Named Entity Classification

M. Huvar, Ph. Richter-Pechanski, S. Safdel

February 25, 2017

Inhaltsverzeichnis

NER in der Forschung

Named Entity Recognition seit 1990er Jahren aktives Forschungsfeld.
(Überblick: Borthwick, 1999, Tjong Kim Sang 2003, Marrero 2013)

Grundlage für weitere Forschungsfelder im Bereich Information Retrieval, z.B. Semantic Annotation, Question Answering, Opinion Mining, usw.
(Marrero 2013)

Was sind Named Entities?

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

Beispiele für Named Entities

The Speaker of the [ORG U.N.] ..
President [PER Obama] ...
The price of the [MONEY Dollar] lost ...
[LOC Moscow] is the capital of Russia.

Unser Projekt

- Typischerweise werden Named Entity Recognition und Named Entity Classification (NEC) zusammen betrachtet.

Unser Projekt

- Typischerweise werden Named Entity Recognition und Named Entity Classification (NEC) zusammen betrachtet.
- Wenige Untersuchungen beschäftigen sich nur mit NEC. (Primadhanty, Carreras 2014, He, Spangler 2016)

Unser Projekt

- Typischerweise werden Named Entity Recognition und Named Entity Classification (NEC) zusammen betrachtet.
- Wenige Untersuchungen beschäftigen sich nur mit NEC. (Primadhanty, Carreras 2014, He, Spangler 2016)
- Dieses Projekt konzentriert sich auf NEC und stellt die Frage, **welchen Einfluss Feature Selection auf die Klassifikationsergebnisse eines Named Entity Klassifizierers hat**.

Unser Projekt

- Typischerweise werden Named Entity Recognition und Named Entity Classification (NEC) zusammen betrachtet.
- Wenige Untersuchungen beschäftigen sich nur mit NEC. (Primadhanty, Carreras 2014, He, Spangler 2016)
- Dieses Projekt konzentriert sich auf NEC und stellt die Frage, **welchen Einfluss Feature Selection auf die Klassifikationsergebnisse eines Named Entity Klassifizierers hat**.
- Nutzung einfacher syntaktischer und lexikalischer 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
- GitHub
- ICL-Wiki

Korpus

- Für Named Entity Klassifikation wird OntoNotes Korpus 2012 genutzt. *(OntoNotes Release 5.0 2012)*
- Englische Nachrichtentexte des 'The Wall Street Journal'. Für die Entwicklungsphase bereits vorgefertigtes Developmentset.
- Für die Klassifikation der Named Entities werden die bereits vorgefertigten Trainings- und Testdatensets aus dem Goldstandard genutzt.

Table: Anzahl an 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, Kontextwörter (ne-1, ne+1), und ordnet ihnen Klassen zu.
- Beispielextraktion aus dem Satz:
  Says Peter Mokaba, President of the South African Youth Congress: "We will ...

Extrahierte Instanz der Korpusreader-Klasse

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

Korpusklassenbalancierung

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

Klassen	Trainingset
ORG	5788
DATE	4080
PERSON	3756
GPE	3601
CARDINAL	1852
MONEY	1509
NORP	1484
PERCENT	1061
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 CARDINAL 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
GPE_NORP	732	5134	588
PERCENT_CARDINAL_MONEY	564	4672	529
DATE	613	4254	601
PERSON	486	3759	413

Beispielinstanz

Beispielinstanz zur Veranschaulichung der Features

[['North', 'NNP'],['-', HYPH], ['America', 'NNP'], 'NP', (',', 'and')]

Features für den Baseline-Klassifizierer

Anzahl der Features: 1317

Table: Features für den Baseline-Klassifizierer

Feature	Wert	Beschreibung
Unigram	numerisch	Vorkommenshäufigkeit der Unigramme (lemmatisiert) in der NE, die mindestens fünfmal im Trainingscorpus vorkommen. *(Mayfield, McNamee 2003)* (america : 1, north : 1)

Erweitertes Featureset I

Anzahl der Features: 1716

Table: Features für den Klassifizierer I

Feature	Wert	Beschreibung
Unigram	numerisch	Häufigkeit der Unigramme (lemmatisiert), die mindestens fünfmal im Trainingscorpus vorkommen. *(Mayfield, McNamee 2003)* (america : 1, north : 1)
POS	numerisch	Häufigkeit von 36 POS-Tags aus der Penn Treebank *(Chieu 2003)* NNP: '2'
isAllCaps	boolean	Wörter nur in Großschreibung *(Nadenau, Turney 2006)* (0)
Context	numerisch	Häufigkeit der Kontexttokens. Beinhaltet Vorgänger- und Nachfolgetoken der NE. *(Munro, Ler 2003)* (,_and : 1)
containsDigit	boolean	Vorkommen von Nummern. (0)

