Finding Predominant Word Senses in Untagged Text

McCarthy &al. (2004)

Proseminar „Word Sense Disambiguation“

Stefan Thater

Torsten Jachmann

17. Januar 2011

Überblick

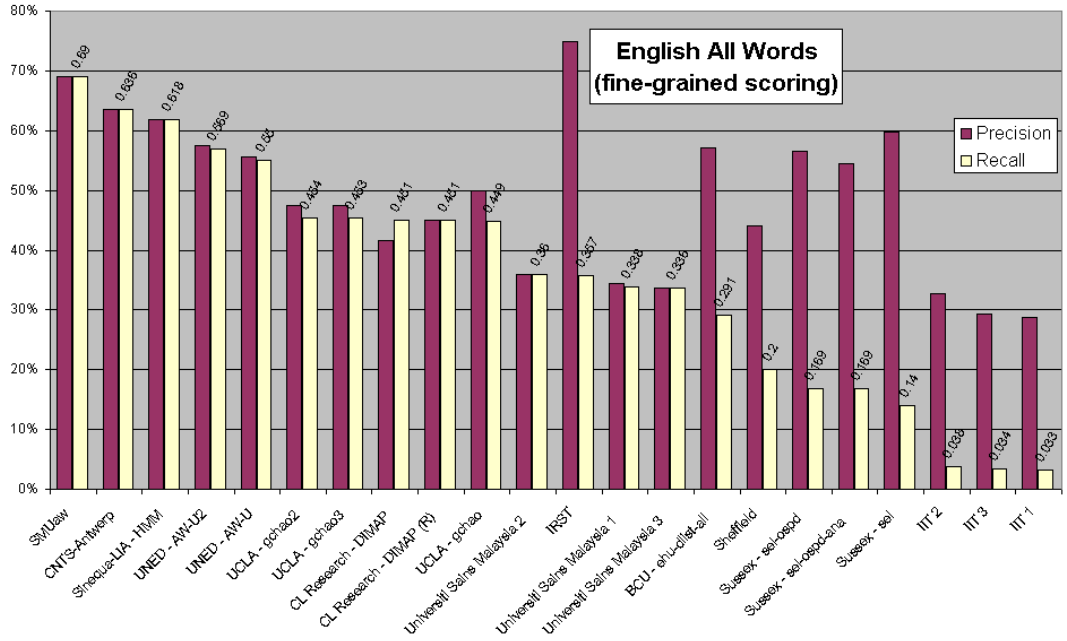
- Einführung
- Methode
  - Automatischer Thesaurus
  - WordNet Similarity Package
  - Algorithmus
- Auswertung
- Zusammenfassung

Einführung

- Kurzer Rückblick:
  - „word sense disambiguation“
    - Zutreffende Bedeutung eines ambigen Wortes im Kontext
    - Die meisten Systeme basieren auf der Anwendung von WordNet
  - „WordNet“
    - Datenbank mit semantischen und lexikalischen Beziehungen zwischen Wörtern
    - Reihenfolge basiert auf der Häufigkeit einer Bedeutung in SemCor (handannotiert)

Einführung

- „first sense heuristic“ vs. WSD-Systeme



  - First Sense liegt bei ca. 60% für Precision und Recall

Einführung

- Abstrakt:
  - „first sense heuristic“ beachtet den Kontext nicht
  - Benötigt handannotierte Daten
    - → Keine handannotierten Daten
    - → Dennoch gute Erfolge (precision: 64%)
    - → Passende „vorwiegende Bedeutungen“ für unterschiedliche Bereiche

Methode

- Erstellen eines automatischen Thesaurus
- WordNet Similarity Package
- Der eigentliche Algorithmus

Erstellen eines automatischen Thesaurus

- Annahme:
  - Wörter in gleichem Kontext haben gleiche (oder ähnliche) Bedeutung
  - Wörter die sich ähnlich sind können nach Frequenz geordnet werden
  - Die Anordnung der ähnlichen Wörter (Nachbarn) gibt den wahrscheinlichsten Sinn eines Wortes wieder

