Visually Guided Spatial Relation Extraction from Text

Taher Rahgooy* **Umar Manzoor*** **Parisa Kordjamshidi***†
*Tulane University, Computer Science Department, New Orleans, LA, USA
†Florida Institute for Human and Machine Cognition (IHMC), Pensacola, FL, USA
{trahgooy,umanzoor,pkordjam}@tulane.edu

Abstract

Extraction of spatial relations from sentences with complex/nesting relationships is very challenging as often needs resolving inherent semantic ambiguities. We seek help from visual modality to fill the information gap in the text modality and resolve spatial semantic ambiguities. We use various recent vision and language datasets and techniques to train inter-modality alignment models, visual relationship classifiers and propose a novel global inference model to integrate these components into our structured output prediction model for spatial role and relation extraction. Our global inference model enables us to utilize the visual and geometric relationships between objects and improves the state-of-art results of spatial information extraction from text.

1 Introduction

Significant progress has been made in spatial language understanding by mapping natural language text to spatial ontologies (Kordjamshidi and Moens, 2015). The research results show that spatial entities can be extracted with a good accuracy, however, spatial relation extraction is still challenging (Kordjamshidi et al., 2017a; Pustejovsky et al., 2015). Particularly, when the sentences convey more than one relationship, finding the right links between the spatial objects and spatial prepositions becomes difficult. For example, the spatial meaning of *There is a car in front of the house on the left*, can be interpreted in different ways: *(A car in front of the house) on the left* or *A car in front of (the house on the left)*. This issue is related to the well-known *prepositional phrase attachments* (pp-attachments) syntactic ambiguity which is problematic for our goal of spatial semantic extraction too. The previous research shows some of these ambiguities can be resolved by simultaneously reasoning from the associated image (Christie et al., 2016; Delecraz et al., 2017). Consider the scene in Figure 1, we can easily resolve the ambiguity and choose the correct interpretation with the help of the associated image.

Although we do not directly tackle the task of pp-attachment here, resolving this issue will help our task to find the accurate link between the spatial prepositions (i.e. *spatial indicators*) and spatial objects/roles (*trajector* and *landmark*). The spatial semantic links can go beyond the syntactic links/attachments, therefore merely fixing the preposition attachments is not sufficient for our task. We exploit the image to find the right preposition that describes the relationships between the spatial roles, for example *on the left* can be a relationship between *the house* and implicit landmark *picture* as well as *a car* and implicit landmark *picture*. There are many recent works on combining vision and language for domains such as image captioning (Karpathy and Fei-Fei, 2017), visual image retrieval (Hu et al., 2016), visual question answering (Krishna et al., 2017; Faghri et al., 2017), activity recognition (Gupta and Malik, 2015; Yatskar et al., 2016; Yang et al., 2016), visual relation extraction (Lu et al., 2016; Xu et al., 2017; Haldekar et al., 2017; Peyre et al., 2017; Liao et al., 2017) and object localization (Kazemzadeh et al., 2014; Schlangen et al., 2016). We aim at exploiting models from visual modality to boost the models trained by the text modality and improve spatial role labeling task (SpRL) (Kordjamshidi et al., 2011). The most related work to ours is (Kordjamshidi et al., 2017a) in which they connected phrases to ground-truth labeled segments using word embedding similarity to generate additional visual features, whereas, in this work, we train actual inter-modality alignment models to include visual information in our model. The challenges are 1) existing textual datasets for SpRL does not have enough examples to train such

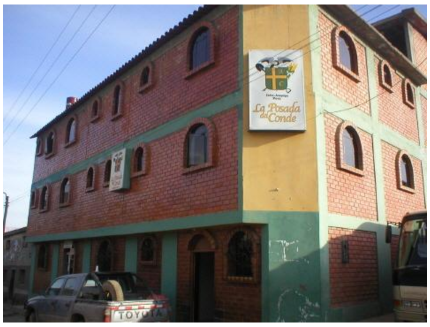

Figure 1: A captioned scene from CLEF IAPR TC-12 dataset: *There is a car in front of the house on the left.*

visual models, therefore, such models need to be trained on external datasets and later incorporated in our multi-modal setting, 2) Aligning text entities with image entities is a complex and challenging task itself. Each modality in isolation represents spatial relations imperfectly, however, each one can reflect different types of spatial relation better than the other. If we can handle the mentioned challenges and combine the two modalities then vision modality fills the information gap of the text modality and improves the information extraction.

