Grounded Language Understanding using Crowd-Sourced Data

Rishabh Sachdeva
University of Maryland Baltimore County
Baltimore, Maryland
rishabs1@umbc.edu

ACM Reference Format:
Rishabh Sachdeva. Grounded Language Understanding using Crowd-Sourced Data,. 5 pages.

1 INTRODUCTION

Grounded Language acquisition is concerned with learning the meaning of language as it applies to physical world[8]. A lot of work is being done in this area to make robots understand natural language in the world to which it refers. Data is required to train machine learning models, to develop a mapping of the natural language sentences, or tokens, to physical world context. Many research projects use crowd-sourced platforms like Amazon Mechanical Turk to gather text data. The major differences between spoken data and written data lie in spontaneity and organization. Written data is deliberately planned and is not as spontaneous when compared with spoken data. When trained on spoken data, there exists a potential to enhance human-robot interaction experience when the robot is deployed in real-world scenarios dealing with voice data. The aim of this paper is to develop grounded language acquisition system using image data set of everyday objects. We collect natural language descriptions both in text and speech of these objects using Amazon Mechanical Turk. This work is inspired by and designed to complement the UW RGB-D dataset: this dataset extends beyond its scope by augmenting the data collection process to include natural language descriptions. We evaluate the descriptions for mentions of color, shape, and object name descriptors as well as length and tokens used. We collect textual as well as spoken data using Amazon MTurk and develop Grounded Language System using "Word as a classifier" approach.

2 BACKGROUND

To enhance the interaction of humans with physically situated agents, it is necessary that robots can understand or relate the physical world with the language. Matuszek* et al. [6] presents a way using a joint language and perception model to acquire groundings for language. Active learning offers a promising approach to extend the idea and has the potential for more efficient learning [7]. The mentioned work uses text data describing object categories gathered using Amazon Mechanical Turk. The work presented in this paper is inspired from [6][7][13] and [12]. We use a similar strategy on more enhanced and complex data set.

2.1 Logistic Regression

Logistic regression is the most popular predictive analysis technique which is widely used for classification problems. Logistic Regression provides a way to predict the probability of certain feature set belonging to a particular class. Logistic regression tends to limit the predictions in range of 0 to 1. In order to map the predictions to probabilities, **Sigmoid Function** is used (Fig. 3).$h_\theta(x)$ is defined as the estimated probability that y belongs to positive class on given input x, and θ are the controlling weights.[9] If model provides out y = 0.7, it means that the model predicts that feature set x belongs to a positive class (y=1) with probability 0.7. In the grounded learning joint model, we are using, logistic regression is used to train the tokens towards the positive corresponding images. For example, positive images corresponding to a token "yellow" could be banana, lime or some yellow colored block. Related equations are as follows:

