Ó Séaghdha & Korhonen (2011)

Probabilistic Models of Similarity in Syntactic Context

Seminar zur distributionellen Semantik

Die Problematik

- Schnee 

Niederschlag in Form von Schneeflocken

Droge, die als weißes Pulver gehandelt wird.

- Eis 

gefrorenes Wasser; (Sport) Eisfläche eines Eisstadions

Speiseeis

Die Problematik

Wie ähnlich sind *Schnee* und *Eis* in folgenden Sätzen?

- „Schnee/Eis auf der Fahrbahn sorgte vielerorts für Verkehrschaos.“
- „Der Italiener um die Ecke macht den besten Schnee/das beste Eis.“

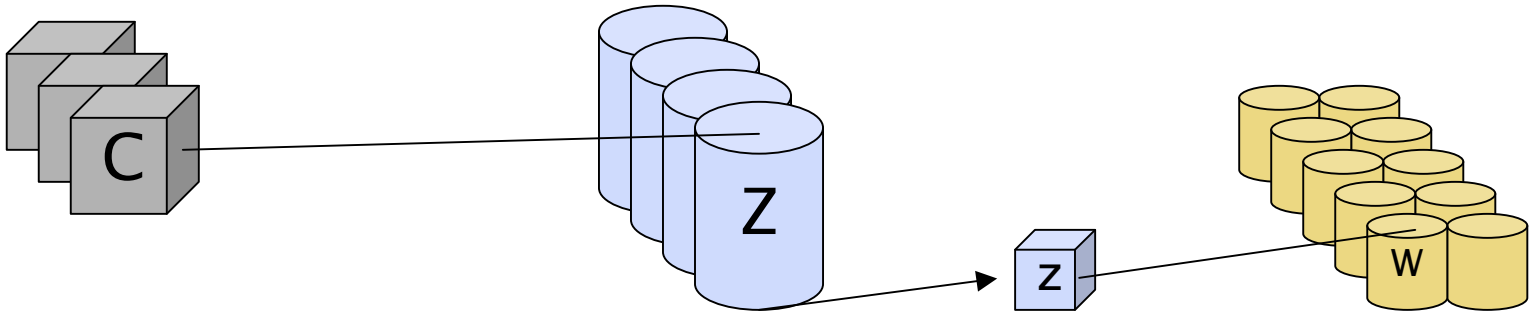
Rückblick

- Mitchell & Lapata (2008)
  - Vektoraddition & -multiplikation
- Erk & Padó (2008)
  - SVS mit Selektionspräferenzen
- Thater, Fürstenau & Pinkal (2011)
  - Kontextwolken
- Reisinger & Mooney (2010)
  - Multi-Prototypen

Rückblick

- Dinu & Lapata (2010)
  - Probabilistisch
  - Latentes Variablenmodell

Rückblick



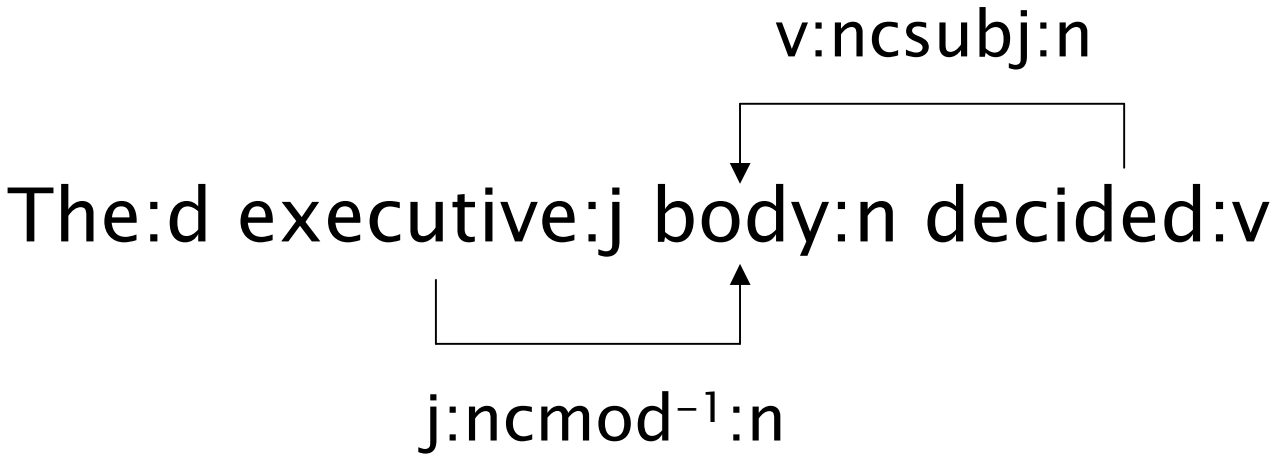
Rückblick

- Dinu & Lapata (2010)
  - Bag of words – Kontext

Neue Modelle

- Einbringen von syntaktischem Kontext
- Differenzierung der Abhängigkeiten
  - Context ⟶ Target
  - Target ⟶ Context
- Verwendung eines effizienteren Ähnlichkeitsmaßes

