Distributionelle Semantik

Theorie und Anwendung

Fritz Günther

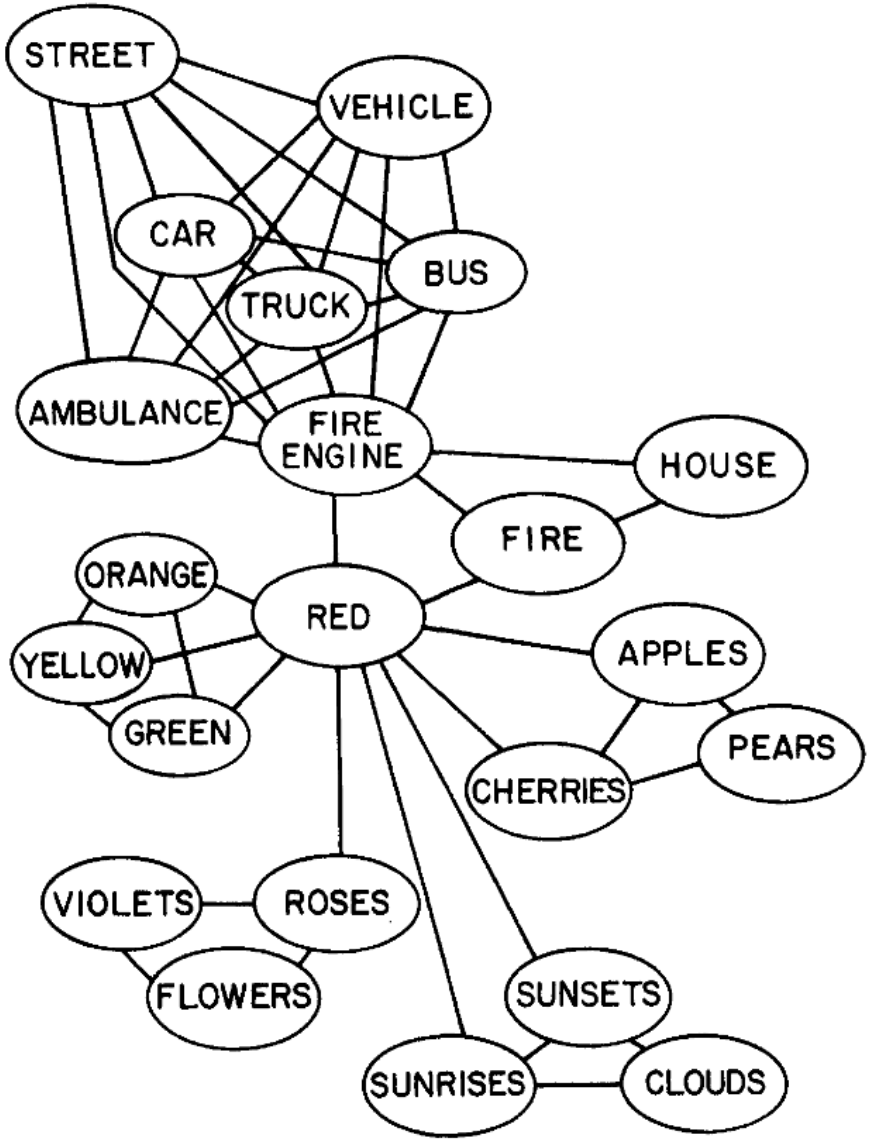
Universität Tübingen

December 8, 2015

Theoretischer Teil

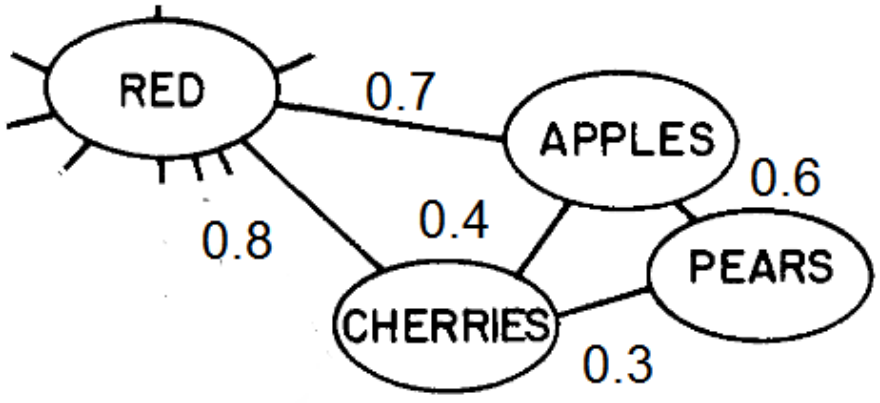
Die Bedeutung von Wörtern

Collins & Loftus, 1975



Wortähnlichkeiten

Kintsch, 1988



Assoziationsnormen

n* × *n* Associative Matrix for *Dinner

Cue	DIN	SUP	EAT	LUN	FOO	MEA
DINNER	1.00	0.54	0.11	0.10	0.09	0.09
SUPPER	0.55	–	0.02	0.03	0.17	0.01
EAT			–		0.40	0.02
LUNCH	0.27	0.02	0.08	–	0.21	0.06
FOOD			0.41	0.01	–	0.02
MEAL	0.21	0.06	0.06	0.06	0.49	–
Summed strength	2.03	.62	.68	.20	1.36	.20

Die Bedeutung von Wörtern

Pantel & Lin, 2002

- A bottle of *tezgüno* is on the table.
- Everyone likes *tezgüno*.
- *Tezgüno* makes you drunk.
- We make *tezgüno* out of corn.

Die distributionelle Hypothese

Harris, 1954

"You shall know the meaning of a word by the company it keeps."

Die distributionelle Hypothese

Deerwester et al., 1990; Lund & Burgess, 1996

Die Bedeutung eines Wortes ergibt sich aus den *Kontexten*, in denen es auftritt.

Wie oft kommt ein Wort in welchem Kontext vor?

Also:
Die Bedeutung eines Wortes ergibt sich aus dessen Verteilung in der Sprache.

Sprachlicher Kontext

Sahlgren, 2008

Was ist Kontext?

Sprachlicher Kontext
Sahlgren, 2008

Was ist Kontext?

- Kontext als *Dokument*, in dem ein Wort vorkommt (LSA-artige Modelle)

Sprachlicher Kontext
Sahlgren, 2008

Was ist Kontext?

- Kontext als *Dokument*, in dem ein Wort vorkommt (LSA-artige Modelle)

- Kontext als *n* Wörter vor und nach einem Wort (HAL-artige Modelle)

Latent Semantic Analysis (LSA)

Deerwester et al., 1990; Landauer & Dumais, 1997

Kontext ist definiert als das Dokument, in dem ein Wort auftritt.

$\Longrightarrow$ Zählen, wie oft ein Wort in welchen Dokumenten auftritt.

