Maschinelles Übersetzen natürlicher Sprachen

6. Übungsblatt

2016-12-01

Die vorliegende Aufgabe soll etwas "praxisorientierter" als die bisherigen sein. Deshalb wollen wir uns näher mit zwei Implementierungen von PCFG-Sprachmodellen beschäftigen, dem *Stanford Parser* sowie dem *Berkeley Parser*.

Die Übung findet im **Rechnerkabinett E042** statt!

Stanford Parser (Web-Interface): `http://nlp.stanford.edu:8080/parser/`

Berkeley Parser: `http://code.google.com/p/berkeleyparser/`

Aufgabe 1 (NLTK)
Installieren Sie nltk. Laden Sie, wie hier unter `http://www.nltk.org/data.html` beschrieben, über NLTK einen kleinen, frei verfügbaren Teil der Penn Treebank [MSM93] herunter. (Das entsprechende Paket heißt `treebank`.) Untersuchen Sie die Bestandteile der Treebank. Nutzen Sie das bereitgestellte Python Skript um einige Bäume zu visualisieren.

Aufgabe 2 (Einführung: Stanford-Parser)
Nutzen Sie den Stanford-Parser zum Parsen einiger kurzer Beispielsätze aus dem Satzkorpus unten. Interpretieren Sie die ausgegebenen Parsebäume!

Aufgabe 3 (Berkeley-Parser)
Laden Sie nun den Berkeley-Parser (sowie die PCFGs für Englisch und Deutsch) herunter und machen Sie sich mit seiner Bedienung vertraut. Der Aufruf

```
java -jar BerkeleyParser-1.7.jar
```

gibt eine kurze Übersicht über die möglichen Kommandozeilenparameter.

- Vergleichen Sie die Ausgabe des Berkeley-Parsers mit der des Stanford-Parsers.
- Der Parser bietet es an, die Eingabesätze vor der Verarbeitung zu *tokenisieren*. Machen Sie sich die Bedeutung dessen an verschiedenen Sätzen anschaulich. Verwenden Sie für die nachfolgenden Aufgaben möglichst diese Tokenisierung.
- Lassen Sie sich für mehrdeutige Sätze (z.B. Sätze 1, 2, 12 unten) die k besten Parsebäume ausgeben! Interpretieren Sie diese und stellen Sie sie zueinander in Bezug!
- Intern verwendet der Berkeley-Parser *binarisierte* Regeln, die entsprechenden binären Parsebäume werden vor der Ausgabe jedoch in Bäume von beliebigem Rang umgewandelt. Nutzen Sie die Option, diese Umwandlung zu deaktivieren, und analysieren Sie so die Funktionsweise der Binarisierung. Beschreiben Sie die Vorteile der Verwendung binarisierter Regeln!

Aufgabe 4 (Max-Rule vs. Viterbi Parsing)
Im Standardmodus versucht der Berkeley-Parser nicht den Parsebaum mit der höchsten Wahrscheinlichkeit (sog. Viterbi Parsing), sondern betreibt *Max-Rule Parsing*.

- Erschließen Sie sich die Bedeutung dessen anhand von [PK07].
- Stellen Sie für ausgewählte Beispielsätze die Viterbi- und die Max-Rule-Parsebäume gegenüber.
- Warum verwendet der Berkeley-Parser standardmäßig Max-Rule als Zielfunktion statt der Baumwahrscheinlichkeit?

Aufgabe 5 (Laufzeit)
Machen Sie sich ein Bild von der Laufzeit des Parsevorgangs in Abhängigkeit von der Satzlänge. Verwenden Sie dazu u.a. die Sätze 16 und 17.

Bonusaufgabe: Schreiben Sie ein Skript, welches das Verhältnis zwischen Satzlänge und Laufzeit für einen größeren Korpus (z.B. den deutschen Teil des Europarl-Korpus, `www.statmt.org/europarl/`) berechnet und in einen Graphen plottet.

