Building Language-Agnostic Grounded Language Learning Systems

Caroline Kery,[1] Nisha Pillai,[2] Cynthia Matuszek,[2] and Francis Ferraro[2]

***Abstract*— Learning the meaning of *grounded* language—language that references a robot's physical environment and perceptual data—is an important and increasingly widely studied problem in robotics and human-robot interaction. However, with a few exceptions, research in robotics has focused on learning groundings for a single natural language pertaining to rich perceptual data. We present experiments on taking an existing natural language grounding system designed for English and applying it to a novel multilingual corpus of descriptions of objects paired with RGB-D perceptual data. We demonstrate that this specific approach transfers well to different languages, but also present possible design constraints to consider for grounded language learning systems intended for robots that will function in a variety of linguistic settings.**

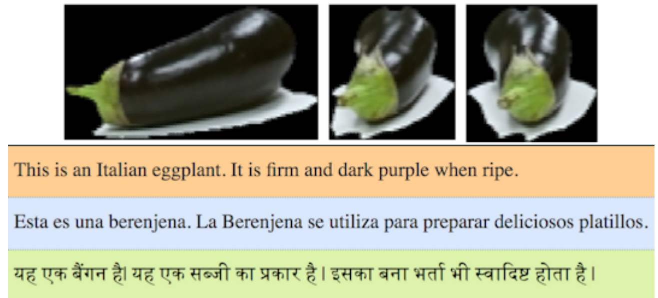

Fig. 1. An example of data collected for grounded language learning. *Top:* Several images of an object in the dataset. *Below:* Descriptions of the object in unconstrained English, Spanish, and Hindi (respectively).

I. Introduction

As robots become less expensive and more capable, it is becoming possible to imagine them being deployed in a variety of human-centric settings such as homes, schools, or workplaces. However, as robots become more accessible, it becomes more critical that they can be communicated with and controlled by non-specialists. Robotic assistants could be extremely helpful for a wide variety of people, so for maximal accessibility it would be ideal to make the use of such robots as intuitive as possible. *Grounded language acquisition,* in which robots learn to understand language in the context of the sensed physical world around them, is a major focus of research for building robot systems that can interact with human partners in a natural way. At the same time, studying language learning and interaction with a physically situated agent offers a mechanism for advancing natural language understanding [23], [12].

Although learning shared embeddings between sensor data and language is a rich and varied field of research in robotics, most of the work has focused on systems that operate on only a single language, frequently English. In the natural language processing community, there is work on building *multilingual* systems, which learn from [10], [13] or can be used with [36], [6] more than one language simultaneously. In this work, we address a complementary problem: Can systems that take advantage of physical percepts to learn meanings in a particular language be deployed in a setting where a different language is used?

In this work, we describe the application of an existing grounded language learning system that uses the words-as-classifiers model [21], [30], [24] to a novel corpus containing two additional languages. We discuss methods of acquiring suitable training data, the overall performance of the system on English, Spanish, and Hindi, and some characteristics of a grounded language process that should be considered when linguistic flexibility is desired. In practice, such flexibility should be considered highly desirable to avoid limiting the benefits of ubiquitous, collaborative robots to English-speaking settings.

This paper is organized as follows: In Section III-A, we describe our novel trilingual dataset, which is comprised of approximately 17,000 descriptions, in three languages, of objects collected with an RGB-D sensor. In Section III-B and Section III-C, we describe applying a grounded language acquisition system designed for English [26] to this corpus. Finally in Section IV, we demonstrate the need for human-provided training data, analyze specific sources of performance degradation across language pairs, and offer design suggestions for perceptually grounded language learning.

To the best of our knowledge, this is the first work that seeks to apply a grounded language learning system designed for one language to other languages to evaluate its transferability. Our main contributions are: an evaluation of how effectively language-agnostic this learning system is, along with a detailed analysis of sources of error in the language transfer; a dataset containing images and trilingual descriptions of those objects; and a set of suggestions and considerations for future research in this area. We will make our dataset available upon publication.

