Recovering gaps in stories

Erin Bennett (erindb@stanford.edu)
Department of Psychology, Stanford University

Abstract

Humans have a great deal of commonsense knowledge – e.g. about causal relations, abstract roles, and common coocurrances – that we implicitly use in understanding the world. This knowledge allows us to fill in gaps in the stories that people tell, inferring other events that likely happened given what a speaker chooses to mention. Some recent script induction models have been developed to learn commonsense knowledge from natural language corpora. One common evaluation metric for script induction models is the "narrative cloze task" in which one event from a chain of events described in a corpus is left out for the model to recover. While this is a reasonable starting point, based in the intuition that humans are able to infer unmentioned events in stories, several sources of information suggest that the task may not be an ideal metric for commonsense knowledge acquisition. In this paper, we present a pilot experiment which to demonstrate that recovering events that were mentioned in *actual* stories may be very difficult for people to do with much accuracy, even with rich linguistic information to draw on. Our pilot experiments also suggest that agreement among human participants about how to complete a narrative is similarly low, and that even when participants agree with one another, the inferences of an existing script induction model does not appear to track their choices. In an appendix section, we provide illustrative simulations of very simple domains with handwritten "scripts" and show that events mentioned in an informative story tend to be much less recoverable by a rational listener with domain knowledge than events that are true but not mentioned.

Keywords: commonsense knowledge ; script induction ; narrative cloze ; narrative understanding

Introduction

Humans have a great deal of commonsense knowledge that we implicitly use in understanding the world. This commonsense knowledge includes knowledge about causal relations between events. For example if a glass falls on the floor, most people would agree that it will probably break. Commonsense knowledge also involves knowledge about abstract roles that people, places, and things can play in a situation. For example, if someone is a server at a restaurant, we expect that they ask for orders, bring food, and are paid to be there. Commonsense knowledge can also refer to the myriad of facts and features that tend to coocur with particular contexts. For example, when you walk into a restaurant, you expect to see tables, silverware, and possibly a bar. Understanding and representing this kind of human knowledge in machines is a long-standing problem in artificial intelligence, and many researchers have thought about this problem and made useful contributions.

In this paper we do some stuff.

Background

Scripts in Psychology Researchers have identified many subtle ways in which human commonsense knowledge comes to play in understanding stories, from which events of a story are more or less memorable to whether a definite (the) or indefinite (a) article is more appropriate in a given context. One way of articulating commonsense knowledge is that it can be organized into scripts, "structure that describe appropriate sequences of events in a particular context, ... made up of slots and requirements about what can fill those slots," (Schank & Abelson, 1977). These knowledge structures resemble theatrical scripts in that they the include roles that different agents (or "actors") can play, props that are likely to be used, and basic events that combine together to form cohesive "scenes". Also specified in a script as a knowledge representation are various preconditions and consequences of the different actions that might take place. Individual handwritten scripts for domains like restaurants, movie theatres, etc. have been created and used to predict many psychological phenomena.

For example, Schank and Abelson (1977) note that when a script is in play, its typical props and characters can be referenced by definite articles (the as opposed to a). For example, the sentence When Anna was at a restaurant, she got into an argument with the waitress, and the sentence When Anna was at a restaurant, she got into an argument with a police officer, sound natural to most speakers of English but When Anna was at a restaurant, she got into an argument with the police officer, sounds much less natural. They also note that events that are typical of a script but that dont happen in a particular story are still counterfactually salient. For example, going to a restaurant and not ordering food is something very natural to comment on, whereas going to a restaurant and not singing the Star-Spangled Banner is not.

