The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task

Anonymous EMNLP submission

Abstract

A writer's style depends not just on personal traits but also on her intent and mental state. In this paper, we show how variants of the same writing task can lead to measurable differences in writing style. We present a case study based on the *story cloze task* (Mostafazadeh et al., 2016a), where annotators were assigned similar writing tasks with different constraints: (1) writing an entire story, (2) adding a story ending for a given story context, and (3) adding an incoherent ending to a story. We show that a linear classifier informed by stylistic features is able to successfully distinguish among the three cases, without even looking at the story context. In addition, our style-based classifier performs on a par with state-of-the-art models on the story cloze task. Our results demonstrate that different task framings can dramatically affect the way people write.

1 Introduction

Writing style is expressed through a range of linguistic elements such as words, sentence structure, and rhetorical devices. It is influenced by personal factors such as age and gender (Schler et al., 2006), by personality traits such as agreeableness and openness (Ireland and Mehl, 2014), as well as by mental states like sentiment (Davidov et al., 2010) and deception (Feng et al., 2012). In this paper, we study the extent to which writing style is affected by the nature of the writing task the writer was asked to perform, as different tasks likely engage different cognitive processes (Campbell and Pennebaker, 2003; Banerjee et al., 2014).[1]

We show that similar writing tasks with different constraints on the author can lead to measurable differences in her writing style. As a case study, we present experiments based on a recently introduced commonsense challenge—the story cloze task (Mostafazadeh et al., 2016a). In this task, a system is presented with two short stories that differ only in their last sentence: one ending makes the entire story *coherent* while the other makes it *incoherent*. The system then decides which is the coherent one.

Interestingly, the story cloze task annotation process introduced three different variants of the task of writing an ending to a short story (Section 2). In this paper, we show that a linear classifier informed with stylistic features can distinguish among these endings to a large degree, even without looking at the rest of the story.

Our results allow us to make a few key observations. First, people adopt a different writing style when asked to write coherent vs. incoherent story endings. Second, people change their writing style when writing the entire story on their own compared to writing only the final sentence for a given story context written by someone else.

In order to further validate our method, we directly tackle the story cloze task. Adapting our classifier to the task, we obtain 72.4% accuracy, only 2% below state-of-the-art results. We also show that the style differences captured by our model can be combined with neural language models to make a better use of the story context. Our final model that combines context with stylistic features achieves an additional 2.8% gain—75.2%—a new state-of-the-art result.

The contributions of our study are threefold. First, findings from our study can potentially shed light on how different kinds of cognitive load influence the style of written language. Second, our results indicate that when designing new NLP

[1] For the purposes of this paper, *style* is defined as content-agnostic writing characteristics, e.g., sentence length.

Story Prefix	Ending
John liked a girl at his work. He tried to get her attention by acting silly. She told him to grow up. John confesses he was trying to make her like him more.	She feels flattered and asks John on a date.
	The girl found this charming, and gave him a second chance.
	John was happy about being rejected.

Table 1: Examples of stories from the story cloze task. The table shows a story prefix with three contrastive endings: The original ending, a coherent ending and a incoherent one.

tasks, special attention needs to be payed to the instructions given to authors. Third, we establish a new state-of-the-art result on the story cloze task.

2 Background: The Story Cloze Task

To understand how different writing tasks affect writing style, we focus on the *story cloze task* (Mostafazadeh et al., 2016a). While this task was developed to facilitate representation and learning of commonsense story understanding, its design included a few key choices which make it ideal for our study. We describe the task below.

ROC stories. The ROC story corpus consists of 49,255 five-sentence stories, collected on Amazon Mechanical Turk (AMT). Workers were instructed to write a coherent self-contained story, which has a clear beginning and end. No subject was imposed for the stories, which resulted in a wide range of different topics.

