

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Automatische Desambiguierung von Wortbedeutungen

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

27. März 2010

WSD: Aufgabe

- **Word Sense Disambiguation (WSD)**: automatische Identifikation der Wortbedeutung eines Wortes in einem bestimmten Kontext
- Klassifikationsaufgabe: Wortbedeutungen entsprechen Klassen, Kontext liefert Anhaltspunkte, und jedes Vorkommen eines Wortes wird auf der Basis der Anhaltspunkte einer oder mehrere Klassen zugeordnet
- intermediäre Aufgabe, kein End-Ziel
- Nähe zur Lexikographie; lexikalische Semantik ist tendenziell zu komplex für WSD

WSD: Aufgabe

- **Aufgabenteile:**
 - ① Definition eines Bedeutungsinventars
 - ② Modell für Zuweisung eines Wort-Vorkommens zu Bedeutung
- **Informationsquellen:**
 - Kontext → datenbasiert
 - externe Informationsquellen → wissensbasiert
- **Integration** eines WSD-Moduls:
 - als Black Box; Ergebnisse werden in Anwendung integriert
 - als genuiner Teil einer Anwendung
- wenige Erfolge von WSD in computerlinguistischen Anwendungen
- aber reiche Forschungsergebnisse in computationeller Semantik

WSD: Bedeutungsinventar

- Anzahl von Bedeutungen pro Wort
- Definition und Abgrenzung der Bedeutungen
- Probleme:
 - Subjektivität und Uneinigkeit
 - Abhängigkeit der Definitionen von der Aufgabe

WSD: Bedeutungsinventar

- Kilgarriff (1997): *I don't believe in word senses*
- There is no reason to expect a single set of word senses to be appropriate for different NLP applications.
- In the absence of sense purposes, word senses do not exist.
- Different corpora and different purposes will lead to different senses.
- If we are to know what word senses are, we need operational criteria for distinguishing them.

WSD: Bedeutungsinventar

- Möglichkeiten:
 - Homonymie als Bedeutungsgrundlage
 - Übersetzungen eines Wortes in eine zweite Sprache
 - Anwendungs-gesteuerte Definition, z.B. Granularität von Wortbedeutungen in Information Retrieval vs. maschineller Übersetzung
 - Jedes Vorkommen eines Wortes repräsentiert eine eigene Bedeutung; aber: Vorkommen clustern sinnvoll
- Anforderungen: Klarheit, Konsistenz, Vollständigkeit
- in groben Unterteilungen fehlen Bedeutungen, in feinen Unterteilungen sind unnötige Unterteilungen oder Fehler

WSD: Bedeutungsinventar

- Bedeutungsinventare und Frequenz (Zipf, 1949)
- **Anzahl Bedeutungen in Bezug auf Wortfrequenz:** hochfrequente Wörter haben mehr Bedeutungen als niedrigfrequente Wörter
- **Anzahl Vorkommen von Bedeutungen:** wenige Bedeutungen eines Wortes kommen häufig vor; viele Bedeutungen eines Wortes kommen selten vor

WSD: Klassifikation der Methode

- nach der Hauptquelle des Wissens: Wörterbuch-basiert vs. wissensbasiert
- nach der Trainingsgrundlage: überwacht, halb-überwacht, unüberwacht (supervised, semi-supervised, unsupervised)
- Kombinationen

WSD: wissensbasierte Methoden

- große lexikalische Ressourcen
- **Maschinen-lesbare Wörterbücher:**
 - gebrauchsfertige Informationsquellen zu Wortbedeutungen
 - Text der Wörterbucheinträge enthält lexikalisch-semantisches Wissen
- **Thesauri:**
 - enthält Informationen zu Beziehungen zwischen Wörtern
 - Wörter in gemeinsamer Kategorie haben eine semantische Relation
- Probleme:
 - Inkonsistenzen
 - Unvollständigkeit
 - für menschliche und nicht für maschinelle Nutzung erstellt