Satzkorpus (deutsch und englisch):

1. I saw her duck.
2. I saw the man with the telescope.
3. Colorless green ideas sleep furiously.
4. Furiously sleep ideas green colorless. [Cho57]
5. Awkward! Kristen Stewart bombarded with break-up quiz!
6. But the midfielder concedes that all England teams have failed to live up to Sir Alf Ramsey's World Cup winners.
7. Police snaps reveal suspects guilty of the world's worst moustaches. [alle The Sun (Onlineausgabe), 15. Nov.]
8. The Spanish unions are protesting austerity cuts and an unemployment rate at 25% of the workforce.
9. Mr. Ballmer has been stressing to employees and investors his vision of making Microsoft's products, such as Windows for computers and smartphones, Xbox and Office software suite work more seamlessly together.
10. The Valemax fight offers a glimpse of one of the biggest battles China's new leaders will face as they take the reins of the world's No. 2 economy this week. [alle Wall Street Journal (Onlineausgabe), 15. Nov.]
11. Der Mann biss den Hund.
12. Ich traf den Sohn des Nachbarn mit dem Gewehr.
13. "Meine Exzesse waren der Versuch auszuloten, wie weit ich gehen konnte, wenn ich jahrelang zwei Flaschen, manchmal auch drei Flaschen Whisky am Tag trank", schilderte Maffay der "Zeit".
14. Demnach hätte Steinbrück für Fahrten, die mit dem Bundestagsmandat nichts zu tun haben, wie jeder normale Bahnkunde sein Ticket aus eigener Tasche bezahlen müssen. [beide `bild.de`, 15. Nov.]

15. Das teleogische Tun ist ein Schluss, worin dasselbe Ganze in subjektiver Form mit seiner objektiven Form, der Begriff mit seiner Realität durch die Vermittlung der zweckmäßigen Tätigkeit zusammengeschlossen wird und der Begriff Grund einer durch ihn bestimmten Realität ist. [Heg40]

16. Und die kleine Antonie, achtjährig und zartgebaut, in einem Kleidchen aus ganz leichter changierender Seide, den hübschen Blondkopf ein wenig vom Gesichte des Großvaters abgewandt, blickte aus ihren graublauen Augen angestrengt nachdenkend und ohne etwas zu sehen ins Zimmer hinein, wiederholte noch einmal: "Was ist das", sprach darauf langsam: "Ich glaube, daß mich Gott", fügte, während ihr Gesicht sich aufklärte, rasch hinzu: "– geschaffen hat samt allen Kreaturen", war plötzlich auf glatte Bahn geraten und schnurrte nun, glückstrahlend und unaufhaltsam, den ganzen Artikel daher, getreu nach dem Katechismus, wie er soeben, anno 1835, unter Genehmigung eines hohen und wohlweisen Senates, neu revidiert herausgegeben war. [Man03]

17. Die anderen Menschen fand ich in der entgegengesetzten Richtung, indem ich nicht mehr in das gehaßte Gymnasium, sondern in die mich rettende Lehre ging, gegen alle Vernunft in der Frühe nicht mehr mit dem Sohn des Regierungsrats in die Mitte der Stadt durch die Reichenhaller Straße, sondern mit dem Schlossergesellen aus dem Nachbarhaus an ihren Rand durch die Rudolf-Biebl-Straße, nicht auf dem Weg durch die wilden Gärten und an den kunstvollen Villen vorbei in die Hohe Schule des Bürger- und des Kleinbürgertums, sondern an der Blinden- und Taubstummenanstalt vorbei und über die Eisenbahndämme und durch die Schrebergärten und an den Sportplatzplanken in der Nähe des Lehener Irrenhauses vorbei in die Hohe Schule der Außenseiter und Armen, in die Hohe Schule der Verrückten und der für verrückt Erklärten in der Scherzhauserfeldsiedlung, in dem absoluten Schreckensviertel der Stadt, an der Quelle fast aller Salzburger Gerichtsprozesse und im Keller als Lebensmittelgeschäft des Karl Podlaha, der ein zerstörter Mensch und ein empfindsamer Wiener Charakter gewesen war und der Musiker hatte werden wollen und dann immer ein kleiner Krämer geblieben ist. [Ber73]

References

[Ber73] Thomas Bernhard. *Der Keller: eine Entziehung*. Residenz Verlag, 1973.

[Cho57] Noam Chomsky. *Syntactic Structures*. Mouton & Co., Den Haag, 1957.

[Heg40] Georg Wilhelm Friedrich Hegel. *Philosophische Propädeutik*. Duncker und Humblot, 1840.

[Man03] Thomas Mann. *Buddenbrooks*. S. Fischer Verlag, 1903.

[MSM93] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[PK07] Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. *Proceedings of NAACL HLT 2007*, 2007.