Towards Using Word Embeddings for Estimating the Perceived Interestingness of Narrative

Anonymized

Abstract

With the emergence of interactive agents, narrative generation is developing a new paradigm of use case, one in which estimating the quality of the perception of narrative is a more important evaluation metric than other ones commonly used, such as text coherence or domain-specific criteria. In order to increase the chances of creating a narrative that is perceived as interesting, a narrative generator needs to consider the cognitive processes involved in the perception of the narrative. Using theories of cognitive interest, we explore the usage of word embedding vectors to introduce a proxy measure for cognitive interest, without relying on a semantic model of the story domain.

Background

Use cases of narrative generation can guide and inform the methods with which narratives are created. Both in terms of method constraints and the factors for which the generated stories are optimized, the paradigm in which a narrative is delivered to the user is extremely consequential. For instance, entertainment has been a prominent use case of narrative and story generation. Simulation games benefit from having versatile storylines and game characters can cause more immersion and believability when they are part of a tailored sequence of events (Mateas and Stern 2003; McCoy et al. 2011). As another example, in games involving interactive narrative, story generators give the user the option to influence the storyline (Riedl and Bulitko 2012).

In games, whether stories are interactive or not, it is often possible to infer the quality or interestingness of the generated story, and by extension, how it is likely to be perceived, using known domain semantics. Therefore, these semantics (e.g. game state) can guide the story generation process. Even when the use case of story generation is not games, many story modeling and narrative generation approaches rely on such similar semantic model of a given domain (e.g. characters, goals, relationships, etc.) which allows the derivation of a sequence of events and ultimately a narrative, such as in (Elson 2012).

Other narrative generation methods have less reliance on a specific use case in which the narrative is delivered to a user, and hence, in order to assess the quality of the generated story, focus more on the general properties of the generated text, such as coherence or the causal plausibility of the sentence ordering. Sometimes referred to as *open story generation*, such methods involve generating narratives without a priori domain knowledge (Martin et al. 2017; Swanson and Gordon 2008). In (Purdy et al. 2018), a set of proxy measures are introduced to assess the story quality, in an open story generation task; these measures are: correct grammar use (“grammaticality”), complexity of the language (“narrative productivity”), similarity of adjacent sentences (“local contextuality”) and adherence to the usual ordering of events in similar stories (e.g. “eat” comes after “order”, “temporal ordering”). These measures are shown in (Purdy et al. 2018) to correlate with human judgment of story quality, and hence, can be used towards evaluation of generated text and easier fine-tuning of generative approaches.

Proxy measures introduced above are a useful start to assessing narrative quality when it is not tied to a specific domain of semantics. However, with the emergence and rapid adoption of interactive agents in our lives, it is plausible to assume that a major use case of storytelling, and hence a prominent paradigm of delivering generated narratives to the users, will involve interactive agents such as virtual agents and sociable robots (Goodrich, Schultz, and others 2008; Fong, Nourbakhsh, and Dautenhahn 2003) or those in smart speakers (NPR 2017). In such cases, a language-based narrative (textual or vocal) need not only be of high quality in terms of the proxy measures recited above, but it also has to be *interesting* to the user.

Interestingness

Storytelling, as an intuitive, natural and commonplace human behavior, seems deceptively simple to judge in terms of interestingness: *“was that an interesting story?”* However, similar to some other intuitive, natural and commonplace behaviors (such as nodding or gazing), it is extremely complicated to evaluate a story’s interestingness, both subjectively and objectively.

The perceived interestingness of a story is hard to predict or model; it can be subjective, it is often cultural and it can also change over time. Moreover, the subtleties and art of authorship makes the ways in which a narrative can seem interesting to humans incredibly diverse, nuanced and hard to model. This is especially true when the domains in which the stories are told are unpredictable and can change based on the nature of the interaction, events in the environment, the parties involved, and other factors; and hence, any agent attempting to sustainably tell social and interesting stories to humans will face this challenge.

Despite such difficulties, there are ways in which we can start to develop proxy measures for perceived story interestingness; for instance, using the literature of cognitive science and by developing methods that can draw from them (Behrooz et al.). The first category ("Experiential Interests") consists of more subjective reasons of why a story could be interesting, such as *instinctive interests* (e.g. danger, sex, or power) and *topic interests* (i.e. a specific topic that a certain person enjoys). The second category, "Cognitive Interests", is less subjective, and is caused by factors such as *unexpectedness* (e.g. expectation violation) or *predictive inference* (e.g. foreshadowing). Indeed, such cases are not an exhaustive list of what can make a story interesting, but can be a basis for developing story interestingness proxy measures.

