???

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 the author's choice of words, spelling, grammar and punctuation.[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 its level of sarcasm (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). In this task, authors were asked to write five-sentence self-contained stories. Following, the stories were 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. Table 1 shows an example of an original story, a *correct* story and a *wrong* story.

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?

Our experiments indicate that the answer to both questions is positive. We train a linear classifier, using simple stylistic features, such as sentence length, character n-grams and PoS counts. First, we show that on a balanced dataset (random guess is 50%) our classifier distinguishes 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. Second, when trained to distinguish between original endings and new (*correct*) endings,

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

Type	Example
Original story	My mother loves clocks that chime. Her house is full of them. She sets them each a little different so she can hear them chime. It sounds like a bell tower during a wedding in her house all day. When I visit I stop them or I'd never be able to sleep at night.
Correct story	Kathy went shopping. She found a pair of great shoes. The shoes were $300. She bought the shoes. She felt buyer's remorse after the purchase.
Wrong story	Kathy went shopping. She found a pair of great shoes. The shoes were $300. She bought the shoes. Kathy hated buying shoes.

Table 1:
Examples of stories from the story cloze task. The first row shows an original story written by one author. The second and third row show two endings of the same story: a *correct* ending and a *wrong* one.

the classifier obtains 70.9% accuracy.

In order to estimate the quality of our results, we turn back to the story cloze task. Using our classifier, we are able to obtain 71.5% accuracy on the task, a 11.6% improvement compared to the published state-of-the-art results (Salle et al., 2016).[2] An ablation study shows that the style differences are realized in syntactic features (such as the over/under use of coordination words like "and" and "but"), but that sentiment also 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" in the latter.

Our results may have a wide impact on a range of fields. First, they have the potential to shed light on cognitive processes that take place in the brain during writing. Second, our results might have a more practical value in an era when fake news are becoming prominent, as they suggest that these might have different style compared to real news.

[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.

Third, the results presented here also provide valuable lessons for designing new NLP tasks, both in terms of the potential impact of even the smallest details, and the need to carefully run baseline models.

Finally, one interesting question that remains open is whether state-of-the-art machine reading tools, for which this task was designed, capture in fact the same stylistic features as our linear classifier. We show that this is not the case. We train a 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 3.6% improvement (75.1%).

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

2 The Cloze Story Task

[3] Writing tasks are plentiful on Amazon Mechanical Turk (AMT) {*cite various examples?*}$_{ms}$[4]. As a testbed for our analyses, we ~~chose~~focus on the *Story Cloze Task* (Mostafazadeh et al., 2016), during which[5] workers were asked to write intuitive and counter-intuitive endings to short stories. This task is particularly interesting in that the two endings were written by the same author, for the same story, allowing for careful comparison.[6]

Originally, the task was developed as an effort to facilitate representation and learning of commonsense knowledge. For that purpose, the creators crowdsourced the creation[7] of two types of

[3] **Roy: High level on first paragraph: much better, but still not enough connection to our story. You need to be explicit as to why we chose this task. Our work is currently not mentioned at all**

[4] **Roy: I would not start with AMT, as it is not important to our discussion. Think of the message that a first sentence in a paragraph gives: this is the topic/main focus of this paragraph. AMT is only a technical detail that should be mentioned later**

[5] **Roy: not sure "during which" is right here. consider rephrasing**

[6] **Roy: This is indeed particularly interesting for us, but currently this is out of context. I would say that this task has a few interesting properties that we exploit.**

[7] **Roy: Don't use "creators . . . creation". Regardless, *creators* is not the right word here**

datasets: the *ROC Stories* and the *Story Cloze test sets*.

ROC Stories consist of $98,163$[8] five-sentence commonsense stories, collected on AMT. Workers were instructed to write a coherent story where something happens, and with[9] a clear beginning and end. To collect a broad spectrum of commonsense knowledge, there was no imposed subject for the stories.

Story Cloze [10] 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. Both endings had to complete the story using ~~a character~~one of the characters in it, and when read out of context, had to be "realistic and sensible" (Mostafazadeh et al., 2016).

