Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6709–6717
Marseille, 11–16 May 2020

Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment

Lorenzo De Mattei[123], Michele Cafagna[23], Felice Dell'Orletta[1] and Malvina Nissim[3]
ItaliaNLP Lab, ILC-CNR, Pisa, Italy[1],
Department of Computer Science, University of Pisa, Italy[2],
University of Groningen, The Netherlands [3]
{l.de.mattei,m.cafagna,m.nissim}@rug.nl
felice.dellorletta@ilc.cnr.it

Abstract

We automatically generate headlines that are expected to comply with the specific styles of two different Italian newspapers. Through a data alignment strategy and different training/testing settings, we aim at decoupling content from style and preserve the latter in generation. In order to evaluate the generated headlines' quality in terms of their specific newspaper-compliance, we devise a fine-grained evaluation strategy based on automatic classification. We observe that our models do indeed learn newspaper-specific style. Importantly, we also observe that humans aren't reliable judges for this task, since although familiar with the newspapers, they are not able to discern their specific styles even in the original human-written headlines. The utility of automatic evaluation goes therefore beyond saving the costs and hurdles of manual annotation, and deserves particular care in its design.

Keywords: Natural Language Generation, Stylistic variations, Evaluation

1. Introduction

Automatic headline generation is conceptually a simple task which can be conceived as a form of extreme summarisation (Rush et al., 2015): given an article, or a portion of it, generate its headline.

Generation of headlines though is not just a matter of summarising the content. Different newspapers report the news in different ways, depending on their policies and strategies. For example, they might exhibit some topic-biases, such as writing more about gossip vs more about politics. But even when reporting on the same topics, they might exhibit specific stylistic features related to word choices, word order, punctuation usage, etc. This might be even more evident when newspapers are positioned at opposite ends of the political spectrum (Cafagna et al., 2019b). Such newspaper-specific style is likely to be exhibited not only in the articles' body, but also in the headlines, which are a prime tool to capture attention and make clear statements about the newspaper's position over a certain event.

Can this newspaper-specific style be distinguished? And is it preserved in automatically generated headlines? To answer such questions, we train newspaper-specific headline generation models, and evaluate how style-compliant the generated headline is for a given newspaper. How such evaluation can be performed though is yet another research question of its own.

Evaluating generated text just using standard metrics based on lexical overlap is normally not accurate enough (Liu et al., 2016). In machine translation, for example, the decisive, final system evaluation is typically human-based, as the lexically-based BLEU score is not exhaustive. Automatic evaluation strategies are still used because human evaluation is expensive, not always available, and complex to include in highly iterative developments. However, human evaluation is not always a decisive and accurate strategy, since there might be aspects of text that for people are not so easy to grasp. For example, in profiling, where differently from the assessment of the goodness of translated text, evaluation can be performed against discrete gold labels, several studies found that humans are definitely not better than machines in identifying the gender of a writer (Koppel et al., 2002; Flekova et al., 2016; van der Goot et al., 2018). Similarly, humans failed to outperform automatic systems in recognising the native language of non-English speakers writing in English (Malmasi et al., 2015). Baroni and Bernardini (2005) also find that seven out of ten subjects, including professional translators, performed worse than a simple SVM at the task of telling apart original from translated texts.

More generally, Gatt and Krahmer (2018) have observed that it is difficult to ascertain if readers can perceive subtle *stylistic variations*, and past human-based evaluations of style have indeed shown very low inter-rater agreement (Belz and Kow, 2010; Cahill and Forst, 2009; Dethlefs et al., 2014). In spite of a recent surge of works focusing on style in generation (Ficler and Goldberg, 2017; Hu et al., 2017; Keskar et al., 2019, e.g.), and on attempts to define best practices for human and automatic evaluation (van der Lee et al., 2019), reliable and shared evaluation metrics and strategies concerning style-aware generation are still lacking (Fu et al., 2018).

As a contribution to this aspect, we develop style-aware headline generation models, and discuss an evaluation strategy based on text classification, which is particularly useful given that human judgement for this task is found to be unreliable. While the strategy of using classification as evaluation is in itself not new, this work has a series of innovative aspects which we discuss in the context of related work (Section 2.).