To overcome the above challenges, we 1) trained two visual models namely word-segment alignment, trained on ImageClef Referring Expression Dataset[1] to connect the two modalities, and preposition classifier, trained on Visual Genome dataset (Krishna et al., 2017) to help in link disambiguation, and 2) generated a unified graph, based on both image and text data and proposed a global machine learning model to exploit the information from the companion images.

The contribution of this paper includes a) exploiting the visual information to solve the SpRL task and improving the state-of-the-art results significantly b) forming a global inference model that imposes the consistency constraints on the decisions made based on the two modalities c) exploiting external vision and language datasets to inject external knowledge into our models d) augmenting an existing dataset which is annotated by spatial semantics with the image segment alignments, this dataset will help the evaluation of the existing methods for combining vision and language for fine-grained spatial semantic extractions.

2 Model Description

Given a piece of text, S, here a sentence - split into number of phrases, and an accompanying image, I segmented into number of segments represented by bounding boxes, the goal is to identify the textual phrases that have spatial roles and detect the relationships between them. The spatial roles included in this task are defined as:

(a) **Spatial indicators (sp):** these are triggers indicating the existence of spatial information in a sentence;

(b) **Trajectors (tr):** these are the entities whose location are described;

(c) **Landmarks (lm):** these are the reference objects for describing the location of the trajectors.

In the textual description of Figure 1, the location of *car (trajector)* has been described with respect to *house (landmark)* using the preposition *in front of*. Furthermore, spatial relationships and their types are defined as follows:

(a) **Spatial relations:** these indicate a link between the three above mentioned roles $(sp.tr.lm)$, forming spatial triplets.

(b) **Coarse-grained relation types:** these indicate the coarse-grained type of relations in terms of spatial calculi formalisms including *region*, *direction*, and *distance* types.

(c) **Fine-grained relation types:** these indicate the fine-grained type of relations in terms of each specific spatial calculi formalism. *Region connection calculi (RCC8)* types (e.g. disconnected (DC), externally connected (EC), etc.), a closed set of directional relations (e.g. left, right), and an open set of distal relations (e.g. close, far) are defined for regional, directional, and distal relationships respectively.

For example, given the sentence and its accompanying image in Figure 1, the goal is to identify the spatial relations, $\langle$[*A car*]$_{\text{tr}}$, [*in front of*]$_{\text{sp}}$, [*the house*]$_{\text{lm}}\rangle$ and $\langle$[*the house*]$_{\text{tr}}$, [*on the left*]$_{\text{sp}}$, [*None*]$_{\text{lm}}\rangle$ and also determine their coarse-grained types (*direction* and *direction* respectively) and fine-grained types (*front* and *left* respectively).

We formulate this problem as a structured output prediction problem. Given a set of input-out pairs as training examples, $E = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} : i = 1..N\}$, a scoring function $g(x, y; W) = \langle W, \phi(x, y)\rangle$ will be learned. Where W is the weight vector and $\langle,\rangle$ is dot product between two vectors. This function is a linear discriminant function defined over combined feature representation of inputs and outputs denoted by $\phi(x, y)$

[1] http://tamaraberg.com/referitgame/

(Tsochantaridis et al., 2005). However, for this work we use a piece-wise training model in which independent models are trained per concepts in the output and the predictions are done based on global inference (Punyakanok et al., 2005).

