Latent Vector Weighting

for

Word Meaning in Context

Tim Van de Cruys, Thierry Poibeau, Anna Korhonen (2011)

Präsentation von Jörn Giesen

Distributionelle Semantik 16.01.2012

Distributational Hypothesis

Wörter die im selben Kontext stehen ,
haben “gleiche“ Bedeutungen.

frei übersetzt nach Harris, 1954

Beispiele

Hans fährt das Boot durch die Wellen.

Hans steuert das Boot durch die Wellen.

(ähnliche Bedeutung, da selber Kontext)

Maria sitzt auf der Bank.

Maria geht in die Bank.

(trotz gleicher Wörter, unterschiedliche Bedeutungen aus dem Kontext heraus)

Rückblick

Nicht probabilistische Ansätze

- Mitchel & Lapata (2008)

  Kontext als Vektor der Kookkurrenzen eines Wortes
  Wortbedeutungen als Komposita einzelner Kontextvektoren

- Pado & Lapata (2007)

  Kontext als Syntaktische Dependenzstruktur

Rückblick

Probabilistische Ansätze

- Dinu & Lapata (2010)

  Wortbedeutung als Wahrscheinlichkeitsverteilung über versteckte Faktoren

- Thater, Fürstenau, Pinkal (2011)

  Neugewichtung des Kontextes durch Einbeziehung von Verteilungsinformationen des Zielwortes

Gliederung und Inhalt

- Methodik
  - NMF
  - Kombination von Syntax und Kontextwörtern
  - Probabilistische Auswertung der Daten
- Evaluation
  - Paraphrasen ranking vs induction
- Ausblick

Nicht Negative Matrix Zerlegung NMF

- NMF ist der Versuch eine Matrix so in 2 Teilmatrizen zu zerlegen, dass :

$$A \approx W H$$

$$\begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 4 \end{pmatrix} \times \begin{bmatrix} 1 & 2 \end{bmatrix}$$

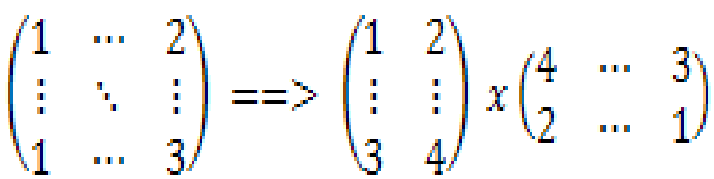
NMF

- Maß für „Ähnlichkeit“ zwischen Matrix A und Faktorisierung W x H

„Kullback-Leibler Divergenz“

Dependenzen

Zielwörter (50+)



A W H

NMF

Das Kernstück der NMF:

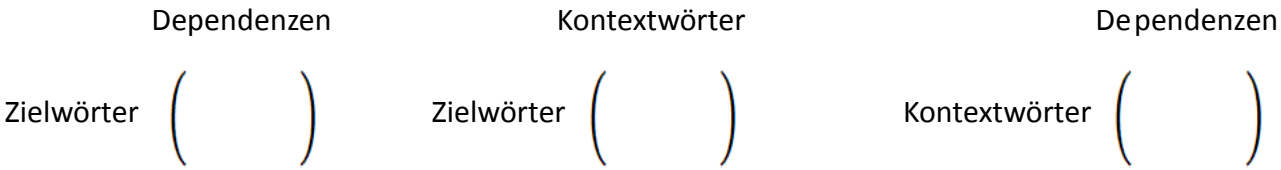
$$\mathbf{H}_{a\mu} \leftarrow \mathbf{H}_{a\mu} \frac{\sum_i \mathbf{W}_{ia} \frac{\mathbf{A}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_k \mathbf{W}_{ka}}$$

$$\mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia} \frac{\sum_\mu \mathbf{H}_{a\mu} \frac{\mathbf{A}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_v \mathbf{H}_{av}}$$

Welche Daten benötigen wir?

