Proseminar: Distributionelle Semantik

Nouns are Vectors, Adjectives are Matrices:

Representing Adjective-Noun Constructions in Semantic Space

Marco Baroni & Roberto Zamparelli (2010)

Referent: Tobias Mayer

Saarbrücken, den 23.01.2012

Bisherige Entwicklung

Bedeutung von Wörtern als Vektor

Vektorerstellung durch Kookkurrenz

Für Nomen, Verben und Adjektive

	Tier	Frau	fahren	krank
groß	8	4	1	2
Mann	4	7	4	5
husten	6	6	2	9

groß = [8 4 1 2]

Mann = [4 7 4 5]

husten = [6 6 2 9]

Bisherige Entwicklung

Fokus:

Disambiguierung bzw. Kontextualisierung von Wörtern

- Algebraische Methoden (z.B. Vektoraddition)
- Clustering und ähnliche Verfahren
- Statistische Ansätze

Neuer Fokus

Kompositionalität:

Berechnung von Vektoren komplexerer Ausdrücke durch ihre Teilausdrücke.

Geht das auch mit den bisherigen Methoden?

Schwierigkeiten

ABER:

falscher Hase?

angeblicher Logiker?

grüne Ampel?

taube Nuss?

Lösungsansatz
Grundidee

Adjektive nicht als Vektoren, sondern als lineare Funktionen:

Vektor 1(N) abgebildet auf Vektor 2(AN)

Vorteil: rekursiv anwendbar

Lösungsansatz
Umsetzung

Abbildung eines Adjektivs erlernt durch
Lineare Regression eines Vektorpaares (N, AN)
aus dem Corpus

> Lineare Regression liefert zu einem Paar die Funktion mit der eine abhängige Variable (AN) durch eine unabhängige (N) beschrieben wird.

Lösungsansatz
Umsetzung

Vektor (N) * n x n Matrix (Adj) = Vektor (AN)

"Zwei Matrizen können multipliziert werden, wenn die Spaltenanzahl der linken mit der Zeilenanzahl der rechten Matrix übereinstimmt." *Wikipedia*

1 x n * n x n = 1 x n

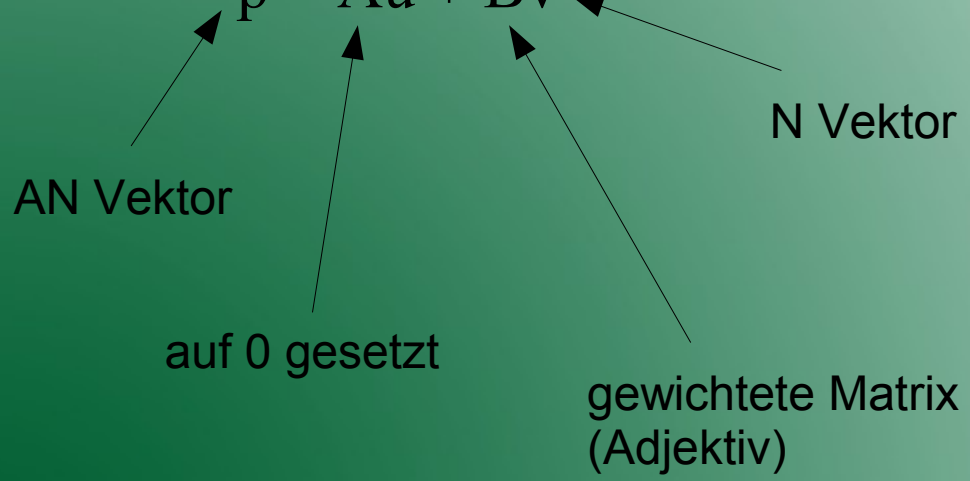
Das Produkt einer l x m –Matrix und einer m x n –Matrix ist eine l x n Matrix.

Lösungsansatz
Vergleich zu altem Modell

Mitchell & Lapata (2008)

gewichtetes Additionsmodell:

$$p = Au + Bv$$



Testsetup
Corpus

Konkatenierter Corpus:

ukWac corpus + 2009 Wikipedia + British National Corpus

→ tokenisiert, POS getaggt und lemmatisiert

→ 2,83 Milliarden Tokens

→ 8000 häufigsten Nomen und 4000 häufigsten Adjektive zu Kernvokabular zusammengefasst (ohne die 50 häufigsten Nomen/Adjektive)

Testsetup
AN test set

Auswahl von 36 Adjektiven aus verschiedenen

Klassifizierungen:

size(7), denominal(6), colors(4),
positive evaluation(4), temporal(5),
modal(2), common abstract(8)

Darunter auch klassenübergreifende (electronic)
und fast-funktionswort-Adjektive (different)

Testsetup
AN test set

Häufigsten Nomen (mind. 300 Vorkommen in post-adjektivischer Stellung)

Ausgenommen häufig vorkommende Zeit- und Maßausdrücke (z.B. time, range)

→ Auswahl von 26440 ANs, die auch unter den ersten 100M Tokens des ukWac Corpus sind.