We construct a graph using the phrases $\{p_1, ..., p_n\}$ and bounding boxes $\{b_1...b_m\}$, and link these nodes to make composed concepts (like relations, roles, etc.). We associate a classifier to each concept in the graph and encode the domain knowledge as global constraints over these concepts. We use global reasoning by imposing these constraints over various (node and edge) classifiers to produce the final outputs. The input of our structured output prediction model is the aforementioned graph and the output is the concepts assigned to the nodes and edges of this graph. In the following we describe the information that we use from each modality (i.e. text and image) and from inter-modality relationships and describe the relevant classifiers, features and constraints.

2.1 Text

We use binary classifiers to identify spatial roles and relations. The spatial roles of trajector, landmark and spatial indicator are denoted by tr, lm, and sp. $sp.tr.lm$, $sp.tr.lm.\gamma$, and $sp.tr.lm.\lambda$ denote spatial relations, coarse-grained relation types, and fine-grained relation types respectively. Additionally, we denote candidate fine-grained types related to coarse-grained type γ by Λ_γ.

Features: We use phrase-based features $\phi_{phrase}(p_i)$ for role classifiers in which p_i is the identifier of i^{th} phrase in the sentence which include several linguistically motivated features such as lexical form of the words in the phrases, lemmas, pos-tags, etc. In addition, motivated by features used in (Roberts and Harabagiu, 2012) and (Kordjamshidi et al., 2017a), we use a combination of phrase-based features like concatenation of headwords of the roles, concatenation of their pos, and other relative features such as distance between roles, dependency relations, sub-categorization, etc., to represent the relations and this is referred to as $\phi^{text}_{triplet}(p_i, p_j, p_k)$.

Constraints: The constraints over spatial concepts expressed in text are as follows,

$\sum_i \sum_k sp_i tr_j lm_k \geq tr_j$	Each tr candidate at least should appear in one relation
$\sum_i \sum_j sp_i tr_j lm_k \geq lm_k$	Each lm candidate at least should appear in one relation
$\sum_j \sum_k sp_i tr_j lm_k = sp_i$	Each sp candidate should appear in one relation
$\sum_j tr_j \geq sp_i$	For each sp we should have at-least one tr
$\sum_k lm_k \geq sp_i$	For each sp we should have at-least one lm. Including null landmarks
$sp_i tr_j lm_k \gamma \leq sp_i tr_j lm_k$	is-a constraints between relations and coarse-grained types
$sp_i tr_j lm_k \lambda \underset{\lambda\in\Lambda_\gamma}{\leq} sp_i tr_j lm_k \gamma$	is-a constraints between coarse-grained and corresponding fine-grained types.

2.2 Image

In the image modality, we have two types of classifiers, 1) for localization of an object in the image given a referring expression, and 2) for extraction of spatial relations, called *Word-Segment Alignment* and *Preposition Classifier* respectively.

Word-Segment alignment: motivated by (Schlangen et al., 2016), we trained a set of binary object localization classifiers to link words and image segments. These per word classifiers are trained using *ImageClef Referring Expression Dataset*.

Preposition classifier: is a multi-class classifier that takes two bounding boxes and returns the spatial relation (preposition here) between them. This classifier is trained on a subset of visual genome dataset (Krishna et al., 2017) described in section 3.1, this classifier provides the external knowledge from visual resources and help in disambiguation of ambiguous links (i.e. finding the correct link between spatial preposition and spatial roles).

Features: A deep convolutional neural network, "GoogLeNet" is used to extract features for bounding boxes that are used by *Word-Segment Alignment*, for details see (Schlangen et al., 2016). For the *Preposition Classifier* we use bounding box features $\phi_{box}(b) = [l, x_{min}, y_{min}, w_b, h_b]$ where l is the label of the box, (x_{min}, y_{min}) is the top-left point of the box, w_b, h_b are the width and height of the box respectively. In addition we use pair features $\phi^{visual}_{pair}(b_i, b_j)$ including, label of each box, distance between the center of the two boxes, a vector from the center of first box to the center of second box, aspect ratio of each box, word to vector representation of each box's label, the normalized area of each box, intersection, union, intersection over union of the two boxes, and four directional (above, below, left and right) features calculated with reference to the two boxes. Box and pair features are adopted from (Ramisa et al., 2015).