Story cloze task. After compiling the story corpus, the *story cloze task*—a task based on the corpus—was introduced. A subset of the stories was selected, and only the first four sentences of each story were presented to AMT workers. Workers were asked to write a pair of new story endings for each story context: one *right*, making the entire story coherent, and one *wrong*, making it incoherent. Importantly, both endings were required to be “realistic and sensible” (Mostafazadeh et al., 2016a) when read out of context (see Table 1).

Based on the new stories, the *story cloze task* was proposed: given a pair of stories that differ only in their endings, the system decides which is *right* and which is *wrong*. The training data contains the original stories (without alternative endings), while development and test data consist of the revised stories with alternative endings (for a set of original stories not included in the training set). The task was suggested as a commonsense understanding task and as a testbed for vector-space evaluation (Mostafazadeh et al., 2016b).

Interestingly, until very recently, one year after the task was first introduced, the published benchmark on this task was still below 60%. This comes in contrast to other recent similar machine reading tasks such as CNN/DailyMail (Hermann et al., 2015), SNLI (Bowman et al., 2015) and SQuAD (Rajpurkar et al., 2016), for which results improved dramatically over much shorter periods of time. This suggests that this task is challenging and that high performance is hard to achieve. In this paper, we show that this is not necessarily the case, by demostrating that a simple linear classifier informed with style features reaches near state-of-the-art results for the task—72.4%.

Different writing tasks in the story cloze task. Several design decisions make the task an interesting testbed for the purpose of this study. First, the training set for the task (ROC Stories corpus) is not a training set in the usual sense,[2] as it contains only positive (*right*) samples, and not negative (*wrong*) ones. On top of that, the *original* endings, which serve as positive training samples, were generated differently from the *right* samples, which serve as the positive samples in the development and test sets. While the former are part of a coherent story written by the same author, the latter were generated by asking an author to read four sentences, and then generate a fifth *right* ending.

Finally, the tasks for generating the pairs of endings were quite different from each other: in one case, the author was asked to write a *right* ending, which would create a coherent five-sentence story along with the other four sentences. In the other case, the author was asked to write a *wrong* ending, which would result in an incoherent story.

3 Model

To what extent do different writing constraints lead authors to adopt different writing styles? In order to answer this question, we first use simple methods that do not capture content, and have been shown to be effective for recognizing style.

We train a logistic regression classifier to categorize an ending, either as *right* vs. *wrong* or as *original* vs. new (*right*). Each feature vector is

[2] I.e., the training instances are drawn from a different population than the one testing instances will be drawn from.

computed using the words in one ending, without considering earlier parts of the story. We use the following style features: (a) ***Length:*** the number of words in the sentence; (b) ***Word n-grams:*** sequences of 1-5 words;[3] and (c) ***Character n-grams:*** sequences of 4 characters.[4]

4 Experiments

We design two experiments to answer our research questions. The first is an attempt to distinguish between *right* and *wrong* endings, the second between *original* endings and new (*right*) endings.[5]

Exp. 1: right/wrong endings. We measure the extent to which style features capture differences between the *right* and *wrong* endings. Our classification task is slightly different from the story cloze task. Instead of classifying pairs of endings, our classifier decides about each ending individually, whether it is *right* (positive instance) or *wrong* (negative instance). By ignoring the coupling between *right* and *wrong* pairs, we decrease the impact of author-specific style differences, and focus on the difference between the styles accompanied with *right* and *wrong* writings.

Exp. 2: original/new endings. We measure whether writing the ending as part of a story imposes different style compared to writing a new (*right*) ending to an existing story. We use endings of the ROC stories (*original* samples) and *right* endings from the story cloze task (*new* samples).

Results. In both experiments, our model performs well above what would be expected under chance (50% by design): 64.5% accuracy for Exp. 1, and 68.5% for Exp. 2. Noting again that our model ignores the story context (the preceding four sentences), our model is unable to capture any notion of coherence. This finding provides strong evidence that the authors' style was affected by the writing task they performed.