Winnograd schemas are another place where people's commonsense knowledge is made apparent (Winograd, 1972). A Winnograd schema (Winnograd, 1972) is a short, sentence-long story involving anaphora resolution to one of two characters or objects. By changing a single word in a Winnograd schema, the correct referent for the anaphora changes from one to the other. For example, in the sentence Anna was dining in a restaurant, and she got into an argument with the waitress because she didnt bring any (money/food), when the last word of the sentence is money, the referent of she is likely to be Anna, but when the last word of the sentence is food, the referent of she is likely to be the waitress. Winnograd schemas highlight in particular a discrepency between human's language understanding and that of artificial intelligence systems. Winnograd schemas are blatantly obvious to humans, but completely opaque to our current technology in automatic language understanding. A collection of sentences like these has been proposed as an alternative to the Turing test as an evaluation metric of intelligent AI (Levesque, Davis, & Morgenstern, 2011).

A very key feature of script-like knowledge that Schank and Abelson (1977) point out is the fact that humans can fill in gaps in stories. For example, John went to a restaurant. He ordered chicken. He left a large tip, implicates a lot of other events, like John sat down, and John read the menu, and John ate chicken.

Bower, Black, and Turner (1979) demonstrated that peoples memory of the contents of a story depend very strongly on the relevant scripts and how typical or atypical the events of this particular story were with respect to the general script. These authors showed that when participants read about script-relevant events that occurred in an atypical sequence (e.g. a customer ordering their food first and then sitting down), they later misremembered the events as having happened in the more typical order (the customer sitting down and then ordering food). They also showed that participants frequently mistakenly believed that the stories they read mentioned events that were not in fact mentioned when those events were typical events for the relevant script.

These observations that people can (and automatically do) infer missing information motivated the narrative cloze task for script induction models (Chambers & Jurafsky, 2008), which we explore in detail this paper.

Script Induction in Natural Language Processing Handwritten scripts are somewhat bounded since they can be time-consuming and require expert annotation, and because people seem to know many more scripts than researchers could reasonably write down. And so researchers in natural language processing have recently been trying to learn scripts from natural language corpora.

In early work on this task by Chambers and Jurafsky (Chambers & Jurafsky, 2008), data from a training corpus were processed into narrative event chains: chains of events (essentially main verbs) that share a coreference to the same person or object (e.g. "Amy went to a restaurant. She ordered chicken. The waitress told Amy they were out of chicken." would yield the event chain "[A] go, [A] order, [A] is told" where [A] represents the referring entity (Amy) between those events). The authors attempted to learn scripts by calculating the point-wise mutual information between events (given their coocurrance in a narrative event chain) and clustering on those values. They showed that clustering on coocurrance in a document alone did less well on their evaluation metric.

Pichotta and Mooney (Pichotta & Mooney, 2014) later expanded on the single event chain models to include multiargument events and track the structure of coreference across all of them. This type of model captures information about the roles that multiple agents play with respect to one another, e.g. If one agent asks the other, the other is likely to respond. Over this representation, the authors articulated a statistical coocurance probability for these events with abstract role structure. Using the same evaluation metric as Chambers and Jurafsky's 2008 model, they achieved slightly higher performance.

These algorithms require evaluation metrics to determine successful learning and inspire extentions. For this purpose, Chambers and Jurafsky (2008) suggested an initial evaluation metric in which sentences could be removed from chains of events and models could be tested on whether they can appropriately fill in the left-out event. This task, the "narrative cloze task" has been used to evaluation a variety of script induction models (e.g. Chambers, 2013; Pichotta & Mooney, 2014; Rudinger, Demberg, Modi, Van Durme, & Pinkal, 2015; Pichotta & Mooney, 2016).

Narrative Cloze Task

We would like to know whether narrative cloze task is an appropriate measure of human-like commonsense.

Chambers and Jurafsky (2008) proposed an evaluation metric for script induction. Inspired by the fact that people can fill in gaps in stories based on their knowledge of a domain, they reasoned that a model with sufficient domain knowledge could recover a left-out event from a held-out event chain from the original corpus. They took a subset of their held-out event chains and tested the model on a Narrative Cloze Task where one event was deleted from the chain and the model was tasked with inferring what event was likely to have been left out of that chain. In their original paper, they chose the event with the highest average PMI with the rest of the events in the chain.

