GAILA Milestone 4 Report

Brown University
March 24, 2020

Milestone Description

A live demonstration of software capable of using English text and/or speech to describe the entities, relations, and events in the Phase 1 material, working with a blank starting state. A report describing the prototype, learning and inference algorithms, and an assessment of acquisition capabilities. A report or paper describing the model that the prototype implements. Evidence of the applicability of the software output to other human language technology tasks.

Summary of Progress and Results

- Prototype 1: Learning from Demonstration with 3D data (§1)
  - Implemented a prototype which represents actions as sequences of automatically detected discrete "skills", using prior work on learning from demonstration (LfD) in robotics
  - LfD prototype produces decent results for identifying nouns but is still at random for identifying verbs (macro-average P@1 for assigning words to actions of an unseen user is 30% for nouns, 11% for verbs). Note noun model still has access to ground truth object classes (details in §**??**).
- Prototype 2: Learning from 2D video data (§2)
  - Implemented transformer-based "query expansion" model which provided strong results in TREC 2019 evaluation, and an efficient listless ranking model (under submission at SIGIR).
  - These models are being adapted to the multimodal setting and will serve as the basis for our video retrieval/captioning prototype.
- Human Subjects Studies of Function Word Learning (§3)
  - Finished data collection on 200+ children for understanding of quantifiers and negative polarity items and are working on data analysis.
  - Nearly finished preparing of SayCAM corpus (videos to be used in adult function word studies), including automatically transcribing the corpus, processing/filtering low-quality video and audio data, and preregistration of methods for MTurk study.
- Live Demo delayed by COVID
  - Demo will involve captioning of actions in real time for novel user in VR; dummy implementation is working but needs to be integrated with trained models
  - Video demonstration will be provided once we are cleared to work in person on campus again.

Attachments

- Videos of detected "skills": `http://dylanebert.com/nbc_actions_analysis` (password: lunar)

1 Learning from Demonstration Prototype

1.1 Model

The high-level goal of our model is to use the standard distributional-semantics approach of representing words by their contexts. However, rather than text context, we want to model a word using the state of the environment in which that word is used. The majority of our technical effort is spent on how to represent the environment such that simple techniques from distributional semantics will work well for word representation learning.

1.1.1 Environment Representation

As a reminder, our raw environment data consists of 3D spatial data (xyz position and rotation for person's hands plus all objects in the environment) recorded at a rate of 90 frames per second. Our previous attempts to apply deep learning models to this raw data proved unsuccessful, likely due to the highly noisy input signal in conjunction with the small amount of training data we have collected (18 participants, $\sim$20K words). Thus, we have decided to explore more structured models for preprocessing the spatial data in order to discover concepts to which words can potentially refer, prior to observing any language. Specifically, we are using a Beta-Process Auto-Regressive Hidden Markov Model (BP-AR-HMM) in order to segment continuous spatial data into discrete actions or "skills". This method is introduced by [**?**] and has been applied for learning from demonstration in robotics [**?**]. The input to the model is a sequence of absolute xyz coordinates of the object(s) in the environment, and the output is a sequence of discrete latent symbols that can be used to describe the time series.

We are using existing implementation of BP-AR-HMM[1] and are leaving all of the parameters at their default values. In our initial explorations, we have focused only on the position of the right hand. This setup results in skills which reflect the trajectory of the hand but are not sensitive to the objects that the hand interacts with (i.e. moving a hand upward may be the same "skill" regardless of whether or not the hand is holding something). Thus, we follow the procedure used by [**?**] and optionally post-process the output of the model by clustering skills instances of skills based on the objects that are closes to the hand at the skill's endpoint. Thus, depending on what knowledge we want to assume prior to word learning, we can break a single skill output from the BP-AR-HMM into multiple skills. E.g. if we assume just that the agent recognizes objects we might decompose `skill_19`) into `skill_19_ending_at_object` and `skill_19_not_ending_at_object`; if we assume the agent differentiates object categories we might get `skill_19_ending_at_apple`, `skill_19_ending_at_table`, etc.

1.1.2 Word Representation

By running the BP-AR-HMM and optionally subdividing skills based on object endpoints as just described, we can now represent our environment using sequences of discrete symbols. We therefore can approach the word learning problem using traditional NLP techniques. To do this, we look at each "frame" (i.e. each $\frac{1}{90}th$ of a second) and record the environment state and language observed at that instant. We then collapse consecutive frames if they contain identical observations. The result is aligned "parallel texts" of the form show in Table 1 for each of the 109 transcripts (18 participants $\times$ 6 tasks each) in our collected data.

