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

Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

LREC 2010, Valletta, Malta
May 19-21, 2010

Outline

Selectional Restrictions and Selectional Preferences

- Selectional Restriction: a predicate cannot be combined with arbitrary complements $\longrightarrow$ restriction to semantic categories

- Famous example: Chomsky (1957)

  Colorless green ideas sleep furiously

  Syntactically well-formed but not semantically meaningful

Selectional Restrictions and Selectional Preferences

- Selectional Restriction: a predicate cannot be combined with arbitrary complements → restriction to semantic categories

- Famous example: Chomsky (1957)

  Colorless green ideas sleep furiously

  Syntactically well-formed but not semantically meaningful

- Selectional Preference:
  - degree of acceptability
  - probabilistic models

Computational Motivation

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

  drink {*coffee, tea, beer, wine*}
  → *drink* ⟨*beverage*⟩
  → *drink* *regina* (German regional type of lemonade)

- 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)

Idea

- Distributional approach: contexts of a linguistic unit provide information about the meaning of the linguistic unit, cf. Firth (1957), Harris (1968)
- 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*?

Idea

- Distributional approach: contexts of a linguistic unit provide information about the meaning of the linguistic unit, cf. Firth (1957), Harris (1968)
- 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.

Idea

- Distributional approach: contexts of a linguistic unit provide information about the meaning of the linguistic unit, cf. Firth (1957), Harris (1968)
- 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.
  $\longrightarrow$ second-order co-occurrence

Idea: Example

Example: *backen* 'bake' $\langle$NPnom,NPacc$\rangle$

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'

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

- Selectional preference description: rates second-order properties according to their contribution to selectional preference description

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

  with $func = freq, log(freq), prob, tf - idf$

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

  cosine, skew divergence, Kendall's τ, jaccard index

Latent Semantic Clusters (LSC)

- Instance of the Expectation-Maximisation algorithm (Baum 1972) for unsupervised training on unannotated data
- Two-dimensional soft clusters (Rooth et al. 1999)

$$\begin{aligned} prob(p,n) &= \sum_{c \in cluster} prob(c,p,n) \\ &= \sum_{c \in cluster} prob(c)\ prob(p,c)\ prob(n,c) \end{aligned}$$

- Clusters can be considered as generalisations over (seen und unseen) members of the two inter-dependent dimensions
- Selectional preference fit: probabilities of verb–noun pairs
- Same corpus data as for the distributional model
- One model for each relation, plus one model with all relations
- Parameters: 20, 50, 100, 200, 500 clusters; 50, 100 iterations

LSC: Example

cluster, prob(c) = 0.015 (range: 0.004-0.035)

entwickeln	'develop'	Konzept	'concept'
vorstellen	'introduce'	Angebot	'offer'
erarbeiten	'work out'	Vorschlag	'suggestion'
geben	'give'	Idee	'idea'
umsetzen	'realise'	Projekt	'project'
ansehen	'look at'	Plan	'plan'
erstellen	'create'	Programm	'program'
präsentieren	'present'	Strategie	'strategy'
diskutieren	'discuss'	Modell	'model'
darstellen	'demonstrate'	Lösung	'solution'

Predicate Argument Clustering (PAC)

- Extension of LSC approach (Schulte im Walde et al. 2008)
- Combination of EM algorithm and Minimum Description Length principle (Rissanen, 1978)
- Incorporates explicit, WordNet-based selectional preferences

$$prob(p, f, n_1, ..., n_k) = \sum_c prob(p)\ prob(p, c)\ prob(f, c) * \prod_{i=1}^{k} \sum_{r \in wn} prob(r|c, f, i)\ prob(n_i|r)$$

- Selectional preference fit: probabilities of verb–noun pairs
- Same corpus data as for the distributional model
- One model for each relation, plus one model with all relations
- Parameters: 20, 50, 100, 200, 500 clusters; 50, 100 iterations

PAC: Example

cluster, prob(c) = 0.069 (range: 0.014-0.085)

leisten	'perform'	Geschehen	'event'
geben	'give'	Aktivität	'activity'
fordern	'demand'	Veränderung	'change'
bedeuten	'mean'	Handlungssequenz	'action sequence'
ermöglichen	'enable'	Realisierung	'realisation'
verhindern	'prevent'	Anschlag	'attack'
feiern	'celebrate'	Straftat	'criminal act'
darstellen	'demonstrate'	Gerichtsverfahren	'lawsuit'
bringen	'bring'	Verbesserung	'improvement'
vornehmen	'carry out'	Optimierung	'optimisation'

Questions

1. Distributional approach:

   How well does 2nd-order co-occurrence model selectional preferences?

