Stating the Obvious: Extracting Visual Common Sense Knowledge

Mark Yatskar[1], **Vicente Ordonez**[2], **Ali Farhadi**[1,2]
[1]Computer Science & Engineering, University of Washington, Seattle, WA
[2]Allen Institute for Artificial Intelligence (AI2), Seattle, WA
`[my89, ali]@cs.washington.edu, vicenteor@allenai.org`

Abstract

Obtaining common sense knowledge using current information extraction techniques is extremely challenging. In this work, we instead propose to derive simple common sense statements from fully annotated object detection corpora such as the Microsoft Common Objects in Context dataset. We show that many thousands of common sense facts can be extracted from such corpora at high quality. Furthermore, using WordNet and a novel submodular k-coverage formulation, we are able to generalize our initial set of common sense assertions to unseen objects and uncover over 400k potentially useful facts.

1 Introduction

How can we discover that bowls can hold broccoli, that if a knife touches a cake then a person is probably cutting cake, or that cutlery can be on dining tables? We propose to leverage the effort of computer vision researchers in creating large scale datasets for object detection and use these resources instead to extract symbolic representations of visual common sense. The knowledge we compile is physical, not commonly covered in text and more exhaustive than what people can usually produce.

Our focus is particularly on visual common sense, defined as the information about spatial and functional properties of entities in the world. We propose to extract three types of knowledge from the Microsoft Common Objects in Context dataset (Lin et al., 2014) (MS-COCO), consisting of 300,000 images, covering 80 objects, with object segments and natural language captions. First, we find spatial relations, e.g. *holds*(bed, dog), from outlines of co-occurring objects. Next, we construct entailment rules like *holds*(bed, dog) $\Rightarrow$ *laying-on*(dog, bed) by associating spatial relations with text in captions. Finally, we uncover general facts such as *holds*(furniture, domestic animal), applicable to object types not present in MS-COCO by using WordNet (Miller, 1995) and a novel submodular k-coverage formulation.

Evaluations using crowdsourcing show our methods can discover many thousands of high quality explicit statements of visual common sense. While some of this knowledge can be potentially extracted from text (Vanderwende, 2005), we found that from our top 100 extracted spatial relations, e.g. *holds*(bed, dog), only 4 are present in some form in the *AtLocation* relations in the popular ConceptNet (Speer and Havasi, 2013) knowledge base. This shows that the knowledge we derive provides complimentary information for other more general knowledge bases. Such common sense facts have proved useful for query expansion (Kotov and Zhai, 2012; Bouchoucha et al., 2013) and could benefit entailment (Dagan et al., 2010), grounded entailment (Bowman et al., 2015), or visual recognition tasks (Zhu et al., 2014).

2 Related Work

Common sense knowledge has been predominately created directly from human input or extracted from text (Lenat et al., 1990; Liu and Singh, 2004; Carlson et al., 2010). In contrast, our work is focused on visual common sense extracted from images anno-