II. Related Work

Grounding natural language to the physical world [11] is highly relevant to robotics research, and researchers are addressing the question from a variety of perspectives. There

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

[1]The first author is affiliated with RTI International, 3040 East Cornwallis Road, Durham, NC 27709, USA. ckery@rti.org

[1]Remaining authors are with the University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250, USA. pillai1|cmat|ferraro@umbc.edu

have been a number of successful approaches in different areas, such as navigation [33], [20], understanding commands and directions [2], [1] or action words [5], grounding spatial relations and concepts [24], and referring expressions for objects in images [8], [38]. Other work has explored interactively grounding additional non-visual properties like sound and weight [34].

Our experiments focus on how to learn labels for color, shape, and semantic object labels from unconstrained descriptions of real-world objects. Not constraining the language people may use sets the stage for learning models idiosyncratic to a particular user. Understanding such attributes is critical for tasks such as grasping [18] and collaboration [14]. In this work, rather than learning connections between novel language and an existing formulation, we follow the approach of [21], extending the formal representation as novel words are encountered.

Many grounded learning systems associate words with the perceived world [32], [19]. In this paper, we demonstrate the importance of selecting natural language processing techniques carefully when improving the efficiency of such grounded learning systems. Language pre-processing techniques have been studied when measuring document clustering and retrieval [15], [3], but not in visual-linguistic grounding. This paper further examines visual classification tasks with lemmatized and/or stemmed tokens of various categories and proposes approaches to improve learning.

Some systems that ground language attributes apply complex, structured linguistic analysis techniques to utterances to examine how particular phrases might relate to other parts of the sentences [24], [21]. Applying these structured linguistic approaches to an arbitrary new language is a complex undertaking, sometimes requiring costly expert annotation. In this paper, we consider each token in isolation and do not use any additional language features from the descriptions when learning the meaning of the tokens, allowing the model to scale more easily across languages.

Most research has focused on learning English-based robotic language acquisition systems, although there are some non-English corpora available (e.g., [7], [35], among others). In contrast, we concentrate on developing language grounding models from descriptions of objects gathered via a low-cost robotic sensor. In our experiments, we extended the monolingual public image dataset of Pillai and Matuszek [26] by collecting (non-paired) Hindi and Spanish language descriptions from native speakers.

Frank *et al.* [9] evaluate the preference of descriptions generated by native language speakers over the descriptions translated from a different language, whereas we evaluate how well a system trained on either translated data or Amazon Mechanical Turk language descriptions collected from native speakers could then complete the object recognition task using AMT descriptions.

Our work is most similar to that of [13] and [10], who improve the performance of learning visual-semantic embeddings by training over multiple languages. However, our work differs in two important ways. First, rather than using descriptions of perceptual data from several languages matched to one another, we explore how a core learning architecture designed for a particular language transfers to a new language; we do not assume that *parallel* (aligned) multilingual descriptions are available. Second, we collect unconstrained language about particular objects as seen by a real sensor, rather than relying on caption data. This is consistent with the goal of deploying robots for use by non-specialists in different environments.

III. Approach

There is a substantial body of research on learning mappings between a natural language and robot-usable representations. We follow [30] and [26] in using a *words-as-classifiers* model, in which a robot learns a mapping from each word in a language to classifiers trained on perceptual data over objects or actions. (For example, obj-referred-to-as-``red'' or obj-referred-to-as-``banana'' classifiers might be trained, using as positive examples every object someone referred to using that word.) These classifiers can then be run against objects in the environment to test whether they match a person's description at test time (for example, "hand me a banana.")

This process consists of three stages. First, perceptual data and descriptions of objects are collected in multiple languages, simulating inputs a robot might receive when being taught about a new object. Second, the language and percepts are pre-processed to extract features, and relevant language tokens (words) are extracted and paired with the features of the objects they describe, along with suitable negative examples [26]. Third, for each term, binary classifiers that describe whether an object is a correct grounding for a particular word are trained and evaluated against a held-out set of objects. Fig. 3 shows this architecture.

A. Data Corpus

The system is trained and tested using a novel corpus consisting of RGB-D images paired with multilingual natural language descriptions of the objects in the images. This corpus extends the public dataset of [26], tripling the number of language descriptions. The dataset itself has 18 object categories (such as *carrot* and *arch*), with four instances

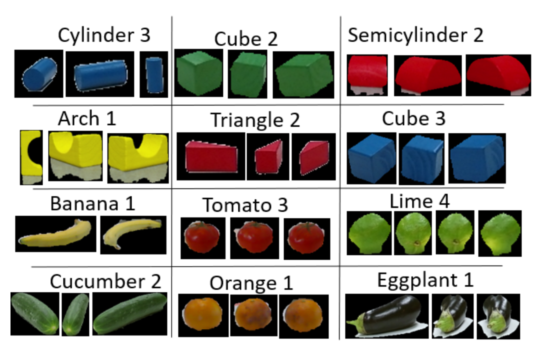

Fig. 2. Cropped Kinect2 images of twelve objects. All categories in the data set were toy blocks, fruit, or vegetables of various colors and shapes.

