Perceptually Grounded Selectional Preferences

Ekaterina Shutova
Niket Tandon
Gerard de Melo

Ekansh Gupta
Senior undergraduate, EE
IIT Kanpur

Introduction

- Selectional preferences (SPs) are the semantic constraints that a predicate places onto its arguments.

- SPs provide generalisations about word meaning and use and are used for including word sense disambiguation, resolving ambiguous syntactic attachments, natural language inference and figurative language processing.

- Automatic acquisition of SPs from linguistic data has thus become an active area of research.

Motivation

- Little research has been concerned with the sources of knowledge that underlie the learning of SPs.
- Text-based models of SPs which suffer from two problems: topic bias and figurative uses of words.
- There is ample evidence in cognitive and neurolinguistics that word meanings are also acquired from our experiences in the physical world.
- There has not yet been a multimodal semantic approach performing extraction of predicate-argument relations from visual data.

Resources and Outline

- British National Corpus (BNC) is used as an approximation of linguistic knowledge and a large collection of tagged images and videos from Flickr as an approximation of perceptual knowledge.

- The experiments focus on verb preferences for their subjects and direct objects.

- The 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

Experimental Data

- The corpus is parsed using the RASP parser and subject–verb and verb–object relations from its dependency output are extracted.

- 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, the Yahoo! Flickr-100M dataset is mined.

- Flickr-100M contains 99.3 million images and 0.7 million videos with language tags annotated by users, to generalize SPs at a large scale.

visual verb-noun co-occurrence

- For a given word *i* and one of its candidate WordNet senses *j*, 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.

- Then it computes 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'* using WordNet's taxonomic pathbased similarities 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 for verb-verb pairs.

- It maximizes the coherence of the senses of the words in the set as an Integer Linear Program, using the Gurobi Optimizer

visual verb-noun co-occurrence

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

- Verb-noun co-occurrence information is then extracted from the PoS-tagged sets.

Selectional preference model

- To address the issue of data sparsity, selectional preferences over argument classes is generalized, as opposed to individual arguments
- *Jensen-Shannon divergence* is used to measure the similarity between feature vectors for two nouns, w_i and w_j, defined as follows:

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

- Where d_{KL} is the Kullback-Leibler divergence and m is the average of w_i and w_j.
- A similarity matrix S is computed with $S_{ij} = \exp(-d_{JS}(w_i, w_j))$
- *S is transformed* into a stochastic matrix *P* containing transition probabilities between the vertices in the graph as $P = D^{-1}S$ 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.
- Clustering is done separately on both linguistic and visual data.

Selectional preference model

- *Selectional preference strength* (SPS) of a verb is computed 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, $p(c)$ as:

$$SPS_R(v) = \sum_c p(c|v) \log\left(\frac{p(c|v)}{p(c)}\right)$$

- 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

$$Ass_R(v,c) = \frac{1}{SPS_R(v)} p(c|v) \log\left(\frac{p(c|v)}{p(c)}\right)$$

- To combine the two models, two interpolation techniques are used: simple linear interpolation and predicate-driven linear interpolation.

Selectional preference model

- The probabilities $p(c)$ and $p(c/v)$ in the linguistic (LM) and visual (VM) models are interpolated, as follows:

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

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

- For each predicate v, the interpolation weights based on its prominence in the respective corpus are computed, as follows:

$$\lambda_i(v) = \frac{\mathrm{rel}_i(v)}{\sum_k \mathrm{rel}_k(v)}$$ where $$\mathrm{rel}_i(v) = \frac{f_i(v)}{\sum_V f_i(v)}$$

- The interpolation weights for LM and VM are then computed as

$$\lambda_{\mathrm{LM}}(v) = \frac{\mathrm{rel}_{\mathrm{LM}}(v)}{\mathrm{rel}_{\mathrm{LM}}(v) + \mathrm{rel}_{\mathrm{VM}}(v)}$$

$$\lambda_{\mathrm{VM}}(v) = \frac{\mathrm{rel}_{\mathrm{VM}}(v)}{\mathrm{rel}_{\mathrm{LM}}(v) + \mathrm{rel}_{\mathrm{VM}}(v)}$$

Evaluation

- The predicate-argument scores assigned by these models is evaluated against a dataset of human plausibility judgements of verb-direct object pairs collected by Keller and Lapata (2003) in terms of Pearson correlation coefficient (*r*) and Spearman rank correlation coefficient (*ρ*)

	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**

(λ_{LM} = 0.9) outperforms all of these methods, as well as LSP, on the *Seen* dataset, confirming the positive contribution of visual features.

Evaluation

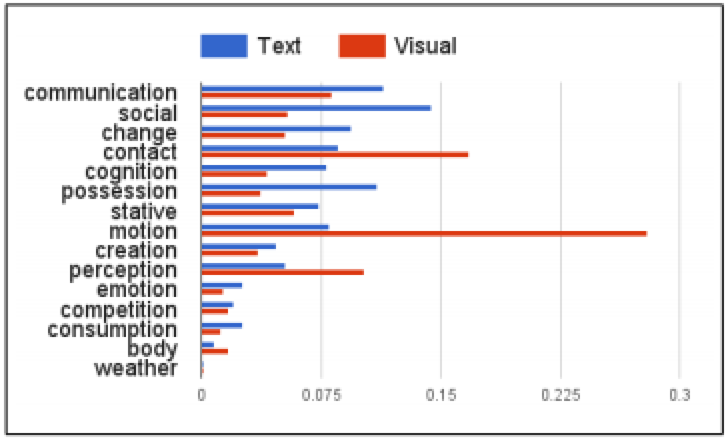

Figure 6: WordNet top level class distributions for verbs in the visual and textual corpora

- This suggests that integrating this ISP model (that currently outperforms others on more common pairs) with such techniques is likely to improve SP prediction across frequency bands.

- 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.

Conclusion

- The experiments show that it outperforms linguistic and visual models in isolation, as well as the previous approaches to SP learning.

- Human-annotated image and video descriptions allow 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.

- In the future, SP interpolation can be applied 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 where combining the visual and linguistic knowledge is likely to outperform text-based models on their own.

THANK YOU