Further Analysis. Table 2 shows the most salient features for *right* and *wrong* endings in Exp. 1. *Right* endings tend to end with past tense verbs, while *wrong* endings with common nouns. Further, *right* endings make wider use of adjectives, which might indicate that incoherent writing inspires authors to write less descriptive text, as is the case in deceptive text (Ott et al., 2011).

Right	Wt.	Freq.	***Wrong***	Wt.	Freq.
'ed .'	0.17	6.5%	START NNP	0.21	54.8%
'and '	0.15	13.6%	NN .	0.17	47.5%
JJ	0.14	45.8%	NN NN .	0.15	5.1%

Table 2: The top 3 most heavily weighted features for predicting *right* vs. *wrong* endings, along with their weights (Wt.) and their frequencies in our story cloze training set (Freq.).

Model	**Acc.**
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
cogcomp	0.744
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Table 3: Results on the story cloze task. The middle block are our results. *cogcomp* results and human scores are taken from (Mostafazadeh et al., 2017). Methods marked with (†) do not use the story context to make a prediction.

Story cloze task. To further estimate the quality of our results, we tackle the story cloze task. We apply the classifier from Exp. 1 as follows: if it assigns different labels to the two given endings, we keep them. Otherwise, the label whose posterior probability is lower is reversed. This task is easier than Exp. 1, as two endings are given and the question is which is *right* and which is *wrong*.

Table 3 shows our results on the story cloze test set. Our classifier obtains 72.4% accuracy, only 2% lower than state-of-the-art results, and equal or higher to all other published results. Importantly, unlike previous approaches, our classifier does not require the story corpus training data, and doesn't even consider the story context in question. These numbers further support the claim that the styles of *right* and *wrong* endings are indeed very different.

Combination with a neural language model. We investigate whether our model can benefit from state-of-the-art text comprehension models. We experiment with an LSTM (Hochreiter and Schmidhuber, 1997) recurrent neural network language model (LM; Mikolov et al., 2010). Unlike the model in this paper, this LM harnesses the story cloze training set, which consists of single-

[3] In order to focus on style, we replace content words (nouns, verbs, adjectives, and adverbs) with their POS tags.

[4] Character 4-grams are very common features in style detection (Stamatatos, 2009).

[5] Experimental details are found in the supplementary file.

ending stories, as well as the story context for each pair of endings.[6] We show that adding our features to this powerful LM gives improvements over our classifier as well as the LM. To apply this LM to the classification problem, we select as *right* the ending with the higher value of

$$\frac{p_\theta(\text{ending} \mid \text{story})}{p_\theta(\text{ending})} \tag{1}$$

Selecting endings based on the ratio in Equation 1 achieves 67.7% on the story cloze task.

We combine our model with the LM by adding three features to our classifier: the numerator, denominator, and ratio in Equation 1, all in log space. We retrain our model with the new feature set, and gain 2.8% absolute. Our final result—75.2%—is a new-state-of-the-art. This result indicates that context-ignorant style features can be used to obtain high accuracy on the task, adding value even when context and a large training dataset are used.

5 Discussion

The effect of writing tasks on mental states. Writing tasks can affect mental states. For instance, writing deceptive text leads to a cognitive burden accompanied by a writing style that is different from truthful text (Newman et al., 2003; Banerjee et al., 2014). Writing tasks can even have a long-term effect, as writing emotional texts was observed to benefit both physical and mental health (Lepore and Smyth, 2002; Frattaroli, 2006). Campbell and Pennebaker (2003) also showed that the health benefits of writing emotional text are accompanied by changes in writing style.

Another line of work has shown that writing style is affected by mental state. First, an author's personality traits (e.g., depression, neuroticism, narcissism) affect her writing style (Schwartz et al., 2013; Ireland and Mehl, 2014). Second, temporary changes, such as a romantic relationship (Ireland et al., 2011; Bowen et al., 2016), work collaboration (Tausczik, 2009; Gonzales et al., 2009), or negotiation (Ireland and Henderson, 2014) may also affect writing style.

