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

Sabine Schulte im Walde

Institute for Natural Language Processing
University of Stuttgart, Germany
schulte@ims.uni-stuttgart.de

Abstract

This paper presents a comparison of three computational approaches to selectional preferences: (i) an intuitive distributional approach that uses second-order co-occurrence of predicates and complement properties; (ii) an EM-based clustering approach that models the strengths of predicate–noun relationships by latent semantic clusters; and (iii) an extension of the latent semantic clusters by incorporating the MDL principle into the EM training, thus explicitly modelling the predicate–noun selectional preferences by WordNet classes. We describe various experiments on German data and two evaluations, and demonstrate that the simple distributional model outperforms the more complex cluster-based models in most cases, but does itself not always beat the powerful frequency baseline.

1. Introduction

Predicates impose selectional restrictions on the realisation of their complements, as first illustrated by Chomsky (1957) through his famous example *"Colorless green ideas sleep furiously"*. Though the sentence is syntactically well-formed, it is not semantically meaningful, unless interpreted metaphorically. Compare also examples (1) and (2), where most people would agree that a chocolate cake is highly acceptable as the patient of the verb *bake*, but a stone is less typical (though, again, it might be subsumed by the context, e.g., as metaphorical when the baking result was unenjoyably hard).

(1) Elsa baked *a chocolate cake*.

(2) ?Elsa baked *a stone*.

Approaches in computational linguistics that model selectional restrictions commonly refer to them as *selectional preferences*. This illustrates the fact that selectional restrictions refer to a certain degree of acceptance, rather than a binary decision, and furthermore that this degree of acceptance is typically represented by probabilistic models.

Selectional preferences are of great interest to research in Computational Linguistics. From a lexicographic perspective, they constitute a crucial part of the lexical semantic knowledge and thus are to be included in the lexicon, cf. the word sketches in the *Sketch Engine* (Kilgarriff et al., 2004) as a prominent lexicographic example. Furthermore, one can rely on lexical selectional preferences to detect accordance with or discrepancies to regularities. This feature allows, e.g., to determine the degree of compositionality of multi-word expressions (McCarthy et al., 2007). From a more practical perspective, selectional preferences can help with the pervasive problem of data sparseness: They can be regarded as generalisations of semantic realisations (e.g., if typical direct objects of the verb *drink* are *coffee, tea, beer, wine*, one can describe the selectional restrictions of this complement by their hypernym *beverage*), and use the generalisations to induce properties for sparse nominal instances, such as *regina*, a regional brand of German lemonade, which is probably not or only sparsely captured by corpus data. Example applications in this direction are word sense disambiguation relying on selectional preferences (McCarthy and Carroll, 2003), and semantic role labelling (Erk, 2007; Zapirain et al., 2009).

This paper presents a comparison of three computational approaches to selectional preferences: (i) an intuitive distributional approach that uses second-order co-occurrence of predicates and complement properties; (ii) an EM-based clustering approach that models the strengths of predicate–noun relationships by latent semantic clusters (Rooth et al., 1999); and (iii) an extension of the latent semantic clusters by incorporating the MDL principle into the EM training, thus explicitly modelling the predicate–noun selectional preferences by WordNet classes (Schulte im Walde et al., 2008).

The motivation of our work was driven by two main question: First, concerning the distributional approach, we are interested not only in how well the model describes selectional preferences, but moreover which second-order properties are most salient. For example, a typical direct object of the verb *drink* is usually fluid, might be hot or cold, can be bought, might be bottled, etc. So are adjectives that modify nouns, or verbs that subcategorise nouns salient second-order properties to describe the selectional preferences of direct objects? Our second interest lies in the actual comparison of the models: How does a very simple distributional model compare to much more complex approaches, especially with respect to model (iii) that explicitly incorporates selectional preferences? And which representation of selectional preferences is more appropriate, using (i) second-order properties, (ii) an implicit generalisation of nouns (by clusters), or (iii) an explicit generalisation of nouns by WordNet classes within clusters? All experiments in this paper were carried out for German, but can be transferred to other languages, given that sufficient corpus data is available to extract predicate–complement pairs, plus assuming a WordNet for (iii).

In the remainder of the paper, we describe the three selectional preference models and the respective experiments (Section 2), the evaluation (Section 3), and the results (Section 4).

