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

Anonymous ACL submission

Abstract

A writer's style depends not just on personal traits but also on her intent and cognitive state. In this paper, we show how writing tasks involving different cognitive processes 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 on their own vs. adding only a story ending for a given story context and (2) writing a story ending that makes the overall story coherent or incoherent.[MANY FELT THESE 2 CONTRAINTS LOOKS LIKE 3 CONSTRAINTS. CLARIFY –CLINIC] We show that a simple linear classifier informed with stylistic features obtains state-of-the-art results on the story cloze challenge, substantially higher than deep learning models, even without looking at the story context. Our results demonstrate that different task framings can dramatically affect the way people write.[SIMILARLY TO PREVIOUS COMMENT: WHICH RESULTS DOES THIS COMMENT ADDRESS? –CLINIC]

1 Introduction

Writing style is expressed through a range of linguistic elements such as words, sentence structure, and rhetorical devices. It is influenced by both personal factors such as age (Schler et al., 2006), gender (Argamon et al., 2003), and personality (Stamatatos, 2009), as well as the cognitive states[AVOID USING "COGNITIVE"? –CLINIC] of the authors such as sentiment (Davidov et al., 2010), sarcasm (Tsur 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, since different tasks likely engage different cognitive processes (Banerjee et al., 2014).

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.
Coherent 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.
Incoherent 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 (Mostafazadeh et al., 2016a). The first row shows an Original story written by one author. The second and third row show revised stories with two contrastive endings: a Coherent ending and a Incoherent one.[THEY SAID IT WOULD BE GREAT TO HAVE THREE ENDINGS TO THE SAME STORY. UNFORTUNATELY NASRIN ET AL. DIDN'T PUBLISH THE ORIGINAL ENDINGS FOR THE DEV/TEST DATA. WE COULD ASK HER, BUT NOT SURE IF THIS IS WORTH IT –CLINIC]

As a case study, we present experiments based on the recently introduced ROC story cloze task (Mostafazadeh et al., 2016a). In this task, crowd-

sourced authors were asked to write five-sentence self-contained stories, henceforth *original* stories. Then, each original story was given to a different author, who was shown only the first four sentences as a story context, and asked to write contrastive story endings: a *right* (coherent) ending, and a *wrong* (incoherent) ending. Framed as a story cloze task, the goal of this dataset was to serve as a commonsense challenge for NLP and AI research. Table 1 shows an example of an *original* story, a *coherent* story, and an *incoherent* story.

While originally designed to be a story understanding challenge, the annotation process triggers interesting questions that may go beyond the original intent of the designers. First, do people maintain the same writing style when asked to write both coherent and incoherent story endings? Second, do people maintain the same writing style[WHAT IS STYLE? (LUKE) –CLINIC] when writing the entire story on their own compared to writing only the final sentence for a given story context written by someone else?

Our analysis indicates that different framings of similar writing tasks can lead to measurable differences in people's writing style.[THIS IS THE MAIN MESSAGE – SHOULD COME EARLIER – CLINIC] Using a simple classifier informed with stylistic features, we show that our classifier can distinguish the *right* and *wrong* endings with 64.5% accuracy, even without looking at the story context, where chance performance is 50%. Further, our classifier can also distinguish between the *original* endings and the new (*right*) endings with 68.5% accuracy, again without looking at the story context. [WHY DO WE CARE? (LUKE) – CLINIC]

In order to further estimate the quality of our results, we also directly tackle the story cloze challenge. Adapting our classifier to the task, we obtain 72.4% accuracy, a 12.5% increase over the previously reported state of the art (Salle et al., 2016).[TOO MANY NUMBERS IN THESE PARAGRAPHS –CLINIC] We analyze the results of our classifier, and show that when distinguishing between *right* and *wrong* endings, sentiment plays a significant role: authors writing the wrong endings often engage negative sentiment.[BETTER WORDING? WE KEEP IT DEPENDING ON WHETHER WE KEEP OUR METHODOLOGY –RS][MORE DETAILS REQUIRED? –CLINIC]

Finally, we 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 75.2%—an additional 2.8% gain, 15.3% better than the best published 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 provide valuable insights for designing new NLP tasks,[UNCLEAR –CLINIC] both in terms of the potential impact of even the smallest details, and the need to carefully run baseline models. Third, we establish a new state-of-the-art result on the commonsense story cloze challenge.

