7. Übung zur Vorlesung "NLP – Analyse des Wissensrohstoffes Text" im Sommersemester 2008

Dr. Andreas Hotho, Dipl.-Inform. Dominik Benz, Wi.-Inf. Beate Krause

2. Juli 2008

1 Word Sense Disambiguation - Grundlagen

Dieser Abschnitt befasst sich mit den Grundlagen der Disambiguierung von Wortbedeutungen.

1. Erklären Sie die Problemstellung der Wortsinnunterscheidung. Beschreiben Sie zwei praktische Anwendungen, in denen Wortsinnunterscheidung eine zentrale Rolle spielt.
2. Wenn man als Mensch einen Text liest oder hört, denkt man normalerweise nicht lange über Wortsinnunterscheidung eines bestimmten Wortes nach. Überlegen Sie, welche Hinweise und Informationen man dabei implizit beachtet. Welche Menge und welche Art von Informationen halten Sie für notwendig, um den Sinn eines Wortes (als Mensch) bestimmen zu können? Erklären Sie dabei den Begriff "Kontext" in diesem Zusammenhang.
3. Welche Arten von "mehrdeutigen" Worten kennen Sie? Welche sind für Menschen schwerer / leichter zu unterscheiden? Welche Schlussfolgerungen ergeben sich daraus für die bestmögliche Performanz einer automatischen Wortsinnerkennung? Welche anderen Faktoren spielen eine Rolle bei der Bewertung dieser oberen und unteren Performanz-Schranken?
4. Zur Evaluierung von Wortsinnunterscheidungs-Algorithmen werden häufig *Pseudowörter* eingesetzt. Aus welchem Grund? Erklären Sie das Vorgehen hierbei. Diskutieren Sie den Unterschied von Pseudowörtern zu "richtigen" mehrdeutigen Wörtern.

2 Word Sense Disambiguation - Ansätze

Ansätze zur Wortsinnerkennung lassen sich zunächst grob in überwachte (supervised) und unüberwachte (unsupervised) Ansätze einteilen. Im Buch finden sich zusätzlich noch

Wörterbuch-basierte (*dictionary-based*) Ansätze. Handelt es sich hierbei um überwachte oder unüberwachte Verfahren?

2.1 Überwachte Verfahren

1. Erklären Sie das Grundprinzip des Bayesschen Ansatzes zur Wortsinnerkennung. Welche Annahme wird hierbei gemacht?

2.2 Wörterbuch-basierte Verfahren

1. Auf welcher Art von Information beruht Lesk's Algorithmus zur Wortsinnunterscheidung? Skizzieren Sie kurz dessen Funktionsweise.
2. Wo liegt eine Schwachstelle von Lesk's Algorithmus? (*Hinweis*: Beachten Sie, welche Worte im Überlapp zwischen den Sinndefinitionen vorkommen können.) Fällt Ihnen eine Strategie ein, diesem Schwachpunkt zu begegnen?

Betrachten Sie folgenden Text:

```
1 Oma hatte Peter 5 Euro gegeben. Er wollte das Geld sparen, um sich später
2 eine Gitarre kaufen zu können. Darum brachte er es auf die Bank. Am
3 Schalter zahlte er das Geld ein, und war stolz auf seine ersten Ersparnisse.
4 Danach ging er zu seiner Oma, die auf einer Bank im Park sass, und erzählte
5 ihr, welche Lieder er mit seiner Gitarre einmal spielen wollte.
```

Nehmen Sie an, folgende semantische Kategorieren wären in einem Thesaurus vorhanden:

```
Oma - RELATIVE
Euro - FINANCE
Geld - FINANCE
Gitarre - MUSIC
Bank - RECREATION
     - FINANCE
Schalter - FINANCE
         - INTERFACE
Ersparnisse - FINANCE
Park - RECREATION
Lieder - ART
zahlen - FINANCE
sitzen - RECREATION
```

Disambiguieren Sie alle Vorkommen des Wortes *Bank* im Text mittels des Algorithmus von Walker. Nehmen Sie für alle Worte, für die keine Sinndefinition im Thesaurus exi-

stiert, die leere Menge an Sinnen an. Betrachten Sie als Kontext eines Wortes ein Fenster der Grösse 3 in beide Richtungen.

2.3 Unüberwachte Verfahren

1. Grenzen Sie die Begriffe *sense tagging* und *sense discrimination* gegeneinander ab. Welches von beiden liegt bei unüberwachten Verfahren zur Wortsinnunterscheidung vor?
2. Welches Modell liegt dem EM-Algorithmus auf Seite 254 zugrunde?
3. Ein wichtiger Parameter des EM-Algorithmus ist K. Welche Funktion hat er? Erklären Sie die Folgen, wenn man diesen Parameter zu gross oder zu klein wählt. Was wird in der Praxis getan, um diesen Parameter richtig einzuschätzen?