[1] `https://github.com/michaelchughes/NPBayesHMM/blob/master/doc/QuickStartGuide.md`

Frame	Word	Skill	Object
11922	okay	3	None
11952	okay	1	None
11997	okay	3	None
12006	okay	5	None
12060	okay	3	None
12147	–	3	None
12192	start	3	None
12204	start	6	None
12228	by	6	None
12237	clearing	6	None
12249	clearing	2	Lamp
12300	off	2	Lamp
12335	the	2	Lamp
12354	table	2	Lamp

Table 1: Example of aligned language and environment data. Shown is participant 1_1a; task 1 (setting the table for lunch). Note we experiment with if/how to use the information about the object at which the skill ends; see text for details.

1.2 Evaluation

1.2.1 Clustering

1.2.2 Human Eval

2 Video Retrieval

To aid in bottom-up language and scene understanding, we have been working towards scene memorization (retrieval) and generalization (captioning) functionality. Concretely, there have been two major lines of work to support better cross-modal retrieval performance:

1) To avoid pitfalls in negative sample selection/generation processes, we are training our rankers on raw list-wise preference data that allows for significant improvements in ranking effectiveness and drastically reduced model training time for arbitrary existing retrieval models. See Figure 1 for evaluation results on a standard retrieval collection. This work is currently under review at ACM SIGIR.

2) To improve generality and coverage of given scenes and narrations, we have been working on transformer-based query expansion schemes in the latent representation space. Expansion candidates are obtained via a separate transformer model trained on scene-internal frame re-ordering and are then re-ranked based on semantic similarity to the original example. After promising initial results at the TREC 2019 benchmark, we are now working on translating this performance to more resource-constrained settings.

Moving forward, we are combining the above techniques to design a joint textual-visual embedding space via which we will facilitate retrieval and captioning of scenes and that will eventually feed back into our language understanding efforts.

