Named Entity Classification

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

January 20, 2017

Inhaltsverzeichnis

NER in der Forschung

Named Entity Recognition seit 1990er Jahren aktives Forschungsfeld. (Überblick: Borthwick, 1999, Tjong Kim Sang 2003, Marrero 2012)

Grundlage für weitere Forschungsfelder im Bereich Information Retrieval, z.B. Semantic Annotation, Question Answering, Opinion Mining, usw. (Marrero 2012)

Was sind Named Entities

Named Entities sind Phrasen die Namen von Personen, Organisationen, Währungen, usw enthalten. Beispiele:

[ORG U.N.], [PER Obama], [MONEY Dollar], [LOC Moscow]

Unser Projekt

- Typischerweise werden Named Entity Recognition und Named Entity Classification (NEC) zusammen betrachtet.
- Wenige Untersuchungen beschäftigen sich nur mit NEC. (Primadhanty 2014, He 2016, Spangler 2016)
- Dieses Projekt konzentriert sich auf NEC und stellt die Frage, welchen Einfluss Feature Selection auf die Klassifikationsergebnisse eines Nemed Entity Klassifizierers hat.
- Untersuchung konzentriert sich auf einfache syntaktische und lexikalische Features, die in fast allen Forschungsarbeiten in ähnlicher Form genutzt wurden. *(Toral, Munoz, 2006; Kazama, Torisawa, 2007; Ratinov, Roth 2009)*

Tools

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

Korpus

- Für Named Entity Klassifikation wird OntoNotes Korpus 2012 genutzt. *(OntoNotes Release 5.0 2012)*
- Englischen Nachrichtentexte des 'The Wall Street Journal'. Für die Entwicklungsphase bereits vorgefertigtes 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

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

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

Korpusklassenbalancierung

Table: Klassen im OntoNotes Korpus *(OntoNotes Release 5.0 2012)*

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

Verteilung Korpusklassen

- Zehn Klassen enthalten nur wenige NE-Instanzen. Diese werden aus dem balancierten Korpus entfernt.
- Semantisch ähnliche Klassen NORP und GPE werden zusammengefasst.
- Numerische Klassen MONEY, PERCENT und CARDINAI werden ebenfalls zusammengefasst.

Beschreibung neuer 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 nach Balancierung

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

Anzahl der Features: 1317

Table: Features für den Baseline-Klassifizierer

Feature	Wert	Beschreibung
Unigram	numerisch	Vorkommen der Unigramme, die mindestens fünfmal im Trainingscorpus vorkommen. *(Mayfield 2003)*

Erweitertes Featureset I

Anzahl der Features: 1716

Table: Features für den Klassifizierer I

Feature	Wert	Beschreibung
Unigram	numerisch	Vorkommen der Unigramme (lemmatisiert), die mindestens fünfmal im Trainingscorpus vorkommen. *(Mayfield 2003)*
POS	numerisch	Häufigkeit von 36 POS-Tags aus der Penn Treebank *(Florian, Chieu 2003)*
isAllCaps	boolean	Wörter nur in Großschreibung *(Nadeau 2006)*
Context	numerisch	Vorkommen der Kontexttokens, die mindestens fünfmal im Trainingskorpus vorkommen. Das Kontextfenster beinhaltet das Vorgänger- und Nachfolgetoken der NE. *(Munro 2003)*

Erweitertes Featureset II

Table: Features für den Klassifizierer II

Feature	Wert	Beschreibung
isInWiki	boolean	Vorkommen der NE in der Wikipedia. *(Toral and Munoz, 2006; Kazama and Torisawa, 2007)*
isTitle	boolean	Prüft, ob Titelbezeichnungen (z.B. Mr. MA) vorkommen. *Ratinov, Roth 2009*
isNP	boolean	Ist NE eine Nominalphrase. *Sánchez, Cuadrado 2009*
isName	boolean	Prüft, ob Vornamen vorkommen.*Ratinov, Roth 2009*
containsDash	boolean	Vorkommen von Viertelgeviertstrichen. *Mayfield 2003*

Klassifizierertyp

Zur Klassifizierung der NE wird eine Support Vector Maschine mit linearem Kernel verwendet.

