Optimal Semantic Distance for Negative Example Selection in Grounded Language Acquisition

Nisha Pillai, Francis Ferraro, Cynthia Matuszek
npillai1 | ferraro | cmat @ umbc.edu
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County, Baltimore, Maryland 21250

I. Overview and Related Work

Grounded language acquisition, in which the meanings of utterances are learned from and with respect to the physical world, is often treated as a data-driven machine learning problem. For a robot, obtaining negative examples of language referents is a challenging problem: people tend to describe things that are true of a situation, rather than negatives about it [15, 4, 3]. For example, humans may are unlikely to describe an apple as "not a banana." Previous methodologies to acquire negative examples include explicit prompting [13, 2] or crowd-sourcing [14, 5]. Other work selects negatives randomly from the dataset, sometimes omitting those with descriptions that overlap those of the object being learned [12, 1].

Our prior research [9] addressed this in an unsupervised system that learned language using a "words-as-classifiers" approach [7], using semantic similarity to automatically choose negative examples from a corpus of perceptual and linguistic data. This joint model of grounded language acquisition is based on the idea that descriptions of physically similar objects should be nearby in semantic space. We used the well-known paragraph vector (PV) [6, 8] encoding to embed these object descriptions in a semantic vector space. Cosine similarity was then used to discover the semantic distance between vector representations: the angle between the PV representations of descriptions was treated as an object similarity metric. Objects with the largest cosine distance were chosen as negative data points. This work is comparable to the unsupervised label identification of Roy [11], but uses document similarity instead of clustering.

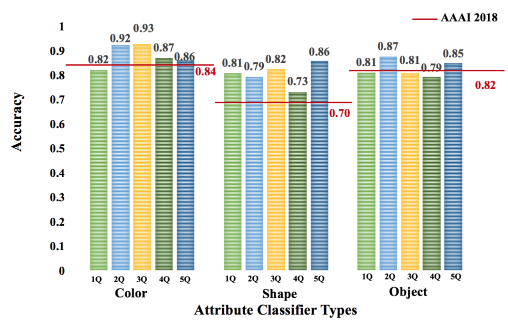

Fig. 1. Language acquisition performance of different attributes using negative examples drawn from quintiles of semantic similarity distance (near to far). The horizontal line represents learning performance using a fixed threshold[9].

II. Preliminary Results

The most immediate outstanding question regarding this approach is how to choose an appropriate distance (angle) to select negative examples; when training classifiers, the least-similar object is often not the ideal choice. Previously, it was chosen empirically; our current research aims to more rigorously find and describe a suitable way of selecting negative samples in the semantic embedding space. Our initial findings show that objects which are semantically closer but with non-overlapping characteristics give better results in our grounded language learning experiments.

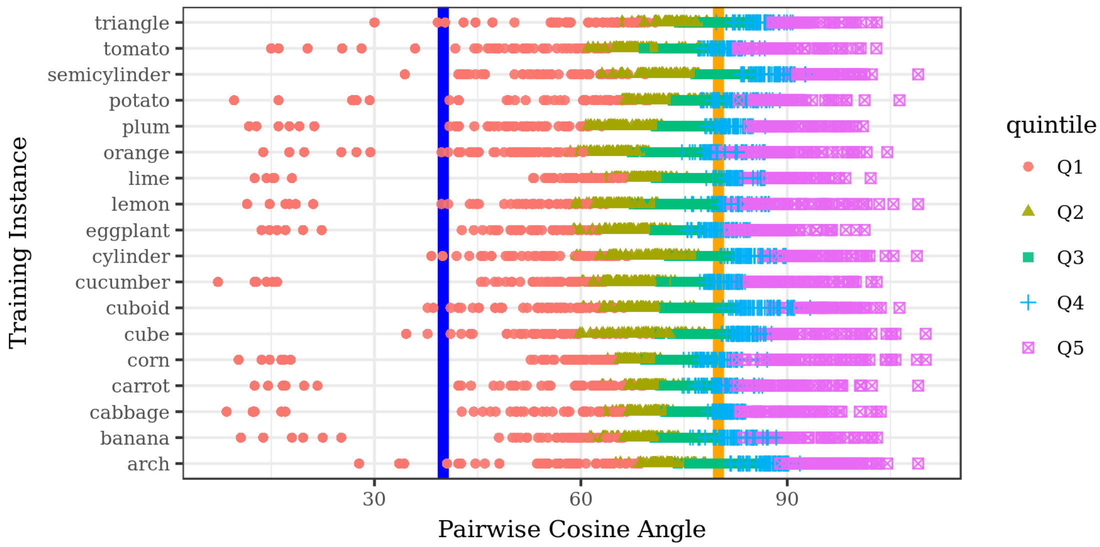

Fig. 2. The distribution of cosine angle between pairs of all training instance semantic document representations. The orange line represents an empirical threshold [9], while the blue line indicates a threshold that divides the space by density.

Negative example selection was conducted on a dataset containing 72 objects in 18 distinct categories (such as food and toys). We collected 6,000 descriptions of these objects from Amazon Mechanical Turk and chose as negative examples objects whose cosine distance was greater than 80 degrees, an empirically chosen threshold [9]. Language groundings were created using on a joint model of language and vision [7, 10] that jointly trained classifiers to predict linguistic descriptions from an object's visible features.

In this work, we aim to understand the implications of drawing negative examples from different parts of the semantic embedding space. We conducted experiments on selecting negative examples from different areas of the cosine distance space, which we initially divided into quintiles. Figure 2 shows the cosine angle distance between a selected object on the left to the remaining objects in the dataset (dots along the line). The orange line in the figure marks the threshold selected as part of previous research, where everything to the right of the orange line was selected as a negative example. In contrast, the blue line suggests a possibly more optimal threshold for selecting best negative examples for language acquisition tasks. The objects to the right of the blue line represent the "most different" objects in semantic space.

We trained our visual classifiers on color, shape, and object features with every section of the negative training data selected from different quintiles. Figure 1 depicts our initial results in which color, shape, and object language acquisition show promising predictive performance when selecting the 2nd and 3rd quintiles of objects as negative samples. These results suggest a more dynamically-chosen threshold may yield improved performance.

III. Discussion and Future Work

Our aim is to build a model which selects the most informative negative data points from the complete negative dataset. Efficient selection of semantic distance will underpin this grounded language learning by providing negative examples without prompting the user explicitly, reducing the number of (possibly repetitive) questions. A thorough evaluation including Mechanical Turk user studies will be conducted to ensure the effectiveness of the model. We are also considering alternative methods of semantic similarity to further improve the performance of our model.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant No. 1657469.

References

[1] Grzegorz Chrupala, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Association for Computational Linguistics*, 2017.

[2] Haris Dindo and Daniele Zambuto. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010.

[3] Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1021. URL http://www.aclweb.org/anthology/D15-1021.

[4] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge extraction. In *AKBC*, 2013.

[5] Ross A Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. Recovering from failure by asking for help. *Autonomous Robots*, 2015.

[6] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML-14)*, 2014.

[7] Cynthia Matuszek*, Nicholas FitzGerald*, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, June 2012.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[9] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *In the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[10] Nisha Pillai, Karan K Budhraja, and Cynthia Matuszek. Improving grounded language acquisition efficiency using interactive labeling. In *Robotics: Science and Systems workshop on Model Learning for Human-Robot Communication*, 2016.

[11] Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 2002.

[12] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[13] Stefanie Tellex, Pratiksha Thaker, Joshua Joseph, and Nicholas Roy. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 2013.

[14] Stefanie Tellex, Ross A Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In *Robotics: Science and systems*, 2014.

[15] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.