   Which 2nd-order properties are most salient?

2. Comparison of models:

   How does a simple distributional model compare with more complex, cluster-based approaches?

Data

- Human judgements on selectional preference fit for German verb–noun pairs, cf. Brockmann & Lapata (2003)
- 30 subjects, 30 direct objects and 30 pp objects (10 verbs each)

- Brockmann & Lapata (BL) compared WordNet-based selectional preference models and a combination of models
- BL normalised system scores and human judgements by $log10$

Data

- Human judgements on selectional preference fit for German verb–noun pairs, cf. Brockmann & Lapata (2003)
- 30 subjects, 30 direct objects and 30 pp objects (10 verbs each)

- Brockmann & Lapata (BL) compared WordNet-based selectional preference models and a combination of models
- BL normalised system scores and human judgements by $log 10$

Correlation of system scores with human judgements, using

1. linear regression
2. Spearman rank-order correlation coefficient

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 and by ranking
- Two baselines: raw frequencies and frequencies transformed by $log10$
- Upper bound: inter-subject agreement (isa) on selectional preference judgements

Overview (Linear Regression)

Models:

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
Distrib.	**.494	verb, prob	***.713	union, freq	***.602	prep, tf-idf	***.517	union, prob
LSC	*.450	20c, 50i	***.569	100c, 100i	**.562	200c, 100i	***.453	50c, 50i
PAC	***.651	20c, 100i	***.795	500c, 100i	**.481	500c, 50i	***.543	100c, 50i
BL	*.408 (Resnik)		***.611 (Clark/Weir)		***.597 (Clark/Weir)		***.400 (comb)	

Baselines and Upper Bound:

f	.274	.343	.384	.313
log10(f)	.652	.559	.565	.574
BL	.386	.360	.168	.301
isa	.790	.810	.820	.810

Significance levels: $*p \leq .05$, $**p \leq .01$, and $***p \leq .001$

Results

- PAC > 2nd-order > LSC
- Similar but not identical results with two evaluations
- Best results vary according to functional relation (and approach)
- High baseline values; strong differences in BL and our baselines
- *log*10 transformations better than original scores

Results

- PAC $>$ 2nd-order $>$ LSC
- Similar but not identical results with two evaluations
- Best results vary according to functional relation (and approach)
- High baseline values; strong differences in BL and our baselines
- $log10$ transformations better than original scores
- Second-order co-occurrence:
  - properties: prepositions and union of properties are best
  - property scoring function: prob and tf-idf $>$ freq and log(freq)
  - selectional preference fit: cosine $> \tau >$ skew $>$ jaccard

Results

- PAC $>$ 2nd-order $>$ LSC
- Similar but not identical results with two evaluations
- Best results vary according to functional relation (and approach)
- High baseline values; strong differences in BL and our baselines
- $log10$ transformations better than original scores
- Second-order co-occurrence:
  - properties: prepositions and union of properties are best
  - property scoring function: prob and tf-idf $>$ freq and log(freq)
  - selectional preference fit: cosine $> \tau >$ skew $>$ jaccard
- Clustering approaches:
  - better when all functions are trained in one model
  - no clear tendency towards an optimal parameter setting

Summary

- Three computational approaches to selectional preferences: intuitive 2nd-order co-occurrence vs. latent semantic clusters
- High correlations between models and human judgements, but powerful frequency baseline is not met
- Answers to questions:
  1. Distributional approach: How well does 2nd-order co-occurrence model selectional preferences?
     $\rightarrow$ highly significant correlations (.494/.713/.602/.517)

     Which 2nd-order properties are most salient?
     $\rightarrow$ prepositions and union of properties
  2. Comparison of models: How does a simple distributional model compare with more complex, cluster-based approaches?
     $\rightarrow$ better than LSC but worse than PAC

Second-order Co-Occurrence: 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'

Second-order Co-Occurrence: 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'

Second-order Co-Occurrence: Example

Example: *bebauen* 'build' $\langle \ldots, \mathrm{PP}_{mit}, \ldots \rangle$

Verb	Properties: $\text{Verb}_{NPacc\,/\,PP}$		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'

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.

Leonard E. Baum.
An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes.
Inequalities, III:1–8, 1972.

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.

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.

John R. Firth.
Papers in Linguistics 1934-51.
Longmans, London, UK, 1957.

Zellig Harris.
Distributional Structure.
In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press, 1968.

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.

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.

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.

Jorma Rissanen.
Modeling by Shortest Data Description.
Automatica, 14:465–471, 1978.

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.