Multi-way Tensor Factorization for Unsupervised Lexical Acquisition

Tim Van de Cruys[1,3] Laura Rimell[1,2]
Thierry Poibeau[1,4] Anna Korhonen[1,2]
(1) DTAL, University of Cambridge, UK
(2) NLIP, University of Cambridge, UK
(3) IRIT, UMR 5505, CNRS, Toulouse, France
(4) LaTTiCe, UMR 8094, CNRS & ENS, Paris, France
tim.vandecruys@irit.fr, laura.rimell@cl.cam.ac.uk,
thierry.poibeau@ens.fr, anna.korhonen@cl.cam.ac.uk

ABSTRACT

This paper introduces a novel method for joint unsupervised aquisition of verb subcategorization frame (SCF) and selectional preference (SP) information. Treating SCF and SP induction as a multi-way co-occurrence problem, we use multi-way tensor factorization to cluster frequent verbs from a large corpus according to their syntactic and semantic behaviour. The method extends previous tensor factorization approaches by predicting whether a syntactic argument is likely to occur with a verb lemma (SCF) as well as which lexical items are likely to occur in the argument slot (SP), and integrates a variety of lexical and syntactic features, including co-occurrence information on grammatical relations not explicitly represented in the SCFS. The SCF lexicon that emerges from the clusters achieves an F-score of 68.7 against a gold standard, while the SP model achieves an accuracy of 77.8 in a novel evaluation that considers all of a verb's arguments simultaneously.

TITLE AND ABSTRACT IN FRENCH

Factorisation de tenseurs à plusieurs dimensions pour l'acquisition lexicale non supervisée

Cet article présente une méthode originale pour l'acquisition simultanée de cadres de sous-catégorisation (*subcategorization frames*) et de restrictions de sélection (*selectional preferences*) appliquée au lexique verbal. L'induction simultanée de ces deux types d'information est vue comme un problème de cooccurrence à plusieurs dimensions. On introduit donc une méthode de factorisation de tenseurs, afin de classer les verbes fréquents d'un grand corpus suivant leur comportement syntaxique. L'approche est fondée sur un ensemble de traits de nature syntaxique et lexicale, y compris des informations de cooccurrence au sein des relations grammaticales qui ne sont pas explicitement représentées dans les schémas de sous-catégorisation. Le dictionnaire de sous-catégorisation produit par la méthode de classification obtient une F-mesure de 68,7 lors de l'évaluation face à un dictionnaire de référence tandis que les restrictions de sélection ont une exactitude (*accuracy*) de 77,8 en tenant compte de tous les arguments simultanément.

KEYWORDS: subcategorization frames, selectional preferences, lexical acquisition, tensor factorization, unsupervised machine learning.

KEYWORDS IN FRENCH: cadre de sous-catégorisation, restriction de sélection, acquisition lexicale, factorisation de tenseurs, apprentissage non supervisé.

Proceedings of COLING 2012: Technical Papers, pages 2703–2720,
COLING 2012, Mumbai, December 2012.

1 Introduction

Verb subcategorization lexicons and selectional preference models capture two related aspects of verbal predicate-argument structure, with subcategorization describing the syntactic arguments taken by a verb, and selectional preferences describing the semantic preferences verbs have for their arguments. Each type of information can support NLP tasks requiring information about predicate-argument structure. For example, subcategorization has proved useful for parsing (Carroll and Fang, 2004; Arun and Keller, 2005; Cholakov and van Noord, 2010), semantic role labeling (Bharati et al., 2005; Moschitti and Basili, 2005), verb clustering, (Schulte im Walde, 2006; Sun and Korhonen, 2011) and machine translation (Han et al., 2000; Hajič et al., 2002), while selectional preferences have benefited parsing (Zhou et al., 2011), semantic role labeling (Gildea and Jurafsky, 2002; Zapirain et al., 2009), and word sense disambiguation (Resnik, 1997; Thater et al., 2010; Seaghdha and Korhonen, 2011).

Verb subcategorization frame (SCF) induction involves identifying the arguments of a verb lemma in a corpus, and generalizing about the *frames* taken by the verb, where each frame includes a number of arguments and their syntactic types. Consider e.g. sentence (1), where the verb *show* takes the frame SUBJ-DOBJ-CCOMP (subject, direct object, and clausal complement).

(1) $[\text{Our October review}]_{\text{SUBJ}}$ comprehensively $[\text{shows}]_{\text{VERB}}$ $[\text{you}]_{\text{DOBJ}}$ $[\text{what's in store in next month's magazine}]_{\text{CCOMP}}$.

Predicting the set of SCFs for a verb can be viewed as a multi-way co-occurrence problem of a verb and its different arguments. One of the main challenges is distinguishing arguments from adjuncts (e.g. temporal, locative, or manner modifiers). Most SCF induction work to date considers only the co-occurrences of verb lemmas with different grammatical relation types (subject, object, prepositional phrase, etc.). Taking SCF acquisition to the next level requires consideration of the lexical fillers of potential argument slots for more accurate argument-adjunct discrimination.

Selectional preference (SP) induction involves predicting the likelihood of a given lexical item occurring in an argument slot, and generalizing about the lexical classes which occur in the slot, which may be dependent on the SCF. In sentence (2), for example, the verb *show* takes the frame SUBJ-DOBJ, and the direct object of *show* in this frame is likely to be inanimate.

(2) $[\text{Stalin}]_{\text{SUBJ}}$, who must have been well informed through his network of spies, $[\text{showed}]_{\text{VERB}}$ $[\text{no emotion}]_{\text{DOBJ}}$.

Most SP induction work to date has focused on discovering lexical preferences within the direct object slot alone, or at most three-way co-occurrences between verb, subject, and direct object, and has not considered the full range of potential argument slots for which verbs subcategorize, thus losing some of the contextual information which may be helpful in learning SPs. Moreover, the ability of SP acquisition methods to model the full range of verbal arguments, including e.g. clausal complements, has not been tested.