The remainder of this paper is organized as follows. In Section 2 we describe the story cloze task. We then present our model, experiments and results in Sections 4, 5 and 6 respectively. Sections 7 and 8 present a further analysis of our results and a discussion, followed by related work and conclusions.[OMIT THIS PARAGRAPH? –CLINIC]

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, it introduces a few interesting properties[WHICH PROPERTIES? –CLINIC] which make it ideal for our study. We describe the task below.

ROC Stories. The ROC Story Corpus consists of 49,255 five-sentence commonsense stories, collected on Amazon Mechanical Turk (AMT).[1] Workers were instructed to write a coherent self-contained story, which has a clear beginning and end. To collect a broad spectrum of commonsense knowledge, there was no imposed subject 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

[1] Recently, an additional 53K stories were released, which results in roughly 100K stories.

endings for each story context: one *right* and one *wrong*. Both endings are required to complete the story using one of the characters in the story context. Additionally, the ending is required to be "realistic and sensible" (Mostafazadeh et al., 2016a) when read out of context.

The resulting stories, both *right* and *wrong*, were then individually rated for coherence and meaningfulness by additional AMT workers. Only stories rated as simultaneously coherent with a *right* ending and neutral with a *wrong* ending were selected for the task. It is worth noting that workers rated the stories as a whole, not only the endings.

Based on the new stories, Mostafazadeh et al. (2016a) proposed the *story cloze task*. The task is simple: given a pair of stories that differ only in their endings, the system decides which ending is *right* and which is *wrong*. The official training data contains only the original stories (without alternative endings), while development and test data consist only of the revised stories with alternative endings (for a different set of original stories that are included in the training set). The task was suggested as an extensive evaluation framework: as a commonsense story understanding task, as the shared task for the Linking Models of Lexical, Sentential and Discourse-level Semantics workshop (LSDSem 2017), and as a testbed for vector-space evaluation (Mostafazadeh et al., 2016b).

A Hard Task. [WAS "STATE-OF-THE-ART RESULTS". LUKE MENTIONED THAT THIS PUTS THE FOCUS ON OUR RESULTS RATHER THAN ON THE NATURE OF THE TASK, AS WE INTENDED –RS] Interestingly, at the time of this submission, 10 months after the task was first introduced, the published benchmark on this task is still below 60% (Salle et al., 2016).[2] This comes in contrast to other recent similar machine reading tasks such as CNN/DailyMail (Hermann et al., 2015), SQuAD (Rajpurkar et al., 2016) and SNLI (Bowman et al., 2015), for which results improved dramatically over a similar or shorter period of time. This suggests that this task is challenging and that high performance is hard to achieve.

Different Writing Tasks in the Story Cloze Task. Mostafazadeh et al. (2016a) made substantial efforts to ensure the quality of this dataset. First, each pair of endings was written by the same author, which ensured that author style differences could not be used to solve the task. Furthermore, the authors implemented nine baselines for the task, using surface level features as well as narrative-informed ones, and showed that each of them reached roughly chance-level. These results indicate that real understanding of text is required in order to solve the task.

Several key design decisions make this task an interesting testbed for our purpose.[WHICH PURPOSE? –CLINIC] First, the training set for the task (ROC Stories corpus) is not a training sample in the usual sense,[3] 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 single coherent story written by the same author, the latter were generated by letting an author read four sentences, and then asking her to generate a fifth *right* ending. This raises the question of whether these different writing setups impose different writing styles.

Finally, although the *right* and *wrong* sentences were generated by the same author, the tasks for generating them were quite different: 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 five-sentence story.

