## Lexikalische Semantik

# Lexikalisch-semantische Disambiguierung mit WordNet

**Conrad Steffens**

Paper:
Rada Mihalcea & Dan I. Moldovan: A Method for Word Sense Disambiguation of Unrestricted Text

# Übersicht

**Word Sense Disambiguation (WSD)**

- Was ist WSD?
- Wozu braucht man WSD?
- Wie kann man WSD realisieren?

# Was ist WSD?

**Word Sense Disambiguation (WSD)**

- Offenes Problem in NLP
- Klärung der Bedeutung eines (polysemischen) Wortes in einem bestimmten Kontext

# Was ist WSD?

## Word Sense Disambiguation (WSD)

Beispiel: “bank”

**Noun**

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

# Wozu braucht man WSD?

**Word Sense Disambiguation (WSD)**

- Lösung beeinflusst andere Bereiche, z.B. Diskurs, Kohärenz, Inferenz
- Anwendungsgebiete z.B. in maschineller Übersetzung, Dialogsystemen, Suchmaschinen, Question Answering, usw

# Lösungsansätze

- WSD mit machine readable dictionaries (MRD – z.B. WordNet)
- WSD durch Training auf semantisch annotierten Korpora (z.B. SemCor)
- WSD durch Training auf nicht annotierten Korpora
  **(Yarowsky-Algorithmus)**
- WSD mit der Methode von R. Mihalcea & D. Moldovan

# Übersicht

**Word Sense Disambiguation (WSD)**

- Was ist WSD?
- Wozu braucht man WSD?
- Lösungsansätze

- Rada Mihalcea & Dan I. Moldovan
- Wie es funktioniert
- Algorithmus 1
- Algorithmus 2
- Evaluation
- Probleme / Erweiterungen

# Rada Mihalcea & Dan I. Moldovan



- http://www.cs.unt.edu/~rada/pictures/io-05-1.jpg
- http://www.hlt.utdallas.edu/~moldovan/danm.jpeg

# Wie es funktioniert

- Betrachtet werden Wortpaare ($W_1$ – $W_2$)
- $W_2$ wird im Kontext von $W_1$ disambiguiert
  - Internet-Suche mit $W_1$ und verschiedenen Bedeutungen von $W_2$
  - *Ranking* der Bedeutungen von $W_2$ anhand der Treffer
    ► Algorithmus 1
  - Berechnung der semantischen Dichte zwischen $W_1$ und $W_2$
    ► Algorithmus 2
  - Ergebnis: *Ranking* der Bedeutungen (z.B. Top4)

# Algorithmus 1

SCHRITT 1: Erstellen einer *similarity list* für jede Bedeutung von $W_2$ mit WordNet

- $(W_2^1, W_2^{1(1)}, W_2^{1(2)}, ..., W_2^{1(k_1)})$
- $(W_2^2, W_2^{2(1)}, W_2^{2(2)}, ..., W_2^{2(k_2)})$
- ...
- $(W_2^m, W_2^{m(1)}, W_2^{m(2)}, ..., W_2^{m(k_m)})$

# Algorithmus 1

SCHRITT 1: Erstellen einer *similarity list* für jede Bedeutung von $W_2$ mit WordNet

BEISPIEL: *present paper*

**Noun**

- S: (n) **paper** (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
- S: (n) composition, **paper**, report, theme (an essay (especially one written as an assignment))
- S: (n) newspaper, **paper** (a daily or weekly publication on folded sheets; contains news and articles and advertisements)
- S: (n) **paper** (a medium for written communication)
- S: (n) **paper** (a scholarly article describing the results of observations or stating hypotheses)
- S: (n) newspaper, **paper**, newspaper publisher (a business firm that publishes newspapers)
- S: (n) newspaper, **paper** (the physical object that is the product of a newspaper publisher)

# Algorithmus 1

SCHRITT 1: Erstellen einer *similarity list* für jede Bedeutung von $W_2$ mit WordNet

BEISPIEL: *present paper*

**Noun**

- S: (n) composition, **paper**, report, theme (an essay (especially one written as an assignment))
- S: (n) newspaper, **paper** (a daily or weekly publication on folded sheets; contains news and articles and advertisements)

# Algorithmus 1

SCHRITT 1: Erstellen einer *similarity list* für jede Bedeutung von $W_2$ mit WordNet