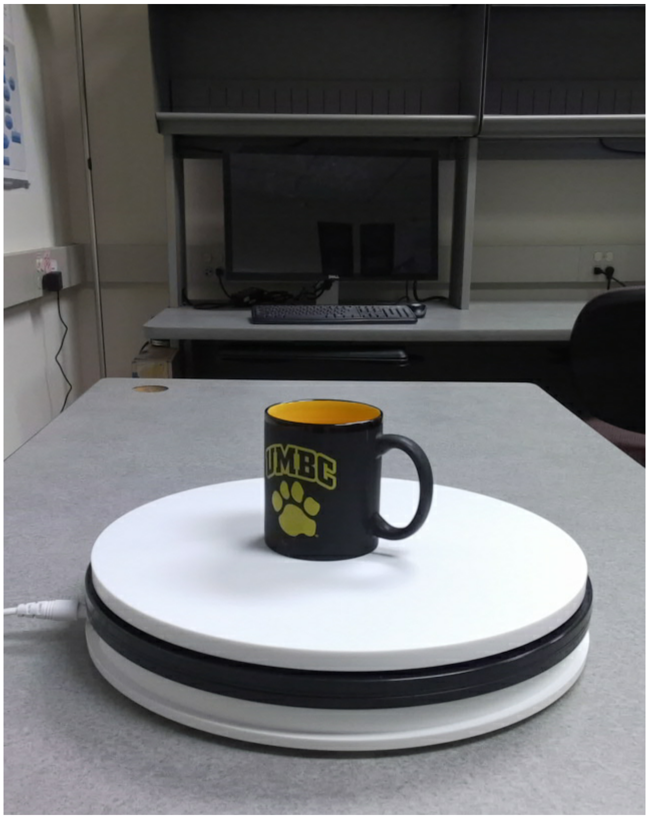

Figure 1: Sample image frame

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

,
. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

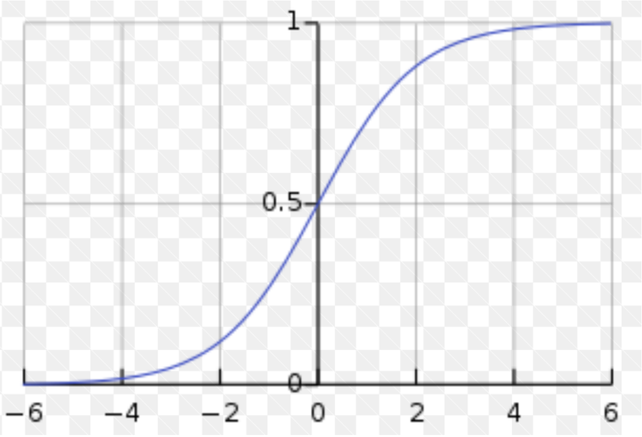

Figure 2: Sample Sigmoid Function (https://en.wikipedia.org/wiki/Sigmoid_function)

Topic	Classes of Objects
food	*potato, soda bottle, water bottle, apple, banana, bell pepper, food can, food jar, lemon, lime, onion*
home	*book, can opener, eye glasses, fork, shampoo, sponge, spoon, toothbrush, toothpaste, bowl, cap, cell phone, coffee mug, hand towel, tissue box, plate*
medical	*band aid, gauze, medicine bottle, pill cutter, prescription medicine bottle, syringe*
office	*mouse, pencil, picture frame, scissors, stapler, marker, notebook*
tool	*allen wrench, hammer, measuring tape, pliers, screwdriver, lightbulb*

Table 1: The classes of objects that appear in our dataset, grouped by their high-level category.

2.2 TF*TDF

The concept of TF*IDF is also necessary to understand the working of the model. Keeping the focus on our study, TF or Term Frequency refers to the frequency of word or token used to describe an object, or, how often a token is used to describe a particular object. Let's say, a token "red" is used many times to describe an object like tomato, then "red" should be closely related to the corresponding object. IDF or Inverse Document Frequency defines how often a token appears in all the descriptions of all the objects. The more often token appears, IDF parameter decreases. Some terms like "the", "or", "and" does not contribute much to the meaning or description of an object. These terms are called stop words and they may vary according to the corpus. So, both TF and IDF parameters are necessary to judge the relevance of token.

2.3 Lemmatization

Lemmatization and Stemming are important techniques when data pre-processing is involved. Lemmatization is utilized to remove conjugations from the word. The goal of lemmatizers is to replace the various variants of a word (baking, baked) with their corresponding root word (bake).

2.4 Stemming

Stemmers does not aim to replace various variants of a word by a correctly spelled root word. Instead, it removes the affixes attached to the word. Reduction of a word to their stem/root/base form is the core idea. One example can be converting "baking,baked,bake" to "bak".

