Unsupervised Large Vocabulary Word Sense Disambiguation with Graph-based Algorithm for Sequence Data Labeling

Rada Mihalcea(2005)

Christian Meyer

03.01.2011

Proseminar „Word Sense Disambiguation“

Stefan Thater

Inhalt

- Einführung
- Der Algorithmus
  - Graph Building
  - Graph-based Ranking
  - Label Assignment
  - Pseudocode
  - Laufzeit
- Evaluation
- Zusammenfassung

Einführung

Ambiguität:

- Wort hat mehrere Bedeutungen
- Finden der richtigen Bedeutung im gegeben Kontext

„Der Maskierte sitzt auf der Bank und zählt seine Kohle.“

Einführung

Wechselseitig desambiguierende Wörter:

The bell

- S: (n) **bell** (a hollow device made of metal that makes a ringing sound when struck)
- S: (n) doorbell, **bell**, buzzer (a push button at an outer door that gives a ringing or b
- S: (n) **bell**, toll (the sound of a bell being struck) *"saved by the bell"; "she heard*
- S: (n) **bell**, ship's bell ((nautical) each of the eight half-hour units of nautical time sign
- S: (n) **bell**, bell shape, campana (the shape of a bell)
- S: (n) **Bell**, Melville Bell, Alexander Melville Bell (a phonetician and father of Alexa
- S: (n) **Bell**, Vanessa Bell, Vanessa Stephen (English painter; sister of Virginia Wool
- S: (n) **Bell**, Alexander Bell, Alexander Graham Bell (United States inventor (born in
- S: (n) chime, **bell**, gong (a percussion instrument consisting of a set of tuned bells tha
- S: (n) **bell** (the flared opening of a tubular device)

(to) ring

- S: (v) **ring**, peal (sound loudly and sonorously) *"the bells rang"*
- S: (v) resound, echo, **ring**, reverberate (ring or echo with sound)
- S: (v) **ring**, knell (make (bells) ring, often for the purposes of musi
- S: (v) call, telephone, call up, phone, **ring** (get or try to get into co
- S: (v) surround, environ, **ring**, skirt, border (extend on all sides of
- S: (v) **ring**, band (attach a ring to the foot of, in order to identify)

The bell rings.

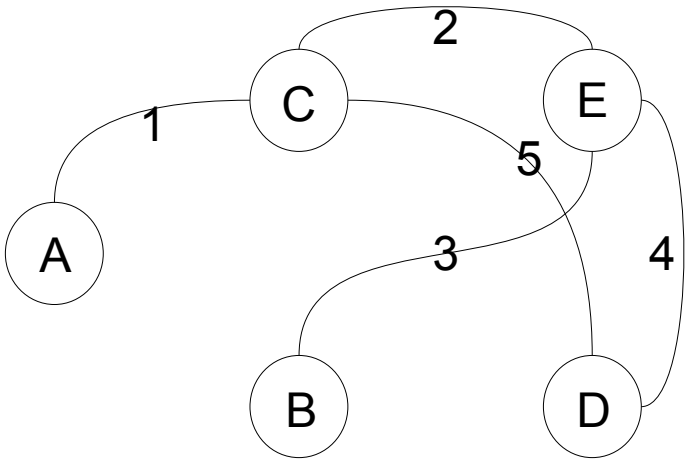
- S: (n) **bell** (a hollow device made of metal that makes a ringing sound when struck)
- S: (v) **ring**, peal (sound loudly and sonorously) *"the bells rang"*

Einführung

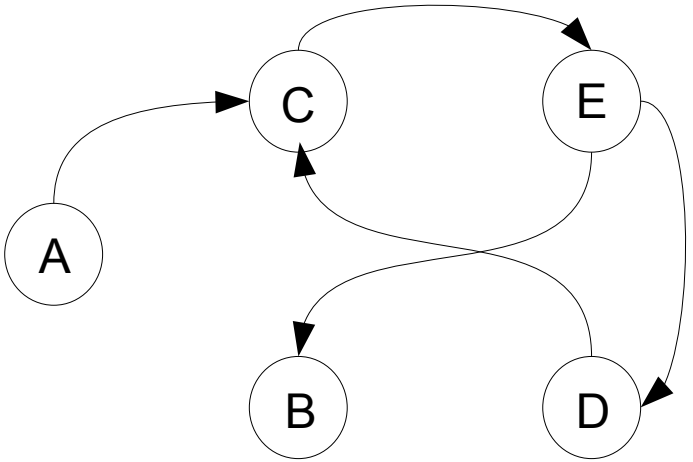
Was ist ein Graph?

- Graph G ist ein Tupel (V, E)
- Knotenmenge V
- Kantenmenge E
- Knoten sind durch die Kanten verbunden
- Gerichtete und ungerichtete Graphen
- Gewichtung
- Multigraphen
- Hypergraphen
- Zyklen

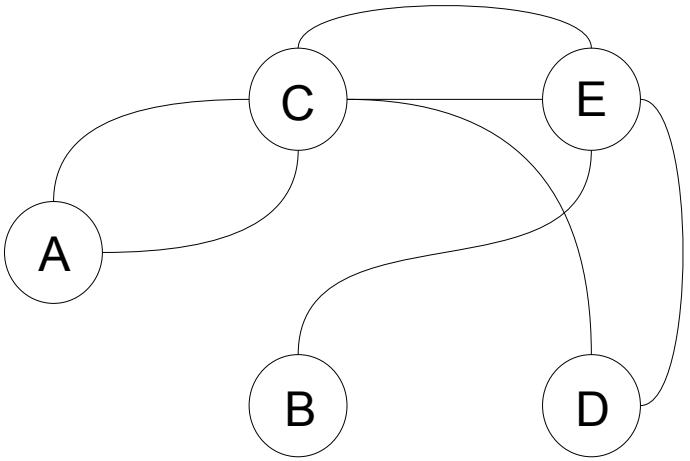
Ungerichteter Graph



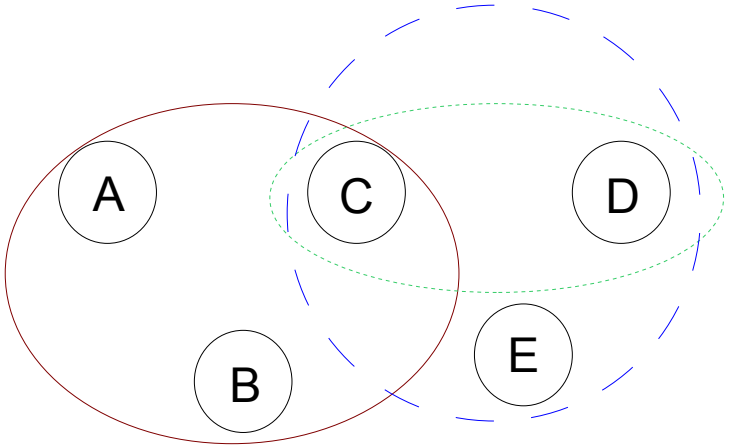
Gerichteter Graph



Multigraph



Hypergraph



Der Algorithmus

- Berücksichtigung aller Bedeutungen der Wörter im Kontext
- Gewichtung der Abhängigkeiten verschiedener Bedeutungen voneinander
- Repräsentation der Bedeutungen und der Abhängigkeiten durch einen Graph
- Berechnung der wahrscheinlichsten Bedeutungen in 3 Teilschritten

Der Algorithmus

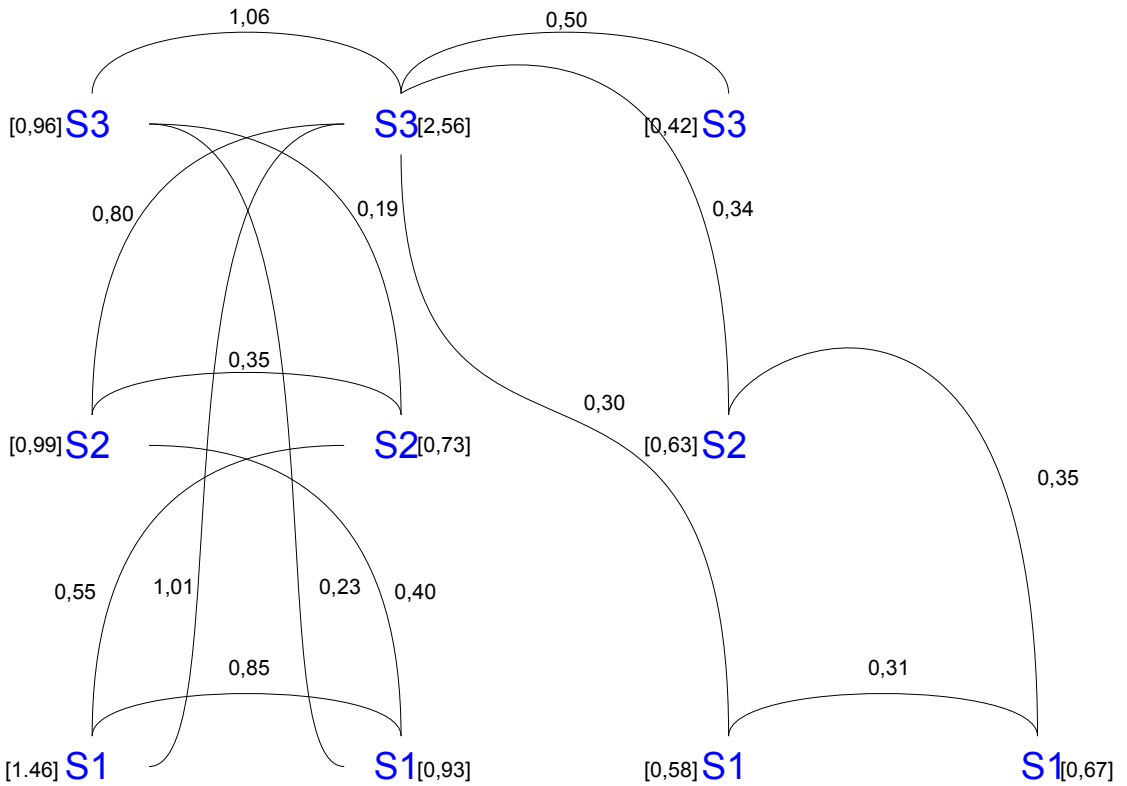
3 Teilschritte:

· Graph Building

· Graph-based Ranking

· Label Assignment

„The church bells ring on Sunday.“



bell (S1) ring(S3) church(S2) sunday(S1)

Der Algorithmus

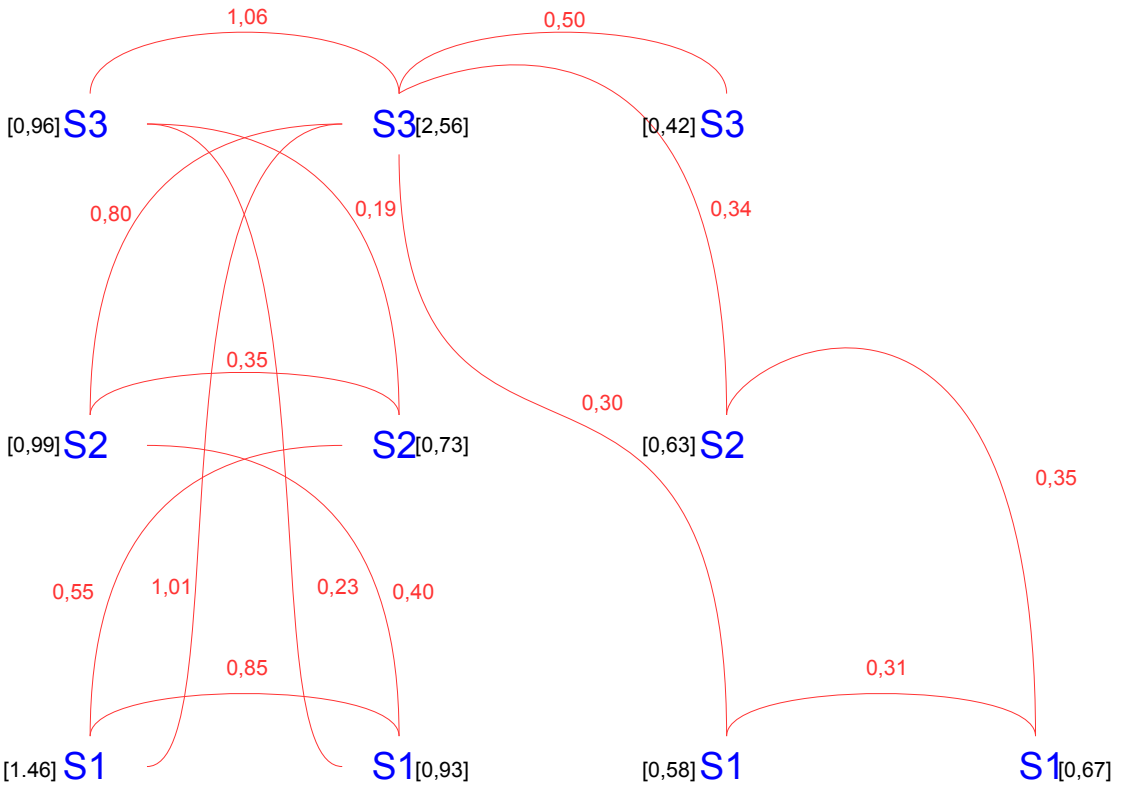
3 Teilschritte:

