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

Anonymous ACL submission

Abstract

People's writing style depends not just on the authors' personal traits but also on the intent and the cognitive states of the authors. In this paper, we show how similar writing tasks involving different cognitive processes can lead to measurable differences in people's writing style. We present a case study based on the *story cloze task* (Mostafazadeh et al., 2016a), where annotators were assigned with 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 vs. incoherent. 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 sophisticated deep learning models, even without looking at the story context. Our results demonstrate that different task framings can dramatically affect the way people write. They also provide important lessons for designing new NLP tasks.

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), and the cognitive states 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 people's writing style is affected by intricate factors, such as the nature of the writing task that would involve different cognitive processes.

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. The first row shows an *original* story written by one author. The second and third row show revised stories with two contrastive endings: a *right* ending and a *wrong* one.

As a case study, we present experiments based on the recently introduced ROC story cloze task (Mostafazadeh et al., 2016a). In this task, crowdsourced 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. Table 1 shows an example of an *original* story, a *coherent* story

and an *incoherent* story.

While originally designed to be a commonsense 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 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. 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.

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 previously reported state-of-the-art (Salle et al., 2016). Our analysis reveals that sentiment plays a significant role, such that authors writing the wrong endings often engage negative sentiment.

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 an additional 2.8% gain – 75.2% – which is 15.3% better than the best published result.[1]

The contributions of our study are threefold. First, findings from our study can potentially shed light on how different cognitive overloads influence the style of language people use in writing. Second, our results provide valuable lessons for designing new NLP tasks, both in terms of the potential impact of even the smallest details, and the need to carefully run baseline models. Third, we establish a new state-of-the-art result on the commonsense story cloze challenge.

[1]Recently, a shared task for the story cloze task has been published (https://competitions.codalab.org/competitions/15333). At the time of submission, the leading results was 71.1%, but no details are available for their results.

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 3, 4 and 5 respectively. Sections 6 and 7 present an ablation study and a discussion, followed by related work in Section 8. We conclude at Section 9.

2 The Cloze Story Task

In this paper, we seek to understand how different writing tasks affect writing style. To do so, 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 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).[2] 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. 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

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

in their endings, the system is required to determine 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).

State-of-the-Art Performance. Interestingly, at the time of submission, 10 months after the paper was first published, the published benchmark on this task is still below 60% (Salle et al., 2016).[3] 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 period of time. This suggest that this task is challenging and that high performance is hard to achieve. *{SQuAD was officially published in Nov 2016. by the time the official EMNLP talk was presented, the leaderboard performance (still unpublished) was 30% beyond the best results in the paper. Given that, the statement of this paragraph seems a bit of overkill...?}$_{yc}$*

Different Writing Tasks in the Story Cloze Task. Mostafazadeh et al. (2016a) made substantial efforts to ensure the quality of this task. 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.

Despite these efforts, several key aspects were not controlled for. First, the training set for the task (ROC Stories corpus) is not really a training set, as it contains only positive (*right*) samples, and not negative (*wrong*) ones. *{As far as I understand, this was intentional — they didn't want people to pick up on random superfluous cues that can inevidently get into the data creation process, exactly the kind that our work picks up. Given that, stating this can be viewed as misunderstanding, I'm afraid.}$_{yc}$*

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. We show that this is indeed the case.

Second, 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. In this work, we show that these differences are significant and impose different writing styles on authors.

Initial 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 much longer (11 words per sentence) than *right* endings (8.75 words), which are in turn also longer (though not as much) than *wrong* ones (8.47 words). *{Is there psycholinguistic literature that we can cite that may have shown that cognitive burden causes writers to write shorter sentences...?}$_{yc}$* 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...?}$_{yc}$*

[3] The LSDSem'17 shared task website does report higher results (71.1%), which are still unpublished.

Finally, Figure 1b presents the distribution of five frequent words across the different groups. The figure shows that *original* endings prefer to use coordinations ("and") more than *right* endings, and substantially more than *wrong* ones. Furthermore, *original* and *right* endings seem to prefer positive words (e.g., "better"), while *wrong* endings prefer negative ones ("hates"). {*Again, any cogsci literature on this?*}$_{yc}$ Next we show that these style differences are not anecdotic, but can be used to distinguish between the different groups of text.

3 Model

The 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, classic NLP tools, which have been shown to be very effective for recognizing style (see Section 8). We describe our model below.

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

- ***Length***. The number of words in the sentence.
- ***Word n-grams.*** We use sequences of 1-5 words. Following (Tsur et al., 2010; Schwartz et al., 2013b), we distinguish between high frequency and low frequency words. Specifically, we replace content words, which are often low frequency, with their part-of-speech tags (Nouns, Verbs, Adjectives and Adverbs).
- ***Character n-grams.*** Character n-grams are one of the most useful features in identifying author style (Stamatatos, 2009). We use character 4-grams.