3 RELATED WORK

Grounded Language acquisition is an active field and lot work is happening in this area. Mooney talked about a system that is able to connect words, phrases, and sentences to its perception of objects in the world [11]. Thompson et al. came up with Human-Robot Dialogue to improve Grounded Natural Language Understanding [16]. They presented how clarification dialogues as a potential approach to enhance the end-user experience. She and Chai show that interactive learning approach leads to more reliable models for grounded verb semantics. [4] combines the linguistic and auditory information into multi-modal representations. Motivated by human concept acquisition, Kiela and Clarke [4] came up with the approach of learning the meaning of the object(say, some musical instrument) not only by visual properties but also by auditory information like sound, pitch, and timber. Fleischman and Deb describe a way to improve speech recognition using Grounded Language Modelling [1]. Their contribution was combining visual context with speech recognition to set up a relationship between the words and non-linguistic context. Chen and Ballard present a way to accomplish a complex task of word learning by utilizing speaker's intentional body movements like head movements, gaze, and hand movements, to establish relations between word and their grounded meanings [17]. Matuszek et al. came up with the joint model of Language and Perception for Grounded Attribute Learning [6]. They extended this work by incorporating the advantages of interactive labeling [7]. In [7], an initial pilot is being described that uses active learning to ask annotations from the human subject. Work is also being done to obtain negative examples of natural language annotations[13]. [12] describes an unsupervised system that learns visual classifiers associated with the words using semantic similarity to choose negative examples from the corpus of perceptual and linguistic data. Thompson et al. came up with an interesting approach of utilizing haptic, auditory and proprioceptive data, along with visual details to learn the groundings of natural language sentences and words [15]. In [15], the system learns to ground natural language words describing objects via human-robot I spy game. Considerable amount of work is done in this area where robot learns under human supervision [7] [16] [15]. In [7] and [12], robot aims to learn groundings via description provided by humans incorporating Active Learning.

4 METHOD

4.1 Data Corpus

Before diving into details, it's necessary to understand the data corpus used. Our data set contains a total of 47 categories that belongs to high level topics like food items, medical and tool equipment. Each category has around 5 instances. For example, 'plate' category can contain different instances like blue plate and red plate. And, each instance has 4 or 5 images taken from different angles. In total, there are 826 images for which data was collected. This image dataset is prepared by members of IRAL lab at UMBC.

Amazon Mechanical Turk which is a crowd sourced data collection platform, is used for this purpose. We collected Textual as well as Spoken data for training. This data is collected by me and other researchers at IRAL lab at UMBC.

Text Description Collection

Text descriptions are gathered from Amazon Mechanical Turk. We split the raw image videos into frames and select frames that capture the object from different angles. Workers are then asked to give a description of the object as if they are describing it to another person in one to two short sentences. Workers are asked not to describe the turntable itself or the background. The multiple angles diversify what each worker sees of each object and gives us a chance to collect descriptions from objects not oriented in a conventional manner. It is important to gather such descriptions since a person talking to a robot may have a partial view or understanding of the object and thus need to ground it using atypical language.

Each Mechanical Turk task includes five images and we assign each task a total of ten times. Therefore, each object has a total of 40 text descriptions from different orientations that can be used for grounded language learning.

Speech Description Collection

The motive of collecting audio data is to capture the nuances between spoken and written language. It is common for people to restructure sentences before writing them, but while speaking, we do not have the liberty to re-frame or restructure them. Therefore, spoken sentences may be less well framed or be grammatically incorrect. We support speech with body gestures, eye gaze, expressions or pitch of the voice, details that are missing in writing. Experienced writers may be able to overcome these differences while communicating. However, these people usually hold formal education [10]. So, to enhance human-robot interaction for a broader group of end users, it will be necessary to train robots with spoken data. Moreover, it is possible that this unorganized and spontaneous obtained data can prepare the robot even better for real-world scenarios. We develop a user interface to collect spoken natural language data using MediaStream recording API.[1] A similar approach is reported in recent work [18, 19] to collect data using web-based and mobile application-based systems. We embed the interface into Amazon Mechanical Turk, and the recorded audio files are collected from these tasks and stored in AWS S3 buckets.

Text Data Example	Speech Data Example
This coffee mug bears a logo evocative of university athletics teams.	that is a black and gold coffee mug
It is a dark blue coffee cup with a college symbol on it.	a coffee cup with a paw print on it

Table 2: Examples of the text and speech descriptions for the object in Figure 1.