Input in den Algorithmus

- die Zielwörter
- ihre Kontextwörter
- die jeweiligen Dependenzstrukturen
  - Jill updated the record → {update obj $^{-1}$}



Die besondere Faktorisierung

- Die 3 Ausgangsmatrizen werden nicht getrennt faktorisiert
- Das Ergebnis der ersten Faktorisierung dient als Grundlage der nächsten Faktorisierung
- Am Ende der Faktorisierung werden die 3 Ausgangsmatrizen durch eine „kleine“ Zahl „versteckter Faktoren“ repräsentiert

Faktorisierung Beispiel

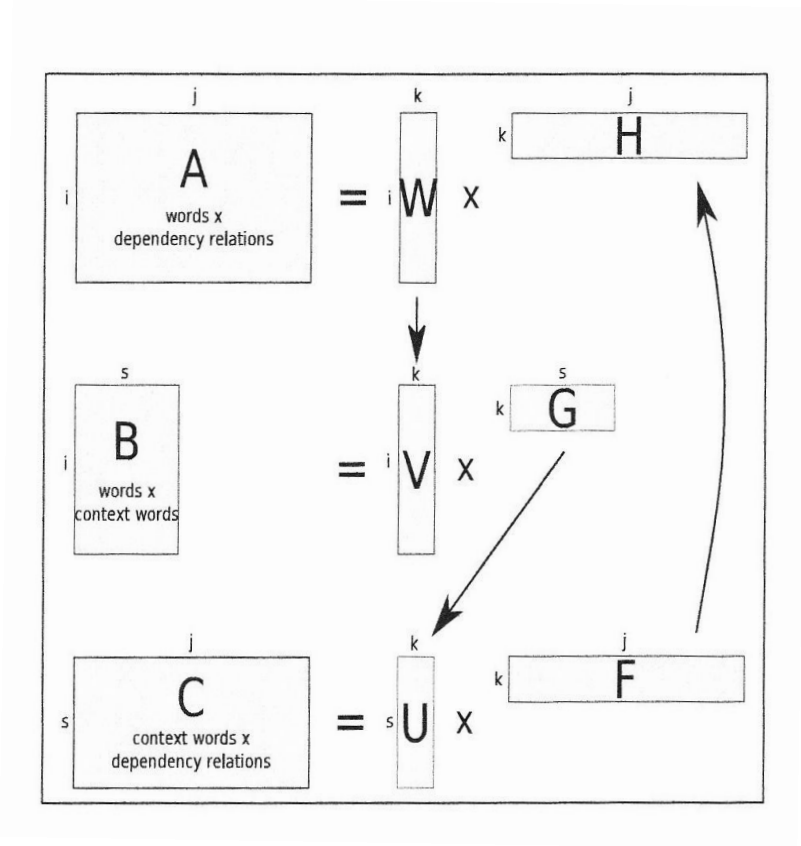
- Zerlegung einer Matrix in ihre „versteckten Faktoren“

	Virus	Disease	Test	Study
Spread	4	8	0	0
Transmit	6	8	0	0
Show	0	0	3	4
Reveal	0	0	2	3
Shed	3	1	2	4

Z1	Z2
1	0
1	0
0	1
0	1
0,4	0,6

- Z1: versteckter Faktor (Bedeutung) der mit Virus und Krankheit in Verbindung steht
- Z2: versteckter Faktor der mit Test und Lernen in Verbindung steht
- Beispiel aus Dissertation Georgiana Dinu Kapitel 4

NMF



- Ergebnisse der 1. Faktorisierung bilden die Grundlage der nächsten

- Nach der Faktorisierung:
probabilistische Interpretation

Matrizenrechnung Rückblick

- K Matrix
- K^{-1} inverse Matrix von K
- $K \times K^{-1}$ = Einheitsmatrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 2 & 1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0{,}1 & 0{,}2 & 0 \\ 0 & 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 & 0{,}1 \end{pmatrix}$ relative Häufigkeiten

