???

Anonymous ACL submission

Abstract

People's writing style is affected by many factors, including topics, sentiment, and individual personality. In this paper we show that writing tasks that impose constraints on the writer result in the author adopting a different writing style compared to tasks that do not. As a case study, we experiment with a recently published machine reading task: the story cloze task (Mostafazadeh et al., 2016). In this task, annotators were asked to generate two sentences: one which makes sense given a previous paragraph and another which doesn't. We show that a linear classifier, which applies only simple style features, such as sentence length and character n-grams, obtains state-of-the-art results on the task, substantially higher than sophisticated deep learning models. Importantly, our model doesn't even look at the previous paragraph, just the two candidate sentences, which, out of context, differ only in the constraint put on the authors. Our results indicate that such constraints dramatically affect the way people write. They also suggest that careful attention to the instructions given to the authors needs to taken when designing new NLP tasks.

1 Introduction

Writing style is defined as the choice of words, spelling, grammar and punctuation made by the author.[1] It is often affected by inter-writer factors such as age (Schler et al., 2006), gender (Argamon et al., 2003), native language (Koppel et al., 2005), or mere personality (Stamatatos, 2009), but also by other parameters such as the sentiment of the text (Davidov et al., 2010) and whether it is sarcastic (Tsur et al., 2010). In this paper we study to what extent is writing style affected by more intricate factors, such as the type of constraints put on the author.

As a testbed, we experiment with the story cloze task (Mostafazadeh et al., 2016). This task was created by having authors write five-sentence self-contained stories. Following, a subset of the the stories was given to another group of authors, who were shown only the first four sentences of each story, and were asked to write two one-sentence endings for it: a *correct* ending, and a *wrong* ending. The goal of the task is to determine which of the endings is the correct one.

Interestingly, although originally intended to be a machine reading task, the compilation of this task raises several research questions which seem to differ from the original intent of the designers. First, do authors use different style when asked to write a *correct* sentence, compared to a *wrong* sentence? Second, do authors use different style when writing the ending as part of their own five sentence story, compared to reading four sentences, and then writing a standalone (correct) ending?

We show that the answer to both of these questions is positive. We train a linear classifier, using simple stylistic features, such as sentence length, character n-grams and PoS counts. We show that on a balanced dataset (random guess is 50%) our classifier can distinguish between *correct* and *wrong* sentences in 64.5% of the cases. Importantly, the classifier is trained **only** on the last sentences, and does not consider the four input sentences. It is also trained on a set of positive samples and a set of negative ones, rather than pairs of (positive, negative) pairs, as in the original story cloze task. Furthermore, when trained to distin-

[1] https://en.wikipedia.org/wiki/Writing_style

guish between original endings and new endings, the classifier obtains 70.9% accuracy.

In order to estimate the quality of our results, we turn back to the story cloze task. We show that using our classifier, we are able to obtain 71.6% accuracy on the task, a 12% improvement compared to the published state-of-the-art results (Speer et al., 2016).[2]

We present an ablation study which shows that the style differences are realized in syntactic features (such as the over/under use of coordination words like "and" and "but"). Furthermore, we also show that sentiment plays an important role in the writing style differences. For instance, one of the key features for distinguishing between correct and wrong sentences is the over-representation of the word "hate" by the latter.

Our results suggest that writing style is affected by the the writer's state of mind. Writing a sentence intended to be *wrong* turns out quite differently than a sentence intended to be *correct*. Similarly, writing a sentence as part of the story is different from reading a story, and then writing the final sentence. These differences can be distinguished to a large extent by simple machine learning tools.

The results presented here also provide valuable lessons for designing new NLP tasks. Although (Mostafazadeh et al., 2016) seem to have put a lot of effort into designing the task, addressing many potential methodological flaws (see Section 2), the importance of a few allegedly minor details were underestimated.

Finally, we show that our stylistic features can benefit from combining with a machine-reading model, for which this task was designed. We train an neural language model on the original five sentence training corpus, and then compute the language probability of each of the candidates answers. We add the numbers as features in our linear classifier, and get an additional 4% improvement (75.6%).

the reminder of this paper is organized as follows. In Section 2 we introduce the cloze story task. We present our model, our experiments and our results at sections 3, 4 and 5, respectively. Sections 6 and 7 present an ablation study and a discussion, while Section 8 surveys related work. We conclude at Section 9.

2 The Cloze Story Task

[3] Developed as an effort to simplify[4] the representation and learning of commonsense knowledge, the *Cloze Story Task* (Mostafazadeh et al., 2016) has become the Shared Task for the LSDSem workshop.[5] The task provides two types of datasets: the *ROC Stories* and the *Story Cloze test sets*.

ROC Stories consist of 98,163 five-sentence commonsense stories, collected on AMT[6]. Workers were instructed to write a coherent story where something happens, and with a clear beginning and end. To collect a broad spectrum of commonsense knowledge, there was no imposed subject for the stories.

Story Cloze test sets were created on AMT, using a subset of ROC Stories. Presented with the first four sentences of a story, workers were asked to write a "right" and a "wrong" ending. ~~The ending~~Both endings had to complete the story using a character in it, and when read out of context, had to be "realistic and sensible" (Mostafazadeh et al., 2016).

