Perceptually grounded selectional preferences

Ekaterina Shutova
Computer Laboratory
University of Cambridge, UK
es407@cam.ac.uk

Niket Tandon
Max Planck Institute
for Informatics, Germany
ntandon@mpi-inf.mpg.de

Gerard de Melo
IIIS
Tsinghua University, China
gdm@demelo.org

Abstract

Selectional preferences (SPs) are widely used in NLP as a rich source of semantic information. While SPs have been traditionally induced from textual data, human lexical acquisition is known to rely on both linguistic and perceptual experience. We present the first SP learning method that simultaneously draws knowledge from text, images and videos, using image and video descriptions to obtain visual features. Our results show that it outperforms linguistic and visual models in isolation, as well as the existing SP induction approaches.

1 Introduction

Selectional preferences (SPs) are the semantic constraints that a predicate places onto its arguments. This means that certain classes of entities are more likely to fill the predicate's argument slot than others. For instance, while the sentences "*The authors wrote a new paper.*" and "*The cat is eating your sausage!*" sound natural and describe plausible real-life situations, the sentences "*The carrot ate the keys.*" and "*The law sang a driveway.*" appear implausible and difficult to interpret, as the arguments do not satisfy the verbs' common preferences. SPs provide generalisations about word meaning and use and find a wide range of applications in natural language processing (NLP), including word sense disambiguation (Resnik, 1997; McCarthy and Carroll, 2003; Wagner et al., 2009), resolving ambiguous syntactic attachments (Hindle and Rooth, 1993), semantic role labelling (Gildea and Jurafsky, 2002; Zapirain et al., 2010), natural language inference (Zanzotto et al., 2006; Pantel et al., 2007), and figurative language processing (Fass, 1991; Mason, 2004; Shutova et al., 2013; Li et al., 2013). Automatic acquisition of SPs from linguistic data has thus become an active area of research. The community has investigated a range of techniques to tackle data sparsity and to perform generalisation from observed arguments to their underlying types, including the use of WordNet synsets as SP classes (Resnik, 1993; Li and Abe, 1998; Clark and Weir, 1999; Abney and Light, 1999; Ciaramita and Johnson, 2000), word clustering (Rooth et al., 1999; Bergsma et al., 2008; Sun and Korhonen, 2009), distributional similarity metrics (Erk, 2007; Peirsman and Padó, 2010), latent variable models (Ó Séaghdha, 2010; Ritter et al., 2010), and neural networks (Van de Cruys, 2014).

Little research, however, has been concerned with the sources of knowledge that underlie the learning of SPs. There is ample evidence in cognitive and neurolinguistics that our concept learning and semantic representation are grounded in perception and action (Barsalou, 1999; Glenberg and Kaschak, 2002; Barsalou, 2008; Aziz-Zadeh and Damasio, 2008). This suggests that word meaning and relational knowledge are acquired not only from linguistic input but also from our experiences in the physical world. Multi-modal models of word meaning have thus enjoyed a growing interest in semantics (Bruni et al., 2014), outperforming purely text-based models in tasks such as similarity estimation (Bruni et al., 2014; Kiela et al., 2014), predicting compositionality (Roller and Schulte im Walde, 2013), and concept categorization (Silberer and Lapata, 2014). However, to date these approaches relied on low-level image features such as color histograms or SIFT keypoints to represent the meaning of isolated words. To the best of our knowledge, there has not yet been a multimodal semantic approach performing extraction of

predicate-argument relations from visual data. In this paper, we propose the first SP model integrating information about predicate-argument interactions from text, images, and videos. We expect it to outperform purely text-based models of SPs, which suffer from two problems: topic bias and figurative uses of words. Such bias stems from the fact that we typically write about abstract topics and events, resulting in high coverage of abstract senses of words and comparatively lower coverage of the original physical senses (Shutova, 2011). For instance, the verb *cut* is used predominantly in the domains of economics and finance and its most frequent direct objects are *cost* and *price*, according to the British National Corpus (BNC) (Burnard, 2007). Predicate-argument distributions acquired from text thus tend to be skewed in favour of abstract domains and figurative uses, inadequately reflecting our daily experiences with *cutting*, which guide human acquisition of meaning. Integrating predicate-argument relations observed in the physical world (in the form of image and video descriptions) with the more abstract text-based relations is likely to yield a more realistic semantic model, with real prospects of improving the performance of NLP applications that rely on SPs.

We use the BNC as an approximation of linguistic knowledge and a large collection of tagged images and videos from Flickr (`www.flickr.com`) as an approximation of perceptual knowledge. The human-annotated labels that accompany media on Flickr enable us to acquire predicate-argument co-occurrence information. Our experiments focus on verb preferences for their subjects and direct objects. In summary, our method (1) performs word sense disambiguation and part-of-speech (PoS) tagging of Flickr tag sequences to extract verb-noun co-occurrence; (2) clusters nouns to induce SP classes using linguistic and visual features; (3) quantifies the strength of preference of a verb for a given class by interpolating linguistic and visual SP distributions. We investigate the impact of perceptual information at different levels – from none (purely text-based model) to 100% (purely visual model). We evaluate our model directly against a dataset of human plausibility judgements of verb-noun pairs, as well as in the context of a semantic task: metaphor interpretation. Our results show that the interpolated model combining linguistic and visual relations outperforms the purely linguistic model in both evaluation settings.

2 Related work

2.1 Selectional preference induction

The widespread interest in automatic acquisition of SPs was triggered by the work of Resnik (1993), who treated SPs as probability distributions over all potential arguments of a predicate, rather than a single argument class assigned to the predicate. The original study used WordNet to define SP classes and to map the words in the corpus to those classes. Since then, the field has moved toward automatic induction of SP classes from corpus data. Rooth et al. (1999) presented a probabilistic latent variable model of verb preferences. In their approach, verb-argument pairs are generated from a latent variable, which represents a cluster of verb-argument interactions. The latent variable distribution and the probabilities that a latent variable generates the verb and the argument are learned from the data using Expectation Maximization (EM). The latent variables enable the model to recognise previously unseen verb-argument pairs. Ó Séaghdha (2010) and Ritter et al. (2010) similarly model SPs within a latent variable framework, but use Latent Dirichlet Allocation (LDA) to learn the probability distributions, for single-argument and multi-argument preferences respectively.

Padó et al. (2007) and Erk (2007) used similarity metrics to approximate selectional preference classes. Their underlying hypothesis is that a predicate-argument combination (p, a) is felicitous if the predicate p is frequently observed in the data with the arguments a' similar to a. The systems compute similarities between distributional representations of arguments in a vector space.

Bergsma et al. (2008) trained an SVM classifier to discriminate between felicitous and infelicitous verb-argument pairs. Their training data consisted of observed verb-argument pairs (positive examples) with unobserved, randomly-generated ones (negative examples). They classified nominal arguments of verbs, using their verb co-occurrence probabilities and information about their semantic classes as features. Bergsma and Goebel (2011) extended this method by incorporating image-driven noun features. They extract color and SIFT keypoint features from images found for a particular noun via Google image searches and add them to the feature vectors to classify nouns as felicitous or infelicitous arguments of a given verb. This method is the closest in spirit to ours and the only one so far to investigate the relevance of visual fea-

tures to lexical preference learning. However, our work casts the problem in a different framework: rather than relying on low-level visual properties of nouns in isolation, we explicitly model interactions of predicates and arguments within an image or a video frame.

Van de Cruys (2014) recently presented a deep learning approach to SP acquisition. He trained a neural network to discriminate between felicitous and infelicitous arguments using the data constructed of positive (observed) and negative (randomly-generated) examples for training. The network weights were optimized by requiring the model to assign a higher score to an observed pair than to the unobserved one by a given margin.

2.2 Multi-modal methods in semantics

Previous work has used multimodal data to determine distributional similarity or to learn multimodal embeddings that project multiple modalities into the same vector space. Some studies rely on extensions of LDA to obtain correlations between words and visual features (Feng and Lapata, 2010; Roller and Schulte im Walde, 2013). Bruni et al. (2012) integrated visual features into distributional similarity models using simple vector concatenation. Instead of generic visual features, Silberer et al. (2013) relied on supervised learning to train 412 higher-level visual attribute classifiers.

Applications of multimodal embeddings include zero-shot object detection, i.e. recognizing objects in images without training data for the object class (Socher et al., 2013; Frome et al., 2013; Lazaridou et al., 2014), and automatic generation of image captions (Kulkarni et al., 2013), video descriptions (Rohrbach et al., 2013), or tags (Srivastava et al., 2014). Other applications of multimodal data include language modeling (Kiros et al., 2014) and knowledge mining from images (Chen et al., 2013; Divvala et al., 2014). Young et al. (2014) apply simplification rules to image captions, showing that the resulting hierarchy of mappings between natural language expressions and images can be used for entailment tasks.

3 Experimental data

Textual data. We extract linguistic features for our model from the BNC. In particular, we parse the corpus using the RASP parser (Briscoe et al., 2006) and extract subject–verb and verb–object relations from its dependency output. These relations are then used as features for clustering to obtain SP classes, as well as to quantify the strength of association between a particular verb and a particular argument class.

Visual data. For the visual features of our model, we mine the Yahoo! Webscope Flickr-100M dataset (Shamma, 2014). Flickr-100M contains 99.3 million images and 0.7 million videos with language tags annotated by users, enabling us to generalise SPs at a large scale. The tags reflect how humans describe objects and actions from a visual perspective. We first stem the tags and remove words that are absent in WordNet (typically named entities and misspellings), then identify their PoS based on their visual context and extract verb–noun co-occurrences.

4 Identifying visual verb-noun co-occurrence

In the Flickr-100M dataset, tags are assigned to images and videos in the form of sets of words, rather than grammatically coherent sentences. However, the roles that individual words play are still discernible from their visual context, as manifested by the other words in a given set. In order to identify verbs and nouns co-occurring in the same images, we propose a *list sense disambiguation* method that first maps each word to a set of possible WordNet senses (accompanied by PoS information) and then performs a joint optimization on the space of candidate word senses, such that their overall similarity is maximized. This amounts to assigning those senses and PoS tags to the words in the set that best fit together.

For a given word i and one of its candidate WordNet senses j, we consider an assignment variable x_{ij} and compute a sense frequency-based prior for it as $P_{ij} = \frac{1}{1+R}$, where R is the WordNet rank of the sense. We then compute a similarity score $S_{ij,i'j'}$ between all pairs of sense choices for two words i,i' and their respective candidate senses j,j'. For these, we rely on WordNet's taxonomic path-based similarities (Pedersen et al., 2004) in the case of noun-noun sense pairs, the Adapted Lesk similarity measure for adjective-adjective pairs, and finally, WordNet verb-groups and VerbNet class membership (Kipper-Schuler, 2005) for verb-verb pairs. Note that even parts of speech that are disregarded later on can still be helpful at this stage, as we aim at a joint optimization over all words. After the similarities have been obtained for all rel-

evant sense pairs, we maximize the coherence of the senses of the words in the set as an Integer Linear Program, using the Gurobi Optimizer (Gurobi Optimization, 2014) and solving

maximize
$\sum_i P_{ij} x_{ij} + \sum_{ij} \sum_{i'j'} S_{ij,i'j'} B_{ij,i'j'}$
subject to
$\sum_j x_{ij} \leq 1 \; \forall i, \;\; x_{ij} \in \{0,1\} \; \forall i,j,$
$B_{ij,i'j'} \leq x_{ij}, \;\; B_{ij,i'j'} \leq x_{i'j'},$
$B_{ij,i'j'} \in \{0,1\} \;\; \forall i,j,i'j'.$

The binary variables $B_{ij,i'j'}$ are 1 iff $x_{ij} = 1$ and $x_{i'j'} = 1$, indicating that both senses were simultaneously chosen. The optimizer disambiguates the input words by selecting sense tuples $x_{1j}, x_{2j}, \ldots$, from which we can directly obtain the corresponding PoS information. Verb-noun co-occurrence information is then extracted from the PoS-tagged sets.

5 Selectional preference model

5.1 Acquisition of argument classes

To address the issue of data sparsity, we generalise selectional preferences over argument classes, as opposed to individual arguments. We obtain SP classes by means of spectral clustering of nouns with lexico-syntactic features, which has been shown effective in previous lexical classification tasks (Brew and Schulte im Walde, 2002; Sun and Korhonen, 2009).

Spectral clustering partitions the data, relying on a similarity matrix that records similarities between all pairs of data points. We use *Jensen-Shannon divergence* to measure the similarity between feature vectors for two nouns, w_i and w_j, defined as follows:

$$d_{\mathrm{JS}}(w_i, w_j) = \frac{1}{2} d_{\mathrm{KL}}(w_i || m) + \frac{1}{2} d_{\mathrm{KL}}(w_j || m), \quad (1)$$

where d_{KL} is the Kullback-Leibler divergence, and m is the average of w_i and w_j. We construct the similarity matrix S computing similarities S_{ij} as $S_{ij} = \exp(-d_{\mathrm{JS}}(w_i, w_j))$. The matrix S then encodes a similarity graph G (over our nouns), where S_{ij} are the adjacency weights. The clustering problem can then be defined as identifying the optimal partition, or *cut*, of the graph into clusters, such that the intra-cluster weights are high and the inter-cluster weights are low. We use the multiway normalized cut (MNCut) algorithm of Meila and Shi (2001) for this purpose. The algorithm transforms S into a stochastic matrix P containing transition probabilities between the vertices in the graph as

$$P = D^{-1} S, \quad (2)$$

where the degree matrix D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} S_{ij}$. It then computes the K leading eigenvectors of P, where K is the desired number of clusters. The graph is partitioned by finding approximately equal elements in the eigenvectors using a simpler clustering algorithm, such as *k-means*. Meila and Shi (2001) have shown that the partition I derived in this way minimizes the MNCut criterion:

$$\mathrm{MNCut}(I) = \sum_{k=1}^{K} (1 - P(I_k \rightarrow I_k | I_k)), \quad (3)$$

which is the sum of transition probabilities across different clusters. Since k-means starts from a random cluster assignment, we run the algorithm multiple times and select the partition that minimizes the cluster distortion, i.e. distances to cluster centroid.

We cluster nouns using linguistic and visual features in two independent experiments.

Clustering with linguistic features: We first cluster the 2,000 most frequent nouns in the BNC, using their grammatical relations as features. The features consist of verb lemmas appearing in the subject, direct object and indirect object relations with the given nouns in the RASP-parsed BNC, indexed by relation type. The feature vectors are first constructed from the corpus counts, and subsequently normalized by the sum of the feature values.

Clustering with visual features: We also cluster the 2,000 most frequent nouns in the Flickr data. Since our goal is to create argument classes for verb preferences, we extract co-occurrence features that map to verb-noun relations from PoS-disambiguated image tags. We use the verb lemmas co-occurring with the noun in the same images and videos as features for clustering. The feature values are again normalised by their sum.

SP classes: Example clusters produced using linguistic and visual features are shown in Figures 1 and 2. Our cluster analysis reveals that the image-derived clusters tend to capture scene-like relations (e.g. *beach* and *ocean*; *guitar* and *concert*), as opposed to types of entities, yielded by the linguistic features and better suited to generalise over

desire hostility anxiety passion doubt fear curiosity enthusiasm impulse instinct emotion feeling suspicion
official officer inspector journalist detective constable police policeman reporter
book statement account draft guide advertisement document report article letter

Figure 1: Clusters obtained using linguistic features

pilot aircraft plane airline landing flight wing arrival departure airport
concert festival music guitar alternative band instrument audience event performance rock benjamin
cost benefit crisis debt credit customer consumer

Figure 2: Clusters obtained using visual features

predicate-argument structure. In addition, the image features tend to be sparse for abstract concepts, reducing both the quality and the coverage of abstract clusters. We thus use the noun clusters derived with linguistic features as an approximation of SP classes.

5.2 Quantifying selectional preferences

Once the SP classes have been obtained, we need to quantify the strength of association of a given verb with each of the classes. We adopt an information theoretic measure proposed by Resnik (1993) for this purpose. Resnik first measures *selectional preference strength* (SPS) of a verb in terms of Kullback-Leibler divergence between the distribution of noun classes occurring as arguments of this verb, $p(c|v)$, and the prior distribution of the noun classes, $p(c)$.

$$\text{SPS}_R(v) = \sum_c p(c|v) \log \frac{p(c|v)}{p(c)}, \qquad (4)$$

where R is the grammatical relation for which SPs are computed. SPS measures how strongly the predicate constrains its arguments. Selectional association of the verb with a particular argument class is then defined as a relative contribution of that argument class to the overall SPS of the verb.

$$\text{Ass}_R(v, c) = \frac{1}{\text{SPS}_R(v)} p(c|v) \log \frac{p(c|v)}{p(c)} \qquad (5)$$

We use this measure to quantify verb SPs based on linguistic and visual co-occurrence information. We first extract verb-subject and verb-direct object relations from the RASP-parsed BNC, map the argument heads to SP classes and quantify selectional association of a given verb with each SP class, thus acquiring its *base* preferences. Since visual verb-noun co-occurrences do not contain information about grammatical relations, we rely on linguistic data to provide a set of base arguments of the verb for a given grammatical relation. We then interpolate the verb-argument probabilities from linguistic and visual models for the base arguments of the verb, thus preserving information about grammatical relations.

5.3 Linguistic and visual model interpolation

We investigate two model interpolation techniques: simple linear interpolation and predicate-driven linear interpolation.

Linear interpolation combines information from component models by computing a weighted average of their probabilities. The interpolated probability of an event e is derived as $p^{\text{LI}}(e) = \sum_i \lambda_i p_i(e)$, where $p_i(e)$ is the probability of e in the model i and λ_i is the interpolation weight defined such that $\sum_i \lambda_i = 1$; and $\lambda_i \in [0, 1]$. In our experiments, we interpolate the probabilities $p(c)$ and $p(c|v)$ in the linguistic (LM) and visual (VM) models, as follows:

$$p^{\text{LI}}(c) = \lambda_{\text{LM}} p_{\text{LM}}(c) + \lambda_{\text{VM}} p_{\text{VM}}(c) \qquad (6)$$

$$p^{\text{LI}}(c|v) = \lambda_{\text{LM}} p_{\text{LM}}(c|v) + \lambda_{\text{VM}} p_{\text{VM}}(c|v) \qquad (7)$$

We experiment with a number of parameter settings for λ_{LM} and λ_{VM}.

Predicate-driven linear interpolation derives predicate-specific interpolation weights directly from the data, as opposed to pre-setting them universally for all verbs. For each predicate v, we compute the interpolation weights based on its prominence in the respective corpus, as follows:

$$\lambda_i(v) = \frac{\text{rel}_i(v)}{\sum_k \text{rel}_k(v)}, \qquad (8)$$

where rel is the relevance function of model i for verb v, computed as its relative frequency in the respective corpus: $\text{rel}_i(v) = \frac{f_i(v)}{\sum_V f_i(v)}$. The interpolation weights for LM and VM are then computed as

$$\lambda_{\text{LM}}(v) = \frac{\text{rel}_{\text{LM}}(v)}{\text{rel}_{\text{LM}}(v) + \text{rel}_{\text{VM}}(v)} \qquad (9)$$

$$\lambda_{\text{VM}}(v) = \frac{\text{rel}_{\text{VM}}(v)}{\text{rel}_{\text{LM}}(v) + \text{rel}_{\text{VM}}(v)}. \qquad (10)$$

The motivation for this approach comes from the fact that not all verbs are represented equally well in linguistic and visual data. For instance, while concrete verbs, such as *run, push* or *throw*, are more likely to be prominent in visual data, abstract verbs, such as *understand* or *speculate*, are best

represented in text. Relative linguistic and visual frequencies of a verb provide a way to estimate the relevance of linguistic and visual features to its SP learning.

6 Direct evaluation and data analysis

We evaluate the predicate-argument scores assigned by our models against a dataset of human plausibility judgements of verb-direct object pairs collected by Keller and Lapata (2003). Their dataset is balanced with respect to the frequency of verb-argument relations, as well as their plausibility and implausibility, thus creating a realistic SP evaluation task. Keller and Lapata selected 30 predicates and matched each of them to three arguments from different co-occurrence frequency bands according to their BNC counts, e.g. *divert attention* (high frequency), *divert water* (medium) and *divert fruit* (low). This constituted their dataset of *Seen* verb-noun pairs, 90 in total. Each of the predicates was then also paired with three randomly selected arguments with which it did not occur in the BNC, creating the *Unseen* dataset. The pairs in both datasets were then rated for their plausibility by 27 human subjects, and their judgements were aggregated into a gold standard. We compare the verb-argument scores generated by our linguistic (LSP), visual (VSP) and interpolated (ISP) SP models against these two datasets in terms of Pearson correlation coefficient, r, and Spearman rank correlation coefficient, ρ. The selectional association score of the cluster to which a given noun belongs is taken to represent the preference score of the verb for this noun. If a noun is not present in our argument clusters, we match it to its nearest cluster, as determined by its distributional similarity to the cluster centroid in terms of Jensen-Shannon divergence.

We first compare LSP, VSP and ISP with static and predicate-driven interpolation weights. The results, presented in Table 1, demonstrate that the interpolated model outperforms both LSP and VSP used on their own. The best performance is attained with the static interpolation weights of $\lambda_{\text{LM}} = 0.8$ ($r = 0.540; \rho = 0.728$) and $\lambda_{\text{LM}} = 0.9$ ($r = 0.548; \rho = 0.699$). This suggests that while linguistic input plays a crucial role in SP induction (by providing both semantic and syntactic information), visual features further enhance the quality of SPs, as we expected. Figure 3 shows LSP- and VSP-acquired direct object preferences of the verb

	Seen		Unseen	
	r	ρ	r	ρ
VSP	0.180	0.126	0.118	0.132
ISP: $\lambda_{\text{LM}} = 0.1$	0.279	0.532	0.220	0.371
ISP: $\lambda_{\text{LM}} = 0.2$	0.349	0.556	0.278	0.411
ISP: $\lambda_{\text{LM}} = 0.3$	0.385	0.558	0.305	0.423
ISP: $\lambda_{\text{LM}} = 0.4$	0.410	0.571	0.320	0.428
ISP: $\lambda_{\text{LM}} = 0.5$	0.448	0.579	0.329	0.430
ISP: $\lambda_{\text{LM}} = 0.6$	0.461	0.591	0.330	0.431
ISP: $\lambda_{\text{LM}} = 0.7$	0.523	0.713	0.335	0.431
ISP: $\lambda_{\text{LM}} = 0.8$	0.540	**0.728**	0.339	0.430
ISP: $\lambda_{\text{LM}} = 0.9$	**0.548**	0.699	0.342	0.429
ISP: Predicate-driven	0.476	0.597	0.391	0.551
LSP	0.512	0.688	**0.412**	**0.559**

Table 1: Model comparison on the plausibility data of Keller and Lapata (2003)

LSP: (1) 0.309 expenditure cost risk expense emission budget spending; **(2)** 0.201 dividend price rate premium rent rating salary wages; **(3)** 0.088 employment investment growth supplies sale import export production [..]

ISP predicate-driven $\lambda_{\text{LM}} = 0.65$
(1) 0.346 expenditure cost risk expense emission budget spending; **(2)** 0.211 dividend price rate premium rent rating salary wages; **(3)** 0.126 tail collar strand skirt trousers hair curtain sleeve

VSP: (1) 0.224 tail collar strand skirt trousers hair curtain sleeve; **(2)** 0.098 expenditure cost risk expense emission budget spending; **(3)** 0.090 management delivery maintenance transport service housing [..]

Figure 3: Top three direct object classes for *cut* and their association scores, assigned by different models

cut, as well as the effects of merging the features in the interpolated model – the verbs' experiential arguments (e.g. *hair* or *fabric*) are emphasized by the visual features.

However, the model based on visual features alone performs poorly on the dataset of Keller and Lapata (2003). This is partly explained by the fact that a number of verbs in this dataset are abstract verbs, whose visual representations in the Flickr data are sparse. In addition, VSP (as other visual models used in isolation from text) is not syntax-aware and is unable to discriminate between different types of semantic relations. VSP thus acquires sets of verb-argument relations that are closer in nature to scene descriptions and semantic frames than to lexico-syntactic paradigms. Figure 4 shows the differences between linguistic and visual arguments of the verb *kill* ranked by LSP and VSP. While LSP produces mainly semantic objects of *kill*, VSP output contains other types of arguments, such as *weapon* (instrument) and *death* (consequence).

Taking the argument classes produced by the linguistic model as a basis and then re-ranking

LSP: **(1)** 0.523 girl other woman child person people; **(2)** 0.164 fleet soldier knight force rebel guard troops crew army pilot; **(3)** 0.133 sister daughter parent relative lover cousin friend wife mother husband brother father; **(4)** 0.048 being species sheep animal creature horse baby human fish male lamb bird rabbit [..]; **(5)** 0.045 victim bull teenager prisoner hero gang enemy rider offender youth killer thief [..]

VSP: **(1)** 0.180 defeat fall death tragedy loss collapse decline [..]; **(2)** 0.141 girl other woman child person people; **(3)** 0.128 abuse suicide killing offence murder breach crime; **(4)** 0.113 handle weapon horn knife blade stick sword [..]; **(5)** 0.095 victim bull teenager prisoner hero gang enemy rider offender youth killer thief [..]

Figure 4: Top five arguments of *kill* and their association scores, assigned by LSP and VSP

(1) 0.442 drink coffee champagne pint wine beer; **(2)** 0.182 mixture dose substance drug milk cream alcohol chemical [..]; **(3)** 0.091 girl other woman child person people; **(4)** 0.053 sister daughter parent relative lover cousin friend wife mother husband brother father; **(5)** 0.050 drop tear sweat paint blood water juice

Figure 5: Error analysis: Mixed subjects and direct objects of *drink*, assigned by the predicate-driven ISP

them to incorporate visual statistics helps to avoid the above problem for the interpolated models, whose output corresponds to grammatical relations. However, static interpolation weights (emphasizing linguistic features over the visual ones for all verbs equally) outperformed the predicate-driven interpolation technique, attaining correlations of $r = 0.548$ and $r = 0.476$ respectively. This is mainly due to the fact that some verbs are over-represented in the visual data (e.g. the predicate-driven interpolation weight for the verb *drink* is $\lambda_{\text{LM}} = 0.08$). As a result, candidate argument classes (selected based on syntactically-parsed linguistic input) are ranked predominantly based on visual statistics. This makes it possible to emphasize incorrectly parsed arguments (such as subject relations in the direct object SP distribution and vice versa). The predicate-driven ISP output for direct object SPs of *drink*, for instance, contains a mixture of subject and direct object classes, as shown in Figure 5. Using a static model with a high λ_{LM} weight helps to avoid such errors and, therefore, leads to a better performance.

In order to investigate the composition of the visual and linguistic datasets, we assess the average level of concreteness of the verbs and nouns present in the datasets. We use the concreteness ratings from the MRC Psycholinguistic Database (Wilson, 1988) for this purpose. In this database, nouns and

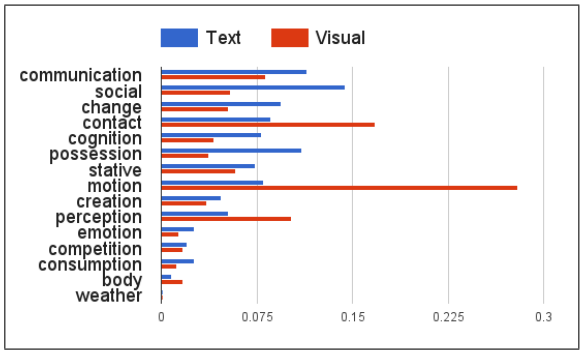

Figure 6: WordNet top level class distributions for verbs in the visual and textual corpora

	Seen		Unseen	
	r	ρ	r	ρ
Rooth et al. (1999)*	0.455	0.487	0.479	0.520
Padó et al. (2007)*	0.484	0.490	0.398	0.430
O'Seaghdha (2010)	0.520	0.548	**0.564**	**0.605**
VSP	0.180	0.126	0.118	0.132
ISP (best)	**0.548**	**0.699**	0.342	0.429
LSP	0.512	0.688	0.412	0.559

Table 2: Comparison to other SP induction methods. * Results reported in O'Seaghdha (2010).

verbs are rated for concreteness on a scale from 100 (highly abstract) to 700 (highly concrete). We map the verbs and nouns in our textual and visual corpora to their MRC concreteness scores. We then calculate a dataset-wide concreteness score as an average of the concreteness scores of individual verbs and nouns weighted by their frequency in the respective corpus. The average concreteness scores in the visual dataset were 506.4 (nouns) and 498.1 (verbs). As expected, they are higher than the respective scores in the textual data: 433.1 (nouns) and 363.4 (verbs). In order to compare the types of actions that are common in each of the datasets, we map the verbs to their corresponding top level classes in WordNet. Figure 6 shows the comparison of prominent verb classes in visual and textual data. One can see from the Figure that the visual dataset is well suited for representing *motion, perception* and *contact*, while abstract verbs related to e.g. *communication, cognition, possession* or *change* are more common in textual data.

We also compare the performance of our models to existing SP induction methods: the EM-based clustering method of Rooth et al. (1999), the vector space similarity-based method of Padó et al. (2007) and the LDA topic modelling approach of Ó Séaghdha (2010)[1]. The best ISP configuration

[1]Since Rooth et al.'s (1999) and Padó et al.'s (2007) models were not originally evaluated on the same dataset, we use the

($\lambda_{\text{LM}} = 0.9$) outperforms all of these methods, as well as our own LSP, on the *Seen* dataset, confirming the positive contribution of visual features. However, it achieves less success on the *Unseen* data, where the methods of Ó Séaghdha (2010) and Rooth et al. (1999) are leading. This result speaks in favour of latent variable models for acquisition of SP estimates for rarely attested predicate-argument pairs. In turn, this suggests that integrating our ISP model (that currently outperforms others on more common pairs) with such techniques is likely to improve SP prediction across frequency bands.

7 Task-based evaluation

In order to investigate the applicability of perceptually grounded SPs in wider NLP, we evaluate them in the context of an external semantic task – that of metaphor interpretation. Since metaphor is based on transferring imagery and knowledge across domains – typically from more familiar domains of physical experiences to the sphere of vague and elusive abstract thought – metaphor interpretation provides an ideal framework for testing perceptually grounded SPs. Our experiments rely on the metaphor interpretation method of Shutova (2010), in which text-derived SPs are a central component of the system. We replace the SP component with our LSP and ISP ($\lambda_{\text{LM}} = 0.8$) models and compare their performance in the context of metaphor interpretation.

Shutova (2010) defined metaphor interpretation as a paraphrasing task, where literal paraphrases for metaphorical expressions are derived from corpus data using a set of statistical measures. For instance, their system interprets the metaphor "a carelessly *leaked* report" as "a carelessly disclosed report". Focusing on metaphorical verbs in subject and direct object constructions, Shutova first applies a maximum likelihood model to extract and rank candidate paraphrases for the verb given the context, as follows:

$$P(i, w_1, ..., w_N) = \frac{\prod_{n=1}^{N} f(w_n, i)}{(f(i))^{N-1} \cdot \sum_k f(i_k)}, \quad (11)$$

where $f(i)$ is the frequency of the paraphrase on its own and $f(w_n, i)$ the co-occurrence frequency of the paraphrase with the context word w_n. This model favours paraphrases that match the given context best. These candidates are then filtered based on the presence of shared features with the metaphorical verb, as defined by their location and distance in the WordNet hierarchy. All the candidates that have a common hypernym with the metaphorical verb within three levels of the WordNet hierarchy are selected. This results in a set of paraphrases retaining the meaning of the metaphorical verb. However, some of them are still figuratively used. Shutova further applies an SP model to discriminate between figurative and literal paraphrases, treating a strong selectional preference fit as a likely indicator of literalness. The candidates are re-ranked by the SP model, emphasizing the verbs whose preferences the noun in the context matches best. We use LSP and ISP scores to perform this re-ranking step.

We evaluate the performance of our models on this task using the metaphor paraphrasing gold standard of Shutova (2010). The dataset consists of 52 verb metaphors and their human-produced literal paraphrases. Following Shutova, we evaluate the performance in terms of mean average precision (MAP), which measures the ranking quality of GS paraphrases across the dataset. MAP is defined as follows:

$$\text{MAP} = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ji},$$

where M is the number of metaphorical expressions, N_j is the number of correct paraphrases for the metaphorical expression j, P_{ji} is the precision at each correct paraphrase (the number of correct paraphrases among the top i ranks). As compared to the gold standard, ISP attains a MAP score of 0.65, outperforming both the LSP (MAP = 0.62) and the original system of Shutova (2010) (MAP = 0.62), demonstrating the positive contribution of visual features.

8 Conclusion

We have presented the first SP induction method that simultaneously draws knowledge from text, images and videos. Our experiments show that it outperforms linguistic and visual models in isolation, as well as the previous approaches to SP learning. We believe that this model has a wide applicability in NLP, where many systems already rely on automatically induced SPs. It can also benefit image caption generation systems, which

results for their re-implementation reported by O'Seaghdha (2010), who conducted a comprehensive evaluation of SP models on the plausibility data of Keller and Lapata (2003).

typically focus on objects rather than actions, by providing information about predicate-argument structure.

In the future, it would be interesting to derive the information about predicate-argument relations from low-level visual features directly. However, to our knowledge, reliably mapping images to actions (i.e. verbs) at a large-scale is still a challenging task. Human-annotated image and video descriptions allow us to investigate what types of verb–noun relations are in principle present in the visual data and the ways in which they are different from the ones found in text. Our results show that visual data is better suited for capturing physical properties of concepts as well as containing relations not explicitly described in text.

The presented interpolation techniques are also applicable outside multi-modal semantics. For instance, they can be generalised to acquire SPs from unbalanced corpora of different sizes (e.g. for languages lacking balanced corpora) or to perform domain adaptation of SPs. In the future, we would like to apply SP interpolation to multilingual SP learning, i.e. integrating data from multiple languages for more accurate SP induction and projecting universal semantic relations to low-resource languages. It is also interesting to investigate SP learning at the level of semantic predicates (e.g. automatically inducing FrameNet-style frames), where combining the visual and linguistic knowledge is likely to outperform text-based models on their own.

Acknowledgements

Ekaterina Shutova's research is funded by the University of Cambridge and the Leverhulme Trust Early Career Fellowship. Gerard de Melo's work is funded by China 973 Program Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003, 61450110088. We are grateful to the ACL reviewers for their insightful feedback.

References

Steven Abney and Marc Light. 1999. Hiding a Semantic Hierarchy in a Markov Model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, ACL*, pages 1–8.

Lisa Aziz-Zadeh and Antonio Damasio. 2008. Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology – Paris*, 102(1-3).

Lawrence W. Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *Proceedings of RANLP*.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of EMNLP 2008*, EMNLP '08, pages 59–68, Honolulu, Hawaii.

Chris Brew and Sabine Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proceedings of EMNLP*, pages 117–124.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in Technicolor. In *Proceedings of ACL 2012*, pages 136–145, Jeju Island, Korea, July. ACL.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting Visual Knowledge from Web Data. In *Proceedings of ICCV 2013*.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of COLING 2000*, pages 187–193.

Stephen Clark and David Weir. 1999. An iterative approach to estimating frequencies over a semantic hierarchy. In *Proceedings of EMNLP/VLC 1999*, pages 258–265.

Santosh Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of CVPR 2014*.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of ACL 2007*.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of NAACL 2010*, pages 91–99. ACL.

Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of NIPS 2013*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28.

Arthur M. Glenberg and Michael P. Kaschak. 2002. Grounding language in action. *Psychonomic Bulletin and Review*, pages 558–565.

Gurobi Optimization. 2014. Gurobi optimizer reference manual, version 5.6. Houston, TX, USA.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL 2014*, Baltimore, Maryland.

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

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Multimodal neural language models. In *Proceedings of ICML 2014*, pages 595–603.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2891–2903.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL 2014*, pages 1403–1414. ACL.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics*, 24(2):217–244.

Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.

Zachary Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of AISTATS*.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL 2010*.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP-CoNLL*.

P. Pantel, R. Bhagat, T. Chklovski, and E. Hovy. 2007. Isp: Learning inferential selectional preferences. In *Proceedings of NAACL 2007*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

Y. Peirsman and S. Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of NAACL 2010*, pages 921–929.

Philip Resnik. 1993. Selection and information: A class-based approach to lexical relationships. Technical report, University of Pennsylvania.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington, D.C.

Alan Ritter, Mausam Etzioni, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings ACL 2010*, pages 424–434.

Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *Proceedings of ICCV 2013*.

Stephen Roller and Sabine Schulte im Walde. 2013. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of EMNLP 2013*, pages 1146–1157, Seattle, WA.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL 1999*, pages 104–111.

David Shamma. 2014. One hundred million Creative Commons Flickr images for research. `http://labs.yahoo.com/news/yfcc100m/`.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing. *Computational Linguistics*, 39(2).

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL 2010*, pages 1029–1037, Los Angeles, USA.

Ekaterina Shutova. 2011. *Computational Approaches to Figurative Language*. Ph.D. thesis, University of Cambridge, UK.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL 2014*, Baltimore, Maryland.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of ACL 2013*, pages 572–582.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Proceedings of NIPS 2013*, pages 935–943.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP 2009*.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of EMNLP 2014*.

Wiebke Wagner, Helmut Schmid, and Sabine Schulte Im Walde. 2009. Verb sense disambiguation using a predicate-argument clustering model. In *Proceedings of the CogSci Workshop on Semantic Space Models (DISCO)*.

M.D. Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, 20:6–11.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 67–78.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of COLING/ACL*, pages 849–856.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2010. Improving semantic role classification with selectional preferences. In *Proceedings of NAACL HLT 2010*, pages 373–376.