Search for Specificities

A particular angle that adds to the value of interestingness proxy measures is the potential role that such measures can play in a search problem that can arise in narrative generation. If a narrative generation system, for instance one used by an agent operating in the real world, attempts to build a narrative from events that have happened previously, there would be a search problem involved to choose which non-essential details should be included in the story. At a minimum, a sequence of events can be described as a mundane narrative (that only describes what happened with simple sentences); however, the inclusion of certain specificities about the events can potentially make the narrative more interesting. The "Chekhov's gun" principle says: "every element in a story must be necessary, and irrelevant elements should be removed."

This search problem can also arise when Recurrent Neural Networks (RNNs) are used to generate stories, such as in (Martin et al. 2017). In order to increase the chances of convergence in RNNs, such methods often group the semantically related words and verbs in a story corpus to a generalized concept (e.g. through semantic trees (Schuler 2005; Miller 1995)) before training a model. However, the generated narratives would then also include such generalized concepts, and hence can be more mundane and less specific. Having proxy measures to find the more interesting specificities may offer a solution, and particularly, word vectors can help with choosing a specific instance of the word. This lack of specificity can occur in any generative method that results in stories that lack specificity, such as Plot Graphs (Li et al. 2013).

Interestingness Proxy Measures

In the absence of a domain's semantic model (as explained in Sec.), we explore the idea of using word embedding vectors, such as GloVe (Pennington, Socher, and Manning 2014), with the goal of developing proxy measures for narrative interestingness. Word vectors introduce a way to estimate the semantic similarity and relationships between the words. In this paper, we focus on proxy measures of cognitive interests, such as predictive inference.

Consider the sample story in Table. 1, which contains a case of predictive inference (or foreshadowing), and hence, a potential to cause cognitive interest in the reader, based on cognitive science research (Behrooz et al. ; Campion 2004). We use part-of-speech tags to focus on the words that capture most of the meanings in the story. In this case, we have focused on **verbs** (extracted as verb root via VerbNet (Schuler 2005)), **nouns** (excluding the named entities, such as "Sam") and **adjectives**. Table 2 shows the extracted words for the story in Table. 1 (using Stanford CoreNLP (Manning et al. 2014)).

Table 1: A sample story which contains a case of foreshadowing. The numbers on the left are story event indexes.

1	Sam and Judy went out for dinner at their favorite restaurant.
2	While driving to the restaurant, Judy's favorite song played on the radio.
3	Sam found a parking space at the very front of the restaurant.
4	Sam and Judy were seated immediately and ordered their favorite food to the waiter.
5	She looked distracted and tired but was polite while taking their order.
6	Sam's favorite song played on the radio while they waited for their food.
7	When the waiter returned with their food it was all wrong!
8	The waiter apologized and returned a few minutes later with the correct order.
9	Sam and Judy enjoyed their meal.
10	They paid their tab, left a tip for the waiter, and drove back home.

Table 2: Extracted keywords from the story in Table. 1.

waiter, return, pay, song, seat, order, radio, look, go, apologize, dinner, take, home, wrong, favorite, find, space, leave, minutes, restaurant, food, enjoy, parking, tired, drive, distracted, front, correct, meal, tip, tab, play, wait

In order to visualize the word vectors corresponding to the keywords in Table 2, we used the T-SNE algorithm (Maaten and Hinton 2008), and the result can be seen in Fig. 1.

The T-SNE visualization shows that certain clusters of words can be distinguishable from others. These clusters can categorize the major sub-sequences of events in the story,