As the two types of lexical information – SCFs and SPs – are closely interlinked and can complement each other, it would make sense to acquire them jointly. However, to the best of our knowledge, no previous work has developed a model for their joint acquisition.

Unsupervised machine learning is attractive for lexical acquisition because it works where little labeled data is available, and ports easily between tasks and languages. Increasingly sophisticated techniques have been applied to SP induction (Rooth et al., 1999; Van de Cruys, 2009; Ó Séaghdha, 2010; Ritter and Etzioni, 2010; Reisinger and Mooney, 2011) while work

on unsupervised SCF acquisition has been limited (Carroll and Rooth, 1996). In this paper we present a largely unsupervised method for the joint acquisition of SCFs and SPs, adapting a method that has been successfully used for SP induction (Van de Cruys, 2009) so that it learns *whether* a verb subcategorizes for a particular argument slot together with *which* lexical items occur in the slot.

Our method uses a co-occurrence model augmented with a factorization algorithm to cluster verbs from a large corpus. Specifically, we use non-negative tensor factorization (NTF) (Shashua and Hazan, 2005), a generalization of matrix factorization that enables us to capture latent structure from multi-way co-occurrence frequencies. The factors that emerge represent clusters of verbs that share similar syntactic and semantic behaviour. To evaluate the performance on SCF acquisition, we identify the syntactic behaviour of each cluster. The SCF lexicon that emerges from the clusters achieves a promising F-score of 68.7 against a gold standard. We further introduce a novel SP evaluation in which we investigate the model's ability to induce preferences for the co-occurrence of a particular verb lemma and all of its arguments *at the same time*. The model achieves a high accuracy of 77.8 on this new evaluation. We also perform a qualitative evaluation which shows that the joint model is capable of learning rich lexical information about both syntactic and semantic aspects of verb behaviour in data.

2 Related Work

Recent SCF acquisition approaches use the output of an unlexicalized parser to generate SCF hypotheses, followed by statistical filtering and/or smoothing to remove noise. Briscoe and Carroll (1997); Korhonen (2002); Preiss et al. (2007) use handcrafted rules to match parser output to a pre-defined set of SCFs, achieving an F-measure of about 70 against a manually annotated gold standard, while O'Donovan et al. (2005); Chesley and Salmon-Alt (2006); Ienco et al. (2008); Messiant (2008); Lenci et al. (2008); Altamirano and Alonso i Alemany (2010); Kawahara and Kurohashi (2010) induce the inventory of SCFs from parsed corpus data. Candidate frames are identified by grammatical relation (GR) co-occurrences, often aided by language-specific heuristics. Statistical filtering or empirically-tuned thresholds are used to select frames for the final lexicon. These 'inductive' approaches have achieved respectable accuracy (60-70 F-measure against a dictionary) and are more portable than earlier methods. However, their ability to improve in accuracy is limited by their inability to incorporate information beyond the GR co-occurrences and heuristics that identify candidate SCFs on a per-sentence basis. Such cues provide no capacity for learning further from the data, e.g. from the lexical content of verbal arguments or from other GRs which are not part of the SCF.

Unsupervised machine learning has been applied to tasks where portability is equally important (Blei et al., 2003; Dinu and Lapata, 2010) but its application to SCF acquisition remains limited. Carroll and Rooth (1996) combined a head-lexicalized context-free grammar with an expectation-maximization (EM) algorithm to acquire an SCF lexicon. Dębowski (2009) used a filtering method based on the point-wise co-occurrence of arguments in parsed data to acquire a Polish SCF lexicon, but this method does not take the semantics of the verb's arguments into account. Lippincott et al. (2012) developed a graphical model for inducing verb frames in corpus data. The model identifies argument types of verbs but not *sets* of SCFs taken by a verb, as full scale SCF systems do.

Recent SP acquisition approaches use latent semantic information to model SPs, making use of probabilistic models, such as latent Dirichlet allocation (LDA) (Ó Séaghdha, 2010; Ritter and Etzioni, 2010; Reisinger and Mooney, 2011), or non-negative tensor factorization (NTF)

(Van de Cruys, 2009). Other approaches solely make use of distributional similarity methods (Bhagat et al., 2007; Basili et al., 2007; Erk, 2007). All approaches model two-way verb-argument co-occurrences, with the exception of Van de Cruys (2009) which models three-way verb-subject-object co-occurrences.

To our knowledge, no previous method has learned SCFs and SPs jointly. Scheible (2010) used SCFs as features in a Predicate-Argument Clustering (Schulte im Walde et al., 2008) approach to SP acquisition, but did not evaluate the resulting clusters for SCFs and found that the SP method did not outperform previous methods. Abend et al. (2009) used co-occurrence measures to perform unsupervised argument-adjunct discrimination for PPs, but not full SCFs.

Our method makes use of non-negative tensor factorization (NTF) (Shashua and Hazan, 2005). Tensor factorization is the multilinear generalization of matrix factorization. It has been extensively studied in the field of statistics (Kolda and Bader, 2009), and has yielded promising results on SP acquisition (Van de Cruys, 2009). We introduce a novel way of considering SCFs with an arbitrary number of arguments, and SPs as multi-way co-occurrences in the context of these larger SCFs. The resulting model provides an ideal framework for joint acquisition of SCF and SP information. The only form of supervision in the model is parameter estimation and choice of the best feature set via cross-validation.

3 Subcategorization Frame Inventory

To facilitate thorough qualitative evaluation (Section 5.6), we defined our SCFs in terms of syntactic slots, and in the form of common GRs. Finer-grained inventories including lexicalized elements and semantic interpretation were left for future work (see Section 7).

We use the GR types produced by the RASP parser (Briscoe and Carroll, 2002). Altogether we experimented with combinations of nine GR types out of the 13[1] which can be headed by verbs, selected on the basis of their frequency in the parsed BNC corpus and relevance for subcategorization. For this initial experiment, we focused on higher-frequency arguments since they will have the greatest impact on downstream applications.