The resulting stories were then individually rated for coherence and meaningfulness by AMT workers. Only stories with a coherent and meaningful "right" ending and a neutral "wrong" ending were selected for the test, yielding 3,744 test stories. {*Hope the "neutral" thing makes sense, I didn't know how else to explain 4.2.2 of the Story*

[2] Recently, a shared task for the story cloze task has been published (https://competitions.codalab.org/competitions/15333). At the time of submission, the leading results is 71.1%, which is much closer to our results, although still inferior. No details about the methods used to generate this result are available.

[3] **Roy: We should start with a general paragraph reminding the reader of the problem and why we are using this corpus.**

[4] **Roy: "simplify" is not the right term here.** *Facilitate* **maybe?**

[5] **Roy: Not sure that this sentence is so terrible, but I would not put the focus on the shared task. This is a rather minor detail and has very little relevance to our story. Think about what we want to reader who has very little time or energy to know. Specifically, I would start with few very general details and motivation (as you did), but then directly move to mention the interesting design decisions they took, which actually raise very different type of research questions, which we tackle in this paper. Then you can move to describe it more formally, as you do below (you can mention the shared task there if you want, not sure it's even necessary).**

[6] **Roy: I would use "Amazon Mechanical Turk (AMT)" here, and AMT from now on**

cloze paper.$\}_{ms}$[7]
[8]

3 Model

3.1 Neural Language Model

{Below is a messy dump of all the hyper parameters... Didn't know how many details were needed}$_{ms}$[9]

We trained a simple RNN Language model (Mikolov et al., 2010) using a single-layer LSTM (Hochreiter and Schmidhuber, 1997) of hidden dimension $h = 512$. We used the ROC Stories for training, setting aside 10% for validation of the language model. We replaced all words occuring less than 3 times by a special out-of-vocabulary character, yielding a vocabulary size of $|V| = 21,582$. Only during training, we applied a dropout rate of 60% while running the LSTM over all 5 sentences of the stories. Using AdamOptimizer (Kingma and Ba, 2014) and a learning rate of $\eta = .001$, we trained with backpropagation on cross-entropy.

To construct features for the Story Cloze task, we scored each of the two endings using our neural language model. We computed the probability of each ending given the first four sentences $p_\theta(ending|story)$, as well as the probability of the endings out of context $p_\theta(ending)$. We also included the likelihood ratio $\frac{p_\theta(ending|story)}{p_\theta(ending)}$ into our classifier.[10]

4 Experiments

5 Results

6 Ablation Study

7 Discussion

8 Related Work

- Different style application (as in introduction). Also include deception works (Yejin has 1-2 papers on it).
- Machine reading papers?
- *{Pennebaker's work on function words?}*$_{ms}$

9 Conclusion

[7] **Roy: Reading 4.2.2, it seems that they were not targeting badly phrased "wrong" sentences, but just the stories as a whole, and validated that all "right" stories were valid and all "wrong" stories were invalid (unless this is how you read the -1 option in their description, which is not addressed anywhere else in the text). If I am correct, I would write this and explicitly mention that no validation was performed on the ending level.**

[8] **Roy: The following might belong in the related work section, but probably more relevant here: I would add a section about performance on this task, stating that the authors tried quite a few baselines and showed that all pretty much fail on the task, and other than that there was just one paper that got slightly better results. This is a place to make two more arguments: (a) the low results obtained by the various baseline methods, as well as the low number of works (only one) that improved results on this task, compared to other similar tasks published in recent years (SQUAD, SNLI, CNN/Daily Mail., etc.), indicate that this is task is relatively hard, even for machine reading systems that has made a huge progress in recent years. (b) To the best of our knowledge, this is the first work to consider the style differences between the two endings, rather than the setup suggested by the author.**
Btw, this could be a good place to mention the shared task.

[9] **Roy: The level of description seems fine (we might remove some of them if space is an issue, but for now it's good. What I'm missing here are (a) an introductory 1-2 sentences to what are we doing here (I realize that it is not 100% clear from the intro what is the role of these experiments other than showing that we have STOA results, so this can come later**
(b) A one-sentence description of your model. Currently, there is no clear separation between the model (RNN using LSTM) and the technical details. Some general description of the model, and specifically the motivation behind (nothing fancy, just saying that we applied state-of-the-art LM tools, which resemble the tools the authors had in mind when they designed the task).

[10] **Roy: This will probably change once we have the rest of the paper, but generally a few things that are missing: (a) performance of the model as a standalone (both how we evaluated it and how much it got). (b) modifying this paragraph to indicate that in order to combine between the style features presented in this paper and this model, we did the above. (c) results of the combination (75.1%). These numbers might best fit in a small table.**

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-* 23(3):321–346.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, pages 241–249.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, pages 624–628.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*. volume 2, page 3.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 839–849. http://www.aclweb.org/anthology/N16-1098.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*. volume 6, pages 199–205.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975* .

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*. pages 162–169.