Erweitertes Featureset II

Table: Features für den Klassifizierer II

Feature	Wert	Beschreibung (Beispielwert)
isInWiki	boolean	Vorkommen der NE in der Wikipedia. *(Toral, Munoz 2006)* (1)
isTitle	boolean	Prüft, ob Titelbezeichnungen (z.B. Mr., MA) vorkommen. *(Ratinov, Roth 2009)* (0)
isNP	boolean	Ist NE eine Nominalphrase. *(Sánchez, Cuadrado 2009)* (1)
isName	boolean	Prüft, ob Vornamen vorkommen.*(Ratinov, Roth 2009)* (0)
containsDash	boolean	Vorkommen von Viertelgeviertstrichen. *(Mayfield, McNamee 2003)* (1)
isComName	boolean	Prüft auf kommerzielle Bezeichner (Corp., Inc.) (0)

Klassifizierertyp

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

Klassifizierertyp

- Zur Klassifizierung der NE wird eine Support Vector Maschine mit linearem Kernel verwendet.
- SVM (sklearn.svm.LinearSVC(loss='squared_hinge', penalty='l2'))

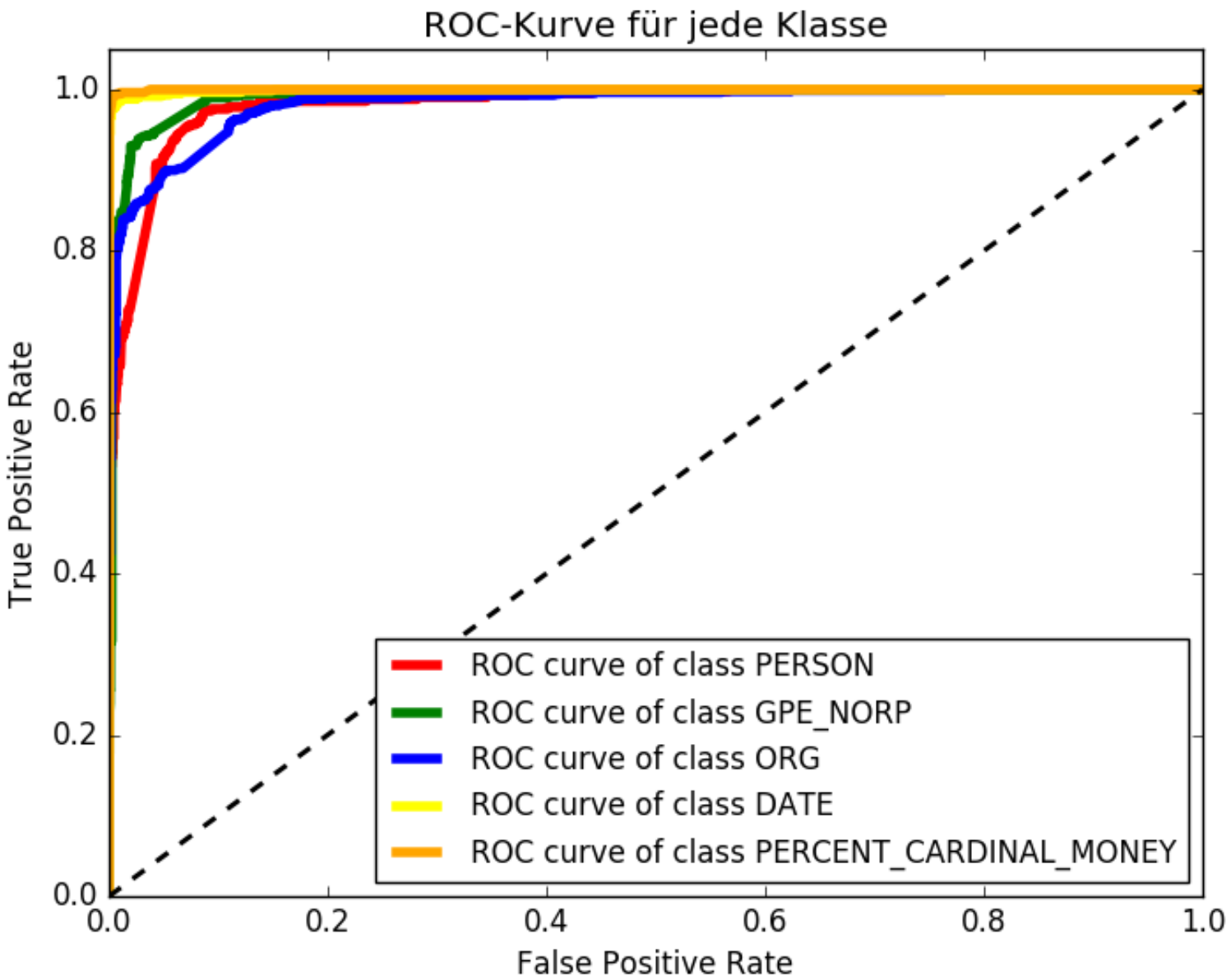
Klassifizierertyp

- Zur Klassifizierung der NE wird eine Support Vector Maschine mit linearem Kernel verwendet.
- SVM (sklearn.svm.LinearSVC(loss='squared_hinge', penalty='l2'))
- Featurevektoren haben sehr viele Features daher linearer Kernel. Mapping in höheren Featurespace eines nicht-linearen Kernels bringt kaum Klassifizierungsverbesserungen. *(Chih-Wei Hsu 2003)*