Our first eight basic GR types are as follows. In subject position we included non-clausal subjects (SUBJ)[2], ignoring sentences with clausal subjects, which are much less frequent. Since objects are key arguments for subcategorization, we included all three object types – direct objects (DOBJ), second objects of ditransitive constructions (OBJ2), and prepositional arguments (IOBJ). Although OBJ2 is less frequent than other objects, it is important for identifying ditransitive frames. We included both types of clausal complements – XCOMP (infinitival/unsaturated) and CCOMP (finite/saturated) – and also PCOMP, which often signifies a *wh*-object of a preposition. We also included particles (PRT). Together, these eight GR types account for 62% of the GRs in the parsed BNC corpus. Using these GRs, there are 23 SCFs in our gold standard (see Section 5.1), of which the 15 with the highest type frequency are shown in Table 1.

Although modifiers are generally not included in SCFs (and are also excluded from our gold standard) we experimented with using them as features, to determine whether their distribution could help reach a better generalization. We focused on non-clausal modifiers (NCMOD). Counting them, the nine GR types account for 95% of the GRs in the BNC corpus.

[1] We count particles (here PRT) as a separate type, though RASP classifies them as a subtype of non-clausal modifiers.
[2] NCSUBJ in RASP.

Frame	Example sentence	Frame	Example sentence
SUBJ-DOBJ	Susan *found* the book.	SUBJ-XCOMP	Susan *wanted* to find the book.
SUBJ-DOBJ-IOBJ	Susan *put* the book on the table.	SUBJ-DOBJ-XCOMP	Susan *asked* Peter to attend.
SUBJ	Susan *knocked*.	SUBJ-DOBJ-IOBJ-PRT	Susan *filled* Peter in on the class.
SUBJ-IOBJ	Susan *appealed* to Peter.	SUBJ-CCOMP	Susan *believed* that Peter had found the book.
SUBJ-PRT	Susan *gave* up.	SUBJ-DOBJ-CCOMP	Susan *told* Betty that Peter had found the book.
SUBJ-DOBJ-PRT	Susan *picked* up the book.	SUBJ-DOBJ-OBJ2	Susan *gave* Betty a book.
SUBJ-PCOMP	Susan *thought* about whether she wanted to go.	SUBJ-IOBJ-XCOMP	Susan *appeared* to Peter to be worried.
SUBJ-IOBJ-PRT	Susan *gave* up on the project.		

Table 1: Fifteen SCFs with highest type frequency in our gold standard, with example sentences.

4 Methodology

4.1 Non-negative tensor factorization

Distributional co-occurrence data is usually represented in the form of a *matrix*. Matrices are perfectly suited for the representation of two-way co-occurrence data, but are unable to cope with multi-way co-occurrence data. We therefore make use of the generalization of a matrix, which is called a *tensor*. Tensor objects are able to encode co-occurrence data beyond two modes. Figure 1 shows a graphical comparison of a matrix and a tensor with three modes. Note that a tensor need not be restricted to three modes; in fact, our model requires tensors of up to 12 modes. Such tensors are difficult to represent visually, but the mathematical machinery remains unchanged.

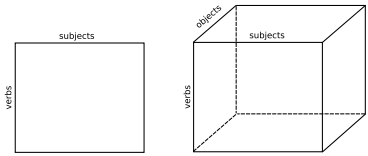

Figure 1: Matrix representation vs. tensor representation.

In order to create a succinct and generalized model of the extracted data, a statistical factorization technique called non-negative tensor factorization (NTF) is applied to the data. The NTF model is similar to parallel factor (PARAFAC) analysis – popular in areas such as psychology and bio-chemistry – with the constraint that all data needs to be non-negative (i.e. ≥ 0). PARAFAC is a multilinear analogue of the singular value decomposition (SVD), used e.g. in latent semantic analysis (Landauer and Dumais, 1997). The key idea is to minimize the sum of squares between the original tensor and the factorized model of the tensor. For an N-mode tensor $T \in \mathbb{R}^{D_1 \times D_2 \times \ldots \times D_N}$ this gives objective function (1), where k is the number of dimensions in the factorized model and $\circ$ denotes the outer product.

$$\min_{x_i \in \mathbb{R}^{D_1}, y_i \in \mathbb{R}^{D_2}, \ldots, z_i \in \mathbb{R}^{D_N}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i \|_F^2 \quad (1)$$

With non-negative tensor factorization, the non-negativity constraint is enforced, which yields a model with objective function (2).

$$\min_{x_i \in \mathbb{R}^{D_1}_{\geq 0}, y_i \in \mathbb{R}^{D_2}_{\geq 0}, \ldots, z_i \in \mathbb{R}^{D_N}_{\geq 0}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i \|_F^2 \quad (2)$$

The algorithm results in N matrices, indicating the loadings of each mode on the factorized dimensions. The model for the three-mode case is represented graphically in figure 2, visualizing the fact that the NTF decomposition consists of the summation over the outer products of N (in this case three) vectors.

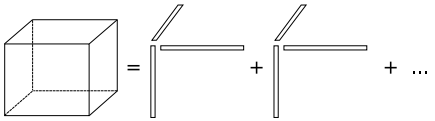

Figure 2: Graphical representation of the NTF as the sum of outer products.

Computationally, the NTF model is fitted by applying an alternating least-squares algorithm. In each iteration, two of the modes are fixed and the third one is fitted in a least squares sense. This process is repeated until convergence.[3]

4.2 Construction of verb-argument tensors

In order to discover SCFs and SPs, we construct a tensor that contains the multi-way co-occurrences of a verb and its different arguments.

4.2.1 Corpus data

We used a subset of the corpus of Korhonen et al. (2006), which consists of up to 10,000 sentences for each of approximately 6400 verbs, with data taken from five large British and American cross-domain corpora. To ensure sufficient data for each verb, we included verbs with at least 500 occurrences, yielding a total of 1993 verbs. The corpus data was tokenized, POS-tagged, lemmatized, and parsed with the RASP system (Briscoe and Carroll, 2002). RASP uses a tag-sequence grammar, and is unlexicalized, so that the parser's lexicon does not interfere with SCF acquisition. RASP produces output in the form of GRs. Passive sentences and those with clausal subjects were ignored.

