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

1 Selectional Preferences

2 Selectional Preference Models and Experiments
   Second-order Co-Occurrence
   Latent Semantic Clusters
   Latent Semantic Clusters integrating Selectional Preferences

3 Evaluation

4 Results
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 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.,
  
  \[ \text{drink} \{ \text{coffee, tea, beer, wine} \} \]
  
  \[ \rightarrow \text{drink} \langle \text{beverage} \rangle \]
  
  \[ \rightarrow \text{drink regina} \text{ (German regional type of lemonade)} \]

- 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)
• **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.  
  → **second-order co-occurrence**
Idea: Example

Example: *backen* 'bake' \(\langle\text{NPnom},\text{NPacc}\rangle\)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Properties: Adj</th>
<th>Realisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>backen</td>
<td>frisch 'fresh'</td>
<td>Keks 'cookie'</td>
</tr>
<tr>
<td></td>
<td>lecker 'delicious'</td>
<td>Brötchen 'roll'</td>
</tr>
<tr>
<td></td>
<td>klein 'small'</td>
<td>Torte 'tart'</td>
</tr>
<tr>
<td></td>
<td>trocken 'dry'</td>
<td>Kuchen 'cake'</td>
</tr>
<tr>
<td></td>
<td>süß 'sweet'</td>
<td>Brot 'bread'</td>
</tr>
<tr>
<td></td>
<td>warm 'warm'</td>
<td>Pizza 'pizza'</td>
</tr>
<tr>
<td></td>
<td>fett 'fat'</td>
<td>Waffel 'waffle'</td>
</tr>
<tr>
<td></td>
<td>eingeweicht 'soaked'</td>
<td>Pfannkuchen 'pancake'</td>
</tr>
</tbody>
</table>
Data

- Corpus-based joint frequencies $freq(p, r_1, n)$ of predicates $p$ and nouns $n$ with respect to some functional relationship $r_1$; $r_1$: subjects, direct object, pp objects

- Corpus-based joint frequencies $freq(n, r_2, prop)$ of nouns $n$ and noun properties $prop$ with respect to some functional relationship $r_2$; $r_2$: 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

\[
score(p, r1, prop) = \sum_{n \in (p, r1)} func(p, r1, n) \ast func(n, r2, prop)
\]

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 \( \tau \), 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)

\[
prob(p, n) = \sum_{c \in \text{cluster}} prob(c, p, n)
\]

\[
= \sum_{c \in \text{cluster}} prob(c) \cdot prob(p, c) \cdot prob(n, c)
\]

- 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

<table>
<thead>
<tr>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>entwickeln</td>
<td>'develop'</td>
</tr>
<tr>
<td>vorstellen</td>
<td>'introduce'</td>
</tr>
<tr>
<td>erarbeiten</td>
<td>'work out'</td>
</tr>
<tr>
<td>geben</td>
<td>'give'</td>
</tr>
<tr>
<td>umsetzen</td>
<td>'realise'</td>
</tr>
<tr>
<td>ansehen</td>
<td>'look at'</td>
</tr>
<tr>
<td>erstellen</td>
<td>'create'</td>
</tr>
<tr>
<td>präsentieren</td>
<td>'present'</td>
</tr>
<tr>
<td>diskutieren</td>
<td>'discuss'</td>
</tr>
<tr>
<td>darstellen</td>
<td>'demonstrate'</td>
</tr>
</tbody>
</table>

**cluster**, $\text{prob}(c) = 0.015$ (range: 0.004-0.035)

- Konzept: 'concept'
- Angebot: 'offer'
- Vorschlag: 'suggestion'
- Idee: 'idea'
- Projekt: 'project'
- Plan: 'plan'
- Programm: 'program'
- Strategie: 'strategy'
- Modell: 'model'
- 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

\[
prob(p, f, n_1, \ldots, n_k) = \sum_c prob(p) \, prob(p, c) \, prob(f, c) \ast \\
\prod_{i=1}^{k} \sum_{r \in wn} prob(r|c, f, i) \, prob(n_i|r)
\]

- 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

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

- 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 $\log_{10}$
Data

- 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 $\log_{10}$

- **Upper bound**: inter-subject agreement (isa) on selectional preference judgements
Overview (Linear Regression)

Models:

<table>
<thead>
<tr>
<th></th>
<th>SUBJ</th>
<th>DIR-OBJ</th>
<th>PP-OBJ</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distrib.</td>
<td>**.494</td>
<td>**.713</td>
<td>**.602</td>
<td>**.517</td>
</tr>
<tr>
<td>LSC</td>
<td>* .450</td>
<td>**.569</td>
<td>**.562</td>
<td>**.453</td>
</tr>
<tr>
<td>PAC</td>
<td>***.651</td>
<td>**.795</td>
<td>**.481</td>
<td>**.543</td>
</tr>
<tr>
<td>BL</td>
<td>*.408</td>
<td>***.611</td>
<td>***.597</td>
<td>***.400</td>
</tr>
</tbody>
</table>