One problem for script induction models has been finding an appropriate evaluation metric. The kinds of commonsense knowledge that we seek to characterize and that computers have had difficulty representing is necessarily abstract and nontrivial to communicate. The narrative cloze test, although a useful metric in that it does not require any additional human annotation, does not obviously track the desired commonsense knowledge that people have. For one thing, as Chambers and Jurafsky (2008) pointed out when they proposed the task, humans do in fact have access to the kind of commonsense knowledge these models aim to learn, but the authors predicted that, at least given the sparse linguistic information provided to their model, humans might do very poorly at recovering the left-out event in a narrative event chain that corresponds to the original text The authors explain that while they do not think a models success at the Narrative Cloze Task implies it has successfully learned commonsense knowledge, it is a reasonable starting point, since one of the phenomena associated with scripts is the ability to fill in (or even misremember) left-out events from a story (Schank & Abelson, 1977; Bower et al., 1979). In later work, Rudinger, Rastogi, Ferraro, and Van Durme (2015) essentially confirmed Chambers and Jurafskys speculation that success at the Narrative Cloze Task does not imply successful script learning. Rudinger, Rastogi, et al. (2015) ran several different classes of models, which they trained on the Gigaword corpus and tested on Narrative Cloze Tests held out from the same corpus. They showed that models that had been developed as language models significantly outperformed models that had been developed for acquiring commonsense knowl-

edge. The conclude that either language models are better at learning commonsense knowledge than models developed for that purpose, or success at the Narrative Cloze Task does not map onto successful commonsense knowledge induction. Given that they also showed qualitative evidence that the narrative event-chain script induction model actually did learn clusters of events that intuitively belong to the same script, it seems unlikely that language modeling per se is actually the most promising way to learn commonsense knowledge.

In our pilot experiment, we provide support for the speculation that the narrative cloze task is in fact difficult for humans, and show that this remains true (though perhaps slightly less so) even when people are given detailed linguistic context.

Pilot Experiment

Participants

For our pilot experiment, we recruited 24 participants from Amazon Mechanical Turk. All participants were from the U.S. Due to technical errors, a number of other participants (approximately 5) participated but we were unable to record their results.

Materials

Rudinger, Demberg, et al. (2015) demonstrated that reasonable performance could be achieved on domain-specific natural text. Using "Dinners from Hell" (Rottler, 2007), a blog about negative experiences in restaurants, Rudinger et al. trained several coreference-chain models and evaluated their performance using a narrative cloze task. We used the same blog as our training corpus for the single-protagonist narrative chain PMI model and as a source of our narrative cloze tests. We scraped 273 stories from the blog, and uniformly selected 17 documents for the narrative cloze tests in our experiment, excluding a few off-topic documents, e.g. letters to the editor. For each document in the corpus, we used CoreNLP annotators (Manning et al., 2014) to get dependency parses (Chen & Manning, 2014) and coreference chains (Recasens, Marneffe, & Potts, 2013; Lee et al., 2013, 2011; Raghunathan et al., 2010). Documents varied in the number of coreference chains they contained. We chose up to 3 sentences from up to 2 coreference chains to create our 3 cloze tasks per document (or occassionally fewer if a document had only one coreference chain, or only short coreference chains). In each cloze task, we showed all the sentences from the coreference chain except for the chosen sentence. At that index in the coreference chain, we left a blank text input for people to freely fill in their own guess about the full sentence (Figure 1). There were a total of 46 cloze tasks, and each participant saw a random set of 17 (one for each document).

The models tested by Rudinger, Demberg, et al. (2015) had access to only a small amount of linguistic information in the narrative cloze test, since each event was reduced to a single verb and the syntactic role of protagonist. We wanted to test people's performance on the task under this constraint, but we also wanted to test them on more complex versions of the task, since other more complex models can be tested on narrative cloze tasks with more linguistic information.

