals that do not fall under another type

Table: Verteilung der Klassen

Klassen	Developmentset	Trainingset	Testset
ORG	930	5857	859
PERSON	486	3759	413
GPE_NORP	732	5134	588
PERCENT_CARDINAL_MONEY	564	4672	529
DATE	613	4254	601

Table: Features für den Baseline-Klassifizierer

Feature	Wert	Beschreibung
Unigram	numerisch	Vorkommen der Unigramme, die mindestens fünfmal im Trainingscorpus vorkommen

Erweitertes Featureset I

Anzahl der Features: 1716

Table: Features für den Baseline-Klassifizierer I

Feature	Wert	Beschreibung
Unigram	numerisch	Vorkommen der Unigramme (lemmatisiert), die mindestens fünfmal im Trainingscorpus vorkommen
POS	numerisch	Häufigkeit von 36 POS-Tags aus der Penn Treebank
isAllCaps	boolean	Wörter nur in Großschreibung
Context	numerisch	Vorkommen der Kontexttokens, die mindestens fünfmal im Trainingskorpus vorkommen. Das Kontextfenster beinhaltet das Vorgänger- und Nachfolgetoken der NE.
containsDigit	boolean	Vorkommen von Nummern.

Table: Features für den Baseline-Klassifizierer II

Feature	Wert	Beschreibung
isInWiki	boolean	Vorkommen der NE in der Wikipedia.
isTitle	boolean	Prüft, ob Titelbezeichnungen (z.B. Mr. MA) vorkommen.
isNP	boolean	Ist NE eine Nominalphrase
isName	boolean	Prüft, ob Vornamen vorkommen..
containsDash	boolean	Vorkommen von Viertelgeviertstrichen

Zur Klassifizierung der NE werden zwei verschiedene Klassifizierer genutzt.

- SVM (`sklearn.svm.LinearSVC`)
Parameter: Dual: True, loss: squared hinge, class weight: balanced
- *DecisionTree* (`sklearn.tree.DecisionTreeClassifier`)
Parameter: class weight: balanced

Probleme

Erfahrungen mit den Korpusklassen

Evaluation

- Cho, Han-Cheol; Okazaki, Naoaki; Miwa, Makoto; Tsujii, Jun'ichi (2013): Named entity recognition with multiple segment representations. In: Information Processing & Management 49
- Derczynski, Leon; Maynard, Diana; Rizzo, Giuseppe; van Erp, Marieke; Gorrell, Genevieve; Troncy, Raphaël et al. (2015): Analysis of named entity recognition and linking for tweets. In: Information Processing & Management 51 (2), S. 32–49.
- Konkol, Michal; Brychcín, Tomáš; Konopík, Miloslav (2015): Latent semantics in Named Entity Recognition. In: Expert Systems with Applications 42 (7), S. 3470–3479.
- Agerri, Rodrigo; Rigau, German (2016): Robust multilingual Named Entity Recognition with shallow semi-supervised features. In: Artificial Intelligence 238, S. 63–82.
- Erik F. Tjong Kim Sang and Fien De Meulder (2003): Language-Independent Named Entity Recognition.

- Marrero, Mónica; Urbano, Julián; Sánchez-Cuadrado, Sonia; Morato, Jorge; Gómez-Berbís, Juan Miguel (2013): Named Entity Recognition. Fallacies, challenges and opportunities. In: Computer Standards & Interfaces 35 (5), S. 482–489.
- Mayfield, James; McNamee, Paul; Piatko, Christine (2003): Named entity recognition using hundreds of thousands of features. In: Walter Daelemans und Miles Osborne (Hg.): Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -. the seventh conference. Edmonton, Canada. Morristown, NJ, USA: Association for Computational Linguistics, S. 184–187.
- Mónica Marrero, Sonia Sánchez-Cuadrado (2009): Evaluation of Named Entity Extraction Systems.
- Weischedel, Ralph M. (2013): OntoNotes release 5.0. [Philadelphia, Pa.]: Linguistic Data Consortium.