4.2.2 Tensor construction

The corpus data is used to construct an N-mode tensor, where N represents the number of GRs. Each mode contains a different GR to the verb. Given the eight GRs from Section 3 plus the verb itself, this yields a 9-mode tensor (up to 12-mode when modifiers and split clausal modifiers are included; see Section 4.2.3).

For any particular verb instance (i.e. sentence), not every GR type will be instantiated. However, to model the multi-way co-occurrences in a tensor framework, each instance must have a feature for every mode to be incorporated into the tensor. Previous applications of non-negative tensor

[3]The algorithm has been implemented in MATLAB, using the Tensor Toolbox for sparse tensor calculations (Bader and Kolda, 2007).

factorization in NLP have not needed a representation for the non-instantiation of a mode. We introduce an empty, *void* (–) feature when a particular mode is not instantiated. For example, sentence (1) from Section 1 would be encoded as the tuple in (3):

(3) $\langle show_V, review_N, you_P, –, –, –, be_V, –, –\rangle$

indicating that the VERB, NCSUBJ, DOBJ, and CCOMP slots are filled with respectively $show_V$, $review_N$, you_P, *and* be_V, and that the remaining slots (IOBJ, OBJ2, PCOMP, XCOMP, PRT) are empty. (See Section 4.2.3 for alternative feature sets for each mode.)

Our final tensor then records how many times the tuple is attested in the corpus (i.e. how many times these particular features for the various grammatical relations occur together with the verb in question). The constructed tensor is then factorized to a limited number of latent dimensions, minimizing objective function (2). We normalize the factorization matrices to 1, to ensure a proper probability distribution.

Initially, we experimented with the number of latent dimensions of the factorization model (in the range 50–200). In further experiments, we retained the number of 150 dimensions, as this gave us the best results, and the model did not improve beyond 150 dimensions.

4.2.3 Feature sets

We constructed the feature sets for each mode in a number of different ways. Our base model uses the POS tag of the argument and no other features. We then experimented with a variety of additional features, based on linguistic intuitions about SCFS and SPS, as follows.

head The lexical head of the argument as well as the POS tag is used;

extpp PPS are extended to include the head of the PP's object, e.g. to_London_N (for the head models) or *to_N* (for the POS models) instead of simply *to*;

split both XCOMP and CCOMP are split up into two different modes to differentiate between null and lexicalized complementizers (e.g. for CCOMP, whether the complementizer is null or *that*);

mod modifiers (NCMOD) are included as an extra mode in the tensor.

Only the models with head features are relevant for SP acquisition. The head features also test how sensitive the learning of SCFS is to lexical-semantic as opposed to purely syntactic generalizations. The extended PP features provide additional lexical-semantic information. The clausal complement subtypes are available in the RASP output and offer a finer-grained syntactic analysis of these GRS. Finally, we used modifiers to test whether modifier co-occurrences, although (by definition) not part of the SCFS, might still be helpful in generalizing about subcategorization (i.e., maybe verbs taking similar frames also take similar modifiers). For each mode, we included the features that occurred with frequency ≥ 500 in the corpus, to maintain tractability.

For example, sentence (1) from Section 1 would be encoded as the tuple in (4) in the base POS-only model, and the tuple in (5) in the model with head and modifier features.

(4) $\langle show_V, N, P, –, –, –, V, –, –\rangle$

(5) $\langle show_V, review_N, you_P, –, –, –, be_V, –, –, comprehensively_R\rangle$

5 Experiment 1: SCF Induction

5.1 Evaluation method

SCF lexicons are traditionally evaluated against gold standards. We took the gold standard of Korhonen et al. (2006), which is a superset of SCFs in large dictionaries, and created a version using our eight basic GR types to define the SCFs. The resulting gold standard contains 183 general language verbs, with an average of 7.4 SCFs per verb. No attempt is made to distinguish between multiple senses of polysemous verbs; SCFs belonging to all senses are included for each lemma in the gold standard.

We evaluated the acquired SCF lexicons using type precision (percentage of SCF types that the system proposes which are correct), type recall (percentage of SCF types in the gold standard that the system proposes), and F-measure (the harmonic mean of type precision and recall).

We have two baselines. For baseline 1, we adopt the baseline of O'Donovan et al. (2005) which uniformly assigns to all verbs the two SCFs known to be most frequent in general language, transitive (SUBJ-DOBJ) and intransitive (SUBJ). This is a challenging baseline for SCF acquisition because of the Zipfian nature of SCF distributions: a small number of frequent SCFs are taken by the majority of verbs. For baseline 2, as described in Section 4.2.3, we use the base model with only POS features and none of the additional lexical or modifier features.

5.2 Mapping latent dimensions to SCFs

In order to evaluate this technique for SCF acquisition, we need to characterize each latent dimension according to its syntactic behaviour, i.e. map each dimension to a characteristic SCF.

Each latent dimension $\mathbf{z}$ is represented by a set of N vectors, indicating the loadings of each mode on $\mathbf{z}$. Because the loadings were normalized, each vector contains a probability distribution, over verbs or features. For a dimension $\mathbf{z}$ and a given mode (i.e. GR slot) we use the probability $p(\text{–}|\mathbf{z})$ of a void appearing in that slot to decide whether that slot is characteristically empty or filled for that dimension. For the verb mode, we use the probability $p(\mathbf{v}|\mathbf{z})$ to decide whether a verb $\mathbf{v}$ takes that dimension's characteristic SCF.

The mapping thus has two parameters. The first, θ_{verb}, represents the minimum $p(\mathbf{v}|\mathbf{z})$ for $\mathbf{v}$ to be assigned the characteristic SCF of $\mathbf{z}$. Based on early experiments, we chose to test three values for θ_{verb}, 0.001, 0.002, and 0.003.

