A Second-order Co-Occurrence Model for Selectional Preferences

Linguistic Evidence 2010
Universität Tübingen

Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

February 13, 2010

Outline

Selectional Restrictions

- Predicates impose selectional restrictions on their complements
- Famous example: Chomsky (1957)

  Colorless green ideas sleep furiously
- Syntactically well-formed but not semantically meaningful
- Further example:

  Elsa baked a chocolate cake.
  ?Elsa baked a stone.
- Realisation of complement with reference to thematic role

Selectional Restrictions vs. Selectional Preferences

- Restriction: a predicate cannot be combined with arbitrary complements → restriction to semantic categories
- Preference:
  - degree of acceptability
  - probabilistic models

Motivation

- Generalisation over specific complement heads helps with data sparseness, e.g.,

  drink {*coffee, tea, beer, wine*}
  $\longrightarrow$ *drink* ⟨*beverage*⟩
  $\longrightarrow$ *drink* *regina*

- Requires knowledge of semantic categories:
  - clusters
  - WordNet
  - distributional information

Overview

- Cluster-based selectional preferences:
  EM-based clusters generalise over seen and unseen data
  - Pereira et al. (1993)
  - Rooth et al. (1999)
  - Schulte im Walde et al. (2008)
- WordNet-based selectional preferences:
  WordNet classes generalise over subordinate instances
  - Resnik (1997): association strength
  - Li & Abe (1998): MDL cut
  - Abney & Light (1999): HMM
  - Ciaramita & Johnson (2000): Bayesian belief network
  - Clark & Weir (2002): MDL cut
  - Light & Greiff (2002): summary of approaches
  - Brockmann & Lapata (2003): comparison of approaches
- Distributional selectional preferences:
  distributional descriptions as abstractions over specific complements
  - Erk (2007)

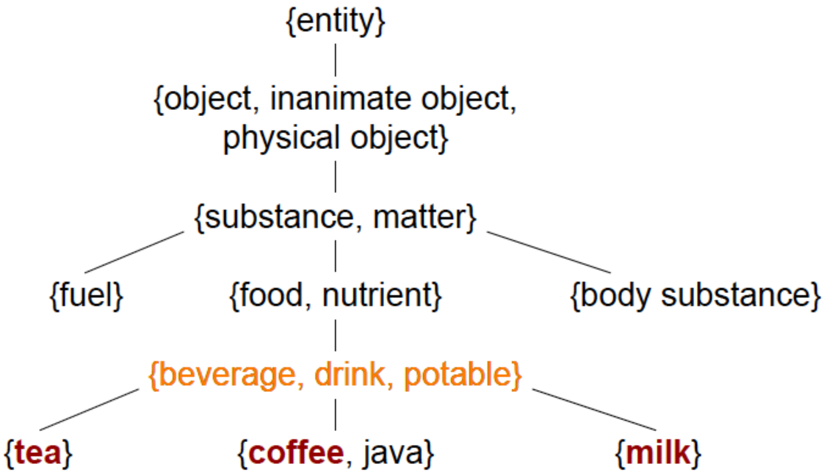
Cluster: Example

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

entwickeln	0.127
vorstellen	0.071
erarbeiten	0.053
geben	0.046
umsetzen	0.043
ansehen	0.022
erstellen	0.020
präsentieren	0.020
diskutieren	0.019
darstellen	0.018

Konzept	0.064
Angebot	0.052
Vorschlag	0.048
Idee	0.044
Projekt	0.037
Plan	0.024
Programm	0.024
Strategie	0.024
Modell	0.023
Lösung	0.018

WordNet: Example



Comparison (of WordNet Approaches)

- Data: German verb-argument pairs with 30 subjects, 30 direct objects, 30 prepositional objects (10 verbs each)
- Models: Resnik (1997), Li & Abe (1998), Clark & Weir (2002), co-occurrence frequency, conditional probability
- Comparison of models against human judgements on acceptability
- All five models are significantly correlated with human judgements
- Inter-subject agreement is higher than correlations
- No model performs best; different methods are suited for different argument functions
- Combination of models by multiple linear regression outperforms individual models

Distributional Approach

- Contexts of a linguistic unit tell us something about the meaning of the linguistic unit
- Example: corpus can tell us that one can *buy, peal, and eat an apple*
- Distributional hypothesis:

  You shall know a word by the company it keeps. (Firth, 1957)

  Each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts. (Harris, 1968)
- Basis for selectional preference model:
  co-occurrence of triples $\langle \textit{predicate}, \textit{relation}, \textit{complement} \rangle$

2nd-Order Co-Occurrence: Idea

- Selectional preferences with respect to a predicate's complement are defined by the properties of the complement realisations
- Example question: what characterises the direct objects of *drink*?

