Unsupervised Word Sense Disambiguation Rivaling Supervised Methods (David Yarowsky)

Kokou Valentin

29.11.2010

Proseminar Word Sense Disambiguation

Word Sense Disambiguation

- Vorkommen mehrdeutiger Worte
- Bestimmung aller möglichen Bedeutungen
- Auswahl der richtigen Bedeutung

Überwachte Lernverfahren

- Klassifikator für jedes Wort
- Sehr hohe Präzision
- Sehr zeitaufwendig („teuer“)
- Benötigt großen annotierten Korpus

Unüberwachtes Lernverfahren

- konkurriert mit Performanz überw. Lernverfahren
- „günstiger“ als Überwachte Lernverfahren
  - benötigt sehr kleinen annotierten Korpus
- basierend auf 2 Eigenschaften natürlicher Sprache
  - One Sense per Collocation (OSPC)
  - One Sense per Discourse (OSPD)
- selbst korrigierend

One Sense per Collocation

- eine Bedeutung innerhalb einer Kollokation
- enge semantische Beziehung
- enger Zusammenhalt des Sachverhalts

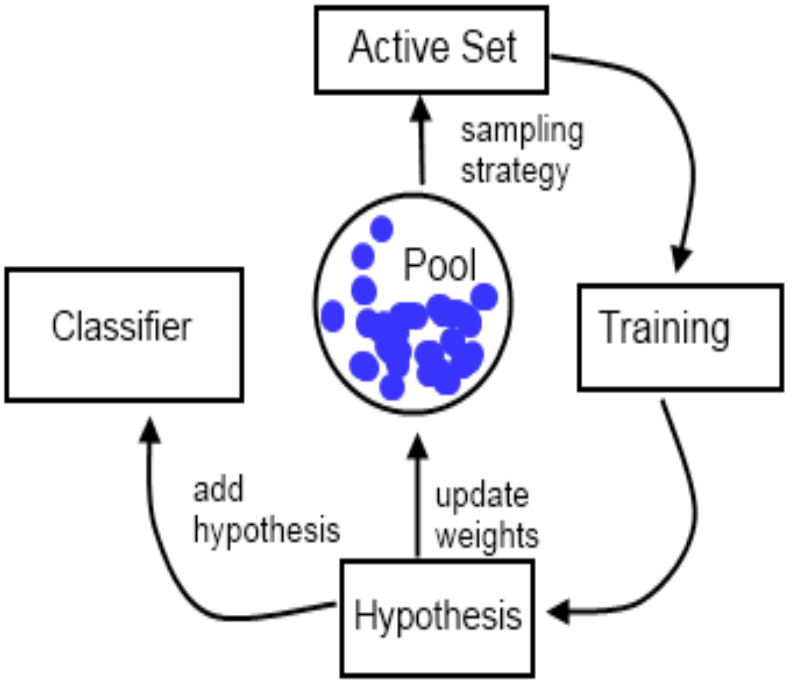
One Sense per Discourse

- Tendenz nur eine Bedeutung pro Diskurs/Dokument
- Beeinflusst Klassifikationen

Wort	Bedeutung	Genauigkeit	Anwendbar
plant	living/factory	99,8%	72.8%
bass	fish/music	100%	58.8%
palm	tree/hand	99.8%	35.5%

Bootstrapping

- Iterativer Prozess
- Vermehrung Startmenge
- Reduzierung der Restmenge



Schritt 1

- Identifiziere alle Beispiele im großen Korpus
- Speichere den Kontext als Zeilen
- Anfangs in nicht annotiertes Trainingsset

Beispiele für Trainingsset

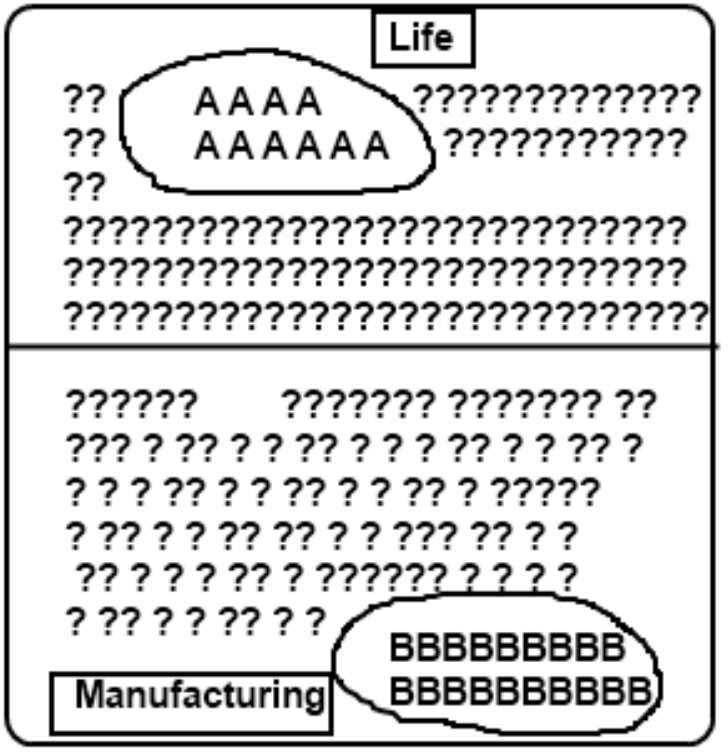
Sense	Trainings Beispiele
?	… company said the *plant* is still operating ....
?	...Computer disk drive *plant* located...
?	… divide life into *plant* and animal kingdom
?	Although thousands of *plant* and animal species...

