Linking WordNet to 3D Shapes

Angel X Chang, Rishi Mago, Pranav Krishna, Manolis Savva, and Christiane Fellbaum
Department of Computer Science, Princeton University
Princeton, New Jersey, USA
angelx@cs.stanford.edu, rmago19@lawrenceville.org,
pranavskrishna@gmail.com, msavva@cs.stanford.edu, fellbaum@princeton.edu

Abstract

We describe a project to link the Princeton WordNet to 3D representations of real objects and scenes. The goal is to establish a dataset that helps us to understand how people categorize everyday common objects via their parts, attributes, and context. This paper describes the annotation and data collection effort so far as well as ideas for future work.

1 Introduction

The goal of this project is to connect WordNet (Fellbaum, 1998) to 3D representations of real objects and scenes. We believe that this is a natural step towards true grounding of language, which will shed light on how people distinguish, categorize and verbally label real objects based on their parts, attributes, and natural scene contexts.

Our main motivation is to establish a dataset connecting language with realistic representations of physical objects and scenes using 3D computer-aided design (CAD) models to enable research in computational understanding of the human cognitive process of categorization.

Categorization is the process by which we group entities and events together based on salient similarities, such as shared attributes or functions. For example, the category "furniture" includes tables, chairs and beds, all of which are typical parts of a room or house and serve to carry out activities inside or around the house. Subcategories are "seating furniture," which includes chairs and sofas and "sleeping furniture," which includes beds, bunkbeds and futons. Note that some categories have a simple verbal label (a name, like "furniture"), but often category names are compounds (like "sleeping furniture"). Compounding, a universal feature of human language that accounts in part for its infinite generativity, allows us to make up names on the fly whenever we feel the need to distinguish finer-grained categories, such as "college dorm room furniture." Of course, not all languages share the same inventory of simple labels.

We form and label categories all the time. Categories help us to recognize never-before-seen entities by perceiving and assessing their attributes and functions and, on the basis of similarity to known category members, assign them to a category (Rosch, 1999). Young children in particular learn to form categories by being exposed to an increasing number of different category members and gradually learning whether they belong to one category or another. Importantly, categories allow us to reason: if we know that beds are made for lying down on and sleeping, encountering a new term like "sleigh bed" will tell us that such a bed is likely to have a flat surface on which a person can lie down. Conversely, seeing a sleigh bed for the first time and identifying this salient feature will prompt us to call it a "bed." Categorization is so fundamental to human cognition that we are not consciously aware of it; however, it remains a significant challenge for computational systems in tasks such as object recognition and labeling.

Parts, attributes, and natural contexts of objects are all involved in category formation. Objects are made of parts, and parts often imbue functionality, especially in the broad category of "artifacts." Thus, seat surfaces are a necessary part for functioning chairs. Parts and the functionality they enable are fundamentally intertwined with categorization (Tversky and Hemenway, 1984).

Beyond their concrete parts, objects are perceived to have a general set of attributes. For example, the distinction between a cup, a goblet and a mug relies less on the presence or absence of specific parts and more on the geometric differences in aspect ratio of the objects themselves.

Lastly, real objects occur in real scenes, meaning that they possess natural contexts within which they are observed. In language, context reflects a variety of aspects beyond functionality, including syntactic patterns and distributional properties. In contrast, physical context is concrete and defined by the sets of co-occurring objects and their relative arrangements in a given scene.

To study these three aspects of category formation, we will connect WordNet at the part, attribute, and contextual level with a 3D shape dataset. The longer term goal of this project is to ask how people distinguish and categorize objects based on how they name and describe the parts and attributes of the objects. We will focus primarily on the category "furniture."

2 Existing datasets

There has been much prior work linking WordNet (Fellbaum, 1998) to 2D images. The most prominent effort in this direction is ImageNet (Deng et al., 2009), which structures a large dataset of object images in accordance with WordNet hierarchies. Our project differs from ImageNet because even though Imagenet is based on the WordNet hierarchy, the focus of our project is on annotating parts and attributes on 3d models rather than the labeling of images. Following this, the SUN (Xiao et al., 2010) dataset focuses on scene images, and the VisualGenome (Krishna et al., 2016) dataset defines object relations and attributes in a connected scene graph representation within each image. Another line of work focuses on detailed annotation of objects and their parts — a prominent recent example is the Ade20K dataset (Zhou et al., 2016). However, a fundamental assumption of all this work is that objects and their properties can be adequately represented in the 2D image plane. This assumption does not generally hold, as many object parts, spatial relations between objects and a full view of object context are hard to infer from the limited field of view of a 2D image.