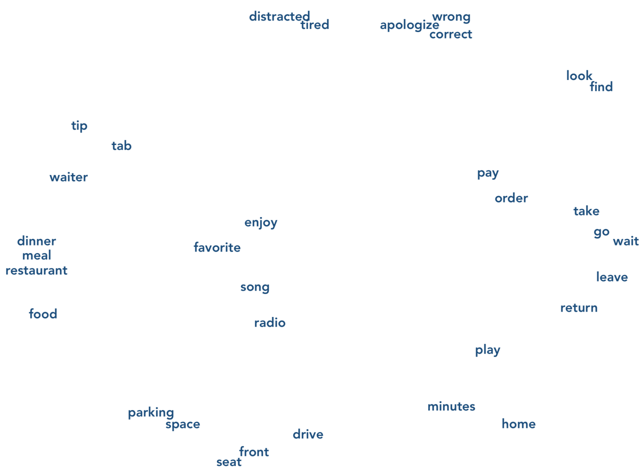

Figure 1: 2D T-SNE visualization of the GloVe vectors representing the keywords in Table 2.

and their distance from each other can signal how much of an anomaly a certain cluster is. For instance, we see clusters about dining, about car and parking, or about music. We also find a cluster containing the words “distracted”, “tired”, “apologize”, “wrong” and “correct”, which we will refer to as cluster X.

The significance of cluster X is that it contains the words that form the foreshadowing in our story. We will explain two observations that can guide us towards estimating the presence of foreshadowing:

1. We can observe that the words of cluster X belong to 2 different non-adjacent areas of the story (event 5 and events 7-8 in Fig. 1).
2. We first define the average of a cluster as the mean value of the word vectors that it contains (e.g. $mean(X)$), and the average of all of the story keywords as $mean_{story}$. We then observe that $mean(X)$ has the lowest cosine similarity score with $mean_{story}$ than any other cluster.

These two observations provide us with a basis to estimate the existence of foreshadowing or predictive inference. To that end, this approach can help us build proxy measures towards estimating the perceived cognitive interest in a generated narrative.

Furthermore, this approach may enable the development of other proxy measures for cognitive interest too. For instance, unexpected events can cause cognitive interest in stories (Schank 1979), and the *unexpectedness* of an event can be defined using the cosine distance of sentence vectors (e.g. using (Pagliardini, Gupta, and Jaggi 2017)).

Conclusion

Narrative generation will evolve in its use cases over time, and a growing category of these use cases will relate to interaction with agents. In such cases, the perceived interestingness of narrative is important to consider in the generation process. Embedding spaces of constituent words offer a basis for estimating the perception of a narrative, and can consequently enable the development of proxy measures that help improve the generation methods. In this paper, we outlined an approach that can estimate the presence of predictive inference (foreshadowing) and hence cognitive interest (Campion 2004).

We plan to expand this approach and seek to find correlations between proxy measures of cognitive interest and judgments of human subjects, similar to story quality measures introduced in (Purdy et al. 2018). Such evaluations should involve a corpus of stories, such as a corpus of movie plot summaries (Bamman, OConnor, and Smith 2013).

References

Bamman, D.; OConnor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 352–361.

Behrooz, M.; Mobramaein, A.; Jhala, A.; and Whitehead, J. Cognitive and experiential interestingness in abstract visual narrative.

Campion, N. 2004. Predictive inferences are represented as hypothetical facts. *Journal of Memory and Language* 50(2):149–164.

Elson, D. K. 2012. Detecting story analogies from annotations of time, action and agency. *Proceedings of the Third Workshop on Computational Models of Narrative* (1981):91–99.

Fong, T.; Nourbakhsh, I.; and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42(3-4):143–166.

Goodrich, M. A.; Schultz, A. C.; et al. 2008. Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction* 1(3):203–275.

Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. 2013. Story generation with crowdsourced plot graphs. In *AAAI*.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.

Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Martin, L. J.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. O. 2017. Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331*.

Mateas, M., and Stern, A. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference*, volume 2.

McCoy, J.; Treanor, M.; Samuel, B.; Mateas, M.; and Wardrip-Fruin, N. 2011. Prom week: social physics as gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, 319–321. ACM.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

NPR, E. R. 2017. *The Smart Audio Report*. Accessed: 2018-07-20.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Purdy, C.; Wang, X.; He, L.; and Riedl, M. 2018. Predicting generated story quality with quantitative measures.

Riedl, M. O., and Bulitko, V. 2012. Interactive narrative: An intelligent systems approach. *AI Magazine* 34(1):67.

Schank, R. C. 1979. Interestingness: controlling inferences. *Artificial intelligence* 12(3):273–297.

Schuler, K. K. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Swanson, R., and Gordon, A. S. 2008. Say anything: A massively collaborative open domain story writing companion. In *Interactive Storytelling*. Springer. 32–40.