Schritt 2

- Für jede mögliche Bedeutung des Wortes
  - Finde kleine Anzahl an Trainingsbeispielen
  - mögliche Untermenge manuell taggen

Schritt 2

- kleine Menge Kollokationen pro Bedeutung
- Alle Trainingssätze mit Label versehen
  - nicht annotierter Rest (85-90%)



A= Sense A Trainingsbeispiel
B= Sense B Trainingsbeispiel
?= nicht klassifiziertes Beispiel
Life= Menge Trainingsbeispiel mit Kollokoation „life“

Sense	Training Examples
A	… divide life into *plant* and animal kingdom
A	....Although thousands of *plant* and animal species...
?	… molecules found in *plant* and animal tissue
?	...Nissan car and truck *plant* in Japan is...
B	...Computer disk drive *plant* located...
B	… company said the *plant* is still operating ...

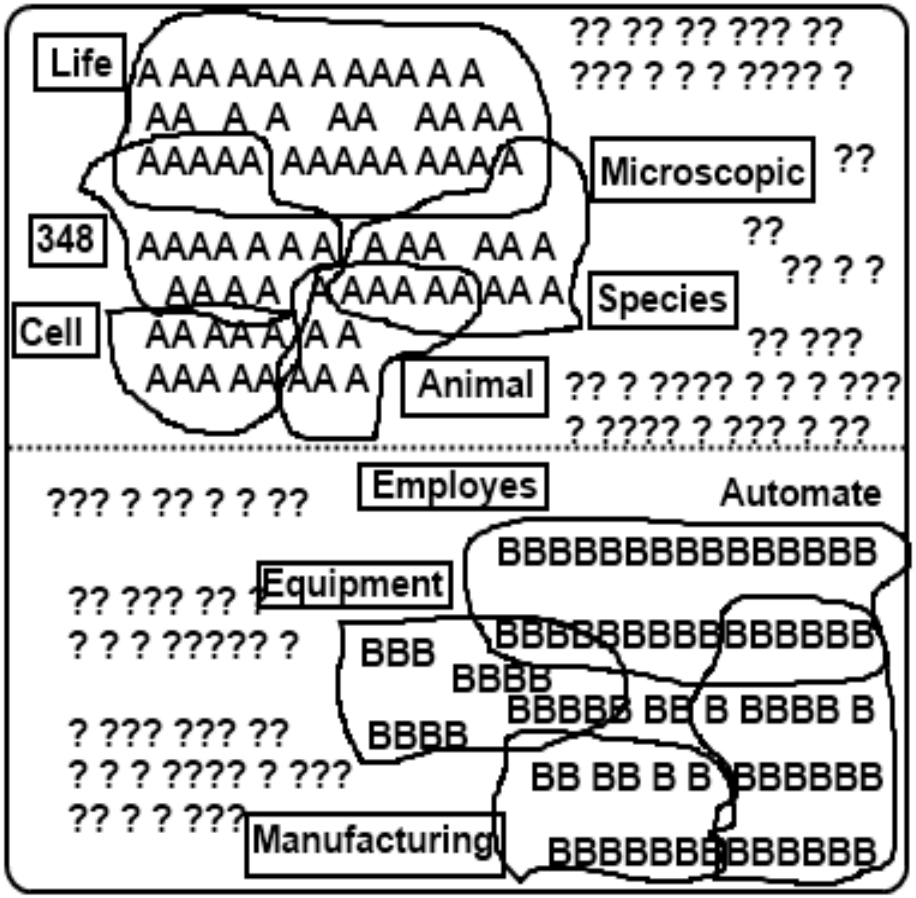
Schritt 3a

- Trainiere Klassifikationsalgorithmus auf Seed Set (Seed=Keim)
- Decision-list Algorithmus findet Kollokationen