2nd-Order Co-Occurrence: Idea

- Selectional preferences with respect to a predicate's complement are defined by the properties of the complement realisations
- Example question: what characterises the direct objects of *drink*?
- Example: typical direct object of *drink* is fluid, might be hot or cold, can be bought, might be bottled, etc.

2nd-Order Co-Occurrence: Idea

- Selectional preferences with respect to a predicate's complement are defined by the properties of the complement realisations
- Example question: what characterises the direct objects of *drink*?
- 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

Idea: Example

Example: *backen* 'bake' ⟨NPnom,NPacc⟩

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'

Idea: Example

Example: *anbraten* 'fry' ⟨NPnom,NPacc⟩

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'

Data

- Corpus-based joint frequencies $freq(p, r1, n)$ of predicates p and nouns n with respect to some functional relationship $r1$;
  $r1$: subjects, direct object, 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)

Scoring

1. Selectional preference description:

$$score_1(p, r1, prop) = \sum_{n \in (p, r1)} freq(p, r1, n) \ * \ freq(n, r2, prop)$$

Scoring

1. Selectional preference description:

$$score_1(p, r1, prop) = \sum_{n \in (p,r1)} freq(p, r1, n) \; * \; freq(n, r2, prop)$$

$$score_2(p, r1, prop) = \sum_{n \in (p,r1)} log(freq(p, r1, n)) \; * \; log(freq(n, r2, prop))$$

Scoring

1. Selectional preference description:

$$score_1(p, r1, prop) = \sum_{n \in (p,r1)} freq(p, r1, n) \; * \; freq(n, r2, prop)$$

$$score_2(p, r1, prop) = \sum_{n \in (p,r1)} log(freq(p, r1, n)) \; * \; log(freq(n, r2, prop))$$

$$score_3(p, r1, prop) = \sum_{n \in (p,r1)} prob(p, r1, n) \; * \; prob(n, r2, prop)$$

Scoring: Example

$$freq(drink, dir\_obj, coffee) = 50$$
$$freq(drink, dir\_obj, tea) = 5$$

$$freq(coffee, n\_mod, hot) = 100$$
$$freq(coffee, n\_mod, fluid) = 30$$

$$freq(tea, n\_mod, hot) = 60$$
$$freq(tea, n\_mod, fluid) = 15$$

Scoring: Example

$freq(drink, dir\_obj, coffee) = 50$
$freq(drink, dir\_obj, tea) = 5$

$freq(coffee, n\_mod, hot) = 100$
$freq(coffee, n\_mod, fluid) = 30$

$freq(tea, n\_mod, hot) = 60$
$freq(tea, n\_mod, fluid) = 15$

$score_1(drink, dir\_obj, hot) = 50 * 100 \ + \ 5 * 60 = 5,300$
$score_1(drink, dir\_obj, fluid) = 50 * 30 \ + \ 5 * 15 = 1,575$

Scoring: Example

$$freq(drink, dir\_obj, coffee) = 50$$
$$freq(drink, dir\_obj, tea) = 5$$

$$freq(coffee, n\_mod, hot) = 100$$
$$freq(coffee, n\_mod, fluid) = 30$$

$$freq(tea, n\_mod, hot) = 60$$
$$freq(tea, n\_mod, fluid) = 15$$

$$score_1(drink, dir\_obj, hot) = 50 * 100 \ + \ 5 * 60 = 5,300$$
$$score_1(drink, dir\_obj, fluid) = 50 * 30 \ + \ 5 * 15 = 1,575$$

$$score_2(drink, dir\_obj, hot) = log(50) * log(100) + log(5) * log(60) = 24.61$$
$$score_2(drink, dir\_obj, fluid) = log(50) * log(30) + log(5) * log(15) = 17.66$$