We use the FFmpeg library[2] to add the missing metadata from the audio files to make them compatible with ASR systems. The audio files are then converted to text using Google's Speech to Text API. Each Mechanical Turk task includes one image and we assigned five assignments for each task. We collect 4059 audio descriptions in total.

4.2 Feature Extraction of Images-Transfer Learning

Transfer learning is a strategy where already trained models are used as a starting point for the second model. This strategy is utilized for collecting image features for this project. VGG 16 is a popular convolutional neural network model which accepts fixed size 224 x 224 RGB image to be classified [20]. The pre-trained VGG 16 is used to collect RGB features of 826 images by not including the top layer which is actually a softmax prediction layer. From the input layer to the last max pooling layer is regarded as the feature extraction part of the model, while the rest of the network is regarded as the classification part of the model.

4.3 Data Processing

Once all the data is collected, nltk stopwords are removed from the object descriptions as they does not actually convey any crucial information about the specific object. Further, the data is stemmed for further experiments. Snowball stemmer is used for this purpose. Then, it is important to identify the crucial tokens which can be informative groundings for the objects. So, TF-IDF strategy is used to find out the most crucial tokens. Most relevant 25 tokens were identified for experiments. Most of them were color, or related to other properties of the objects. It turns out that people color as most common attribute to describe any object. Some such tokens are:

```
yellow
red
light
box
metal
pink
```

After cleaning and pre-processing of this data, inverted index is developed for the purpose of identifying positive and negative data points. The 826 images were randomly split in 80-20 ratio for training and testing. Implementation-wise, inverted index is a

[1] https://developer.mozilla.org/en-US/docs/Web/API/MediaStream_Recording_API

[2] https://www.ffmpeg.org/

dictionary which has token as key and value as list of images for which corresponding token is used in description. Let's look at one of such entries in inverted index, red token was used in descriptions of images like:

```
potato_3\potato_3_7.png
food_jar_1\food_jar_1_1.png
toothpaste_4\toothpaste_4_11.png
apple_5\apple_5_7.png
hand_towel_3\hand_towel_3_6.png
allen_wrench_1\allen_wrench_1_1.png
potato_3\potato_3_1.png
```

Inverted index is used to prepare positive and negative training images for tokens. Positive images for each of the crucial tokens are the ones in which that corresponding token is used in their description. Similarly, negative points for a token refers to images where that token was never used by MTurkers.

4.4 GLS Machine Learning Model - Word as a Classifier

After data processing, we have positive and negative data images for each of the identified crucial tokens according to their TF-IDF scores. A Logistic Regression model is trained corresponding to each of the crucial tokens. So, we have 25 classifiers to find out the groundings of test images. Also, we have the ground truth descriptions for these images. In testing time, we find out the intersecting tokens between their ground truth and the identified crucial tokens. Ideally, our models should give positive classification for each such overlapping token. We following 'Ask Questions' strategy while testing. Lets' say we have classifiers corresponding to tokens - red, metal, round, small. Then we ask 4 questions for the testing image (say apple).

```
Is it red?
Is it related to metal?
Is it round?
Is it small?
```

If our model give positive classifications (probability>0.5) for red, round and small, then these are the associated groundings for the test image according to the model.

5 PERFORMANCE METRIC AND RESULTS

We find out accuracy according to the number of overlapping tokens identified as associated grounding by our model. All the classifiers which misclassify the image are counted as False Positives. Some examples are below:

```
    image: can_opener_4/can_opener_4_16.png
Identified Groundings: silver, open, handle

    image: lemon_2/lemon_2_16.png
Identified Groundings:  small, yellow, round

    image : food_jar_3/food_jar_3_16.png
Identified Groundings: pepper, yellow, red, glass
```

The achieved accuracy is 68.67%. Out of 418, 287 grounded tokens are classified correctly.