2. Selectional Preference Models

Existing approaches to the automatic induction of selectional preferences fall into three categories. The majority of the approaches models selectional preferences by exploiting the hypernym/hyponym hierarchy in WordNet (Fellbaum, 1998). Relying on corpus-based *predicate–relation–noun* frequencies, they aim to find the optimal generalisation of the nouns as selectional preference characterisation with respect to the predicate and the predicate–noun relation. As the result, the selectional restrictions are expressed by WordNet classes, or sets over WordNet classes (most commonly a disjunctive set of classes represented by a *cut* through the hierarchy). Referring to the above example, having seen *coffee, tea, beer* and *wine* as direct objects of the verb *drink*, the hypernym *beverage* generalises over the seen nouns and thus represents a suitable WordNet label for the verb–object selectional preference. Approaches that fall into this category are Resnik (1997), Li and Abe (1998), Abney and Light (1999), Ciaramita and Johnson (2000), and Clark and Weir (2002).

An alternative to WordNet-based models are cluster-based models such as Pereira et al. (1993) and Rooth et al. (1999). Also relying on corpus-based *predicate–relation–noun* frequencies, cluster-based approaches represent selectional preferences by noun clusters that generalise over the seen nouns, without specific generalisation labels other than the cluster numbers. Two of our models fall into this category, with the more complex one refining the selectional preference description by WordNet categories.

Last but not least, Erk (2007) suggested a distributional, similarity-based model for selectional preferences that uses the corpus-based input data to first define a selectional preference representation, and then use vector-based similarity metrics to determine selectional preference scores for unseen nouns. Our second-order co-occurrence model is an instance of a distributional model, in many respects similar to Erk's model.

While WordNet-based approaches are attractive models of selectional preferences in that they explicitly provide preference categories, cluster-based and similarity-based approaches are attractive in that they are independent of such a manual resource which is not available for all languages and is costly to build.

We present and compare three approaches to selectional preference induction: (i) an intuitive distributional approach that uses second-order co-occurrence of predicates and complement properties; (ii) an EM-based clustering approach that models the strengths of predicate–noun relationships by latent semantic clusters (Rooth et al., 1999); and (iii) an extension of the latent semantic clusters by incorporating the MDL principle into the EM training, thus explicitly modelling the predicate–noun selectional preferences by WordNet classes (Schulte im Walde et al., 2008). The three models have been implemented for German, but can be transferred to other languages, given that sufficient corpus data is available to extract predicate–complement pairs, plus assuming a WordNet for (iii). The models are described in some detail in the following subsections.

2.1. A second-order distributional model

According to the distributional hypothesis, the sum of contexts of a linguistic unit is a crucial indicator of the meaning of the linguistic unit (Firth, 1957; Harris, 1968). In this vein, we define a distributional approach to selectional preference induction that is both intuitive and cheap. The underlying idea is that selectional preferences of a predicate's complement are defined by the properties of the complement realisations. For example, a typical direct object of the verb *drink* is usually fluid, might be hot or cold, can be bought, might be bottled, etc. So –referring to this example– are adjectives that modify nouns, or verbs that subcategorise nouns salient properties to describe the selectional preferences of direct objects? The general question we ask is: what characterises the realisations of selectional preferences? We thus suggest a second-order co-occurrence model for selectional preferences: a predicate's restrictions to the semantic realisation of its complements are expressed through the properties of the complements.

The basis of the distributional approach are corpus-based co-occurrences of triples $\langle predicate, relation, complement\rangle$, i.e., joint frequencies of predicate–complement pairs with respect to a specific functional relation. Being of second-order co-occurrence, the model combines two types of co-occurrences: (1) Corpus-based joint frequencies $freq(p, r1, n)$ of predicates p and nouns n with respect to some functional relationship $r1$. These co-occurrences refer to the functional relationships whose selectional preferences we address. We concentrate on German verb–noun relationships $r1$, namely subjects, direct object, and pp objects. This choice was motivated by the language we work on (German), plus the available data for evaluation, cf. Section 3. The approach can easily be expanded to other predicates and relations, but in order to incorporate the latent semantic class model with WordNet generalisations, the complement choice is necessarily nouns. (2) Corpus-based joint frequencies $freq(n, r2, prop)$ of nouns n and noun properties $prop$ with respect to some functional relationship $r2$. These co-occurrences refer to the properties of the selectional preferences we address. We concentrate on modifying adjectives, subcategorising verbs (for direct object and pp object), and subcategorising prepositions, because these properties were expected to shed light on complementary semantic properties of the nouns (and thus the selectional preference descriptions). We tested the properties by themselves, and also in combinations. The set of properties can easily be enlarged, as the experiments will demonstrate. The joint frequencies were estimated on approx. 560 million words from the German web corpus *deWaC* (Baroni and Kilgarriff, 2006), after the corpus was preprocessed by the Tree Tagger (Schmid, 1994) and by a dependency parser (Schiehlen, 2003).