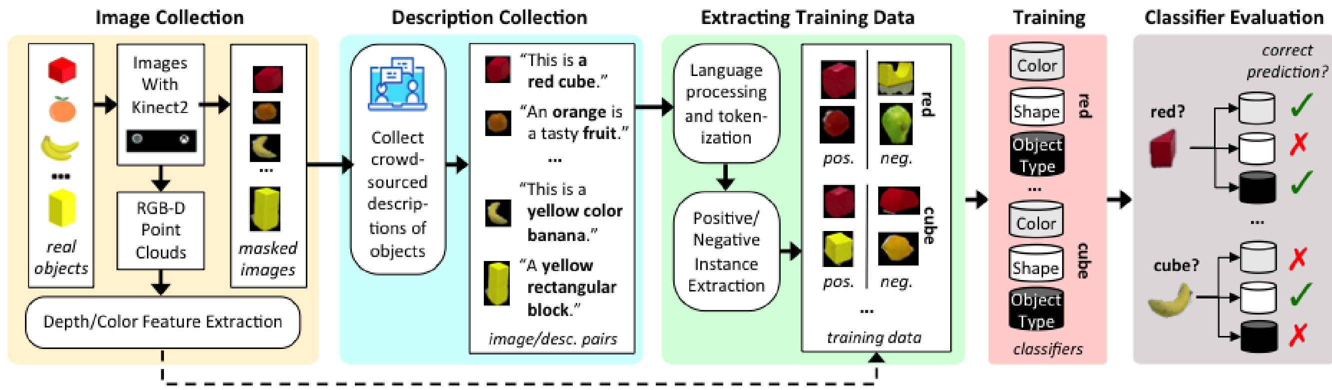

Fig. 3. This diagram shows the data flow of the grounded language acquisition system, from sensor and language data collection, through selection of training data and learning of classifiers, to eventual testing of those classifiers. Individual sections indicate the elements of the approach, as detailed in Section III. Note that while this diagram only shows classifiers and examples for "red" and "cube," classifiers are learned for all tokens.

(individual objects) in each category and an average of five images of each instance (see Fig. 2 for examples). Each image was taken from various angles with a Kinect2 depth sensor, yielding both RGB-D point clouds and regular color images. The goal in using these images is to collect images that accurately represent how a robot might perceive objects in its environment. The RGB images are masked so only the objects are visible to participants.

In existing and previous work, the English natural language descriptions of each instance were obtained using the Amazon Mechanical Turk crowd-sourcing platform. We chose to mimic this setup when collecting fluent speaker descriptions in Hindi and Spanish. These languages are chosen for their high number of speakers, as well as their varying dissimilarities to English. Our motivation for collecting new language data is to determine whether machine-translated training data is sufficient to handle language use by native speakers, or if it is necessary to involve speakers of that language to train an adequate model. As we discuss later in this paper, we find that the translated data is not sufficient for learning the language used by native speakers, resulting in an approximately 20% reduction in F_1 score across all languages (see Fig. 5).

Workers are presented with cropped and masked images of objects and asked to provide descriptions. We chose to provide the workers with no sample descriptions, in order to maximize the variation in the descriptions. The purpose of this is to evaluate our model with data that could accurately represent the ways in which speakers of each language might talk about objects when presenting them to a robot. Despite the lack of priming, many workers in all three languages chose to describe most objects with simple descriptions like "This is a red cube." However, we also see noticeable variation in all three languages, where workers chose to provide extra information about objects, such as describing cucumbers as healthy or being good in a salad.

