Lexical preferences in an automated story writing system

Melissa Roemmele and Andrew S. Gordon
Institute for Creative Technologies, University of Southern California
roemmele@ict.usc.edu, gordon@ict.usc.edu

1 Introduction

The field of artificial intelligence has long envisioned the ability of computers to automatically write stories (Dehn [1981], Lebowitz [1985], Meehan [1977], Turner [1993]). For a long time, progress on this task was limited by the difficulty of encoding the vast narrative knowledge needed to produce stories with diverse content. The rise of data-driven approaches to AI introduced the opportunity to acquire this knowledge automatically from story corpora. Since then, this approach has been utilized to generate narratives for different domains and genres (Li et al. [2013], McIntyre and Lapata [2009]), which has in turn made it possible for systems to collaborate with human authors in developing stories (Khalifa et al. [2017], Manjavacas et al. [2017], Swanson and Gordon [2012]).

Roemmele and Gordon [2015] introduced a web-based application called Creative Help that provides automated assistance for writing stories. The interface consists of a text box where users type "\help\" to automatically generate a suggestion for the next sentence in the story. One novelty of the application is that it tracks users' modifications to the suggestions, which enables the original and modified form of a suggestion to be compared. This enables sentences generated by different models to be comparatively evaluated in terms of their influence on the story.

We examined a dataset of 1182 Creative Help interactions produced by a total of 139 authors, where each interaction consists of the generated suggestion and the author's corresponding modification. The suggestions were generated by a Recurrent Neural Network language model (RNN LM), as described in Roemmele et al. [2017], which generates sentences by iteratively sampling words according to their observed probability in a corpus. The training corpus for the model analyzed here was 8032 books (a little over half a billion words) in the BookCorpus[1], which contains freely available fiction from a variety of genres. This paper briefly characterizes the generated sentences by highlighting their most prominent words and phrases and showing examples of them in context.

2 Analysis and Discussion

Table 1 shows the most common lemmatized words generated by the model according to grammatical (part-of-speech) categories[2], as well the most common two and three word phrases (bigrams and trigrams), with the total count of each item in parenthesis. The total number of words generated across all suggestions was 16025. Table 2 shows some examples of suggestions that feature these lexical items.

The model has an interesting writing style and preference for topics. It commonly conveys both narrated and quoted dialogue, as indicated by the frequency of the verbs *say, tell, ask, whisper* and *talk*. It is surprisingly eager to introduce a *monster* into a story. It refers frequently to body parts (*hand, eye, body, face, head, blood*), elements of a residence (*home, door, room, bed*), and *time* (along with specifically *day* and *night*). It likes abstract nouns (*way, idea, mind, thing, life, truth*). It often

[1] yknzhu.wixsite.com/mbweb
[2] Part-of-speech tags and word lemmas were automatically detected using the spaCy library

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Verbs	be(147), know(110), go(105), would(73), want(70), say(69), think(63), look(56), tell(44), stay(41), feel(37), ask(33), come(30), happen(28), get(24), wait(23), have(22), find(21), whisper(20), talk(18), leave(18), take(17), see(17), pull(16), open(16), sit(15), need(15), hold(15), expect(14)
Nouns	time(59), way(42), monster(36), thing(33), hand(29), man(27), door(27), world(25), eye(24), night(23), life(22), father(22), girl(21), body(20), person(19), voice(18), day(18), face(17), room(16), blood(15), bed(15), mind(14), idea(14), head(14), chance(14), truth(13), word(12)
Pronouns	I(638), she(346), it(212), you(203), he(145), her(113), they(77), me(77), him(74), we(53)
Adjectives	sure(58), able(45), good(28), little(20), sorry(19), right(17), ready(17), real(15), glad(14), troubled(12), beautiful(12), afraid(12), wrong(11), long(11), new(10), strong(9), true(8), alive(8)
Adverbs	not(218), maybe(27), away(10), probably(6), home(6), forward(6), especially(6), right(5)
Bigrams	not know(26), not sure(25), be go(22), not want(18), be sorry(14), not go(11), feel like(10), not expect(8), look like(7), not think(6), monster say(6), not handle(5), good idea(5)
Trigrams	not be able(8), able to stay(8), want to know(7), be a good(6), want to stay(5), able to handle(5)