The second parameter, θ_{void}, represents the maximum value of $p(\text{–}|\mathbf{z})$ at which the argument slot will be considered part of the SCF of $\mathbf{z}$. For example, if $p(\text{–}|\mathbf{z}) > \theta_{void}$ in the vector representing the DOBJ mode for $\mathbf{z}$, then the characteristic SCF of $\mathbf{z}$ does not include a direct object. We did not apply the θ_{void} threshold to subjects, but rather assumed that all characteristic SCFs include subjects; early experiments showed that subjects were otherwise sometimes erroneously excluded from the SCFs because the data contained high numbers of subjectless embedded clauses. For all other modes, we tested θ_{void} values from 0.1 to 0.8 in increments of 0.1.

The mapping process can be thought of as labeling the clusters produced by the tensor factorization. E.g. for a latent dimension $\mathbf{z}$ with a void value below θ_{void} for the DOBJ and IOBJ modes, its label is simply SUBJ-DOBJ-IOBJ. This label is assigned as an SCF to all the verbs with probabilities over θ_{verb} in $\mathbf{z}$.

If a dimension's characteristic SCF does not correspond to an SCF in the gold standard, that

Frame	# dims	Frame	# dims
SUBJ-DOBJ	29	SUBJ-XCOMP	17
SUBJ-DOBJ-IOBJ	9	SUBJ-DOBJ XCOMP	5
SUBJ	24	SUBJ-DOBJ-IOBJ-PRT	0
SUBJ-IOBJ	12	SUBJ-CCOMP	26
SUBJ-PRT	7	SUBJ-DOBJ-CCOMP	0
SUBJ-DOBJ-PRT	5	SUBJ-DOBJ-OBJ2	5
SUBJ-PCOMP	3	SUBJ-IOBJ-XCOMP	0
SUBJ-IOBJ-PRT	3		

Table 2: SCFs in order of type frequency in gold standard, with number of latent dimensions mapped to them (model features: POS, modifiers).

cluster is excluded from the evaluation. This typically happens with high values of θ_{void} because too many argument slots are simultaneously included in the SCF.

Note that multiple dimensions may be mapped to each SCF, because we chose the number of latent dimensions to be greater than the number of SCFs in the inventory. This decision allows the system to detect semantic structure in the data at a finer-grained level, which we hypothesized would improve overall accuracy on subcategorization acquisition, and to discover multiple lexical classes within a single argument slot. The relationship between number of dimensions mapping to an SCF and the frequency of the SCF is somewhat complex. To the extent that different verbs appear in different clusters, we expect that a larger number of dimensions mapping to an SCF roughly corresponds to higher type frequency of the SCF. However, some clusters contain more verbs than others; also, more clusters may indicate more semantic variability in argument slot fillers, without corresponding to higher frequency. A general relationship between type frequency and number of dimensions can be seen in Table 2, although note the high number of clusters mapped to the clausal complement frames SUBJ-XCOMP and SUBJ-CCOMP, possibly because these relations are semantically variable and used for adjuncts as well as arguments.

5.3 Tuning parameters

We used ten-fold cross-validation to tune the parameters θ_{verb} and θ_{void}, as well as to select the best feature combination (see Section 4.2.3). We randomly divided our test verbs into ten sets, each containing either 18 or 19 verbs. For each fold, we selected the parameters that gave the highest accuracy on the remaining nine-tenths of the verbs agaist the gold standard, and used those settings to acquire the lexicon for the 18 or 19 verbs in the fold.

For all ten folds, the best result was achieved with $\theta_{verb} = 0.001$ and $\theta_{void} = 0.4$, and with modifier features, but without extended PPs or split clause types. For seven of the folds, the best result was achieved with POS features, and for the other three with head features.

5.4 Results

Table 3 shows the results for our system after tuning with cross-validation. The parameters are: $\theta_{verb} = 0.001$, $\theta_{void} = 0.4$, POS and modifier features. Precision and recall are averaged over the ten folds. The standard deviation for precision was 4.3 and for recall 5.9. The final system achieves an F-measure of 68.7, well above the baseline 1 F-measure of 36.9, and nearly four

	P	R	F
Baseline 1	86.3	23.5	36.9
Baseline 2 (POS features)	53.1	83.3	64.8
Final system	61.0	78.5	68.7

Table 3: Results of cross-validation experiment. Precision and recall averaged over ten folds. F-score calculated as harmonic mean of the average P and R.

points better than the baseline 2 F-measure of 64.8. All of the improvement over baseline 2 is in precision, which shows that adding features beyond simple GR co-occurrences is beneficial to accurate SCF acquisition. Because of the Zipfian nature of SCF distributions, the system does not match the precision of baseline 1.

Direct comparison against previous unsupervised SCF acquisition methods on English was not possible because of the use of different data and frame inventories. However, best current methods involving handcrafted rules have reached a ceiling at an F-measure of about 70 (Korhonen et al., 2006; Preiss et al., 2007). Our results are promising considering the challenges of less supervised lexical acquisition.

5.5 Investigation of features

We also investigated the contribution of the different feature sets on the entire gold standard, using the values for θ_{verb} and θ_{void} which emerged from the cross-validation. The results of the different models are shown in Table 4 (note that the best result is slightly different from that in Table 3 because it is on the entire gold standard, not averaged over folds).

	Model							
	head	PP	split	mod	P	R	F	cov
1				•	61.4**,††	81.1**,††	**69.9**††	183
2	•			•	63.9**,††	76.4**,††	69.6††	183
3	•		•		67.2**,††	70.4**,††	68.8††	183
4		•	•		59.3††	80.9††	68.4††	183
5		•			58.7**,††	81.2**,††	68.2††	183
6		•		•	60.5*,††	77.9**,††	68.1††	183
7			•	•	58.7**,††	81.2**,††	68.1††	182
8		•	•	•	61.2**,††	76.0**,††	67.8††	183
9	•	•	•		**67.5****,††	67.7**,††	67.6††	183
10			•		56.1**,††	83.1**	67.0††	183
11	•	•			60.2††	74.3**,††	66.5†	182
12	•	•		•	61.8*,††	71.4*,††	66.3	183
13	•				59.8**,††	73.6**,††	66.0	183
14					53.1**	**83.3****	64.8*	183
15	•		•	•	65.1††	60.3**,††	62.6**,†	183
16	•	•	•	•	63.3††	52.6††	57.5††	181

Table 4: Results for each feature set, with 150 dimensions, $\theta_{verb} = 0.001$, $\theta_{void} = 0.4$. ** significant difference from next row with $p < 0.01$, * with $p < 0.05$. †† significant difference from baseline (row 14) with $p < 0.01$, † with $p < 0.05$.