LOGL	Kollokationen	Sense
8.10	*Plant* **life** →	**A**
7.58	Manufacturing plant →	**B**
7.39	life(within **±** 2-10 words) →	**A**
7.20	manufacturing(in **±** 2-10 words) →	**B**

Schritt 3b

- Klassifikator anwenden auf gesamtes Testset
- Sense A/B mit WK über bestimmter Schwelle
- Füge A/B zu wachsendem Seed Set hinzu
- Neue Kollokationen verweisen auf vorheriges Set



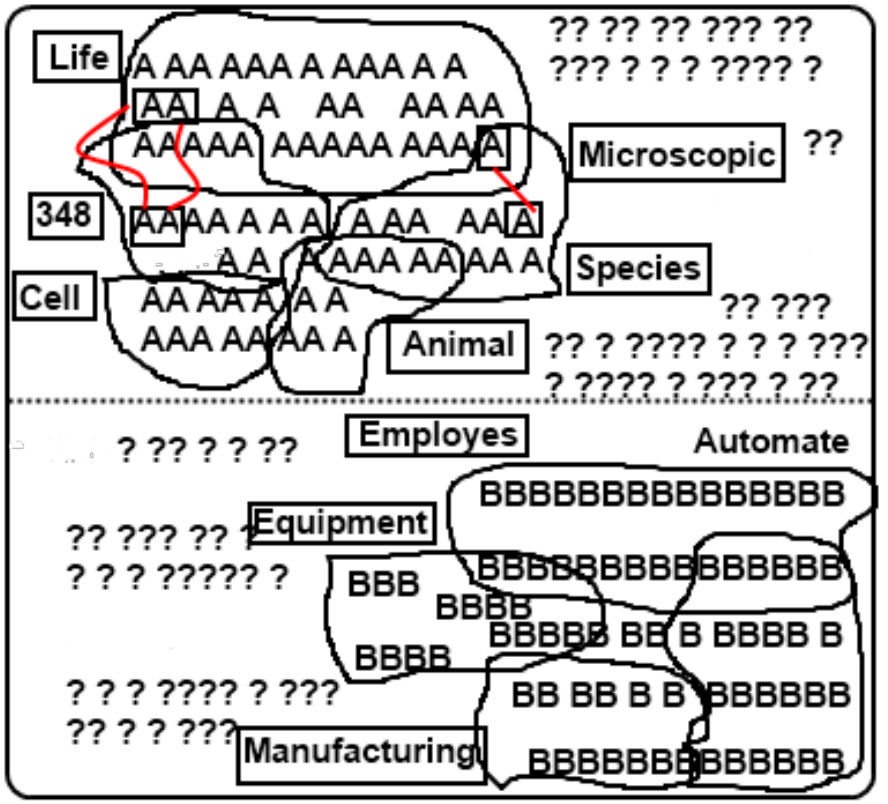
Schritt 3c

- OSPD Bedingung zum filtern und erhöhen
  - Mehrere Instanzen eines Wortes im Diskurs gleicher Sense Tag
  - Erweitern des Tag zu allen Beispielen im Diskurs
  - Bedingt durch relative Anzahl und WK
  - Erhöhung kann Verbindungen zu neuen Kollokationen bilden
  - Korrigiert fehlerhafte annotierte Beispiele

Change in tag	Diskurs Nummer	Training Samples (Von gleichem Diskurs)
A-> A	742	… the existence of *plant* and animal life...
A->A	742	....classified as either *plant or animal*...
?->A	742	…Altough bacterial and *plant* cells are enclosed
?->A	348	...an aspect of *plant life, for example*...
A->A	525	...slight within Arctic *plant* species...
B->A	525	*… are protected by plant* parts remaining from ....