This large body of work indicates a tight connection between writing tasks, mental states, and writing style. This connection hints that the link discovered in this paper, between different writing tasks and resulting variation in writing style, involves differences in mental state. Further investigation is required to validate this hypothesis.

Design of NLP tasks. Our study provides important insights for the future design of NLP tasks. The story cloze task was very carefully designed, controlling for factors such as topic diversity and temporal and causal relation diversity. Nonetheless, several significant style differences can be found between the training and the test set, as well as between the positive and negative labels.

Our findings suggest that careful attention must be paid to instructions given to authors, especially in unnatural tasks such as writing a *wrong* ending. The COPA dataset (Roemmele et al., 2011), which was also designed to test commonsense knowledge, explicitly addressed potential style differences in its annotation instructions. In this task, systems are presented with premises like *I put my plate in the sink*, and then decide between two alternatives, e.g.: (a) *I finished eating* and (b) *I skipped dinner*. Importantly, when writing the alternatives, annotators were asked to be as brief as possible and avoid proper names, as well as slang.

Applying our story cloze classifier to this dataset yields 53.2% classification accuracy—higher than a random baseline, but substantially lower than our story cloze task results. While this is partially explained by the smaller data size of the COPA dataset (1,000 samples compared to 3,742 in the story cloze task), our findings also indicate that simple instructions may help alleviate the effects of writing style found in this paper.

6 Conclusion

Different writing tasks assigned to an author result in different writing styles for that author. We experimented with the story cloze task, which introduces two interesting comparison points: the difference between writing a story on one's own and continuing someone else's story, and the difference between writing a coherent and an incoherent story ending. In both cases, a simple model reveals measurable differences in writing styles, which in turn allows our final model to achieve state-of-the-art results on the story cloze task.

The findings presented in this paper have cognitive implications, as they motivate further research on the effects that a writing prompt has on an author's mental state, and also her concrete response. They also provide valuable lessons for designing new NLP datasets.

[6] The LM training details are in the supplementary file.

References

Ritwik Banerjee, Song Feng, Jun S. Kang, and Yejin Choi. 2014. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proc. of EMNLP*.

Jeffrey D. Bowen, Lauren A. Winczewski, and Nancy L. Collins. 2016. Language style matching in romantic partners? conflict and support interactions. *Journal of Language and Social Psychology* 0(0):1–24.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Michael Bugert, Yevgeniy Puzikov, Andreas Rckl, Judith Eckle-Kohler, Teresa Martin, Eugenio Martnez-Cmara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych. 2017. Exploring data generation methods for the story cloze test. In *Proc. of LSDSem*.

R. Sherlock Campbell and James W. Pennebaker. 2003. The secret life of pronouns flexibility in writing style and physical health. *Psychological Science* 14(1):60–65.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. of COLING*.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proc. of ACL*.

Joanne Frattaroli. 2006. Experimental disclosure and its moderators: a meta-analysis. *Psychological bulletin* 132(6):823.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2009. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37(1):3–19.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NIPS*.

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

Molly E. Ireland and Marlone D. Henderson. 2014. Language style matching, engagement, and impasse in negotiations. *Negotiation and Conflict Management Research* 7(1):1–16.

Molly E. Ireland and Matthias R. Mehl. 2014. *Natural language use as a marker of personality*, Oxford University Press, USA, pages 201–237. The Oxford Handbook of Language and Social Psychology.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science* 22(1):39–44.

Stephen J. Lepore and Joshua M. Smyth. 2002. *The writing cure: How expressive writing promotes health and emotional well-being*. American Psychological Association.

Todor Mihaylov and Anette Frank. 2017. Simple story ending selection baselines. In *Proc. of LSDSem*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. of Interspeech*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proc. of NAACL*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F. Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proc. of LSDSem*.

Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. 2016b. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *Proc. of RepEval*.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proc. of ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Andrew H. Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Unger. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

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

Yla Rebecca Tausczik. 2009. *Linguistic analysis of workplace computer-mediated communication*. Master's thesis, University of Texas.