The differences in F-measure between the top few models are rather small, but the models show wide variance in precision and recall. Using the head words of the arguments as features seems to favor precision (rows 2, 3, 9, 15, 16), while using POS tags favors recall. This is probably because evidence for different arguments is less sparse using POS tags, making less frequent frames easier to identify, but finer-grained distinctions more difficult. The highest F-scores are achieved with modifier features (rows 1, 2); however, these models strongly favor recall over precision, suggesting that the general applicability of modifiers to many verb classes interferes with accurate identification of SCFs. More balanced models have head features and split clausal complement types (row 3), or head features, extended PPs, and split clausal types (row 9), without losing out on F-score. This suggests that lexical-semantic features are valuable for SCF acquisition. Another trend is towards more accurate models with fewer additional features; individual features and pairs of features seem to provide the most improvement (rows 1-7) over the base model (row 14), but the model with all additional features (row 16) has markedly worse performance, which may indicate a data sparsity problem.

We carried out significance tests for the mentioned model differences using stratified shuffling (Yeh, 2000). These tests indicate that most of the models (rows 1-11) have significantly higher F-score than the baseline, and most show significant pairwise differences in precision and recall.

Parameter tuning with cross-validation resulted in a θ_{void} of 0.4 (though exploration of the models in Table 4 showed that some models performed better with even lower values). This means that the model only needs to assign a relatively low confidence score to the void feature to infer that a slot is not part of an SCF. This is probably because adjuncts and other noise in the data means that these slots are filled some of the time. We observed many cases of void probabilities nearly equal to 1 in various dimensions – most verbs never occur with an OBJ2, for example. However, void probabilities tend to be fairly low for CCOMP and XCOMP.

5.6 Qualitative evaluation

Table 5 shows the accuracy by SCF for the fifteen most frequent frames, using the final model that resulted from cross-validation. The system performs very well on a number of SCFs, especially the most frequent ones such as SUBJ-DOBJ, SUBJ-DOBJ-IOBJ, and SUBJ, but also on some SCFs involving the semantically important particle verbs, such as SUBJ-DOBJ-PRT and SUBJ-IOBJ-PRT. Precision is lower on frames involving clausal complements (XCOMP and CCOMP), possibly because these GRs are used frequently for adjuncts. Accuracy is also poor on SUBJ-PCOMP and SUBJ-DOBJ-OBJ2. These GRs are rarer and may be subject to parser errors (e.g. OBJ2).

6 Experiment 2: SP Induction

6.1 Introduction

Our second experiment looks at the model's ability to induce SPs. We investigate the model's ability to induce *multi-way* SPs, i.e. the preference of the model for the co-occurrence of a particular verb and all of its particular arguments *at the same time*.

The calculation of a SP value according to our NTF model is fairly straightforward. Recall that our model yields probabilities $p(\mathbf{v}|\mathbf{z})$, i.e. the probability of a verb given a latent dimension, and, for each argument to the verb, $p(\mathbf{g}|\mathbf{z})$, i.e. the probability of an argument given a latent dimension. The final SP value $SP(v, GR)$ for a particular verb v and a list of arguments GR then amounts to calculating the product of the probabilities of the verb and the various GRs given a particular latent dimension, and summing over all dimensions (equation 3).

Frame	P	R	F	Frame	P	R	F
SUBJ-DOBJ	95.4	98.8	97.0	SUBJ-XCOMP	44.0	98.6	60.9
SUBJ-DOBJ-IOBJ	89.6	88.5	89.0	SUBJ-DOBJ-XCOMP	45.9	79.4	58.1
SUBJ	82.7	98.7	90.0	SUBJ-DOBJ-IOBJ-PRT	0.0	0.0	0.0
SUBJ-IOBJ	80.6	91.5	85.7	SUBJ-CCOMP	35.9	100.0	52.8
SUBJ-PRT	75.2	87.1	80.7	SUBJ-DOBJ-CCOMP	33.3	71.1	45.4
SUBJ-DOBJ-PRT	72.8	83.0	77.6	SUBJ-DOBJ-OBJ2	20.0	90.3	32.8
SUBJ-PCOMP	56.9	45.7	50.7	SUBJ-IOBJ-XCOMP	0.0	0.0	0.0
SUBJ-IOBJ-PRT	71.9	83.1	77.1				

Table 5: Results by SCF for fifteen most frequent frames in gold standard with best-performing model.

$$SP(v, GR) = \sum_{i=1}^{k} p(v|z_i) \prod_{g \in GR} p(g|z_i) \quad (3)$$

We evaluate our method's ability to induce SPs using the lexicalized (HEAD) model that achieves the best score in our first experiment, i.e. model 2 in Table 4.

6.2 Evaluation method

To evaluate the results of the NTF model with regard to SPs, we make use of a pseudo-disambiguation task (similar to the one used by Rooth et al. (1999)). The task allows us to evaluate the generalization capabilities of the model. For a particular tuple (viz. a verb and its various arguments) that appears in a held-out test corpus, we generate random instances in which one or several arguments are substituted by random instantiations. We exhaustively substitute every individual argument, as well as the various random combinations.[4] For the sentence in (1), this yields instances like:

(6) $\langle show_V, rabbit_N, you_P, –, –, –, be_V, –, –\rangle$

(7) $\langle show_V, consumption_N, tunnel_N, –, –, –, dream_V, –, –\rangle$

We then calculate SP values according to our model, both for the corpus instance and the random instances. A tuple is considered correct if our model prefers the corpus instance over all random instances. Accuracy is then calculated by averaging over all instances that are part of the test corpus.