Scoring: Example

$$freq(drink, dir\_obj, coffee) = 50$$
$$freq(drink, dir\_obj, tea) = 5$$

$$freq(coffee, n\_mod, hot) = 100$$
$$freq(coffee, n\_mod, fluid) = 30$$

$$freq(tea, n\_mod, hot) = 60$$
$$freq(tea, n\_mod, fluid) = 15$$

$$score_1(drink, dir\_obj, hot) = 50 * 100 \; + \; 5 * 60 = 5,300$$
$$score_1(drink, dir\_obj, fluid) = 50 * 30 \; + \; 5 * 15 = 1,575$$

$$score_2(drink, dir\_obj, hot) = log(50) * log(100) \; + \; log(5) * log(60) = 24.61$$
$$score_2(drink, dir\_obj, fluid) = log(50) * log(30) \; + \; log(5) * log(15) = 17.66$$

$$score_3(drink, dir\_obj, hot) = 0.91 * 0.77 \; + \; 0.09 * 0.80 = 0.77$$
$$score_3(drink, dir\_obj, fluid) = 0.91 * 0.23 \; + \; 0.09 * 0.20 = 0.23$$

Scoring

2. Selectional preference fit of a specific noun by standard distributional measures: compares noun's contribution to overall preference

Scoring

2. Selectional preference fit of a specific noun by standard distributional measures: compares noun's contribution to overall preference

   - *cosine*, standard measure in linear algebra

     $$cos(x, y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Scoring

2. Selectional preference fit of a specific noun by standard distributional measures: compares noun's contribution to overall preference

- *cosine*, standard measure in linear algebra

$$cos(x,y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- *skew divergence*, information-theoretic measure and variant of the Kullback-Leibler divergence, cf. Lee (2001)

$$KL(x||y) = \sum_{i=1}^{n} x_i \ * \ log \ \frac{x_i}{y_i}$$

$$skew(x,y) = KL(x||w * y \ + \ (1-w) * x), w = 0.9$$

Scoring

- *Kendall's τ*, a measure for rank correlation, cf. Hatzivassiloglou & McKeown (1993), Lapata (2006)

$$\tau(x, y) \;=\; \frac{f_{agree}}{f_{agree} + f_{disagree}} \;-\; \frac{f_{disagree}}{f_{agree} + f_{disagree}}$$

Scoring

- *Kendall's* τ, a measure for rank correlation, cf. Hatzivassiloglou & McKeown (1993), Lapata (2006)

$$\tau(x, y) \;=\; \frac{f_{agree}}{f_{agree} + f_{disagree}} \;-\; \frac{f_{disagree}}{f_{agree} + f_{disagree}}$$

- *jaccard index*, a binary distance measure, cf. Manning & Schütze (1999)

$$jaccard(x, y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Data

- Human judgements on German subjects, direct objects and pp objects, cf. Brockmann & Lapata (2003)
- Correlation of system scores with human judgements, by linear regression
- Brockmann & Lapata normalised system scores and human judgements by $log10$

Baselines and Upper Bound

- Baseline: correlation of joint corpus-based predicate-noun frequencies of subjects, direct objects and pp objects with human judgements, also by linear regression
- Two baselines: raw frequencies and frequencies transformed by $log10$
- Upper bound: inter-subject agreement on selectional preference judgements

Results

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
	log(f)	prob	log(f)	prob	log(f)	prob	log(f)	prob
adj (a)	.447	.430	.200	.399	.185	.266	.173	.327
verb (v)	.461	.438	.142	.221	.226	.171	.171	.234
prep (p)	.344	.433	.220	.657	.403	.505	.265	.492
v+vp	.472	.433	.202	.318	.310	.373	.218	.310
v+vp+a	.468	.428	.205	.414	.288	.297	.214	.335
v+vp+a+p	**.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	

Results

- Best scoring: probabilities
- Best measure: cosine;
  skew and τ are similar; jaccard is lowest
- Normalising system scores by $log 10$ decreases results
- Most successful features: v+a+prep, or prep only
- Direct objects are modelled better than subjects or pp objects
- Large difference in baseline results (BL vs. ours);
  probably due to corpus size