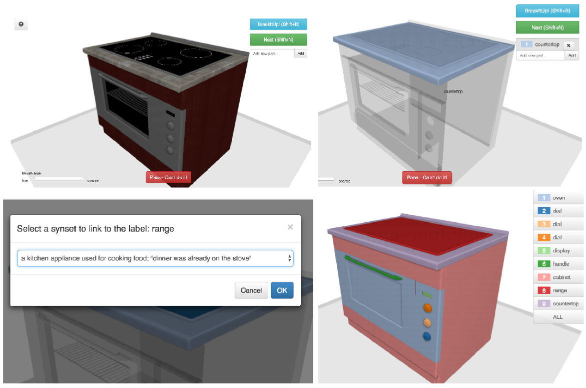
More recently, there has been some work that links 3D CAD models to WordNet. The ShapeNet (Chang et al., 2015) dataset is a large collection of CAD representations (curating close to 65,000 objects in approximately 200 common WordNet synsets), whereas the SUNCG (Song et al., 2017) dataset contains CAD representations of 45,000 houses, composed of individual objects (in 159 WordNet synsets). These two datasets are both "synthetic" in the sense that the 3D CAD representations are designed virtually by a human expert. A different form of 3D representation is obtained by scanning and 3D reconstruction of real world spaces. Recent work introduced ScanNet (Dai et al., 2017) and the Matterport3D (Chang et al., 2017) dataset, which both contain 3D reconstructions of various public and private interior spaces (containing 409 and 430 object synsets respectively).

Though both synthetic and reconstructed 3D data are increasingly available, no effort currently exists to connect such 3D representations to WordNet at the part, attribute, and contextual levels. Such a link of WordNet entries to 3D data can provide much richer information than 2D image datasets. Naturally, 3D representations allow us to reason about unoccluded parts and symmetries, arrangement of objects in a physically realistic three dimensional space, and to account for empty space, a critical property of real scenes which is not observable in 2D images. Moreover, 3D representations are appropriate for computationally simulating real spaces and the actions that can be performed within them. The ability to do this is a powerful tool for investigating and understanding actions (Pustejovsky et al., 2016). Therefore, our project aims to annotate 3D models in one of the existing datasets and link them to the appropriate synset in the WordNet database at the part, attribute, and contextual level.

3 Project description

Our project has so far focused on annotating part and attribute information on 3D CAD objects in SUNCG (Song et al., 2017) and linking them to the corresponding WordNet synsets. We are working with a preliminary categorization of the objects performed in prior work, which establishes their connection to WordNet synsets denoting physical objects. However, we plan to refine the granularity of this categorization by introducing finer-grained categories — e.g. partitioning "doors" into "garage doors" and "screen doors" among others.

We chose this dataset because the 3D objects in SUNCG (approximately 2,500) are used across a large number of 3D scenes (more than 45,000). This means that for each object, we can automatically establish many contextual observations. This property of the SUNCG dataset differs from other common 3D CAD model research datasets such



Select a synset to link to the label: range

a kitchen appliance used for cooking food; "dinner was already on the stove"

Figure 1: Interface for labeling object parts. Top left: a stove and range object is displayed to a user. Top right: the user paints the "countertop" part. Bottom left: the user links the "range" part to the corresponding WordNet synset. Bottom right: the fully annotated object with parts in different colors.

as ShapeNet (Chang et al., 2015) where each object is de-contextualized. For the latter dataset we would have to additionally compose scenes using available objects in order to acquire observation contexts for the objects.

We first augment the SUNCG objects with part annotations that are linked to WordNet synsets. We defer the assignment of attributes to the same objects as it is an easier annotation task in terms of interface design. To perform the part annotation, we designed an interface with a "paint and name" interaction where the user paints parts of the surface of an object corresponding to a distinct part and assigns a name to that part. The details of the interface and annotation task are described in the following section.