Erstellen eines automatischen Thesaurus

- Formel zum erstellen eines automatischen Thesaurus (über Ähnlichkeit der Worte):

$$\frac{\sum_{(r,x)\in T(w)\cap T(n)} (I(w,r,x)+(I(n,r,x))}{\sum_{(r,x)\in T(w)} I(w,r,x)+\sum_{(r,x)\in T(n)} I(n,r,x)}$$

mit:

$$I(w,r,x)=\log\frac{P(x\mid w\cap r)}{P(x\mid r)}$$

Formelerläuterung:

- „w" und „n" sind Worte, deren Ähnlichkeit festgestellt werden soll
- „r" entspricht einem grammatischen Verhältnis zwischen „w" und „x"
- „x" ist ein Wort, das in Relation „r" zu „w" und „n" steht
- „T(w), bzw. „T(n)" stellt eine Menge von gemeinsamen Auftreten von „x" unter „r" für das entsprechende Wort dar

Erstellen eines automatischen Thesaurus

- Beispiel:

Wir erstellen automatisch einen Thesaurus. Natürlich kann jeder einen Thesaurus auf diese Weise erstellen. Wenn wir ein Wörterbuch erstellen wollen, brauchen wir gegebenen Falls andere Daten. […]

w = Thesaurus

n = Wörterbuch

r = ist Objekt des Verbes

x = erstellen

Erstellen eines automatischen Thesaurus

- Berechnung:
  - $I(w,r,x) = \log (1 / 0{,}75) = 0{,}125$
  - $I(n,r,x) = \log (1 / 0{,}75) = 0{,}125$

  - Da dies die einzige Relation ist, und bei beiden Worten vorkommt:

    $I(w,r,x) + I(n,r,x) = 0{,}25$

  ➔ Die beiden Worte sind identisch (bzw. Der dss beträgt 1)

Erstellen eines automatischen Thesaurus

- Thesaurus beinhaltet zu (fast) jedem Wort x die *k* ähnlichsten Worte
  - Einschränkung: Berechnung nur für Worte (bzw. Relationen) deren <w, r, x> Frequenz über 10 liegt
    - Grund: Siehe Beispielberechnung
- ACHTUNG! Beinhaltet Wortähnlichkeit nicht Bedeutungsähnlichkeit

WordNet Similarity Package

- Beinhaltet WordNet Similatity scores (wnss)
- Diese bestehen zwischen zwei Bedeutungen in WordNet
- Allgemeine Formel:

$$wnss(ws_i, n_j) = \max_{ns_x \in senses(n_j)} (wnss(ws_i, ns_x))$$

- Die beiden mit der besten Performanz sind lesk und jcn

WordNet Similarity Package

- lesk
  - Betrachten der Schnittmenge von Wörtern in beiden Definitionen
    - In der gesuchten Bedeutung (ws_i)
    - In jeder Bedeutung des verglichenen Wortes (n)

WordNet Similarity Package

- jcn
  - Nutzt Korpus-Daten um Klassen (Synsets) anzulegen mit Anzahl der Auftreten
  - Die Häufigkeit der Auftreten beinhaltet alle Auftreten eines Wortes aus dem Synset im Korpus (Direkt oder über Hyponymrelation)

WordNet Similarity Package

- jcn
  - Der Informationsgehalt einer solchen Klasse ist:
    $IC(s) = -\log(p(s))$
  - Die Distanz zweier Klassen:
    $D_{jcn}(s_1,s_2) = IC(s_1) + IC(s_2) – 2 * IC(s_3)$
    s_3 = informativste Superklasse von s_1 und s_2

Algorithmus

- Formel:

$$\sum_{n_j \in N_w} dss(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)}$$

N_w = Alle Nachbarn aus dem Thesaurus zu w
ws_i = Eine Bedeutung aller Bedeutungen von w

- Für jede Bedeutung (ws_i) von w erhalten wir einen Prevalence Score. Der höchste wird als first sense betrachtet.