4 Experiments

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

Experiment 1: Correct/Wrong Endings. The goal of this experiment is to measure to what extent style features capture differences between the *right* and *wrong* endings. As the cloze story 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. We split it (90/10) into our training and development set. We keep the story cloze test set as is. The final size of our training/development/test sizes are 3,366/374/3,742 endings, respectively.

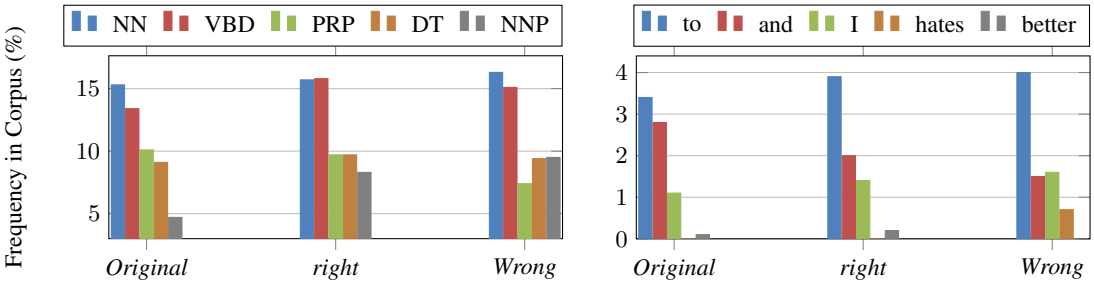
It is worth nothing that our classification task is slightly different from the story cloze task. Instead of classifying pairs of endings, one which is *right* and another which is *wrong*, we take the set of *right* endings as positive samples and the set of *wrong* endings as our negative examples. By ignoring the coupling between *right* and *wrong* pairs, we are able make a more general claim about the style used when writing each of the tasks.

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 (*right*) ending to an existing story. We use the endings of the ROC stories as our *original* training samples and *right* endings from the story cloze task development and test sets as *new* training samples. As there are far more *original* samples than *new* ones, we randomly select N *original* samples, where N is the number of *new* samples, such that we have balanced labels in our training, validation and test sets (sizes are the same as in Experiment 1). We randomly sample 5 training sets and 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 L2 loss. We grid search the regularization parameter on our development set.

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

[5] spacy.io/



(a) PoS tags (b) Words

Figure 1: The distribution of five frequent POS tags (1a) and words (1b) across different endings: *Original* – endings of stories from the ROC story corpus (the story cloze training set). *right* – correct endings, *Wrong* – wrong endings (both from story cloze task).

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

5 Results

Figure 2 shows our results. Results show that for Experiment 1, our model obtains results well above a random baseline – 64.5% classification accuracy. For Experiment 2, we get even higher results – 68.5%. These numbers indicate that these different writing tasks clearly impose different writing styles on authors.

As a complementary experiment, we measured whether these style differences are additive. That is, whether the style differences between *right* endings and *wrong* ones are different from the differences between *right* and *original* endings. We repeated Experiment 2, this time comparing between *original* and *wrong* sentences. Our hypothesis is that differences would be even clearer. Our results (Figure 2) show that this is indeed the case: the classifier's accuracy jumps to 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 here the task is given two endings, which one is *right* and which is *wrong*. In order to solve the task, we use the output of our trained classifier. For each pair of endings in our test set, if the classifier assigns different labels to them, we keep these labels. If they share the same label, then one of them must be wrong, so we use the classifier confidence level as a deciding factor: the label of the one with the lower confidence is reversed.

Table 2 shows our results on the story cloze test set. Our classifier obtains 72.4% accuracy, which is 12.5% higher than the published state-of-the-art result on the task. Importantly, unlike previous approaches, our classifier does not require the story corpus training data, and in fact doesn't even look at the first four sentences of the story in question. These numbers indicate that the styles of *right* and *wrong* endings are indeed very different.

Other than the published works that tackled this task, a few recent works published their results in the LSDSem shared task website.[6] At the time of submission, our results are still state-of-the-art (though by a smaller margin). No other information other than the name of the groups and their results is available.

Combination with an RLM. 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 language model (RLM, Mikolov et al. (2010)). Unlike the model in this paper, which

[6] https://competitions.codalab.org/competitions/15333

Model	Results
DSSM (Mostafazadeh et al., 2016a)	0.585
LexVec (Salle et al., 2016)	0.599
Niko (shared task)	0.7
tbmihaylov (shared task)	0.711
RNN	0.677
Our Model	**0.724**
Combined (Our model, RNN)	**0.752**
Human judgment	1

Table 2: Results on the test set of the cloze story task. Upper part are published results. LexVec results are taken from (Speer et al., 2016). Results in the middle part are taken from the LSDSem 2017 shared task webpage. Bottom part is our results. RNN is our implementation of an RNN.

only considers the story endings, this language model follows the protocol suggested by the story cloze task designers, and harnesses the ROC Stories training set, which consists of single-ending stories. We show that adding the language model features further boosts our performance on the story cloze task.