WSD: wissensbasierte Methoden

- große lexikalische Ressourcen
- **computationelle Lexika:**
 - aufzählende Lexika: explizite Wortbedeutungen (z.B. WordNet, Cyc)
 - generative Lexika: semantische Information ist unerspezifiziert; durch Generierungsregeln wird Information zu spezifischer Wortbedeutung erzeugt (z.B. Pustejovskys generatives Lexikon)

WSD: datenbasierte Methoden

- Kontext und Kookkurrenz
- *Bag-of-Words* vs. syntax-basierte Kookkurrenz
- Probleme: Frequenz und Ambiguität
- überwachtes Lernen von Korpusdaten, die mit Wortbedeutungen annotiert sind
- unüberwachtes Lernen: Wortbedeutungen für Cluster vs. Wortvorkommen
- Bilinguale Korpora stellen Übersetzungen zur Verfügung

Senseval und Semeval

- Evaluierung und Vergleich von Ergebnissen verschiedener Systeme schwierig, da Testdaten, Annotatoren, Bedeutungsinventare und Korpora typischerweise unterschiedlich sind
- **Senseval**: Evaluierungsaufgabe für WSD; 1998-2004
- **Semeval**: semantische Evaluierung im Allgemeinen; schließt Senseval ein; seit 2007
- [Senseval Homepage](#)

Senseval

- stellt Gold-Standard(s) für Wortbedeutungen zur Verfügung, z.B. *SemCor*
- Annotation mit WordNet-Bedeutungen
- Aufgabe: Entwicklung eines (un)überwachten Systems für WSD
- Evaluierung und Vergleich von Systemen an gemeinsamen Daten

Senseval

- **Ziele:**
 - Einigung über explizite und detaillierte Definition der Aufgabe
 - Erstellung eines Gold-Standards mit “richtigen” Wortbedeutungen
 - Evaluierung und Vergleich von Systemen in Bezug auf verschiedene Wörter und verschiedene Sprachen
 - verbessertes Verständnis von lexikalischer Semantik, speziell Ambiguitäten
- **Sprachen:** Englisch, Italienisch, Baskisch, Katalanisch, Chinesisch, Rumänisch, Spanisch

Beispielansätze

- 1 Lesk (1986): Wörterbuch-basiert
- 2 Gale et al. (1992): Rolle von Diskurs
- 3 McCarthy et al. (2004, 2007): dominante Wortbedeutungen

Desambiguierung durch Wörterbuch-Definitionen

- Idea: automatische Desambiguierung der Bedeutung eines Wortvorkommens auf der Basis von Wörterbuchdefinitionen
- Vergleich von Wörterbuchdefinitionen für konkurrierende Wörter im Kontext
- Maß: Überlappung von Wörtern in Definitionen
- nicht-syntaktischer Ansatz, der mit syntaktischer Information verknüpft werden kann
- Genauigkeit: 50-70%

Beispiele: Wörterbuchdefinitionen

- pine** 7 kinds of evergreen tree with needle-shaped evergreen tree
0 waste away through sorrow or illness
0 pine for sth; pine to do sth; have a pine
- cone** 0 solid body which narrows to a point ...
0 something of this shape whether solid or hollow
8 fruit of certain evergreen trees (fir, pine, ...)

Beispiele: Wörterbuchdefinitionen

- coal**
1. a piece of glowing carbon or charred wood: ember
 2. charcoal
 3. a black or brownish black solid combustible substance formed by the partial decomposition of vegetable matter without free access of air and under the influence of moisture and often increased pressure and temperature that is widely used as a natural fuel pieces or a quantity of the fuel broken up for burning
- coal**
1. to burn to charcoal
 2. to supply with coal
 3. to take in coal
- ash**
1. any of a genus of trees of the olive family with pinnate leaves, thin furrowed bark, and gray branchlets
 2. the tough elastic wood of an ash
- ash**
1. the solid residue left when combustible material is thoroughly burned or is oxidized by chemical means; fine particles of mineral matter from a volcanic vent
 2. ruins
 3. the remains of the dead human body after cremation or disintegration
 4. something that symbolizes grief, repentance, or humiliation
 5. deathly pallor
- ash**
1. to convert into ash