Testsetup

AN test set

Erweitertes Vokabular :

41K Wörter (darunter unter anderem

12K Kernvokabular und 26440 ANs)

Testsetup

Semantischer Raum

10K Grundformen, die am häufigsten mit dem Kernvokabular vorkommen, als Dimensionen.

(Keine Kookkurrenzwerte, sondern LMI)

→ 41K x 10K Matrix

Durch SVD (Singular Value Decomposition) auf 41K x 300 reduziert.

Testsetup
Methode alm

Adjective-specific linear map (*alm*):

AN-Generierung durch multiplizieren des N-Vektors mit der Adj-Matrix

Testsetup
Die Konkurrenten

Additive AN vectors (*add*):

Addieren der normalisierten Vektoren

Multiplicative vectors (*mult*):

Multiplizieren der Adj.- und Nomenvektoren

single linear mapping model (*slm*):

zu *alm* ähnliches Modell mit nur einer Adj.-Matrix

Testsetup
Auswertung

Was vergleichen wir überhaupt?

→ Berechnung des Cosinus' der generierten ANs mit den 41K Vektoren des erweiterten Vokabulars

→ Vergleich mit den gesehenen ANs im Corpus anhand der Cosinusrangliste

Testsetup
Auswertung

Methode	25%	Median	75%
alm	17	170	≥1K
add	27	257	≥1K
noun	72	448	≥1K
mult	279	≥1K	≥1K
slm	629	≥1K	≥1K
adj	≥1K	≥1K	≥1K

→ *alm* mit signifikantem Abstand an erster Stelle

→ nur noch *add* besser als baseline

→ *mult* hinter *add* ↔ bisherige Studien

Testsetup
Auswertung

alm im Detail:

- Besten Median bei hochfrequenten, polysemantischen Adjektiven (z.B. new,large, etc.)

  → inverse Korrelation zwischen Median und Frequenz

  → je mehr Information desto bessere Ergebnisse

- gesehene und pred. Nachbarn oftmals ähnlich oder gleich (z.B. für *recent request* beides mal *recent enquiry*)

Testsetup
Auswertung

- Allerdings 27% aller pred. ANs 1K Nachbarn von den gesehenen entfernt
- Probleme vorallem bei kontextabhängigen/seltenen ANs wie z.B. *white profile* (Nachbar: *white snow*)
- pred. Nachbarn zum Teil näher am AN als gesehene (z.B. *young image* vs. *important song* für *young photo*)

Zwischenfazit

- neuer Ansatz zur Kompositionalität
- Adjektive als Funktionen nicht als Vektoren
- effektiver als bisherige Methoden
- dennoch verbesserungsbedürftig

Vergleichbarkeit

Sind Adjektive als Funktionen noch vergleichbar?

2 Varianten in diesem Framework möglich:

1. Durchschnittsvektor aller AN-Vektoren mit dem selben Adjektiv
2. Auffaltung der 300 x 300 Matrix eines Adjektivs in einen 90K Vektor

→ Vergleich mit klassischen Kookkurrenzvektor

Vergleichbarkeit

- Clusterbildung mit jeweils 19 Beispiel Adjektiven
- Clusterqualität ermittelt durch *percentage purity*
- Baseline sind 10K zufällig verteilte Adjektive

input	*purity*
matrix	73.7(68.4-94.7)
centroid	73.7(68.4-94.7)
vector	68.4(63.2-89.5)
random	45.9(36.8-57.9)

→ "indirekte Repräsentationen" sind äquivalent

→ beide einen Tick besser als das Vektormodell

Überlegung

ANs im Semantischen Raum

Intuitiv:

Zentren der ANs mit gemeinsamen Adj (z.B. *red N*) in der Nähe des Adjektives oder des dazugehörigen Nomens (Bsp. s*üße N* nahe bei *Süße* bzw. *Süß*)

Nachbarn des Kompositum sollten die Bedeutung stärker ausdrücken als Adj. bzw. Nomen alleine

Überlegung

ANs im Semantischen Raum

Im Modell:

AN-Zentren sind in der Nähe des Adj. (*nice N* bei *nice*) oder des korrespondieren Nomens (*green N* in der Nähe von *green(n)*)

AN-Nachbarn zeigen stärker auf das AN als Nomen bzw. Adj. (*small son/daughter* ist näher bei *young husband* als *young* oder *husband*)

Fazit

- Adjektive als Funktionen
- Vorteile bei AN-Generierung
- besonders geeignet für mehrdeutige Adj.
- kommt der Vorstellung über den Semantischen Raum nahe
- Rekursion bei größeren Ausdrücken möglich
- noch stark verbesserungsbedürftig

Ausblick

- Anwendung bei gebunden Morphemen (z.B. re- oder andere Affixe)
- Ausbau zu abstrakteren Konstruktionen

  z.B. determiner N
- Generierung von immer größer werdenden Komponenten
- Verbesserung des jetztigen Modells

Fragen?