Contributions. (i) We provide a dataset of news from two major Italian newspapers, one left-oriented and one right-oriented containing a partially topic-aligned subset which could be exploited in further style transfer experiments; (ii) we develop and share models based on a pointer network with coverage attention to generate newspaper-specific headlines for two Italian newspapers given the article; (iii) we show that an automatic, classification-based methodology can be used to evaluate style-compliance in Natural Language Generation (NLG), and can successfully substitute human judgement which proves to be unreliable for this task. The dataset and the models are available at this repository: `https://github.com/LoreDema/RepGioDataset`.

2. Related Work

The focus of this contribution is not on investigating best models for style-compliant headline generation. Rather, we want to test an automatic evaluation strategy that can overcome the limitation of unreliable human judgement. Besides the works mentioned in the Introduction to frame the problem, we will not discuss further related work on style modelling or summarisation. Rather, we concentrate on discussing previous works that make use of automatic classification for the evaluation of NLG systems, also to show in what sense our approach differ from existing ones.

Using a classifier to assess the goodness of generated texts in connection to a broad definition of style-aware generation has been used in several previous works (Hu et al., 2017; Tian et al., 2018; Prabhumoye et al., 2018; John et al., 2018; Li et al., 2018, e.g.). However, these works tend to focus on sentiment aspects (transforming a positive review into a negative one, for example), which are usually mostly associated to a lexical problem (only a small part of *style*). Indeed, the problem of style transfer is usually addressed within the Variational Autoencoder framework and/or trough lexical substitution. Lexical substitution was also the key element of a system developed for obfuscating gender-related stylistics aspects in social media texts (Reddy and Knight, 2016), where a classification-based evaluation was used.

In addition, Li et al. (2018) compared the automatic classification-based evaluation with human evaluation. They find a high correlation between human and automatic evaluation in two out of their three data-sets, showing the validity of the automatic approach. However, the task of sentiment analysis, though subjective, is not too hard for humans, who are usually able to perceive sentiment encapsulated in text. Rao and Tetreault (2018) also exploited human and automatic classification as benchmarks for a machine translation system that translates formal texts into informal texts and vice-versa. Also in this case, usually text register is something that humans are quite able to grasp.

Our work differs from the above in at least two respects. One is that we want to evaluate the capabilities of an NLG system to learn (different) stylistics aspects from (different) training data sets, rather than evaluating the capabilities of style transfer systems mostly based on lexical substitution. The other is that the stylistic aspects that we attempt to model are not easily identified by human annotators. Therefore, relying on human-based evaluation in a real setting is not an option, and even the classification-based method cannot be easily validated against human judgement for this task. Also because of this, we devised a quite fine-grained evaluation setting, carefully selecting training and testing conditions.

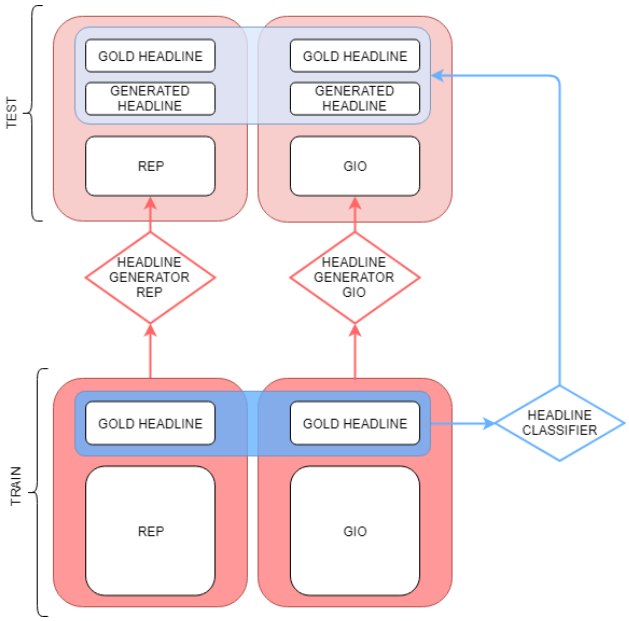

Figure 1: Red: generation task. Blue: classification task. Darker: training. Lighter: testing.

3. Approach and Models

The principle behind our approach is using a classifier to assess the style-compliance of automatically generated text.

Specifically, we train two models to generate headlines for newspaper articles coming from two (politically) different newspapers, namely *La Repubblica* (left-wing), and *Il Giornale* (right-wing), and expect that the generated headlines will carry some newspaper-specific characteristics (see also (Potash et al., 2015; Tikhonov and Yamshchikov, 2018)).