The dataset contains 6,045 descriptions in English, 5,735 descriptions in Hindi, and 5,104 in Spanish. Originally, over 6,000 descriptions were collected for each of Hindi and Spanish, but in both cases we exclude descriptions from workers who explicitly did not follow the directions, such as by responding in the wrong language or consistently responding with text unrelated to the images they were being asked to describe; we did not exclude for other reasons. In the analysis section, we account for the smaller number of descriptions collected in Spanish and Hindi by randomly subsetting all datasets in such a way that each instance is trained on an equal amount of descriptions in each language. The results are averaged from several of such subsets.

B. Semantic Processing of Descriptions

Once collected, descriptions are put through a series of preprocessing steps to extract relevant tokens. We define a relevant token as a word the robot might want to learn to recognize. For example, in the description "This object is a large yellow banana!", a preprocessed version might be "object large yellow banana" (or "large yellow banana"). We conducted initial experiments to determine methods that extracted the most appropriate tokens in the three languages.[1]

1) Stemming and Lemmatizing Tokens: The overall design of this grounded language learning system hinges upon gaining an understanding about the meaning of a new object or attribute descriptor by examining the objects it was used to describe. Given limited training data, the system should be able to recognize when the same word is being used across different examples. In English, nouns have only a singular or plural form, and adjectives are rarely conjugated (with the exception of comparative terms like "large*r*"). In contrast, in both Spanish and Hindi nouns can be gendered and adjectives are often conjugated to match the gender of the nouns. For example, in Spanish, "The red table" is "La mesa roja," where "roj" is the stem of the adjective "red," and "roja" is the feminine singular form. A robot may therefore unnecessarily learn both a masculine and feminine word

[1]Following standard language approaches, we lower-case all words (which is only a valid step for Spanish and English) and remove punctuation, including language-specific punctuation like the Spanish upside-down question mark and the Hindi full-stop.

describing the same concept, subdividing the training data and weakening the resulting classification.

Past work applied a *lemmatizer* to the English data to remove conjugation from words. Lemmatizers take words that have been conjugated in some way and reduce them to their unconjugated forms—for example, reducing "running" to "run." When examining this step in the context of Hindi and Spanish, due to a lack of readily available non-English lemmatizers, we replace lemmatization with the simpler but related step of *stemming*. Stemming is very similar to lemmatization in that both tools take conjugated words and remove the conjugation. The difference is that stemmers simply chop off conjugations instead of attempting to find the original unstemmed form. This simplification makes them potentially less effective (e.g., a stemmer might reduce "running" to "runn" instead of "run"), but also much easier to implement. For this paper, we use the NLTK Snowball Stemmer [27] for Spanish and English text while For Hindi we used the simple stemmer described in [28].

2) Stop Word Identification and Removal: An additional challenge in using unconstrained natural language is identifying what words the system should attempt to learn physical meanings for and what words it should ignore. In natural language processing, it is common practice to remove "stop words" from language data as a preprocessing step. Stop words are defined as words that are necessary to form grammatical sentences, but do not contribute to the overall meaning of the sentence. We consider two kinds of stop words: general and domain-specific. There are publicly available lists of general stop words, such as "the" or "and." For English and Spanish, we use NLTK's stop word lists in each language. Since NLTK does not have a list for Hindi, we instead utilize the stop word list compiled in [31].

In our data, terms such as "object" or "item," are broadly used as generic terms to describe most elements in the dataset. We identify these domain-specific stop words by their Inverse Document Frequency (IDF), an effective method for information retrieval in which each word is scored according to the log of the number of instances in the dataset divided by the total number of instances where that word appears in some description. Words with low IDF scores appear in many different descriptions and are removed. We discuss this further in later sections.

C. Model Learning

This section discusses the grounded language learning system's components, including feature extraction and identification of "ground truth" positive and negative examples.

1) Tokenization and Positive/Negative Example Identification: After the descriptions are preprocessed, the next step is to extract relevant tokens and identify positive and negative examples of these tokens. As the only human guidance we have is the collection of unprompted and unguided descriptions (as opposed to annotations over pre-defined labels), the "ground truth" for what instances are and are not representative of various concepts have to be extracted from the descriptions.

