Proseminar
Word Sense Disambiguation

Stefan Thater
FR 4.7 Allgemeine Linguistik (Computerlinguistik)
Universität des Saarlandes

Wintersemester 2010/11

Ambiguität

- **Ambiguität:** Eigenschaft von Ausdrücken natürlicher Sprachen (Wort, Satz, etc.), denen mehrere Bedeutungen zugewiesen werden können.
- **Lexikalische Ambiguität**
  - *Absatz* – Teil des Schuhs, Verkauf, Textabschnitt
  - *Brücke* – Bauwerk, Zahnersatz
- **Strukturelle Ambiguität**
  - *Umschulung zum Erzieher für Männer*
  - *Die Wahl des Vorsitzenden fand Zustimmung*
  - *Maria möchte einen Prinzen heiraten*

Ambiguität

- Mehrdeutige Ausdrücke werden typischerweise durch den Kontext desambiguiert ⇒ eine Bedeutung ist wesentlich wahrscheinlicher als die andere(n).
  - *In Europa ist der Absatz auch ohne Werbung gesichert*
  - *Beginnen Sie Ihre Briefe mit einem kurzen Absatz, …*
- **Kontext:**
  - Sprachlicher Kontext
  - Hintergrundwissen
  - Intonation
  - […]

Lexikalische Ambiguität

- **Homonymie** – kein Bezug zwischen Bedeutungen
  - *Strauß* – Vogel, Blumen
  - *Kiefer* – Baum, Teil des Schädels
  - *Arm* – Körperteil, mittellos
- **Polysemie** – gemeinsamer „Bedeutungskern“
  - *Brücke* – Bauwerk, Zahnersatz
- **Systematische Polysemie** (betrifft Klassen)
  - *Die {Schule, Uni, ...} wird renoviert* (Gebäude)
  - *Die {Schule, Uni, ...} muss sparen* (Institution)
  - *Die {Schule, Uni, ...} beginnt um 8:30 Uhr* (Prozess)

Weitere Phänomene

- **Metonymie** – Ersatz eines Ausdrucks für einen sachlich verwandten Ausdruck
  - *Goethe ist schwer zu verstehen* (Autor - Werk)
  - *Ein Glas trinken* (Produkt - Material)
  - *Berlin hat beschlossen, dass ...* (Ort – Institution)
- **Vagheit** – Kontinuum von Bedeutungsschattierungen (statt mehrere, klar unterscheidbare Bedeutungen)
  - *Berg* (vs. *Hügel*)
  - *groß* (vs. *klein*)
  - *rot* (vs. *orange*)

Word Sense Disambiguation

- **Word sense disambiguierung:**
  Weise den Vorkommen mehrdeutiger Worte abhängig vom Kontext seine Bedeutung (word sense) zu.
  - Bestimmung aller möglichen Bedeutungen
  - Auswahl der richtigen Bedeutung

Sense

- *"Not in the largest **sense** of the words", I said.*
- *The body, **senses** and brain, in common with all matter, have their counterpart on each of a countless number of frequencies.*
- *We will achieve a more vivid **sense** of what it is by realizing what it is not.*
- *He soon came back to his **senses**.*
- *It made no **sense** to him.*

Sense in WordNet

- sense#1: a general conscious awareness – "a sense of security", "sense of happiness"
- sense#2: the meaning of a word or expression – "the dictionary gave several senses for the word"
- sense#3: the faculty through which the external world is apprehended – "in the dark he had to depend on touch and on his senses of smell and hearing"
- sense#4: sound practical judgment – "fortunately she had the good sense to run away"
- sense#5: a natural appreciation or ability – "a keen musical sense", "a good sense of timing"

Word Sense Discrimination

- **Word sense disambiguierung:**
  Weise den Vorkommen mehrdeutiger Worte abhängig vom Kontext seine Bedeutung (word sense) zu.
- **Word sense discrimination:**
  Entscheide, ob zwei Vorkommen eines Wortes die gleiche Bedeutung haben.

[Ide & Veroni, 1998]

Einige Anwendungen

- **Maschinelle Übersetzung**
  - Der Absatz ist eingebrochen – The paragraph is broken
- **Information retrieval (IR)**
- **Speech processing**
- **Allgemein:** sprachverstehende Systeme

Word Sense Disambiguation

- **Was ist Wortbedeutung?**
  - Die Bedeutung eines Wortes w zu kennen bedeutet, zu wissen, wann/ob etwas ein w ist.
- **Wie können Bedeutungen repräsentiert werden?**
  - Wörterbucheinträge
  - Thesaurus
  - Übersetzungen in andere Sprachen
  - etc.
  - ⇒ Abhängig von der konkret betrachteten Aufgabe

Methoden

- Welche Wissensressourcen werden verwendet?
  - Sind Bedeutungen durch ein (maschinenlesbares) Wörterbuch bereits definiert?
  - Knowldegde lean vs. knowledge rich
- Werden manuell annotierte (bereits disambiguierte) Trainingsdaten verwendet?
  - Supervised WSD vs. unsupervised WSD