(D1) The dog and the cat are playing.
(D2) I love my dog.
(D3) My cat is great. I love the cat more than your dog.
(D4) Snoopy is clearly a dog.

Latent Semantic Analysis (LSA)

Deerwester et al., 1990; Landauer & Dumais, 1997

Kontext ist definiert als das Dokument, in dem ein Wort auftritt.

$\Longrightarrow$ Zählen, wie oft ein Wort in welchen Dokumenten auftritt.

(D1) The dog and the cat are playing.
(D2) I love my dog.
(D3) My cat is great. I love the cat more than your dog.
(D4) Snoopy is clearly a dog.

	D1	D2	D3	D4
dog	1	1	1	1
cat	1	0	2	0

Hyperspace Analogue to Language (HAL)

Lund & Burgess, 1996

Kontext ist definiert als die n Wörter vor und nach dem Zielwort (n-word window).
$\Longrightarrow$ Zählen, wie oft ein Wort mit welchen anderen Wörtern auftritt.

Beispiel: $n = 2$ (Inhaltswörter)

(D1) During night, the moon can be seen in the sky.
(D2) Men land on the moon last night and begin to explore it.

Hyperspace Analogue to Language (HAL)

Lund & Burgess, 1996

Kontext ist definiert als die n Wörter vor und nach dem Zielwort (n-word window).
$\Longrightarrow$ Zählen, wie oft ein Wort mit welchen anderen Wörtern auftritt.

Beispiel: $n = 2$ (Inhaltswörter)

(D1) During night, the moon can be seen in the sky.
(D2) Men land on the moon last night and begin to explore it.

	night	sun	seen	sky	men	land	last	explore
moon	2	0	1	1	1	1	1	0

Unterschiede zwischen LSA und HAL

Sahlgren, 2008; Jones, Kintsch, & Mewhort, 2006

In LSA haben Wörter dann sehr ähnliche Vektoren, wenn sie in den gleichen Dokumenten zusammen vorkommen

$\Longrightarrow$ **Assoziativ** ähnliche Wörter

Beispiele: *Arzt - Krankenhaus*, *Spinne - Netz*, *Straße - Auto*

In HAL haben Wörter dann sehr ähnliche Vektoren, wenn sie von den gleichen Wörtern umgeben werden