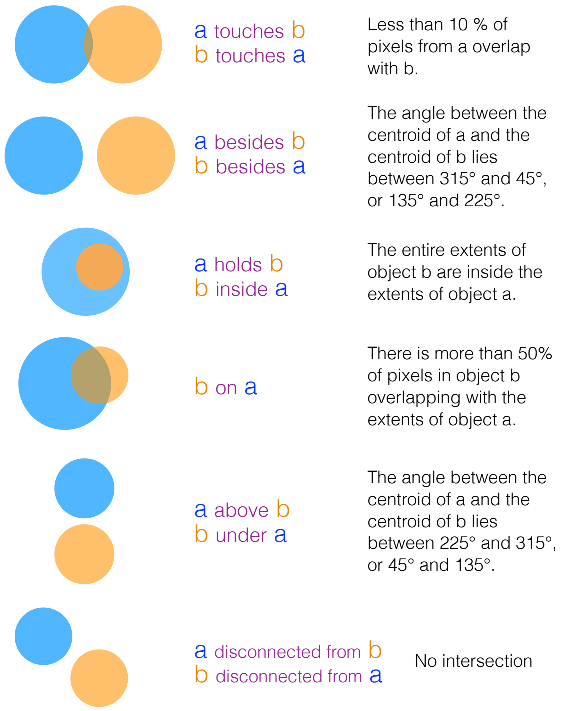

a touches b
b touches a
Less than 10 % of pixels from a overlap with b.
a besides b
b besides a
The angle between the centroid of a and the centroid of b lies between 315° and 45°, or 135° and 225°.
a holds b
b inside a
The entire extents of object b are inside the extents of object a.
b on a
There is more than 50% of pixels in object b overlapping with the extents of object a.
a above b
b under a
The angle between the centroid of a and the centroid of b lies between 225° and 315°, or 45° and 135°.
a disconnected from b
b disconnected from a
No intersection

Figure 1: We define 6 types of unique relationships: {touches, above, besides, holds, on, disconnected}.

tated with regions and descriptions.

There has also been recent interest in the vision community to build databases of visual common sense knowledge. Efforts have focused on a small set of relations, such as *similar to* or *part of* (Chen et al., 2013). Webly supervised techniques (Divvala et al., 2014; Chen et al., 2013) have also been used to test whether a particular object-relation-object triplet occurs in images (Sadeghi et al., 2015). In contrast, we use seven spatial relations and allow natural language relations that represent a larger array of higher level semantics. We also leverage existing efforts on annotating large scale image datasets instead of relying on the noisy outputs of a computer vision system.

On a technical level, our methods for extracting common sense facts from images rely on Pointwise-Mutual Information (PMI), analogous to other rule extraction systems based on text (Lin and Pantel, 2001; Schoenmackers et al., 2010). We view objects as an analogy for words, images as documents, and object-object configurations as typed bigrams. Our methods for generalizing relations are inspired by work that tries to predict a class label for an image given a hierarchy of concepts (Deng et al., 2012; Ordonez et al., 2013; Ordonez et al., 2015). Yet our work is the first to deal with visual relations between pairs of concepts in the hierarchy by using a sub-modular formulation that maximizes the amount of coverage of subordinate categories while avoiding contradictions with an initial set of discovered common-sense assertions.

3 Methods

We assume the availability of an object-level annotated image dataset D containing a set of images with textual descriptions. Each object in an image must be annotated with: (1) a mask or polygon outlining the extents of the object, and (2) the category of the object from a set of categories V and (3) an overall description of the image.

We produce three types of common sense facts, each with an associated scoring function: (1) Object-object relationships implicitly encoded in the relative configurations between objects in the annotated image data, i.e. *on*(bed, dog) (sec 3.1) , (2) Entailment relations encoded in the relationships between object-object configurations and textual descriptions i.e. *on*(bed, dog) $\Rightarrow$ *laying-on*(bed, dog) (sec 3.2), and (3) Generalized relations induced by using the semantic hierarchy of concepts in WordNet, i.e. *on*(furniture , domestic-animal) (sec 3.3).

3.1 Mining Object-Object Relations

Our objective in this section is to score and rank a set of relations $S_1 = \{r(o_1, o_2)\}$, where r is a object-object relation and $o_1, o_2 \in V$, using a function $\gamma_1 : S_1 \rightarrow \mathbb{R}$. First, we define a vocabulary R of object-object relations between pairs of annotated objects. Our relations are inspired by Region Connection Calculus (Randell et al., 1992), and the Visual Dependency Grammar of (Elliott et al., 2014; Elliott and de Vries, 2015), details in Figure 1.

For every image, we record the instances of each of these object-object relations $r(o_1, o_2)$ between all co-occurring objects in D[1]. We use Point-wise Mutual Information (PMI) to estimate the evidence for

[1] For symmetric relations like *above*(o_1, o_2), and *under*(o_1, o_2) we only record one of the relations.

Quality

r(o₁, o₂)	holds(person, o₂)	holds(o₁, person)	r(o₁, frisbee)
holds(pizza, broccoli)	holds(person, tie)	holds(bus, person)	touches(dog, frisbee)
holds(person, tie)	holds(person, toothbrush)	holds(train, person)	touches(person, frisbee)
holds(dining table, sandwich)	holds(person, cellphone)	holds(airplane, person)	holds(dog, frisbee)
holds(dining table, broccoli)	holds(person, baseball glove)	holds(boat, person)	holds(person, frisbee)
holds(dining table, pizza)	holds(person, remote)	holds(tv, person)	besides(umbrella, frisbee)
…	…	…	…
holds(cell_phone, person)	holds(person, bench)	holds(dining table, person)	besides(person, frisbee)
above(person, bus)	holds(person, dining table)	holds(cell phone, person)	above(car, frisbee)
above(bicycle, car)	holds(person, car)	holds(chair, person)	above(person, frisbee)

Figure 2: Example of our extracted object-object relations. The first column contains the overall 3 best and worst relations ranked by PMI, the following columns show similar results for the queries: what does a person hold? what holds a person?, and what interacts with a frisbee?

each relationship triplet:

$$\gamma_1(r(o_1, o_2)) = log\frac{p[r(o_1, o_2)]}{p[r]p[(o_1, o_2)]}, \qquad (1)$$

We estimate these probabilities by counting object-object-relation co-ocurrences using existential quantifiers for every image. This means every image can at most contribute one to the count of $r(o_1, o_2)$ so that we do not exacerbate the results by images with many identical object types taken from unusual viewpoints. In Figure 2, we provide examples of our extracted object-object relations.

3.2 Mining Entailment Relations

In this section we combine our relation-based tuples mined from visual annotations (section 2) with more than 400k textual descriptions included in MS-COCO. We generate a set of entailments $S_2 = \{r(o_1, o_2) \Rightarrow z\}$, where $r(o_1, o_2)$ is an element from S_1 and z is a consequent obtained from textual descriptions. Similarly as in the previous section, we rank the relations in S_2 using a function $\gamma_2 : S_2 \to \mathbb{R}$.

We start by generating an exhaustive list of candidate consequents z. We first pre-process the image captions with the part-of-speech tagger and lemmatizer from the Stanford Core NLP toolkit (Manning et al., 2014), and remove stop words. Then we generate a list of n-length skipgrams in each caption. The set of n-skipgrams are filtered based on predefined lexical patterns[2], and redundancies are removed[3]. Skipgrams, z, are then paired with co-occurring relations, $r(o_1, o_2)$, removing pairs with the disconnected-from spatial relation (see Figure 1). Pairs are scored with the conditional probability:

$$\gamma_2(r(o_1, o_2) \Rightarrow z) = \frac{P[z, r(o_1, o_2)]}{P[r(o_1, o_2)]} \qquad (2)$$

The consequent z can take the form q, $q(o_1)$, $q(o_2)$, or $q(o_1, o_2)$, by performing a simple alignment with the arguments in the antecedent. We perform this alignment by mapping the object categories in the antecedent $r(o_1, o_2)$ to WordNet synsets, and matching any word in z to any word in the gloss set of the predicate arguments o_1 and o_2. The unmatched words in z form the relation, whereas matched words form arguments. We produce the form q if there are no matches, $q(o_1)$, or $q(o_2)$ when one argument word matches, and $q(o_1, o_2)$ when both match. Examples of discovered entailments are in Figure 3.

3.3 Generalizing Relations using WordNet

In this section we present an approach to generalize an initial set of relations, S, to objects not found in the original vocabulary V. Using WordNet we construct a superset G containing all possible parent relations for the relations in S by replacing their arguments o_1, o_2 by all their possible hypernyms. Our objective is to select a subset T from G that contains high quality and diverse generalized relations.

[2] ⟨noun, verb⟩, ⟨noun,*, verb,*, noun⟩, ⟨noun,*, preposition, *, noun⟩, ⟨noun,*, verb, preposition,*,noun⟩

[3] ⟨noun,*, verb, *, noun⟩ are collapsed to ⟨noun,*, verb, preposition,*,noun⟩.

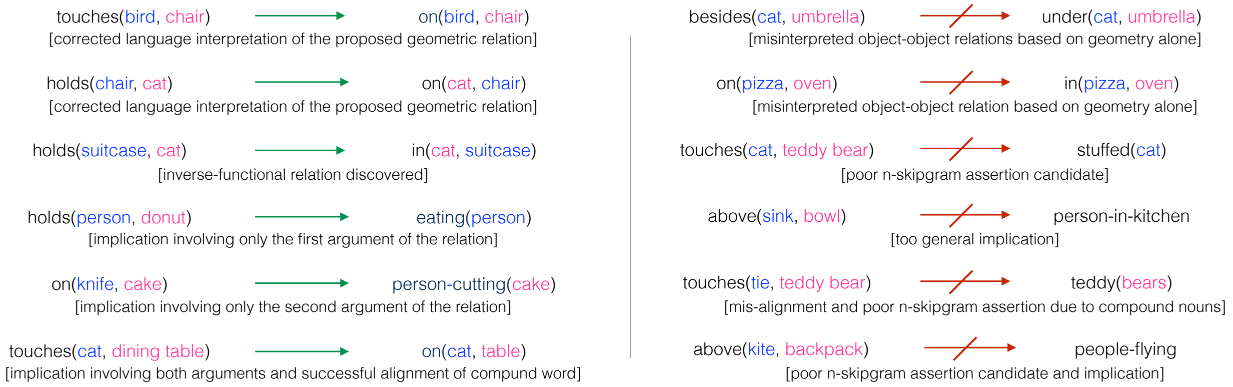

touches(bird, chair)
on(bird, chair)
[corrected language interpretation of the proposed geometric relation]
holds(chair, cat)
on(cat, chair)
[corrected language interpretation of the proposed geometric relation]
holds(suitcase, cat)
in(cat, suitcase)
[inverse-functional relation discovered]
holds(person, donut)
eating(person)
[implication involving only the first argument of the relation]
on(knife, cake)
person-cutting(cake)
[implication involving only the second argument of the relation]
touches(cat, dining table)
on(cat, table)
[implication involving both arguments and successful alignment of compund word]
besides(cat, umbrella)
under(cat, umbrella)
[misinterpreted object-object relations based on geometry alone]
on(pizza, oven)
in(pizza, oven)
[misinterpreted object-object relation based on geometry alone]
touches(cat, teddy bear)
stuffed(cat)
[poor n-skipgram assertion candidate]
above(sink, bowl)
person-in-kitchen
[too general implication]
touches(tie, teddy bear)
teddy(bears)
[mis-alignment and poor n-skipgram assertion due to compound nouns]
above(kite, backpack)
people-flying
[poor n-skipgram assertion candidate and implication]

Figure 3: Left: correctly identified entailment relations and right: failure cases.

Note that elements in G can be too general and contradict statements in S while others could be correct but add little new knowledge. To balance these concerns, we formulate the selection as an optimization problem by maximizing a fitness function $\mathcal{L}$:

$$\max_{T} \mathcal{L}(T), \text{ such that } |T| = k, \text{ and } T \subseteq G, \qquad (3)$$

$$\mathcal{L}(T) = \lambda \log(1+\psi(T)) + \sum_{t \in T} \log(1+\phi(t,S)), \qquad (4)$$

