First year project proposal

Erin Bennett

Introduction

Psychologists study what people believe and how this influences their perception, actions, goals, etc. To determine what theory a person or group of people hold, we often have to design subtle experiments, careful surveys, or coding schemes. This is effective at isolating people's theories, but it requires a lot of effort on the part of experimentors and participants. If we were able to automatically elicit a structured, quantitative representation of people's beliefs, this could help make psychological research more efficient.

Learning commonsense knowledge from humans is a problem that artificial intelligence researchers have been working on for a long time.

Different forms of representation have been proposed for representing commonsense knowledge in computer systems. Some candidates include causal bayes nets (Pearl 2003; and see Sloman and Lagnado 2015), neural networks, fist-order logic, probabilistic programs, schemas, and scripts (Schank and Abelson 1977). Scripts are sequences of events that are connected to eachother via the entities that are shared across them. They represent stereotypical knowledge of the kinds of things that happen in a given domain.

Recently, natural language processing techniques have been developed that can actually learn scripts for some domains from text corpora. We will build off of this work in the proposed project. First we will test the psychological validity of learned scripts using a variety of dependent measures on Amazon's Mechanical Turk. Based on these results, we will turn to strategies for improving the representation and learning.

Methods

Script induction

Rudinger et al. (2015) used a representation of scripts from Chambers and Jurafsky (2008) and Jans et al. (2012). In this representation, events are related to one another whenever their subjects corefer to one another, e.g. an instance of "eat" is related to "pay" if the same person does both actions. Rudinger et al. (2015) ran several script induction techniques on a relatively small dataset of stories about restaurants (pulled from the blog "Dinners from Hell") and have made their code publicly available online. They express concern that a richer representation that includes subject *and* object information, such as that of Pichotta and Mooney (2014) might face trouble with data sparsity given the size of the "Dinners from Hell" dataset. Chambers (2013) provide a generative model approach to script induction, but their model includes much less information about sequences of events, which will prove important for eventually integrating causal relationships into the learned theories. In this project, we will minimally reproduce Rudinger et al. (2015)'s results and attempt to extend them to Pichotta and Mooney (2014)'s representation. We will also attempt to implement Chambers (2013)'s model, which though limited in sequential information has been shown to require much less training data than a pipeline approach.

The results of these replications will be groups of events that share subjects (and possibly objects) and some information about the sequence of those events.

Experiments

To further test the psychological validity of these learned scripts, we will run survey experiments on Amazon's Mechanical Turk. We will confirm the coherence of the learned groups of event groups, coreferences among their arguments, and their sequential order, and we will also try to elicit what other commonsense information might be relevant to these situations. Because we hypothesize that the model will perform worse for events that frequently occur but are not frequently talked about, we will also try to elicit from participants their estimates of the probabilities of the event occuring and the event being talked about.

Validation of learned scripts

For the following tasks, it is relatively straigtforward to determine what the script would predict as a response, and so we use these questions to assess the accuracy of the information learned.

- **Coherence & Sequence:** Following Li et al. (2013), we will produce simple "stories" from the learned script. Participants will be able to edit the story to make them more coherent. Using a simple drag-and-drop environment, they will be able to delete events, reorder them, and add new events. The more edits participants make, the less "coherent" the events in the story are. The sequence that participants chose for the events in the script will be taken as the correct ordering. Any added events indicate lower coherence (because they are edits), and they can additionally supplement the responses to the narrative cloze task, described in more detail below. As an attention check, we will insert one obviously out-of-place event that must be deleted.
- **Sequence:** Previous work has evaluated the sequence learned in a script induction by comparing the learned sequence to a randomly ordered sequence. We will instead use a drag-and-drop ranking measure such that people can arrange the events in order spatially on a computer screen. We will compare participants' orderings to the script's.
- **Coreference:** We can ask participants whether the coreferences that link the script events actually do co-refer. We can also ask the frequency with which a particular coreference holds acrros events. E.g. "At a restaurant, someone serves something and then someone eats something. How likely is it that the same person does both the eating and the serving?" or "How likely is it that the same thing is both served and eaten?" The coreference information in scripts gives them the ability to potentially solve Winnegrad schemas, which have been posed as a good challenge for successful artificial intelligence (Levesque, Davis, and Morgenstern 2012).
- **Human narrative cloze:** In much of the literature on script induction, the narrative cloze task is used to evaluate performance. In this task a subset of documents is held out from training, and at test the model is given some of the events that occurred and asked to guess another event that occurred at some time relative to the known events. The general script knowledge is used to try to fill in the missing information. Although neither humans nor machines can perform this task perfectly, both the model and humans can make guesses. We will compare the responses of the model to responses that humans give, determining the amount of overlap.

