Improved Lexical Acquisition through DPP-based Verb Clustering

Roi Reichart
University of Cambridge, UK
rr439@cam.ac.uk

Anna Korhonen
University of Cambridge, UK
alk23@cam.ac.uk

Abstract

Subcategorization frames (SCFs), selectional preferences (SPs) and verb classes capture related aspects of the predicate-argument structure. We present the first unified framework for unsupervised learning of these three types of information. We show how to utilize Determinantal Point Processes (DPPs), elegant probabilistic models that are defined over the possible subsets of a given dataset and give higher probability mass to high quality and diverse subsets, for clustering. Our novel clustering algorithm constructs a joint SCF-DPP DPP kernel matrix and utilizes the efficient sampling algorithms of DPPs to cluster together verbs with similar SCFs and SPs. We evaluate the induced clusters in the context of the three tasks and show results that are superior to strong baselines for each [1].

1 Introduction

Verb classes (VCs), subcategorization frames (SCFs) and selectional preferences (SPs) capture different aspects of predicate-argument structure. SCFs describe the syntactic realization of verbal predicate-argument structure, SPs capture the semantic preferences verbs have for their arguments and VCs in the Levin (1993) tradition provide a shared level of abstraction for verbs that share many aspects of their syntactic and semantic behavior.

These three of types of information have proved useful for Natural Language Processing (NLP) tasks which require information about predicate-argument structure, including parsing (Shi and Mihalcea, 2005; Cholakov and van Noord, 2010; Zhou et al., 2011), semantic role labeling (Swier and Stevenson, 2004; Dang, 2004; Bharati et al., 2005; Moschitti and Basili, 2005; zap, 2008; Zapirain et al., 2009), and word sense disambiguation (Dang, 2004; Thater et al., 2010; Ó Séaghdha and Korhonen, 2011), among many others.

Because lexical information is highly sensitive to domain variation, approaches that can identify VCs, SCFs and SPs in corpora have become increasingly popular, e.g. (O'Donovan et al., 2005; Schulte im Walde, 2006; Erk, 2007; Preiss et al., 2007; Van de Cruys, 2009; Reisinger and Mooney, 2011; Sun and Korhonen, 2011; Lippincott et al., 2012).

The task of SCF induction involves identifying the arguments of a verb lemma and generalizing about the frames (i.e. SCFs) taken by the verb, where each frame includes a number of arguments and their syntactic types. For example, in (1), the verb "show" takes the frame SUBJ-DOBJ-CCOMP (subject, direct object, and clausal complement).

(1) [A number of SCF acquisition papers]SUBJ [show]VERB [their readers]DOBJ [which features are most valuable for the acquisition process]CCOMP.

SP induction involves identifying and classifying the lexical items in a given argument slot. In sentence (2), for example, the verb "show" takes the frame SUBJ-DOBJ. The direct object in this frame is likely to be inanimate.

(2) [Most SCF and SP acquisition papers]SUBJ,

[1] The source code of the clustering algorithms and evaluation is submitted with this paper and will be made publicly available upon acceptance of the paper.

[show]VERB [no evidence to the usefulness of joint learning leaning for these tasks]DOBJ.

Finally, VC induction involves clustering together verbs with similar meaning, reflected in similar SCFs and SPs. For example, "show" in the above examples could get clustered together with "demonstrate" and "indicate".

Because these challenging tasks capture complementary information about predicate argument structure, they should be able to inform and support each other. Recently, researchers have begun to investigate the benefits of their joint learning. Schulte im Walde et al. (2008) integrated SCF and VC acquisition and used it for WordNet-based SP classification. Ó Séaghdha (2010) presented a "dual-topic" model for SPs that induces also verb clusters. Both works reported SP evaluation with promising results. Lippincott et al. (2012) presented a joint model for inducing simple syntactic frames and VCs. They reported high accuracy results on VCs. de Cruys et al. (2012) introduced a joint model for SCF and SP acquisition. They evaluated both the SCFs and SPs, obtaining reasonable result on both tasks.

In this paper, we present the first unified framework for unsupervised learning of the three types of information - SCFs, SPs and VCs. Our framework is based on Determinantal Point Processes (DPPs, (Kulesza, 2012; Kulesza and Taskar, 2012c)), elegant probabilistic models that are defined over the possible subsets of a given dataset and give higher probability mass to high quality and diverse subsets.

We first show how individual-task DPP kernel matrices can be naturally combined to construct a joint kernel. We use this to construct a joint SCF-SP kernel. We then introduce a novel clustering algorithm based on iterative DPP sampling which can (contrary to other probabilistic frameworks such as Markov random fields) be performed both accurately and efficiently. When defined over the joint SCF and SP kernel, this new algorithm can be used to induce VCs that are valuable for both tasks.