where ψ is a *coverage* term that computes the total number of facts implied through hyponym relationships by the elements in T. The second term ϕ is a *consistency* term that measures the compatibility of a generalized relation t with the relations in S. We assume that if a relation is missing from S, then it is false (this corresponds to a closed world assumption over the domain of S). Thus, ϕ is the ratio of the scores of relations in S consistent with relation t (i.e. evidence for t based on S), and a value that is proportional to the number of missing relations from S (i.e. the amount of counter-evidence). More concretely:

$$\phi(t,S) = \frac{\sum_{s:t \Rightarrow s \wedge s \in S} \gamma(s)}{\mu \cdot (1 + \sum_{s:t \Rightarrow s \wedge s \notin S} 1) \cdot d(t,S)}, \qquad (5)$$

where μ is a constant and d is the product of the WordNet distances of the synsets involved in t to their nearest synset in S. This penalizes relations that are far away from categories in S. The optimization defined in Equation 3 is an instance of the submodular k-coverage problem. We use a greedy algorithm that adds elements to T that maximize $\mathcal{L}$, which due to the submodular nature of the problem approximates the solution up to a constant factor.

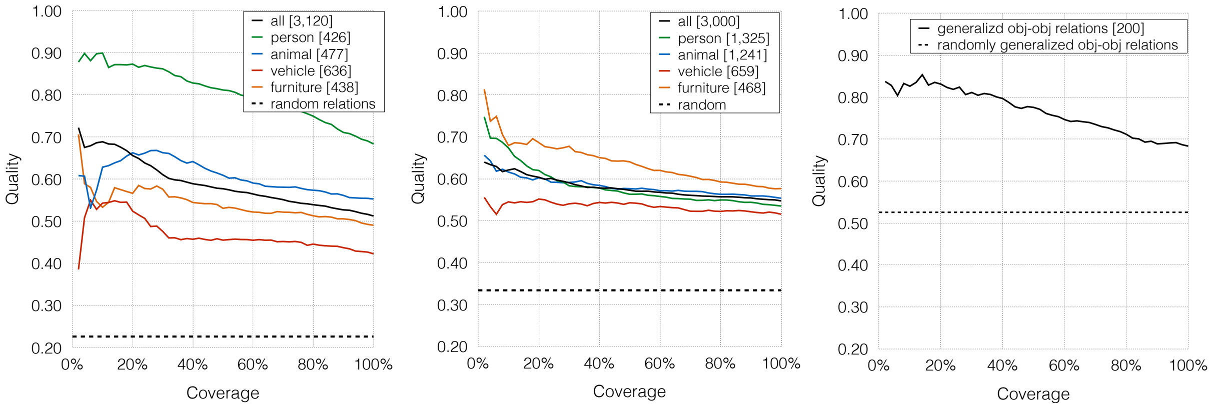
4 Experimental Setup

Object-Object Relations: We filter out from the initial set of candidate relations the ones that occur less than 20 times. We extract more than 3.1k unique statements (6k including symmetric spatial relations). *Entailment Relations*: We use skipgrams of length 2-6 allowing at most 6 skips, filter candidates such that they occur at least 5 times, and return the top 10 most likely entailments per spatial relation. Overall, 6.3k unique statements are extracted (10k including symmetric relations). *Generalized Relations*: We optimize Equation 4 only for object-object relations because the closed world assumption makes counts for implications sparse. The parameter μ is set to the average of the scores, $\lambda = 0.05$ and $k = 200$.

5 Evaluation

We evaluated the quality of the common sense we derive on Amazon Mechanical Turk. Annotators are presented with possible facts and asked to grade statements on a five point scale. Each fact was evaluated by 10 workers and we normalize their average responses to a scale from 0 to 1. Figure 4 shows plots of quality vs. coverage, where coverage means the top percent of relations sorted by our predicted quality scores.

Object-Object Relations As a baseline, 1000 randomly sampled relations have a quality of 0.225. Figure 4a shows our PMI measure ranks many high quality facts at the top, with the top quintile of the ranking being rated above 0.63 in quality. Facts about persons are higher quality, likely because this category is in over 50% of the images in MS-COCO.



