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 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 simple linear classifier informed with stylistic features is able to successfully distinguish between the three cases, without even looking at the story context. In addition, our style-based classifier obtains state-of-the-art results on the story cloze challenge, substantially higher than deep learning models. 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 both personal factors such as age (Schler et al., 2006) and gender (Argamon et al., 2003), by personality traits such as agreeableness and openness (Ireland and Mehl, 2014), as well as by mental states 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 (Campbell and Pennebaker, 2003; Banerjee et al., 2014a).

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 rows show revised stories with two contrastive endings: a Coherent ending and a Incoherent one.

We show that similar writing tasks with different constraints on the author can lead to measurable differences in people's writing style. As a case study, we present experiments based on the recently introduced ROC story cloze task (Mostafazadeh et al., 2016a). In this task, 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 end-

ings: a *right* (coherent) ending, and a *wrong* (incoherent) ending. Framed as a story cloze task, the goal of this dataset is 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 the story cloze task was originally designed to be a story understanding challenge, its annotation process introduced three variants of the same writing task: writing an *original*, *right* or *wrong* ending to a short story. In this paper, we show that a linear classifier informed with stylistic features can distinguish between the different endings to a large degree, even without looking at the story context (64.5%–75.6% binary classification results).

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 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). 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 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 indicate that when designing new NLP 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 commonsense story cloze challenge.

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 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 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 of the revised stories with alternative endings (for a different set of original stories that are not included in the training set).[ADDED *not* HERE –RS] 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).

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

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), SNLI (Bowman et al., 2015), LAMBADA (Paperno et al., 2016) and SQuAD (Rajpurkar et al., 2016), 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.

In addition, 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.

Different writing tasks in the story cloze Task. Several key design decisions make this task an interesting testbed for the purpose of this paper. 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.

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.

[2]The LSDSem 2017 shared task website (https://competitions.codalab.org/competitions/15333) does report higher results (71.1%)[TRICKY. WHICH NUMBER SHOULD WE REPORT HERE? LEAVE EMPTY? –RS], which are still unpublished along with the underlying methodology.

[3]I.e., the training instances are not drawn from a population similar to the one that future testing instances will be drawn from.[I ADDED **NOT** HERE, AS OTHERWISE IT IS CONFUSING –RS]

3 Surface Analysis of the Story Cloze Task

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). Previous work have shown that sentence length is also indicative of whether a text was deceptive (Yancheva and Rudzicz, 2013; Qin et al., 2004). Although writing *wrong* sentences is not the same as lying, it is not entirely surprising to observe similar trends in both tasks.

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. Once again, deceptive text is also characterized by fewer pronouns compared to truthful text (Newman et al., 2003). In contrast, *wrong* and *right* endings favor proper nouns 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...?}*$_{yc}$ *{The proper noun thing is prolly related to the task design, which stated that people had to use at least one of the characters; research shows people usually use more other references when lying, but it's not clear whether there's a pronoun/proper noun distinction.}*$_{ms}$[GOOD POINT! THIS IS TRICKY. ON THE ONE HAND, IT HURTS OUR STORY BY SAYING THAT THE DIFFERENCE IS NOT REALLY THE GENERAL TASK BY ITSELF, BUT ALSO THE CONSTRAINTS PUT BY NASRIN ET AL. IT ALSO FORCES US TO BASH THEM SOME MORE. OR WE CAN JUST IGNORE IT... –RS]

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. Furthermore, *original* and *right* endings seem to prefer enthu-

siastic language (e.g., "!"), while *wrong* endings tend to use more negative language ("hates"), as is deceptive 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

Our 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 the extent to which 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* writings.

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.

