Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Learning Names For RFID-Tagged Objects in Activity Videos

Ian Perera and **James F. Allen**
University of Rochester
Computer Science Department
734 Computer Studies Bldg.
P.O. Box 270226 Rochester, NY 14627
{iperera,james}@cs.rochester.edu

Abstract

We describe a method for determining the names of RFID-tagged objects in activity videos using descriptions which have been parsed to provide anaphoric reference resolution and ontological categorization.

Introduction

Exophoric reference resolution, the mapping of linguistic utterances to corresponding event or object instances in an environment, typically makes use of objects that already have some semantic label attached to them. In fields such as object detection, virtual environment interaction, and activity recognition, only certain types of objects are expected to be observed, and this technique is acceptable. However, the domains of these research efforts could be expanded if new objects could be identified by their mention in descriptive text, without any prior knowledge or mapping of the object instance to a concept.

We propose a method of learning names for objects in activity videos with the activities described in natural language by the person performing the task. This description provides utterances which are sense-disambiguated according to concepts in the TRIPS ontology (Allen, Swift, and de Beaumont 2008), and are then mapped to RFID tags representing instances in the environment.

Our techniques could be applied not only to RFID data, but to any data where object instances are given a set of consistent identifications and intervals over which they are salient and likely to be referred to by the speaker. Specifically, we hope to use these techniques with the object detection results from the Kinect data we collect.

Related Work

Previous work on exophoric references and learning object names is rather limited due to the dependence on a controlled physical or virtual environment to support the references and the limited number of multi-modal datasets with labeled referring expressions to train on. (Schlangen, Baumann, and Atterer 2009) provide a metric for evaluating the quality of reference resolution, as well as develop a system for resolving references using textual features to match descriptions to puzzle pieces in a cooperative game environment. Their work shows the success one can achieve in resolving references given enough textual clues, even without the timestamp matching we perform. Work has also progressed towards using visual features to detect object instances rather than RFID tags. (Kruijff, Kelleher, and Hawes 2006) use groups of related SIFT features to represent instances of an object that can be referred to by text, in a similar way to how we group RFID tags on the same object.

Environment and Data

We collect video, audio, Kinect, RFID, and other sensor data simultaneously in a kitchen environment. A person demonstrates how to perform a task, such as making tea, by describing the actions he or she carries out in front of the camera and Kinect. RFID tags are placed on all relevant objects that can accept them, and the subject wears an iBracelet on each wrist to detect the objects currently being interacted with. In our current data, there is only one type of each object in the environment. In some cases, multiple tags are placed on a single object to increase detection rate, although we do not know in advance how many sensors are attached to each object. Other sensors are attached to kitchen appliances and cabinets to detect the subject's interaction with the environment.

The audio is transcribed and parsed using the TRIPS parser. Each utterance is labeled with the timestamp interval over which it was spoken. Likewise, sensor data is collected as polled data which is then converted to timestamps using the polling interval of the bracelets.

The RFID system suffers from a poor detection rate and from a limitation of detecting only one tag at any given time. Although having multiple tags on each object increases the detection rate, it introduces an additional complexity in resolving references - we now must consider the possibility that multiple IDs map to the same object.

Utterance Intervals

Each object mention is processed by using the ontological concept corresponding to the noun instance while considering temporal data only at the sentence level. Each concept is treated as occurring during the entire interval, for purposes

of Concept-ID cooccurrence. We make use of anaphoric coreference information from gold-standard parses of the transcripts to replace pronouns with their referents. We also use the parse data to only resolve concepts with definite articles and remove mass nouns, like "water", which could not have an ID tag.

Reference Resolution

Since we are looking for the names of object instances, and we know that each detected tag is attached to an object, we are trying to find the most probable mapping from ID's to concepts. By generating mappings in this manner, we ensure coverage over all sensors in the environment and are more likely to resolve references that have a corresponding object.

For each RFID, we find the most likely assignment of a concept to that ID by finding the highest score for each concept that cooccurs at least once with that ID. Concept-ID cooccurrance is defined as some overlap between the interval over which the ID is sensed and the interval of an utterance containing the concept.