- $(W_2^1, W_2^{1(1)}, W_2^{1(2)}, ..., W_2^{1(k_1)})$
- $(W_2^2, W_2^{2(1)}, W_2^{2(2)}, ..., W_2^{2(k_2)})$
- ...
- $(W_2^m, W_2^{m(1)}, W_2^{m(2)}, ..., W_2^{m(k_m)})$

BEISPIEL: *present paper*

- *(paper, composition, report, theme)*
- *(paper, newspaper)*

# Algorithmus 1 (2)

SCHRITT 2: Bildung von Wortpaaren ($W_1$ - $W_2^{i(s)}$)

- ($W_1$ - $W_2^{1}$, $W_1$ – $W_2^{1(1)}$, $W_1$ - $W_2^{1(2)}$ , $W_1$ – $W_2^{1(k_1)}$)
- ($W_1$ - $W_2^{2}$, $W_1$ – $W_2^{2(1)}$, $W_1$ - $W_2^{2(2)}$ , $W_1$ - $W_2^{2(k_2)}$)
- ...
- ($W_1$ - $W_2^{m}$, $W_1$ – $W_2^{m(1)}$, $W_1$ - $W_2^{m(2)}$ , $W_1$ - $W_2^{m(k_m)}$)

BEISPIEL: *present paper*

- *(present-paper, present-composition, present-report, present-theme)*
- *(present-paper, present-newspaper)*

# Algorithmus 1 (3)

SCHRITT 3: Internet-Suche und Sense-Ranking

- Suchmaschinen-Anfragen mit folgenden Mustern:

  1) (“$W_1\ W_2^{i}$” OR “$W_1\ W_2^{i\ (1)}$” OR “$W_1\ W_2^{i\ (2)}$“ OR “$W_1\ W_2^{i\ (k_i)}$”

  2) ((“$W_1$ NEAR $W_2^{i}$) OR ($W_1$ NEAR $W_2^{i\ (1)}$) OR ($W_1$ NEAR $W_2^{i\ (2)}$) OR ($W_1$ NEAR $W_2^{i\ (k_i)}$))

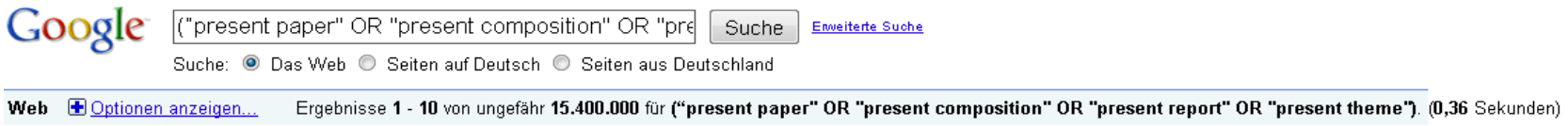
- Ranking nach Anzahl der Treffer

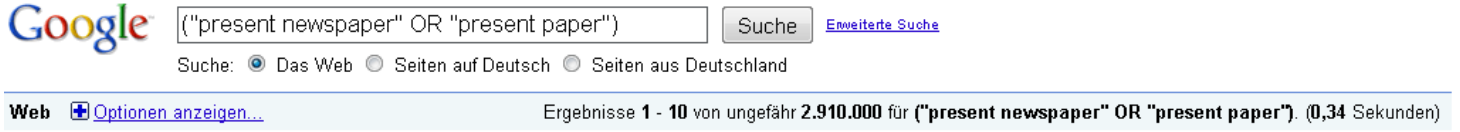
BEISPIEL: *present paper*

- *(“present paper" OR "present composition" OR "present report" OR "present theme")*
- *("present paper" OR "present newspaper")*

# Algorithmus 1 (3)

SCHRITT 3: Internet-Suche und Sense-Ranking



# Algorithmus 1 (3)

SCHRITT 3: Internet-Suche und Sense-Ranking

- Suchmaschinen-Anfragen mit folgenden Mustern:

  1) (“$W_1\ W_2^{i}$” OR “$W_1\ W_2^{i\ (1)}$” OR “$W_1\ W_2^{i\ (2)}$“ OR “$W_1\ W_2^{i\ (k_i)}$”

  2) ((“$W_1$ NEAR $W_2^{i}$) OR ($W_1$ NEAR $W_2^{i\ (1)}$) OR ($W_1$ NEAR $W_2^{i\ (2)}$) OR ($W_1$ NEAR $W_2^{i\ (k_i)}$))

- Ranking nach Anzahl der Treffer

BEISPIEL: *present paper*

- *(“present paper" OR "present composition" OR "present report" OR "present theme") (15.400.000)*
- *("present paper" OR "present newspaper") (2.910.000)*