Beispiele: Wörterbuchdefinitionen

- coal**
1. a piece of glowing carbon or charred wood: ember
 2. charcoal
 3. a black or brownish black **solid combustible** substance formed by the partial decomposition of vegetable matter without free access of air and under the influence of moisture and often increased pressure and temperature that is widely used as a natural fuel pieces or a quantity of the fuel broken up for **burning**
- coal**
1. to burn to charcoal
 2. to supply with coal
 3. to take in coal
- ash**
1. any of a genus of trees of the olive family with pinnate leaves, thin furrowed bark, and gray branchlets
 2. the tough elastic wood of an ash
- ash**
1. the **solid** residue left when **combustible** material is thoroughly **burned** or is oxidized by chemical means; fine particles of mineral matter from a volcanic vent
 2. ruins
 3. the remains of the dead human body after cremation or disintegration
 4. something that symbolizes grief, repentance, or humiliation
 5. deathly pallor
- ash**
1. to convert into ash

Zusammenfassung

- Offene Fragen:
 - Welches Wörterbuch ist geeignet?
→ möglichst lange Einträge
 - Stellen Beispiele in Wörterbucheinträgen geeignete Information dar?
 - Sollen die Überlappungen Token- oder Typ-basiert berechnet werden?
 - Was ist ein geeignetes Wortfenster für Kookkurrenz?
 - Soll die Überlappung in Bezug auf Nähe im Wortfenster gewichtet werden?

Rolle von Diskurs

- Annahme: wenn ein ambiges Wort mehrfach in einem Diskurs auftaucht, ist es extrem wahrscheinlich, dass die Vorkommen dieselbe Bedeutung haben
- Experimenteller Nachweis der Annahme
- Ergebnis: Tendenz liegt bei 98%
- Folgerung: Kennzeichnung von Bedeutungen auf Diskurs-Ebene

WSD-System

- Experiment-Umgebung: WSD-System mit zwei Komponenten:
 - automatischer heuristischer Ansatz mit einem parallelen Korpus
 - Bayesian-Discrimination Ansatz
- Vor-Experiment:
 - Zufallsauswahl von 108 Nomen
 - drei Beurteiler (Autoren)
 - 100 Mengen Konkordanz-Zeilen, die jeweils alle Vorkommen eines Wortes in einem bestimmten Grolier-Artikel enthalten
 - Aufgabe: Entscheidung, ob Menge von Konkordanz-Zeilen dieselbe Bedeutung eines Wortes ansprechen
 - Ergebnis: nur sechs von 300 Beurteilungen beziehen sich auf unterschiedliche Bedeutungen

WSD-System

- Haupt-Experiment (blind):
 - Auswahl von neun ambigen Nomen: *antenna, campaign, deposit, drum, hull, interior, knife, landscape, marine*
 - fünf Beurteiler
 - Information: Definitionen vom OALD und Paare von Konkordanz-Zeilen aus der Grolier-Enzyklopädie
 - Entscheidung: Sprechen die Paare von Konkordanz-Zeilen jeweils dieselbe Bedeutung an?
 - 54 Paare aus identischen Diskursen, 28 Paare aus verschiedenen Diskursen (mit nachweislich verschiedenen Bedeutungen)
 - durchschnittliche Übereinstimmung der Beurteiler: 96.8%
 - durchschnittlich 51 von 54 Paaren als identische Bedeutung beurteilt (94%)
 - Annahme: durchschnittlich 60% nicht-ambige Wörter und 40% ambige Wörter in einem Text → ca. 98% der ambigen Wörter innerhalb eines Diskurses beziehen sich auf dieselbe Bedeutung