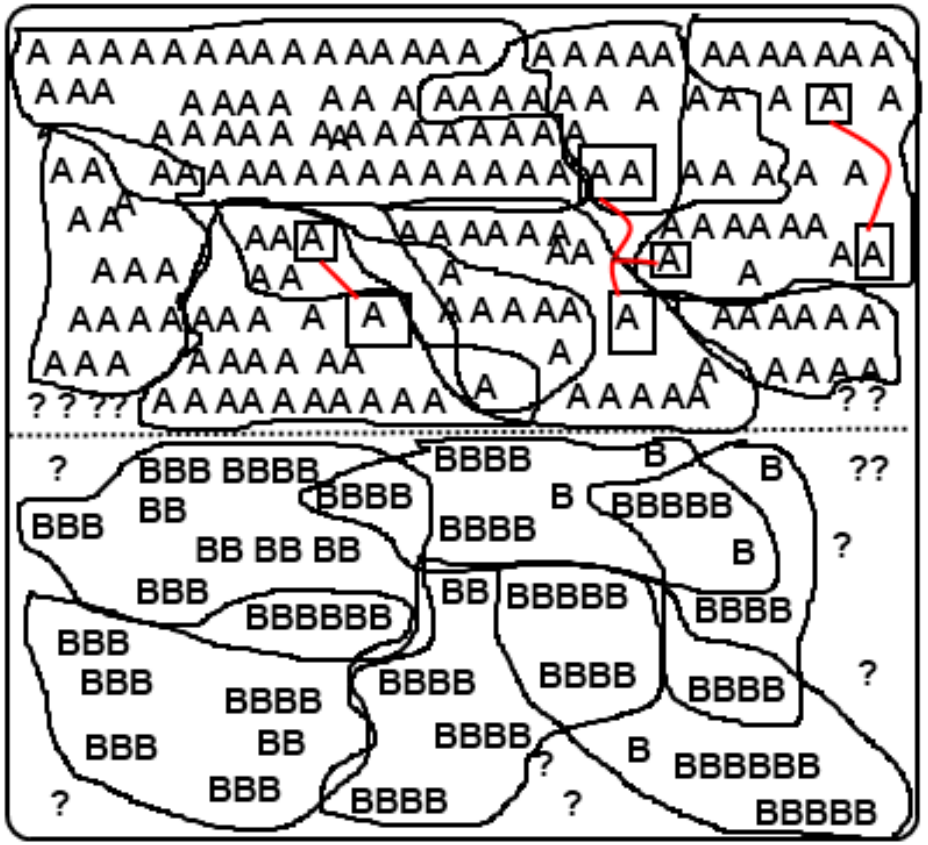
Schritt 3d

- Wiederhole Schritt 3 iterativ
- Trainings Set wird wachsen
- Restmenge wird schrumpfen



Schritt 4

- Trainings Parameter konstant halten
- Algorithmus nähert sich stabiler Restmenge
- Meisten Beispiele zeigen auf multiple Kollokationen mit der selben Bedeutung
- Decision List Alg nutzt nur höchsten Eintrag
- Keine Kombination aller zutreffenden Kollokationen



Schritt 5

- Klassifikator aus finalem Trainingsschritt
  - Anwendung auf neue Daten
  - Annotiert original nicht annotierten Korpus
  - Sense Tag und Wahrscheinlichkeiten annotieren
  - Ursprüngliche Seeds nicht länger an Spitze der Liste

Schritt 5

- Möglich, dass ein Sense A zu einem Sense B wird
  - Bei höherer Kompatibilität zur 2. Klasse
- Störanfälligkeit durch fehlerhafte Ausgangswörter nicht fatal
- Korrigiert, wenn Mehrheit der Ausgangs-WK koherenten Kollokationsraum bilden

Final decision list for *plant*

LOGL	Kollokationen	Sense
10.12	**plant growth →**	**A**
9.68	**Car (within ± k words) →**	**B**
9.64	**plant height →**	**A**
9.61	**Union (± k words) →**	**B**
9.54	**Equipment (within ± k words) →**	**B**
9.51	**assembly plant →**	**B**
9.02	**Plant species →**	**A**

Beispiel

… the loss of animal and plant species trough extenction...

- **Höchste Kollokation im Bsp. Satz „species“ wird genutzt zum Klassifizieren als Sense A (living plant)**

Woher kommen Seed Patterns?

- Nutze Wörter aus Wörterbuch Definitionen
  - Automatisch auslesen von Wörtern mit hoher Frequenz relativ zum gesamten Wörterbuch
- Nutze einzelne abgrenzende Anordnung für jede Klasse
  - Nur diese Anordnung nutzen mit ihrem Kontext
  - Bsp.: flora und factory für plant
- Markiere hervorstechende Korpusanordnungen
  - Wörter mit ungewöhnlich hohem Vorkommen
  - Menschliche Entscheidung über Bedeutung

Korrektur

- Beispiele aus Restmenge nur solange WK über dem Grenzwert
- Klassifikation kann sich ändern durch neue Beispiele
- Rückkehr zur nicht annotierten Restmenge
- Neuklassifikation durch iterativen Trainingsschritt
- Prozess ist selbstkorrekturfähig

Rettung

- Trotzdem Möglichkeit der falschen Klassifikation
  - starke Kollokationen Indikator für falsche Klasse
- 2 Techniken im Trainingsalgorithmus sollen dies verhindern