Significance levels: *p ≤ .05, **p ≤ .01, and ***p ≤ .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
- log10 transformations better than original scores
• 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
• 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
• $\log_{10}$ 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?
     → highly significant correlations (.494/.713/.602/.517)
     Which 2nd-order properties are most salient?
     → prepositions and union of properties
  2. Comparison of models: How does a simple distributional model compare with more complex, cluster-based approaches?
     → better than LSC but worse than PAC
Example: *anbraten* 'fry' \(\langle\text{NPnom, NPacc}\rangle\)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Properties: Verb(_{NPacc})</th>
<th>Realisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>anbraten</td>
<td>schälen 'peel'</td>
<td>Champignon 'mushroom'</td>
</tr>
<tr>
<td></td>
<td>schneiden 'cut'</td>
<td>Zwiebel 'onion'</td>
</tr>
<tr>
<td></td>
<td>essen 'eat'</td>
<td>Kartoffel 'potatoe'</td>
</tr>
<tr>
<td></td>
<td>zugeben 'add'</td>
<td>Gemüse 'vegetable'</td>
</tr>
<tr>
<td></td>
<td>anschwitzen 'sweat'</td>
<td>Knoblauch 'garlic'</td>
</tr>
<tr>
<td></td>
<td>pellen 'peel'</td>
<td>Hackfleisch 'minced meat'</td>
</tr>
<tr>
<td></td>
<td>riechen 'smell'</td>
<td>Roulade 'roulade'</td>
</tr>
<tr>
<td></td>
<td>waschen 'clean'</td>
<td>Keule 'haunch'</td>
</tr>
</tbody>
</table>
**Second-order Co-Occurrence: Example**

Example: *abflauen* 'calm down' \(\langle \text{NPnom,} \ldots \rangle\)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Properties: Adj</th>
<th>Realisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>abflauen</td>
<td>frisch 'cool'</td>
<td>Interesse 'interest'</td>
</tr>
<tr>
<td></td>
<td>stark 'strong'</td>
<td>Sturm 'storm'</td>
</tr>
<tr>
<td></td>
<td>heftig 'strong'</td>
<td>Begeisterung 'enthusiasm'</td>
</tr>
<tr>
<td></td>
<td>kalt 'cold'</td>
<td>Wind 'wind'</td>
</tr>
<tr>
<td>öffentlich</td>
<td>öffentlich 'public'</td>
<td>Protest 'protest'</td>
</tr>
<tr>
<td>wirtschaftlich</td>
<td>wirtschaftlich 'economic'</td>
<td>Wachstum 'increase'</td>
</tr>
<tr>
<td>national</td>
<td>national 'national'</td>
<td>Kampf 'fight'</td>
</tr>
</tbody>
</table>
Second-order Co-Occurrence: Example

Example: *bebauen* 'build' ⟨..., PP_{mit}, ...⟩

<table>
<thead>
<tr>
<th>Verb</th>
<th>Properties: Verb_{NPacc/PP}</th>
<th>Realisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>bebauen</td>
<td>errichten</td>
<td>Familienhaus</td>
</tr>
<tr>
<td>mit</td>
<td>wohnen in</td>
<td>'family home'</td>
</tr>
<tr>
<td></td>
<td>handeln um</td>
<td>Gebäude</td>
</tr>
<tr>
<td></td>
<td>zerstören</td>
<td>'building'</td>
</tr>
<tr>
<td></td>
<td>erwerben</td>
<td>Geschäftshaus</td>
</tr>
<tr>
<td></td>
<td>verlassen</td>
<td>'business house'</td>
</tr>
<tr>
<td></td>
<td>einbrechen in</td>
<td>Mietshaus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'apartment building'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Villa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'villa'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wohngebäude</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'residential building'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wohnung</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'apartment'</td>
</tr>
</tbody>
</table>
Steven Abney and Marc Light.
Hiding a Semantic Class Hierarchy in a Markow Model.

Marco Baroni and Adam Kilgarriff.
Large Linguistically-processed Web Corpora for Multiple Languages.

Leonard E. Baum.
An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes.

Carsten Brockmann and Mirella Lapata.
Evaluating and Combining Approaches to Selectional Preference Acquisition.


John R. Firth.
*Papers in Linguistics 1934-51.*

Zellig Harris.
Distributional Structure.

Hang Li and Naoki Abe.
Generalizing Case Frames Using a Thesaurus and the MDL Principle.


Marc Light and Warren R. Greiff.
Statistical Models for the Induction and Use of Selectional Preferences.


Michael Schiehlen.
A Cascaded Finite-State Parser for German.

Helmut Schmid.
Probabilistic Part-of-Speech Tagging using Decision Trees.

Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences.