A vision-grounded dataset for predicting typical locations for verbs

Nelson Mukuze[1], **Anna Rohrbach**[3,4], **Vera Demberg**[2], **Bernt Schiele**[3]
[1]Visual Meta GmbH
[2]Saarland University
[3]Max Planck Institute for Informatics, Saarland Informatics Campus
[4]EECS, UC Berkeley
nelson.mukuze@visual-meta.com, vera@coli.uni-saarland.de, {arohrbach, schiele}@mpi-inf.mpg.de

Abstract

Information about the location of an action is often implicit in text, as humans can infer it based on common sense knowledge. Today's NLP systems however struggle with inferring information that goes beyond what is explicit in text. Selectional preference estimation based on large amounts of data provides a way to infer prototypical role fillers, but text-based systems tend to underestimate the probability of the most typical role fillers. We here present a new dataset containing thematic fit judgments for 2,000 verb/location pairs. This dataset can be used for evaluating text-based, vision-based or multimodal inference systems for the typicality of an event's location. We additionally provide three thematic fit baselines for this dataset: a state-of-the-art neural networks based thematic fit model learned from linguistic data, a model estimating typical locations based on the MSCOCO dataset and a simple combination of the systems.

Keywords: human judgments, thematic fit, vision

1. Introduction

Most of automatic language understanding today is based on information that is explicitly stated in text. However, tasks that require the ability to make additional common sense inferences are increasingly coming into focus (Levesque et al., 2011; Roemmele et al., 2011; Mostafazadeh et al., 2016). We here propose a dataset which addresses a sub-task towards this goal, i.e. inferring typical locations for a given verb. This task is related to selectional preference tasks, and taps into the kind of inferences humans make when comprehending language. The task consists of predicting a location, given other information from the text (here: the predicate), see also Baroni and Lenci (2010; Sayeed et al. (2015; Tilk et al. (2016). Given the verb *to eat*, a good thematic role filling model would for instance prefer the location *restaurant* over *office*. Previous work has found that locations are a particularly difficult role to predict (Sayeed et al., 2016). This could be due to the fact that location modifiers are often omitted in text when the location is inferable from common sense knowledge, known as reporting bias (Gordon and Van Durme, 2013; Misra et al., 2016). It is hence a logical next step to use multimodal data, and in particular, data from vision. However, there are to date no datasets which can be used to evaluate the contribution of a vision model, as the only existing location fit dataset is very small (Ferretti et al., 2001), comprising only 277 verb/location pairs, and includes many rare verbs.

We here present a new dataset based on 20,000 human judgments for a total of 2,000 verb/location pairs. The dataset was specifically constructed to evaluate the contribution of both visual and linguistic information for learning common sense knowledge about locations. It builds on MS COCO (Lin et al., 2014), a dataset of images with captions; locations for these pictures were labeled using a scene classifier trained to distinguish 365 locations of the Places365 dataset (Zhou et al., 2016). Along with the dataset, we provide performance baselines for a language-based, a vision-based and a simple multimodal model for a common sense inference task using this dataset.

2. Related work

Datasets. Datasets with human ratings on a scale of 1 (least common) to 7 (most common) for *agent* and *patient* roles were made available as part of McRae et al. (1997), Padó et al. (2009), Vandekerckhove et al. (2009). Ferretti et al. (2001) created a dataset of 277 *location* ratings (**Ferretti-Loc**) using questions like "How common is it for someone to *eat* at the following locations"? 40 participants provided ratings on a 7 point scale, e.g.: eat/restaurant: 7. For their study, Ferretti et al. (2001) chose 40 transitive verbs. Many highly frequent verbs like *move* were not included in their study due to the selection criterion they used (the verb should activate a "distinct prototype"), and hence performance on such verbs was not evaluated. Our proposed dataset includes the 100 most frequent verbs in MS COCO image captions (Lin et al., 2014), and is hence intended to be more representative, and makes it possible to evaluate multimodal systems on the task.