The distributional model comprises two parts: (1) the selectional preference description with respect to a specific verb–noun relationship, i.e., the second-order properties of the relationships, and (2) the selectional preference fit of a specific noun with respect to the verb–noun relationship. Part (1) is a simple scoring that combines the two types of corpus-based joint frequencies, $freq(p, r1, n)$ and

$freq(n, r2, prop)$, cf. Equations (3) to (6). The second-order selectional preference of the verb–noun relationship $r1$ is represented by the joint noun–property corpus frequencies across the nominal complements, cf. Equation (3) for the most basic version. I.e., the feature vector of the predicate is a union of the properties of the nouns. For example, if the predicate is the verb *drink*, the verb–noun relation is a direct object, and the property is adjectives that modify nouns; then, the verb's selectional restrictions are defined by an adjective feature vector, where the set of adjectives is the union of the adjectives modifying the nouns subcategorised by *drink*. The feature values $score_1(drink, dir\_obj, adj)$ thus rely on the frequencies of all nouns that appeared as direct objects of *drink*, $freq(drink, dir\_obj, n)$, and on the frequencies of the adjectives those nouns appeared with (not necessarily in the same context with the verb, $freq(n, n\_mod, adj)$). For example, if *coffee* appeared 50 times as direct object of *drink*, and *tea* appeared 5 times, and if *coffee* was modified by the adjective *hot* 100 times and by *fluid* 30 times, and if *tea* was modified by *hot* 60 times and by *fluid* 15 times, then $score_1(drink, dir\_obj, hot) = 50*100 + 5*60 = 5,300$, and $score_1(drink, dir\_obj, fluid) = 50*30 + 5*15 = 1,575$. The scoring provided in Equation (3) is the most simplest, using raw corpus frequencies. Alternative versions rely on log-transformed frequencies (Equation (4)), probabilities (Equation (5)), and tf-idf values (Equation (6), where $tf\_idf(triple) = tf(triple) * idf(triple)$, with $tf(triple) = prob(triple)$, and $idf(triple) = log\frac{|p,r1|}{|p,r1,n|}$ for $r1$, and $idf(triple) = log\frac{|n,r2|}{|n,r2,prop|}$ for $r2$, i.e., determining the "inverse document frequency" of nouns ($r1$) or properties ($r2$) by incorporating the number of different predicates ($r1$) or nouns ($r2$) a noun ($r1$) or property ($r2$) occurred with).

Tables 1 to 3 present examples of second-order properties, for the direct objects of the verb *backen* 'bake' with adjective properties, *anbraten* 'fry' with verb properties, and *abschmecken* 'taste' with preposition properties, respectively. The tables list the eight most probable properties and also the eight most probable nominal realisations, according to some of the most successful distributional models. The information is rather for intuitive purposes; therefore, the system scores are omitted. The prepositions in Table 3 are the most difficult to grasp intuitively, but at the same time the most successful system features, cf. Section 3.

