Semantic Frames and Visual Scenes: Learning Semantic Role Inventories from Image and Video Descriptions

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

Andreas Wundsam
Big Switch Networks
Santa Clara, CA
andi@wundsam.net

Helen Yannakoudakis
ALTA Institute
University of Cambridge, UK
hy260@cl.cam.ac.uk

Abstract

Frame-semantic parsing and semantic role labelling, that aim to automatically assign semantic roles to arguments of verbs in a sentence, have recently become an active strand of research in NLP. However, to date these methods have relied on a predefined inventory of semantic roles. In this paper, we present a method to automatically learn argument role inventories for verbs from large corpora of text, images and videos. We evaluate the method against manually constructed role inventories in FrameNet and show that the visual model outperforms the language-only model and operates with a high precision.

1 Introduction

The theory of frame semantics (Fillmore, 1976) postulates that our interpretation of word meanings is not limited to isolated concepts, but rather instantiates complex knowledge structures about events and their participants, known as *semantic frames*. For instance, the COMMERCIAL TRANSACTION frame includes elements such as *a seller, a buyer, goods* and *money* which can be mapped to higher-level semantic roles such as *agent, patient, instrument* etc. The verbs linked to this frame are *buy, sell, pay, cost* and *charge*, each evoking different aspects of the frame.

This theory has been implemented in a lexical-semantic resource called FrameNet (Fillmore et al., 2003). Each semantic frame is encoded in FrameNet as a list of lexical units that evoke this frame (typically verbs) and the roles that their semantic arguments may take given the scenario represented by the frame. FrameNet has inspired a direction in NLP research known as semantic role labelling (Gildea and Jurafsky, 2002; Màrquez et al., 2008) and frame-semantic parsing (Das et al., 2014), whose goal is to assign semantic roles to arguments of the verbs in a sentence. However, these works point out the coverage limitations of the hand-constructed FrameNet database, suggesting that a data-driven frame acquisition method is needed to enable the integration of frame semantics into real-world NLP applications. In this paper, we propose such a method, experimenting with semantic frame induction from linguistic and visual data. Our system first performs clustering of verb arguments to identify their possible semantic roles and then computes the level of association between a given argument role and the verb, thus deriving the structure of the semantic frame in which the verb participates.

Frame semantics emphasizes the relation between our lexical semantic knowledge and our experience in the world, suggesting that semantic frames are not merely a linguistic construct but also a result of our sensory-motor and perceptual experience. However, frame semantic approaches in NLP typically rely on textual data. Our method, in contrast, induces semantic frames from both a text corpus and a corpus of tagged images and videos. We evaluate the method against hand-constructed frames in FrameNet. Our results show that the visual model outperforms the language-only model and achieves a high precision. This frame induction method can be used to complement existing FrameNets or to construct a new resource of automatically mined semantic frames, free from manual annotation bias.

2 Experimental Data

Textual data. We extracted linguistic features for our model from the British National Corpus (BNC) (Burnard, 2007). We parsed the corpus using the RASP parser (Briscoe et al., 2006) and

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 149–154,
Vancouver, Canada, August 3-4, 2017.

extracted subject–verb and verb–object relations from its dependency output. These relations were then used as features for clustering to obtain arguments classes, which we then used as proxies for frame elements, i.e. argument roles.

Image and video data. We used the Yahoo! Webscope Flickr-100M dataset (Shamma, 2014) to extract visual relations between verbs and their arguments. Flickr-100M contains 99.3 million images and 0.7 million videos with natural language tags for scenes, objects and actions annotated by users. We first stem the tags and remove words that are absent in WordNet (e.g. named entities and misspellings). We then identify their part of speech based on their visual context using the method of Shutova et al. (2015) and extract verb–noun co-occurrences.

3 Frame Induction Model

3.1 Argument Clustering

We use a clustering method to obtain semantic classes of arguments of verbs, thus generalising from individual arguments to their semantic types which correspond to frame roles. We obtain argument classes by means of spectral clustering of nouns with lexico-syntactic features, which has been shown effective in previous lexical classification tasks (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 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 multi-way 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 $P = D^{-1}S$, 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:

$$\mathrm{MNCut}(I) = \sum_{k=1}^{K} (1 - P(I_k \rightarrow I_k | I_k)), \quad (2)$$

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

We clustered the 2,000 most frequent nouns in the BNC, using their grammatical relations as features. The features consisted 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 were first constructed from the corpus counts, and subsequently normalized by the sum of the feature values.

Our use of linguistic dependency features for argument clustering is motivated by the results of previous research (Sun and Korhonen, 2011; Shutova et al., 2015), that has shown that such features lead to clusters of nouns belonging to the same semantic type, as opposed to topic or scene as it is the case with linguistic window-based features or image-derived features (Shutova et al., 2015). Since the argument roles in semantic frames correspond to semantic types (such as *location* or *instrument*), the linguistic dependency features are best suited to generalise the predicate-argument structure in semantic frames. Example clusters produced by our method are shown in Fig. 1. The resulting clusters represent frame elements, i.e. argument roles, in our model.