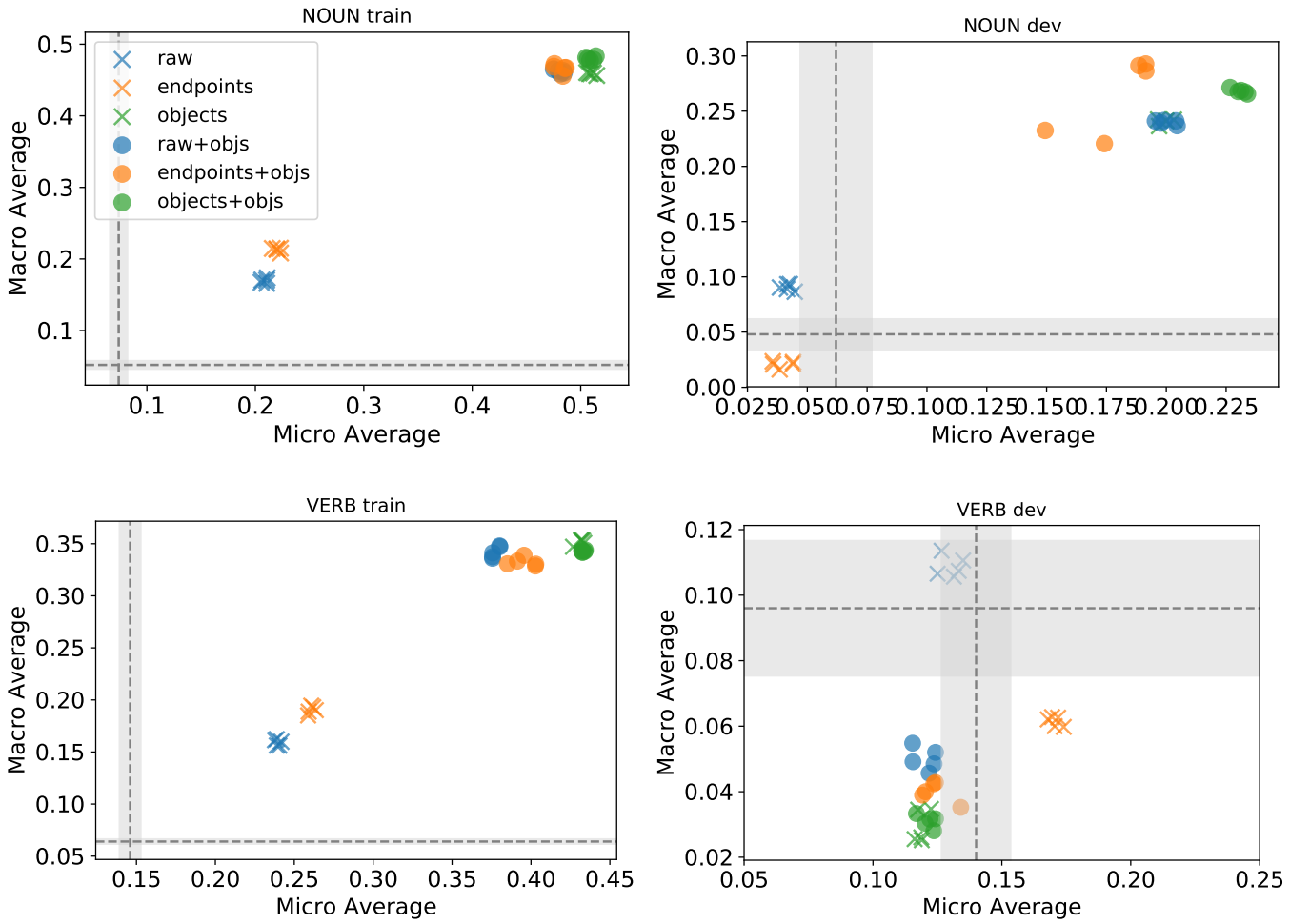


Figure 1: Noun representations which include object information perform (unsurprisingly) well. No verb representations perform above chance on dev.

3 Human Simulation for Function Word Acquisition

On studies with children: Just before the COVID-19 pandemic hit, we completed data collection on two projects looking at children's learning of logical words. One project has tested over 200 children between the ages of 3 and 7 years old on their understanding of novel quantifier words (how do children learn words like "all" and "some", and how might they extend this procedure to new words in the same syntactic position and sentence frames, like "give me dax of the toys". The other has tested 60 children between ages 3-5 on their understanding of Negative Polarity Items (words like "any" and "much", which are only grammatical in the context of negative sentences). These projects involve two undergraduate thesis students, who are currently processing the data, analyzing it, and writing up the results for their undergraduate theses.

On studies with adults: We have identified a video corpus of parents' speech to a child (SayCAM corpus, entries for Asa) that we can use to develop materials for online studies. This corpus has longitudinal data: approximately 1-2 hours of video recorded from the child's perspective every week between the child's ages of 6 to 23 months (although months 18-23 have not yet been made available). It also has the filmed parents' permissions for others to view video clips from the corpus, including permission to show videos to participants on Amazon Mechanical Turk or other online platforms, which will form the participant pool for the studies we do with these materials. We have machine-transcribed the entire