  For the original narrative cloze task, only events mentioned in the held-out script could be tested. But asking humans means that we can elicit events that were not mentioned in the text but that probably occurred between two sequential events that *were* mentioned. This will hopefully start to probe people's intuitions about events so commonplace they don't initially think to mention it.
- **Conditional probabilities:** For pairs of related events, we can ask participants the probability of one event happening given the other occurred. We can also estimate this value from the PMI or the bigram frequency of the events.
- **Categorization:** We can categorize entities in the script by what event frames they occur in. If "rice" and "soup" both occur as the objects of "order" and "eat", then it is possible that they are semantically related. We can calculate a similarity metric between entities with this information and also ask people how similar they think the different entities in the dataset are. We can compare these distances as a way of evaluating the knowledge encoded in the script. In doing this, we will leave out proper nouns and only use coreference-resolved noun phrases.

Additional information for evaluating and extending scripts

We can also ask participants some additional questions that the script will not be able to address because of the limitations of our current representation. These answers will be useful in understanding places where the model differs from people on the above tasks and for extending these scripts to a richer representation.

- **Counterfatual dependence:** We can present a sequence of events in a script as having occurred. Then we can ask participants "If it had not been the case that *[event]*, how likely would each of the other events have been?" We can provide participants with sliders for each of the other events in the script, from "extremely unlikely" to "extremely likely". This paradigm is very similar to that used by Lucas and Kemp (2015) in validating their model of counterfactual reasoning.
- **Causal dependence:** For each pair of connected events, we can ask whether participants believe the first event is a cause of the second event.
- **Actual frequency:** For each event, we can ask participants, "How often do you think *[event]* occurs at a restaurant?" and allow participants to respond by adjusting a slider between "almost never" to "almost all the time".

  We can also ask this questions about the responses that people give to the narrative cloze task, allowing us to compare people's estimates for real-world frequency with the ability of the learned script to infer its occurrance.

  We predict that the highest frequency events in the real world (or those that are completely dependent on other events) would consequently have few mentions in language and so be more difficult for the learned script to predict.
- **Discourse frequency:** For each event, we can ask participants, "If *[event]* happens at a restaurant, how often do you think someone will talk about it?" and allow participants to respond by adjusting a slider between "almost never" to "almost all the time". We can compare participants' responses to this question to the actual frequencies of mentioning this event in the corpus. Having both measures will provide us with richer information about relative frequencies and allow us to assess people's awareness of the frequency of mentioning these events in language.

  As is the case for the actual frequency events, we can also ask this question for the responses in the narrative cloze task.

Followup directions

Once we have a sense of which script induction techniques were successful on this dataset and in what ways their results differed from people's intuitions, we can revise and extend the representation and algorithms for learning scripts. We can also extend to other corpora, e.g. other collections of blog posts. We can start to include explicity categorical and causal information into the scripts.

References

Chambers, Nathanael. 2013. *Event Schema Induction with a Probabilistic Entity-Driven Model.*

Chambers, Nathanael, and Daniel Jurafsky. 2008. "Unsupervised Learning of Narrative Event Chains." In *ACL*, 94305:789–797. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.3427&rep=rep1&type=pdf#page=833.

Jans, Bram, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. "Skip N-Grams and Ranking Functions for Predicting Script Events." In *Proceedings of the 13th Conference of the European Chapter* *of the Association for Computational Linguistics*, 336–344. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2380858.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. 2012. "The Winograd Schema Challenge." In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492.

Li, Boyang, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. "Story Generation with Crowdsourced Plot Graphs." In *AAAI*. http://boyangli.co/paper/aaai13.pdf.

Lucas, Christopher G., and Charles Kemp. 2015. "An Improved Probabilistic Account of Counterfactual Reasoning." *Psychological Review* 122 (4): 700. http://psycnet.apa.org/journals/rev/122/4/700/.

Pearl, Judea. 2003. "Causality : models, Reasoning, and Inference." *Econometric Theory* 19: 675–685. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.646.322&rep=rep1&type=pdf.

Pichotta, Karl, and Raymond J. Mooney. 2014. "Statistical Script Learning with Multi-Argument Events." In *EACL*, 14:220–229. http://www.aclweb.org/anthology/E/E14/E14-1.pdf#page=246.

Rudinger, Vera Demberg, Modi Ashutosh, Van Durme Benjamin, and Pinkal Manfred. 2015. "Learning to Predict Script Events from Domain-Specific Text." *acl*.

Schank, Roger C., and Robert P. Abelson. 1977. "Scripts, Plans, Goals, and Understanding: an Inquiry into Human Knowledge Structures." https://halshs.archives-ouvertes.fr/hal-00692030/.

Sloman, Steven A., and David Lagnado. 2015. "Causality in Thought." *Annual Review of Psychology* 66 (1): 223–247. doi:10.1146/annurev-psych-010814-015135. http://dx.doi.org/10.1146/annurev-psych-010814-015135.