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 an experiment which 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. We also show 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 existing script induction models does not appear to track their choices. Finally, in simulations of very simple domains with handwritten "scripts", we confirm 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.

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: stereotypical sequences of events for a particular domain (Schank & Abelson, 1977). Individual handwritten scripts for domains like restaurants, movie theatres, etc. have been created and used to predict phenomena in story understanding.

end with: people can fill in gaps

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.

(Chambers & Jurafsky, 2008) (Chambers, 2013) (Pichotta & Mooney, 2014)

end with: narrative cloze task

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, et al., 2015; Pichotta & Mooney, 2016).

Narrative Cloze Task

we would like to know whether narrative cloze task is a good measure of human commonsense.

Rudinger, Rastogi, et al. (2015) showed that narrative cloze tasks tend to be more effectively solved by shallow language models than by models aimed at acquiring commonsense knowledge. In response to 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. 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.

how does the fact that natural stories are generated informatively affect this task?

Experiment

Participants

We recruited N participants from Amazon Mechanical Turk. All participants were from the U.S. N participants were excluded because they were not native English speakers.

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 N 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). how we chose cloze tasks for documents

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.

Procedure

in pilots, people did not understand the task. we blurred out the original text and inserted a textbox with instructions near it we used corenlp on the webserver to check that all responses were parsable before participants' responses would be accepted and they could 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 et al., 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. 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.

PMI model comparison Rudinger, Demberg, et al. (2015) helpfully provided code that we were able to use with only slight modification

Results

people are not good at this task.

The agreement metric is similar to that of Pichotta and Mooney (2016), 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. They showed that their LSTM's candidate responses received an average rating of 3.67 on a 0 to 5 rating scale.

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.

Discussion

Future Directions

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 et al., 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.

References

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

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).

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).

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*.