(a) Object-object relations (b) Entailment relations (c) Generalizations

Figure 4: Quality of extracted common sense, as judged by people. Legends show total relations covered at 100% for a few high level types in MS-COCO.

Entailment Relations Turkers were instructed to assign the lowest score when they could not understand the consequent of the entailment relation. As a baseline, 1000 randomly sampled implications that meet our patterns have a quality of 0.33. Figure 4b shows that extracting high quality entailment is harder than object-object relations likely because supposition and consequent need to coordinate. Relations involving furniture are rated higher and manual inspection revealed that many relations about furniture imply stative verbs or spatial terms.

Generalized Relations To evaluate generalizations, Figure 4c, we also present users with definitions[4]. As a baseline, 200 randomly sampled generalizations from our 3k object-object relations have a quality of 0.53. Generalizations we find are high quality and cover over 400k objects facts not present in MS-COCO. Examples from the 200 we derive include: *holds*(dining-table, cutlery), *holds*(bowl, edible fruit) or *on*(domestic animal, bed).

6 Conclusion

In this work, we use an existing object detection dataset to extract 16k common sense statements about annotated categories. We also show how to generalize using WordNet and induced hundreds of thousands of facts about *unseen* objects. The information we extracted is visual, large scale and good quality. It has the potential to be useful for both visual recognition and entailment applications.

7 Acknowledgments

The authors would like to thank Hannaneh Hajishirzi, Yejin Choi, and Luke Zettlemoyer for helpful comments on the draft and the reviewers for useful feedback. We also thank Ani Kembhavi, and Roozbeh Mottaghi for helping to formalize visual common sense and Mechanical Turkers who helped evaluate output.

References

Arbi Bouchoucha, Jing He, and Jian-Yun Nie. 2013. Diversified query expansion using conceptnet. In *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management*, CIKM '13, pages 1861–1864, New York, NY, USA. ACM.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI Conference on Artificial Intelligence*, volume 5, page 3.

Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. 2013. Neil: Extracting visual knowledge from web data. In *International Conference on Computer Vision (ICCV)*, pages 1409–1416. IEEE.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(01):105–105.

[4] sometimes rules involve abstract concepts, for example *vessel*, any object that can be used as a container

Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3450–3457. IEEE.

Santosh Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277.

Desmond Elliott and Arjen P de Vries. 2015. Describing images using inferred visual dependency representations. *Association for Computational Linguistics (ACL)*.

Desmond Elliott, Victor Lavrenko, and Frank Keller. 2014. Query-by-example image retrieval using visual dependency representations. In *International Conference on Computational Linguistics (COLING)*, pages 109–120, August.

Alexander Kotov and ChengXiang Zhai. 2012. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 403–412, New York, NY, USA. ACM.

Douglas B Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. Cyc: toward programs with common sense. *Communications of the ACM*, 33(8):30–49.

Dekang Lin and Patrick Pantel. 2001. Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Hugo Liu and Push Singh. 2004. Conceptnet: A practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara Berg. 2013. From large scale image categorization to entry-level categories. In *International Conference on Computer Vision (ICCV)*, pages 2768–2775. IEEE.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2015. Predicting entry-level categories. *International Journal of Computer Vision*, pages 1–15.

David A Randell, Zhan Cui, and Anthony G Cohn. 1992. A spatial logic based on regions and connection. In *International Conference on Knowledge Representation and Reasoning*.

Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1456–1464.

Stefan Schoenmackers, Oren Etzioni, Daniel S Weld, and Jesse Davis. 2010. Learning first-order horn clauses from web text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1098.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer.

Lucy Vanderwende. 2005. Volunteers created the web. In *Proceedings of the 2005 AAAI Spring Symposium, Knowledge Collection from Volunteer Contributors*. American Association for Artificial Intelligence, March.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about Object Affordances in a Knowledge Base Representation. In *European Conference on Computer Vision*.