Multiple ID's may share a best concept match. While we want to allow this to an extent, since multiple ID's may be attached to the same object, it is unlikely that a large number of tags will be attached to one object. Therefore, we condition the probability of assigning a concept to a tag on the number according to a prior distribution of how many different tags that concept is likely to be assigned to. We want to find the most probable RFID to concept mapping by solving the following expression:

$$\arg\max_{R \to C} \sum_{R} (P(c_i|r_i) \times P_{assign}(c_i))$$

where R is the set of ID's, C is the set of concepts, and P_{assign} is the probability of assigning that concept given its other assignments. However, since the probability of assigning a concept depends on all other concept assignments, finding the best solution could quickly become intractable with a greater number of concepts or ID's. To avoid this, we take a greedy approach of starting with the ID's that have the shortest detection interval over all of the data, mapping the most probable concept to that ID, and updating the assignment probabilities as we continue to the next ID.

For P_{assign}, we use a binomial distribution with a mean at the expected number of ID's per object. When evaluating the probability of assigning a concept that has been assigned n times, P_{assign} returns the probability of $n+1$ according to the distribution. Such a distribution provides an intuitive parameter of the data set that is likely to be known in advance, and can be generalized to any situation where multiple ID's might refer to the same object.

Using Bayes' Rule, $P(c_i|r_i)$ expands to $P(c_i) \times P(r_i|c_i)$, with the normalization $P(r_i)$ omitted. $P(c_i)$ is simply the probability of the concept among the other observed concepts in the transcripts, while $P(r_i|c_i)$ is the duration of r_i that overlapped with all instances of c_i divided by the total duration of the utterances containing c_i. Note that this is an unnormalized score, and not strictly a probability.

Results

We ran this algorithm on eight videos demonstrating the same procedure for making tea, each lasting about two to four minutes. There were eight RFID tags detected in total and 177 ontological concepts mentioned. Correct matchings are determined by agreement with an annotator with access to the video and sensor data.

Without the determiner, sense disambiguation, and coreference data provided by the gold-standard TRIPS annotations, none of the objects were correctly labeled. "Tea" was assigned to the teacup's tags, because the subjects often use tea in many different senses - the end product of the activity, the liquid in the cup, and the tea bags.

Only resolving references for the 25 ontological concepts preceded by "the", but without using coreference data, yielded two exact matchings and two matchings of "bag" to the teabox. We consider these to be close matches, given that the teabags themselves cannot take RFID tags and that subjects usually refer to the teabags themselves in describing the task.

Finally, using the knowledge from the ontology with the coreference information, the algorithm exactly matched four of the eight objects. Two of the four missed assignments again labeled the RFID tags of the teabox as teabags.

	Word correlation	Ontology, no coref	Coref
Exact	0	2	4
Close	1	2	2
Missed	7	4	2

Table 1: Concept-ID matches on the tea-making data

Future Work

We plan to use these same techniques with ID's assigned by both trained and untrained object detection algorithms running on the Kinect data. Initial results with trained algorithms are promising and show both that this method generalizes to visual sensor data and that errors are primarily caused by noise in the RFID data.

Acknowledgments

This work was supported in part by NSF grants IIS-1012205 and IIS-1012017, "Activity Recognition and Learning for a Cognitive Assistant".

References

Allen, J.; Swift, M.; and de Beaumont, W. 2008. Deep Semantic Analysis of Text. In *Symposium on Semantics in Systems for Text Processing*, volume 2008, 343–354. Morristown, NJ, USA: Association for Computational Linguistics.

Kruijff, G.; Kelleher, J.; and Hawes, N. 2006. Information Fusion for Visual Reference Resolution in Dynamic Situated Dialogue. *Perception and Interactive Technologies* 117–128.

Schlangen, D.; Baumann, T.; and Atterer, M. 2009. Incremental Reference Resolution. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 30–37. Association for Computational Linguistics.