As mentioned before, the resulting selectional preference descriptions are predicate vectors over complement properties. In part (2), the natural fit of a specific noun can then be specified by standard distributional similarity measures, comparing a specific noun's contribution to the overall preference: In order to determine the selectional preference for a specific (seen or unseen) noun, we calculate the vector-based similarity between the predicate's preference vector and the specific noun's vector. The measures to calculate the similarities and thus the natural fit of a specific noun to a selectional preference description can be varied. We experimented with four standard measures that were expected to provide different perspectives on the selectional preference fit, due to their mathematical nature: the cosine of the vector's angle (a standard measure in linear algebra), the skew divergence, an information-theoretic measure and variant of the Kullback-Leibler divergence (Lee, 2001), Kendall's τ, a measure for rank correlation (Hatzivassiloglou and McKeown, 1993), and jaccard, a binary distance measure (Manning and Schütze, 1999).
Our method is similar to Erk's approach who also used complements' corpus-based properties to describe selectional preferences. We addressed the task from a different direction, though, and the result is a simplified version of her approach. The models with a single nominal property are specific cases of her model, and only the models with combined nominal properties come close to a general distributional description. Furthermore, our goal is different from hers in that we were interested in the contributions of the various properties, in addition to determining the natural fit of nouns to selectional preferences.

Properties: adjectives		Example realisations	
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'

Table 1: Direct objects of *backen* 'bake'.

Properties: verbs$_{NPacc}$		Example realisations	
schälen	'peel'	Champignon	'mushroom'
schneiden	'cut'	Zwiebel	'onion'
essen	'eat'	Kartoffel	'potato'
zugeben	'add'	Gemüse	'vegetable'
anschwitzen	'sweat'	Knoblauch	'garlic'
pellen	'peel'	Hackfleisch	'minced meat'
riechen	'smell'	Roulade	'roulade'
waschen	'clean'	Keule	'haunch'

Table 2: Direct objects of *anbraten* 'fry'.

Properties: prepositions		Example realisations	
mit	'with'	Soße	'sauce'
in	'in'	Salat	'salad'
für	'for'	Brühe	'stock'
zu	'for'	Gemüse	'vegetables'
von	'from'	Eintopf	'stew'
unter	'under'	Suppe	'soup'
auf	'on'	Püree	'puree'
als	'as'	Essen	'food'

Table 3: Direct objects of *abschmecken* 'taste'.

2.2. Latent Semantic Classes

The Latent Semantic Cluster *(LSC)* approach is an instance of the Expectation-Maximisation (EM) algorithm (Baum, 1972) for unsupervised training on unannotated data. It has previously been applied to model the selectional dependencies between two sets of words participating in a grammatical relationship (Rooth et al., 1999). The cluster analyses

$$score_1(p, r1, prop) = \sum_{n \in (p,r1)} freq(p, r1, n) \ * \ freq(n, r2, prop) \tag{3}$$

$$score_2(p, r1, prop) = \sum_{n \in (p,r1)} log(freq(p, r1, n)) \ * \ log(freq(n, r2, prop)) \tag{4}$$

$$score_3(p, r1, prop) = \sum_{n \in (p,r1)} prob(p, r1, n) \ * \ prob(n, r2, prop) \tag{5}$$

$$score_4(p, r1, prop) = \sum_{n \in (p,r1)} tf\_idf(p, r1, n) \ * \ tf\_idf(n, r2, prop) \tag{6}$$

Figure 1: Second-order selectional preference description.

define two-dimensional soft clusters which are able to generalise over hidden data. LSC training learns three probability distributions, one for the probabilities of the clusters, and a joint probability distribution for each lexical class participating in the grammatical relationship, (e.g., predicates and subcategorised nouns) with respect to cluster membership, thus the two dimensions. LSC was chosen because the clusters can be considered as generalisations over the members of the two inter-dependent dimensions. The LSC approach therefore fits selectional preferences, by generalising over seen and unseen lexical items.

Concerning our task, the semantically smoothed probability of a predicate–noun pair (p, n) with respect to some functional relation is defined by Equation (7). Our experiments with LSC rely on the same corpus data as the distributional model; we used the same verb–subject, verb–direct-object, and verb–pp-object data. We trained three LSC models, one for each functional relation, and a forth model that contained all relations, using a relation marker at the verb (e.g., replacing the verb *backen* with *backen-subj*) to distinguish between the relations. The resulting analyses were used to calculate the probabilities of verb–noun pairs as the natural fit of the nouns to the selectional preferences the clusters incorporate. The training parameters were varied, producing cluster analyses with 20, 50, 100, 200, and 500 clusters, over 50 and 100 iterations.

$$\begin{aligned} prob(p, n) &= \sum_{c \in cluster} prob(c, p, n) \\ &= \sum_{c \in cluster} prob(c)\ prob(p, c)\ prob(n, c) \qquad (7) \end{aligned}$$

Table 4 presents an LSC example of a cluster containing verbs and their direct objects, as taken from a 100-cluster analysis. The left-hand column contains the most probable predicates within this cluster; the right-hand column contains the most probable nouns within this cluster. The nouns are assumed to represent the selectional preferences of the direct objects of the verb dimension.