We therefore manipulated the original text in three different ways. In one condition, we gave people the full text of every sentence that contained a mention of the protagonist, which is much more information than the coreference chain models had access to in their version of the narrative cloze task. In a second condition, somewhat resembling the structure of the narrative cloze task used by Pichotta and Mooney (2014), we provided people with a "caveman-speak" version of the sentence that contained the lemmatized (or participle-form in the case of passive sentences) main verb and a few of its principle arguments: the head of any subject or object noun phrases and the preposition and head noun of any prepositional phrase. For example, the sentence "Just to spite him, we remained at our table for nearly 3 hours." was reduced to "We remain at table for hours." in this condition. In a third condition, we provided only the main verb and its subject.

Procedure

The format of the task was similar across conditions. We blurred the original text behind each sentence and provided the text for participants on top of this, dependent on condition (Figure 1). Above the text box in small letters, we reminded participants to provide a full sentence that they think might have been at that position in the text. In earlier pilots we found that participants sometimes misunderstood the task and gave phrases or fragments instead of sentences. We therefore confirmed that each sentence could be parsed by the CoreNLP tools (Manning et al., 2014) by running a webserver and parsing participants' responses in real time before accepting each of their responses and allowing them to continue to the next question.

Analyses

Response processing We extracted main verbs from the responses using CoreNLP, and used NLTK's interface to WordNet (Bird, Klein, & Loper, 2009; Miller & Fellbaum, 1998) to find all synonyms of that word in all of its synsets. If any verbs or synonym that participants mentioned was the same as the original verb, we recorded that participant's response on that cloze task as correct. We call this matching method "automatic matching". As this resulted in very low scores, we also manually determined a single word "gloss" of each response using the experimenter's subjective judgement. These glosses were chosen generously such that for any set of responses with similar meanings, they were labeled with the same gloss. We consider this a reasonable upper bound on how well participants might be doing on this task. We call this matching method "manual matching". Given these two metrics for successfully matching the original event, we measured overal performance as the average success for participants across tasks.

Agreement Metric Regardless of whether people are able to recover the actual underlying text, we were also interested

(a) (b)

Figure 1: Two example trials, one (a) of the original text condition, the other (b) of the "caveman" condition.

in how consistent people's responses might be in tasks like this.

To do this we need to a quantity to represent "agreement". We calculated an empirical entropy of the sample distribution of human responses for each cloze task. We first find a gloss for each response, which we determine by finding the root verb and then choosing the name corresponding to that verb's first WordNet synset's. Given these responses, we caculate the spread by use a boostrap correction to a naive estimate of entropy (DeDeo, Hawkins, Klingenstein, & Hitchcock, 2013), which is less biased than a naive estimate alone (in which $P(response) \approx \hat{P}(response) := \frac{N_{response}}{N_{total}}$). However, since different conditions had different total numbers of participants, this quantity alone would be a poor representation of the spread of responses. For example, if 2 participants gave a total of 2 unique responses our sample distribution of responses would have the same calculated entropy as if 5 of 10 participants gave one response and the other 5 gave another. Intuitively, the latter represents clearer agreement among participants than the former, and so we divide this empirial entropy by its maximum possible value for that set of participants. Our measure of agreement is the additive inverse of this scaled entropy, namely:

$$\text{agreement} := -\frac{\hat{H}(R)}{\log(N)}$$

where N is the number of participants for that task and $\hat{H}(R)$ represents the entropy estimate across responses for the task. This metric is not defined if N=1, and so we exclude the task and linguistic condition combinations in the pilot for which we collected only 1 participant's response.

The agreement metric tracks humans' intuition rather than the original text, and in this respect it is similar to the metric used by Pichotta and Mooney (2016) for evaluating their model's performance, since they also elicited people's judgements about held-out events. But rather than ask participants to generate narrative cloze answers, they asked participants to judge the *model's* top-ranking answers, and so they did not get as much information about what human performance might look like on this task.