Thematic role filling. Many thematic role filling models have been proposed (Baroni and Lenci, 2010; Baroni et al., 2014; Greenberg et al., 2015; Lenci, 2011; Sayeed and Demberg, 2014). Typically, they are evaluated by computing a correlation to human judgments, as motivated by Padó et al. (2009). A related line of work focuses on selectional preference estimation (Erk, 2007; Van de Cruys, 2014). Recently, Tilk et al. (2016) proposed a neural network model for thematic role filling. The system distributes probability over the possible role fillers of specific missing roles. Their model learns the interactions between different roles and achieves state-of-the-art performance on multiple datasets. As a baseline for this dataset, we use their best model, denoted as **Language baseline**. The **Vision baseline** system presented in this work leverages visual information from the captioned images as opposed to purely relying on text.

Vision helps language. Language and vision are the two primary human communication channels. They have a lot of complementary information, which was successfully used by prior work. Regneri et al. (2013) showed that video features help to improve the similarity estimation for sentences describing actions. Yatskar et al. (2016a) relied on images to extract common sense knowledge about objects and spatial relations between them. Tandon et al. (2016) used image tags to learn the *part-of* relation between objects. Yatskar et al. (2016b) proposed a dataset of images and "situations" (a verb with e.g. agent, tool), based on FrameNet (Baker et al., 1998). Prior work has shown that vision can benefit various linguistic similarity tasks (Bruni et al., 2014; Silberer and Lapata, 2014). For such work, it is however a crucial precondition to have a dataset on which system performance can be meaningfully evaluated. Hence, the dataset proposed in this work allows to study the impact of vision for the task of thematic role filling.

3. Dataset

In order to be able to measure progress on the task of learning typical locations of a given event (here approximated via a verb), it is necessary to have a dataset that contains a representative set of verbs and locations together with judgments of how typical these locations are. In order to enable evaluations that correlate automatic fit estimates with judgments, it is particularly important to create the dataset such that there is a good range of well fitting to badly fitting verb/location pairs.

Verb/Location pairs selection. As the existing evaluation set by (Ferretti et al., 2001) is quite heavily biased against verbs that can be grounded in pictures, we here chose to base our corpus on the 100 most frequent verbs found in the captions of the MS COCO (Common Objects in Context) (Lin et al., 2014) dataset (training/validation sets). The MS COCO dataset contains 123,287 images of people, animals and other objects "in context", i.e. in realistic environments. For each image, five captions are provided (see Figure 1).

As verbs are hard to recognize automatically from an image, the availability of captions, together with the situational context of the images, makes this corpus a good choice for studying relations between verbs and locations. We processed the available captions with the Natural Language Toolkit (NLTK) (Bird, 2006) to extract the verbs. Each of the extracted 100 most frequent verbs occurred in the captions of at least 100 images. For each of those images, we ran a state-of-the-art neural network based scene classifier, ResNet (He et al., 2016), trained to distinguish 365 locations of the Places365 dataset (Zhou et al., 2016)[1]. The classifier applies a Softmax function to the last fully connected layer representation distributing a probability of 1 over all 365 locations. Figure 1 provides an example MSCOCO image with associated captions, that e.g. mention the verbs *to sit* and *to work*, as well as the top 5 predicted locations from the ResNet classifier. In order to identify common likely locations for each verb, for each image,

Top-5 predicted locations:
Office
Home office
Waiting room
Reception
Computer room

MSCOCO Captions:
1. A man in a tuxedo working on a laptop.
2. There is a man sitting on a two seat bench in a tuxedo with a laptop in front of him.
3. A man in a suit using his laptop.
4. A man in formal attire sitting on bench using laptop computer.
5. A man in a tuxedo sits at a table and uses a laptop.

Figure 1: Example MSCOCO image with 5 captions and top-5 predicted locations from the ResNet classifier.

with at least two captions containing the verb, we extracted the top 5 predicted locations. This resulted for each verb in a list of locations, out of which some occur very frequently, while others are only seen once.

In order to obtain a balanced dataset, from the list of predicted locations per verb, we selected a subset of these by randomly choosing 4 top ranked (from the top 10), 2 middle ranked (from rank 11 to 20), and 4 low ranked (from rank 21 onwards) locations. The ranking of a location for a particular verb was based on increasing average probability (predicted score) of the location across all images relevant for that verb (see Equation 1). Next we selected a set of locations present in language data for the same set of 100 verbs as follows. From a chosen vocabulary of 50,000 most frequent words in the ukWaC corpus (Ferraresi et al., 2008), we selected candidate words, labeled as locations by the SENNA role labeler (Collobert and Weston, 2007), similar to Tilk et al. (2016), and excluded proper names using the Stanford Named Entity Recognizer (Finkel et al., 2005). We manually removed all non-physical locations (e.g. the Web) from the resulting set, which resulted in a total set of 423 candidate physical locations. We again selected for each verb 4 likely, 2 middle ranked and 4 unlikely locations, relying on the probabilities from the best model of Tilk et al. (2016). Overall, this resulted in a total of 2,000 verb/location pairs (100 verbs, with about 10 locations selected based on the visual domain and about 10 locations selected based on the language system).

Some of the most frequent verbs from MS COCO, which were included in our proposed dataset, are: *move*, *open*, *ride*, *walk*, *dress*, *eat*, *wait*, *serve* etc. Compared to FerrettiLoc, the verbs in the proposed dataset activate locations

[1]The ResNet classifier achieves a top-5 accuracy of 85% on the test set of Places365 dataset.

with little overlap in features. For example one could *walk* at various places which have little in common (e.g. at a beach, at home, in an office, in a restaurant, at a plaza). Additionally there was no assumption made about the sense of the verb for polysemous verbs like *serve* (*serving food*, *serving the ball during a tennis or volley ball match*). This choice of verbs made the locations per verb in our dataset more varied than those in Ferretti-Loc. Last but not the least, the locations in our dataset include general locations like *office*, *house*, *park* as well as more specific locations like *home office*, *courthouse*, *amusement park*. The vision-based subset has more fine-grained (specific) locations (from Places365 categories) and fewer general locations while the language-based subset has mainly general locations (all vocabulary words used by the language system are single words).

Human ratings. We collected 10 human judgments for each verb/location pair (20,000 judgments in total) via crowd-sourcing on Amazon Mechanical Turk using the LingoTurk software (Pusse et al., 2016). We provided Turkers with a verb V and location L and asked them: *Assume "X" (X can be a human, an animal or an object) is doing V, how common is it to happen at location L?* We followed the rating strategy of Ferretti et al. (2001), so the Turkers could rate on a scale from 1 ("extremely uncommon") to 7 ("extremely common"), and additionally introduced option 0 for "impossible". The rating "impossible" was chosen 1,227 times out of the 20,000 judgments, with the majority of these ratings falling on a set of 130 verb/location pairs including e.g. *graze/flat*, *dock/airplane cabin*. As some locations are rare words that may not be familiar to all Turkers, we allowed them to indicate whether they were unsure about any words in the question. This option was selected for 2% of ratings; exclusion of these ratings did not change results. We found that on average 46% of ratings (around 5 out of 10) for a given verb/location pair were identical. Standard deviations for verb/location ratings varied from 0.0 to 3.14, with the mean standard deviation being 1.55. Example ratings, averaged over 10 humans, for the verb *play* are shown in Table 1. As expected, the locations such as a *soccer field* or a *playground* are rated higher than e.g. a *repair shop* or a *hangar*.