We compare our NTF model to a simple non-negative matrix factorization (NMF) model, comparable to the unsupervised model presented by Rooth et al. (1999). For this model, a matrix was constructed that contains the pairwise co-occurrence frequencies of verbs and their various arguments. As noted before, a matrix is only able to represent two modes; hence, the first mode consists of the verbs, while the second mode contains the concatenated list of the different argument features. We used the same number of features as with the NTF model, and also factorized to 150 dimensions. According to the NMF model, a tuple is considered correct if, for each argument to the verb, the model prefers the verb-argument pair containing the attested argument over the verb-argument pair containing the random substitute. As a baseline, we

[4] We do not substitute empty argument slots with lexical arguments; neither do we substitute filled arguments slots with *void* values. This experiment solely focuses on the induction of selectional preferences; the induction of SCFs is evaluated in experiment 1.

include an uninformed random model, which makes a random choice among the various possibilities.[5]

The models are evaluated using ten-fold cross-validation: the corpus is divided into 10 equal parts; in each fold, models are trained on nine tenths of the corpus, and tested on the remaining tenth.

6.3 Results

The results of the ten-fold cross-validation are shown in table 6. The NTF model clearly outperforms the matrix factorization model with regard to the reconstruction of SPS, with the NTF model reaching a score about 10% higher than its NMF counterpart. These results indicate that the use of multi-way data leads to a richer and more accurate representation of SPS. For comparison, (Van de Cruys, 2009) achieved accuracy of 90.89 on a three-way pseudo-disambiguation task, which is less complex than our eight-way task.

	accuracy (%)
baseline	29.21 ± .08
NMF	69.71 ± .28
NTF	**77.78** ± .17

Table 6: Selectional preference accuracy using ten-fold cross-validation (mean accuracy and standard deviation)

6.4 Qualitative evaluation

Additionally, we performed a qualitative evaluation of the 150 latent dimensions yielded by our NTF model. This evaluation shows that our model is indeed able to capture semantic information from the data. Recall from Section 5.2 that multiple dimensions map to a single SCF. Our cluster analysis shows that such dimensions reflect semantic information. Below are three example dimensions (denoted by the top 10 verbs with highest value on each dimension) that all map to a simple transitive SCF.

dim 29 buy, sell, use, collect, produce, handle, remove, purchase, obtain, eat

dim 38 kill, love, see, like, marry, know, meet, visit, help, say

dim 44 examine, identify, see, consider, assess, investigate, discuss, study, determine, explore

The three different transitive SCFS clearly exhibit different semantic properties. Dimension 29 seems to represent a general 'trading' dimension, in which the DOBJ argument contains inanimate objects, largely goods. The DOBJ argument has nouns such as *thing*, *material*, *food*,... as its top features. Dimension 38, on the other hand, is a transitive frame where the DOBJ argument takes animate objects. The last dimension 44 represents a transitive frame in which the DOBJ argument takes abstract objects.

Among the dimensions that map to the SUBJ-IO SCF, i.e. a single PP argument, there are also some interesting semantic and syntactic distinctions. Dimension 91 clearly represents a 'travel' cluster with a location complement; the IO slot for this dimension is mostly PPs lexicalized with *to*. Dimension 122 is a 'communication' cluster, and again most of the prepositions in the IO slot are *to*. Dimension 123 consists of verbs that occur with *at*, largely vision and non-verbal

[5] Note that the number of possibilities for both tensor and matrix model is exactly the same.

communication verbs. Finally, dimension 134 is interesting, because there is no clear semantic cohesion, but it represents cluster of verbs that take PP *for*. This indicates that the model is learning both semantic and syntactic regularities.

dim 91 go, come, return, move, walk, get, run, rush, travel, fly

dim 122 talk, speak, listen, write, belong, happen, appeal, come, say, lie

dim 123 look, stare, smile, laugh, shout, gaze, glance, glare, grin, scream

dim 134 wait, pay, look, care, work, ask, vote, call, prepare, apply

The results presented here indicate that our model is able to capture syntactic as well as semantic properties. On a coarse-grained level, our model is able to induce a verb's different SCF frames. When we zoom in to the level of individual clusters, we notice that these clusters are often semantically cohesive, expressing the selectional preferences of the verb's argument slots. The ability to capture both syntax and semantics is an important advantage of our method.

7 Conclusion

We have presented a novel method for joint unsupervised SCF and SP acquisition which allows the incorporation of a range of features (syntactic, lexical and semantic) in the acquisition process. Although SCFs and SPs are closely related and can complement each other, to the best of our knowledge, no previous work has proposed a joint model for them.

Applying NTF to the multi-way co-occurrence tensor of verbs and their arguments, we are able to cluster verbs from a large corpus according to their syntactic and semantic behaviour. The SCF lexicon that emerges from the clusters yields an F-score of 68.7 against a gold standard, outperforming lexicons produced by our baseline methods. This performance is promising for a largely unsupervised method. The model yields an accuracy of 77.8 on a new pseudo-disambiguation evaluation for SPs, in which all arguments of the verb are considered at once, clearly outperforming a matrix factorization model. Our qualitative evaluation reveals that the method is indeed capable of learning rich lexical information about both syntactic and semantic aspects of verb behaviour in corpus data.

In the future, we plan to improve our approach in several directions. In addition to improving the detection of low accuracy SCFs through the use of lexical features that may help to distinguish arguments from adjuncts in clausal complements, we plan to improve precision by using e.g. statistical filtering. We also plan to extend the model to acquire finer-grained SCFs for English. This will involve e.g. refining SCFs with lexicalized elements and including semantically-based SCFs in the inventory, making use of the factorization method's ability to induce latent structure, as demonstrated by the SP evaluation. Finally, we intend to improve our SP acquisition through the use of a more extensive feature set.

