Named Entity Classification

Madita Huvar, Phillip Richter-Pechanski, Sanaz Safdel

27. Februar 2017

Inhaltsverzeichnis

1 Einführung **3**
1.1 Was sind Named Entities? 3

2 Unser Projekt **3**
2.1 Ziel 3
2.2 Organisation 3

3 Tools **3**

4 Korpus **4**

5 Korpusreader **4**

6 Korpusklassenbalancierung **4**

7 Baselineklassifizierer **4**

8 Erweitertes Featureset **4**

9 Klassifizierertyp **5**

10 Featureselection **5**

11 Evaluation **5**

12 Probleme und Lösungsvorschläge **5**

13 Zusammenfassung **6**

14 Anhang **7**

1 Einführung

Named Entity Recognition ist seit den 1990er Jahren ein aktives Forschungsfeld. Es stellt die Grundlage für weitere Forschungsfelder dar. Zum Beispiel im Bereich Semantic Annotation, Question Answering, Opinion Mining und viele mehr.

1.1 Was sind Named Entities?

Named Entities sind Phrasen, die Namen von Personen, Organisationen, Währungen, usw. enthalten. In unserem Projekt werden sie wie folgt dargestellt:

2 Unser Projekt

2.1 Ziel

Typischerweise werden Named Entity Recognition und Named Entity Classification zusammen betrachtet und nur wenige Untersuchungen beschäftigen sich nur mit Named Entity Classification. In unserem Projekt konzentrieren wir uns auf Named Entity Classification und stellen vor allem die Frage, welchen Einfluss Feature Selection auf die Klassifikationsergebnisse eines Named Entity Klassifizierers haben. Wir verwenden hierzu einfache syntaktische und lexikalische Features, die in fast allen Forschungsarbeiten in ähnlicher Form genutzt werden.

2.2 Organisation

Zur Organisation unseres Projekts benutzen wir GitHub. Zusätzlich haben wir eine WhatsApp Gruppe, in der wir uns besprechen und unsere Treffen planen können. Diese finden ca. ein bis zwei Mal pro Woche statt. Unsere Ergebnisse halten wir auf einer speziell angelegten Wikiseite fest.

3 Tools

Für unser Projekt verwenden wir folgende Tools:

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

- GitHub
- ICL-Wiki

4 Korpus

Wir verwenden für unser Projekt das OntoNotes Korpus 2012. In diesem sind englische Nachrichtentexte des 'The Wall Street Journal' enthalten. Es existieren bereits ein Development-/Trainings- und Testset. Siehe Tabelle 1.

5 Korpusreader

Für die Extraktion der Named Entities haben wir einen Korpusreader erstellt. Dieser Reader extrahiert alle Named Entities, inklusive POS-Tags der einzelnen Token, Phrasenart, Kontextwörter (ne-1, ne+1) und ordnet ihnen Klassen zu.

6 Korpusklassenbalancierung

Die Anzahl der Named Entites im Trainingsset ist für die einzelnen Klassen sehr unterschiedlich, wie in Tabelle 2 zu sehen. Die zehn Klassen mit der geringsten Anzahl Named Entities, werden aus dem balancierten Korpus entfernt. Die Klassen NORP und GPE werden zusammengefasst, da sie semantisch ähnlich sind. Die Klassen MONEY, PERCENT und CARDINAL werden ebenfalls zusammengefasst, da sie alle numerische Klassen sind. Daraus ergeben sich folgende neue Korpusklassen, zu sehen in Tabelle 3. Die Verteilung der Named Entities auf die Klassen ist nun wesentlich ausgeglichener, wie in Tabelle 4 zu sehen.

7 Baselineklassifizierer

Für unseren Baselineklassifizierer verwenden wir nur das Feature 'Unigram', welches die Vorkommenshäufigkeit, der Unigramme in der Named Entity,welche mindestens fünfmal im Trainigskorpus vorkommen, beschreibt.

8 Erweitertes Featureset

Für unseren Named Entity Klassifizierer, haben wir weitere Features hinzugefügt. Diese sind in Tabelle 5 beschrieben. Insgesamt hatten wir eine Anzahl von 1716 Features.