3 Surface Analysis.

We computed several characteristics of the three types of endings: *original* endings (from the ROC Story Corpus training set), *right* endings and *wrong* endings (both from the story cloze task development set). Our analysis reveals several style differences between different groups. First, *original* endings are on average longer (11 words per sentence) than *right* endings (8.75 words), which are in turn slightly longer than *wrong* ones (8.47 words). This is inline with findings by **??**, which found that sentence length was indicative of

[2] The LSDSem 2017 shared task website (https://competitions.codalab.org/competitions/15333) does report higher results (71.1%), which are still unpublished along with the underlying methodology.

[3] I.e., the training instances are drawn from a population similar to the one that future testing instances will be drawn from.

whether a text was deceptive. Furthermore, **?** suggests lying increases one's cognitive load, which in turn could affect sentence length. Second, Figure 1a shows the distribution of five frequent POS tags in all three groups. The figure shows that both *original* and *right* endings use pronouns more frequently than *wrong* endings, which in turn prefer proper nouns, especially compared to *original* endings. {*Is there literature that may have shown that (1) pronouns correlate with coherent text, and/or (2) referencing characters by proper nouns shows a way of cognitive distancing...?*}$_{yc}$ Finally, Figure 1b presents the distribution of five frequent words across the different groups. The figure shows that *original* endings use coordinations ("and") more than *right* endings, and substantially more than *wrong* ones. {*(Newman et al., 2003) found that liars use less "contrast" coordination, is there anything there?*}$_{ms}$. Furthermore, *original* and *right* endings seem to prefer positive words (e.g., "better"), while *wrong* endings prefer negative ones ("hates"). This is in line with previous findings that dececptively written text expresses more negative emotion compared to truthful text (Newman et al., 2003). Next we show that these style differences are not anecdotal, but can be used to distinguish between the different groups of text.

4 Model

One goal of this paper is to determine the extent to which different writing constraints lead the authors to adopt different writing styles. In order to answer these questions, we use simple methods that have been shown to be very effective for recognizing style (see Section 9). We describe our model below.

We train a logistic regression 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) and Schwartz et al. (2013b), we distinguish between high frequency and low frequency words. Specifically, we replace content words (nouns, verbs, adjectives, and adverbs), which are often low frequency, with their part-of-speech tags.
- ***Character n-grams.*** Character n-grams are one of the most useful features in identifying author style (Stamatatos, 2009). We use character 4-grams.

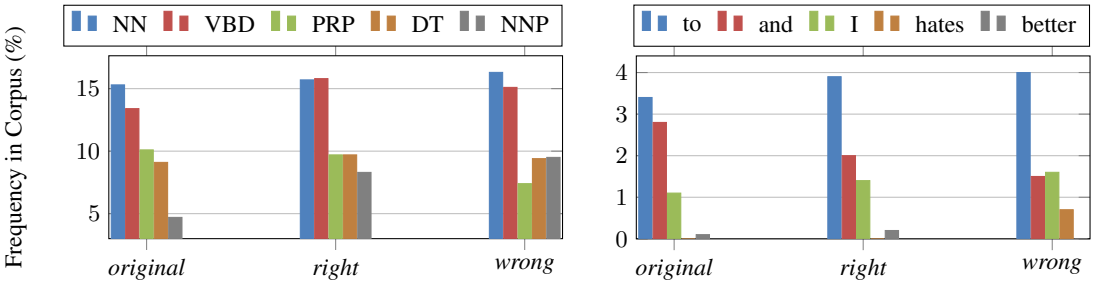
5 Experiments

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. We describe both experiments below.

Experiment 1: right/wrong endings. The goal of this experiment is to measure to what extent style features capture differences between the *right* and *wrong* endings. As the story cloze task doesn't have a training corpus for the *right* and *wrong* endings (see Section 2), we use the development set as our training set, holding out 10% for development (3,366 training endings, 374 for development). We keep the story cloze test set as is (3,742 endings).