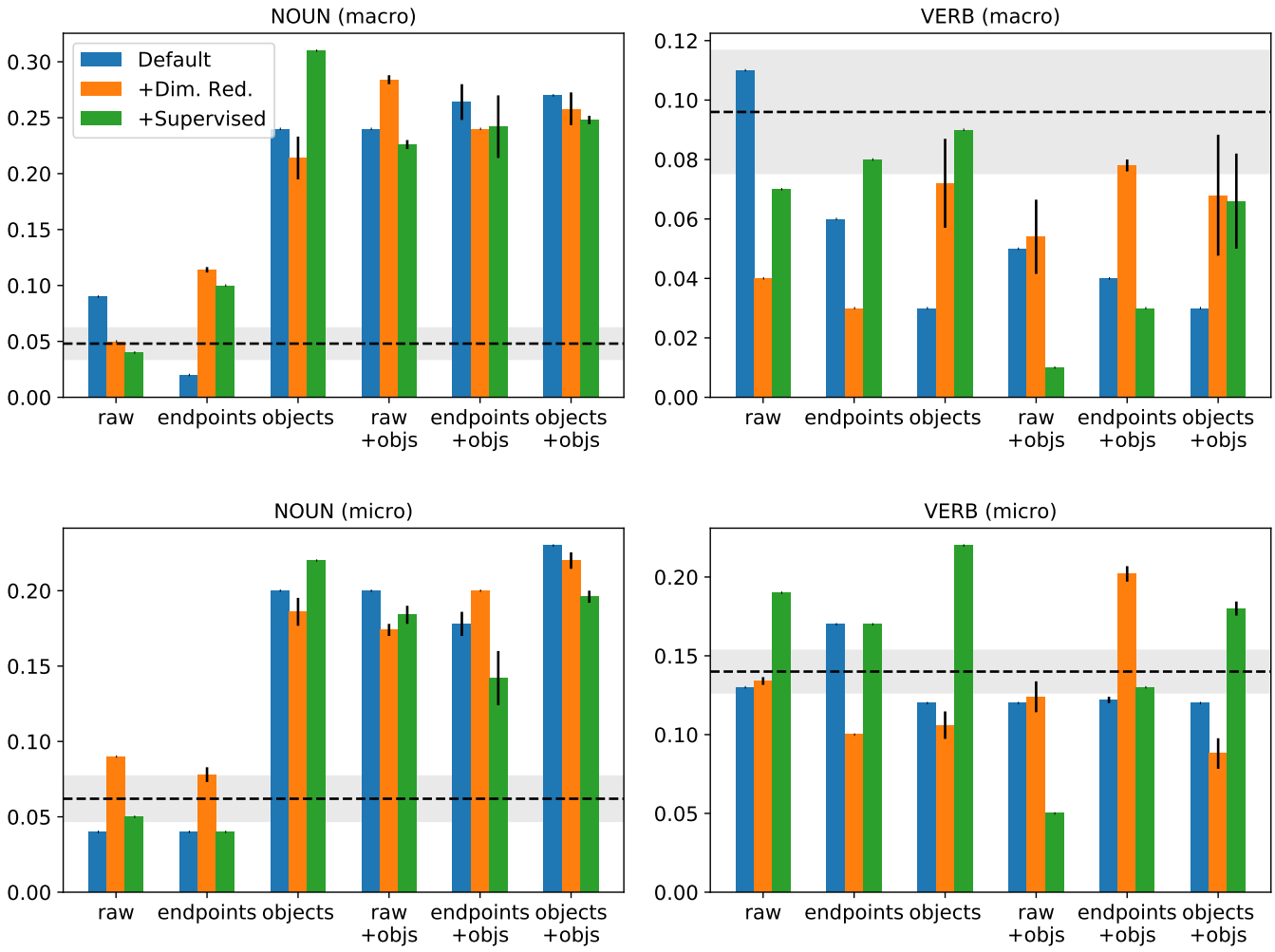
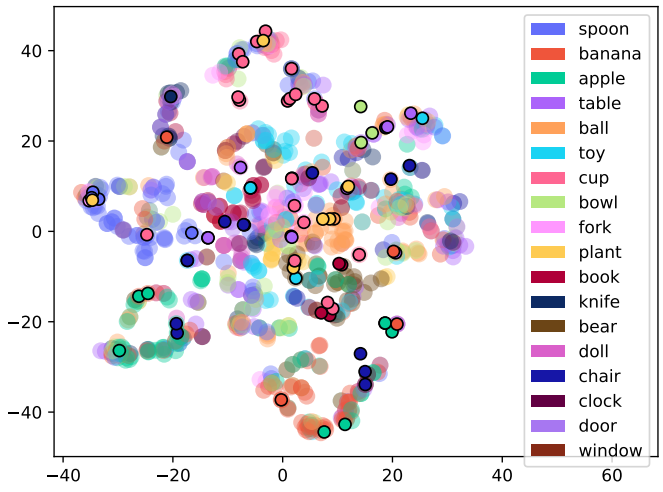
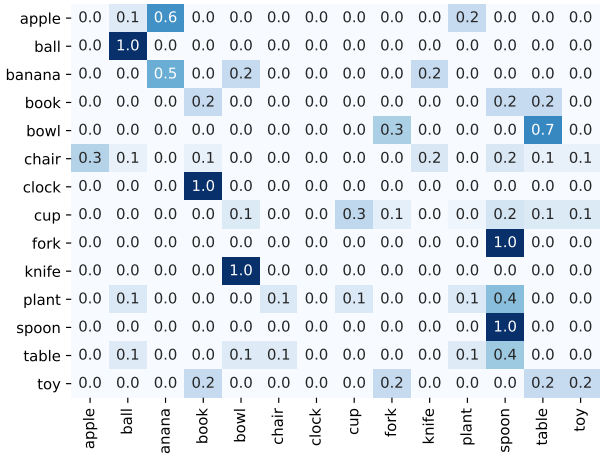


Figure 2: TODO

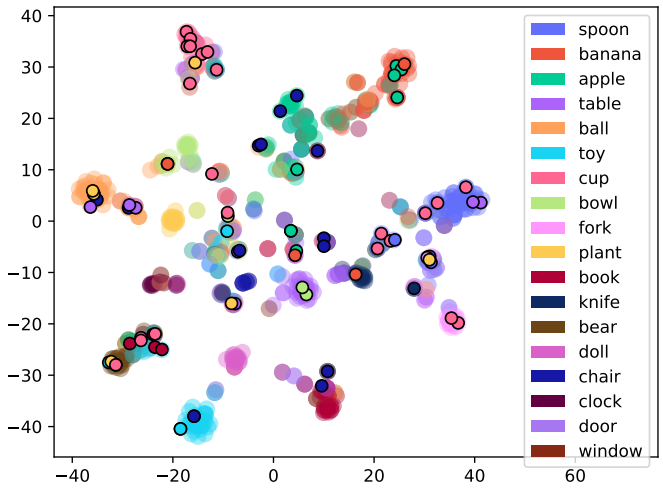
available corpus to identify instances of specific words being used, paired with timestamp information that helps find the corresponding points in the videos. After that, each instance of the words ”no” and ”not” (over 3,500) has been manually coded by two research assistants to identify the function of negation that the speaker meant to express (e.g. rejection of an offer being made, denial of the truth of another statement, etc.) Currently, we are in the process of identifying videos in which the negation words are spoken by the parents (rather than the kids), which are spoken to the children (rather than to other adults or in another room), and which both coders have agreed on the function of negation being used. We have also completed an extensive preregistration of this project’s materials and hypotheses, which lays out this complex data processing pipeline in detail for compliance within our own research team, and transparency and replication by others in the future. The next steps are to trim and process these video clips to create multiple versions for the different experimental conditions we plan to compare to each other. These activities have recently been delayed by research assistants’ limited availability as they have had to move off-campus due to the COVID-19 pandemic. However, online data collection should not be impeded by the pandemic once stimulus creation is complete.



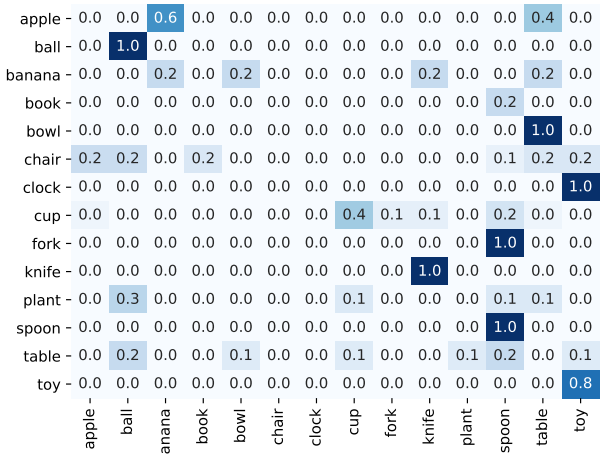
(a) Clusters, Unsupervised



(b) Confusion Matrix, Unsupervised

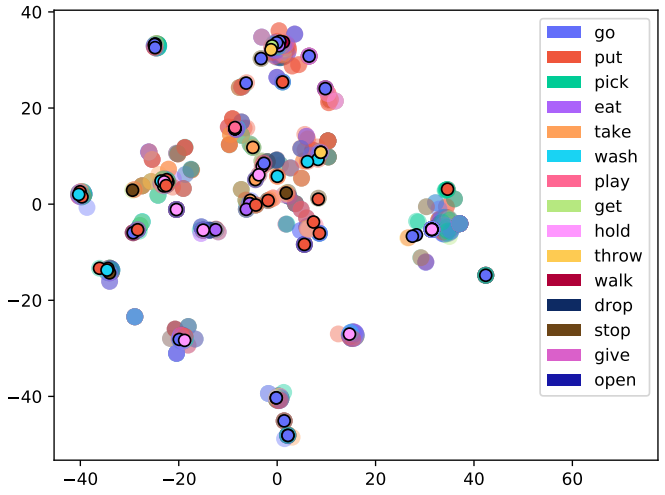


(c) Clusters, Supervised

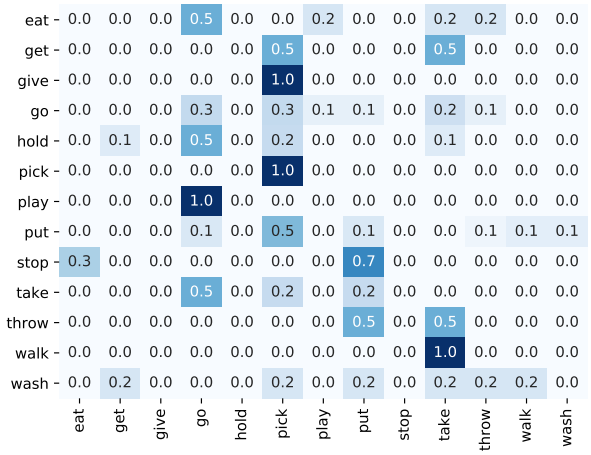


(d) Confusion Matrix, Supervised

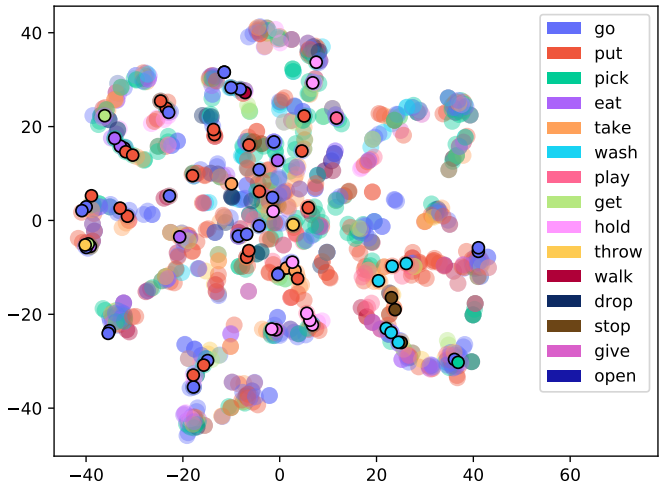
Figure 3: Side by side of noun clustering, with vs. without supervision. These are results for the object-based state symbols, but no additional objects added as features. I.e. the blue and green bars under the group labeled "objects" in Figure **??**.