Klassifizierertyp

- Zur Klassifizierung der NE wird eine Support Vector Maschine mit linearem Kernel verwendet.
- SVM (sklearn.svm.LinearSVC(loss='squared_hinge', penalty='l2'))
- Featurevektoren haben sehr viele Features daher linearer Kernel. Mapping in höheren Featurespace eines nicht-linearen Kernels bringt kaum Klassifizierungsverbesserungen. *(Chih-Wei Hsu 2003)*
- *Alternativ wurde ein Decisiontree getestet, dieser hatte allerdings mit allen Featurekombinationen tendenziell schlechtere Evaluationsergebnisse. Zudem trainiert der SVM deutlich schneller.*

ROC Curve


ROC-Kurve für jede Klasse
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

Featureselektion I

- Insgesamt wurden elf Features eingesetzt.

Featureselektion I

- Insgesamt wurden elf Features eingesetzt.
- Um die Performance der einzelnen Features zu testen, wurde die Potenzmenge des Featuresets gebildet.

Featureselektion I

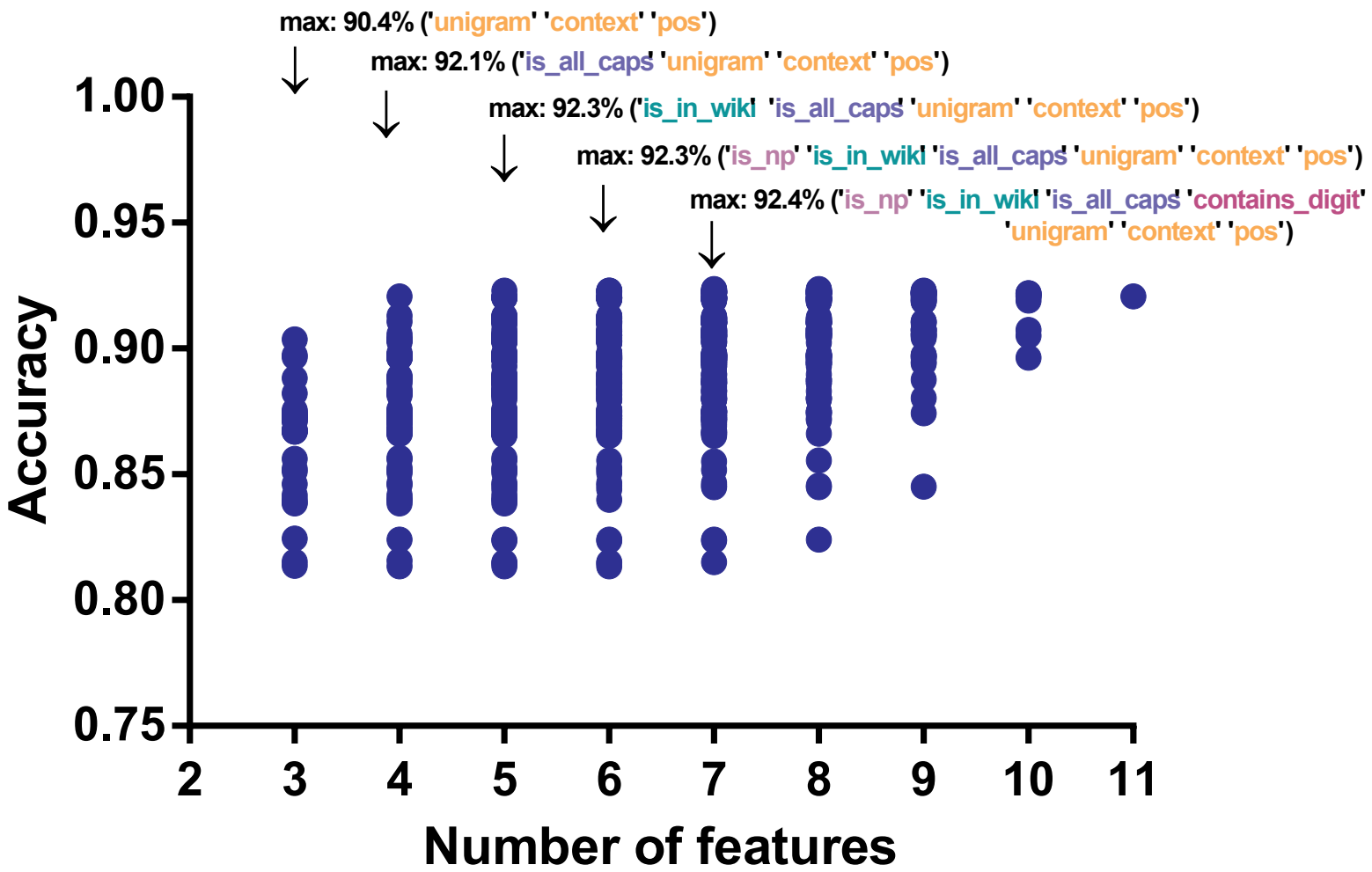
- Insgesamt wurden elf Features eingesetzt.
- Um die Performance der einzelnen Features zu testen, wurde die Potenzmenge des Featuresets gebildet.
- Schließlich wurde der Klassifizierer auf allen 1013 Teilmengen durchgeführt.