Algorithmus

- Beispiel:

Senses	Neighbors of *star* (*dss*) *actor* (0.22)	*footballer* (0.12)	*planet* (0.08)	*circle* (0.03)	ps
celebrity	0.42	0.53	0.02	0.01	0,3145
celestial body	0.01	0.01	0.68	0.10	0,0071
shape	0.0	0.0	0.02	0.78	0,00001
zodiac	0.03	0.03	0.21	0.01	0,0207
Total	0.46	0.57	0.93	0.90	

$$ps(celebrity) = 0,22 \times \frac{0,42}{0,46} + 0,12 \times \frac{0,53}{0,57} + 0,08 \times \frac{0,02}{0,93} + 0,03 \times \frac{0,01}{0,90}$$

$$= 0,2009 + 0,1116 + 0,0017 + 0,0003$$

$$= 0,3145$$

Auswertung

- SemCor
  - Thesaurus basiert auf BNC
  - Thesaurus mit allen polysemen Worten die in BNC über 10x und in SemCor über 2x vorkommen
  - K = 50

  Ergebnisse:

measure	PS_{acc} %	WSD_{sc} %
lesk	54	48
jcn	54	46
baseline	32	24

Auswertung

- SemCor
  - Warum passen einige first senses nicht?
    → Die Grundlage für den Thesaurus
    → Inhalt des Corpus ist unterschiedlich

pipe :	1. tobacco pipe	(Richtig für SemCor)
	2. tube made of metal	(Richtig für BNC)

Auswertung

- SENSEVAL-2 English all Words Data
  - Nicht Grundlage für WordNet
  - 5000 Worte aus Artikeln von Penn Treebank II
  - Thesaurus wie zuvor

  Ergebnisse:

	precision	recall
Automatic	64	63
SemCor	69	68
SENSEVAL-2	92	72

Es wurde jeweils der first sense betrachtet

Auswertung

- SENSEVAL-2 English all Words Data

  – Warum ist das gut?

    → Ähnliche Ergebnisse wie SemCor

      ➔ ABER! Keine handannotierten Daten

    → Nicht gedeckte Worte wegen mangelnder grammatischer Relation

      ➔ Kann durch ergänzen der zu beachtenden Regeln verbessert werden

Auswertung

- Reuters Korpus (Themen spezifische Korpora)
  - 810000 Reuters (englische Zeitungsartikel)
  - Beschränkung auf SPORT- und FINANZ-Artikel
  - Thesaurus auf dem jeweiligen Korpus erstellt
    →First sense unterschiedlich in den Thesauren
  - Zusätzliche Verwendung der Subject Field Codes
  - Alle Worte aus WortNet, die ein **economy** oder **sports** label besitzen