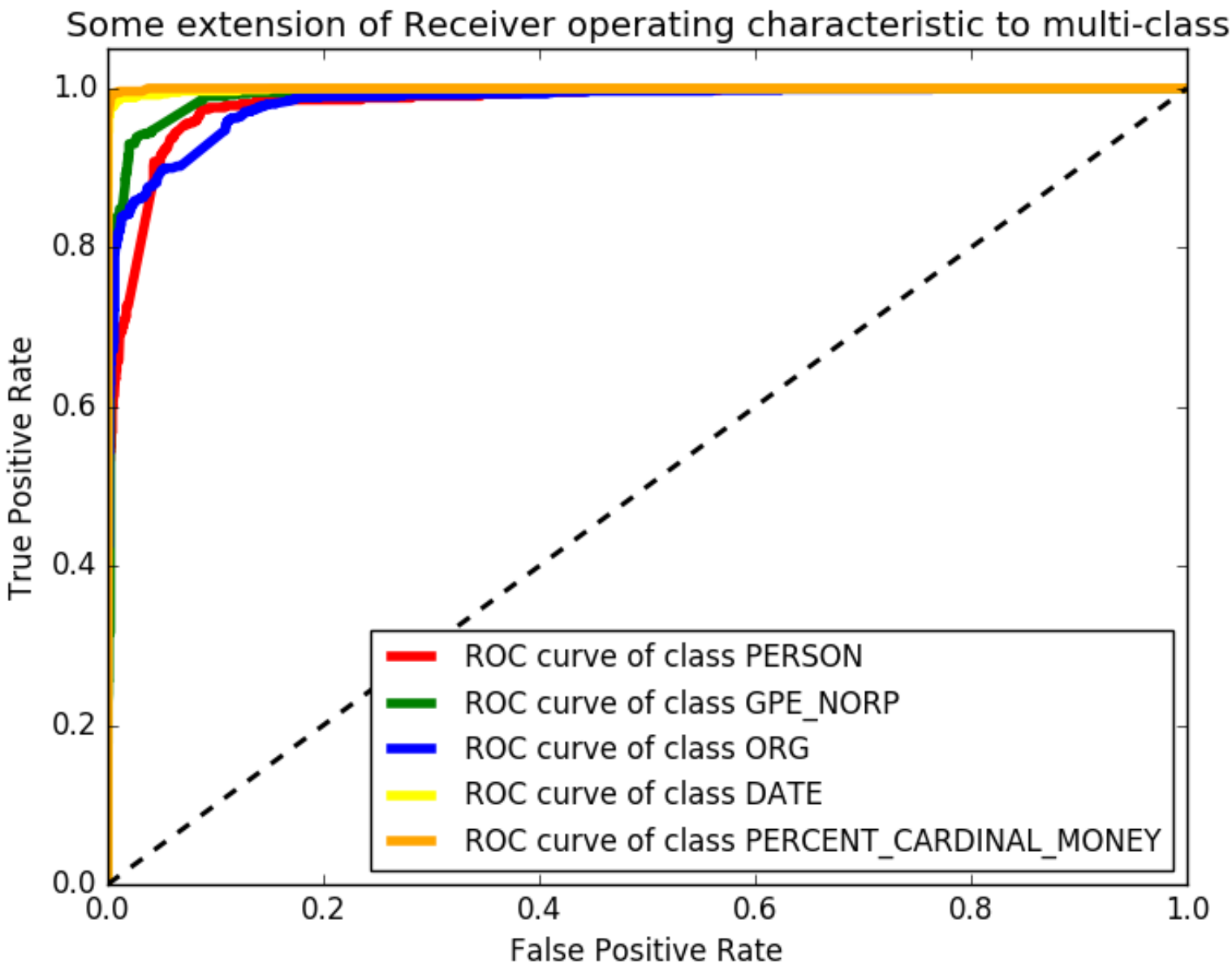
Alternativ wurde ein Decisiontree getestet, dieser hatte allerdings mit allen Featurekombinationen tendenziell schlechtere Evaluationsergebnisse. Zudem trainiert der SVM deutlich schneller.