Maschinenlesbare Wörterbücher

- **Lesk (1986)**
  Entscheide, welche Bedeutung eines Wortes *w* in einem bestimmten Kontext intendiert ist:
  - betrachte die Wörter, die in der Definition von *w* in einem maschinenlesbaren Lexikon vorkommen
  - und berechne die „Überlappung“ dieser Wörter mit den Worten im Kontext (nähere Umgebung) von *w*.
- *The **bank** can guarantee that **deposits** will …*
  - bank_1 – sloping land / the slope beside a body of water
  - bank_2 – a financial institution that accepts **deposits** and …
- Funktioniert überraschend gut

Überwachte Lernverfahren

- Überwachte Lernverfahren (Klassifikation)
  - Für jedes Wort wird ein separater Klassifikator trainiert
- Zutaten:
  - Trainingsdaten – bereits disambiguierte Textkorpora
  - Repräsentationen des Kontexts (Features)
  - Lernverfahren (Algorithmus, Modell)
- Sehr hohe Präzision, aber sehr „teuer“ (Trainingsdaten)

Decision Lists (Yarowski, 1994)

- Geordnete Liste von Regeln (Tests)
- *bass* (fish) vs. *bass* (instrument)
  - *fish* in ±k Wortfenster ⇒ FISH
  - *striped bass* ⇒ FISH
  - *guitar* in ±k Wortfenster ⇒ INSTRUMENT
  - *bass player* ⇒ INSTRUMENT
  - …

Bootstrapping (Yarowski, 1995)

- **Yarowsky (1995)**
  - Starte mit einem (sehr) kleinen Teilkorpus („seed“)
  - Lerne Klassifikator
  - Disambiguiere restliches Korpus
  - Behalte Teilkorpus (mit hoher Konfidenz)
  - Wiederholen …
- **Heuristiken:**
  - One sense per discourse
  - One sense per collocation

Weitere Methoden

- **Mihalcea & Moldovan (1999)**
  - Web-basierten Wort-Wort-Kookkurrenzen + WordNet-basiertem Maß für „semantische Dichte“
- **Hinrich Schütze (1998)**
  - Word-sense discrimination, Clustering
- **Mihalcea (2005)**
  - Graph-basiertes Verfahren, PageRank
- **Diab & Resnik (2002)**
  - Paralleles Korpus (Englisch-Französisch)

Weitere Methoden

- **McCarthy &al. (2004)**
  - Wie findet man die häufigste Bedeutung eines Wortes in unannotierten Daten?
  - Kombiniert Thesaurus mit WordNet-Ähnlichkeit
- **Mihalcea (2007)**
  - Betrachtet Wikipedia als (Sense-) annotiertes Korpus
- **Erk & McCarthy (2009)**
  - Keine klaren Grenzen zwischen verschiedenen Wortbedeutungen sondern Kontinuum
  - ⇒ *graded word senses*

Wie schwierig ist WSD?

- **SensEval 3** (WordNet senses)
  - F-score 0.65 (alle Wörter disambiguieren)
  - Most frequent sense baseline: 0.62
  - Inter-Annotator Agreement: 0.72
- **SemEval 2007** (WordNet senses, coarse-grained)
  - F-score 0.82 (unsupervised 0.70)
  - Most frequent sense baseline: 0.79

Zeitplan

2010-10-25	Einführung	Thater
2010-11-08	Einführung	Thater
2010-11-15	Fellbaum (1998). Wordnet.	N.N.
2010-11-22	Yarowsky (1994). Decision lists for lexical ambiguity resolution: Application to Accent Restoration in Spanish and French.	N.N.
2010-11-29	Yarowsky (1995). Unsupervised word sense disambiguation rivaling supervised methods.	N.N.
2010-12-06	Mihalcea & Moldovan (1999). A Method for Word Sense Disambiguation of Unrestricted Text.	N.N.
2010-12-13	Hinrich Schütze (1998). Automatic Word Sense Discrimination.	N.N.
2011-01-03	Mihalcea (2005). Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling.	N.N.
2011-01-10	Diab & Resnik (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora.	N.N.
2011-01-17	McCarthy &al. (2004). Finding predominant senses in untagged text.	N.N.
2011-01-24	Mihalcea (2007). Using Wikipedia for Automatic Word Sense Disambiguation	N.N.
2011-01-31	Erk & McCarthy (2009). Graded Word Sense Assignment.	N.N.
2011-02-07	Abschlussdiskussion	Thater

Organisatorisches

- **Prüfungsleistungen**
  - Vortrag (etwa 45 Minuten)
  - Seminararbeit (etwa 10 Seiten)
  - Aktive Teilnahme an den Diskussionen
- **Gewichtung**
  - Vortrag und Seminararbeit je 40%
  - Diskussionsbeiträge 20%

Organisatorisches

- Vorbesprechung zwei Wochen vor dem Vortrag
  - ⇒ Klärung inhaltlicher Fragen
- Vorbesprechung eine Woche vor dem Vortrag
  - ⇒ Feedback zu den Folien

Nächste Sitzung

- Themenvergabe
- Wie halte ich einen guten Vortrag?