Dataset statistics. Figure 2 presents a distribution of obtained human ratings over 2,000 verb/location pairs. We first average 10 human ratings for each pair, and then show how many verb/location pairs fall into each rating "category". We see that all levels of "fitness" are covered, from 0 ("impossible) to 7 ("extremely common"), as well as the intermediate levels. We also analyze agreement within the human ratings. Human judgments correlate with each other at Spearman's $\rho = 0.63$, which reflects the difficulty of the task. This number was obtained by computing the correlation between an individual rater vs. the average of the other 9. We repeated this for each of 10 raters and averaged the obtained correlations; the standard deviation was 0.02.

Table 2 highlights the key differences between our proposed dataset and Ferretti-Loc (Ferretti et al., 2001): the size (the higher number of verbs and verb/location pairs) and the fact that our verbs can be grounded in images. Our dataset is publicly available[2].

Location	Rating
Cockpit	0.6
Repair shop	1.8
Hangar	1.9
Landing deck	2.2
Market	2.3
Fishpond	3.6
Hall	4.9
Porch	5.4
Room	5.7
Street	5.8
Martial arts gym	6.0
School	6.0
Music studio	6.3
House	6.6
Field	6.7
Playground	6.9
Basketball court	6.9
Soccer stadium	7.0
Soccer field	7.0
Football stadium	7.0

Table 1: Example ratings, averaged over 10 humans, for the verb *play* and respective 20 locations, on a scale from 0 ("impossible") to 7 ("extremely common").

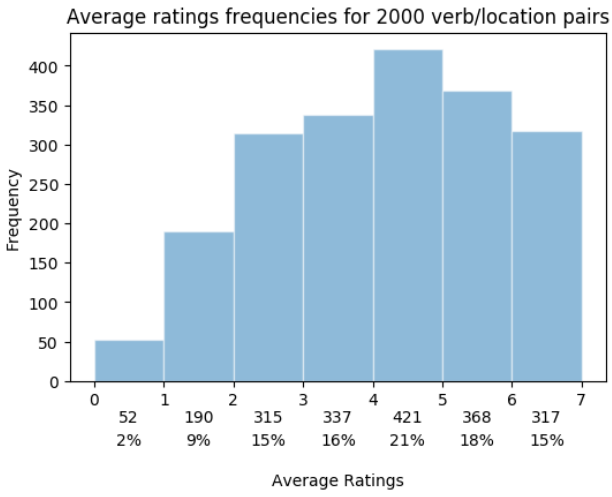

Figure 2: Histogram of the distribution of ratings, averaged over 10 humans, for all the 2,000 verb/location pairs in our dataset.

Use cases. In Section 2. we stated that our dataset allows to study the impact of vision for the task of thematic role filling, or typicality estimation. Here we discuss some of the cases where verb/location typicality can in its turn aid vision tasks. Specifically, consider the tasks of activity and scene recognition. If the visual classifiers are noisy, typicality ratings can serve as a post hoc verification to filter out untrustworthy predictions (e.g. it is unlikely to *ski* on an *ice skating rink*). If the visual classifiers are reliable

[2] *http://datasets.d2.mpi-inf.mpg.de/arohrbach/datasetV1.csv* The columns in the .csv correspond to id, verb, location and an average human rating.

Dataset	#Verbs	#Verb/Location Pairs	Vision Grounded
Ferretti-Loc	40	277	no
Our	100	2,000	yes

Table 2: Comparison of our dataset and Ferretti-Loc.

and we trust their predictions, typicality estimates could be used for anomaly detection. For instance, if we confidently recognize an action "eat" at the location "synagogue", we may decide that it is worth reporting when generating a textual description of the scene. Another way of using typicality ratings would be to incorporate them in the model at training time when learning to predict activities and locations jointly (e.g., for situational recognition; (Yatskar et al., 2016b)). We leave the experimental validation of these use cases to future work.

4. Baseline systems for location typicality estimation

We have built our dataset in a way that allows us to study whether visual information can help to improve location typicality estimates over language-only models. In this section we introduce three baseline systems which we benchmark on our proposed dataset.

