Collecting Spatial Information for Locations in a Text-to-Scene Conversion System

Masoud Rouhizadeh[1], Daniel Bauer[2], Bob Coyne[2], Owen Rambow[2], and Richard Sproat[1]

[1] Oregon Health & Science University, Portland OR, USA,
[2] Columbia University, New York NY, USA

Abstract. We investigate using Amazon Mechanical Turk (AMT) for building a low-level description corpus and populating VigNet, a comprehensive semantic resource that we will use in a text-to-scene generation system. To depict a picture of a location, VigNet should contain the knowledge about the typical objects in that location and the arrangements of those objects. Such information is mostly common-sense knowledge that is taken for granted by human beings and is not stated in existing lexical resources and in text corpora. In this paper we focus on collecting objects of locations using AMT. Our results show that it is a promising approach.

Keywords: Text-to-Scene Systems, Amazon Mechanical Turk, Lexical Resources, VigNet, Location Information, Description Corpora

1 Introduction

Our aim is to populate **VigNet**, a comprehensive semantic resource that we will use in a text-to-scene generation system. This system follows in the footsteps of Coyne and Sproat's WordsEye [2], but while WordsEye did only support a very limited number of actions in a static manner and mostly accepted low-level language as input (*John is in front of the kitchen table. A cup is on the table. A plate is next to the cup. Toast is on the plate*) the new system will support higher-level language (*John had toast for breakfast*).

VigNet is based on FrameNet[1] and contains lexical, semantic and spatial/graphical information needed to translate text into plausible 3D scenes. In VigNet frames are decomposed into subframes and eventually into primitive spatial relations between frame participants (frame elements), describing one way a frame can be depicted graphically. We call a frame that is decomposable into such primitives a **vignette**. Even though the technical details are not crucial to understand this paper we refer the interested reader to [4].

This paper deals with the collection of spatial information to populate VigNet. Even though VigNet contains vignettes for actions and other events, complex objects and situations, this paper focuses only on the induction of **location vignettes**. Knowledge about locations is of great importance to create detailed scenes because locations define the context in which an action takes place. For instance when someone takes a shower he usually does so in the bathroom, interacting with the 'affordances' provided by this room (i.e. shower cabin, curtain, shower head, shower tap etc.) in a specific way. Note that location vignettes can, but do not have to be evoked by lexical items. We can say *John took a shower in the bathroom*, but this seems redundant because bathrooms are the preferred location for *taking a shower*. VigNet records knowledge of this type that can be accessed in the text-to-scene generation process.

In this paper we propose a methodology for collecting semantic information for locations vignettes using Amazon Mechanical Turk (AMT). The next section first discusses location vignettes in more detail. We then review related work in section 3. We describe how we use AMT to build an image description corpus and collect semantic information for locations in section 4 and compare different methods in an evaluation. Section 5 concludes.

2 Location Vignettes

As mentioned before, location vignettes are important because they provide the context in which actions can take place. Locations involve the spatial composition of several individual objects. For example, in *'John sat in the living room'*, we might expect the living room to contain objects such as a sofa, a coffee table, and a fireplace. In addition, these objects would be spatially arranged in some recognizable manner, perhaps with the fireplace embedded in a wall and the coffee table in front of the sofa in the middle of the room. In order to represent such locations graphically we are adding knowledge about the typical arrangements of objects for a wide variety of locations into VigNet.

Any given location term can potentially be realized in a variety of ways and hence can have multiple associated vignettes. For example, we can have multiple location vignettes for a *living room*, each with a somewhat different set of objects and arrangement of those objects. This is analogous to how an individual object, such as a *couch*, can be represented in any number of styles and realizations. Each location vignette consists of a list of constituent objects (its frame elements) and graphical relations between those objects (by means of frame decomposition). For example, one type of living room (of many possible ones) might contain a couch, a coffee table, and a fireplace in a certain arrangement.

```
LIVING-ROOM_42(left_wall, far_wall, couch,
                         coffee_table, fireplace)

TOUCHING(figure:couch, ground:left_wall)
FACING(figure:couch, ground:right_wall)
FRONT-OF(figure:coffee_table, ground: sofa)
EMBEDDED(figure:fire-place, ground:far_wall)
```

The set of graphical primitives used by location vignettes control surface properties (color, texture, opacity, shininess) and spatial relations (position, orientation, size). This set of primitive relations is sufficient to describe the basic spatial layout of most locations (and scenes taking place in them). Generally we do not record information about how the parts of a location can be used in an action, but rather consider this knowledge to be part of the action.