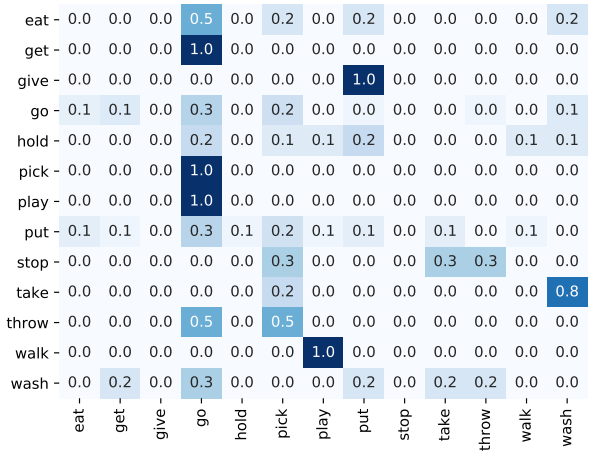
(a) Clusters, Raw



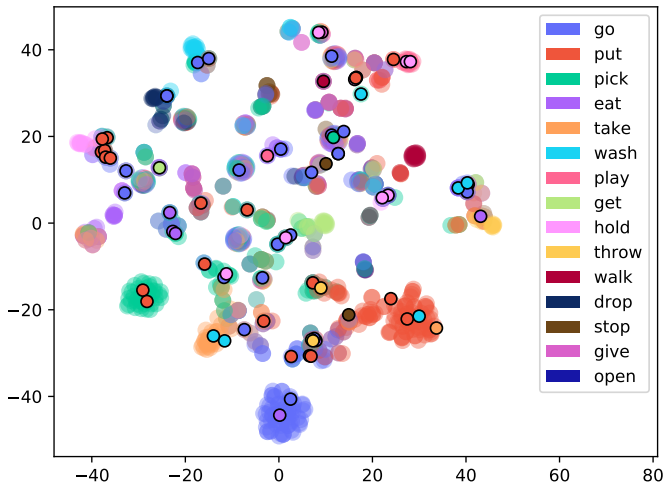
(b) Confusion Matrix, Raw



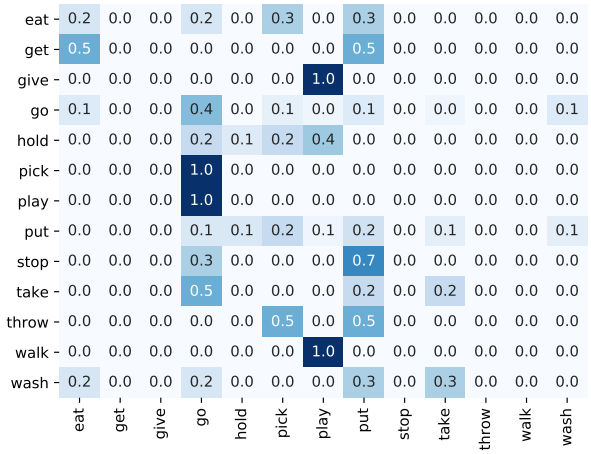
(c) Clusters, Objects



(d) Confusion Matrix, Objects

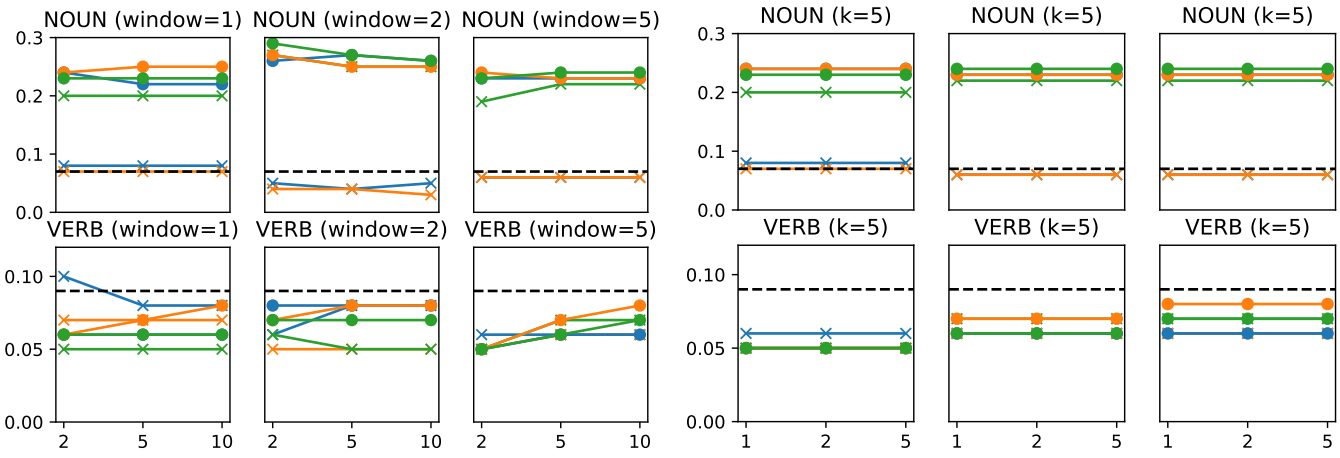


(e) Clusters, Objects Supervised



(f) Confusion Matrix, Objects Supervised

Figure 4: Side by side of verb clustering. Results are raw, objects, and objects-supervised.



(a) Effect of K (b) Effect of Window

Figure 5: No meaningful effect of window size or ngram length

Model	MRR	nDCG	MAP	Time
DRMM_pairwise	0.14723	0.27916	0.14414	7.96h
DRMM_PoolRank	**0.16146**	**0.28851**	**0.15832**	**3.46h**
KNRM_pairwise	0.20885	0.34811	0.20475	6.41h
KNRM_PoolRank	**0.22604**	**0.36233**	**0.22202**	**4.81h**
Conv-KNRM_pairwise	0.23255	0.36923	0.22828	20.51h
Conv-KNRM_PoolRank	**0.26233**	**0.39659**	**0.25767**	**10.10h**

Figure 6: TODO