Auswertung

- Reuters Korpus

  Ergebnisse anhand von 10 Beispiel Worten

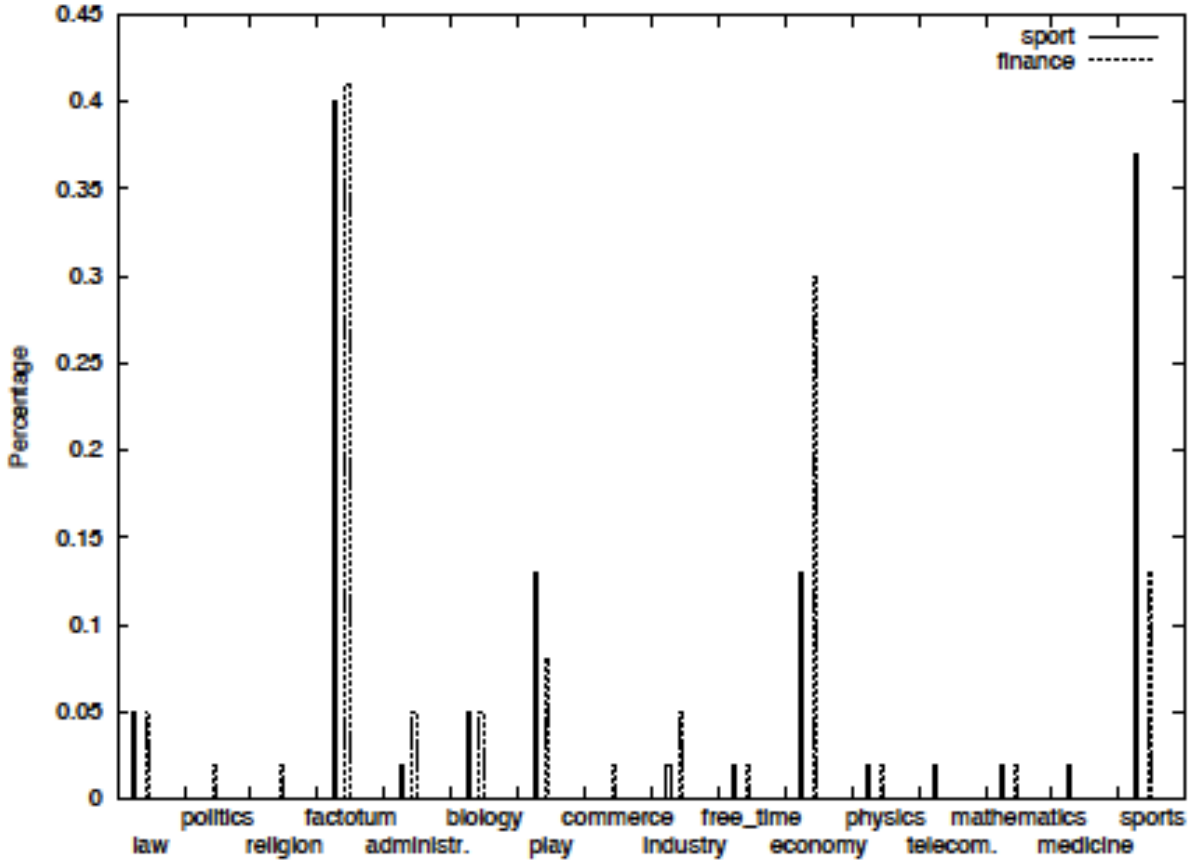
Word	PS BNC	PS FINANCE	PS SPORTS
pass	1 (**accomplishment**)	14 (**attempt**)	15 (**throw**)
share	2 (**portion, asset**)	2	2
division	4 (**admin. unit**)	4	6 (**league**)
head	1 (**body part**)	4 (**leader**)	4
loss	2 (**transf. property**)	2	8 (**death, departure**)
competition	2 (**contest, social event**)	3 (**rivalry**)	2
match	2 (**contest**)	7 (**equal, person**)	2
tie	1 (**neckwear**)	2 (**affiliation**)	3 (**draw**)
strike	1 (**work stoppage**)	1	6 (**hit, success**)
goal	1 (**end, mental object**)	1	2 (**score**)

Die Zahl entspricht der Bedeutung laut WordNet

Auswertung

- Reuters Korpus

Welche domain lables wurden den Bedeutungen zugewiesen (Beispiel anhand von 38 Worten)



Zusammenfassung

- Warum?
  - → Kein eigentliches WSD-System, da kein Kontext beachtet wird
  - → Kann auf nicht annotierten Korpora arbeiten
  - → first sense ist als Grundlage für andere Systeme verwertbar
  - → Kann auf verschiedene Themengebiete angewandt werden

    ➔ Kann auch auf neue Gebiete angewandt werden

Zusammenfassung

- Ausbaufähigkeit

→Mehr grammatische Regeln

➔ Verfeinerung des Thesaurus

→lesk als Grundlage

➔ lesk benötigt nicht zwingend WordNet

Jedes maschinenlesbare Wörterbuch ist anwendbar

Vielen Dank für Eure Aufmerksamkeit!