3 Related work

Existing lexical and common-sense knowledge resources do not contain the spatial and semantic information required to construct location vignettes. In a few cases, WordNet [5] glosses specify location-related information, but the number of such entries with this kind of information is very small, and they cannot be used in a systematic way. For example, the WordNet gloss for *living room* (*a room in a private house or establishment where people can sit and talk and relax*) defines it in terms of its function, not its constituent objects and spatial layout. Similarly, the WordNet gloss for *sofa* (*an upholstered seat for more than one person*) provides no location information. FrameNet [1] is focused on verb semantics and thematic roles and provides little to no information on the spatial arrangement of objects.

More relevant to our project is OpenMind [8] where online crowd-sourcing is used to collect a large set of common-sense assertions. These assertions are normalized into a couple dozen relations, including the typical locations for objects. The list of resulting objects found for each location, however, is noisy and contains many peripheral and spurious relations. In addition, even the valid relations are often vague and represent different underlying relations. For example, a *book* is declared to be located *at a desk* (the directly supporting object) as well as *at a bookstore* (the overall location). In addition, like most existing approaches, it suffers from having objects and relations being *generalized* across all locations of a given type and hence is unable to represent the dependencies that would occur in any given *specific* location. As a result, there's no clear way to reliably determine the main objects and disambiguated spatial relations needed for location vignettes.

LabelMe [7] is a large collection of images with annotated 2D polygonal regions for most elements and objects in a picture. It benefits from the coherence of grounding the objects in specific locations. It suffers, though, from the lack of differentiation between main objects and peripheral ones. Furthermore, it contains no 3D spatial relations between objects.

One of the well-known approaches for building lexical resources is automatic extracting lexical relations from large text corpora. For a comprehensive review of these works see [6]. However, a few works focus specifically on extracting semantic information for locations, including [10] and [11], which use the vector-space model and a nearest-neighbor classifier to extract locations of objects. Also directly relevant to this paper is work by Sproat [9] which attempts to extract associations between actions and locations from text corpora. This approach provides some potentially useful information, but the extracted data is noisy and requires hand editing. In addition, it extracts locations for actions rather than the objects and spatial relations associated with those locations.

Furthermore, much of the information that we are looking for is commonsense knowledge that is taken for granted by human beings and is not explicitly stated in corpora. Although structured corpora like Wikipedia do mention associated objects, they are often incomplete. For example in the Wikipedia entry for *kitchen* there is no mention of a *counter* or other surface on which to prepare food but the picture that goes with the definition paragraph (labeled "A modern Western kitchen") clearly has one.

In this paper we investigate using Amazon Mechanical Turk (AMT) for building a low-level description corpus for locations and for directly collecting objects of locations vignettes. We will compare the accuracy of collected data to several gold standard vignettes generated by an expert. We show that we can tune our information collection method to scale for large number of locations.

4 Using AMT to build location vignettes

In this section we discuss how we use Amazon Mechanical Turk (AMT) to build a *location description corpus* and for collecting the *typical objects of location vignettes*. AMT is an online marketplace to co-ordinate the use of human intelligence to perform small tasks such as image annotation that are difficult for computers but easy for humans. The input to our AMT experiments are pictures of different rooms. By collecting objects and relations grounded to specific rooms we capture coherent sets of dependencies between objects in context and not just generalized frequencies that may not work together. In each task we collected answers for each room by five workers who were located in the US and had previous approval rating of 99%. Restricting the location of the workers increases the chance that they are native speakers of English, or at least have good command of the language. We carefully selected input pictures from the results of image searches using the Google and Bing search engines. We selected photos that show 'typical' instances of the room type, e.g. room instances which include typical large objects found in such rooms. Photos should show the entire room. We then defined the following task:

Task 1: Building low-level location description corpus: In this task, we asked AMT workers to provide simple and clear descriptions of 85 pictured room. We explicitly asked AMT workers that their descriptions had to be in the form of naming the main elements or objects in the room and their positions in relation to each other, using verbs such as *is* or *are* (i.e. *linking verb*). Each description had to be very precise and 5 to 10-sentence long. Our collected description corpus contains around 11,000 words.