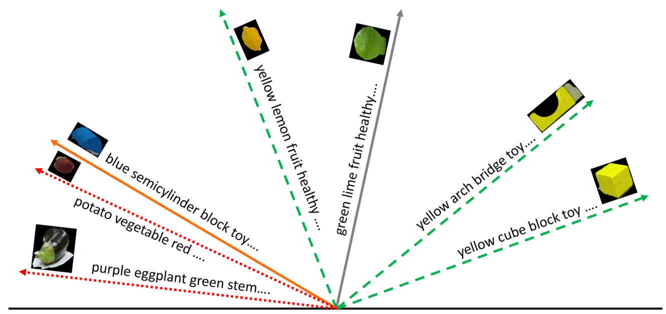

Fig. 4. A visualization of negative example selection for the token "yellow." Note that the negative examples selected (shown with red dotted arrows) are the ones that are the farthest away from the most positive instances (shown with green dashed arrows).

To identify relevant tokens, we concatenate all descriptions for each instance (approximately 80 descriptions per instance) and count how many times each unique word is used. If a word is used more than five times for a particular instance, the instance is deemed a "positive example" of that word. Words that do not have at least three positive examples are excluded from consideration, as it is deemed that the robot has not seen this concept enough times to learn it. The cutoff of five appearances is found empirically and is consistent with our own previous work [21], [26]. By extracting the tokens in this way, the system ignores any context in which the token was used outside of the image it was paired with.

Once positive examples are found for each token, our system also needs negative examples. Negative examples are rarely given in descriptions: users are much more likely to describe an object with positive properties, rather than enumerate negative properties. To address this, we use the approach of [26]. We take the combined descriptions of each object instance and represent these descriptions in vector space using the Distributed Memory Model of Paragraph Vectors (PV-DM) [22], [17]. We use the PV-DM model to measure dissimilarity between descriptions; in this model, semantically similar documents are similar in vector space. We use the cosine similarity statistic measure to find the dissimilar vectors and used those objects as negative examples in our language learning model. The intuition behind this approach is that instances that are negative examples of other instances are likely to have been described using very different language.

One consideration when choosing negative instances is that a token may have positive instances that are dissimilar to each other (see Fig. 4). We minimize this risk by choosing negative examples that are the farthest away in space from the most positive instances. This is decided using a weighted vote. For a term, all objects that had ever been described using that term are discarded. Remaining objects are sorted by the cosine similarity of their descriptions in vector space, and the last 2/3rds of the list are retained; we have established in previous work that below that threshold, any negative examples perform consistently [25]. Each positive instance then returns this set of candidates, weighted by similarity. The final scores for each negative instance candidate come from the sum of its weighted vote. The candidates are sorted by this score, and the top 25% of the candidates are selected as negative examples. This proportion was chosen empirically. Tokens with no identified negative examples were removed from consideration.

2) Extracting Image Features: For the classification task, two kinds of features are extracted from the RGB-D images. For generating color features, the RGB images of the objects are used. The RGB features of all pixels in the images are clustered using *k*-means based on RGB values. Then histograms are created based on the density of all clusters. The 'R, G, B' values of the centroid with the highest density are selected as the color features. We use HMP-extracted kernel descriptors [16], [4] to take the location and depth data for each image and extract shape features.

3) Training and Testing Classifiers: For each remaining token, three binary classifiers are trained using logistic regression on positive instances (objects described using that token) and negative instances selected as above. The underlying idea behind training all three types of classifiers per token is that a new word might describe a color, shape, or object and a robot with no previous knowledge would not know which category the token should belong to. Color classifiers are trained using RGB features, shape classifiers are trained using HMP features, and object classifiers are trained using a combination of the above. During evaluation, all three possible classifiers for a word are applied to held-out instances; these scored are reported separately in Fig. 7.

IV. Experimental Results

In this work, we expand the language learning system previously tested with English data to additional languages. To assess the quality of the language groundings, we apply the learned token classifiers to an object selection task. One instance of every object category is held out during training. For each token used in the descriptions of test instances, positive and negative examples are found from the testing instances, as discussed above. The token classifiers learned in the training phase are then scored by how accurately they could identify which instances were positive and which were negative. Reported scores are averaged over 180 runs.

In addition, as each token has a separate color, shape, and object classifier, three F1-scores are reported for each token. The final scores presented here are averages across all test tokens for each category. No manual sorting of words into categories is performed, which depresses overall scores, but is consistent with our unsupervised training scenario.