· Graph Building

· Graph-based Ranking

· Label Assignment

„The church bells ring on Sunday.“



Der Algorithmus

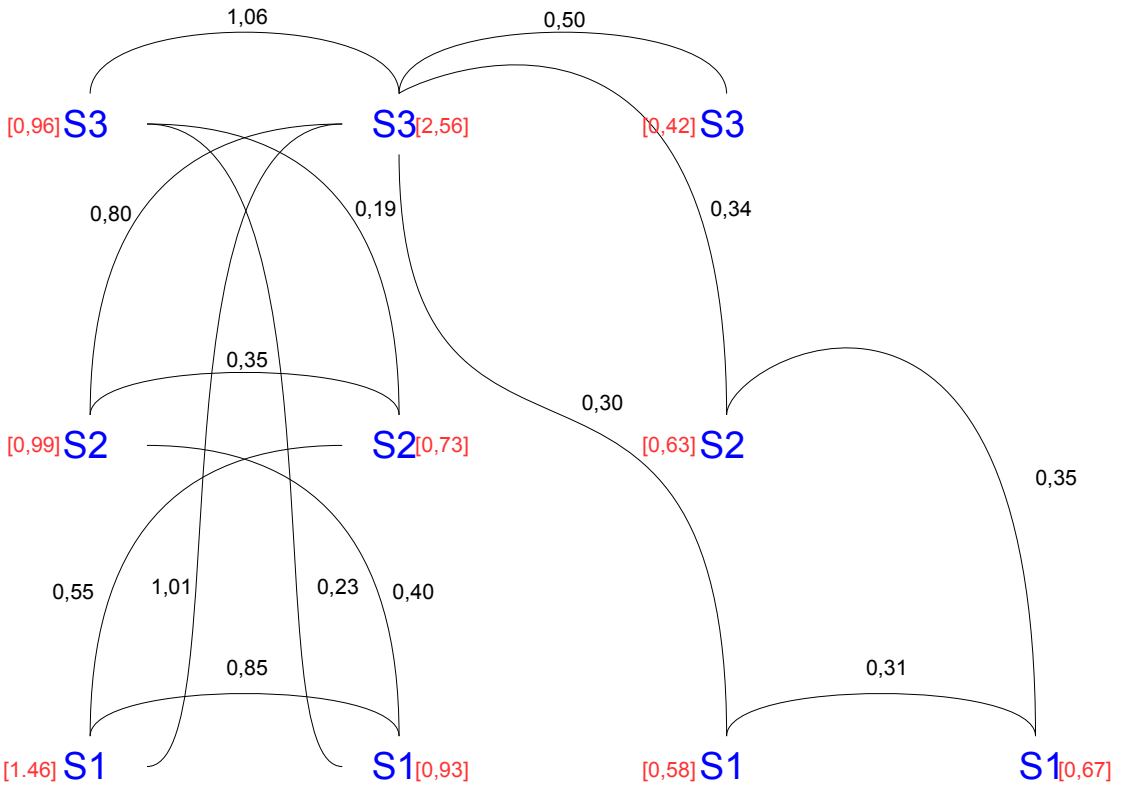
3 Teilschritte:

· Graph Building

· Graph-based Ranking

· Label Assignment

„The church bells ring on Sunday.“



Der Algorithmus

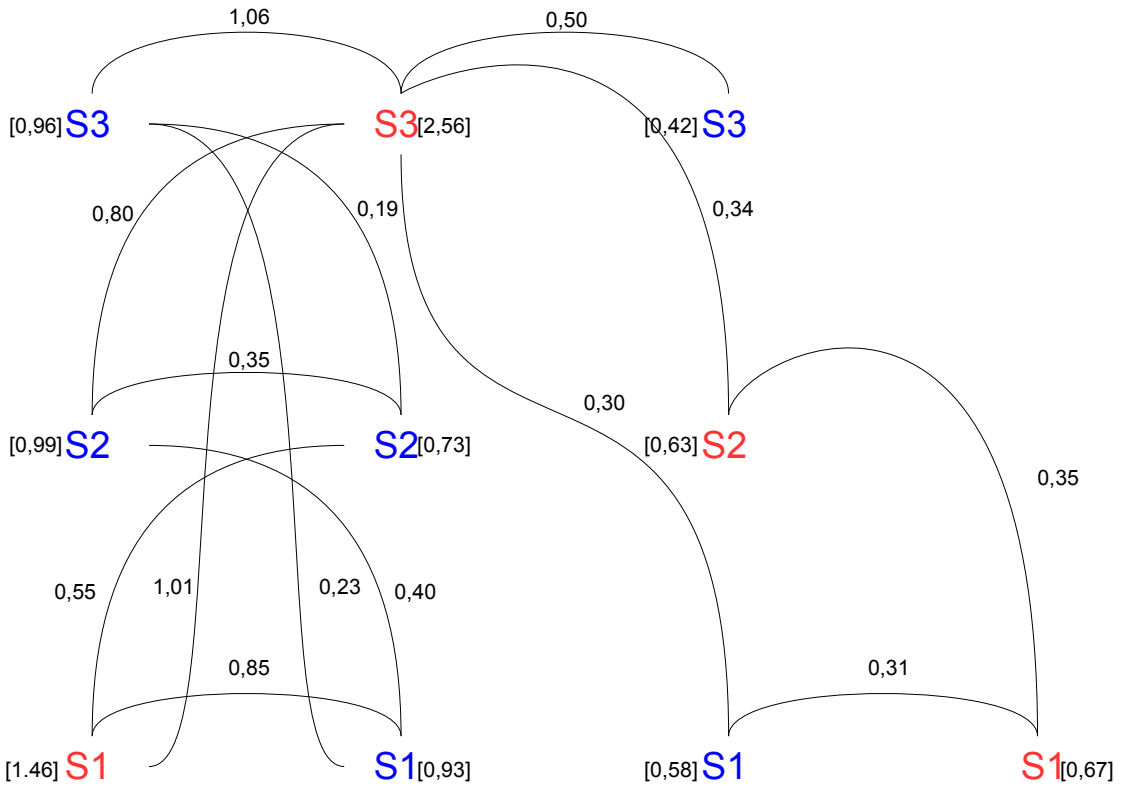
3 Teilschritte:

· Graph Building

· Graph-based Ranking

· Label Assignment

„The church bells ring on Sunday.“



bell (S1) ring(S3) church(S2) sunday(S1)

Graph Building

- Ein Knoten pro Wortbedeutung
- Gewinnung der gewichteten Kanten
  - Wahrscheinlichkeiten für gleichzeitiges Auftreten oder bedingte Wahrscheinlichkeiten mithilfe eines annotierten Korpus ($P(l^t_{wi}, l^t_{wj})$, $P(l^t_{wi} \mid l^t_{wj})$ oder $P(l^t_{wi} \mid l^t_{wj}) \times P(w_i \mid l^t_{wi})$)
  - Ähnlichkeit der Definitionen zweier Wortbedeutungen (A, B)
    - Eliminieren der Stop-words
    - Zählen der gemeinsamen Wörter $CO = |(A \cap B)|$
    - Normalisierung: $CO = CO / |A| + |B|$
    - Optional: Angeben einer maximalen Distanz

S3 S3

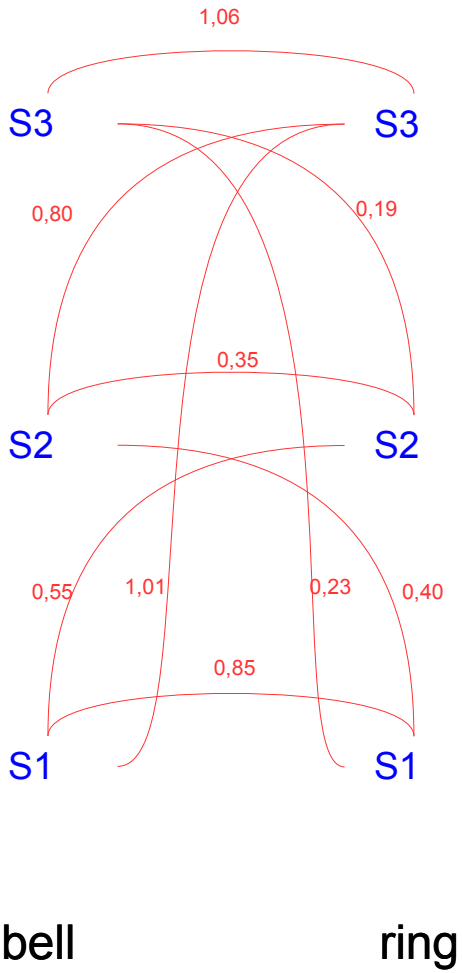
S2 S2

S1 S1

bell ring

Graph Building

- Ein Knoten pro Wortbedeutung
- Gewinnung der gewichteten Kanten
  - Wahrscheinlichkeiten für gleichzeitiges Auftreten oder bedingte Wahrscheinlichkeiten mithilfe eines annotierten Korpus ($P(l^t_{wi}, l^t_{wj})$, $P(l^t_{wi} \mid l^t_{wj})$ oder $P(l^t_{wi} \mid l^t_{wj}) \times P(w_i \mid l^t_{wi})$)
  - Ähnlichkeit der Definitionen zweier Wortbedeutungen (A, B)
    - Eliminieren der Stop-words
    - Zählen der gemeinsamen Wörter CO = |(A ∩ B)|
    - Normalisierung: CO = CO / |A| + |B|
    - Optional: Angeben einer maximalen Distanz



Graph-based Ranking

- Knoten bewerten sich gegenseitig
  - Ein Knoten vergibt Punkte an jeden Zielknoten, mit dem er verbunden ist.
  - Je mehr Punkte ein Knoten bekommt, desto mehr vergibt er auch.

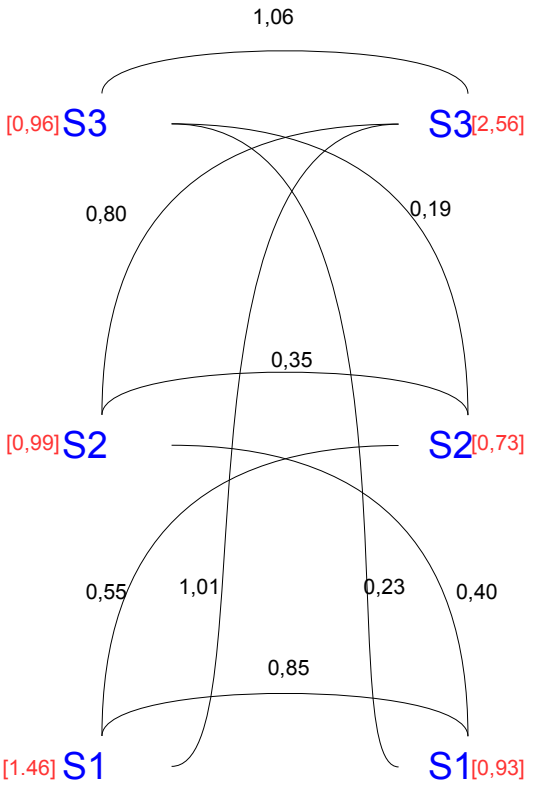
- PageRank (Brin and Page, 1998)