- SVM (sklearn.svm.LinearSVC)
  Featurevektoren haben sehr viele Features daher linearer Kernel. Mapping in höheren Featurespace eines nicht-linearen Kernels bringt kaum Klassifizierungsverbesserungen. *(Chih-Wei Hsu 2003)*

Erfahrungen mit den Korpusklassen I

Wie die ROC-Kurve zeigt, hat der Klassifizierer insbesondere Schwierigkeiten, die Klassen PERSON und ORG und GPE_NORP zu unterscheiden.

ROC Curve


Some extension of Receiver operating characteristic to multi-class
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class PERSON
ROC curve of class GPE_NORP
ROC curve of class ORG
ROC curve of class DATE
ROC curve of class PERCENT_CARDINAL_MONEY

Erfahrungen mit den Korpusklassen II

Table: Confusion Matrix

361	28	22	2	0	PERSON
29	549	10	0	0	GPE_NORP
71	45	736	6	1	ORG
0	2	1	591	7	DATE
0	2	0	2	525	PERCENT_CARDINAL_MONEY

Beobachtungen zur Confusion-Matrix

- Von 361 PERSON Entities, werden 71 als ORG und 29 als GPE_NORP klassifiziert.
- Numerische Klassen und Datum werden fast zu 100% erkannt.

Featureselektion I

Insgesamt haben werden elf Features eingesetzt.

Um die Performance der einzelnen Features zu testen, wurde die Potenzmenge des Featuresets gebildet.