2.3. Latent Semantic Classes integrating Selectional Preferences

While the original LSC approach models selectional preferences only implicitly, by assigning semantically similar words to common classes, an extension of the LSC approach incorporates explicit selectional preferences. The PAC model (Schulte im Walde et al., 2008) provides a combination of the EM algorithm and the Minimum Description Length (MDL) principle (Rissanen, 1978), and thus refines the second, nominal dimension by explicit generalisations based on WordNet and the MDL principle. The model compromises for incorporating a manual resource (differently to the two preceding models).

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'

Table 4: Example LSC cluster.

The PAC model is estimated through the joint probability of a predicate p, a subcategorisation frame type f, and the complement realisations $n_1, ..., n_k$, cf. Equation (8). In addition to the LSC parameters in Equation (7), $prob(r|c, f, i)$ is the probability that the ith complement of frame f in class c is realised by WordNet (wn) concept r, and $prob(n|r)$ is the probability that the WordNet concept r is realised by complement head n. See Schulte im Walde et al. (2008) for detailed explanations of the model.

$$\begin{aligned} prob(p, f, n_1, ..., n_k) = \sum_c prob(p)\ prob(p, c)\ prob(f, c) \\ \prod_{i=1}^{k} \sum_{r \in wn} prob(r|c, f, i)\ prob(n_i|r) \qquad (8) \end{aligned}$$

Our experiments with PAC rely on the same corpus data as the other two models; we used the same verb–subject, verb–direct-object, and verb–pp-object data. We trained four PAC models, one for each functional relation, and one for all data in one model (as PAC incorporates frame types and thus distinguishes between functional relations). As for LSC, we used the resulting analyses to calculate the probabilities of verb–noun pairs as the natural fit of the nouns to the selectional preferences the clusters incorporate. The training parameters were varied as for LSC, producing cluster analyses with 20, 50, 100, 200, and 500 clusters, over 50 and 100 iterations.

Table 5 presents a PAC example of a cluster containing verbs and their direct objects, as taken from a 20-cluster analysis. The left-hand column contains the most probable predicates within this cluster; the right-hand column contains a selection of the most probably WordNet classes from different hierarchical levels. The extensive WordNet hierarchical structure that is part of the second cluster dimension is omitted for space and clarity reasons.

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'

Table 5: Example PAC cluster.

3. Evaluation

The three selectional preference approaches were evaluated against human judgements on German verb–noun pairs. The judgements had been collected by Brockmann and Lapata (2003)[1] whose study compared the WordNet-based selectional preference approaches by Resnik (1997), Li and Abe (1998), and Clark and Weir (2002), plus two distributional models relying on co-occurrence frequency and conditional probability. The human data contains 90 verb–noun pairs, with 30 pairs each for subjects, direct objects and pp objects, and each of the 30 pairs contains 10 different verbs with 3 different nouns. Verbs and nouns were chosen randomly; furthermore, the noun choice was done in accordance with three frequency bands of the verb–relation–noun triples. The participants in the study were asked to provide selectional preference scores for the 90 verb–noun pairs; the scores were then normalised to a common scale, and transformed by taking the decadic logarithm $log10$.

Brockmann and Lapata used the human judgements to compare the above-mentioned selectional preference approaches. Each model provided selectional preference scores for the 90 verb–noun pairs, the system scores were transformed by taking the decadic logarithm and then correlated against the human judgement scores by linear regression. Brockmann and Lapata found that all five models were significantly correlated with the human judgements, but inter-subject agreement was consistently higher than the correlations. Furthermore, no model performed best; different methods were suited for different functional relations. A combination of the models by multiple linear regression outperformed the individual models.