$$P(V_a) = (1-d) + d * \sum_{V_b \in In(V_a)} \frac{P(V_b)}{|Out(V_b)|}$$

- Weighted PageRank

$$WP(V_a) = (1-d) + d \sum_{V_b \in In(V_a)} \frac{w_{ba}}{\sum_{V_c \in Out(V_b)} w_{bc}} WP(V_b)$$

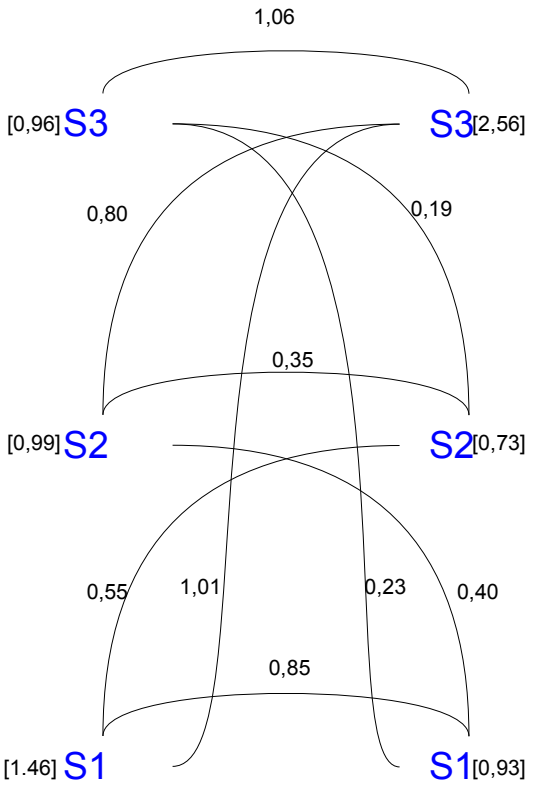
- Erreichen stabiler Werte



bell ring

Label Assignment

- Extrahieren der Bedeutungen mit der höchsten Punktzahl
- Zuweisen der Bedeutungen zu den entsprechenden Wörtern



bell(S1) ring(S3)

Pseudocode

Graph Building

Build graph G of label dependencies

for $i = 1$ to N **do**
  for $j = i + 1$ to N **do**
    if $j - i > MaxDist$ **then**
      $break$
    end if
    for $t = 1$ to N_{w_i} **do**
      for $s = 1$ to N_{w_j} **do**
        $weight \leftarrow Dependency(l^t_{w_i}, l^s_{w_j}, w_i, w_j)$
        if $weight > 0$ **then**
          $AddEdge(G, l^t_{w_i}, l^s_{w_j}, weight)$
        end if
      end for
    end for
  end for
end for

Pseudocode

Graph-based Ranking

Score vertices in G

1: **repeat**
2: **for all** $V_a \in Vertices(G)$ **do**
3: $$WP(V_a) = (1-d) + d* \sum_{V_b \in In(V_a)} w_{ba} WP(V_b) / \sum_{V_c \in Out(V_b)} w_{bc}$$
4: **end for**
5: **until** convergence of scores $WP(V_a)$

Pseudocode

Label Assignment

Label assignment

1: **for** $i = 1$ to N **do**
2: $\quad l_{w_i} \leftarrow argmax\{WP(l^t_{w_i})|t = 1..N_{w_i}\}$
3: **end for**

Laufzeit

· Graph-based Sequence Data Labeling:

$$O(C \sum_{i=1}^{n} \sum_{j=i+1}^{i+MaxDist} (N_{w_i} \times N_{w_j}))$$

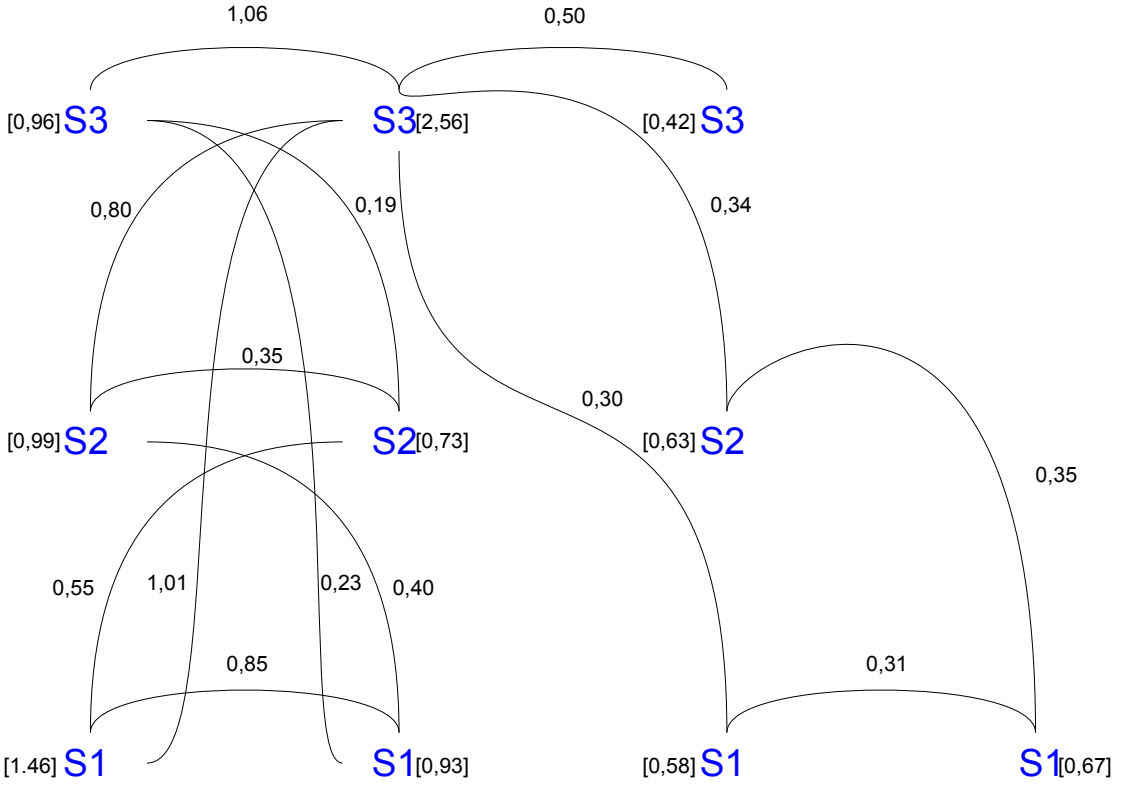
· Basic Sequence Data Labeling:

$$O(\prod_{i=1}^{n} N_{w_i})$$

· Individual Data Labeling:

$$O(\sum_{i=1}^{n} N_{w_i})$$

„The church bells ring on Sunday.“



bell (S1) ring(S3) church(S2) sunday(S1)

Evaluation

Senseval-2 Datenbank

- 3 Dokumente aus Penn Treebank
- 2,456 open-class Wörter
- All-word task mit Bedeutungshierarchie

  → feine und grobe Disambiguierung
- Viele Vergleichsdaten

Evaluation

Einreichungen an Senseval-2 mit besten Ergebnissen

Language	Task	No. of submissions	No. of teams	IAA	Baseline	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40

Evaluation

Vergleich mit Lesk Algorithmus:

Gemeinsamkeiten:

- Basiert auf Lexikondefinitionen
- Content Overlap
- Normalisierung
- Keine Einbeziehung der Ordnung der Definitionen

Einziger Unterschied:

individuelle Disambiguierung

Evaluation

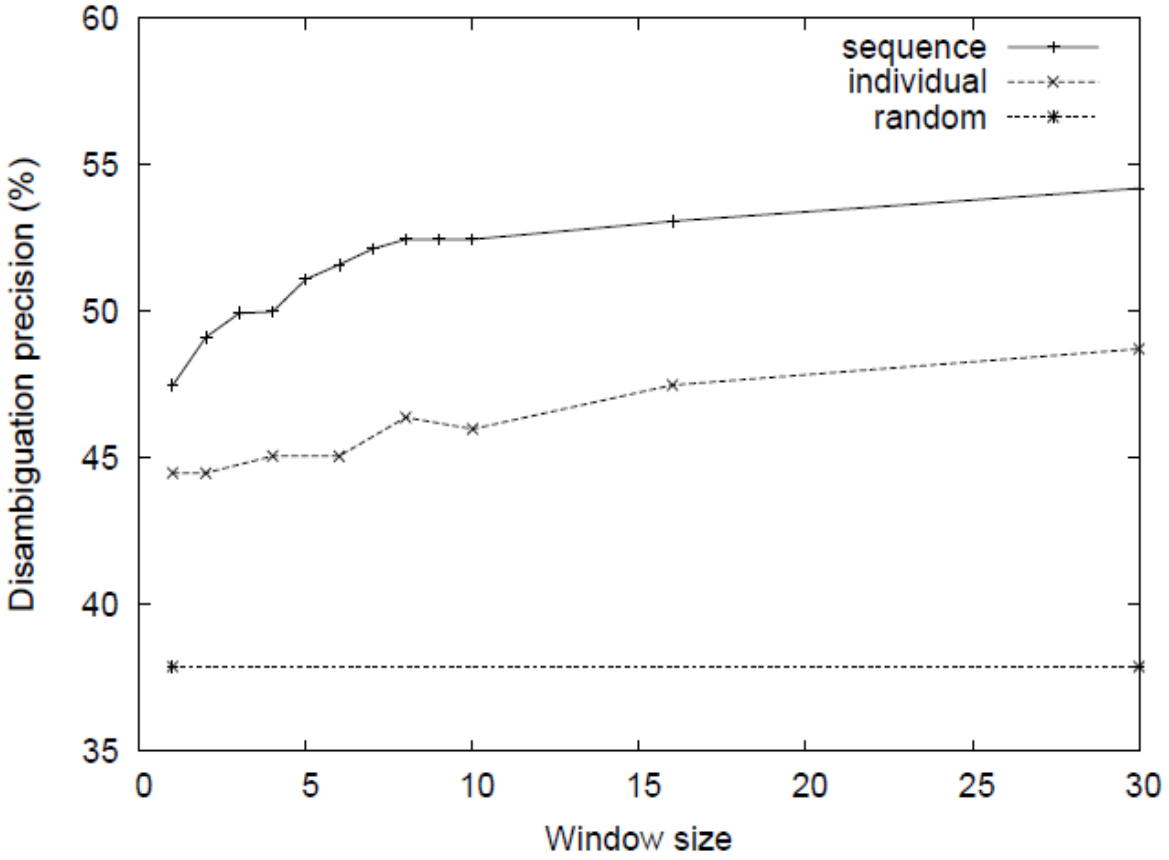
Part-of speech	Fine-grained sense distinctions						Coarse-grained sense distinctions					
	Random baseline		Individual (Lesk)		Sequence (graph-based)		Random baseline		Individual (Lesk)		Sequence (graph-based)	
	P	R	P	R	P	R	P	R	P	R	P	R
Noun	41.4%	19.4%	50.3%	23.6%	57.5%	27.0%	42.7%	20.0%	51.4%	24.1%	58.8%	27.5%
Verb	20.7%	3.9%	30.5%	5.7%	36.5%	6.9%	22.8%	4.3%	31.9%	6.0%	37.9%	7.1%
Adjective	41.3%	9.3%	49.1%	11.0%	56.7%	12.7%	42.6%	42.6%	49.8%	11.2%	57.6%	12.9%
Adverb	44.6%	5.2%	64.6%	7.6%	70.9%	8.3%	40.7%	4.8%	65.3%	7.7%	71.9%	8.5%
ALL	37.9%	37.9%	48.7%	48.7%	**54.2%**	**54.2%**	38.7%	38.7%	49.8%	49.8%	**55.3%**	**55.3%**

Reduzierte Fehlerrate von 10,7% bzw 11%

→ Unabhängigkeit von der Klassifizierungsgranularität

Evaluation

Reduzierte Kontextgröße:



- Beide Algorithmen profitieren von größerem Kontext
- Stabile Verbesserung bei der Präzision
  → unabhängig vom Kontext

Zusammenfassung

- Unüberwachter Algorithmus
- Erzeugung eines Graphen für die Wortsequenz
- Repräsentation der Wortbedeutungen durch Knoten
- Gegenseitiges Empfehlen der Bedeutungen
- Zeitgleiches Zuweisen aller Bedeutungen
- Reduzierung der Fehlerrate

Zusammenfassung

Referenzen:

Paper (Mihalcea 2005):
http://www.aclweb.org/anthology-new/H/H05/H05-1052.pdf

Graphentheorie:
http://de.wikipedia.org/wiki/Graph_(Graphentheorie)

Senseval Homepage:
http://www.senseval.org/

Senseval-2 Homepage:
http://86.188.143.199/senseval2/

Artikel über Senseval-2:
http://www.senseval.org/publications/senseval.pdf

Zusammenfassung

Was offen bleibt:

Brauchen wir den Algorithmus?

- schlechtere Laufzeit
- bessere Ergebnisse

Zusammenfassung

Part-of speech	Fine-grained sense distinctions						Coarse-grained sense distinctions					
	Random baseline		Individual (Lesk)		Sequence (graph-based)		Random baseline		Individual (Lesk)		Sequence (graph-based)	
	P	R	P	R	P	R	P	R	P	R	P	R
Noun	41.4%	19.4%	50.3%	23.6%	57.5%	27.0%	42.7%	20.0%	51.4%	24.1%	58.8%	27.5%
Verb	20.7%	3.9%	30.5%	5.7%	36.5%	6.9%	22.8%	4.3%	31.9%	6.0%	37.9%	7.1%
Adjective	41.3%	9.3%	49.1%	11.0%	56.7%	12.7%	42.6%	42.6%	49.8%	11.2%	57.6%	12.9%
Adverb	44.6%	5.2%	64.6%	7.6%	70.9%	8.3%	40.7%	4.8%	65.3%	7.7%	71.9%	8.5%
ALL	37.9%	37.9%	48.7%	48.7%	**54.2%**	**54.2%**	38.7%	38.7%	49.8%	49.8%	**55.3%**	**55.3%**

Language	Task	No. of submissions	No. of teams	IAA	Baseline	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40