2.3 Inter-Modality

An essential part of having a global inference over multimodal data is to have the connections between the two modalities. *Word-Segment Alignment* classifier is used to align the headword of each phrase p in the sentence to its corresponding bounding box b in the image and this alignment is denoted by $p \rightarrow b$. A binary feature $isAligned$ (that indicates if the phrase is connected to an object in the image) is added to the features of classifiers in the text side.

Constraints: Given two bounding boxes b_1 and b_2 we say that the preposition α is supported by the image and write $iSup^{\alpha}_{b_1 b_2}$ if α is ranked among top N prepositions according to *Preposition Classifier* scores. Using this indicator we define the following inter-modality constraint.

$$\underset{p_i \rightarrow b_1, p_j \rightarrow b_2, \alpha = p_i}{iSup^{\alpha}_{b_1 b_2} \leq sp_i tr_j lm_k}$$

For aligned pairs, the visual relation should support the textual relation

2.4 Global Reasoning

We obtain the output of each classifier in the model holistically by global reasoning that is by considering global correlations among classifiers, when calculating outputs. This goal is achieved by optimizing an objective function that is the summation of classifiers' discriminant functions,

$$\sum_{i \in C_{sp}} \langle W_{sp}, \phi_{sp_i} \rangle . sp_i + \sum_{i \in C_{tr}} \langle W_{tr}, \phi_{tr_i} \rangle . tr_i +$$

$$\sum_{i \in C_{lm}} \langle W_{lm}, \phi_{lm_i} \rangle . lm_i +$$

$$\sum_{i \in C_{sp}} \sum_{j \in C_{tr}} \sum_{k \in C_{lm}} \langle W_{sptrlm}, \phi_{sp_i tr_j lm_k} \rangle . sp_i tr_j lm_k +$$

$$\sum_{\gamma \in \Gamma} \sum_{i \in C_{sp}} \sum_{j \in C_{tr}} \sum_{k \in C_{lm}} \langle W_{sptrlm}, \phi_{sp_i tr_j lm_k \gamma} \rangle . sp_i tr_j lm_k \gamma +$$

$$\sum_{\lambda \in \Lambda} \sum_{i \in C_{sp}} \sum_{j \in C_{tr}} \sum_{k \in C_{lm}} \langle W_{sptrlm}, \phi_{sp_i tr_j lm_k \lambda} \rangle . sp_i tr_j lm_k \lambda +$$

$$\sum_{\alpha \in Prep} \sum_{(i,j) \in C_{iSup}} \langle W_{iSup^{\alpha}}, \phi_{iSup^{\alpha}_{ij}} \rangle . iSup^{\alpha}_{ij}.$$

Each classifier is shown as a binary variable (e.g. tr_i for trajector classifier). Λ, Γ, $Prep$ are the candidates for fine-grained relations, coarse-grained relations, and prepositions from text respectively. C_l denotes the candidates for label l.

The following model variations are designed using combination of text and image modalities for experimentation.

Baseline Model (BM): Independent classifiers are trained only on the textual features described in Section 2.1. This is a learning only model and each classifier makes independent predictions.

Baseline + Constraints (BM+C): The output of the classifiers obtained from the BM model are adjusted by global inference over textual constraints defined in Section 2.1.

Ground-truth alignments (GT): This setting is very similar to the $BM + C$ model except the $isAligned$ feature (see Section 2.3) added to consider the ground-truth alignments.

Alignment Classifier (AC): Similar to the GT model, but instead of ground-truth information we use *Word-Segment Alignment* classifier to align bounding boxes with the phrases in the sentence.

GT + Preposition (GT+P): In this setting, ground-truth alignments alongside *Preposition classifier* is used to enforce all constraints in the global inference over the two modalities.