At the same time, on the gold headlines from the two newspapers we train a prediction model that learns to classify a given headline as coming from one newspaper or the other. Good performance of this classifier indicates that it is able to distinguish the two sources.

In order to test whether the generation is indeed newspaper-specific, we run the classifier on the automatically generated headlines and verify whether it is able to correctly classify their source.

Figure 1 shows an overview of the approach.

3.1. Generation Models

As the focus of this contribution is not on making the best model for headline generation, rather on evaluation strategies, we leverage existing implementations of sequence-to-sequence networks. More specifically, we experiment with the following three models:

- *Sequence-to-Sequence with Attention (S2S)*
  We used a sequence-to-sequence model (Sutskever et al., 2014) with attention (Bahdanau et al., 2014) with the configuration used by See et al. (2017) but we used a bidirectional instead of a unidirectional layer. This choice applies to all the models we used. The final configuration is 1 bidirectional encoder-decoder layer with 256 LSTM cells each, no dropout and shared embeddings with size 128; the model is optimised with Adagrad with learning rate 0.15 and gradient clipped (Mikolov, 2012) to a maximum magnitude of 2.
- *Pointer Generator Network (PN)*
  The basic architecture is a sequence-to-sequence model, but the hybrid pointer-generator network uses a *pointing mechanism* (See et al., 2017) that lets it copy words from the source text, and generate words from a fixed vocabulary. This allows for a better handling of out-of-vocabulary words, providing accurate reproduction information, while retaining the ability to reproduce novel words.
- *Pointer Generator Network with Coverage (PNC)*
  This model is basically a Pointer Generator Network with an additional coverage attention mechanism that is intended to overcome the copying problem typical of sequence-to-sequence models. This is done by penalising the attention over already generated words (See et al., 2017).

In order to assess the quality of the generated headlines, independently of whether they were maintaining or not the style of the source, we ran a human-based evaluation on a variety of criteria, including grammatical correctness and appropriateness to the article's content (for details see (Cafagna et al., 2019a)).

Results showed that while the basic sequence-to-sequence model produces rather low quality headlines, the pointer network, with and without attention, yields headlines whose grammaticality is on par with the gold, human-written headlines.[1] Automatically generated headlines apparently are not as attractive towards reading the whole paper as the gold headlines, but compared to the latter they were evaluated much more appropriate in terms of reflecting the article's content.

For the current evaluation experiments we thus opt for a pointer network with coverage attention, and generate headlines according to different newspapers' styles. We train two pointer network models that, given the first portion of an article (approx. 500 words), learn to generate its respective headline. The first model is trained on articles from *la Repubblica*, while the second model is trained on *Il Giornale*. From an architecture and implementation perspective, the models and their parameters are identical.

[1] Please note that in any case humans do not judge either gold or automatically produced headlines as particularly correct according to grammatical standards, as grammatical correctness per se is not necessarily a requirement of news' titles (Cafagna et al., 2019a).

	Rep F1	Gio F1	AVG F1
classifier	0.813	0.812	0.813
human	0.619	0.640	0.630

Table 1: Classification performance on random split.

3.2. Classifier

We use a Bidirectional LSTM (Bi-LSTM, Hochreiter and Schmidhuber (1997)) which exploits as features the concatenation of word and character embeddings. We used a word embeddings lexicon trained with word2vec (Mikolov et al., 2013) on the ItWac Corpus (Baroni et al., 2009) by Cimino et al. (2018). The character embeddings are extracted by a Convolutional Neural Network (CNN, LeCun et al. (1995) that takes as input a sequence of one-hot encoded characters. The CNN weights are optimised during training. We use a sigmoid layer as classifier.

For each training setting (see Section 5.2.), we extracted a randomly sampled validation set (10% of the training set) which we used for model selection and fine tuning. We use binary cross entropy as a loss function, and the Adam optimiser (Kingma and Ba, 2014) for optimisation.

4. Data

We scraped the websites of two major Italian newspapers, namely *la Repubblica* and *Il Giornale*, collecting a total of approximately 275,000 article-headline pairs. The two newspapers are not equally represented, with *Il Giornale* covering 70% of the data (in terms of documents, though not in terms of tokens). In all of the experiments we create training sets with an equal/comparable number of documents for the two newspapers.