The resulting stories, matched with each ending, were then individually rated for coherence and meaningfulness by AMT workers. Only stories rated as simultaneaously coherent with "right" ending and neutral with a "wrong" ending were selected for the test, yielding $3,744$ test stories. It is worth noting that workers rated the stories as a whole, not the endings; no selection was done on the endings alone. [11]

Benchmarks *{Like you said, maybe this belongs in related work, but for now I'll put it here}$_{ms}$* The creators of the dataset implemented various benchmarks, using surface level features as well as narrative informed ones. Compared to similar tasks (**???**), the baselines perform relatively poorly, indicating that the task is a relatively hard one.

3 Model

The goal of this paper is to determine to what extent does constraining authors in their writing assignments lead to them adopting different writing styles. In order to answer these questions, we use simple, classic NLP tools, which have been shown to be very effective for recognizing style (see Section 7). We describe our model below.

We train a linear SVM classifier to distinguish between different endings. Each feature vector is computed using the words in one ending, without considering earlier parts of the story. We use the following style features.

- ***Length*** The number of words in the sentence.
- ***Word n-grams*** We use sequences of 1-5 words. Following (Tsur et al., 2010; Schwartz et al., 2013), we distinguish between high frequency and low frequency words. Specifically, we replace content words with their part-of-speech tags (Nouns, Verbs, Adjectives and Adverbs).
- ***Character n-grams*** Character n-grams are one of the most useful features in identifying author style (Stamatatos, 2009). We used character 5-grams.

4 Experiments

We design two experiments to answer our research questions. The first is an attempt to distinguish between *correct* and *wrong* endings. The second attempts to distinguish between original endings and new *correct* endings. We describe both experiments below.

Experiment 1: Correct/Wrong Endings. The goal of this experiment is to measure to what extent style features capture differences between *correct* and *wrong* endings. As the cloze story

[8] **Roy: Minor, but still: writing the number inside dollar signs adds a white space in the pdf. This is a bit annoying, and also potentially confusing, as in the next case you use it below (3,744 something)**

[9] **Roy: "and with" doesn't sound right here. rephrase**

[10] **Roy: say something general first before diving into details**

[11] **Roy: Other things to talk about (not necessarily in that order): 1. no training corpus for the task. That is, the training set (a) does not contain negative example (just positive), and (b) was not generated in the same way as the positive test samples. Here you need to explicitly say that we show that we show that the training and test sets use very different styles. 2. Maybe talk about the shared task, as well as their RepEval paper, which suggested this task as a vector-space evaluation measure? 3. Say that the authors did an extensive work in order to assure that the positive and negative samples are not easily distinguishable: for each pair, both endings were written by the same author (you addressed this earlier, but this seems like a better location), they tried an extensive set of baselines, showing all don't surpass a random baseline by much. Maybe now you can briefly mention the benchmarks, saying that several baselines were presented in the papers. All benchmarks attempted to learn the connection between the previous paragraph and the ending, and all performed near chance level.. However, no baseline looked at the endings independently. Then you can add the benchmark text below, and conclude that this indicates that this task is well designed and hard and the one hand, but might also point to potential disagreements between the train and the test data, which we highlight in this paper.**

task doesn't have a training corpus for the *correct* and *wrong* endings, we use the development set as our training set. We split it (90/10) into our training and development set. We keep the story cloze test set as is. The final size of our training/development/test sizes are 3366/374/3742, respectively.

We take *correct* endings as positive samples and *wrong* endings as our negative examples. By ignoring the coupling between *correct/wrong* pairs, we are able make a more general claim about the style used when writing each of the tasks.

We add a special START symbol at the beginning of each sentence. For computing our features, we keep n-gram (character or word) features that occur at least five times in the training set.[12] For the PoS features, we tag each sentence with the Spacy PoS tagger.[13] We use sklearn's LinearSVC SVM implementation[14] with L2 loss. We grid search the regularization parameter on our development set.