4 Annotation interface

Our interface is designed to allow for efficient annotation of parts by inexperienced workers on crowdsourcing platforms such as Amazon's Mechanical Turk. The interface is implemented in javascript using three.js (a WebGL-based graphics library). It can be accessed on the web using any modern browser, and does not require specialized software or hardware.

Figure 1 shows a series of screenshots from the interface illustrating the annotation process. The user first sees a rotating "turn table" view that reveals the appearance of the object from all directions. The user then types the name of a part in a text panel and drags over the surface of the object to select the regions corresponding to the part. The process is repeated for each part and the final result is a multi-colored painting of the object with corresponding part names for each color. The partitioning of the object's geometry into different parts is saved and submitted to a data server upon completion. Figure 2 shows a close-up of the interface while an object is in the process of being annotated.

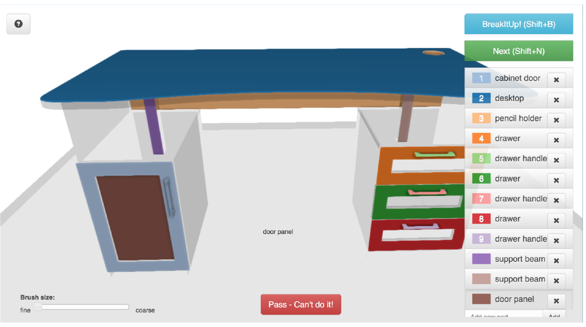


Figure 2: Partially annotated object with parts highlighted in different colors. The "door panel" label is selected, indicating that this part was just annotated.

In order to enable an efficient multi-level painting interaction, the size of the paint brush can be adjusted by the annotator. The object geometry is pre-segmented for several levels of granularity: segmentation into surfaces with the same material assignment, a segmentation with a loose surface normal distance criterion, and finally a segmentation into sets of topologically connected components of the object geometry. This multi-level painting allows the speed of labeling to be adjusted to accommodate both small parts (e.g., door handles) and large parts (e.g., countertops on kitchen islands).

The annotators use freeform text for the part names, requiring that we address the problem of mapping this part name text to a WordNet synset. We implemented a simple algorithm that restricts the candidate synset set to physical objects in the WordNet hierarchy, preferring furniture (since we are dealing with indoor scenes that are predominantly composed of furniture). Given the object category, we can additionally use the meronym relations in WordNet to suggest and rerank possible part synsets.

After applying this algorithm to connect each part name's text to a WordNet synset, we manually verify and if needed fix the inferred link using an interface that displays the WordNet synset assignment for the given part (in the same view as the part annotation view) and allows the user to select a different synset.

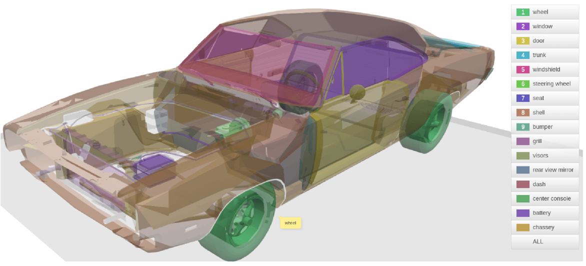


Figure 3: A fully completed part annotation example for a car 3D object. Different colors correspond to distinct parts with corresponding names provided by the annotator on the right.

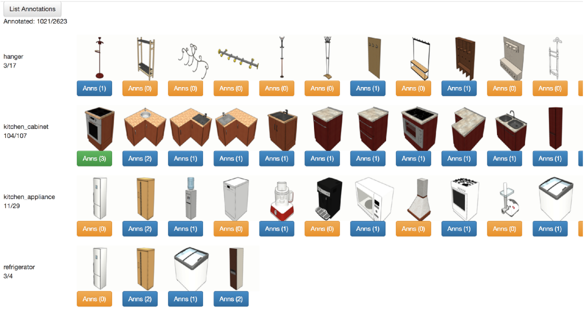

List Annotations
Annotated: 1021/2623
hanger
3/17
Anns (1) Anns (0) Anns (0) Anns (0) Anns (0) Anns (0) Anns (1) Anns (0) Anns (1) Anns (0) Anns (0)
kitchen_cabinet
104/107
Anns (3) Anns (2) Anns (1) Anns (1) Anns (1) Anns (1) Anns (1) Anns (1) Anns (1) Anns (1) Anns (1)
kitchen_appliance
11/29
Anns (0) Anns (2) Anns (1) Anns (0) Anns (1) Anns (0) Anns (1) Anns (0) Anns (1) Anns (0) Anns (1)
refrigerator
3/4
Anns (0) Anns (2) Anns (1) Anns (2)

Figure 4: List of 3D models to be annotated. This is the interface through which one selects a model to annotate or can examine a previously annotated model.