For our experiments we want to account for potential topic biases in the two newspapers, and reduce them as much as possible. This should help us to better disentangle newspaper-specific style from potential newspaper-specific topics. Thus, we create a subset of the data where articles are topic-aligned.

4.1. Alignment

While we work with headlines, the alignment procedure is run over the whole articles. This is exactly because we want the headlines to refer to the same topics, but we know that they might not express the same content in the same way. Thus, we expect that headlines of aligned articles might not necessarily be that similar (see indeed also examples in Table 2.

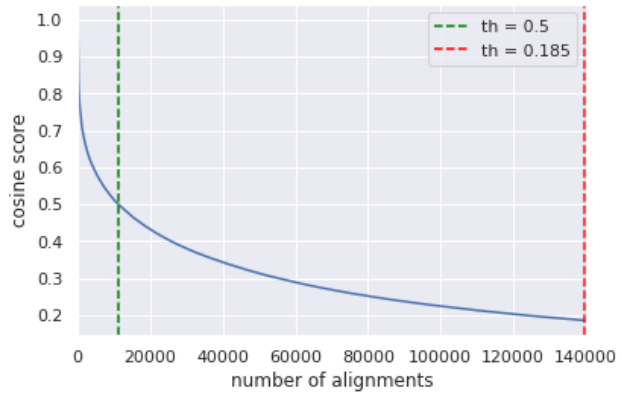

Figure 2: Trend of the number of alignments varying with the cosine similarity threshold. The green vertical dashed line is the stricter threshold, used to get the best alignments, the red one is the looser one.

First, we clean the full articles, removing stop words and punctuation. Second, we compute the tf-idf vectors of all the articles of both newspapers and we create subsets of relevant news filtering by date, i.e. considering only news which were published in approximately the same, short, temporal range for the two sources. Third, on the tf-idf vectors we compute cosine similarities for all news in the resulting subset. Fourth, we rank them, and retain only the alignments that are above a certain threshold.

The threshold is chosen taking into consideration a trade-off between number of documents and quality of alignments. The quality is assessed by manual inspection of random samples. In this experiment we choose two different thresholds: one is stricter (> 0.5) and we use it to select best alignments for the test set; the other one is looser (> 0.185, and $<= 0.5$) and we use it to select a portion of alignments to use in one of the training sets we experiment with (train-M, see Section 4.3. below).

In Figure 2 we show the trade-off between the strictness (in terms of cosine similarity) and the number of alignments. As can be expected, the number of alignments exponentially grows when decreasing the similarity score. Our stricter threshold (the green dashed line, 0.5) guarantees high quality alignments, while the looser one (the red dashed line, 0.185) provides a large number of at least partially aligned news. As quality control, we observe that restricting the considered news to a short time span makes it possible to obtain reliable alignments even with a relatively low similarity thresholds, while preserving some substantial number of instances, which we need to use for training. In Table 2 we report some examples of aligned headlines with varying similarity scores. As mentioned before, while articles might be exhibit high lexical overlap which has indeed led to strict alignment (> 0.5), the *La Repubblica*'s headline might be very different than the one written by *Il Giornale*, highlighting different aspects of the news in different ways.

4.2. Test set

The test set stays the same across all settings.

It contains only aligned headlines (11k total), which are selected after the alignment procedure described in Section 4.1. as having a minimum cosine distance of 0.5, thus ensuring their articles are lexically very similar. The rationale behind this is that testing on aligned data tries to remove a topic factor: if the classifier is able to distinguish generated headlines from the two newspapers in spite of them coming from the lexically aligned dataset, these headlines are likely to carry some characteristics of the two newspapers that are not necessarily topic-related.

4.3. Training sets

We create two different training sets of equal size, each composed of a total of 130K documents: 65K from *la Repubblica* and 65K from *Il Giornale*. These two training sets differ with respect to alignment and therefore potential topic bias:

- **train-D**, where we exclude all aligned data, resulting in a topic biased dataset, since the two newspapers often focus on different topics (*Il Giornale* for example has much more gossip than *la Repubblica*);
- **train-M**, where we include weakly aligned data (cosine distance between 0.185 and 0.5), resulting in a mixed, less topic biased dataset; train-M is therefore more similar than train-D to the test set (which, as explained, only includes strongly aligned texts).