It is worth noting that our classification task is slightly different from the story cloze task. Instead of classifying pairs of endings, one which is *right* and another which is *wrong*, 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 are able to decrease the impact of author-specific style differences, and focus on the difference between the styles accompanied with *right* and *wrong* writing. [BETTER NOW? – RS]

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 new (*right*) ending to an existing story. We use the endings of the ROC stories as our *original* samples and *right* endings from the story cloze task as *new* samples. As there are far more *original* instances than *new* instances, we randomly select the same number of *original* instances as we have *new* instances (3,366 training endings, 374 development endings, and 3,742 test endings). We randomly sample 5 *original* sets and repeat the classification experiments. We report the average classification result.



(a) POS tags

(b) Words

Figure 1: The distribution of five frequent POS tags (1a) and words (1b) across *original* endings (story cloze training set), and *right* and *wrong* endings (from the story cloze task).

Experimental setup. In both experiments, we add a START symbol at the beginning of each sentence.[4] For computing our features, we keep n-gram (character or word) features that occur at least five times in the training set. All feature values are normalized to $[0, 1]$. For the POS features, we tag all endings with the Spacy POS tagger.[5] We use Python's sklearn logistic regression implementation with L_2 regularization, performing grid search on the development set to tune a single hyperparameter – the regularization parameter.

6 Results

Figure 2 shows our results. In Experiment 1, our model obtains 64.5% classification accuracy, well above a (50%-accurate) random baseline.[HOW GOOD IS THIS? –CLINIC] For Experiment 2, our model is even stronger, at 68.5%. These results indicate that an author's style is affected in an easily-detected way when she is prompted to write (1) a *wrong* story ending vs. a *right* one, or (2) finishing her own short story vs. someone else's.

We further measured whether these style effects are additive, by classifying *original* vs. *wrong* endings. The setup is exactly as in Experiment 2, but using *wrong* endings instead of *right* ones. The result is that the effects are somewhat additive: this third classifier achieves 75.2%.

Story Cloze Task. The results of Experiment 1 indicate that *right* and *wrong* endings are characterized by different styles. In order to further estimate the quality of our classification results, we tackle the story cloze task using our classifier. This classification task is more constrained than Experiment 1, as two endings are given and question is which is *right* and which is *wrong*. We apply the classifier from Experiment 1 as follows: if it assigns different labels to the two given endings, we keep them. If not, the label whose posterior probability is lower is reversed.

Experiment	Accuracy
Right/Wrong	64.5%
Original/Right	68.5%
Original/Wrong	75.6%

Table 2: Results of experiments 1 (*Right/Wrong*) and 2 (*Original/Right*). Bottom row shows an additional experiment which classifies *original* endings vs ***wrong*** endings. In all cases, our setup implies a 50% random baseline.

Table 3 shows our results on the story cloze test set. Our classifier obtains 72.4% accuracy, 12.5% higher than the published state-of-the-art result on the task (Salle et al., 2016) . Importantly, unlike previous approaches, our classifier does not require the story corpus training data, and in fact doesn't even consider the first four sentences of the story 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, for which this task was designed. Specifically, we experiment with a recurrent neural network language model (RNNLM; Mikolov et al., 2010). Unlike the model in this paper, which only considers the story endings, this language model follows the protocol suggested by the story cloze task designers, and harnesses their ROC Stories training set, which consists of single-ending stories. We

[4] Virtually all sentences end with a period or an exclamation mark, so we do not add a STOP symbol.

[5] http://spacy.io/

Model	Accuracy
DSSM (Mostafazadeh et al., 2016a)	0.585
LexVec (Salle et al., 2016)	0.599
Niko (shared task)	0.700
tbmihaylov (shared task)	0.711
RNN	0.677
Ours	**0.724**
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Table 3: Results on the test set of the story cloze task. The first block are published results, the second block as-yet-unpublished results from the LSDSem 2017 leaderboard, the third block ours. LexVec results are taken from (Speer et al., 2016). RNN is our implementation. Human judgement scores are taken from (Mostafazadeh et al., 2016a). [ADD A SYMBOL FOR "USES STORY CONTEXT?" –CLINIC]

show that the adding our features to this powerful language model gives improvements over our classifier as well as the language model.