Dominante Wortbedeutungen

- Ausgangspunkt: Frequenzverteilung von Wortbedeutungen ist typischerweise schief (*skewed*)
- Konsequenz für WSD: Die Heuristik, die häufigste Wortbedeutung zu wählen, ist extrem mächtig. Sie wird in den meisten Systemen “nur” als *back-off* eingesetzt, ist aber schwierig zu übertreffen.
- Heuristik ist abhängig von Domäne, Thema und Genre;
Beispiel: *star*
- Überwachte Systeme sind generell besser als unüberwachte Systeme
- Probleme: wenige, teure Trainingsdaten; feinkörnige Bedeutungsunterscheidungen

Vorgehensweise

- Methode: Berechnung und Gewichtung von distributionell nächsten Nachbarn
- Annahme: nächste Nachbarn der häufigsten Wortbedeutung haben ein Übergewicht
- Bedeutungsinventar aus bestehender Ressource, z.B. WordNet;
Beispiel: *star* → celebrity, celestial body, shape, sign of the zodiac
- nächste Nachbarn über Thesaurus
- Gewichtung über Ähnlichkeitsmaß

Vorgehensweise

- Schritte:
 - ➊ Automatische Akquisition eines Thesaurus' (Lin, 1998)
 - ➋ Anwendung von WordNet-basierten Ähnlichkeitsmaßen für semantische Ähnlichkeit, um häufigste Bedeutung herauszufinden
 - ➌ Anwendung der häufigsten Bedeutungen auf WSD-Aufgabe

Parameter

- Parameter:
 - Korpus-Daten: British National Corpus, Newswire
 - Ähnlichkeitsmaße: Lesk (Überlappung in Definitionen) vs. Jiang & Conrath (Frequenz-basierter Vergleich von Synsets)
 - Wortart

Beispiel

Beispiel: *star*

Senses	Neighbours of <i>star</i>			
	<i>actor</i> (0.22)	<i>footballer</i> (0.12)	<i>planet</i> (0.08)	<i>circle</i> (0.03)
<i>celebrity</i>	0.42	0.53	0.02	0.01
<i>celestial body</i>	0.01	0.01	0.68	0.10
<i>shape</i>	0.00	0.00	0.02	0.78
<i>zodiac</i>	0.03	0.03	0.21	0.01
Total	0.46	0.57	0.93	0.90

$\text{score}(\textit{celebrity}): 0.22 * \frac{0.42}{0.46} + 0.12 * \frac{0.53}{0.57} + 0.08 * \frac{0.02}{0.93} + 0.03 * \frac{0.01}{0.90} = 0.3145$

$\text{score}(\textit{celestial body}): 0.0687$

$\text{score}(\textit{shape}): 0.0277$

$\text{score}(\textit{zodiac}): 0.0390$

Ergebnisse

- Ergebnisse:
 - ADJ > N > ADV > V
 - Lesk allgemeiner und meistens besser als Jiang & Conrath
 - Nomen haben stärkere Tendenz zu Domänen-basierter Bedeutung als andere Wortarten
 - Domänenspezifische häufigste Wortbedeutungen sind besser als allgemeine häufigste Wortbedeutungen

Referenzen: WSD allgemein



Nancy Ide and Jean Véronis.

Word Sense Disambiguation: The State of the Art.
Computational Linguistics, 24(1), 1998.



Eneko Agirre and Philip Edmonds, editors.

Word Sense Disambiguation: Algorithms and Applications.
Springer-Verlag, 2006.
URL: <http://www.wsdbook.org/>.



Roberto Navigli.

Word Sense Disambiguation: A Survey.
ACM Computing Surveys, 41(2), 2009.

Referenzen: WSD-Ansätze



Michael Lesk.

Automatic Sense Disambiguation using Machine-Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone.
In Proceedings of the SIGDOC Conference, Toronto, Canada, 1986.



William A. Gale, Kenneth W. Church, and David Yarowsky.
One Sense per Discourse.

In Proceedings of the DARPA Speech and Natural Language Workshop, 1992.



Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll.
Unsupervised Acquisition of Predominant Word Senses.
Computational Linguistics, 33(4):553–590, 2007.