# Algorithmus 2 (2)

SCHRITT 1: Mögliche Bedeutungen von V – N:

$< v_1, v_2, \ldots, v_h >$ und $< n_1, n_2, \ldots, n_l >$

# Algorithmus 2 (3)

SCHRITT 2: Sense-Ranking von N

- Algorithmus 1 anwenden
- Die ersten *t* Möglichkeiten werden beibehalten (z.B. Top 4)
- Alle anderen Möglichkeiten werden verworfen

# Algorithmus 2 (4)

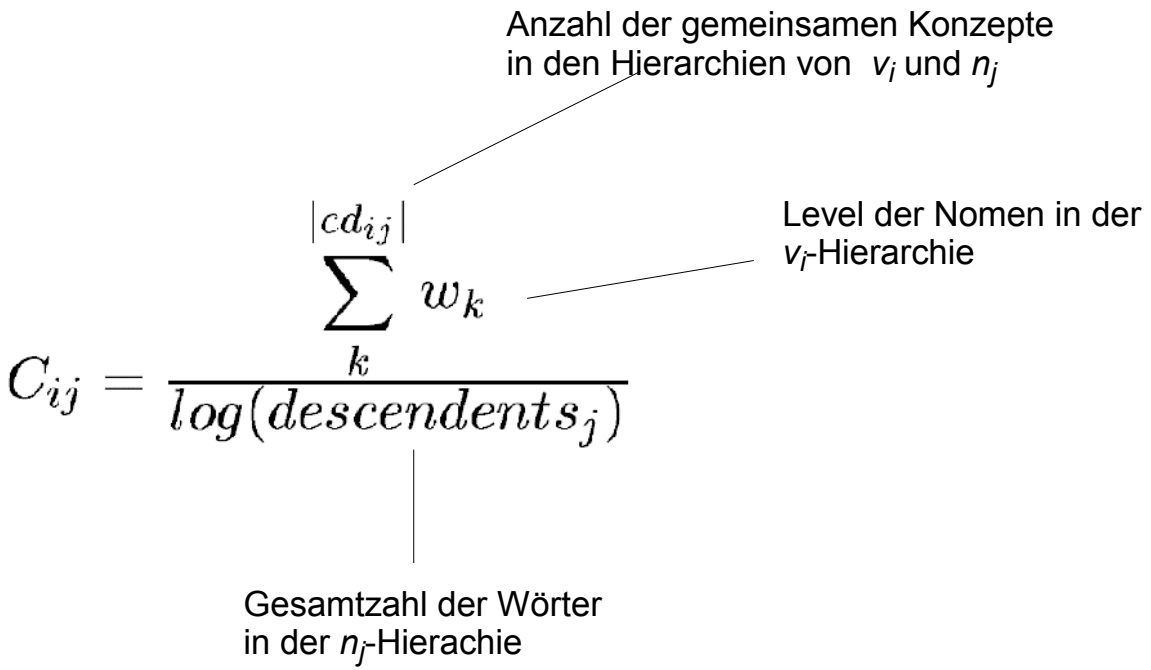
SCHRITT 3: Berechnen der *konzeptuellen Dichte* von jedem möglichen Paar $v_i$ – $n_j$

- Betrachtung der Glossen der Subhierachie des Verbs
- Die Nomen dieser Glossen bestimmen den Nomen-Kontext von *v*
- Bestimmen der Nomen in der Subhierachie von *n*
- Bilden der Schnittmenge $cd_{ij}$ : *common concepts*
- Berechnung der konzeptuellen Dichte $C_{ij}$ mit folgender Formel

$$C_{ij} = \frac{\sum_{k}^{|cd_{ij}|} w_k}{log(descendents_j)}$$

# Algorithmus 2 (5)

SCHRITT 3: Berechnen der *konzeptuellen Dichte* von jedem möglichen Paar $v_i$ – $n_j$



# Algorithmus 2 (6)

SCHRITT 4: Ranking jedes Paares $v_i$ – $n_j$ mit $C_{i\,j}$

# Algorithmus 2 (Beispiel aus Paper)

Gegeben: Verb-Nomen-Kollokation *revise law*

- Algorithmus 1 (mit AltaVista) ergibt folgendes Ranking:
  - *law* #2 (2829)
  - *law* #3 (648)
  - *law* #4 (640)
  - *law* #6 (397)
  - *law* #1 (224)
  - *law* #5 (37)
  - *law* #7 (0)
- $t = 2$
- Behalten von Bedeutung #2 und #3