We train the RNNLM using a single-layer LSTM (Hochreiter and Schmidhuber, 1997) of hidden dimension 512. We use the ROC Stories for training, setting aside 10% for validation of the language model. We replace all words occurring less than 3 times by a special out-of-vocabulary character, yielding a vocabulary size of 21,582. Only during training, we apply 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 train to minimize cross-entropy.

To apply the language model to the classification problem, we select as *right* the ending with the higher value of

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

The intuition is that a *right* ending should be unsurprising (to the model) given the four preceding sentences of the story (the numerator), controlling for the inherent surprisingness of the words in that ending (the denominator).

[PLEASE CHECK THE PARAGRAPH ABOVE. WE NEED TO SAY HOW WE USE THE LM BEFORE WE EVALUATE IT! ALSO, HAS ANYONE TRIED AN APPROACH LIKE THIS BEFORE? EVEN IF WE AREN'T EXACTLY REPLICATING ANOTHER PAPER, IF SOMEONE ELSE USED RNNS FOR THIS TASK, WE SHOULD CREDIT THEM. –NAS]

Feature Type	Accuracy
Word n-grams	0.612
Character n-grams	0.639
Full model	0.645

Table 4: Results on Experiment 1 with different subsets of features.

On its own, our neural language model performs moderately well on the story cloze test. This is not surprising, as Mostafazadeh et al. (2016a) hinted that simple language models are insufficient. Indeed, using our neural language model and selecting endings based on $p_\theta(\text{ending} \mid \text{story})$ (i.e., the numerator of Equation 1), we obtained only 55% accuracy. The ratio in Equation 1 achieves 67.7% (see Table 3).[6]

We combine our linear model with the RNNLM by adding three features to our classifier: the numerator, denominator, and ratio in Equation 1. We retrain our linear model with the new feature set, and gain 2.8% absolute, reaching 75.2% (15.3% better than the published state of the art and 4.1% better than the best unpublished result). These results indicate 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.

7 Further Analysis

Most discriminative feature types. A natural question that follows this study is which style features are most helpful in detecting the underlying task an ending's author was asked to perform. To answer this question, we re-ran Experiment 1 with different sub-groups of features. Table 4 shows our results. Results show that character n-grams are the most effective style predictors, reaching within 0.6% of the full model, but that word n-grams also capture much of the signal, yielding 61.2%. These findings are in line with previous work that used character n-grams along with other types of features to predict writing style (Schwartz et al., 2013b).

[6] Further analysis of this large difference is out of the scope of this paper, but suggests careful study of suitable probabilistic inference methods for such tasks.

Most salient features—experiment 1. Table 5 shows the five features with the highest positive and negative coefficients in the logistic regression classifier for Experiment 1. These correspond to the 5 most salient features for *right* (coherent) and *wrong* (incoherent) endings, respectively. [HAVE TO BE VERY CAREFUL HERE. SEE YANO/SMITH/WILKERSON §4.3 WHERE WE PROPOSE AN ALTERNATIVE TO INSPECTING FEATURE COEFFICIENTS! –NAS]

The table shows a few interesting trends. First, authors tend to structure their sentences differently when writing coherent vs. incoherent endings. For instance, coherent endings are more likely to start with an adverb (e.g., "then", "so", "eventually"), while incoherent ones tend to start with a proper noun. In addition, we find that incoherent endings are more likely to finish the sentence with a common noun. [NOTE TO SELF TO COME BACK TO THIS AFTER WE AGREE ON METHODOLOGY. –NAS]

More interestingly, the different writing tasks seem to impose a specific sentiment on the writer. Three of the top four most salient features for detecting *wrong* endings are variants of the verb "hate". This indicates that when authors are asked to write *wrong* text, they tend to use negative language.