We also contribute by evaluating the value of the clusters induced by our model for the acquisition of the three information types. Our evaluation against a well-known VC gold standard shows that our clustering model outperforms the state-of-the-art verb clustering algorithm of Sun and Korhonen (2009), in our setup where no manually created SCF or SP data is available. Our evaluation against a well-known SCF gold standard and in the context of SP disambiguation tasks shows results that are superior to strong baselines, demonstrating the benefit our approach.

2 Previous Work

SCF acquisition Most current works induce SCFs from the output of an unlexicalized parser (i.e. a parser trained without SCF annotations) using hand-written rules (Briscoe and Carroll, 1997; Korhonen, 2002; Preiss et al., 2007) or grammatical relation (GR) co-occurrence statistics (O'Donovan et al., 2005; Chesley and Salmon-Alt, 2006; Ienco et al., 2008; Messiant et al., 2008; Lenci et al., 2008; Altamirano and Alonso i Alemany, 2010; Kawahara and Kurohashi, 2010).

Only a handful of SCF induction works are unsupervised. Carroll and Rooth (1996) applied an EM-based approach to a context-free grammar based model, Dkebowski (2009) used point-wise co-occurrence of arguments in parsed Polish data and Lippincott et al. (2012) presented a Bayesian network model for syntactic frame induction that identifies SPs on argument types. However, the frames induced by Lippincott et al. (2012) do not capture sets of arguments for verbs so are far simpler than traditional SCFs.

Current approaches to SCF acquisition suffer from lack of semantic information which is needed to guide the purely syntax-driven acquisition process. Previous works have showed the benefit of hand-coded semantic information in SCF acquisition (Korhonen, 2002). We will address this problem in an unsupervised way: our approach is to consider SCFs together with semantic SPs through VCs which generalize over syntactically and semantically similar verbs.

SP acquisition Considerable research has been conducted on SP acquisition, with a variety of unsupervised models proposed for this task that use no hand-crafted information during training. The latter approaches include latent variable models (Ó Séaghdha, 2010; Ritter and Etzioni, 2010; Reisinger and Mooney, 2011), distributional similarity methods (Bhagat et al., 2007; Basili et al., 2007; Erk, 2007) and methods based on non-negative tensor factorization (Van de Cruys, 2009). These works use a variety of linguistic features in the acquisition process but none of them

integrates the three information types covered in our work.

Verb clustering A variety of VC approaches have been proposed in the literature. These include syntactic, semantic and mixed syntactic-semantic classifications (Grishman et al., 1994; Miller, 1995; Baker et al., 1998; Palmer et al., 2005; Schuler, 2006; Hovy et al., 2006). We focus on Levin style classes (Levin, 1993) which are defined in terms of diathesis alternations and capture generalizations over a range of syntactic and semantic properties. Previous unsupervised VC acquisition approaches clustered a variety of linguistic features using different (e.g. K-means and spectral) algorithms (Schulte im Walde, 2006; Joanis et al., 2008; Sun et al., 2008; Li and Brew, 2008; Korhonen et al., 2008; Sun and Korhonen, 2009; Vlachos et al., 2009; Sun and Korhonen, 2011). The linguistic features included SCFs and SPs, but these were induced separately and then feeded as features to the clustering algorithm. Our framework combines together SCF-motivated and SP-motivated kernel matrices , and uses the joint kernel to induce verb clusters which are likely to be highly relevant for both tasks. Importantly, no manual or automatic system for SCF or SP acquisition has been utilized when constructing the kernel matrices, we only consider features extracted from the output of an unlexicalized parser. Our approach hence provides a framework for acquiring valuable information for the three tasks together.

Joint Modeling A small number of works have recently investigated joint approaches to SCFs, SPs and VCs. Each of them has addressed only a subset of the tasks and all but one have evaluated the performance in the context of one task only. Ó Séaghdha (2010) presented a "dual-topic" model for SPs that induces VCs, reporting evaluation of SPs only. Lippincott et al. (2012) presented a Bayesian network model for syntactic frame (rather than full SCF) induction that induces VCs. Only VCs are evaluated. de Cruys et al. (2012) presented a joint unsupervised model of SCF and SP acquisition based on non-negative tensor factorization. Both SCFs and SPs were evaluated. Finally, the model of Schulte im Walde et al. (2008) addresses the three types of information but SP parameters are estimated with a WordNet based method and only the SPs are evaluated. Although evaluation of these recent joint models has been partial, the results have been encouraging and further motivate the development of a framework that acquires the three types of information together.

3 The Unified Framework

In this section we present our unified framework. Our idea is to utilize DPPs for verb clustering that informs both SCF and SP acquisition. DPPs define a probability distribution over the possible subsets of a given set. These models assign higher probability mass to subsets that are both high quality and diverse.

Our novel clustering algorithm makes use of three DPP properties that are appealing for our purpose: (1) The existence of efficient sampling algorithms for these models, which enable tractable sampling of high quality and diverse verb subsets; (2) Such verb subsets form natural high quality seeds for hierarchical clustering; and (3) Given individual-task DPP kernel matrices there are various simple and natural ways to combine them into a new DPP kernel matrix.

Individual task DPP kernels represent (i) the quality of a data point (verb) as its average feature-based similarity with the other points in the data set and (ii) the divergence between a pair of points as the inverse similarity between them. For different tasks, different feature sets are used for the kernel construction. The high quality and diverse subsets sampled from the DPP model are considered good cluster seeds as they are likely to be relatively uniformly spread and to provide good coverage of the data set. The algorithm induces an hierarchical clustering, which is particularly suitable for semantic tasks, where a set of clusters that share a parent consists of pure members (i.e. most of the points in each cluster member belong to the same gold cluster) and together provide good coverage of the verb space.

After a brief description of the Determinantal Point Processes (DPP) framework (Section 3.1), we discuss the construction of the joint DPP kernel, given a kernel for each individual task, In section 3.3 we present the DPP-Cluster clustering algorithm.

3.1 Determinantal Point Processes

Determinantal point processes (DPPs) are elegant probabilistic models of repulsion that offer efficient and exact algorithms for sampling, marginalization, conditioning, and other inference tasks. Recently (Kulesza, 2012; Kulesza and Taskar,

2012c) introduced them to the machine learning community and demonstrated their usefulness for a variety of tasks including document summarization, image search, modeling non-overlapping human poses in images and video and automatically building timelines of important news stories (Kulesza and Taskar, 2010; Kulesza and Taskar, 2012a; Gillenwater et al., 2012; Kulesza and Taskar, 2012b). Below we provide a brief description of the framework, a comprehensive survey can be found in (Kulesza and Taskar, 2012c).

Given a set of items $\mathcal{Y} = \{y_1, \ldots, y_N\}$, a DPP $\mathcal{P}$ defines a probability measure on the set of all subsets of $\mathcal{Y}$, $2^{\mathcal{Y}}$. Kulesza and Taskar (2012c) restricted their discussion of DDPs to L-ensembles, where the probability of a subset $\mathbf{Y} \in \mathcal{Y}$ is defined through a positive semi-definite matrix L indexed by the elements of $\mathcal{Y}$:

$$\mathcal{P}_L(\mathbf{Y} = Y) = \frac{det(L_Y)}{\sum_{Y \subseteq Y} det(L_Y)} = \frac{det(L_Y)}{det(L+I)} \quad (1)$$

Where I is the $N \times N$ identity matrix and $det(L_\phi) = 1$. Since L is positive semi-definite, it can be decomposed to $L = B^T B$. This allows the construction of an intuitively interpretable model where each column B_i is the product of a quality term $q_i \in R^+$ and a vector of (normalized) diversity features $\phi_i \in R^D, ||\phi_i|| = 1$. In this model, q_i measures an inherent quality of the $i-th$ item in $\mathcal{Y}$ while $\phi_i^T \phi_j \in [-1, 1]$ is a similarity measure between items i and j. With this representation we can write:

$$L_{ij} = q_i \phi_i^T \phi_j q_j \quad (2)$$

$$S_{ij} = \phi_i^T \phi_j = \frac{L_{ij}}{\sqrt{L_{ii} L_{jj}}} \quad (3)$$

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto (\prod_{i \in Y} q_i^2) det(S_Y) \quad (4)$$

It can be shown that the first term in equation 4 increases with the quality of the selected items, and the second term increases with their diversity. As a consequence, this distribution places most of its weight on sets that are both high quality and diverse.

Although the number of possible realizations of Y is exponential in N, many inference procedures can be performed accurately and efficiently (i.e. in polynomial time which is very short in practice). In particular, sampling, which NP-hard for alternative models such as Markov Random Fields (MRFs), is efficient, theoretically and practically, for DPPs.

3.2 Constructing a Joint Kernel Matrix

DPPs are particularly suitable for joint modeling as they come with various simple and intuitive ways to combine individual model kernel matrices into a joint kernel. This stems from the fact that every positive-semidefinite matrix forms a legal DPP kernel (equation 1). Given individual model DPP kernels, we would therefore like to combine them into a positive-semidefinite matrix.

While there are various ways to construct a positive-semidefinite matrix from two positive-semidefinite matrices – for example, by taking their sum – in this work we are motivated by the product of experts approach (Hinton, 2002), reasoning that high quality assignments according to a product of models have to be of high quality according to each individual model, and sick for a product combination. [2]

In practice we construct the joint kernel in the following way. We build on the aforementioned property that a matrix L is positive semi-definite iff $L = B^T B$. Given two DPPs, $\mathcal{P}_{L^1}$ defined by $L^1 = A_1^T A_1$ and $\mathcal{P}_{L^2}$ defined by $L^2 = A_2^T A_2$, we construct the joint kernel L^{12}:

$$L^{12} = L^1 L^2 L^2 L^1 = A_1^T A_1 A_2^T A_2 A_2^T A_2 A_1^T A_1 = C^T C \quad (5)$$

Where $C = A_2^T A_2 A_1^T A_1$ and $C^T = A_1^T A_1 A_2^T A_2$.

3.3 Clustering Algorithm

Algorithm (1) and Figure (1) provide a pseudo-code of the algorithm and an example output. Below is a detailed description.

Features Our algorithm builds two DPP kernel matrices (the $GenKernelMatrix$ function), in which the rows and columns correspond to the verbs in the data set, such that the (i, j)-th entry corresponds to verbs number i and j. Following equations 2 and 3 one matrix is built for SCF and

[2] Note that we do not take a product of the individual models but only of their kernel matrices. Yet, if we construct the joint matrix by a multiplication then it follows from a simple generalization of the Cauchy-Binet formula that its principle minors, which define the subset probabilities (equation 1), are a sum of multiplications of the principle minors of the individual model kernels. Still, we do not have guarantees that our choice of kernel combination is the right one. We leave this for future research.

one for SP, and they are then combined into the joint kernel matrix (the $GenJointMat$ function) following equation 5. Each kernel matrix requires a proper feature representation ϕ and quality score q.

In both kernels we represent a verb by the counts of the grammatical relations (GRs) it participates in. In the SCF kernel a GR is represented by the GR type and the POS tags of the verb and its arguments. In the SP kernels the GRs are represented by the POS tags of the verb and its arguments as well as by the argument head word. Based on this feature representation, the similarity (opposite divergence) is encoded to the model by equation 3 as the dot product between the normalized feature vectors. The quality score q_i of the i-th verb is the average similarity of this verb with the other verbs in the dataset.

Cluster set construction In its while loop, the algorithm iteratively generates fixed-size cluster sets such that each data point belongs to exactly one cluster in one set. These cluster sets form the leaf level of the tree in Figure (1). It does so by extracting the T highest probability K-point samples from a set of M subsets, each of which sampled from the joint DPP model, and clustering them by the $cluster$ procedure. The sampling is done by the K-DPP sampling process ((Kulesza and Taskar, 2012c), page 62) [3].

The $cluster$ procedure first seeds a K-cluster set with the highest probability sample. Then, it gradually extends the clusters by iteratively mapping the samples, in decreasing order of probability, to the existing clusters (the $m1Mapping$ function). Mapping is done by attaching every point in the mapped subset to its closet cluster, where the distance between a point and the cluster is the maximum over the distances between the point and each of the points in the cluster. The mapping is many-to-one, that is, multiple points in the subset can be assigned to the same cluster.

Based on the DPP properties, the higher the probability of a sampled subset, the more likely it is to consist of distinct points that provide a good coverage of the verb set. By iteratively extending the clusters with high probability subsets, we thus expect each cluster set to consist of clusters that demonstrate these properties.

[3]K-DPP is a DPP conditioned on the sample size. As shown in ((Kulesza and Taskar, 2012c), Section 2.4.3) this conditional distribution is also a DPP. We could have obtained samples of size K by sampling the DPP and rejecting samples of other sizes but this would have been slower.

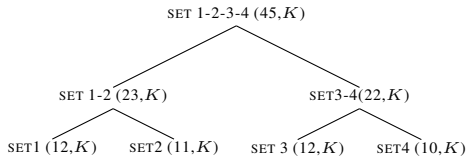

Figure 1: An example output hierarchy of DPP-Cluster for a set of 45 data points. Each set is augmented with the number of points (left number) and clusters (right number) it includes. The iterative DPP-samples clustering (the While loop) generates the lowest level of the tree, by dividing the data set into cluster sets, each of which consists of K clusters. Each point in the data set belongs to exactly one cluster in exactly one set. The agglomerative clustering then iteratively combines cluster sets such that in each iteration two sets are combined to one set with K clusters.

Agglomerative Clustering Finally, the $AgglomerativeClustering$ function builds a hierarchy of cluster sets, by iteratively combining cluster set pairs. In each iteration it computes the similarity between any such pair, defined to be the lowest similarity between their cluster members, which is in turn defined to be the lowest cosine similarity between their point members. The most similar cluster sets are combined such that each of the clusters in one set is mapped to its most similar cluster in the other set. In this step the algorithm generates data partitions at different granularity levels from finest (from the iterative sampling step) to the coarsest set (generated by the last agglomerative clustering iteration and consisting of exactly K clusters). This property is useful as the optimal level of generalization may be task dependent.

4 Evaluation

Data sets and gold standards We evaluated the SCFs and verb clusters on gold standard datasets. We based our set of the largest available joint set for SCFs and VCs - that of (de Cruys et al., 2012). It provides SCF annotations for 183 verbs (an average of 12.3 SCF types per verb) obtained by annotating 250 corpus occurrences per verb with the SCF types of (de Cruys et al., 2012). The verbs represent a range of Levin classes at the top level of the hierarchy in VerbNet (Kipper-Schuler, 2005). Where a verb has more than one VerbNet class, we assign it to the one supported by the

	$\|C\| = 20, 21.6$			$\|C\| = 40, 41$			$\|C\| = 60, 58.6$			$\|C\| = 69, 77.6$			$\|C\| = 89, 97.4$		
Model	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
DPP-cluster	93.1	17.3	29.3	77.9	25.4	**38.3**	63	31.9	**42.3**	43.8	33.6	**38.1**	34.4	40.6	**37.2**
AC	67	17.8	28.2	46.6	24	31.7	40.5	29.4	34	33	34.9	33.9	24.7	41.1	30.9
SC	32.1	27.5	**29.6**	26.6	35.9	30.6	23.7	41.5	30.2	22.8	43.6	29.9	21.6	48.7	29.9

Table 1: Verb clustering evaluation for the last five iterations of our DPP-cluster model and the baseline agglomerative clustering algorithm (AC, see text for its description), and for the spectral clustering (SC) algorithm of (Sun and Korhonen, 2009) with the same number of clusters induced by DPP-cluster. $|C|$ is the number of clusters for DPP-cluster and SC (first number) and for AC (second number). The F-score performance of DPP-cluster is superior in 4 out of 5 cases.

Arg. per verb	P (DPP)	P(AC)	P (B)	P (NF)	R (DPP)	R (AC)	R (B)	R(NF)	ERR DPP	ERR AC	ERR B
$\leq$ 200 (133 verbs)	27.3	23.7	27.3	23.1	9.9	7.6	8	11.3	**3.4**	0.16	1.55
$\leq$ 600 (205 verbs)	26.5	25	27.3	22.6	14.8	11.5	11.9	16.6	**2.3**	0.50	1.1
$\leq$ 1000 (238 verbs)	24.6	23.6	25.6	21.1	17.5	13.8	14.7	19.8	**1.6**	0.42	0.95

Table 2: Performance of the Corpus Statistics SP baseline (non-filtered, NF) as well as for three filtering methods: frequency based (filter-baseline, B), DPP-cluster based (DPP) and AC cluster based (AC). P (method) and R (method) present the precision and recall of the method respectively. The error reduction ratio (ERR) is the ratio between the reduction in precision error achieved by each method and the increase in recall error (each method is compared to the NF baseline). Ratio greater than 1 means that the reduction in precision error is larger than the increase in recall error (see text for exact definition). DPP based filtering provides substantially better ratio.

highest number of member verbs. To ensure sufficient representation of each class, we collected from VerbNet the verbs for which at least one of the possible classes is represented in the 183 verbs set by at least one and at most seven verbs. This yielded 101 additional verbs which we added to the gold standard with the initial 183 verbs.

We parsed the BNC corpus with the RASP parser (Briscoe et al., 2006) and used it for feature extraction. Since 176 out of the 183 initial verbs are represented in this corpus, our final gold standard consists of 34 classes containing 277 verbs, of which 176 have SCF gold standard and has been evaluated for this task. We set the parameters of our algorithm on an held-out data, consisting of different verbs than those used in our experiments, to be $M = 10000$, $K = 20$ and $T = 10$.

Clustering Evaluation We first evaluate the quality of the clusters induced by our algorithm (DPP-cluster) compared to the gold standard VCs (table 1). To evaluate the importance of the DPP component, we compare to the performance of a version of our algorithm where everything is kept fixed except from the sampling which is done from a uniform distribution rather than from the DPP joint kernel (this model is denoted in the table with AC for agglomerative clustering) [4]. We also compare to the state-of-the-art spectral clustering method of Sun and Korhonen (2009) where our kernel matrix is used for the distance between data points (SC) [5].

We evaluated the unified cluster set induced in each iteration of our algorithm and of the AC baseline and induced the same number of clusters as in each iteration of our algorithm using the SC baseline. Since the number of clusters in each iteration is not an argument for our algorithm or for the AC baseline, the number of clusters slightly differ between the two. The AC and SC baseline results were averaged over 5 and 100 runs respectively. DPP-cluster has produced identical output across runs.

The table demonstrates the superiority of the DPP-cluster model. For four out of five conditions its F-score performance outperforms the baselines by 4.2-8.3%. Moreover, in all conditions its recall performances are substantially higher than those of the baselines (by 9.7-26.1%). Note that DPP-cluster runs for 17 iterations while the AC baseline performs only 6. We therefore evaluated only the last 5 iterations of each model [6].

SCF evaluation For this evaluation, we first

[4]Importantly, the kernel matrix L used in the agglomerative clustering process is also used by AC.

[5]Sun and Korhonen (2009) report better results than those we report for their algorithm (on a different data set). Note, however, that they used the output of a rule-based SCF system as a source of features, as opposed to our unsupervised approach.

[6]For the additional comparable iteration the result pattern is very similar to the ($C = 89, 97.4$) case in the table, and is not presented due to space limitations.

Algorithm 1 The DPP-cluster clustering algorithm. K is the size of the sampled subsets, M is the number of subsets sampled at each iteration, $\mathcal{Y}$ is the verb set, T is the number of most probable samples to be used in each iteration

Algorithm DPP-cluster :
Arguments: K,M,$\mathcal{Y}$,T
Return: cluster sets $S = \{S_1, \ldots S_n\}$
$i \leftarrow 1$
$S \leftarrow \emptyset$
while $\mathcal{Y} \neq \emptyset$ **do**
  $(L^1, S^1) \leftarrow GenKernelMatrix(\mathcal{Y}, SCF)$
  $(L^2, S^2) \leftarrow GenKernelMatrix(\mathcal{Y}, SP)$
  $(L^{12}, S^{12}) \leftarrow GenJointMat(L^1, L^2)$
  $samples \leftarrow sampleDpp(L, K, M)$
  $topSamples \leftarrow exTop(samples, T)$
  $S_i \leftarrow cluster(topSamples, L)$
  $\mathcal{Y} \leftarrow \mathcal{Y} - elements(S_i)$
  $S \leftarrow S \cup S_i$
  $i \leftarrow i + 1$
end while
$AgglomerativeClustering(S)$

Function cluster :
Arguments: topSamples,L
Return: S
$S \leftarrow \emptyset$, $topSample \leftarrow \emptyset$
$i \leftarrow 1$
while $(topSample \cap elements(S) = \emptyset)$ **do**
  $topSample \leftarrow topSamples(i)$
  $S \leftarrow m1Mapping(topSample, S)$
  $i \leftarrow i + 1$
  if $(i > size(topSamples))$ **then**
    return S
  end if
end while

built a baseline SCF lexicon based on the parsed BNC corpus. We do this by gathering the GR combinations for each of the verbs in our gold standard, assuming they are frames and gathering their frequencies. Note that this *corpus statistics* baseline is a very strong baseline that performs very similarly to (de Cruys et al., 2012), the best unsupervised SCF model we are aware of, when run on their dataset [7].

As shown in table 3 the corpus statistics baseline achieves high recall (84%) at the cost of low precision (52.5%) (similar pattern has been demonstrated for the system of de Cruys et al. (2012)). On the other extreme, two other commonly used baselines strongly prefer precision. These are the Most Frequent SCF (O'Donovan et al., 2005) which uniformly assigns to all verbs the two most frequent SCFs in general language, transitive (SUBJ-DOBJ) and intransitive (SUBJ) (and results in poor F-score), and a filtering that removes frames with low corpus frequencies (which results in low recall even when trying to provide the maximum recall for a given precision level). The task we address is therefore to improve the precision of the corpus statistics baseline in a way that does not substantially harm the F-score.

To remedy this imbalance, we apply a cluster based filtering method on top of the maximum-recall frequency filter. This filter excludes a candidate frame from a verb's lexicon only if it meets the frequency filter criterion and appears in no more than N other members of the cluster of the verb in question. The filter utilizes the clustering produced by the seventh to last iteration of DPP-cluster that contains seven clusters with approximately 30 members each. Such clustering should provide a good generalization level for the task.

We report results for moderate as well as aggressive filtering ($N = 3$ and $N = 7$ respectively). Table 3 clearly demonstrates that cluster based filtering (DPP-cluster and AC) is the only method that provides a good balance between the recall and the precision of the SCF lexicon. Moreover, the lexicon induced by this method includes a substantially higher number of frames per verb compared to the other filtering methods. While both AC and DPP-cluster still prefer recall to precision, DPP-cluster does so to a smaller extent [8]. This clearly demonstrates that the clustering serves to provide SCF acquisition with semantic information needed for improved performance.

SP evaluation We explore a variant of the pseudo-disambiguation task of Rooth et al. (1999) which has been applied to SP acquisition by a number of recent papers (e.g. (de Cruys et al., 2012)). Rooth et al. (1999) proposed to judge which of two verbs v and $\tilde{v}$ is more likely to take a given noun n as its argument. In their experiments

[7] personal communication with the authors.

[8] We show results for the maximum recall frequency filtering with precision equals to 80 or 90. When the frequency threshold is further reduced from 0.03, the same result pattern hold. We do not give a detailed description due to space limitations.

		Corpus Statistics: [P = 52.5, R = 84, F = 64.6, AF = 12.3] Most Frequent SCF: [P = 86.7, R = 22.5, F = 35.8, AF = 2]							
		Clustering Moderate				Clustering Aggressive			
Maximum Recall Frequency Threshold	Model	P	R	F	AF	P	R	F	AF
threshold = 0.03, Prec. > 80	DPP-cluster	60.8	68.3	64.3	8.7	64.1	64.2	64.2	7.7
[P=88.7,R=52.4,F=65.9,AF=4.5]	AC	58	73.2	64.6	9.7	61.3	68.9	64.7	8.6
threshold = 0.05, Prec. > 90	DPP-cluster	60.1	64.6	62.3	8.7	63.3	59.3	61.3	7.2
[P=92.3,R=44.4,F=59.9,AF=3.7]	AC	57.5	70.6	63.2	9.4	60.7	65.4	62.7	8.3

Table 3: SCF Results for the DPP-cluster model compared to the Corpus Statistics baseline, Most Frequent SCF baseline, maximum-recall frequency thresholding with the maximum threshold values that keep precision above 80 (threshold = 0.03) and above 90 (threshold = 0.05), and the AC clustering baseline. AF is the average number of frames per verb. **All methods except from cluster based filtering (DPP-cluster and AC) induce lexicons with strong recall/precision imbalance. Cluster based filtering keeps a larger number of frames in the lexicon compared to the frequency thresholding baseline, while keeping similar F-score levels**. DPP-cluster provides better recall/precision balance than AC.

the model has to choose between a pair (v, n) that appears only in the test corpus and a pair $(\tilde{v}, n)$ that appears neither in the test nor in the training corpus. Note, however, that this test only evaluates the capability of a model to distinguish a correct unseen verb-argument pair from an incorrect one, but not its capability to identify erroneous pairs when no alternative pair is presented. This last property can strongly affect the precision of the model.

We therefore propose to measure both aspects of the SP task by computing both the recall and the precision between the list of possible arguments a verb can take according to the model and the corresponding test corpus list [9].

We evaluate the value of our clustering for SP acquisition in the particularly challenging scenario of domain adaptation. For each of the verbs in our set we induce a list of possible noun direct objects from the BNC corpus and an equivalent list from the North American News Text (NANT) corpus. Following previous work (e.g. (de Cruys et al., 2012)) arguments are identified using a parser (RASP in our case). Using the verb clusters we create a filtered version of the BNC argument lexicon which includes in the noun argument list of a verb only those nouns that appear in the BNC as arguments of that verb and of one of its cluster members. For each verb we then compare the filtered as well as the non-filtered BNC induced lexicon to the NANT lexicon by computing the average recall and precision between the argument lists and then report the average scores across all verbs. We compare to a baseline which maintains only noun arguments that appear at least twice in BNC [10]. As a final measure of performance we compute the ratio between the reduction in precision error (i.e. $\frac{p_{model}-p_{baseline}}{100-p_{baseline}}$) and the increase in recall error ($\frac{r_{baseline}-r_{model}}{100-r_{model}}$).

Table 2 presents the results for verbs with up to 200, 600 and 1000 noun arguments in the training data. In all cases, the relative error reduction of the DPP cluster filter is substantially higher than that of the frequency baseline. Note that for this task the baseline AC clusters are of low quality which is reflects by an error reduction ratio of up to 0.5.

5 Conclusions and Future Work

In this paper we have presented the first unified framework for the induction of verb clusters, subcategorization frames and selectional preferences from corpus data. Our key idea is to cluster together verbs with similar SCFs and SPs and to use the resulting clusters for SCF and SP induction. To implement our idea we presented a novel method which involves constructing a product DPP model for SCFs and SPs and introduced a new algorithm that utilizes the efficient DPP sampling algorithms to cluster together verbs with similar SCFs and SPs. The induced clusters performed well in evaluation against a VerbNet -based gold standard and proved useful in improving the quality of SCFs and SPs over strong baselines.

Our results demonstrate the benefits of a uni-

[9] In principle these measures can take into account the probability assigned by the model to each argument and the corresponding test corpus frequency. In this work we compute probability-ignorant scores and keep more sophisticated evaluations for future research.

[10] we experimented with other threshold values for this baseline but the recall in those case becomes very low.

fied framework for acquiring lexical information about different aspects of verbal predicate-argument structure. Not only the acquisition of different types information (syntactic and semantic) can support and inform each other, but also a unified framework can be useful for NLP tasks and applications which require rich information about predicate-argument structure. In future work we plan to apply our approach on larger scale data sets and gold standards and to evaluate it in different domains, languages and in the context of NLP tasks such as syntactic parsing and SRL.

In addition, in our current framework SCF and SP information is used for clustering which is in turn used to improve SCF and SP quality. At this stage no further information flows from the SCF and SP models to the clustering model. A natural extension of our unified framework is to construct a joint model in which the predictions for all three tasks inform each other at all stages of the prediction process.

Acknowledgements

The work in this paper was funded by the Royal Society University Research Fellowship (UK).

References

Ivana Romina Altamirano and Laura Alonso i Alemany. 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*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL-98*.

Roberto Basili, Diego De Cao, Paolo Marocco, and Marco Pennacchiotti. 2007. Learning selectional preferences for entailment or paraphrasing rules. In *RANLP 2007*, Borovets, Bulgaria.

Rahul Bhagat, Patrick Pantel, and Eduard Hovy. 2007. Ledir: An unsupervised algorithm for learning directionality of inference rules. In *EMNLP-07*, page 161170, Prague, Czech Republic.

Akshar Bharati, Sriram Venkatapathy, and Prashanth Reddy. 2005. Inferring semantic roles using subcategorization frames and maximum entropy model. In *CoNLL-05*.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *ANLP-97*.

E.J. Briscoe, J. Carroll, and R. Watson. 2006. The second relsease of the rasp system. In *COLING/ACL interactive presentation session*.

Glenn Carroll and Mats Rooth. 1996. Valence induction with a head-lexicalized pcfg. In *EMNLP-96*.

Paula Chesley and Susanne Salmon-Alt. 2006. Automatic extraction of subcategorization frames for french. In *LREC-06*.

Kostadin Cholakov and Gertjan van Noord. 2010. Using unknown word techniques to learn known words. In *EMNLP-10*.

Hoa Trang Dang. 2004. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. Ph.D. thesis, CIS, University of Pennsylvania.

Tim Van de Cruys, Laura Rimell, Thierry Poibeau, and Anna Korhonen. 2012. Multi-way tensor factorization for unsupervised lexical acquisition. In *COLING-12*.

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

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *ACL 2007*, Prague, Czech Republic.

J. Gillenwater, A. Kulesza, and B. Taskar. 2012. Discovering diverse and salient threads in document collections. In *EMNLP-12*.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex syntax: Building a computational lexicon. In *COLNIG-94*.

G.E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Porceedings 0f NAACL-HLT-06 short papers*.

Dino Ienco, Serena Villata, and Cristina Bosco. 2008. Automatic extraction of subcategorization frames for italian. In *LREC-08*.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*.

Daisuke Kawahara and Sadao Kurohashi. 2010. Acquiring reliable predicate-argument structures from raw corpora for case frame compilation. In *LREC-10*.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, June.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *Proceddings of COLING-08*.

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*.

A. Kulesza and B. Taskar. 2010. Structured determinantal point processes. In *NIPS-10*.

A. Kulesza and B. Taskar. 2012a. k-dpps: fixed-size determinantal point processes. In *ICML-11*.

A. Kulesza and B. Taskar. 2012b. Learning determinantal point processes. In *UAI-12*.

Alex Kulesza and Ben Taskar. 2012c. Determinantal point processes for machine learning. In *arXiv:1207.6083*.

A. Kulesza. 2012. *Learning with determinantal point processes*. Ph.D. thesis, CIS, University of Pennsylvania.

Alessandro Lenci, Barbara McGillivray, Simonetta Montemagni, and Vito Pirrelli. 2008. Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *LREC-08*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation.* Chicago, IL.

Jianguo Li and Chris Brew. 2008. Which are the best features for automatic verb classification. In *ACL-08*.

Tom Lippincott, Anna Korhonen, and Diarmuid Ó Séaghdha. 2012. Learning syntactic verb frames using graphical models. In *ACL-12*, Jeju, Korea.

Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem: A large subcategorization lexicon for French verbs. In *LREC-08*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Alessandro Moschitti and Roberto Basili. 2005. Verb subcategorization kernels for automatic semantic labeling. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*.

Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. 2005. Large-scale induction and evaluation of lexical resources from the penn-ii and penn-iii treebanks. *Computational Linguistics*, 31:328–365.

Diarmuid Ó Séaghdha and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *EMNLP-11*, Edinburgh, UK.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *ACL-10*, Uppsala, Sweden.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *ACL-07*.

Joseph Reisinger and Raymond Mooney. 2011. Cross-cutting models of lexical semantics. In *EMNLP-11*, Edinburgh, UK.

Alan Ritter and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *ACL-10*.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *ACL-99*.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

S. Schulte im Walde, C. Hying, C. Scheible, and H. Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *ACL-08*, pages 496–504.

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

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *CICLING-05*.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *EMNLP-09*, Singapore.

Lin Sun and Anna Korhonen. 2011. Hierarchical verb clustering using graph factorization. In *EMNLP-11*.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. *Lecture Notes in Computer Science*, 4919(16).

Robert Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *EMNLP-04*.

Stefan Thater, Hagen Furstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *ACL-10*, Uppsala, Sweden.

Tim Van de Cruys. 2009. A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the workshop on Geometric Models for Natural Language Semantics (GEMS)*.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*.

2008. *Robustness and generalization of role sets: PropBank vs. VerbNet*.

Benat Zapirain, Eneko Agirre, and Lluis Marquex. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *ACL-IJCNLP-09*, Singapore.

Guangyou Zhou, Jun Zhao, Kang Liu, and Li Cai. 2011. Exploiting web-derived selectional preference to improve statistical dependency parsing. In *ACL-11*, Portland, OR.