Vision baseline. To estimate the probability of a location *L* given a verb *V* we rely on the scene classifier predictions (see also Section 3.). For each image where *V* occurs in at least two of its captions, we obtain the probability for the location *L* from the scene classifier, and then average probabilities across all images relevant for that verb, to obtain an estimate of how prototypical *L* is given *V*:

$$P(L|V) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{cat=L\}} \circ score_i, \qquad (1)$$

where cat is a category in Places365, N is the number of images where *V* occurs in at least two captions, $\circ$ is the element-wise vector product, and $score_i$ is the vector output of the Softmax function.

Language baseline. As a baseline for a system trained on language data, we use the current state of the art model for this task, Tilk et al. (2016). To mitigate the domain shift between MS COCO captions and the ukWaC corpus, we used the domain adaptation method of Chelba and Acero (2006). The model gained 2% in performance from this domain adaptation. Domain adaptation benefited in particular animal related locations which are highly frequent in the MS COCO dataset (e.g., *pasture*, *corral*).

Simple multimodal baseline. We also build a **"Vision+Language"** system, which averages the probabilities from both systems, to study their complementarity.

Results. Table 3 shows the correlations between average human judgments and each of the three baseline systems' predictions. Whenever a system failed to make an estimate on the full dataset (because the relevant target location was not part of its vocabulary), we used the average probability as a generic value. This allowed us to estimate performance on the complete dataset. To make a detailed analysis of the baseline systems, we divided the dataset into three subsets: a subset with 1,353 verb/location pairs for which a language system can estimate fit (**"Ratings-language"**), a subset with 1,311 pairs for which a vision system can make predictions (**"Ratings-vision"**), and an overlapping set of 664 verb/location pairs (**"Ratings-overlap"**) for which both systems can make predictions. We can see that the multimodal system consistently improves over the unimodal systems (on **"Ratings-all"** and **"Ratings-overlap"**). A substantial gap to human performance remains.

Dataset **Ratings**	# ratings	Vis. sys.	Lang. sys.	Vis.+Lang. sys.	Human corr.
-all	2,000	0.28	0.25	0.32	0.63
-vision	1,311	0.36	-	-	0.63
-language	1,353	-	0.30	-	0.63
-overlap	664	0.41	0.27	0.43	0.64

Table 3: Spearman's ρ correlation between mean human ratings and systems' estimates, on our dataset, see text for details.

System	all	overlap-vis
Baroni & Lenci, 2010	0.23	
Greenberg et al., 2015	0.29	
Language	0.44	0.50
Vision	0.30	0.42
Vision+Lang	0.47	0.54

Table 4: Spearman's ρ correlation between mean human ratings and systems' probabilities on Ferretti-Loc dataset. The language system used here is identical to Tilk et al., 2016.

Error analysis. We further analyzed our results to see where the complementarity between the language and vision systems comes from. A typical pattern we observed was that there are cases where human ratings are high but language predictions are low, confirming our initial hypothesis that language-only models may down-rate common locations. Some examples of such cases (shown here with the average human rating) are: *cook/delicatessen: 5.7*, *ride/street: 6.4*, *display/supermarket: 6.5* or *graze/farm: 6.9*. In these cases, the vision system provides valuable complementary information.

A common failure case for the vision system is related to a lack in abstraction: more general locations, e.g. *house* are underrated compared to more specific ones which are present in the set of target locations, e.g., *beach house*. In the joint system, the language system can sometimes help to recover from such errors, e.g. for *sleep/house: 6.0*.

Another type of error is when the "Vision+Language" system predicts a high score, but human ratings are low. E.g. *fly/soccer field* is rated low by humans, while the "Vision+Language" rates it high due to a large support in images and captions mentioning the *soccer ball flying*. It appears that the human raters are biased towards animate sub-

jects (or a more agentive interpretation of flying) so they assigned a low rating to these cases.