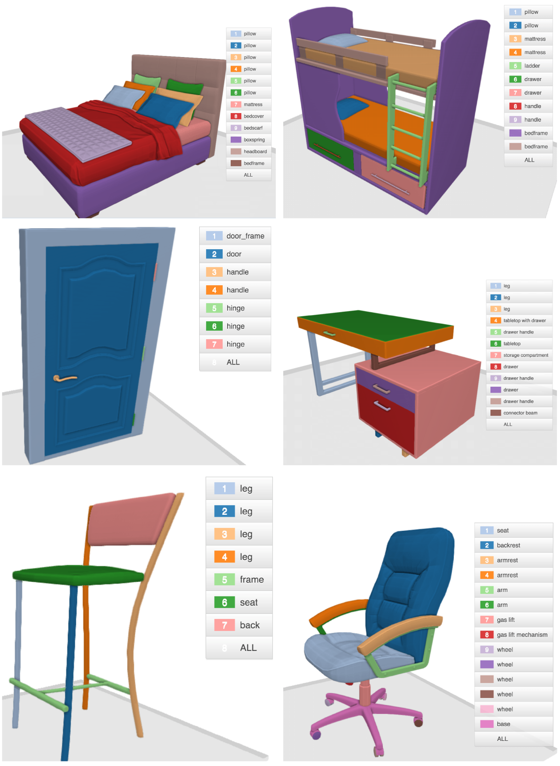


Figure 5: Example objects annotated with parts. Objects are assigned to the following WordNet synset, from top left: `double_bed.n.01`, `bunk_bed.n.01`, `door.n.01`, `desk.n.01`, `straight_chair.n.01`, `swivel_chair.n.01`.

5 Initial annotations and statistics

Five persons have worked on the annotation thus far. Annotating one 3D model takes roughly five minutes on average, with time being mostly a function of the complexity of the particular object. To maintain consistency while annotating the objects in the database, the annotators were instructed to name the geometric and functional parts of the object, not decorations or stylistic elements (e.g., a picture of a fish on a lamp shade).

Using the interface described above, we have so far collected more than 100,000 part instance annotations. An example of a car annotated by a crowd worker is shown in Figure 3. SUNCG includes a total of 2547 models, of which we have so far annotated 1021 during the prototyping process for developing our interface (Figure 4). Figure 5 shows several other example objects with their part annotations visualized.

6 Limitations

The initial stages of the annotations were limited by two major factors: the quality of the 3D models, and language in general. A handful of the models in the database had segmentation issues, i.e., any error in how the model was broken up into regions. For example, a segmentation issue for a table is encountered when more than one leg is connected into one region, thus making it impossible even with the smallest brush size to label the individual legs. To solve this problem, the segmentation algorithm must be improved upon. A more widespread problem with the database, however, was the lack of internal structure in many of the models. For example, in book cases with doors, only the doors would be present in the model, while the internal structures — in this case the shelves — were omitted.

Some linguistic factors can cause the annotation to be less than straightforward. For example, should the link be to the British or the U.S. word? If we annotate a coffee cup, should we annotate the piece of cardboard around the cup that is designed to protect our hand from the heat of the coffee using its specific but rare name "zarf," or should we choose a more common but less specific term such as "sleeve" or even "piece of cardboard?" Moreover, some parts do not have proper labels: what should we call the beams that help stabilize some tables other than "support beams?" See Figure 6 for an example.