$\Longrightarrow$ **Semantisch** ähnliche Wörter

Beispiele: *Wespe - Fliege*, *Gitarre - Geige*, *Uhr - Wecker*

Aufbereitung der Vektoren

Martin & Berry, 2007

1. **Gewichtung der Vektoren**

   z.B. PMI statt Häufigkeiten; $PMI(a, b) = log(\frac{p(a,b)}{p(a)*p(b)})$

Aufbereitung der Vektoren

Martin & Berry, 2007

1. **Gewichtung der Vektoren**

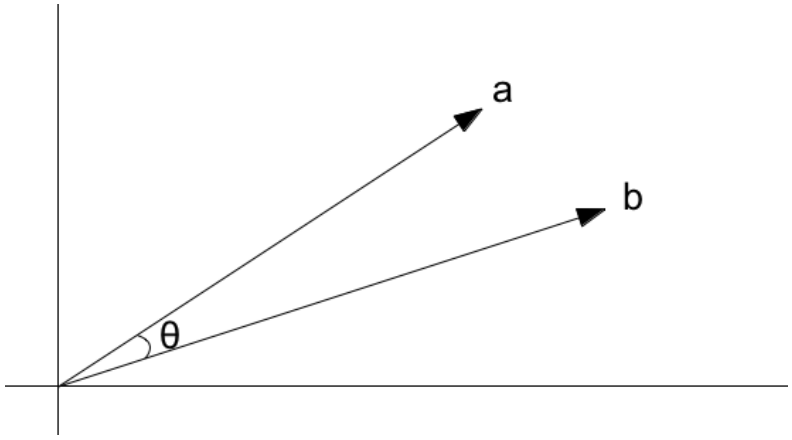
   z.B. PMI statt Häufigkeiten; $PMI(a, b) = log(\frac{p(a,b)}{p(a)*p(b)})$

2. **Dimensionsreduktion**

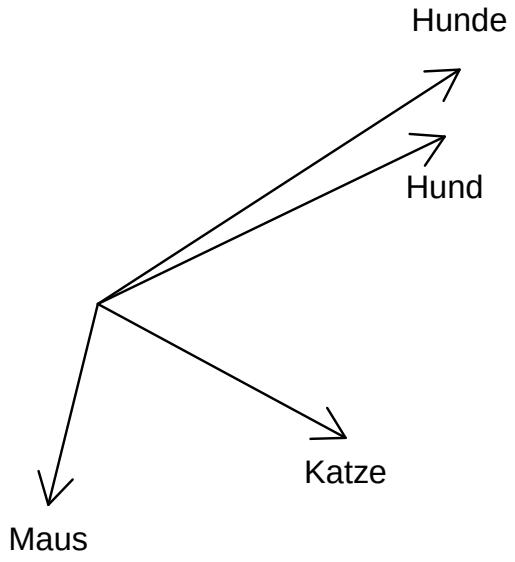
   z.B. SVD oder NMF

Bestimmung von Wortähnlichkeiten

Landauer & Dumais, 1997; Lund & Burgess, 1996



$$similarity = cos(\theta) = \frac{a * b}{||a|| * ||b||}$$



Notiz zu Wortähnlichkeiten

Sahlgren, 2008; Günther, Dudschig & Kaup, 2015

Modelle der distributionellen Semantik beschreiben Wortähnlichkeiten **in der Sprache**

Ob sie auch Wortähnlichkeiten **"im Kopf"** beschrieben ist eine empirische Frage

Praktischer Teil

LSA Homepage

Dennis, 2007

lsa.colorado.edu

R-Package auf CRAN

Installieren in R:

```
install.packages("LSAfun")
library(LSAfun)
```

Semantic Space Repository

Günther, Dudschig & Kaup, 2015

www.lingexp.uni-tuebingen.de/z2/LSAspaces

Enthält LSA- und HAL-artige semantische Räume in

- Englisch
- Deutsch
- Französisch
- usw.

Laden eines semantischen Raumes

R-Package LSAfun

1. Herunterladen des semantischen Raumes in einen Ordner
2. `setwd(PATHTOFILE)` auf diesen Ordner
3. Semantischen Raum mit `load(NAMEOFSPACE.rda)` in R laden
4. Für nicht-englische Räume: Sicherstellen, dass die Enkodierung der Buchstaben klappt mit `Encoding(rownames(NAMEOFSPACE)) <- "UTF-8"`

Berechnung von Wortähnlichkeiten
R-Package LSAfun

- Zwei Wörter

```
Cosine("tiger","katze",tvectors=dewak100k,breakdown=F)
[1] 0.7490589
```