In order to extract location information from the low-level location description corpus, the text is first processed using the NLP module of WordsEye. We extracted the objects and other elements of locations which are mainly in the form of RELATION–GROUND–FIGURE and extract the objects and elements which are represented as FIGURE or GROUND. We then further processed the extracted locations as is explained in sub-section 4.1.

Task 2: Listing functionally important objects of locations: According to this criterion, the important objects for a room are those that are required in order for the room to be recognized or to function in this way. One can imagine a *kitchen* without a *picture frame* but it is rarely possible to think of a *kitchen* without a *refrigerator*. Other functional objects include a *stove*, an *oven*, and a *sink*. We asked workers to provide a list of functional objects using an AMT hit such as the one shown in figure 1. We showed each AMT worker an example room with a list of objects and their counts. We gave the following instructions:

" Based on the following picture of a **kitchen** list the objects that you really need in a **kitchen** and the counts of the objects.

1. In each picture, first tell us how many room doors and room windows do you see.
2. Again, don't list the objects that you don't really need in a **kitchen** (such as magazine, vase, etc). Just name the objects that are absolutely required for this **kitchen**. "

Task 3: Listing visually important objects of locations: For this task we asked workers to list large objects (furniture, appliances, rugs, etc) and those that are fixed in location (part of walls, ceilings, etc). The goal was to know which objects help define the basic structural makeup of this particular room instance. We used the AMT input form shown in figure 1 again, provided a single example room with example objects and and gave the following instruction:

" What are the main objects/elements in the following **kitchen**? How many of each?

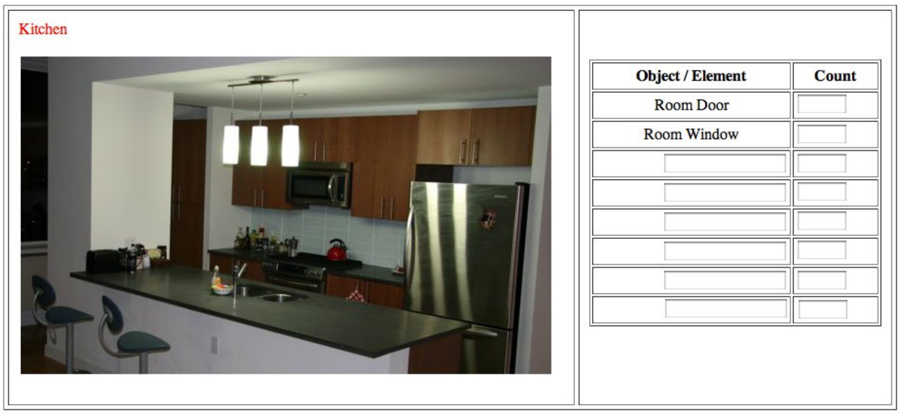
1. In selecting the objects give priority to:
   – Large objects (furniture, appliances, rugs, etc).
   – Objects that are fixed in location (part of walls, ceilings, etc).

   The goal is to know which objects help define the basic makeup and structure of this particular kitchen.
2. In each picture, first tell us how many room doors and room windows do you see. "

4.1 Post-processing of the extracted object names from AMT

We post-processed the extracted objects from the location description corpus and the objects that were listed in tasks 2 and 3 in the following steps:

1. Manual checking of spelling and converting plural nouns to singular.
2. Removing conjunctions like"*and*", "*or*", and "/". For example, we converted "*desk and chair*" to "*desk*" and "*chair*".
3. Converting the objects belonging to the same WordNet synset into the most frequent word of the synset. For example we converted *tub*, *bath*, and *bathtub* into *bathtub* with frequency of three.
4. Finding the intersection from the inputs of five workers and selecting the objects that listed three times or more



Kitchen

Object / Element	Count
Room Door	
Room Window	

Fig. 1. AMT input form to collect functionally important objects (task 2) or visually important (task 3) objects in locations. Workers are asked to enter the name of each object type and the object count.

5. Finding major substrings in common: some input words only differ by a space or a hyphen character such as *night stand*, *night-stand*, and *nightstand*. We convert such variants to the simplest form i.e. *nightstand*.
6. Looking for head nouns in common: if the head of the compound noun input such as *projector screen* can be found in another single-word input i.e. *screen*, we assume that both refer to the same object i.e. *screen*.
7. Recalculating the intersections and selecting the objects with frequency of three or more.