A key advantage of this approach is that it is able to combine syntactic SCF and semantic SP acquisition. In the future, we plan to explore the joint induction of verb syntax and semantics in greater depth and look into modelling additional information about semantic verb classes which tend to capture similar SCF and SP behaviour. This could facilitate inducing a more comprehensive lexical resource that supplements the SCFs and SPs with a verb classification – in the style of VerbNet (Kipper-Schuler, 2005) – providing generalizations that can be useful for a wider range of NLP tasks.

Acknowledgements

The work in this paper was funded by the Royal Society (UK), the Isaac Newton Trust (Cambridge, UK), EPSRC grant EP/G051070/1 (UK) and EU grant 7FP-ITC-248064 'PANACEA'.

References

Abend, O., Reichart, R., and Rappoport, A. (2009). Unsupervised argument identification for semantic role labeling. In *Proceedings of ACL*.

Altamirano, I. R. and Alonso i Alemany, L. (2010). IRASubcat, a highly customizable, language independent tool for the acquisition of verbal subcategorization information from corpus. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 84–91, Los Angeles, CA.

Arun, A. and Keller, F. (2005). Lexicalization in crosslinguistic probabilistic parsing: The case of french. In *Proceedings of ACL*, Ann Arbor, Michigan.

Bader, B. W. and Kolda, T. G. (2007). Matlab tensor toolbox version 2.2. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`.

Basili, R., Cao, D. D., Marocco, P., and Pennacchiotti, M. (2007). Learning selectional preferences for entailment or paraphrasing rules. In *Proceedings of RANLP 2007*, Borovets, Bulgaria.

Bhagat, R., Pantel, P., and Hovy, E. (2007). Ledir: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*, page 161–170, Prague, Czech Republic.

Bharati, A., Venkatapathy, S., and Reddy, P. (2005). Inferring semantic roles using subcategorization frames and maximum entropy model. In *Proceedings of CoNLL*, pages 165–168, Ann Arbor.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, pages 356–363. Association for Computational Linguistics.

Briscoe, T. and Carroll, J. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

Carroll, G. and Rooth, M. (1996). Valence induction with a head-lexicalized pcfg. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP 3)*, pages 36–45.

Carroll, J. and Fang, A. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114, Sanya City, China.

Chesley, P. and Salmon-Alt, S. (2006). Automatic extraction of subcategorization frames for french. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 253–258, Genoa, Italy.

Cholakov, K. and van Noord, G. (2010). Using unknown word techniques to learn known words. In *Proceedings of EMNLP*, pages 902–912, Massachusetts.

Dinu, G. and Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics.

Dębowski, Ł. (2009). Valence extraction using EM selection and co-occurrence matrices. *Language resources and evaluation*, 43(4):301–327.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, page 245–288.

Hajič, J., Čmejrek, M., Dorr, B., Ding, Y., Eisner, J., Gildea, D., Koo, T., Parton, K., Penn, G., Radev, D., and Rambow, O. (2002). Natural language generation in the context of machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

Han, C., Lavoie, B., Palmer, M., Rambow, O., Kittredge, R., Korelsky, T., and Kim, M. (2000). Handling structural divergences and recovering dropped arguments in a korean/english machine translation system. In *Proceedings of the AMTA*.

Ienco, D., Villata, S., and Bosco, C. (2008). Automatic extraction of subcategorization frames for italian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Kawahara, D. and Kurohashi, S. (2010). Acquiring reliable predicate-argument structures from raw corpora for case frame compilation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, PA.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Korhonen, A. (2002). Semantically motivated subcategorization acquisition. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 51–58. Association for Computational Linguistics.

Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*, Genoa, Italy.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychology Review*, 104:211–240.

Lenci, A., McGillivray, B., Montemagni, S., and Pirrelli, V. (2008). Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Lippincott, T., Korhonen, A., and Ó Séaghdha, D. (2012). Learning syntactic verb frames using graphical models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

Messiant, C. (2008). ASSCI: A subcategorization frames acquisition system for French verbs. In *Proceedings of the Association for Computational Linguistics (ACL, Student Research Workshop)*, pages 55–60, Columbus, Ohio.

Moschitti, A. and Basili, R. (2005). Verb subcategorization kernels for automatic semantic labeling. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 10–17, Ann Arbor.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J., and Way, A. (2005). Large-scale induction and evaluation of lexical resources from the penn-ii and penn-iii treebanks. *Computational Linguistics*, 31:328–365.

Ó Séaghdha, D. (2010). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Preiss, J., Briscoe, T., and Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*, Prague, Czech Republic.

Reisinger, J. and Mooney, R. (2011). Crosscutting models of lexical semantics. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proc. of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*

Ritter, A. and Etzioni, O. (2010). A latent dirichlet allocation method for selectional preferences. In *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL*.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via em-based clustering. In *37th Annual Meeting of the ACL*.

Scheible, C. (2010). An evaluation of predicate argument clustering using pseudo-disambiguation. In *Proceedings of LREC*.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Schulte im Walde, S., Hying, C., Scheible, C., and Schmid, H. (2008). Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the ACL*, pages 496–504, Columbus, OH.

Seaghdha, D. O. and Korhonen, A. (2011). Probabilistic models of similarity in syntactic context. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Shashua, A. and Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM.

Sun, L. and Korhonen, A. (2011). Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*, pages 1023–1033, Edinburgh, Scotland.

Thater, S., Furstenau, H., and Pinkal, M. (2010). Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of ACL-10*, Uppsala, Sweden.

Van de Cruys, T. (2009). A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the workshop on Geometric Models for Natural Language Semantics (GEMS)*, pages 83–90, Athens, Greece.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics*, pages 947–953, Saarbrücken, Germany.

Zapirain, B., Agirre, E., and Marquex, L. (2009). Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of ACL-IJCNLP-09*, Singapore.

Zhou, G., Zhao, J., Liu, K., and Cai, L. (2011). Exploiting web-derived selectional preference to improve statistical dependency parsing. In *Proceedings of ACL-11*, Portland, OR.