Neue Modelle



ncmod^{-1} = inverse non-clausal modifier relation

ncsubj = non-clausal subject relation

Neue Modelle

C(body)

{ executive: j:ncmod^{-1}:n

decide: v:ncsubj:n }

Neue Modelle

- Paraphrasierungsmodell

$$P_{C \to T}(n \mid o, C) = \sum_{z} P(n \mid z) P(z \mid o, C)$$

C = Kontext-Set

T = Target (Zielwort)

n = Paraphrase

o = Beobachtungswort

z = Thema

Neue Modelle

- Paraphrasierungsmodell

$$P_{T \to C}(n \mid o, C) = \frac{P(C \mid o, n) P(n \mid o)}{P(C \mid o)}$$

$$P_{C \to T}(n \mid o, C) = \sum_{z} P(n \mid z) P(z \mid o, C)$$

C = Kontext-Set

T = Target (Zielwort)

n = Paraphrase

o = Beobachtungswort

z = Thema

Neue Modelle

- Ähnlichkeitsmaß
  - Nach Bhattacharyya

$$sim(P_x(z), P_y(z)) = \sum_z \sqrt{P_x(z) P_y(z)}$$

  - Werte zwischen 0 und 1

Neue Modelle

- Training über Latent Dirichlet Allocation
  - Nach Blei
  - Unterteilung von Dokumenten in Themen
  - Verteilung der Wörter auf Themen

Modellüberblick

- probabilistische Modelle
  - Paraphrasierungsmodell (Para)
  - Ähnlichkeitsmaß (Sim)

- Kontext-Sets
  - bag of words
  - syntaktische Relationen

Evaluation

- Überblick
  - Exp. 1 – Ähnlichkeit im Kontext
  - Exp. 2 – Lexikalische Substitution

Ähnlichkeit im Kontext

- Ziel
  - Performanz-Check unserer Modelle
  - Vergleich der Modellvorhersagen mit den menschlichen Vorhersagen

Ähnlichkeit im Kontext

- Daten
  - participant20 boom noise prosper 3 low
    - dröhnen Lärm florieren ?
  - participant20 boom export prosper 7 high
    - florieren Export florieren !

Ähnlichkeit im Kontext

- Daten
  - Satzpaare mit statischem Subjekt, aber variablem Verb
    - 120 Paare
    - 15 Verben
    - Balanciert

Ähnlichkeit im Kontext

- Training auf BNC (90 Mio Wörter)
  - Kontext: syntaktische Relationen
    - v:ncsubj:n
    - n:ncsubj^{-1}:v

Ähnlichkeit im Kontext

- Ablauf
  - 60 Versuchspersonen bewerten die Satzpaare auf einer Skala (1–7)
  - Korrelation = 0.40 (rho)
  - Spearman‘s rho (Ranking-Vergleich)
    - –1 umgekehrte Reihenfolge
    - 0 kein Zusammenhang
    - 1 identische Reihenfolge

Ähnlichkeit im Kontext

	Modell	Para	Sim
Keine Optimierung	C > T	0.24	0.34
	T > C	0.36	0.39 ←
	T<>C	0.33	0.39
Optimierung auf DevSet	C > T	0.24	0.35
	T > C	0.41	0.41 ←
	T<>C	0.37	0.41

Erk & Padó (2008)	Mult	0.24
	SVS	0.27

Ähnlichkeit im Kontext

- Ergebnis
  - Werte sind über dem state of the art von Erk & Padó
  - T > C Modelle schlagen C > T

Ähnlichkeit im Kontext

- Doch:
  - geringe Datenmenge (120 Satzpaare)
  - erfundene Testsätze
  - zu wenig syntaktische Relationen

  - Deshalb Experiment 2 ⟶

Lexikalische Substitution

- Hintergrund
  - SemEval (semantische Evaluation)
  - Ranking von Substitutionswörtern durch Probanden

Lexikalische Substitution

- Realizing that strangers have come, the animals charge them and began to fight
  - charge: attack, rush at

- Commission is the amount charged to execute a trade
  - charge: levy, impose, take

levy, impose = erheben

Lexikalische Substitution

- Die Diebe klauen Goldbarren aus dem Tresor.
  - klauen: rauben, stehlen
- Die Kinder klauen Süßigkeiten aus der Speisekammer.
  - klauen: mopsen, stibitzen

Lexikalische Substitution

- Daten
  - 1986 annotierte Sätze
  - 201 Zielwörter
    - 4 POS: Nomen, Verb, Adj, Adv

Lexikalische Substitution

- Ziel
  - Ranking zugewiesener Substitute
  - Vergleich mit Goldstandard
  - Als Maße dienen GAP und τ_b

Lexikalische Substitution

- Generalised Averaged Precision (GAP)
  - Ursprünglich Information Retrieval
  - Durchschnittspräzision von Suchanfragen
  - In %
- Kendall‘s τ_b
  - −1 bis 1 (vgl. rho)