By using the same gold standard data and the same computations as Brockmann and Lapata, we can compare not only our models against each other, but also compare our results to theirs. We therefore calculated system scores for the 90 verb–noun pairs (which had previously been removed from the training data) with respect to our three approaches. As Brockmann and Lapata, we also transformed the system scores by taking the decadic logarithm, before performing the linear regression with their $log10$ human judgements. In comparison, however, we also correlated the original system scores against the human judgements back-transformed by the $log10$ reverse function. The latter procedure seemed reasonable, as we did not agree with a general $log10$ transformation without knowledge about the underlying data distribution.[2]

Furthermore, we added a second type of evaluation, and compared the approaches using the Spearman rank-order correlation coefficient (henceforth: *ranking*). This correlation is a non-parametric statistical test that measures the association between two variables that are ranked in two ordered series. The ranking seemed reasonable, as it looked at the evaluation from a different perspective, assessing how well the systems can distinguish fine-grained rank-order differences across the gold standard pairs.

The baselines of the experiments were calculated by correlating the joint corpus-based predicate–noun frequencies of the subjects, direct objects and pp-objects with the human judgements (also by linear regression, and by ranking). The upper bound of the approaches is referred to as the inter-subject agreement (*isa*) on the selectional preference judgements, as calculated by Brockmann and Lapata, henceforth *BL*.

4. Results

Tables 6 to 8 present an extract of the results of the distributional approach, and the LSC and PAC experiments. All of the results refer to the evaluation by linear regression, where the system scores were not transformed by the decadic logarithm (and, accordingly, correlated with the back-transformed judgements); the results with respect to ranking will be described below. In each table, our best results per column are printed in bold font. The overall best results per relation are in addition printed in blue, and marked by the significance levels $*p \leq .05$, $**p \leq .01$, and $***p \leq .001$, if applicable. The BL results in Table 6 refer to the best results achieved in the Brockmann/Lapata comparison, and provide the respective system in brackets. The baseline and upper bound values are only listed in Table 6 but refer to all linear regression experiments in the three tables. The frequency baseline correlated the joint predicate–noun frequencies against the back-transformed human judgement scores; the $log10(f)$ baseline correlated the frequencies transformed by the decadic logarithm against the BL judgement scores; the BL baseline is taken from their paper and refers to $log10$-transformed frequencies correlated against the $log10$ BL judgement scores.

The distributional results list the *cosine* scores as the measure of selectional preference fit, because it provided the overall best scores. The rows refer to the second-order properties, and the columns to the second-order selectional preference description, cf. Figure 1. As mentioned before,

[1] Thanks to Carsten Brockmann for providing the judgement scores to us.

[2] If the data is normally distributed without transformation, then it needs no transformation to go into a linear regression; if the data is normally distributed after the transformation, then a transformation is reasonable. In any case, the transformation will change the linear regression, as a logarithm imposes a shape on the scores that influences the linearity. The degree of the change depends on the scale of the scores.

we used adjectives, verbs, and prepositions as second-order properties; furthermore, we enlarged and unified the property sets: $v + vp$ adds verb–preposition pairs (subcategorising for the respective nouns), $v + vp + adj$ adds adjectives to this set, and $v + vp + adj + prep$ further adds the prepositions. A number of things are striking in Table 6: (1) Not only with respect to the cosine results but also in more general, the prepositions by themselves, or the union of second-order properties $v + vp + adj + prep$ are in many cases the most successful properties. On the one hand, we can conclude that prepositions are a powerful indicator of selectional preference properties; on the other hand, our largest set of properties comes close to a general distributional description without strong restrictions on the selection of properties, in the vein of Erk (2007), and the question is whether a less careful choice than the properties we provided would be as successful or even more successful. One could try, e.g., window information as a very crude choice. (2) The best results vary quite strongly with respect to the functional relation. Direct objects are modelled best, subjects are modelled worst. (3) Not only with respect to this table but in more general, the probability and tf-idf scores tended to outperform the frequency- and log(f)-based scores. Note that the best overall result in Table 6 is achieved by frequency, though. (4) Quite striking in the table are the large values of the baseline using the log-transformed predicate–noun frequencies, $.652/.559/.565/.574$ for *subj, dir obj, pp-obj, overall* with an upper bound for isa of $.790/.810/.820/.810$, cf. BL. The baseline is so high that it beats some of the best system (and system combination results) in BL ($.408/.611/.597/.400$), and some of our results ($.494/.713/.602/.517$). Furthermore, our baseline is much higher than BL's baseline (calculated identically, as far as we know): $.386/.360/.168/.301$. The only explanation for this is that the results differ because of the different underlying corpora, 560 million words of the *deWaC* vs. 179 million words of the German *Süddeutsche Zeitung* newspaper corpus. To be sure whether the size or the domain differences are the crucial ingredients, one would have to replicate our experiments on a portion of the *deWaC* comparable to BL's portion. We hypothesise, though, that the difference is rather due to the corpus domains, which should arguably provide different frequency counts for verb–noun pairs such as *reward a child*, or *clean the pavement*, whose German translations are among the gold standard pairs. The same reason applies to the fact that our results are all above those of BL's comparison. One would have to re-run the various systems on our data, in order to have a fair comparison. (5) As mentioned above, the cosine measure was the most useful for our purposes. The skew divergence and the jaccard binary measures were always clearly below the cosine-based scores. Only the results with Kendall's τ were in some cases similar to the cosine results; for subjects, τ could even beat the cosine, with a correlation of **.532, using $v + vp$. (6) The cosine results when correlating the $log10$ system scores against the $log10$ judgements were quite below the ones in Table 6, confirming our intuition that a general $log10$ transformation and linear regression do not necessarily fit.