Single-entity Event Chain Model Comparison We compare the human data to the single-entity event chain model of Chambers and Jurafsky (2008). In this model insert math here. Rudinger, Demberg, et al. (2015) helpfully provided code for this model that we were able to use with only slight modification. We run the model with the maximum likelihood parameters from Rudinger, Rastogi, et al. (2015)'s experiments. While this is not the most recent or successful model of its kind, it performed well among the models tested by Rudinger, Rastogi, et al. (2015) and provides an interesting comparison to the human performance.

The model produces ranked lists of candidate events, and various metrics have been adopted for assessing performance. One common metric is recall of the actual event within the top N candidates (different papers use different thresholds). Another metric is to determine the model's ranking of the actual event. Because we want to measure against the original text but also against the varied responses that participants gave, we would like a metric of success that considers a set of correct events, rather than a single correct event. We therefore chose, among the set of "correct" events, the highest ranking event[1] and used its rank to represent the model's performance on that task. For tasks where none of the correct events were recalled, the ranking 100 is used for graphing.

Results & Discussion

Overall, as Chambers and Jurafsky (2008) predicted when they introduced this task, people do not perform very well at recovering the original left-out event from an event chain. On average across tasks, people recover the main verb or one of its synonyms 5% of the time. There is no significant difference between linguistic conditions for this metric (p¿0.05). The experimenter's manual judgements showed that people might be able to recover the gist of the original text as much as 8% of the time. When we use the experimenter's judgement as the metric for success, we actually do find a significant improvement in performance in the condition where the contextual sentences are provided in their original detail (b=0.045, t(2, 126)=3.31, p=0.001). These results are shown in Figure

[1]Taking the average of the rankings yielded similar results.

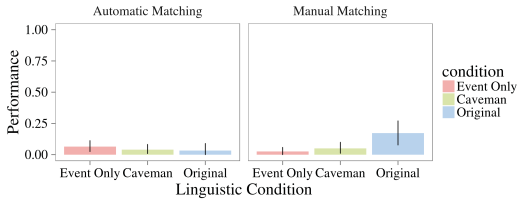

Figure 2: Human performance on the narrative cloze task is quite poor across trials and across all three linguistic conditions. This is the case even when responses are manually glossed into categories by the experimenter, however with this metric performance is slightly higher when the original text is provided.

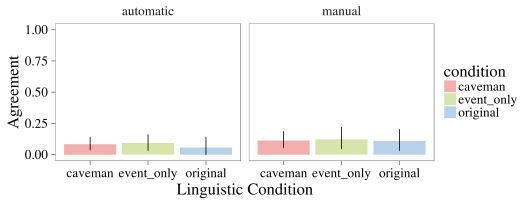

Figure 3: Human agreement on the narrative cloze task is quite low across trials and across all three linguistic conditions. This is the case even when responses are manually glossed into categories by the experimenter.

2.

people don't agree with one another

people's agreement doesn't track model agreement

Discussion

discuss

in response to concerns like these about narrative cloze tasks, researchers have attempted adaptations of the narrative cloze task, e.g. story cloze task.

In response to recent concerns about the effectiveness of narrative cloze tasks for specifically targeting commonsense knowledge, Mostafazadeh et al. (2016) very recently developed a dataset with an alternative evaluation metric, a *story* cloze task in which models must determine the conclusion of a crowdsourced 4-sentence story segement, and Pichotta and Mooney (2016) chose to extend their narrative cloze tasks so that the model's inferences were compared to people's inferences about what might have been the held-out event.

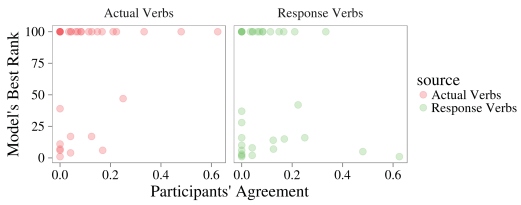

Figure 4: The model does not appear to increase in its ability to recover at least one synonym of the humans' responses as agreement between participants increases.