3.2 Predicate–Argument Association

We then use the verb–noun co-occurrence information extracted from the visual data to quantify the strength of association of a given verb with each of the argument classes, thus identifying the relevant argument roles for the verb. We adopted an information theoretic measure originally proposed by Resnik (1993) in his selectional preference model. Resnik first measures *selectional*

official officer inspector journalist detective constable police policeman reporter
fire pipe torch candle lamp cigarette
potato apple slice food cake meat bread fruit
lifetime quarter period century succession stage generation decade phase interval future
disorder infection illness disease virus cancer
profit surplus earnings income turnover revenue

Figure 1: Clusters representing argument roles

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, $p(c)$:

$$\text{SPS}(v) = \sum_c p(c|v) \log \frac{p(c|v)}{p(c)}. \quad (3)$$

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:

$$\text{Ass}(v, c) = \frac{1}{\text{SPS}(v)} p(c|v) \log \frac{p(c|v)}{p(c)}. \quad (4)$$

We use this measure to quantify the strength of verb–argument association based on the visual co-occurrence information. We extract verb-noun co-occurrences from Flickr-200M, map the nouns to argument classes and quantify selectional association of a given verb with each argument class, thus acquiring its semantic frame structure. An example argument distribution for the verb *kill*, and thus the KILLING frame, is presented in Fig. 2. One can see from the figure that the argument clusters correspond to specific roles in FrameNet, e.g. the *killer* and the *victim*, the *motive*, the *weapon* (instrument) and *death* (result).

4 Evaluation against FrameNet

Baseline. We evaluate the effectiveness of visual information for our task by comparing the model based on vision and language (VIS) to a baseline model using language alone (LING). In the LING system, the predicate-argument association scores are computed based on verb-argument co-occurrence information extracted from verb-subject, verb-direct object and verb-indirect object relations in the BNC. In case of the indirect object relations, the accompanying prepostions were discarded and the noun counts were aggregated.

0.180 defeat fall **death** tragedy loss collapse decline disaster destruction fate
0.141 girl other woman child **person** people
0.128 suicide **killing** offence **murder** breach crime ...
0.113 handle **weapon** horn knife blade stick sword ankle waist neck wrist
0.095 **victim** bull teenager prisoner hero gang enemy rider offender youth **killer** thief driver defender hell
0.086 recession disappointment **shock pain** frustration embarrassment **guilt** sensation depression wound
0.030 sister daughter parent **relative** lover cousin friend wife mother husband brother father
0.020 **motive** self origin meaning **cause** secret truth ...
0.018 official officer inspector journalist detective constable police **policeman** reporter

Figure 2: System output for *kill*

Evaluation setup. In order to investigate the role of visual information for different types of verbs, we selected 25 concrete verbs (e.g. *cut, throw, swim*) and 25 abstract verbs (e.g. *trust, prepare, cheat*), according to the MRC concreteness database (Wilson, 1988). The verb was considered concrete if its concreteness score was $\geqslant 400$ and abstract if it was < 400. We extracted the 10 highest-ranked verb–argument class pairings produced by the system for each verb. Each pairing was then evaluated against the argument roles listed for this verb in FrameNet via manual comparison. This resulted in a dataset of 500 verb-argument pairings for VIS and 500 for LING. The pairing was considered correct if the argument cluster corresponded to the semantic type of the role listed in FrameNet and contained nouns listed in the linguistic examples (if these were provided in FrameNet). We have evaluated the system performance in terms of precision at top 10 argument classes and recall of the Core Frame Elements (FEs) among the top 10 argument classes.

Results The VIS model attained a performance of $P = 0.74$ and $R = 0.78$, outperforming the LING model with $P = 0.72$ and $R = 0.76$. When evaluated on the subsets of concrete and abstract verbs separately, VIS attains a $P = 0.76$; $R = 0.80$ (concrete) and $P = 0.72$; $R = 0.75$ (abstract), and LING attains $P = 0.67$; $R = 0.75$ (concrete) and $P = 0.78$; $R = 0.76$ (abstract).

5 Discussion and Data Analysis

Our results show that the vision-based model outperforms the language-only model on our dataset. The difference in performance is particularly pronounced for the concrete verbs. For the abstract verbs in isolation, however, LING attains a higher

precision and recall. This is not surprising, as the visual information is better suited to capture the properties of concrete concepts than the abstract ones (Kiela et al., 2014). However, our results indicate that integrating linguistic and visual information provides a better overall model than the linguistic information alone.

Our qualitative analysis of the data revealed a number of interesting trends. Some of the errors of both systems can be traced back to the clustering step. Different argument roles according to FrameNet are sometimes found in one cluster. For instance, both the *killer* and the *victim* are in the same cluster, as shown in Figure 2. However, it is also the case that one FrameNet role can be split into several clusters, e.g. the *Victim* role in the *killing* frame is represented by two clusters of *humans* and *animate beings* more generally.