  1) taktweise Erhöhung des Kontextraums

  2) zufällige Störung des Grenzwertes der Klassifikation

OSPD Eigenschaft nutzen

- gute Performanz bei lokalen Kollokationen
  - jedes Token von Zielwort unabhängig behandelt
- Genauigkeit kann verbessert werden
  - Alle Vorkommen in Diskurs haben selbe Bedeutung
- Eigenschaft an 2 Stellen nutzbar
  - nach Schritt 3c (nach jeder Iteration)
  - Ende Schritt 4 nachdem Algo. konvergiert

OSPD Eigenschaft nutzen

- Fehlerkorrektur
  - Bsp. *plant* erscheint mehrmals in einem Diskurs
  - Tokens vom Algo. durch lokale Kollokationen niedriger Wert
    - Kann durch dominanten Tag des Diskurs überschrieben werden
  - Variablen zur Entscheidung
    - n vollständige Anzahl des Auftretens im Diskurs
    - Anzahl Vorkommen von zugeteilten Mehr-/Minderheiten der Bedeutung
    - Summe von LogL-Verhältnissen

Fehlerkorrektur II

- Wenn Hauptbedeutung größer als Nebenbedeutung
  - Grenzwert in Abhängigkeit von n
  - Nebenbedeutung wird neu markiert
- n=2 keine Neuklassifikation da nicht klar, welche dominant
- n>4 meisten lokalen Klassifikationen durch dominanten Tag überschrieben
- Im Diskurs ohne dominante Bedeutung werden Tags entfernt

Tabelle Evaluation

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Word	Sense	Samp. Size	% Major Sense	Supvsd Algrtm	Seed Training Options		
					Two Words	Dict. Defn.	Top Colls.
plant	living/ factory	7538	53.1	97.7	97.1	97.3	97.6
space	volume/ outer	5745	50.7	93.3	89.1	92.3	93.5
tank	vehicle/ container	11420	58.2	97.1	94.2	94.6	95.8
bass	fish/ music	1859	57.5	97.8	96.6	97.2	97.7
palm	tree/ hand	1572	56.1	96.5	93.9	94.7	95.8
poach	steal/ boil	585	74.9	97.1	96.6	97.2	97.7
axes	grid/ tools	1344	84.9	95.5	94.0	94.3	94.7
drug	medicine/ narcotic	1380	50.0	93.0	90.4	91.4	92.6
crane	bird/ machine	2145	78.0	96.6	59.6	93.6	94.2
AVG		3936	63.9	96.1	90.6	94.8	95.5

Tabelle Evaluation

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	**(9)**	**(10)**
Word	Sense	Samp. Size	% Major Sense	Supvsd Algrtm	Seed Training Options			**(7)+ OSDP**	
					Two Words	Dict. Defn.	Top Colls.	**End only**	**Each Iter.**
plant	living/ factory	7538	53.1	97.7	97.1	97.3	97.6	**98.3**	**98.6**
space	volume/ outer	5745	50.7	93.3	89.1	92.3	93.5	**93.3**	**93.6**
tank	vehicle/ container	11420	58.2	97.1	94.2	94.6	95.8	**96.1**	**96.5**
bass	fish/ music	1859	57.5	97.8	96.6	97.2	97.7	**98.5**	**98.8**
palm	tree/ hand	1572	56.1	96.5	93.9	94.7	95.8	**95.5**	**95.9**
poach	steal/ boil	585	74.9	97.1	96.6	97.2	97.7	**98.4**	**98.5**
axes	grid/ tools	1344	84.9	95.5	94.0	94.3	94.7	**96.8**	**97.0**
drug	medicine/ narcotic	1380	50.0	93.0	90.4	91.4	92.6	**93.9**	**93.9**
crane	bird/ machine	2145	78.0	96.6	59.6	93.6	94.2	**95.4**	**95.5**
AVG		3936	63.9	96.1	90.6	94.8	95.5	**96.1**	**96.5**

Schlussfolgerung

- Algorithmus nutzt mehrere Eigenschaften von Sprache aus
  - One Sense per Discourse
  - One Sense per Collocation
- max. Einfluss zur Modellierung großer Vielfältigkeit von Beziehungen zw. Kollokationen
- empfindlicher für größere Spanne an sprachlichen Details
- Fast gleiche Performanz wie überwachtes Lernverfahren 95.5% vs. 96.1%
- in Fällen mit OSPD besser 96.5% vs. 96.1%
- großer Sense-tagged Trainingskorpus nicht notwendig

Zusammenfassung

- Teilweise bessere Performanz (mit/ohne OSPD)
- Gutes Ergebnis bei Homonymen
- Voraussichtlich bei Polyseme weniger gut