Named Entity Classification

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

January 8, 2017

Inhaltsverzeichnis

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

Tools

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

Korpus

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

Korpusreader

-

Korpusklassenbalancierung

-

Table: Klassen im OntoNotes Korpus

Klassen	Trainingset
ORG	4959
PERSON	2885
GPE	2802
NORP	1057
PERCENT	990
CARDINAL	1514
MONEY	1383
DATE	3474
FAC, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, TIME, QUANTITY, ORDINAL	> 1600

Beschreibung Korpusklassen

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

Verteilung Korpusklassen

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

Features für den Baseline-Klassifizierer

Table: Features für den Baseline-Klassifizierer

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

Erweitertes Featureset

Klassifizierertyp

Probleme

Erfahrungen mit den Korpusklassen

Evaluation

Ausblick

Referenzen