AC + Preposition (AC+P): Same as GT+P model but with *Word-Segment Alignment* classifiers instead of ground-truth alignments.

3 Experimental Setup

We report the experimental results of our model and compare it with the state-of-the-art (Kordjamshidi et al., 2017a) model, referred here as *M0* model. A role prediction is considered correct if there is a phrase overlap between the ground-truth and predicted roles and each relation is counted as correct when all three arguments are correct. All the base classifiers described in Section 2.1 are sparse perceptrons. We use *Saul* (Kordjamshidi et al., 2015, 2016) to implement the models and solve the global inference of Section 2.4. The code is publicly available here. [2]

3.1 DataSets

CLEF 2017 mSpRL dataset: This dataset is a subset of IAPR TC-12[3] Benchmark which is annotated for the SpRL task (Kordjamshidi et al., 2017c, 2012). It contains 613 images with descriptions including 1,213 sentences. The standard split of the dataset contains 761 training and 939 testing spatial relations (Kordjamshidi et al., 2017b). Furthermore, we added new annotations

[2] `https://github.com/HetML/SpRL/tree/paper2`

[3] `http://www.imageclef.org/SIAPRdata`

	Visual Genome			CLEF		
	P	R	F1	P	R	F1
above	47.24	21.59	29.63	87.50	22.58	35.90
behind	65.02	22.65	33.56	80.00	22.22	34.78
in	80.99	54.96	65.48	83.80	79.87	81.79
in front of	31.65	6.67	11.02	80.00	16.33	27.12
on	79.76	95.75	87.03	38.39	91.49	54.09

(a) Preposition classifier results on Visual genome and CLEF datasets

	P	R	F1
M0	68.34	57.93	62.71
BM	65.64	60.23	62.82
BM+C	70.04	66.55	68.25
GT	66.37	75.14	70.48
GT+P	67.14	74.80	70.76
AC	71.39	66.55	68.89
AC+P	71.69	66.10	68.78

(b) Spatial relations results on CLEF test set

Table 1: Experimental results, where P and R denote precision and recall respectively.

to this dataset to align phrases in the text with the segments of the related images using brat tool.[4] The alignments are used only for evaluations and are publicly available. [5]

Visual Genome dataset (VG): Visual Genome dataset has seven main components (Krishna et al., 2017), one of them is relationships component which contains the relationships (prepositions) between two bounding boxes. The dataset contains 108077 images and the relationships component contains 2316104 relation instances. We used a subset the relationships that correspond to the most frequent prepositions in CLEF dataset. We used 80% for training (811661 instances) and 20% for testing (202916 instances).

ReferItGame Dataset: It contains $120,000$ referring expressions and covers 99.5 percent of the regions of SAIAPRTC-12 dataset which is a segmented and annotated version of the IAPR TC-12 dataset (Kazemzadeh et al., 2014).

3.2 Experimental Results

Word-Segment Alignment: We implemented and trained classifiers per words as described in Section 2.2 for the most frequent words in ReferItGame dataset using (Schlangen et al., 2016) approach. We evaluated the trained model on both ReferitGame and CLEF testset, and obtained 64% and 45% accuracy respectively. This trained model is used to align words and segments in CLEF dataset. The end-to-end evaluation results show that the models trained by this external dataset are helpful though those are not highly accurate for every referring word.

Preposition Classifier: As described in Section 2.2, these are trained on a subset of Visual Genome dataset described in Section 3.1 and evaluated on both Visual Genome and CLEF test sets. Table 1a shows five best prepositions result whereas the result for other prepositions is less than 20% F1.