Featureselektion II

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

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

Featureselektion III



Featureselektion IV

- Höchste Accuracy: Ab sieben Features. Ab vier Features kaum mehr Verbesserung der Accuracy

Featureselektion IV

- 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'

Featureselektion IV

- 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 optimalem Featureset.

Table: Final Evaluation

Featureset		Accuracy
Baseline	unbalanced	0.7867
'unigram'	balanced	0.8408
Optimales Featureset	unbalanced	0.8728
'pos', 'is_all_caps', 'is_in_wiki', 'is_np', 'contains_digit', 'unigram', 'context'	balanced	0.9237

Probleme und Lösungsvorschläge

- Context bezieht auch Satzzeichen ein (oft ',' oder '.'), dies könnte man auf alphanumerische Strings beschränken.
- Verbesserung bei PERSON-Klassifizierung möglicherweise durch Generierung weiterer PERSON-Instanzen.
- Klassifikationsfehler im Testset, da nur die automatisch annotierte Testsetversion von OntoNotes 5.0 zur Verfügung steht.

Beispiel für "falsch" klassifizierte Instanz

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

Zusammenfassung

- Mehr Features bieten nicht zwangsläufig bessere Evaluationsergebnisse.
- Die Dimensionalität der Features, scheint Einfluss auf Klassifikationsergebnisse zu haben.
- Hochdimensionale Features, wie Unigram und Context, tragen maßgeblich zu besseren Klassifikationsergebnissen bei.
- Semantische Zusammenfassung von Klassen zur besseren Balancierung verbessern die Ergebnisse.

Referenzen

- Borthwick, A. (1999): A Maximum Entropy Approach to Named Entity Recognition, Diss., New York.
- Chieu H. (2003): Named Entity Recognition with a Maximum Entropy Approach. In Proceedings of CoNLL-2003.
- H. Chih-Wei (2003): A Practical Guide to Support Vector Classification.
- He, Q.; Spangler, S. (2016): Semi-supervised data integration model for named entity classification. Google Patents.
- J. Kazama, K. Torisawa (2007): Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. In: Proceedings of ACL-08.
- Marrero, M.; Urbano, J. (2013): Named Entity Recognition. Fallacies, challenges and opportunities. In: Computer Standards & Interfaces 35 (5).
- Marrero, M.; Sánchez-Cuadrado, S. (2009): Evaluation of Named Entity Extraction Systems. In: Advances in Computational Linguistics. Research in Computing Science.
- Mayfield, J.; McNamee, P. (2003): Named entity recognition using hundreds of thousands of features. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003.
- Munro, R.; Ler, D. (2003): Meta-Learning Orthographic and Contextual Models for Language Independent Named Entity Recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL.
- Nadeau, D.; Turney, P. (2006): Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In: Proceedings of the 19th international conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence.
- Primadhanty, A.; Carreras, X. (2014): Low-Rank Regularization for Sparse Conjunctive Feature Spaces: An Application to Named Entity Classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.

- Ratinov, L.; Roth, D. (2009): Design Challenges and Misconceptions in Named Entity Recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).
- Tjong Kim Sang,E.; De Meulder, F. (2003): Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2003.
- Toral, A.; Munoz, R. (2006): A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia.
- Weischedel, R. (2013): OntoNotes release 5.0. [Philadelphia, Pa.]: Linguistic Data Consortium.