Table 7 presents the results for the LSC experiments. The first column for each relation refers to a linear regression of the probabilities of the verb–noun pairs and the back-transformed judgements; the second column refers to the correlation between the $log10$-transformed probabilities against the $log10$-transformed judgements. Although the best LSC correlations are also significant, all of them are below those of the simpler distributional model. Interestingly, though, the correlations based on the $log10$-transformed scores were in most cases above those without transformation. Concerning the number of clusters and training iterations, there is no clear tendency towards an optimal settings. The number of training iterations did not consistently improve the results, and neither did a smaller or larger number of clusters. When training used all relation information at the same time ($all\ func$), relying on relation markers at the verb (cf. Section 2.2), LSC performed better than after individual training on the relation data.

Table 8 presents the results for the PAC experiments. Again, the first column for each relation refers to a linear correlation between the probabilities of the verb–noun pairs against the back-transformed judgements, and the second column refers to the correlation between the $log10$-transformed probabilities against the $log10$-transformed judgements. The PAC results vary quite strongly with respect to the verb–noun relationship: For subjects, the correlation of .507 even beats the distributional model; for all other relations, the results are worse than in the simple model, for pp objects they can even be considered quite poor. When training all relations at the same time, the best PAC correlation is similar (and slightly above) the best LSC score. As for LSC, the correlations based on the $log10$-transformed scores were in most cases above those without transformation. Also similar is the fact that the number of clusters and iterations does not have a clear tendency towards selectional preference prediction. It seems to be the case, though, that smaller numbers of clusters are better.

The evaluation by the Spearman rank-order correlation coefficient provides similar results as the linear regression evaluation. The tables are omitted for space reasons. Again, the distributional model using the cosine is identified as the most successful selectional preference approach. In comparison to a baseline of $.903/.863/.928/.884$ (where the ranking according to the joint predicate–noun frequencies is correlated against the gold standard ranking), the cosine reaches best correlations of $.880/.938/.947/.879$ for subject, direct object, pp object and across relation selectional preferences. It thus beats the baseline in all cases but the subject. In comparison, the distributional model using the skew divergence achieves only correlations of $.758/.739/.773/.772$. LSC and PAC reach correlations of $.872/.872/.896/.873$ and $.882/.877./.795/.850$, respectively. The results of the cluster approaches are therefore in most cases below those of the distributional model, with PAC reaching slightly higher scores than LSC.

The properties of the most successful distributional models, and the number of clusters and training iterations of the most successful cluster models are not the same as those in the linear regression evaluation. Therefore, we cannot conclude about any general optimal settings of the models.

5. Conclusions

This paper presented three computational approaches to selectional preferences, an intuitive second-order co-occurrence model, and two latent semantic cluster models. Quantitative and qualitative analyses of the approaches demonstrated that the simple distributional model outperforms the more complex cluster-based models in most cases. Prepositions played a dominant role among the second-order properties, on an individual basis and in combinations with other properties. Even the best models, though, did not always beat the powerful frequency baseline. Comparing the two cluster-based models, an explicit generalisation of nouns by WordNet classes within clusters provides little help.