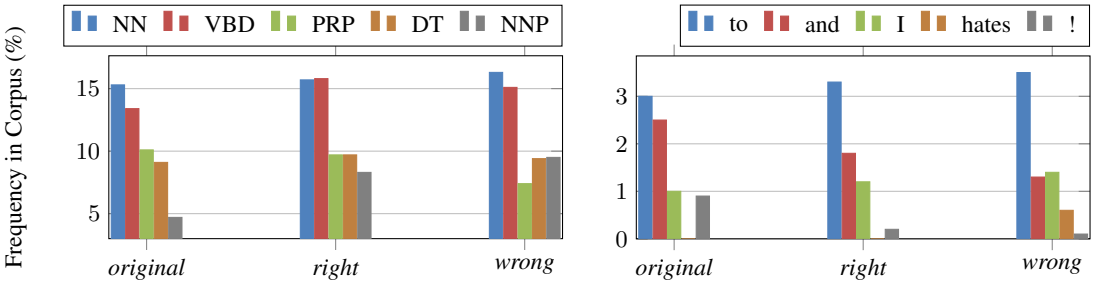
Experimental setup. In both experiments, we add a START symbol at the beginning of each sentence.[4] For computing our features, we keep n-gram (character or word) features that occur at least five times in the training set. All feature values are normalized to $[0, 1]$. For the POS features, we tag 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

Table 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

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

[5] http://spacy.io/



(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).

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.

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 75.6%.

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 the 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.

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

Model	Accuracy
†DSSM (Mostafazadeh et al., 2016a)	0.585
†LexVec (Salle et al., 2016)	0.599
†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 is our results. LexVec results are taken from (Speer et al., 2016). Human judgement scores are taken from (Mostafazadeh et al., 2016a). Methods marked with (†) use the story context in order to make a prediction.

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 an LSTM-based (Hochreiter and Schmidhuber, 1997) 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, as well as the story context for each pair of endings. We show that 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 of hidden dimension 512. We use the ROC stories for training,[6] setting aside 10% for validation of the language model. We replace all words occurring less than 3 times with 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 the Adam optimizer (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] [OURS IS THE FIRST WORK TO PUBLISH LSTM RESULTS ON THIS DATASET. –RS]

On its own, our neural language model performs moderately well on the story cloze test. 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).[7]

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 result).[AS NOW THE CODALAB SHARED TASK TABLE CONTAINS OUR NUMBERS, I REMOVED THE REFERENCES TO THESE WORKS, AND LEFT ONLY THE FOOTNOTE EARLIER WHICH SAYS THAT THERE ARE HIGHER, UNPUBLISHED NUMBERS. –RS] 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.

[6] We use the extended, 100K stories corpus (see Section 2).

[7] 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.

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.

7 Further Analysis

7.1 Most Discriminative Feature Types

A natural question that follows this study is which style features are most helpful in detecting the underlying task an 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%, which is only 3.3% worse than the full model. 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).

7.2 Most Salient Features

A followup question is which individual features contribute most to the classification process, as these could shed light on the stylistic differences imposed by each of the writing tasks.

In order to answer this question, we consider the highest absolute positive and negative coefficients in the logistic regression classifier in both Experiments 1 and 2, an approach widely used as a method of extracting the most salient features (Nguyen et al., 2013; Burke et al., 2013; Brooks et al., 2013).[8] We consider only features appearing in at least 5% of the endings in our training set.

Experiment 1. Table 5 shows the most salient features for *right* (coherent) and *wrong* (incoherent) endings in Experiment 1, along with their cor-

[8] Although it is worth noting that its reliability is not entirely clear (Yano et al., 2012).

Right	Freq.	***Wrong***	Freq.
'ed .'	6.5%	START NNP	54.8%
'and '	13.6%	NN .	47.5%
JJ	45.8%	NN NN .	5.1%
to VB	20.1%	VBG	10.1%
'd th'	10.9%	START NNP VBD	41.9%

Table 5
The top 5 most discriminative features for predicting *right* vs. *wrong* endings, along with their frequency in our story cloze training set.

Original	Freq.	***New***	Freq.
length	100.0%	'.'	93.0%
'!'	6.1%	START NNP	39.2%
NN	78.9%	START NNP VBD	29.0%
RB	44.7%	NN .	42.3%
','	12.7%	the NN .	10.6%

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

pus frequency. The table shows a few interesting trends. First, authors tend to structure their sentences differently when writing coherent vs. incoherent endings. For instance, incoherent endings are more likely to start with a proper noun and end with a common noun, while coherent endings have a greater tendency to end with a past tense verb.