9 Klassifizierertyp

Zur KLassifizierung der Named Entities wir eine Support Vector Maschine mit linearem Kernel verwendet. SVM XXXXXXFOLIEN Unsere Featurevektoren haben sehr viele Features, daher verwenden wir den linearen Kernel. Mapping in höheren Featurespace eines nicht-linearen Kernels bringt kaum Klassifiziereungsverbesserungen. Alernativ haben wir Decisiontree getestet, dieser hatte allerdings mit allen Featurekombinationen tendenziell schlechtere Evaluationsergebnisse. Zudem trainiert der SVM deutlich schneller. Dies wird deutlich in Abbildung 1 und Abbildung **??**

10 Featureselection

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. Für die Evaluation entscheidend waren alle Teilmengen, die die Featured Unigram und Context enthalten und mindestens drei Features besitzen. Die Accuracy ohne diese Features lag nur bei ¡69. Wohingegen die Accuracy nur mit den Features Unigram und Context bei 87.42 lag. Wie in Abbildung 2 zu sehen erreichen wir die höchste Accuracy ab sieben Features. Jedoch schon ab vier Features gibt es kaum noch Verbesserungen der Accuracy. Die Features, die zur Erhöhung der Accuracy beitragen sind: POS, isallcaps, isinwiki, isnp und containsdigit. Die beste Accuracy bei möglichst kleinem Featureset erreichen wir mit Unigram, Context, POS, und isallcaps.

11 Evaluation

Tabelle 6

12 Probleme und Lösungsvorschläge

Ein mögliches Problem ist, dass das OntoNotes Korpus bereits automatisch annotiert ist und dadurch bereits vor unserer Verarbeitung Klassifikationsfehler im Testset vorhanden sind. Um dieses Problem zu beheben, müsste man das Korpus Handannotieren, was jedoch sehr aufwändig ist. Die Klassifizierung der Klasse PERSON könnte möglicherweise durch die Generierung weiterer PERSON-Instanzen verbessert werden. Ein weiterer Punkt, den man genauer betrachten könnte, ist das Feature 'Context', dieses bezieht um Moment auch Satzzeichen mit ein. Man könnte testen, ob eine Verbesserung erzielt wird, wenn man das Feature auf alphanumerische Strings beschränkt.

13 Zusammenfassung

In unserem Projekt haben wir herausgefunden, dass mehr Features nicht zwangsläufig bessere Ergebnisse liefern. Außerdem scheint die Dimensionalität der Features Einfluss auf die Klassifikationsergebnisse zu haben. Hochdimensionale Features wie 'Unigram' oder 'Context' tragen maßgeblich zu besseren Ergebnisse bei. Auch das Zusammenfassen von Klassen, die sich ähneln verbessert die Ergebnisse.

