Proseminar Maschinelles Lernen und Experimentelles Design, UdS, SS11 Hausaufgaben für 16.6.2011

Caroline Sporleder

June 8, 2011

Datenset Unter `http://www.coli.uni-saarland.de/~csporled/feats.deu.testa.undersampled.arff` findet ihr ein Datenset zum Thema 'Named Entity Recognition' (NER). Ladet dieses Datenset herunter und speichert es als Textdatei (d.h. nicht html etc.).

Dieses Datenset wurde aus den CoNLL-2003 NER Shared Task Daten generiert. Die CoNLL-Daten bestanden aus einem laufenden (deutschen) Text, in dem Wörter (genauer "Tokens") mit Named Entity (NE) Klassen annotiert wurden. Es wurden jeweils vier NE Klassen und eine nicht-NE Klasse unterschieden:

- Organisation (ORG)
- Location (LOC)
- Person (PER)
- Miscellaneous (MISC): NE-Klasse anderen Typs
- Other (O): keine NE

Um bei aufeinanderfolgenden NEs gleichen Typs zwischen den NEs unterscheiden zu können, wurden an die NE Klassen ferner "I" oder "B" angehängt, wobei "I" der Default ist und "B" verwendet wird für das erste Token/Wort einer NE, die auf eine NE gleichen Typs folgt (Beispiel 1). (Diese sogenannte BIO-Annotation wird oft in Sequenzlabelling-Aufgaben verwendet.)

(1) Die_O Weltmeisterschaft_O im_O westfälischen_I-LOC Riesenbeck_B-LOC ...

Aus den so annotierten Daten wurden die Instanzen in `feats.deu.testa.undersampled.arff` generiert. Jede Instanz besteht aus einem Attribut-Wert-Vektor für ein Token im

Text (dem Zieltoken) und der dazugehörigen Ausgabeklasse (O, I-LOC, B-LOC, I-PER etc.). Die Attribute sind die folgenden:

1. lemma-left: das Lemma links neben dem Zieltoken.
2. pos-left: der Pos-Tag des Lemmas links neben dem Zieltoken.
3. chunk-left: der Chunk, zu dem das Lemma links neben dem Zieltoken gehört. (Chunking ist eine Art von flachem Parsing und identifiziert NPs, PPs und z.T. VPs.)
4. ne-left: der NE-Tag des Lemmas links neben dem Zieltoken.
5. lemma-right: das Lemma rechts neben dem Zieltoken.
6. pos-right: der Pos-Tag des Lemmas rechts neben dem Zieltoken.
7. chunk-right: der Chunk, zu dem das Lemma rechts neben dem Zieltoken gehört.
8. ne-right: der NE-Tag des Lemmas rechts neben dem Zieltoken.
9. wort: das Zieltoken.
10. lemma: das Lemma des Zieltokens.
11. pos: das Pos-Tag des Zieltokens.
12. chunk: der Chunk, zu dem das Zieltoken gehört.

Der Attribut-Wert-Vektor für *westfälischen* in Beispiel (1) sieh z.B. so aus (dabei APPART=POS:Präposition, I-PC=Chunk:PP, NE=POS:NE, I-NC=Chunk:NP, ADJA=POS:Adj, I-NC=Chunk:NP):

(2) <im,APPRART,I-PC,O,Riesenbeck,NE,I-NC,B-LOC,westfälischen,westfälisch,ADJA, I-NC>

Aufgabe 1: Named Entity Recognition (NER) mit J48 Generiert mit Wekas J48 Classifier einen Decision Tree für das Datenset. Ihr könnt entweder einen 90-10 Trainings-Test-Split nehmen oder 10-fold Cross-Validation. Werft einen Blick auf den generierten Entscheidungsbaum. Sieht er plausibel aus (d.h. befinden sich Attribute, von denen ihr denkt, dass sie wichtig sind, in der Nähe der Wurzel des Baums)? Für welche Ausgabeklassen ist die Performanz am besten, für welche am schlechtesten? Warum?

Aufgabe 2: Auswahl der Attribute Welche Attribute, denkt ihr, sollten bei der Klassifikation helfen, welche eher nicht? Könnt ihr die Performanz verbessern, indem ihr unwichtig Attibute ausschaltet? Mit welchen Attributen bekommt ihr die besten Ergebnisse?