Summary

- 2nd-order co-occurrence provides insights into properties of selectional preferences
- Simple and intuitive distributional model beats WordNet-based preferences in most cases
- Best performing properties are prepositions and general distributional descriptions → compare with larger features sets (e.g., window-based co-occurrence)
- Difficult to outperform frequency baselines
- Evaluation suboptimal → compare with ranking evaluation
- Effect of corpus size

Related Work: Erk (2007)

- Primary corpus: extract tuples $\langle p, r, w\rangle$ of a predicate p, an argument position r, and a seen headword w
- Generalisation corpus: compute a corpus-based semantic similarity metric
- Selectional preference S of a functional relation r for a possible headword w_0 is modelled as a weighted sum (weight: α) of the similarities between w_0 and the seen headwords w:

  $S_{r_p}(w_0) = \sum_{w \in Seen(r_p)} sim(w_0, w) * \alpha_{r_p}(w)$

Idea: Example

Example: *abflauen* 'calm down' ⟨NPnom,...⟩

Verb	Properties: Adj		Realisations	
abflauen	frisch	'cool'	Interesse	'interest'
	stark	'strong'	Sturm	'storm'
	heftig	'strong'	Begeisterung	'enthusiasm'
	kalt	'cold'	Wind	'wind'
	öffentlich	'public'	Protest	'protest'
	wirtschaftlich	'economic'	Wachstum	'increase'
	national	'national'	Kampf	'fight'

Idea: Example

Example: *bebauen* 'build' $\langle \ldots, PP_{mit}, \ldots \rangle$

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'

References

Steven Abney and Marc Light.
Hiding a Semantic Class Hierarchy in a Markow Model.
In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8, College Park, MD, 1999.

Marco Baroni and Adam Kilgarriff.
Large Linguistically-processed Web Corpora for Multiple Languages.
In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.

Carsten Brockmann and Mirella Lapata.
Evaluating and Combining Approaches to Selectional Preference Acquisition.
In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 27–34, Budapest, Hungary, 2003.

Noam Chomsky.
Syntactic Structures.
Mouton, The Hague, 1957.

Massimiliano Ciaramita and Mark Johnson.
Explaining away Ambiguity: Learning Verb Selectional Preference with Bayesian Networks.
In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 187–193, Saarbrücken, Germany, 2000.

References

Stephen Clark and David Weir.
Class-Based Probability Estimation using a Semantic Hierarchy.
Computational Linguistics, 28(2):187–206, 2002.

Katrin Erk.
A Simple, Similarity-based Model for Selectional Preferences.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

Vasileios Hatzivassiloglou and Kathleen R. McKeown.
Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning.
In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus, OH, 1993.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell.
The Sketch Engine.
In *Proceedings of the 11th EURALEX International Congress*, pages 105–111, Lorient, France, 2004.

Mirella Lapata.
Automatic Evaluation of Information Ordering: Kendall's Tau.
Computational Linguistics, 32(4):471–484, 2006.

References

Lillian Lee.
On the Effectiveness of the Skew Divergence for Statistical Language Analysis.
Artificial Intelligence and Statistics, pages 65–72, 2001.

Hang Li and Naoki Abe.
Generalizing Case Frames Using a Thesaurus and the MDL Principle.
Computational Linguistics, 24(2):217–244, 1998.

Marc Light and Warren R. Greiff.
Statistical Models for the Induction and Use of Selectional Preferences.
Cognitive Science, 26(3):269–281, 2002.

Christopher D. Manning and Hinrich Schütze.
Foundations of Statistical Natural Language Processing.
MIT Press, Cambridge, MA, 1999.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi.
Detecting Compositionality of Verb-Object Combinations using Selectional Preferences.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 369–379, 2007.

Fernando Pereira, Naftali Tishby, and Lillian Lee.
Distributional Clustering of English Words.
In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH, 1993.

References

Philip Resnik.
Selectional Preference and Sense Disambiguation.
In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC, 1997.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil.
Inducing a Semantically Annotated Lexicon via EM-Based Clustering.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD, 1999.

Michael Schiehlen.
A Cascaded Finite-State Parser for German.
In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary, 2003.

Helmut Schmid.
Probabilistic Part-of-Speech Tagging using Decision Trees.
In *Proceedings of the 1st International Conference on New Methods in Language Processing*, 1994.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid.
Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences.
In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, 2008.