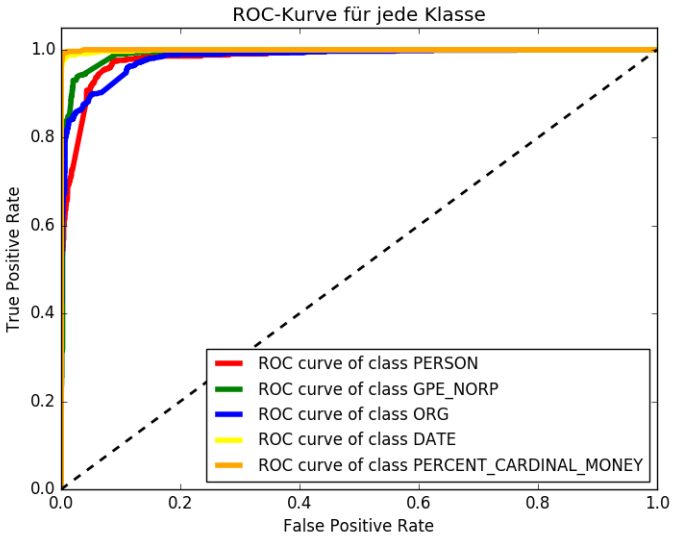

Abbildung 1: ROC Curve

Developmentset	Trainingset	Testset
3325	23686	2996

Tabelle 1: Anzahl an Named Entities

14 Anhang

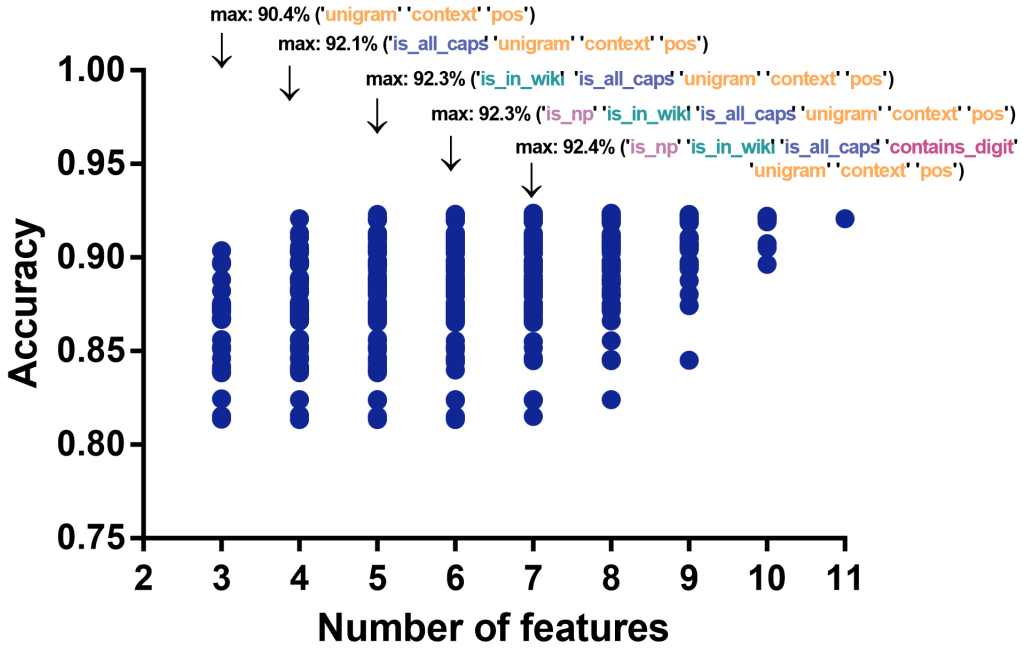

Abbildung 2: ???

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

Tabelle 2: Klassen im OntoNotes Korpus *(OntoNotes Release 5.0 2012)*

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

Tabelle 3: Balancierte Klassen

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

Tabelle 4: Verteilung der Klassen nach Balancierung

Feature	Wert	Beschreibung
Unigram	numerisch	Häufigkeit der Unigramme (lemmatisiert), die mindestens fünfmal im Trainingscorpus vorkommen. *(Mayfield, McNamee 2003)*
POS	numerisch	Häufigkeit von 36 POS-Tags aus der Penn Treebank *(Chieu 2003)*
isAllCaps	boolean	Wörter nur in Großschreibung *(Nadenau, Turney 2006)*
Context	numerisch	Häufigkeit der Kontexttokens. Beinhaltet Vorgänger- und Nachfolgetoken der NE. *(Munro, Ler 2003)*
containsDigit	boolean	Vorkommen von Nummern.
isInWiki	boolean	Vorkommen der NE in der Wikipedia. *(Toral, Munoz 2006)*
isTitle	boolean	Prüft, ob Titelbezeichnungen (z.B. Mr., MA) vorkommen. *(Ratinov, Roth 2009)* (0)
isNP	boolean	Ist NE eine Nominalphrase. *(Sánchez, Cuadrado 2009)*
isName	boolean	Prüft, ob Vornamen vorkommen.*(Ratinov, Roth 2009)*
containsDash	boolean	Vorkommen von Viertelgeviertstrichen. *(Mayfield, McNamee 2003)*
isComName	boolean	Prüft auf kommerzielle Bezeichner (Corp., Inc.)

Tabelle 5: Features für den Klassifizierer

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

Tabelle 6: Final Evaluation