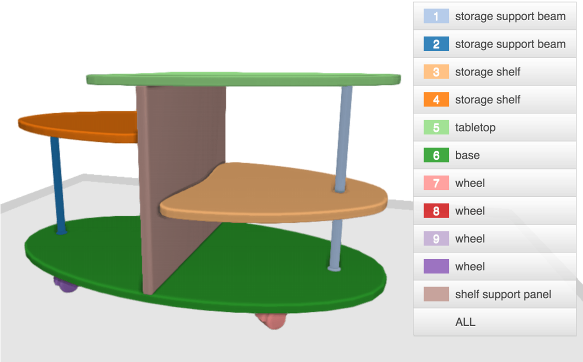


Figure 6: An example demonstrating some of the limitations of our annotation system. Parts of the table are identified as "storage support beam" and "shelf support panel" due to lack of a better term.

7 Future Work

Our work so far has focused on developing the infrastructure and annotation interfaces to collect 3D object part annotations at scale. This part data linked to WordNet is of tremendous potential value, which we plan to investigate as our project continues.

A very interesting direction of work is in building contextual multimodal embeddings. Many object parts and attributes are rarely mentioned in language. For example, a stool doesn't have a back, but people don't refer to stools as "backless chairs." Neither do speakers encode the fact that chairs often have four legs or 5 wheels; only non-default exemplars might be labeled in an ad hoc fashion as "five-legged chairs," for example. Furthermore, the physical contexts of objects (see Figure 7) provides richer information than is found in text. In this regard, the text and 3D modalities are complementary and provide an excellent target for building multimodal distributional representations (Bruni et al., 2014). Multimodal embeddings are a promising semantic representation which has been leveraged for various Natural Language Processing and vision tasks (Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015; Kottur et al., 2016).

Another direction for future work is to leverage the object part and attributes and their correspondences to WordNet to go beyond the set of WordNet synsets and automatically induce new senses, along the lines of recent work on sense induction (Chen et al., 2015; Thomason and J. Mooney, 2017). For example, we have found that WordNet synsets do not have good coverage of some fairly modern categories of objects that we observe in our 3D object datasets, including iPads, iPhones and various electronic devices such as game consoles.

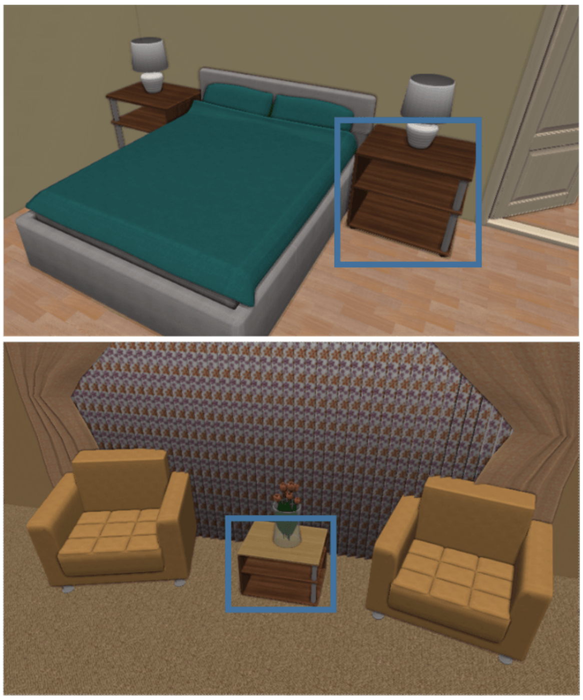


Figure 7: An example of the same nightstand object (outlined in blue) in two different 3D scene contexts. A contextual embedding afforded by the full 3D representation of the scene within which the nightstand is observed would be a powerful way to analyze and disentangle different usage contexts for common objects.

References

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An information-rich 3D model repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. 2015. Sense discovery via co-clustering on images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5298–5306.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.*

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR).*

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.

Satwik Kottur, Ramakrishna Vedantam, Jose M. F. Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

James Pustejovsky, Tuan Do, Gitit Kehat, and Nikhil Krishnaswamy. 2016. The development of multimodal lexical resources. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 41–47.

Eleanor Rosch. 1999. Principles of categorization. *Concepts: core readings*, 189.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pages 721–732.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition.*

Jesse Thomason and Raymond J. Mooney. 2017. Multi-modal word synset induction. In *IJCAI.*

Barbara Tversky and Kathleen Hemenway. 1984. Objects, parts, and categories. *Journal of experimental psychology: General*, 113(2):169.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442.*