- Ein Wort und eine Wortliste (oder zwei Listen)

```
multicos("tiger","katze löwe panther
strauss",tvectors=dewak100k,breakdown=F)
```

	katze	löwe	panther	strauss
tiger	0.749	0.781	0.827	0.294

Berechnung von Dokumentähnlichkeiten
R-Package LSAfun

- Zwei Dokumente (oder ein Wort und ein Dokument)

```
costring("tiger","im urwald wohnt ein
bär",tvectors=dewak100k,breakdown=F)
[1] 0.717126
```

- Ein Dokument und eine Wortliste

```
multicostring("im urwald wohnt ein bär","tiger
pizza ofen kerze",tvectors=dewak100k,breakdown=F)
```

	tiger	pizza	ofen	kerze
expression in x	0.712	0.364	0.378	0.449

Finden von Wörtern in bestimmter Distanz

R-Package LSAfun

- Die *n* nächsten Nachbarn (auch für Dokumente)

```
neighbors("tiger",n=5,tvectors=dewak100k,breakdown=F)
```

tiger	bären	elefant	elefanten	leopard
1.0000000	0.8790901	0.8583553	0.8533323	0.8291466

- *n* zufällige Wörter in bestimmter Distanz

```
choose.target("tiger",n=4,lower=0.1,upper=0.2,
tvectors=dewak100k,breakdown=F)
```

wiederaufbauhilfe	angepflanzt	schematischen	jodmangel
0.1320919	0.1661086	0.1164557	0.1079467

Umlaute ersetzen:

```
dewak2 <- dewak100k
rownames(dewak2) <- breakdown(rownames(dewak2))
```

Plot der Nachbarschaft eines Wortes (oder Dokuments), 2D oder 3D

```
plot_neighbors("rom",n=20,dims=2,
tvectors=dewak2,breakdown=F)
```

```
plot_neighbors("rom",n=100,dims=3,
connect.lines="all",alpha="shade",col="rainbow",
tvectors=dewak2,breakdown=F)
```

Plot der Ähnlichkeiten einer Wortliste

```
words <-
c("karotte","kartoffel","zwiebel","lauch","brhe",
"schwein","kuh","schaf","huhn","ziege")

plot_wordlist(words,dims=3,
connect.lines="all",alpha="shade",col="rainbow",
tvectors=dewak2,breakdown=T)
```

Übungsaufgaben

www.lingexp.uni-tuebingen.de/z2/material

Ein paar Analysen zur Übung

(1) Wortähnlichkeiten im Priming-Experiment
(Smolka et al., 2014)

Aufgabe: Sind transparente Wörter ähnlicher zu ihrem Wortstamm als intransparente Wörter?

Daten: `morphprime.txt`

Hinweis: `for()`-Schleifen sind nützlich
HAL geeigneter

Ein paar Analysen zur Übung

(2) Assoziationen von Material mit einzelnen Wörtern
(Lachmair et al., 2011)

Aufgabe: Sind die oben-Wörter stärker mit dem Wort "oben" ("unten") assoziiert als die unten-Wörter?

Daten: `updown.txt`

Hinweis: `tolower()`
HAL oder LSA

Ein paar Analysen zur Übung

(3) Erstellen von Primingmaterial

Aufgabe: Für jedes Target ein ähnliches und ein unähnliches Prime finden

Daten: `targets.txt`

Hinweis: HAL oder LSA (je nach Art des Primings)

Ein paar Analysen zur Übung

(4) Enkodierung von geographischer Information in der Sprache (Louwerse & Zwaan, 2009)

Aufgabe: Stimmen die "semantischen Ähnlichkeiten" von Ländern mit ihrer geographischen zueinander Nähe überein?

Daten: `europe.txt`

Hinweis: Zweidimensionale Abbildung LSA geeigneter

Ein paar Analysen zur Übung

(5) Polysemie (Mehrere Bedeutungen pro Wort)

Aufgabe: Inwiefern können semantische Räume abbilden, dass manche Wörter mehrdeutig sind

Daten: Explorativ

Hinweis: Gibt es mehrere "Cluster" in einer dreidimensionalen Abbildung?
HAL oder LSA

Nützliche R-Funktionen zur Textverarbeitung

- `as.character()`
- `tolower()`
- `chartr()`
- `gsub()`
- `grep()`
- `breakdown()`

Software zum Erstellen semantischer Räume

Freeware

- S-Space (Jurgens & Stevens, 2010)
- Semantic Vectors (Widdows & Ferraro, 2008)
- gensim (Rehurek & Sojka, 2010)
- DISSECT (Dinu, Pham, & Baroni, 2013)
- Anleitung (für Unix-Systeme): http://clic.cimec.unitn.it/marco/teaching/compling/materials/lab-notes.txt