REVISE
*1. {revise#1}*
*=> { rewrite}*
*2. {retool, revise#2}*
*=> { reorganize, shake up}*

LAW
*1. { law#1, jurisprudence}*
*=> {collection, aggregation, accumulation, assemblage}*
*2. {law#2}*
*=> {rule, prescript} ...*
*3. {law#3, natural law}*
*=> { concept, conception, abstract}*
*4. {law#4, law of nature}*
*=> { concept, conception, abstract}*
*5. {jurisprudence, law#5, legal philosophy}*
*=> { philosophy}*
*6. {law#6, practice of law}*
*=> { learned profession}*
*7. {police, police force, constabulary, law#7}*
*=> { force, personnel}*

# Algorithmus 2 (Beispiel aus Paper)

4 mögliche Kombinationen:

1) $v_1 - n_2$: *revise* #1/2 – *law* #2/7
2) $v_1 - n_3$: *revise* #1/2 – *law* #3/7
3) $v_2 - n_2$: *revise* #2/2 – *law* #2/7
4) $v_2 - n_3$: *revise* #2/2 – *law* #3/7

# Algorithmus 2 (Beispiel aus Paper)

Anzahl der gemeinsamen Konzepte in den Hierarchien von $v_i$ und $n_j$

Level der Nomen in der $v_i$-Hierarchie

$$C_{ij} = \frac{\sum_{k}^{|cd_{ij}|} w_k}{log(descendents_j)}$$

Gesamtzahl der Wörter in der $n_j$-Hierachie

| | $\|cd_{ij}\|$ | | $descendants_j$ | | $C_{ij}$ | |
|---|---|---|---|---|---|---|
| | $n_2$ | $n_3$ | $n_2$ | $n_3$ | $n_2$ | $n_3$ |
| $v_1$ | 5 | 4 | 975 | 1265 | 0.30 | 0.28 |
| $v_2$ | 0 | 0 | 975 | 1265 | 0 | 0 |

größte konzeptuelle Dichte: $C_{12}$ = 0.30 -> $v_1$ – $n_2$: *revise* #1/2 – *law* #2/7

# Evaluation

Test auf SemCor 1.6:

- 200 Verb-Nomen-Paare
- 127 Adjektiv-Nomen-Paare
- 57 Adverb-Verb-Paare

| | top 1 | top 2 | top 3 | top 4 |
|---|---|---|---|---|
| noun | 76% | 83% | 86% | 98% |
| verb | 60% | 68% | 86% | 87% |
| adjective | 79.8% | 93% | | |
| adverb | 87% | 97% | | |

nur Algorithmus 1

| | top 1 | top 2 | top 3 | top 4 |
|---|---|---|---|---|
| noun | 86.5% | 96% | 97% | 98% |
| verb | 67% | 79% | 86% | 87% |
| adjective | 79.8% | 93% | | |
| adverb | 87% | 97% | | |

beide Algorithmen

# Evaluation

Vergleich mit anderen Methoden:

| | Base line | Stetina | Yarowsky | Our method |
|---|---|---|---|---|
| noun | 80.3% | 85.7% | 93.9% | 86.5% |
| verb | 62.5% | 63.9% | - | 67% |
| adjective | 81.8% | 83.6% | - | 79.8 |
| adverb | 84.3% | 86.5% | - | 87% |
| AVERAGE | 77% | 80% | - | 80.1% |

- Base line: meistgebrauchte Bedeutung (erster Eintrag in WordNet)
- Stetina: Diskurs-Kontext, semantische Distanz zwischen Wörtern, benutzt WordNet
- Yarowsky: "one sense per collocation", "one sense per discourse"

# Probleme / Erweiterungen

Probleme:

- SemCor wurde in größerem Kontext annotiert (Satzkontext, Diskurs)
- Wortpaare haben keinen solchen Kontext
  → Betrachtung von allen Wortpaaren im Satz
  (z.B. Subjekt-Verb, Subjekt-Objekt, Verb-Subjekt, Verb-Objekt)
- Nomen-Nomen-Paare, Verb-Verb-Paare
  → Suche mit “NEAR”
- Adjektive und Adverben haben keine Hierarchien (Algorithmus 2 nicht möglich)
- (zu) feine Bedeutungsunterscheidung in WordNet

# Quellen

- http://www.cse.unt.edu/~rada/
- http://www.hlt.utdallas.edu/~moldovan/