Comparing Computational Models of Selectional Preferences — Second-order Co-Occurrence vs. Latent Semantic Clusters

Sabine Schulte im Walde
schulte@ims.uni-stuttgart.de
SFB 732 "Incremental Specification in Context"
Project D4 "Modular Lexicalisation of Probabilistic Context-Free Grammars"
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Second-order Co-Occurrence Selectional Preferences

Idea

- Selectional preferences with respect to a predicate's complement are defined by the properties of the complement realisations
- Example: typical direct object of *drink* is fluid, might be hot or cold, can be bought, might be bottled, etc.
- Second-order co-occurrence: a predicate's restrictions to the semantic realisation of its complements are expressed through the properties of the complements

Scoring

- Selectional preference description:

$$score(p,r1,prop) = \sum_{n \in (p,r1)} freq(p,r1,n) * freq(n,r2,prop)$$

  Variations of frequency: *log(freq)* and *prob*
- Selectional preference fit of a noun by standard distributional measures: compares noun's contribution to overall preference
  → cosine, skew divergence, tau, jaccard

Data

- Corpus-based frequencies *freq(p,r1,n)* of predicates *p* and nouns *n* with respect to some functional relationship *r1;*
  r1: subjects, direct objects, pp objects
- Corpus-based joint frequencies *freq(n,r2,prop)* of nouns *n* and noun properties *prop* with respect to some functional relationship *r2;*
  r2: modifying adjectives, subcategorising verbs (for direct object), subcategorising prepositions
- Corpus source: approx. 560 million words from the German web corpus *deWaC* (Baroni & Kilgarriff, 2006)
- Preprocessing: *Tree Tagger* (Schmid, 1994) and dependency parser (Schiehlen, 2003)

Examples

- *backen,* NP-akk:

Verb	Properties: Adj		Realisations	
backen	frisch	'fresh'	Keks	'cookie'
	lecker	'delicious'	Brötchen	'roll'
	klein	'small'	Torte	'tart'
	trocken	'dry'	Kuchen	'cake'
	süß	'sweet'	Brot	'bread'
	warm	'warm'	Pizza	'pizza'
	fett	'fat'	Waffel	'waffle'
	eingeweicht	'soaked'	Pfannkuchen	'pancake'

- *anbraten,* NP-akk:

Verb	Properties: $Verb_{NPacc}$		Realisations	
anbraten	schälen	'peel'	Champignon	'mushroom'
	schneiden	'cut'	Zwiebel	'onion'
	essen	'eat'	Kartoffel	'potatoe'
	zugeben	'add'	Gemüse	'vegetable'
	anschwitzen	'sweat'	Knoblauch	'garlic'
	pellen	'peel'	Hackfleisch	'minced meat'
	riechen	'smell'	Roulade	'roulade'
	waschen	'clean'	Keule	'haunch'

- *bebauen,* PP-mit:

Verb	Properties: $Verb_{NPacc/PP}$		Realisations	
bebauen	errichten	'build'	Familienhaus	'family home'
mit	wohnen in	'live in'	Gebäude	'building'
	handeln um	'concern'	Geschäftshaus	'business house'
	zerstören	'destroy'	Mietshaus	'apartment building'
	erwerben	'acquire'	Villa	'villa'
	verlassen	'leave'	Wohngebäude	'residential building'
	einbrechen in	'break in'	Wohnung	'apartment'

Evaluation

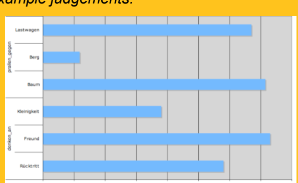
Data:

- Human judgements on 90 German subjects, direct objects, pp objects (across 30 verbs)
- Taken from Brockmann & Lapata (2003)
- Correlation of system scores with human judgements, by linear regression
- Brockmann & Lapata (BL) normalised system scores by *log10*

Baselines and Upper Bound:

- Correlation of joint corpus-based predicate-noun frequencies with judgements
- Two baselines: raw frequencies and frequencies transformed by *log10*
- Upper bound: inter-subject agreement on selectional preference judgements

Example judgements:



Selectional Restrictions

- Predicates impose selectional restrictions on their complements
- Example (Chomsky, 1957): *Colorless green ideas sleep furiously*
- Syntactically well-formed but semantically meaningless
- Realisation of complements with respect to thematic role
- Examples:
  Elsa baked a chocolate cake.
  ?Elsa baked a stone.

Selectional Preferences

- Degree of acceptability
- Probabilistic models

Computational Approaches

Cluster-based vs. WordNet-based vs. distributional

Results

2nd-order co-occurrence:

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
	log(f)	prob	log(f)	prob	log(f)	prob	log(f)	prob
adj	.447	.430	.200	.399	.185	.266	.173	.327
verb	.461	.438	.142	.221	.226	.171	.171	.234
prep	.344	.433	.220	.657	.403	.505	.265	.492
v+vp	.472	.433	.202	.318	.310	.373	.218	.310
v+vp+adj	.468	.428	.205	.414	.288	.297	.214	.335
v+vp+adj+prep	**.504**	**.452**	**.242**	**.695**	**.445**	**.541**	**.337**	**.512**
BL comparison	.408 (Resnik)		.611 (comb)		.597 (comb)		.374 (Resnik)	
baseline: f	.298		.315		.319		.289	
baseline: log10(f)	.652		.559		.565		.574	
baseline: BL	.386		.360		.168		.301	
isa	.790		.810		.820		.810	