Future Directions

larger N

While natural language is readily available in vast quantities on the internet, sentences that people freely generate tend to be very complex and can often mention events or ideas outside the story they are telling. These features of natural language make both script induction and narrative cloze tasks difficult. For this reason, there have been efforts to use crowdsourcing to collect corpora of stories that are more focused, shorter, and contain more simple language (Li, Lee-Urban, Johnston, & Riedl, 2013; Mostafazadeh et al., 2016). Li et al. (2013) collected stories focused on particular topics (a trip to a restaurant, a bank robbery, etc.) for which they asked the story writers to use very simple language and then recruited readers to identify and exclude mentions of unrelated events. Mostafazadeh et al. (2016) crowdsourced a slightly different, highly structure corpus of 5-sentence-long stories. These stories span a huge variety of commonsense domains, as story writers were told to write about anything they thought readers would easily understand.

It could be argued that given sufficiently curated corpora, narrative cloze tasks would be easier for humans and their success reflective of their commonsense knowledge. One possible test of this would be to repeat all of our analyses on these more simplified crowdsourced corpora.

Also, we could use word2vec, glove, or sentence2vec embeddings to check similarity between people's responses, though since manually annotating the responses didn't seem to change the results, that might not be all that worthwhile.

more simulations

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. " O'Reilly Media, Inc.".

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive psychology*, *11*(2), 177–220.

Chambers, N. (2013). Event schema induction with a probabilistic entity-driven model. In *Emnlp* (Vol. 13, pp. 1797–1807).

Chambers, N., & Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Acl* (Vol. 94305, pp. 789–797).

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Emnlp* (pp. 740–750).

DeDeo, S., Hawkins, R. X., Klingenstein, S., & Hitchcock, T. (2013). Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, *15*(6), 2246–2276.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, *39*(4), 885–916.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task* (pp. 28–34).

Levesque, H. J., Davis, E., & Morgenstern, L. (2011). The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning.*

Li, B., Lee-Urban, S., Johnston, G., & Riedl, M. (2013). Story generation with crowdsourced plot graphs. In *Aaai.*

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations* (pp. 55–60). Available from `http://www.aclweb.org/anthology/P/P14/P14-5010`

Miller, G., & Fellbaum, C. (1998). *Wordnet: An electronic lexical database.* MIT Press Cambridge.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., et al. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. *Proceedings of NAACL HLT, San Diego, California, June. Association for Computational Linguistics*.

Pichotta, K., & Mooney, R. J. (2014). Statistical script learning with multi-argument events. In *Eacl* (Vol. 14, pp. 220–229).

Pichotta, K., & Mooney, R. J. (2016). Learning statistical scripts with lstm recurrent neural networks. In *Proceedings of the 30th aaai conference on artificial intelligence (aaai-16).*

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., et al. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 492–501).

Recasens, M., Marneffe, M.-C. de, & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Hlt-naacl* (pp. 627–633).

Rottler, G. (Ed.). (2007). *Dinners from hell.* Available from `http://www.dinnersfromhell.com`

Rudinger, R., Demberg, V., Modi, A., Van Durme, B., & Pinkal, M. (2015). Learning to predict script events from domain-specific text. *Lexical and Computational Semantics (* SEM 2015)*, 205.

Rudinger, R., Rastogi, P., Ferraro, F., & Van Durme, B. (2015). Script induction as language modeling. In *Proceedings of the 2015 conference on empirical methods in natural language processing (emnlp-15).*

Schank, R. C., & Abelson, R. P. (1977). Scripts. *Plans, Goals and Understanding*.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*(1), 1–191.

Appendix: Illustrative Simulations

We did a few different simulations of the narrative cloze task, assuming a tiny little toy domain with "scripts" that we hand wrote.

Unsurprisingly, a model that infers left-out events from a story does worse when the story was generated by an informative storyteller than when the story is just a random subsequence of events also, its pretty easy to infer events that *weren't mentioned* by an informative storyteller, which is the actual thing that people said scripts were good for and which inspired the narrative cloze task.