An alternative explanation to this hypothesis might be that the first four sentences of the stories in the ROC story corpus tend to be positive, and thus in order to make an ending *wrong*, authors adopted a negative approach. A similar idea was suggested in the original story cloze paper, where two sentiment-based baselines were evaluated. These baselines measured the relative sentiment between the ending and the previous sentences. The performance of both these baselines was roughly chance-level, which seems to suggest that this is not entirely the case. [THIS IS DEFINITELY AN OPTION, AND THAT'S PARTLY WHY I USED HEDGING ("SEEMS TO SUGGEST"). TONES IT DOWN A BIT MORE. –RS]

Most salient features—experiment 2. Table 6 shows the same analysis for Experiment 2. Interestingly, here the most salient features are quite different. As noted in Section 2, *original* endings tend to be much longer, which is indeed the most salient feature for them. Another clear style difference is the use of punctuation: *original* endings often end with exclamation marks, while *new* endings end almost exclusively with periods. Finally, syntactic difference appear to play an important role in stylistic differences between the two tasks. Specifically, *original* endings contain more common nouns (NN), adverbs (RB) and gerunds (VBG), while *new* endings seem to contain more past tense verbs (VBD), especially as the second word in the sentence.

Right	***Wrong***
'ally'	' hate'
'VBD the'	' hat'
'START RB'	'START NNP'
'ved '	'ated'
' tim'	'NN .'

Table 5
The top 5 most discriminative features for predicting *right* vs. *wrong* endings.

Original	***New***
length	'.'
'!'	'START NNP'
'NN'	'START NNP VBD'
'RB'	'START I VBD'
'VBG'	'NNS .'

Table 6
The top 5 most discriminative features for predicting *original* vs. *new* (*right*) endings.

8 Discussion

In this paper we have shown that giving a writer a different writing task affects writing style in easily detected ways. Our results indicate that when authors are asked to write the last sentence of a five sentence story, they will use different style to write a *right* ending compared to a *wrong* ending. We have also shown that writing the ending as part of one's own five-sentence story is very different than reading four sentences and then writing the fifth.

Our findings hint that the nature of the writing task imposes a different mental state on the author, which is expressed in ways which are often implicit, but can be observed using extremely simple automatic tools. This is in line with previous cognitive findings. For instance, previous studies have shown that when answering questions, people tend to adopt the writing style of the question (Ireland and Pennebaker, 2010). [THERE MUST

BE CLOSER EXAMPLES? –NAS]

Other studies have shown that writing tasks can even have a long term effect. A range of works have shown that writing emotional texts can benefit both physical and mental health (Lepore and Smyth, 2002; Frattaroli, 2006). Some of these works showed that these health benefits are also accompanied by changes in writing style (Campbell and Pennebaker, 2003).[WHICH KIND OF CHANGES? RELATED TO OUR BY ANY CHANCE? –CLINIC] The current study suggests that even subtle changes to a writing prompt can affect these processes.[BELONGS TO INTRODUCTION? –CLINIC]

This paper also reveals several important lessons for the future design of NLP tasks. The story cloze task was very carefully designed. Many factors, such as the topic diversity, as well as temporal and causal relation diversity, were controlled for (Mostafazadeh et al., 2016a). The authors also made sure each pair of endings was written by the same author, partly in order to avoid author specific style effects. Nonetheless, this paper shows that despite these efforts, several significant style differences are found between the training and the test set, as well as between the positive and negative labels.

The findings in this paper suggest that careful attention must be paid to instructions given to authors, especially in unnatural tasks such as writing a *wrong* ending. One way to avoid such problems is by using shorter text spans, such as the ones used in the Winograd schema (Levesque et al., 2011). A different approach is to use naturally occurring text, as used in recent machine reading tasks (see Section 9). However, these approaches have their shortcomings, and do not capture the same signal as intended by the author of the story cloze task.