LSC:

	SUBJ		DIR-OBJ		PP-OBJ		*all*		*all-func*	
50 training iterations										
20	.259	.550	.075	.318	.046	.387	.118	.411	.118	.402
50	.313	.559	.155	.468	.156	.195	.215	.394	.176	.468
100	.266	.553	.354	.515	.206	.297	.268	.432	.077	.386
200	.255	.546	.262	.493	.272	.266	.254	.363	.143	.341
500	.197	.477	.280	.493	.279	.274	.182	.420	.156	.463
100 training iterations										
20	.250	.441	.105	.399	.042	.386	.117	.423	.126	.461
50	.313	.662	.216	.407	.152	.418	.232	.461	.171	.478
100	.261	.515	.351	.380	.214	.596	.272	.461	.108	.301
200	.266	.546	.294	.350	.294	.602	.278	.439	.179	.328
500	.209	.501	.268	.392	.302	.605	.167	.478	.159	.445
baseline & upper bound										
f	.298		.315		.319		.289			
log10(f)	.652		.559		.565		.574			
isa	.790		.810		.820		.810			

PAC:

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
50 training iterations								
20	.197	.490	.247	.507	.043	.095	.143	.430
50	.140	.249	.266	.476	.043	.095	.110	.508
100	.140	.249	.056	.077	.043	.095	.163	.454
200	.140	.249	.056	.077	.043	.095	.148	.456
500	.140	.249	.056	.077	.043	.095		
100 training iterations								
20	.200	.487	.271	.571	.043	.095	.126	.413
50	.140	.249	.260	.525	.043	.095	.118	.497
100	.140	.249	.056	.077	.043	.095	.149	.505
200	.140	.249	.056	.077	.043	.095	.162	.496
500	.140	.249	.056	.077	.043	.095		
baseline & upper bound								
f	.298		.315		.319		.289	
log10(f)	.652		.559		.565		.574	
isa	.790		.810		.820		.810	

LSC (Latent Semantic Clusters)

- Two-dimensional clustering model
- Soft-clustering approach
- Training by EM algorithm (Baum, 1972)
- Probability model for verb-noun pairs:

$$p(v,n) = \sum_{c \in C} p(c)\ p(v \mid c)\ p(n \mid c)$$

cluster, p(c) = 0.021 (range: 0.007 - 0.065), 50 clusters

stattfinden	0.299		*Veranstaltung*	0.020
beginnen	0.217		*Diskussion*	0.020
laufen	0.075		*Prozess*	0.017
verlaufen	0.041		*Spiel*	0.016
enden	0.038		*Gespräch*	0.016
dauern	0.037		*Entwicklung*	0.013
ablaufen	0.021		*Frist*	0.012
folgen	0.018		*Ermittlung*	0.011
werden	0.017		*Verhandlung*	0.011
andauern	0.012		*Ausbildung*	0.011

cluster, p(c) = 0.015 (range: 0.004 - 0.035), 100 clusters

entwickeln	0.127		*Konzept*	0.064
vorstellen	0.071		*Angebot*	0.052
erarbeiten	0.053		*Vorschlag*	0.048
geben	0.046		*Idee*	0.044
umsetzen	0.043		*Projekt*	0.037
ansehen	0.022		*Plan*	0.024
erstellen	0.020		*Programm*	0.024
präsentieren	0.020		*Strategie*	0.024
diskutieren	0.019		*Modell*	0.023
darstellen	0.018		*Lösung*	0.018

cluster, p(c) = 0.016 (range: 0.003 - 0.036), 100 clusters

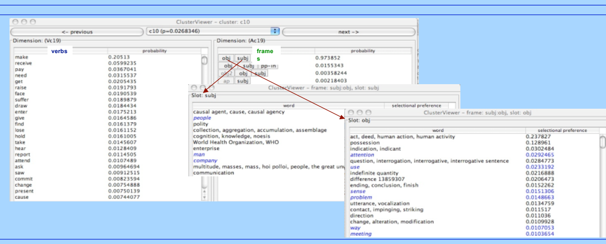
handeln um	0.119		*Entwicklung*	0.030
informieren über	0.081		*Tatsache*	0.024
hinweisen auf	0.037		*Situation*	0.020
berichten über	0.033		*Möglichkeit*	0.018
geben über	0.033		*Ergebnis*	0.017
ergeben aus	0.032		*Inhalt*	0.014
beziehen auf	0.029		*Erfahrung*	0.014
beruhen auf	0.025		*Angebot*	0.013
gelten für	0.024		*Verhalten*	0.013
verweisen auf	0.024		*Ereignis*	0.012

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll and Franz Beil (1999): „Inducing a Semantically Annotated Lexicon via EM-based Clustering". Proceedings of ACL. Maryland, MD.

PAC (Predicate-Argument Clustering)

- Extension of LSC model by selectional preferences
- Incorporates Minimum Description Length (MDL) cuts through WordNet hierarchy
- Probability model for verb-noun pairs via frame types:

$$p(v,f,a_{f1},...,a_{fn}) = \sum_{c \in C} p(c)\ p(v \mid c)\ p(f \mid c) \prod_{i=1}^{n} \sum_{r \in R} p(r \mid c,f,i)\ p(a_i \mid r)$$



Sabine Schulte im Walde, Christian Hying, Christian Scheible and Helmut Schmid (2008): „Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences." Proceedings of ACL. Columbus, OH.