Second, *right* endings will make wider use of coordination structures, as well as adjectives. The latter might indicate that writing coherent stories imposes the authors to write more descriptive text compared to incoherent ones, as is the case in truthful vs. deceptive text (Ott et al., 2011). Finally, we notice a few syntactic differences: *right* endings will more often use infinite verb structure, while *wrong* endings will prefer gerunds (VBG). [NOTE TO SELF TO COME BACK TO THIS AFTER WE AGREE ON METHODOLOGY. –NAS]

Experiment 2. Table 6 shows the same analysis for Experiment 2. As noted in Section 2, *original* endings tend to be much longer, which is indeed the most salient feature for them. This is inline with similar findings in the deception detection literature (Qin et al., 2004). An interesting observation is that exclamation marks are a strong indication for *original* ending. This indicates that authors are more likely to show enthusiasm when writing their own text compared to ending an existing task.

Finally, when comparing the two groups of salient features from both experiments, we find an interesting trend. Several features, such as "START NNP" and "NN .", which indicate *wrong* sentences in Experiment 1, are used to predict *new* (i.e., *correct*) endings in Experiment 2. This indicates that, for instance, incoherent endings have a stronger tendency to begin with a proper noun compared to coherent endings, which in turn are more likely to do so than original endings. This partially explains why distinguishing between *original* and *wrong* endings is an easier task compared to the other pairs (Section 6).

8 Discussion

The effect of writing tasks on mental states. In this paper we have shown that giving a writer different writing tasks 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.

Previous work have shown that writing task could affect mental states. For instance, several works demonstrated that writing deceptive text leads to a significant cognitive burden (Newman et al., 2003; Banerjee et al., 2014a). These works have also shown that this burden is accompanied by different writing style compared to truthful text. Other studies have shown that writing tasks can even have a long term effect, by showing that writing emotional texts can 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, mostly in the use of pronouns.

Another line of works has shown that writing style is affected by the mental state. First, the author's personality traits (e.g., depression,

neuroticism, narcissism) affect her writing style (Schwartz et al., 2013a; 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. Finally, writing style can also change from one sentence to another, for instance between positive and negative text (Davidov et al., 2010) or when writing sarcastic text (Tsur et al., 2010).

This large body of works indicates a tight connection between writing tasks, mental states and different writing styles. This connection hints that the link discovered in this paper, between the different writing tasks and the resulting difference in writing style, also goes through a change in mental state. Nonetheless, further investigation is required in order to further validate this hypothesis.

The story cloze task. This paper also provides important insights 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). A way to avoid the inherent biases people have when writing is to rather than ask them to write text, have them rate sentences for naturally occurring text by parameters such as coherence (or very differently – the level of surprise).

[COMMENTED OUT THE LESSONS ABOUT THE IMPORTANCE OF RUNNING BASELINES. –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).

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 them[THIS IS NOT CLEAR TO ME, WHAT DOES IT MEAN? –NAS] [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 mental state, and also her concrete 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. 2014a. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1469–1473. http://www.aclweb.org/anthology/D14-1155.

Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. 2014b. 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.

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* page 0261927X16666308.

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* .

Michael Brooks, Katie Kuksenok, Megan K Torkildson, Daniel Perry, John J Robinson, Taylor J Scott, Ona Anicello, Ariana Zukowski, Paul Harris, and Cecilia R Aragon. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, pages 317–328.

David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory* 6(3):203–242.

Moira Burke, Lada A Adamic, and Karyn Marciniak. 2013. Families on facebook. In *ICWSM*.

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.

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* .

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 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. *The Oxford handbook of language and social psychology* pages 201–237.

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.

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*.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. " how old do you think i am?" a study of language and age in twitter .

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.

Tiantian Qin, Judee Burgoon, and Jay F Nunamaker. 2004. An exploratory study on promising cues in deception detection and application of decision tree. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, pages 23–32.

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.

Yla Rebecca Tausczik. 2009. Linguistic analysis of workplace computer-mediated communication.

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* .

Maria Yancheva and Frank Rudzicz. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Association for Computational Linguistics*.

Tae Yano, Noah A Smith, and John D Wilkerson. 2012. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 793–802.