4.2 Evaluation

For evaluating the results we manually built a set of gold standard vignettes (GSVs) for 5 rooms which include A) a list of objects in each room, and B) the arrangements of those objects. Selected objects for GSVs are the ones that help define the basic makeup and structure of the particular room. We are comparing the extracted object from AMT tasks against the list of objects in the GSVs.

Table 1 shows the comparison of the AMT tasks against GSVs. The "Extracted Objs" row shows the number of objects we extracted from each AMT tasks for 5 rooms. The "Correct Objs" row shows the number of extracted objects from AMT that are present in our GSVs of each room and the precision score derived based on that. The "Expected Objs" row shows the number of all the objects in GSVs that we expected the workers to list, and the recall score based on that.

AMT Task	Free Description		Functional		Visual	
Extracted Objs	39		32		32	
Correct Objs	26	Pre: 67%	28	Pre: 87%	29	Pre: 91%
Expected Objs	33	Rec: 79%	33	Rec: 85%	33	Rec: 88%

Table 1. The accuracy of each AMT tasks for the objects of 5 rooms compared to GSVs. (See the above paragraph for the definition of rows and columns.)

5 Conclusion and future work

In this paper we explored different approaches to populate VigNet, a resource containing spatially grounded lexical semantics, with locational information (location vignettes) using Amazon Mechanical Turk. In one approach we used AMT to collect a low-level description corpus for locations. We then used the Words-Eye NLP module to extract the objects from each description. For comparison we asked AMT workers to directly list objects of locations shown in photographs, either based on visual or on functional criteria. We then post-processed the extracted objects from each experiment and compared them against gold standard location vignettes.

We have shown that we can extract reasonably accurate objects from processing the description corpus as well as spatial relations and arrangements of objects. The results achieved using the functional and visual object listing tasks approximate the gold standard even better, with the visual elicitation criterion outperforming the functional one.

In current work, due to the good results on the small training set we are using the visual object listing paradigm to induce descriptions of 85 rooms. We are planing to collect vignettes for a variety of other indoor and outdoor locations.

Location vignettes also contain the spatial arrangement of objects. In addition to the extracted relations from the description corpus, we also designed a series of AMT tasks for determining the arrangements of objects in different locations using the objects that we collected in the present work. For each room we ask AMT workers to determine the arrangements of the previously collected objects in that particular room. For each object in the room, workers have to determine its spatial relation with A) *one wall* of the room and B) *one other object* in the room. We did not include the results in this paper since we are still exploring methods to evaluate the *spatial arrangements* task. The gold standard location vignettes include arrangements of objects, but it is difficult to directly compare the gold standard to the AMT workers' inputs as there are different possibilities to describe the same spatial layout.

References

1. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet Project. COLING-ACL (1998)
2. Coyne, B., Sproat, R.: Wordseye: An automatic text-to-scene conversion system. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, Los Angeles, CA, USA, pp. 487- 496, (2001)
3. Coyne, B., Rambow, O., Hirschberg, J., Sproat, R.: Frame semantics in text-to-scene generation. In R. Setchi, I. Jordanov, R. Howlett, and L. Jain (Eds.), Knowledge-Based and Intelligent Information and Engineering Systems, Volume 6279 of Lecture Notes in Computer Science, pp. 375-384. Springer Berlin / Heidelberg (2010)
4. Coyne, B., Bauer, D., Rambow, O.:VigNet: Grounding Language in Graphics using Frame Semantics. In ACL Workshop on Relational Models of Semantics, (2011)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
6. Girju, R., Beamer, B, Rozovskaya, A., Fister, A., Bhat. S.:A knowledge-rich approach to identifying semantic relations between nominals. Information Processing and Management, vol. 46, no. 5, pp. 589-610, (2010)
7. Russell, B. C. , Torralba, A., Murphy, K. P., and Freeman, W. T.: LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, vol. 77, no. 13, pp. 157173, May (2008).
8. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. Proceedings of Recent Advances in Natural Languge Processing (2007)
9. Sproat, R.: Inferring the environment in a text-to-scene conversion system. First International Conference on Knowledge Capture, Victoria, BC (2001)
10. Turney,P.,Littman,M.:Corpus-based Learning of Analogies and Semantic Relations. Machine Learning Journal 60 (1-3), pp. 251-278. (2005)
11. Turney, P.: Expressing implicit semantic relations without supervision. In: Proceedings of COLING-ACL, Australia (2006)