We train the RLM using a single-layer LSTM (Hochreiter and Schmidhuber, 1997) of hidden dimension $h = 512$. 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 $|V|$ =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 with backpropagation on cross-entropy.

On its own, our neural language model performs moderately on the story cloze test. This is not surprising, as Mostafazadeh et al. (2016a) hinted that simple language models do not perform well on the task. Indeed, using our neural LM and selecting endings based on $p_\theta(ending|story)$, we obtain only 55% accuracy. However, when using the likelihood ratio $\frac{p_\theta(ending|story)}{p_\theta(ending)}$ to select endings, performance jumps to 67.7% (see Table 2).[7]

We combine our linear model with the RLM by adding three features to our classifier: $p_\theta(ending|story), p_\theta(ending)$ and $\frac{p_\theta(ending|story)}{p_\theta(ending)}$. We retrain our linear model with the new feature set, and gain a performance boost of 2.8% – up to 75.2%, which is 15.3% better than the published state-of-the-art, and 4.1% better than the best unpublished result. These results indicate that while style features can be used to obtain high accuracy on the task, adding context and training on a large dataset can improve performance even more.

[7] Further analysis of this large performance gain is out of the scope of this paper.

Feature Type	Result
Word n-grams	61.2%
Char n-grams	63.9%
Full Model	64.5%

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

6 Ablation Study

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

Most Salient Features – Experiment 1. Table 4 shows the five features with the highest positive and negative weights in the logistic regression classifier hyperplane for Experiment 1. These correspond to the 5 most salient features for *right* (coherent) and *wrong* (incoherent) endings, respectively.

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.

More interestingly, the different writing tasks

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

Table 4
The top 5 most discriminative features for predicting *right* and *wrong* endings.

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

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

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 the case. *{Is it really because there's no statistical tendency in the original stories to have happy endings, as opposed to the alternative possibility — the sentiment classifier used by Monstafazadeh 2016 didn't quite nail down the optimal feature encodings?}*$_{yc}$

Most Salient Features – Experiment 2. Table 5 shows the same analysis for Experiment 2. Interestingly, here the most salient features are quite different. As noticed on 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.

7 Discussion

In this paper we have shown that different writing tasks affect writing style. 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 a 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 unfolded using 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).

Other studies have shown that writing tasks can even have a long term effect. A range of works have shown that writing emotional texts and benefit both physical and mental health (Lepore and Smyth, 2002; Frattaroli, 2006). Some of these works showed that these changes are also accompanied by changes in writing style (Campbell and Pennebaker, 2003). The results of the current study can shed more light on the processes that people undergo when presented with different writing tasks.

This paper also reveals several important lessons for the future design of NLP tasks. First, it stresses the need to carefully control for even seemingly minor details. Second, it emphasizes the importance of running baseline models. While recent advances in NLP suggest that many classic NLP methods are out-of-date, this is not necessarily the case, especially for small datasets. This is particularly important in order to ensure that new technological improvements are capturing a qual-

itatively different signal. Luckily, the substantial improvement we get when combining our model with the RNN language model (Section 4) suggests that despite the high results obtained by the simple, classic model, the more recent RNN approach captures a different aspect of the task.

8 Related Work

Writing Style. Writing style has been an active topic of research for decades now. The models used to effectively solve such tasks are often linear classifiers with style features such as character and word n-grams (Stamatatos, 2009; Koppel et al., 2009). Previous works have 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), gender (Argamon et al., 2003; Schler et al., 2006) and native language (Koppel et al., 2005). At the extreme case, each individual author adopts a unique writing style (Pennebaker and King, 1999; Schwartz et al., 2013b). Interestingly, previous work have 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 different style used when writing positive vs. negative text (Davidov et al., 2010) and 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 works that most resembles our work is the detection of deceptive text. Several works 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 cloze story task, which is the focus of this paper, is part of a wide set of machine reading works published in the last couple of 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 the NLP community (as well as for the deep learning community), it is often the case that these datasets suffer from methodological problems, such as inherent ambiguity caused by applying automatic tools when generating the datasets (Chen et al., 2016). In this paper we have pointed to other methodological problems in machine reading tasks, namely the different writing tasks used to generated the data, which largely affect the writing style of the positive and negative samples, as well as create a discrepancy between the training and the test data.

9 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 people writing a story on their own and people continuing someone else's story and the difference between the same author writing a coherent and an incoherent story ending. In both cases, a simple linear model reveals measurable differences in people's writing styles, which in turn allows our final model to reach state-of-the-art results on the story cloze task, more than 12% improvement compared to the best published result.

The findings presented in this paper have cognitive implications, as they motivate further research on the exact effect that different writing tasks have on humans. They also provide valuable lessons for designing new NLP datasets. Future work will include testing whether other similar writing tasks, such as fake news, also impose their own unique and identifiable style on their authors.

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.

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.

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.

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

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

H Andrew 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.

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* .