Next we discuss the performance of the system, and how it is impacted by both the origins of the training data and the language processing steps applied to it. In sum, our experimental results are as follows: (1), using machine-translated training data does not lead to good performance on human-provided language; (2), the more NLP preprocessing a system uses, the less well that system can handle a new language; and (3), the grounded language system under consideration handles both Hindi and Spanish well.

A. Performance Across Data Sets and Processing Techniques

1) Learning from Translated Data Yields Poor Results: For a robotic system to learn language data in a new language, it would be convenient if one could train such a system on translations of the language data already collected. We explored this possibility in Fig. 5, which shows the performance of the system when it is trained using either a translated version of the English dataset (using Google Translate's API [37]), or the crowd-sourced descriptions. In both cases, the trained models are tested on the tokens and positive/negative instances identified from the crowd-sourced descriptions, as these most accurately represent how a native speaker might describe the objects to a robot. The figure shows that the classifiers trained using the translated corpora are not sufficient when faced with the native language data. A major contribution to the lower performance is the large number of tokens that were used in the Mechanical Turk descriptions, and not in the translated corpora. Intuitively, a direct translation of a corpus in English is unable to accurately represent the variety of ways that a native speaker might describe objects.

2) Results Across Languages are Comparable: Fig. 6 shows the comparative performance of the learning system across all three languages, highlighting the score differences between English and non-English data. The model performs comparably across the three languages. Because the focus of this paper is *comparative* that is, we seek to know how effectively the system can be transferred between languages), we are less concerned about specific values and more concerned about the delta). Average scores are heavily influenced by the large number of previously unseen words occurring in the test data; this is especially the case for Spanish, with an average of fifty previously unseen tokens per testing run (where English and Hindi each averaged approximately thirty). This is primarily a product of our small dataset; larger initial training data collected from Mechanical Turk

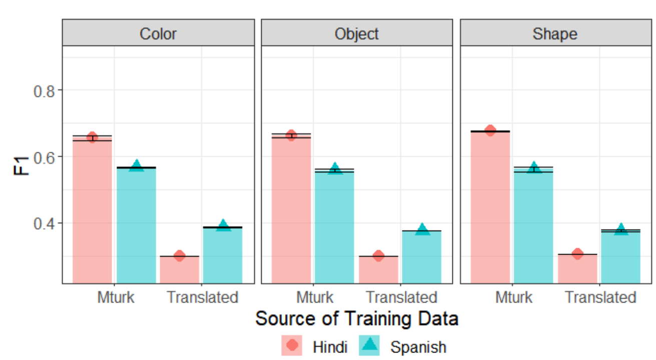

Fig. 5. System performance as compared between training the model on a translated version of the English dataset and training it on using descriptions collected in the target language using crowd-sourcing. Note that the translated dataset is not sufficient for learning many of the tokens used by native speakers when describing the images.

would improve performance by ensuring that more total descriptions occurred in the test data.

3) Language Pre-processing Affects Transferability: When expanding the grounded language system to Hindi and Spanish, the language processing techniques that ensure that relevant tokens are identified and correctly conflated must be considered. As seen in figure 7, stemming has a relatively low impact on the scores. Qualitative analysis of the results shows that stemming does enable the system to correctly conflate different gendered forms of adjectives in Hindi and Spanish. Varying the IDF score threshold for removal demonstrates that the optimal threshold varies by language. For Spanish, several important color words are used often enough that they appear in the bottom 2% of tokens by IDF score, while for Hindi, the bottom 3% of terms are safe to remove. These are comparatively small differences, but suggest that although it can be beneficial to remove unnecessary words, care must be taken with selection. Fig. 7 shows the average F1-scores of the model across the three languages when trained and tested on the Mechanical Turk descriptions.

B. Design Criteria for Language-Agnostic Learning

In the course of this work, we have presented a comparative analysis of the performance of a particular grounded language learning method when applied to two novel languages. Beyond this detailed analysis, an additional contribution of this work is suggested design criteria that groups designing grounded language learning systems may wish to consider.

First, consistent with expectations, we find that the more sophisticated a semantic processing step is, the less likely it is to work in a new language without significant modifications. For example, the simpler approach of stemming is more accessible than lemmatization for the novel languages considered. A robotics audience may then wish to explore learning methods that do not rely heavily on natural language preprocessing, or to focus on methods that are themselves relatively language-agnostic.

Second, training data beyond that provided by end-users should be provided by fluent speakers of the new language whenever possible. Fig. 7 shows that for our test case, using training data drawn from a simple machine translation approach is inadequate to support classifier performance when tested with fluent speakers, who tend to take wide advantage of the rich variety of terms and idioms available.

Finally, in general, the words-as-classifiers approach is well suited to transitioning across languages. The only modifications required were in identifying meaningful words in the data. This reduces possible complications that would be introduced by more sophisticated NLP techniques. For example, disregarding the relative placement of words in the descriptions means that the model is unaffected by the fact that Hindi has much looser rules for word ordering than English or Spanish.

V. Discussion and Future Work

Grounding natural language in perception is an essential task in human robot interaction. In this paper, we take an existing grounded language learning system and demonstrated that it can be easily extended to handle data in new languages. We find that when designing a system with unconstrained language data and noisy perceptual data, it is important to minimize and simplify the natural language processing preprocessing steps. We collect two new language corpora in Spanish and Hindi, and demonstrate that translated data is not sufficient for training. We will make our novel corpus available upon publication. In the future, we will implement the modified system or a similar learning system on a mobile robot in collaborative setting, where it must learn from and then interact with people using one or more novel languages. In addition, while the scores presented in this work are consistent with previous work shown in, [26] we expect in the future to move to a new category-free approach [29] which should improve performance.

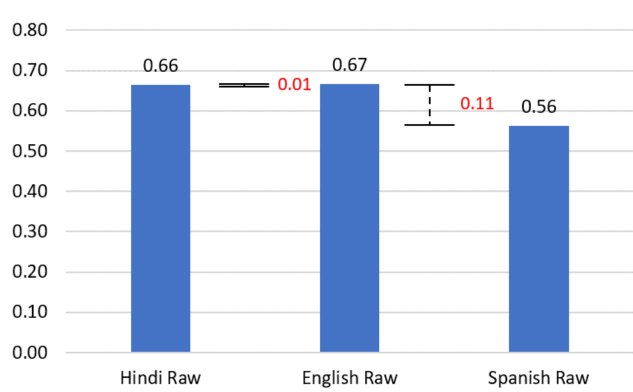

Fig. 6. This shows the differences in the scores of the grounded language system when applied to the three raw Mechanical Turk corpora. Each bar is the average of the raw scores for color, shape, and object for that language. Between English and Hindi, and English and Spanish are bars that highlight the differences in scores between the language pairs.

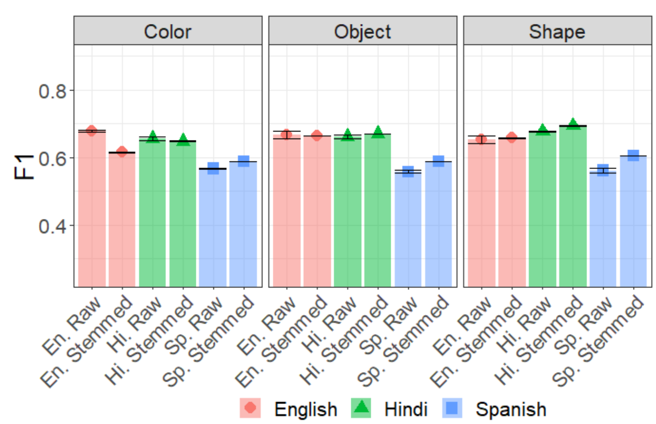

Fig. 7. Overall scores of all three languages. This compares the performance over the un-stemmed "raw" datasets, and the stemmed versions. Note that to allow for fair comparison, the Hindi, English, and Spanish datasets have been subset to have equal amounts of descriptions per instance. The error bars provide the variance in the scores.

REFERENCES

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2018.

[2] Y. Artzi and L. Zettlemoyer, "Weakly supervised learning of semantic parsers for mapping instructions to actions." *Transactions of the Association for Computational Linguistics (TACL)*, vol. 1, pp. 49–62, 2013.

[3] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," *Lecture Notes on Software Engineering*, vol. 2, no. 3, 2014.

[4] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Computer Vision and Pattern Recognition*, 2011.

[5] J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu, "Language to action: Towards interactive task learning with physical agents." in *IJCAI*, 2018, pp. 2–9.

[6] D. L. Chen, J. Kim, and R. J. Mooney, "Training a multilingual sportscaster: Using perceptual context to learn language," *Journal of Artificial Intelligence Research*, vol. 37, pp. 397–435, 01 2010.

[7] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74.

[8] N. FitzGerald, Y. Artzi, and L. S. Zettlemoyer, "Learning distributions over logical forms for referring expression generation." in *EMNLP*, 2013, pp. 1914–1925.

[9] S. Frank, D. Elliott, and L. Specia, "Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices," *Natural Language Engineering*, vol. 24, no. 3, pp. 393–413, 2018.

[10] S. Gella, R. Sennrich, F. Keller, and M. Lapata, "Image pivoting for learning multilingual multimodal representations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2839–2845. [Online]. Available: https://www.aclweb.org/anthology/D17-1303

[11] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[12] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, M. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual storytelling," in *NAACL*, 2016, equal contribution: TH, FF.

[13] Á. Kádár, D. Elliott, M.-A. Côté, G. Chrupała, and A. Alishahi, "Lessons learned in multilingual grounded language learning," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 402–412. [Online]. Available: https://www.aclweb.org/anthology/K18-1039

[14] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Recovering from failure by asking for help," *Autonomous Robots*, vol. 39, no. 3, pp. 347–362, Oct 2015.

[15] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04. New York, NY, USA: ACM, 2004, pp. 625–633. [Online]. Available: http://doi.acm.org/10.1145/1031171.1031285

[16] 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*, 2013, pp. 167–192.

[17] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[18] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[19] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[20] C. Matuszek, N. FitzGerald, E. Herbst, D. Fox, and L. Zettlemoyer, "Interactive learning and its role in pervasive robotics," in *ICRA Workshop on The Future of HRI*, St. Paul, MN, 2012.

[21] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[23] R. Mooney, "Learning to connect language and perception," in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, Chicago, IL, 2008, pp. 1598–1601.

[24] R. Paul, J. Arkin, N. Roy, and T. M Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators," in *Proceedings of Robotics: Science and Systems (R:SS) 2016*. Robotics: Science and Systems (RSS), 2016.

[25] N. Pillai, F. Ferraro, and C. Matuszek, "Optimal semantic distance for negative example selection in grounded language acquisition," *Robotics: Science and Systems Workshop on Models and Representations for Natural Human-Robot Communication*, 2018.

[26] N. Pillai and C. Matuszek, "Unsupervised selection of negative examples for grounded language learning," in *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA, 2018.

[27] M. F. Porter, "Snowball: A language for stemming algorithms," *Retrieved March*, vol. 1, 01 2001.

[28] A. Ramanathan and D. D. Rao, "A lightweight stemmer for hindi," in *the Proceedings of EACL*, 2003.

[29] L. Richards and C. Matuszek, "Learning to understand non-categorical physical language for human robot interactions," in *From the RSS Workshop on AI and its Alternatives in Assistive and Collaborative Robotics*, 06 2019.

[30] D. Schlangen, S. Zarrieß, and C. Kennington, "Resolving references to objects in photographs using the words-as-classifiers model," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[31] S. Siddiqi and A. Sharan, "Construction of a generic stopwords list for hindi language without corpus statistics," *International Journal of Advanced Computer Research*, vol. 8, no. 34, pp. 35–40, 2018.

[32] S. Tellex, P. Thaker, J. Joseph, and N. Roy, "Learning perceptually grounded word meanings from unaligned parallel data," *Machine Learning*, vol. 94, no. 2, pp. 151–167, 2014.

[33] J. Thomason, S. Zhang, R. J. Mooney, and P. Stone, "Learning to interpret natural language commands through human-robot dialog," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[34] J. D. Thomason *et al.*, "Continually improving grounded natural language understanding through human-robot dialog," Ph.D. dissertation, University of Texas at Austin, 2018.

[35] E. van Miltenburg, Á. Kádár, R. Koolen, and E. Krahmer, "DIDEC: The Dutch image description and eye-tracking corpus," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3658–3669. [Online]. Available: https://www.aclweb.org/anthology/C18-1310

[36] G. Wilcock and K. Jokinen, "Multilingual wikitalk: Wikipedia-based talking robots that switch languages." in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 162–164.

[37] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," in *CoRR*, 2016.

[38] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290.