LC-CRF-Wortart-Tagger 1: Training

Laden Sie Trainingsdaten für das Wortart-Tagging von `www.cis.lmu.de/~schmid/lehre/Experimente/data/Tiger.zip` herunter und dekomprimieren Sie die Daten. Jede Zeile enthält ein Wort und ein Wortart-Tag. Auf jeden Satz folgt eine Leerzeile. Die Daten sind in Trainingsdaten, Testdaten und Development-Daten aufgeteilt.

Fügen Sie beim Einlesen zu den Tags jedes Satzes ein Start-Tag und ein Ende-Tag hinzu. Fügen Sie außerdem zu jeder Tokenfolge leere Strings am Anfang und Ende als Start- und Ende-Tokens hinzu.

Trainieren Sie auf den Daten einen LC-CRF-Tagger.

Aufruf: **crf-train.py train.txt param-file**

Verwenden Sie bei der Implementierung eine Klasse CRFTagger. Nehmen Sie für den Gewichtsvektor ein defaultdict(float), welches Merkmalsnamen wie "`WT-Haus+NN`" auf eine Zahl (das Gewicht) abbildet. Merkmalsvektoren repräsentieren Sie als eine Liste von Merkmalsnamen.

Das Training ist sehr langsam (etwa 2 Sätze pro Sekunde). Sie werden den Code in der nächsten Woche noch effizienter machen.

Vorüberlegungen

- Welche Merkmale verwenden Sie am besten?
- Welche Teilaufgaben umfasst das Training?
- Welche Datenstrukturen verwenden Sie?
- Was speichern Sie in der Parameterdatei?
- Wie vermeiden Sie Underflow?

Schicken Sie das fertige Programm an `schmid@cis.lmu.de`.