Another lesson from this paper relates to baseline methods. The authors of the story cloze task experimented with a wide and diverse set of baselines, including n-gram baselines, word embeddings baselines, sentiment baselines, and narrative chain baselines. Nonetheless, running preliminary tests on the data, such as average length or POS and word distribution, could have revealed potentially unwanted phenomena in the data. This is particularly important in order to ensure that new technological improvements are capturing a qualitatively different signal. The substantial improvement we get when combining our model with the RNN language model (Section 5) suggests that despite the high results obtained by the simple model, methods that exploit the wider context capture a different aspect of the task. [MODIFIED THIS TO BE MORE CONCRETE. DID END UP HINTING THAT THEY SHOULD HAVE DONE A BETTER JOB AT ANALYZING THEIR DATA BEFORE PUBLISHING. FEEL FREE TO TONE DOWN. –RS]

9 Related Work

Writing style. Writing style has been an active topic of research for decades. The models used to characterize style are often linear classifiers with style features such as character and word n-grams (Stamatatos, 2009; Koppel et al., 2009). Previous work has shown that different authors can be grouped by their writing style, according to factors such as age (Pennebaker and Stone, 2003; Argamon et al., 2003; Schler et al., 2006; Rosenthal and McKeown, 2011; Nguyen et al., 2011), gender (Argamon et al., 2003; Schler et al., 2006; Bamman et al., 2014), and native language (Koppel et al., 2005; Tsur and Rappoport, 2007; Bergsma et al., 2012). At the extreme case, each individual author adopts a unique writing style (Mosteller and Wallace, 1963; Pennebaker and King, 1999; Schwartz et al., 2013b). Interestingly, previous work has shown that individual style can be affected from coarse-grained factors such as those just described, but also from other less apparent factors such as mental state, or even living in a high-elevation location (Schwartz et al., 2013a).

Unlike the works just described which compare the writing style between different authors, some works have shown that the same author can adopt a different style used when writing positive vs. negative text (Davidov et al., 2010) or when writing sarcastic text (Tsur et al., 2010). In this work, we have shown that the same author can adopt a different style when facing different writing tasks.

The line of work that most resembles our work is the detection of deceptive text. Several researchers have used stylometric features to predict deception (Newman et al., 2003; Hancock et al., 2007; Ott et al., 2011; Feng et al., 2012). Some works even showed that writing style applied when lying is different across genders (Pérez-Rosas and Mihalcea, 2014; Pérez-Rosas and Mihalcea, 2014). In this work, we have shown that

an even more subtle writing task—writing *coherent* and *incoherent* story endings—imposes different styles on the author.

Machine reading. The story cloze task, which is the focus of this paper, is part of a wide set of machine reading works published in the last few years. These include datasets like bAbI (Weston et al., 2015), SNLI (Bowman et al., 2015), CNN/DailyMail (Hermann et al., 2015), SQuAD (Rajpurkar et al., 2016), and LAMBADA (Paperno et al., 2016).

While these works have presented valuable resources for researchers, it is often the case that these datasets suffer from methodological problems caused by applying noisy automatic tools to generate the data [BETTER? –RS](Chen et al., 2016). In this paper we have pointed to another methodological challenge in designing machine reading tasks, namely that different writing tasks used to generated the data affect the writing style of the positive and negative samples, confounding the classification problem.

10 Conclusion

The research question that guided this work is the extent to which different writing tasks result in different writing styles. 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 linear 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 exact effect that a writing prompt has on an author's response. They also provide valuable lessons for designing new NLP datasets.

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.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

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

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 327–337. http://www.aclweb.org/anthology/N12-1033.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* .

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.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2358–2367. http://www.aclweb.org/anthology/P16-1223.

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.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 171–175.

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

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45(1):1–23.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. pages 1693–1701.

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

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology* 99(3):549.

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

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology* 60(1):9–26.

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.

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

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*. volume 46, page 47.

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. 2016a. 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.

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. *ACL 2016* page 24.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association* 58(302):275–309.

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.

Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 309–319.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1525–1534. http://www.aclweb.org/anthology/P16-1144.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: language use over the life span. *Journal of personality and social psychology* 85(2):291.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 440–445. http://www.aclweb.org/anthology/P14-2072.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Gender differences in deceivers writing style. In *Mexican International Conference on Artificial Intelligence*. Springer, pages 163–174.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* .

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 763–772. http://www.aclweb.org/anthology/P11-1077.

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.

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

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013b. 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.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Association for Computational Linguistics, Prague, Czech Republic, pages 9–16. http://www.aclweb.org/anthology/W/W07/W07-0602.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698* .