Matrizenrechnung Rückblick

	d1	d2	d3
w1	0,1	0,2	0
w2	0	0,1	0,2
w3	0,2	0,1	0,1

p(wi, dj)

→

	d1	d2	d3
w1	1/3	2/4	0
w2	0	1/4	2/3
w3	2/3	1/4	1/3

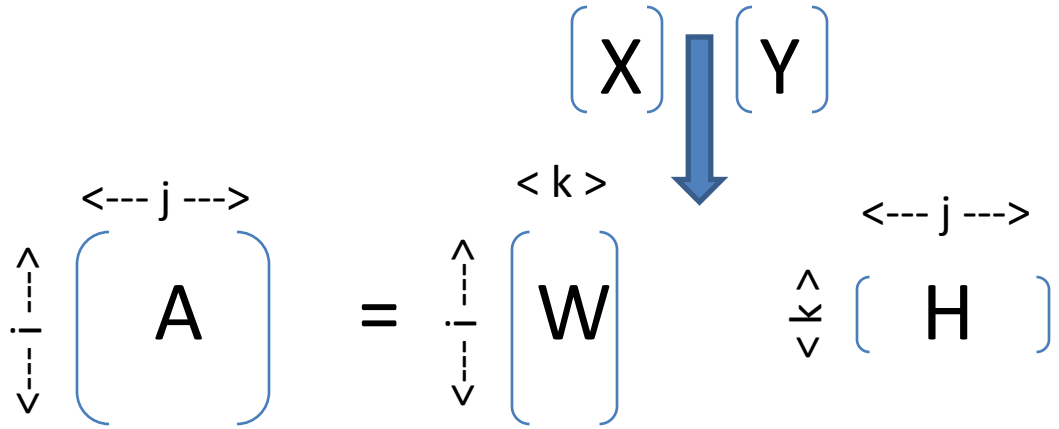
p(wi | dj)

Bedingte Wahrscheinlichkeiten

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Probabilistische Aufbereitung

$$A = W\,H = (W)(X^{-1}X)\,(Y\,Y^{-1})(H)$$



$$= (W\,X^{-1})\,(X\,Y)\,(Y^{-1}H)$$

Probabilistische Aufbereitung

$$(W\ X^{-1})\ (X\ Y)\ (Y^{-1}H)$$

- WX^{-1} = p (wi |z)
- $Y^{-1}H$ = p (dj |z)
- XY = p (z)
- z = „latent factors“

	Z1	Z2
spread	1	0
transmit	1	0
show	0	1
reveal	0	1
shed	0,4	0,6

→ p (spread | Z1) = 1/ 2,4 = 0,41667

Probabilistische Aufbereitung

$$p(\mathbf{z}|w_i) = \frac{p(w_i|\mathbf{z})p(\mathbf{z})}{p(w_i)}$$

$$p(\mathbf{z}|d_j) = \frac{p(d_j|\mathbf{z})p(\mathbf{z})}{p(d_j)}$$

Nur zur Erinnerung: das Bayes Theorem!

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

WK eines latenten Faktors unter einem speziellen Wort p(z|wi) kaum Aussagekräftig

$$p(\mathbf{z}|C) = \frac{\sum_{w_i \in C} p(\mathbf{z}|w_i)}{|C|}$$

Analog für p(z|**d**)

NMF

$$p(\mathbf{z}|C) = \frac{\sum_{w_i \in C} p(\mathbf{z}|w_i)}{|C|}$$

$$p(\mathbf{d}|C) = p(\mathbf{z}|C)p(\mathbf{d}|\mathbf{z})$$

- Jetzt haben wir also einen Zusammenhang zwischen Dependenzen und Kontextwörtern.

Gewichtung

Zuletzt:

$$p(\mathbf{d}|wi,C) = p(\mathbf{d}|wi) * p(\mathbf{d}|C)$$

Diese WK dient als Daumenabdruck, mit dem in der Ausgansmatrix nach Übereinstimmungen gesucht wird!