Lexikalische Substitution

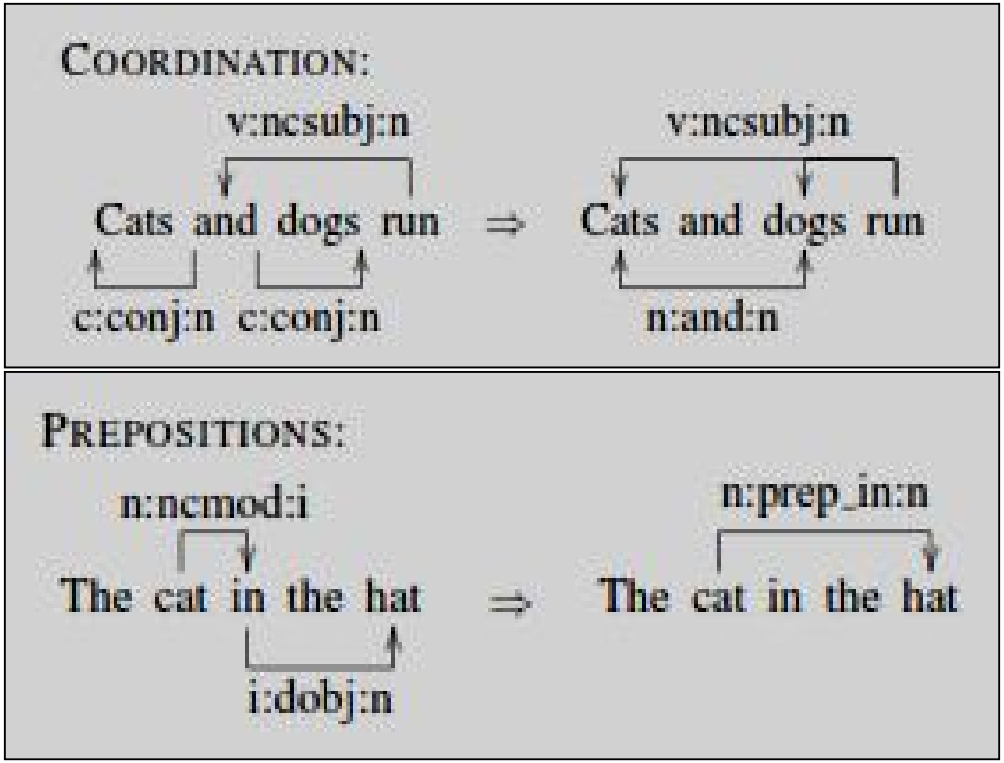
- Training
  - BNC und Wikipedia (45 Mio Sätze)
  - Nur Sim-Modelle

Lexikalische Substitution

- Training
  - 1 bag_of_words Modell (W5)
  - 3 Modelle für syntaktische Relationen
    - C > T T > C T<>C
    - Kontext: alle Relationen zwischen Nomen, Verb, Adj, Adv

Lexikalische Substitution

- Preprocessing



Lexikalische Substitution

	GAP	τ_b	Deckung
W5	44.8	0.17	100.0
C>T	48.7	0.21	86.5
T>C	49.3	0.22	86.5
T<>C	49.1	0.23	86.5
W5 + C>T	48.7	0.21	100.0
W5 + T>C	49.3	0.22	100.0
W5 + T<>C	**49.5**	**0.23**	100.0

Lexikalische Substitution

- Vergleiche Dinu & Lapata 2010
  - GAP = 42.0 / τ_b = 0.15

49.5 0.23

Lexikalische Substitution

	N (GAP)	V	Adj	Adv	Ges
W5 + T<>C	50.7	45.1	48.8	55.9	49.5
Thater et al. 2012	46.4	45.9	43.2	51.4	44.6

Lexikalische Substitution

	N τ_b	V	Adj	Adv	Ges
W5 + T<>C	<u>0.22</u>	<u>0.20</u>	<u>0.24</u>	0.24	<u>0.23</u>
DL 2010 NMF	0.15	0.14	0.16	<u>0.26</u>	0.16

Was wissen wir jetzt?

- Effektivität probabilistischer Modelle
- Einfluss von syntaktischem Kontext
- Bedeutung der Modelle für spezielle Anwendungen
  - Ähnlichkeit von Wörtern
  - Substitution / Paraphrasierung

Was kommt nun?

- Erweiterung des Anwendungsbereiches
  - Word sense disambiguation
  - Gene name normalisation
- Weiterentwicklung der Modelle
  - Correlated topic model
  - Modelling polysemy effects

Vielen Dank

- Zusätzliche Quellen
  - duden.de
  - wikipedia.de
  - openthesaurus.de
  - homepages.inf.ed.ac.uk/mlap/

Diskussion

- Problematik: Ähnlichkeit von Wörtern
- Syntaktischer Kontext
- Variablenmodell
- Para / Sim
- Evaluation
  - Ähnlichkeit im Kontext
  - Lexikalische Substitution