6 CONCLUSION AND FUTURE WORK

The results achieved in this experiment are not very impressive. To achieve the desired results, more sophisticated Machine Learning models can be developed. Deep Learning would be definitely an interesting approach for the Grounded Language Learning task. This will be our next step on the same dataset. Also, we got good amount of corrupt data, especially the audio recording descriptions that affects the model performance. It would be interesting to see how people provide audio descriptions in more controlled environment. An interesting approach would be to incorporate "Confirmation Dialogue Strategy" in the data collection process. Thompson et al. presented Human-Robot Dialogue Dialogue as a potential approach to improve Grounded Natural Language Understanding [16]. Dialogues are difficult to implement, maintaining context of the conversation and speaking right thing at right time is challenging when robots are involved. But, this strategy has a potential to overcome the troublesome behavior of speech-to-text.

7 CITATIONS AND BIBLIOGRAPHIES

REFERENCES

[1] M. Fleischman and D. Roy. "Grounded Language Modeling for Automatic Speech Recognition of Sports Video." In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008.

[2] A. Julius, "Web Speech API," KTH R. Inst. Technology, 2013.

[3] V. Këpuska and G. Bohouta,"Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx),"Int. Journal of Engineering Research and Application, 2017.

[4] D. Kiela and S. Clark, "Multi- and Cross-Modal Semantics Beyond Vision:Grounding in Auditory Perception," Proceedings of Conference on Empirical Methods in Natural Language Processing, 2017.

[5] K. Lai, L. Bo, X. Ren, and D. Fox, "Rgb-d object recognition: Features, algorithms, and a large scale benchmark." In Consumer Depth Cameras for Computer Vision: Research Topics and Applications, pages 167–192, 2013.

[6] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. "Joint Model of Language and Perception for Grounded Attribute Learning," In International Conference on Machine Learning (ICML), 2012.

[7] C. Matuszek, N. Pillai and K.K. Budhraja "Improving Grounded Language Acquisition Efficiency Using Interactive Labeling," In Robotics: Science and Systems Workshop on Model Learning for Human-Robot Communication, 2016.

[8] C. Matuszek, "Grounded Language Learning: Where Robotics and NLP Meet," In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018.

[9] Andrew Ng, http://cs229.stanford.edu/notes/cs229-notes1.pdf

[10] J. Miller, "Spoken and Written English," Bas Aarts , and April McMahon , eds.The Handbook of English Linguistics. Oxford: Blackwell,673-679, 2006.

[11] R.J. Mooney, "Learning to connect Language and Perception." In Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI),2008.

[12] N. Pillai and C. Matuszek, "Unsupervised Selection of Negative Examples for Grounded Language Learning," In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[13] N. Pillai and C. Matuszek, "Identifying Negative Exemplars in Grounded Language Data Sets," Robotics: Science and Systems Workshop on Spatial-Semantic Representations in Robotics, 2017.

[14] G. Redekar, "On differences between spoken and written language," Discource Processes, 1984.

[15] J. Thompson, J. Sinapov, M. Svetlik, P. Stone and R.J. Mooney, "Learning Multi-Modal Grounded Linguistic Semantics by Playing I Spy," IEEE International Conference on Robotics and Automation, 2018.

[16] J. Thompson, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P.Stone and R.J. Mooney, "Improving Grounded Natural Language Understanding through Human-Robot Dialog," To appear in IEEE International Conference on Robotics and Automation (ICRA), 2019.

[17] C. Yu and D.H. Ballard, "A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions." In ACM Transactions on Applied Perception.

[18] I Lane, A. Waibel, M. Eck and K. Rottmann, "Tools for Collecting Speech Corpora via Mechanical-Turk." In Proceedings of the NAACL HLT

[19] K. A. Lee,A. Larcher, G. Wang, P. Kenny, N. Brummer, D. V. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, J. Perez, "The RedDots Data Collection for Speaker Recognition.", In 15th

Annual Conference of the International Speech Communication Association (INTERSPEECH)

[20] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", In ICLR, 2015