The common error of the LING model concerns frame mixing, i.e. both literal and metaphorical arguments of the verb are present in the output. For instance, *eat* has a *disease* cluster as one of its arguments; however, *disease* is not part of the *ingestion* frame, but rather an instance of its metaphorical transfer. A common trend in the LING output is that it is dominated by the *Agent* and *Theme* roles, with situational roles (e.g. *Location*) typically ranked lower or not appearing at all. In contrast, the output of VIS encompases a range of situational roles, such as *Instrument, Location, Time* etc. The two models also sometimes differ in the roles that they identify. For instance, for the verb *risk* the VIS output is dominated by arguments of type *Asset* and the LING output by the arguments related to the *Bad outcome* role in FrameNet.

6 Related Work

6.1 Semantic Role Induction

Approaches most similar in spirit to ours are those concerned with unsupervised semantic role labeling. A number of methods represented semantic roles as latent variables in a graphical model, which related the verb, its semantic roles and their syntactic realisations (Grenager and Manning, 2006; Lang and Lapata, 2010; Garg and Henderson, 2012). The induction process then relied on inferring the state of the latent variable. Other researchers adopted a similarity-based argument clustering framework to derive semantic roles. The investigated methods include graph partitioning algorithms (Lang and Lapata, 2014), Bayesian clustering based on Chinese Restaurant Process (Titov and Klementiev, 2012) and integer linear programming to incorporate semantic and structural constraints during clustering (Woodsend and Lapata, 2015). Titov and Khoddam (2015) proposed a reconstruction-error minimization approach using a log-linear model to predict roles given syntactic and lexical features and a probabilistic tensor factorization model to identify argument fillers based on the role predictions and the predicate. To the best of our knowledge, ours is the first approach to this task exploiting visual data, in the form of image and video descriptions.

6.2 Multi-modal Methods in Semantics

Visual data has been previously used to learn meaning representations that project multiple modalities into the same vector space. Semantic models integrating linguistic and visual information have been shown successful in tasks such as modeling semantic similarity and relatedness (Silberer and Lapata, 2014; Bruni et al., 2012), lexical entailment (Kiela et al., 2015a), compositionality (Roller and Schulte im Walde, 2013), bilingual lexicon induction (Kiela et al., 2015b) and metaphor identification (Shutova et al., 2016).

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) show that large collections of image captions can be exploited for entailment tasks. Shutova et al. (2015) used image and video descriptions to induce verb selectional preferences enhanced with visual information.

7 Conclusion

We have presented a method for semantic frame induction from text, images and videos and shown that it operates with a high precision and recall. Although our experiments relied on manually annotated tags for images and videos, recent research shows that such tags can be generated automatically (Bernardi et al., 2016). In the future, our model can be applied to such automatically generated tags, reducing its dependence on manual annotation. While our current experiments focused on nominal arguments of the verbs for semantic role identification, in principle, our model can be applied to other parts of speech, e.g. adverbs, to better incorporate argument roles such as *Manner*.

Acknowledgment

We are grateful to the *SEM reviewers for their feedback. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *JAIR* .

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

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in Technicolor. In *Proceedings of ACL*. Korea.

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*. http://www.natcorp.ox.ac.uk/XMLedition/URG/.

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

Dipanjan Das, Desai Chen, Andr F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics* 40:1:9–56.

S. Divvala, A. Farhadi, and C. Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of CVPR*.

Charles Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280(1):20–32.

Charles Fillmore, Christopher Johnson, and Miriam Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3):235–250.

Nikhil Garg and James Henderson. 2012. Unsupervised semantic role induction with global role ordering. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*. pages 145–149.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3).

Trond Grenager and Christopher D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06, pages 1–8.

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*.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015a. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015b. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Multimodal neural language models. In *Proceedings of ICML 2014*. pages 595–603. http://jmlr.org/proceedings/papers/v32/kiros14.html.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10, pages 939–947.

Joel Lang and Mirella Lapata. 2014. Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics* 40(3):633–669.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics* 34(2):145–159.

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

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

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

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

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *NAACL HLT 2016*. pages 160–170.

Ekaterina Shutova, Niket Tandon, and Gerard de Melo. 2015. Perceptually grounded selectional preferences. In *Proceedings of ACL 2015*. Beijing, China.

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

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

Lin Sun and Anna Korhonen. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*. Edinburgh, UK, pages 1023–1033.

Ivan Titov and Ehsan Khoddam. 2015. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *HLT-NAACL*. The Association for Computational Linguistics, pages 1–10.

Ivan Titov and Alexandre Klementiev. 2012. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12, pages 12–22.

M.D. Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers* 20(1):6–11.

Kristian Woodsend and Mirella Lapata. 2015. Distributed representations for unsupervised semantic role labeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 2482–2491.

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