???

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 compliation 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 a 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 state of mind the writer is at. 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 Model

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?

9 Conclusion

[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.

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.

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.

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.