Table 1: Most common generated words/phrases

"I'm not mad enough to have you asking me to stay," she whispered.
Some, the monster said I'd be a child, and surely some of the world had been able to capture him.
I thought that's what I was going to be, and I wasn't sure that he would be able to stay here.
I thought about that thing, but I knew I couldn't handle it.
"You are mine and I'm afraid of you, Edgar."
She asked, her voice thick with fear, and she saw no one in there, and her heart squeezed, and she felt the soft breath of her body against his body.
"I'm sorry," he said, but I wasn't sure what to do.
It was a silent view of someone else and everything to be visible to make the deal at home this summer, waiting for the door to go.
Into a new life with her: leaving the house, leaving a troubled walk by a way that she'd never experienced in a long time.

Table 2: Examples of generated suggestions

writes from the first-person perspective, given that *I* is by far the most frequent pronoun (though some of these occurrences might be included in quoted dialogue). In terms of pronouns, it is more likely to refer to female than male characters, indicated by the relative frequency of *she* and *her* versus *he* and *him*. However, *man* and *father* also show up as characters. The model has an interesting preoccupation with being *sure* (or *not sure*), as well as the capacity to do something (*able, not be able, able to handle, not handle*). Apologies are common (*sorry*). References to cognitive and emotional states also appear often, indicated by the adjectives *glad, troubled*, and *afraid* as well as the verbs *want, feel* and *think*. The model makes value judgments about things being *good, right, wrong,* and *true*. Negations (*not*) are overwhelming common, particularly in the context of *not know*, *not sure*, *not want*, *not go*, *not expect*, *not think*, and *not handle*. Moreover, the model tends to hedge with the adverb *maybe* as well as *probably*. The act of transferring or coordinating locations also shows up (*go, leave, stay, come, wait, away*).

These biases are largely indicative of the corpus on which the model was trained. The corpus included several romance novels, for instance, and this appears to be reflected in many of the word choices. However, Roemmele et al. [2017] found that compared to human-written text, the RNN LM was much more constrained in the variety of words/phrases it generated. It tends to repeat itself by disproportionately generating frequently observed words. The probability distribution of the model can be manipulated to influence this diversity (via the temperature variable; see Manjavacas et al. [2017]), but it comes with the trade-off of potentially reduced coherence and grammaticality of the sentences.

As future work, we are focusing on how the linguistic features of a suggestion influence how users modify it. For example, specific lexical items may be more or less likely to be retained in the story. This analysis supports the overall goal of generating suggestions that are maximally helpful to authors for continuing a story.

References

Natalie Dehn. Story Generation After TALE-SPIN. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 16–18, 1981.

Ahmed Khalifa, Gabriella AB Barros, and Julian Togelius. Deeptingle. *International Conference for Computational Creativity 2017*, 2017.

Michael Lebowitz. Story-telling as planning and learning. *Poetics*, 14(6):483–502, 1985.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. Story Generation with Crowd-sourced Plot Graphs. In *27th AAAI Conference on Artificial Intelligence*, 2013.

Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 29–37, 2017.

Neil McIntyre and Mirella Lapata. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

James R Meehan. TALE-SPIN, An Interactive Program that Writes Stories. In *5th International Joint Conference on Artificial Intelligence*, pages 91–98, 1977.

Melissa Roemmele and Andrew S Gordon. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer International Publishing, 2015.

Melissa Roemmele, Andrew S Gordon, and Reid Swanson. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*, 2017.

Reid Swanson and Andrew S. Gordon. Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Transactions on Interactive Intelligent Systems*, 2(3): 1–35, 2012.

Scott R Turner. *MINSTREL: A computer model of creativity and storytelling*. University of California at Los Angeles, 1993.