Experiment 2: Original/New Endings. Here the goal is to measure whether writing the ending as part of a story imposes different style compared to writing a (*correct*) ending to an existing story. We use the endings of the original ROC stories as our *original* training samples and *correct* endings from the ROC stories development set as *new* training samples. As there are far more *original* samples than *new* ones, we randomly select N *original* samples, where N is the number of *new* samples, such that we have balanced labels in our training, validation and test sets. We apply the same experimental decisions as in Experiment 1, and repeat this process 5 times while reporting the average.

Results. Figure 1 shows our results. Results show that for Experiment 1, our model obtains results well above a random baseline – 64.5% classification accuracy. For Experiment 2, we get even higher results – 70.9%. These numbers indicate that these different writing tasks clearly impose a different writing style on authors.

As a complementary experiment, we measured whether these style differences are additive. That is, whether the style differences between *correct* endings and *wrong* ones are different from the differences between *correct* and *original* endings. We repeated Experiment 2, this time comparing between *original* and *wrong* sentences. Our hypothesis is that differences would be even clearer. Our results (Figure 1) show that this is indeed the case: the classifier's accuracy jumps to 78%.

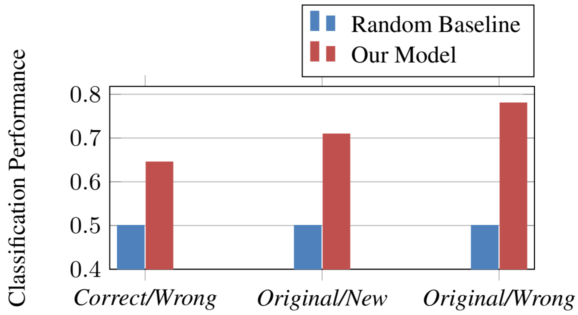

Figure 1: Results of experiments 1 and 2 (left two charts). Rightmost experiment shows a control experiment which classifies *original* endings vs ***wrong*** endings.

Story Cloze. Results of Experiment 1 indicate that *correct* and *wrong* endings are characterized by different styles. In order to estimate the quality of our classification results, we tackle the story cloze task using our classifier. This classification task is much more constrained than Experiment 1, as here the task is given two endings, which one is *correct* and which is *wrong*. In order to solve the task, we use the output of our trained classifier. For each pair of endings e_1, e_2, if both endings have different classification labels, we keep these labels. If they share the same label, we use the classifier confidence level as a deciding factor: the label of the one with the lower confidence is reversed.

Table 2 shows our results on the story cloze test set. Our classifier obtains 71.5% accuracy, which is 11.6% better than the published state-of-the-art result on the task. Importantly, unlike previous approaches, our classifier does not require the story corpus training data, and in fact doesn't even look at the first four sentences of the story in question. These numbers indicate that the styles between *correct* and *wrong* are indeed very different.

Other than the published works that tackled this task, a few recent works published their results in the LSDSem shared task website.[15] As to the time

[12] Virtually all sentences end with a period, so no need for an END token
[13] spacy.io/
[14] scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
[15] https://competitions.codalab.org/competitions/15333

Model	Results
DSSM (Mostafazadeh et al., 2016)	0.585
LexVec (Salle et al., 2016)	0.599
RNN	0.681
Our Model	**0.715**
Combined (Our model, RNN)	**0.751**
Niko (shared task)	0.7
tbmihaylov (shared task)	0.711
Human judgment	1

Table 2:
Results on the test set of the cloze story task. Upper part are published results (Lexvec results are taken from (Speer et al., 2016)). Bottom part are taken from the cloze task shared task webpage.

Feature Type	Result
Word n-grams	0.646
Char n-grams	0.699
Full Model	0.715

Table 3:
Results on Experiment 1 with different types of features.

of submission, our results are still state-of-the-art (though by a much smaller margin). No other information other than the name of the group and their results is available.

5 Ablation Study

Most Discriminative Feature Types. A natural question that follows this study is which features are most helpful in making predictions about writing style. To answer this question, we re-ran Experiment 1 with different sub-groups of features. Table 3 shows our results. Results clearly show that character n-grams are the most effective style predictors, reaching within less than 2% of the full model. These findings are in line with previous works that used character n-grams along with other types of features to predict writing style (Schwartz et al., 2013).