6. References

Steven Abney and Marc Light. 1999. 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.

Marco Baroni and Adam Kilgarriff. 2006. 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.

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

Carsten Brockmann and Mirella Lapata. 2003. 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.

Noam Chomsky. 1957. *Syntactic Structures*.

Massimiliano Ciaramita and Mark Johnson. 2000. 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.

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

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

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

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

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

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. 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.

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

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

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

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

Diana McCarthy and John Carroll. 2003. Disambiguating Nouns, Verbs and Adjectives using Automatically Acquired Selectional Preferences. *Computational Linguistics*, 29(4):639–654.

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

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

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

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

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

Michael Schiehlen. 2003. 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.

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

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. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

Benat Zapirain, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over Lexical Features: Selectional Preferences for Semantic Role Classification. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 73–76, Suntec, Singapore.

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
	f	log(f)	f	log(f)	f	log(f)	f	log(f)
adj	.416	.373	.417	.261	.113	.220	.244	.156
verb	.456	.412	.271	.222	.176	.278	.201	.178
prep	.461	.345	.681	.263	.318	.393	.391	.272
v+vp	**.468**	.425	.345	.295	.344	.369	.295	.235
v+vp+adj	.420	.411	.388	.287	.235	.345	.285	.222
v+vp+adj+prep	.459	**.465**	***.713	**.328**	**.380**	**.476**	**.422**	**.359**
	prob	tf-idf	prob	tf-idf	prob	tf-idf	prob	tf-idf
adj	.430	.420	.352	.301	.339	.373	.309	.311
verb	**.494	.406	.285	.325	.242	.487	.273	.386
prep	.443	**.487**	.625	**.680**	.554	***.602	.481	**.516**
v+vp	.479	.387	.333	.290	.476	.564	.345	.401
v+vp+adj	.435	.383	.402	.307	.401	.478	.345	.364
v+vp+adj+prep	.465	.437	**.705**	.428	**.599**	.581	***.517	.455
BL	*.408 (Resnik)		***.611 (Clark/Weir)		***.597 (Clark/Weir)		***.400 (comb)	
baselines & upper bound								
baseline: f	.274		.343		.384		.313	
baseline: log10(f)	.652		.559		.565		.574	
baseline: BL	.386		.360		.168		.301	
isa	.790		.810		.820		.810	

Table 6: Distributional results.

	SUBJ		DIR-OBJ		PP-OBJ		*all*		*all-func*	
50 training iterations										
20	.253	*.450	.016	.282	.181	.295	.033	.338	.118	.383
50	**.332**	.382	.074	.424	.117	.061	.172	.240	.185	***.453
100	.202	.222	**.313**	.483	.234	.141	.203	.235	.081	.379
200	.310	.308	.285	.469	.243	.189	**.216**	.275	**.226**	.332
500	.261	.210	.258	.393	.318	.189	.157	.242	.155	.339
100 training iterations										
20	.249	.165	.061	.386	.149	.352	.064	.266	.096	.362
50	.320	.317	.184	.420	.069	.042	.194	.241	.181	.439
100	.199	.306	.300	***.569	.232	.276	.198	.264	.082	.245
200	.286	.386	.300	.505	**.366**	**.562	.209	***.407	.220	.363
500	.302	.389	.285	.315	.325	.396	.185	.315	.146	.244

Table 7: LSC results.

	SUBJ		DIR-OBJ		PP-OBJ		*all*	
50 training iterations								
20	**.189**	.503	.209	**.509	**.062**	**.045**	.121	.377
50	.094	.208	**.258**	.360	**.062**	**.045**	.070	.444
100	.094	.208	.041	.074	**.062**	**.045**	.134	.400
200	.094	.208	.041	.074	**.062**	**.045**	.060	.367
500	.094	.208	.041	.074	**.062**	**.045**	.060	.367
100 training iterations								
20	.185	**.507	.229	.495	**.062**	**.045**	.114	.429
50	.094	.208	.222	.478	**.062**	**.045**	.107	***.465
100	.094	.208	.041	.074	**.062**	**.045**	**.141**	.385
200	.094	.208	.041	.074	**.062**	**.045**	.081	.442
500	.094	.208	.041	.074	**.062**	**.045**	.060	.367

Table 8: PAC results.