Spatial Relations: The experimental results in Table 1b show that our baseline model (BM) is as good as the state-of-the-art model (M0). Incorporating $isAligned$ feature (in *GT* and *AC* models) further improves the results because having the phrases visualized in the image increases the confidence scores of the spatial role and relation classifiers and leads to a higher recall. The global inference over constraints in *BM+C* significantly improves the performance of *BM* (about 5% F1). *GT+P* results show that inter-modality constraints help in improving the results (about 2% F1) which indicates some of the visual relations successfully confirmed and boosted their corresponding relations in the text modality. However, this improvement is limited which is expected considering the low performance of *Preposition Classifier*. The *GT+P* results indicate the significance of the visual information in our model when the correct alignments are provided. The alignment classifiers in the *AC* model also slightly improve the *BM+C*. However, as it is visible in *AC+P* results, when we have both noisy alignments and noisy visual relations the results drop slightly compared to *AC*.

4 Conclusion

Our global inference model exploits visual modality classifiers including object localization by referring expressions and spatial relation classifiers between visual objects, as well as classifiers that extract spatial roles and relation from text. The global inference imposes consistency over the two modalities and identifies the spatial relations in text in accordance with their counterparts in the image. The experimental results show the effectiveness of the visual information in resolving the ambiguity of spatial semantics of text. There is still a large room to improve the modality alignments and relation extraction from images to obtain better gains from visual information.

[4] http://brat.nlplab.org/

[5] http://www.cs.tulane.edu/~pkordjam/SpRL.htm#data

References

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1493–1503.

Sebastien Delecraz, Alexis Nasr, Frederic Bechet, and Benoit Favre. 2017. Correcting prepositional phrase attachments using multimodal corpora. In *Proceedings of the 15th International Conference on Parsing Technologies*. pages 72–77.

Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. 2017. VSE++: improved visual-semantic embeddings. *CoRR* abs/1707.05612. http://arxiv.org/abs/1707.05612.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *CoRR* abs/1505.04474. http://arxiv.org/abs/1505.04474.

Mandar Haldekar, Ashwinkumar Ganesan, and Tim Oates. 2017. Identifying spatial relations in images using convolutional neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, pages 3593–3600.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, pages 4555–4564.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4):664–676. https://doi.org/10.1109/TPAMI.2016.2598339.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

P. Kordjamshidi, D. Roth, and H. Wu. 2015. Saul: Towards declarative learning based programming. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. http://cogcomp.cs.illinois.edu/papers/KordjamshidiRoWu15.pdf.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*. volume 2, pages 365–373.

Parisa Kordjamshidi, Daniel Khashabi, Christos Christodoulopoulos, Bhargav Mangipudi, Sameer Singh, and Dan Roth. 2016. Better call saul: Flexible programming for learning and inference in nlp. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 3030–3040.

Parisa Kordjamshidi and Marie-Francine Moens. 2015. Global machine learning for spatial ontology population. *Web Semantics: Science, Services and Agents on the World Wide Web* 30:3–21.

Parisa Kordjamshidi, Taher Rahgooy, and Umar Manzoor. 2017a. Spatial language understanding with multimodal graphs using declarative learning based programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*. pages 33–43.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017b. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 367–376.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing* 8:1–36.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2017c. Spatial role labeling annotation scheme. In N. Ide James Pustejovsky, editor, *Handbook of Linguistic Annotation*, Springer Verlag.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Wentong Liao, Shuai Lin, Bodo Rosenhahn, and Michael Ying Yang. 2017. Natural language guided visual relationship detection. *CoRR* abs/1711.06032. http://arxiv.org/abs/1711.06032.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. Springer, pages 852–869.

Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2017. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 5179–5188.

V. Punyakanok, D. Roth, W. Tau Yih, and D. Zimak. 2005. Learning and inference over constrained output. In *IJCAI'05*. pages 1124–1129. http://dl.acm.org/citation.cfm?id=1642293.1642473.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, pages 884–894. http://cogcomp.org/papers/PustejovskyetalSemEval2015.pdf.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 214–220. http://aclweb.org/anthology/D15-1022.

Kirk Roberts and Sanda M Harabagiu. 2012. Utd-sprl: A joint approach to spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 419–424.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1213–1223.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research (JMLR)*. pages 1453–1484.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426* .

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y Chai. 2016. Grounded semantic role labeling. In *HLT-NAACL*. pages 149–159.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 5534–5542.