$p(wi|dj)$ → $p(wi|\mathbf{d})$ → $p(\mathbf{d}|wi)$

$$\begin{pmatrix} 0,1 & 0,2 & 0 \\ 0 & 0,1 & 0,2 \\ 0,2 & 0,1 & 0,1 \end{pmatrix} \qquad \begin{pmatrix} 0,3 \\ 0,3 \\ 0,4 \end{pmatrix}$$

Ergebnis

- System berechnet für:

Jack is listening to the record.

Jill updated the record.

-record C1: album, song, recording, track, cd

-record C2: file, datum, document, database

Evaluation

- Evaluation in englischer sowie französischer Sprache

- **Für das Englische:**
- trainiert auf dem UKWaC Korpus (500M Wörter)
- Getagged mit Stanford POS Tagger
- Geparst mit MaltParser

Evaluation

- Aus dem Trainingskorpus wurde zur Faktorisierung benutzt:

Pro Wortart (N,V, Adj, Adv)

5k Wörter

80k Dependenzstrukturen

2K Kontextwörter

Evaluation

Im Englischen:

- SEMEVAL 2007 (lexical substitution task)
- 200 Zielwörter (jeweils 50 Nomen, Verben, Adjektive und Adverbien)
- jedes Wort in 10 Sätzen → 2000 Sätze
- 5 Annotatoren erstellen Listen passender Substitute

Wie wurde das Ergebnis gemessen?

Paraphrase ranking:

- Vergleich der Reihenfolge der Substitute mit Goldstandard
- „Kendalls Tau“ vs „**G**eneral **A**verage **P**recission“

Paraphrase induktion

- Erstellung der Substitut-Liste und Vergleich mit Goldstandard

Paraphrase induction

- Maß für die Fähigkeit eines Systems, geeignete Substitute für Zielwort zu finden

$$R_{best} = \frac{\sum_{s \in M \cap G} f(s)}{|M| \cdot \sum_{s \in G} f(s)} \qquad P_{10} = \frac{\sum_{s \in M \cap G} f(s)}{\sum_{s \in G} f(s)}$$

- M = vom System vorgeschlagenen Substitute
- G = Liste der Substitute des Goldstandards
- f (s) = Übereinstimmungen mit den Anotatoren

Paraphrase induction

model	R_{best}	P_{10}
vector$_{dep}$	8.78	30.21
DL	1.06	7.59
KU	20.65	46.15
IRST2	20.33	68.90
NMF$_{context}$	8.81	**30.49**
NMF$_{dep}$	7.73	26.92
NMF$_{c+d}$	**8.96**	29.26

Kein vorgegebenes Bedeutungsinventar,

trotzdem Werte die mit dem dep.based Vektor Modell vergleichbar sind!

Paraphrase ranking

Model	Kendalls Tau	GAP
random	-0,61	29,98
vector dep	16,57	45.08
EP 09	-	32,2*
EP 10	-	39,9*
TFP	-	45,94*
DL 10	16,56	41.68
NMF cont	20,64**	47,60**
NMF dep	22,49**	48,97**
NMF c+d	22,59**	49,02**

Legende:
*: Werte aus den jeweiligen Veröffentlichungen übernommen

**: statistisch signifikant

Fazit

- Wir haben also ein System das versteckte Faktoren aus den Ausgangsdaten extrahiert,

- Verbindungen zwischen Kontextörtern und Dependenzstrukturen erstellt,

- Und für ein Zielwort, selbständig eine Liste möglicher Paraphrasen errechnet, ohne dabei auf ein Bedeutungsinvertar zurückgreifen zu müssen.

Für die Zukunft

Optimale Verbindung zwischen Kontextwörtern und Dependenzstrukturen?

Weitere Test des Systems auf verschiedenen Gebieten

Vielen Dank für die Aufmerksamkeit