Most Salient Features. In order to understand which features were most salient, we repeated Experiment 1, this time running SVM with L1 norm, such that it generates a sparse separating hyperplane. This came at a minor cost of less than 2% in performance compared to our L2 results (0.697 compared to 0.715). Table 4 shows the 5 features

Positive	Negative
' time'	'ed of'
' bett'	'and d'
'found'	'threw'
'but'	'ever '
'ally '	' hate'

Table 4:
The top 5 most discriminative features for predicting *correct* (Positive) and *wrong* (Negative) endings.

with the highest positive and negative weights in the hyperplane. These correspond to the 5 most salient features for *correct* and *wrong* endings, respectively.

The table shows a few interesting trends. First, some syntactic features play become salient when authors are asked to write *wrong* vs. *correct* endings. For instance, *correct* endings are more likely to the coordinator "but", while *wrong* ones would prefer to use "and".

More interestingly, *correct* endings are likely to use positive words (e.g., " bett", which in all cases stands for " bett**er**"), while *wrong* endings would use negative words (e.g., " hate"). The idea that sentiment would play a role in this task was suggested in the original story cloze paper, where two sentiment-based baselines were evaluated. However, these baselines measured the relative sentiment between the ending and the previous sentences, and did not test whether there is a general tendency of the *wrong* endings to have a negative sentiment, and the *correct* one to have a positive sentiment. Indeed, the performance of both these baselines was roughly chance-level.

add control experiments

Neural Language Model. *{This needs to be placed somewhere else LOL}$_{ms}$* *{Below is a messy dump of all the hyper parameters... Didn't know how many details were needed}$_{ms}$*

To leverage the genre of these short stories,[16] we train a recurrent language model (RLM) (Mikolov et al., 2010),[17] which are used to achieve

[16] **Roy: Not sure this is the best introductory sentence, but you can leave it to me if you can't find anything better**

[17] **Roy: don't start with the details, but with the high level. We wanted wether our model can be combined with state-of-the-art text comprehension models, for which this task was designed. Specifically, we train an RLM ...**

state-of-the-art performance on various tasks (e.g. machine translation)[18]. Training this RLM on the ROC stories, we harness the unsupervised, single-ended stories[19], which boosts performance on the Story Cloze task.[20]

We trained an RLM 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 occurring 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.

On its own, our neural language model performs moderately on the Story Cloze test. As hinted at by the creators of this task, simple language models aren't enough to do well. In fact, using our neural LM and selecting endings based on $p_\theta(ending|story)$, we obtain only 55% accuracy. When using the likelihood ratio $\frac{p_\theta(ending|story)}{p_\theta(ending)}$ to select endings, we obtain 68%.[21] This doesn't beat reported SOTA[22], which is why we included these features into our stylistic driven model.[23][24]

6 Discussion

7 Related Work

- Different style application (as in introduction). Also include deception works (Yejin has 1-2 papers on it).

- Machine reading papers?

[18] Roy: citation required

[19] Roy: I would frame it differently. Unlike the model suggested in this paper, which only considers the endings, this model harnesses the training set as intended by the authors, which is composed of unsupervised, single-ended stories

[20] Roy: In what sense does it boost performance? we get STOA results without it...

[21] Roy: (see Table 2). also, its 68.1, isn't it? (just to be consistent with the table)

[22] Roy: SOTA is not a standard abbreviation. Use state-of-the-art

[23] Roy: no, this is not why we include these as features (at least not in this paper:). It's because we want to see whether the different approaches are complementary.

[24] Roy: Now you need to discuss our results and what they mean. You can point to Table 2

- *{Pennebaker's work on function words? they link function words to psychological aspects}*$_{ms}$

8 Conclusion

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.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016. Enhancing the lexvec distributed word representation model using positional contexts and external memory. *arXiv preprint arXiv:1606.01283* .

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.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1880–1891. http://www.aclweb.org/anthology/D13-1193.

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.