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

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.

LÖSUNGSVORSCHLAG:

1. In natürlicher Sprache haben viele Wörter mehrere Bedeutungen oder *Sinne*. Wortsinnunterscheidung bezeichnet den Vorgang, einem Wort in einem bestimmten Kontext den jeweils richtigen Sinn zuzuordnen - z.B. ob in einem Text von einer *Bank* im

Sinne einer Sitzmöglichkeit oder einem Finanzinstitut die Rede ist. Anwendungen wie z.B. Information Retrieval oder automatische Übersetzer hängen stark davon ab, den richtigen Wortsinn zu treffen.

2. Als Mensch beachtet man den Kontext eines Wortes, und wählt dann automatisch den Wortsinn, der "am meisten Sinn" macht oder am plausibelsten erscheint. Dabei verwendet man als Mensch einerseits (lexikalische / syntagmatische) Hinweise direkt aus dem Text (*Ïch setze mich auf die Bank"*), oder aber Hintergrundwissen bzw. Erfahrungswerte (z.B. meinen die meisten Leute wohl das Finanzinstitut mit dem Satz *Ïch gehe zur Bank"*). Letzteres zählt zu den paradigmatischen Informationsquellen. Je nach Kontext kann die Menge der benötigten Informationen zur Unterscheidung variieren.

3. Siehe Kapitel 3 :) Man unterscheidet zwischen *Homonymen* (unterschiedliche Bedeutung) und *Polysemen* (ähnliche Bedeutung). Erstere (Beispiel: *Bank*) werden auch von Menschen leichter und mit höherer Übereinstimmung unterschieden. Letztere (z.B. *Titel*, *Fenster*) sind schwerer zu unterscheiden, was sich auch in geringerer Übereinstimmung bei verschiedenen Personen ausdrückt. Für Polyseme gelten deshalb allgemein niedrigere Schwellwerte der Performanz ($\tilde{6}0\%$) als bei Homonymen ($\tilde{9}5+\%$). Eine weitere Rolle bei der Performanz-Bewertung spielen die Anzahl der verschiedenen Sinne sowie deren Wahrscheinlichkeit (siehe Buch S. 234, vorletzer Abschnitt).

4. Man verwendet Pseudowörter, weil von Hand disambiguierte Texte sehr teuer in der Erstellung sind. Die Methode von Gale et al. hierzu ist, zwei natürlichsprachliche Wörter aneinanderzuhängen (*bananadoor*). Dieses Verfahren stellt zwar eine Hilfe dar, kann aber nur recht grob "wirkliche" mehrdeutige Worte (besonders im Hinblick auf den Kontext) abbilden.

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?

LÖSUNGSVORSCHLAG:

Streng genommen existiert bei Wörterbuch-basierten Ansätzen kein Trainingskorpus, in dem die Wortsinne korrekt annotiert sind. Andererseits basieren diese Ansätze auf externer Information (nämlich genau dem Wörterbuch), die quasi aus einer Menge von annotierten Korpora genommen ist. Deshalb zählt man diese Ansätze zu den überwachten Verfahren.

2.1 Überwachte Verfahren

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

LÖSUNGSVORSCHLAG:

1. Bei diesem Ansatz handelt es sich um einen Klassifikationsansatz. Für ein mehrdeutiges Wort wird diejenige Bedutung ausgewählt, die unter Voraussetzung des gegebenen Kontextes am wahrscheinlichsten ist (Formel 7.2, Seite 236). Diese bedingte Wahrscheinlichkeit wird mittels der Bayes-Regel berechnet. Unter der Annahme, dass die Wörter innerhalb eines Kontextes unabhängig voneinander sind, lassen sich die individuellen Wahrscheinlichkeiten für jedes Kontextwort aufsummieren (Formel 7.4, Seite 237). Die bedingten Wahrscheinlichkeiten für jedes Kontextwort werden mittels Maximum-Likelihood-Abschätzung berechnet (Seite 237 unten).

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?

LÖSUNGSVORSCHLAG:

1. Lesk's Algorithmus basiert auf einem Wörterbuch, das für jedes Wort einen kurzen erklärenden Text (Definition) enthält. Für ein mehrdeutiges Wort (z.B. *Bank*) vergleicht der Algorithmus dann die Definitionen D_k der Sinne (z.B. *Finanzinstitut* und *Sitzgelegenheit*) mit den Definitionen E_{v_j} der Wörter, die im Kontext vorkommen. Hierzu wird der Wortüberlapp der Definitionen verwendet.
2. Der Algorithmus is nicht optimal, da im Überlapp der Definitionen auch z.B. Stopwörter oder sehr häufige Worte vorkommen. Eine Gegenmassnahme wäre hier eine Stopwortliste, oder eine Gewichtung der Worte nach ihrer Spezifität, z.B. durch TF/IDF (siehe Information Retrieval).

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

LÖSUNGSVORSCHLAG:

Der 1. Kontext lautet von *Bank* lautet

```
es auf die Bank. Am Schalter zählte
```

Laut Thesaurus sind die möglichen Sinne $s_1 = \texttt{FINANCE}, s_2 = \texttt{RECREATION}$. Für jeden dieser Sinne wird nun ein score berechnet:

$$\begin{aligned} score(s_1) &= \delta(\texttt{FINANCE}, es) + \delta(\texttt{FINANCE}, auf) + \delta(\texttt{FINANCE}, die) \\ &\quad + \delta(\texttt{FINANCE}, am) + \delta(\texttt{FINANCE}, Schalter) + \delta(\texttt{FINANCE}, zählte) \\ &= 0 + 0 + 0 + 0 + 1 + 0 \\ &= 1 \end{aligned}$$

$$\begin{aligned} score(s_1) &= \delta(\texttt{RECREATION}, es) + \delta(\texttt{RECREATION}, auf) + \delta(\texttt{RECREATION}, die) \\ &\quad + \delta(\texttt{RECREATION}, am) + \delta(\texttt{RECREATION}, Schalter) + \delta(\texttt{RECREATION}, zählte) \\ &= 0 + 0 + 0 + 0 + 0 + 0 \\ &= 0 \end{aligned}$$

Da s_1 den score maximiert, wird im ersten Kontext auf den Wortsinn `FINANCE` entschieden. Der 2. Kontext

```
die auf einer Bank im Park sass
```

wird analog behandelt - hier gilt $score(s_1) = 0, score(s_1) = 2$. Somit wird im zweiten Kontext auf den Wortsinn `RECREATION` entschieden.

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?

LÖSUNGSVORSCHLAG:

1. Beim *sense tagging* werden vorher festgelegte Sinne einem Wort zugeordnet. Bei unüberwachten Verfahren sind diese Sinne im Vorhinein nicht bekannt; deshalb kann hier nur eine *sense discrimination* erfolgen. Hierbei werden die verschiedenen Kontexte eines Wortes zu Ähnlichkeitsgruppen zusammengefasst (bzw. geclustert).
2. Der EM-Algorithmus basiert auf dem Bayes-Modell, das im Buch in Abschnitt 7.2.1 vorgestellt wurde.
3. Der Parameter K stellt die Anzahl der möglichen Sinne für ein Wort dar. Wählt man ihn zu klein, hat das entstehende Modell zuwenig Struktur, um sich den Daten anzupassen; wählt man ihn zu gross, werden möglicherweise zu viele Sinne unterschieden. In der Praxis werden Experimente mit ansteigenden K's gemacht: Ändert sich die Wahrscheinlichkeit des Korpus damit sprunghaft, war die Erhöhung gerechtfertigt; steigt sie nur leicht, so nimmt man an, dass eine weitere Erhöhung keinen Sinn macht.