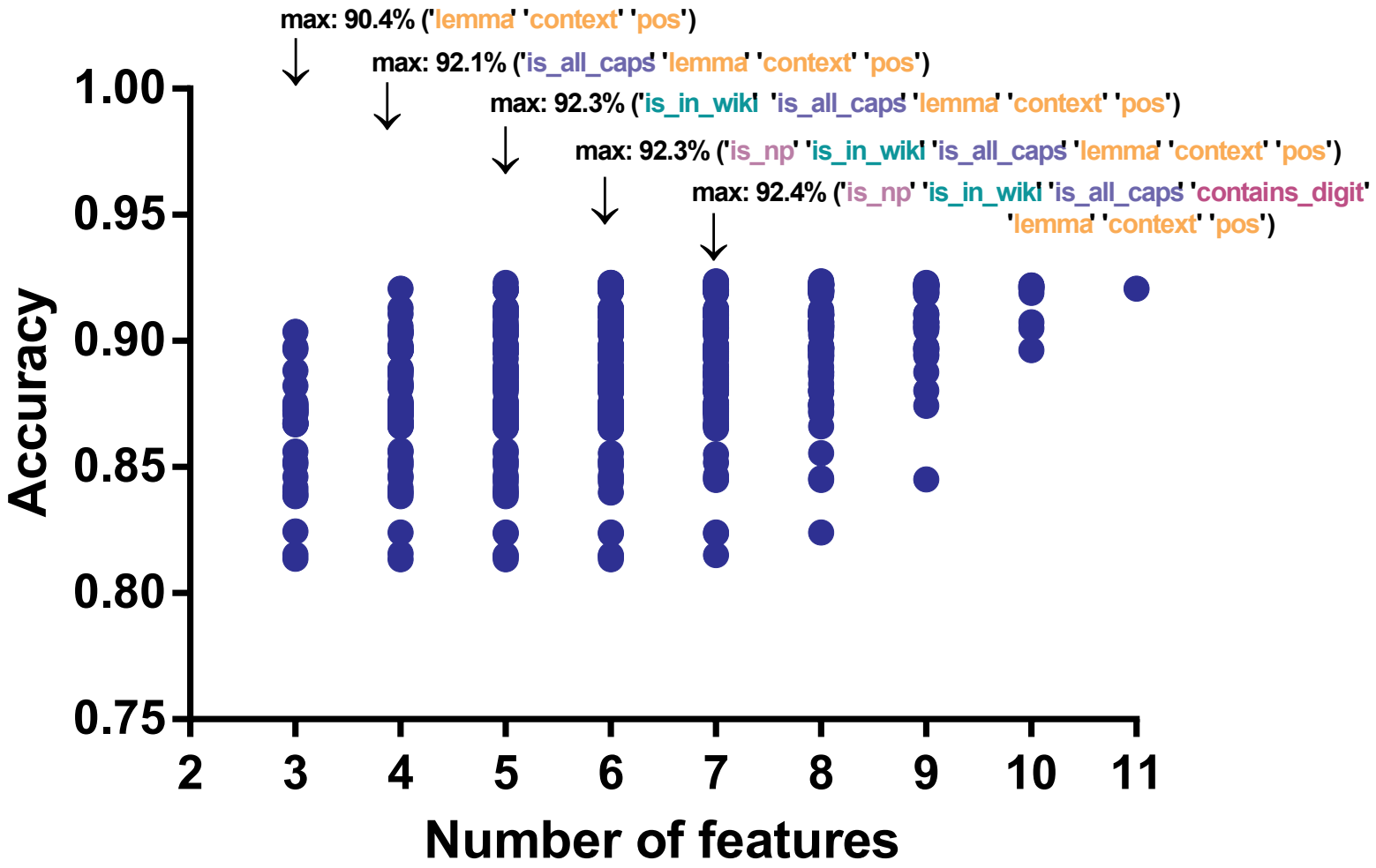
Schließlich wurde der Klassifizierer auf allen 1013 Teilmengen durchgeführt.

Featureselektion II

Für die Evaluation entscheidend waren alle Teilmengen, die die Features 'Unigram' und 'Context' enthalten und mind. drei Features 3 besitzen.

- Accuracy aller Teilnmengen ohne diese Features: $<$69 %.
- Accuracy nur mit Unigram und Context: 87.42%

Featureselektion III



Featureselektion IV

- Beste Features: 'Unigram', 'Context'
- Höchste Accuracy: Ab sieben Features. Ab vier Features kaum mehr Verbesserung der Accuracy
- Features, die zur Erhöhung der Accuracy beitragen: 'POS', 'is_all_caps', 'is_in_wiki', 'is_np', 'contains_digit'
- Das Featureset aus vier Features: 'Unigram', 'Context', 'POS', 'is_all_caps' erreicht die beste Accuracy bei möglichts kleinem Featureset.

Evaluation

Evaluationsergebnisse der Baseline im Vergleich mit optimalen Testset

Table: Final Evaluation

Featureset		Accuracy	F1-Score
Baseline	unbalanced	0.786757848928	0.786781978815
	balanced	0.840802675585	0.854079560307
Optimales Featureset	unbalanced	0.87286291576	0.869301986664
	balanced	0.923745819398	0.924314367424

Probleme

Context bezieht auch Satzzeichen ein (oft ',' oder '.'), dies könnte man auf alphanumerische Strings beschränken.

Klassifikationsfehler im Testset, da nur die automatisch annotierte Testsetversion von OntoNotes v5 zur Verfügung steht.

Beispiel:
{'ORG': [['American', 'JJ'], 'NP', ('to', 'notions')]}
classified as ['GPE_NORP']

Ausblick

Was kann man noch verbessern?

Referenzen

- 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.