Results on Ferretti-Loc dataset. While the previous experiment on our own dataset has served as a proof of concept, providing evidence that the vision system is complementary to the language system, and that the combination of the two may improve performance, we now proceed to test whether this finding holds up for a previously used dataset for this task. Table 4 presents an evaluation on the Ferretti-Loc dataset. We see that the "Vision system" performs well on this general dataset, and that the joint "Vision+Language" system consistently improves over the language-only baseline. Note that only 29 (overlap-vis) out of 40 (all) verbs in this dataset occur in MS COCO captions, and only 9 of them occur frequently. We again used average location probability on MS COCO for out of vocabulary pairs. The "Vision+Language" system relies on averaged probabilities when vision predictions are available, otherwise it uses the language probabilities. On the "overlap-vis", both systems can make predictions. These results provide further support for the idea that the vision and language systems are complementary.

5. Conclusions

We have presented a new dataset for thematic role filling, targeting the *location* role, which is significantly larger than the prior work (Ferretti et al., 2001). Our data collection relies on image captions and visual scene classifiers, and allows for different types of approaches to be evaluated and compared. We show three different baselines for the task of predicting typical locations for a verb. Our experiments support the hypothesis that the visual scene probabilities provide useful cues for typical location prediction, and are complementary to the language estimates. The multimodal baseline performs substantially better than the unimodal baselines.

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *ACL*, pages 86–90.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Bird, S. (2006). Nltk: the natural language toolkit. In *COLING/ACL on Interactive presentation sessions*, pages 69–72.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(2014):1–47.

Chelba, C. and Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech and Language*, 20(4):382–399.

Collobert, R. and Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. In *ACL*, volume 45, page 560.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *ACL*, volume 45, page 216.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370.

Gordon, J. and Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM.

Greenberg, C., Sayeed, A. B., and Demberg, V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *HLT-NAACL*, pages 21–31.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66.

Levesque, H. J., Davis, E., and Morgenstern, L. (2011). The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

McRae, K., Ferretti, and Liane Amyote, T. R. (1997). Thematic roles as verb-specific concepts. *Language and cognitive processes*, 12(2-3):137–176.

Misra, I., Lawrence Zitnick, C., Mitchell, M., and Girshick, R. (2016). Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

Padó, U., Crocker, M. W., and Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Pusse, F., Sayeed, A., and Demberg, V. (2016). Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *TACL*, 1:25–36.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of com-

monsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Sayeed, A. and Demberg, V. (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *First Italian Conference on Computational Linguistics (CLiC-it 2014)*.

Sayeed, A., Demberg, V., and Shkadzko, P. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Computational Linguistics*, 1(1).

Sayeed, A., Greenberg, C., and Demberg, V. (2016). Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 99–105.

Silberer, C. and Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *ACL*, pages 721–732.

Tandon, N., Hariman, C., Urbani, J., Rohrbach, A., Rohrbach, M., and Weikum, G. (2016). Commonsense in parts: Mining part-whole relations from the web and image tags. In *AAAI*, pages 243–250.

Tilk, O., Demberg, V., Sayeed, A. B., Klakow, D., and Thater, S. (2016). Event participant modelling with neural networks. In *EMNLP*.

Van de Cruys, T. (2014). A neural network approach to selectional preference acquisition. In *EMNLP*, pages 26–35.

Vandekerckhove, B., Sandra, D., and Daelemans, W. (2009). A robust and extensible exemplar-based model of thematic fit. In *EACL*, pages 826–834.

Yatskar, M., Ordonez, V., and Farhadi, A. (2016a). Stating the obvious: Extracting visual common sense knowledge. In *NAACL-HLT*, pages 193–198.

Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016b). Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, pages 5534–5542.

Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., and Oliva, A. (2016). Places2: A large-scale database for scene understanding. *arXiv:1610.02055*.

7. Language Resource References

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.