Please note that each training set contains two equally represented portions of the two newspapers. Thus train-D contains a subset of *la Repubblica* and a subset of *Il Giornale*, and likewise for train-M.

5. Classification as Evaluation

Given that we want to train models that are able to generate headlines retaining the specific style of a given newspaper, we will know that we are successful if indeed our automatically generated headlines can be recognised as pertaining to one and not the other source.

In this Section we outline our approach to perform this non-trivial evaluation and the results we obtain.

5.1. Automatic vs Human Classification

A first option is to ask humans to perform this evaluation, but as mentioned, humans have proven not much reliable in capturing stylistic aspects (Belz and Kow, 2010; Cahill and Forst, 2009; Dethlefs et al., 2014; Gatt and Krahmer, 2018). A second option is to do this evaluation automatically, but we need to have reliable models that are able to distinguish the two sources/styles.

In order to assess the classifier's ability to correctly label the headlines from the two newspapers, we randomly split

cosine score	newspaper	alignment
0.96	rep	Estroverso o nevrotico? Lo dice la foto scelta per il profilo social *en:[Extrovert or neurotic? The photo chosen for the social profile says so]*
	gio	L'immagine del profilo usata nei social network rivela la nostra personalità *en:[The profile picture used in social networks reveals our personality]*
0.5 (strict)	rep	Egitto, governo si dimette a sorpresa *en:[Egypt, government resigns surprisingly]*
	gio	Egitto, il governo si dimette *en:[Egypt, government resigns]*
0.185 (loose)	rep	Elezioni presidenziali Francia, la Chiesa non si schiera né per Macron né per Le Pen *en:[Presidential elections France, the Church does not take sides either for Macron or for Le Pen]*
	gio	Il primo voto con l'incubo Isis ma il terrorismo esce sconfitto *en:[The first vote with the Isis nightmare but terrorism comes out defeated]*

Table 2: Example of alignments between *La Repubblica* and *Il Giornale*, extracted with different similarity scores. The second and the third one are respectively the strict and the loose threshold used to split the alignments. The first two headlines are well aligned, the third one has a partial alignment.

	in-generator	**cross-generator**
train-D	setting 1	setting 3
train-M	setting 2	setting 4
test set = same and aligned for all settings		

Table 3: Experimental settings. In **train-D** all of the aligned data is excluded; in **train-M** the data is mixed, thus also including weakly aligned texts (highly aligned data is only used in the test set). The two trainsets are equal in size, and the two corpora therein are balanced, too. In **cross** settings we use the model trained on one newspaper to generate headlines from articles of the other newspaper.

our gold data into 80% training and 20% test (no generated data is involved at this stage, and no information about news alignment is exploited). As a preliminary test, we asked one annotator (largely familiar with one of the two newspapers) to label 100 gold headlines randomly picked to get a first idea of the task's feasibility.

Results for both model and the human judge are reported in Table 1. We take them as general indication that (i) headlines are indeed classifiable automatically with good accuracy, (ii) humans seem not as reliable at the same task.

At this stage though we do not know if the classifier's ability is related to detecting the newspapers' specific styles or rather content. Indeed, the classification model is trained on non-aligned data, and thus potentially topic biased. We therefore design our experiments using different training strategies and splits, but a single testset across all settings, in order to best evaluate newspaper-specific style, rather than content. We also include more humans in the evaluation loop, for comparison and to further verify their ability at this task.

5.2. Settings

We generate and classify headlines under the four different settings shown in the matrix in Table 3.

Training Generation Models For generation, in all settings we always train two distinct generation models: one on the *la Repubblica* data, which learns to generate *la Repubblica*-specific headlines, and one on the *Il Giornale* portion of the documents, learning to generate *Il Giornale*-specific headlines. In setting1-3 the training is done over the topic-biased training sets (train-D), and in setting2-4 over the mixed datasets (train-M).

Applying Generation Models on the Testset When generating headlines, we use two conditions, according to whether the generation model is tested on articles from the same newspaper it was trained on (*in-newspaper*, settings1-2) or not (*cross-newspaper*, settings3-4).

In settings1-2, we use each generator on its own test set: we ran the *la Repubblica* model over *la Repubblica* articles in the test set, and generated the corresponding headline. Likewise for *Il Giornale*.

In settings3-4, instead, we cross-test the models: we run the *la Repubblica* model over *Il Giornale* articles in the test set, and generate the corresponding headline. Even though the articles come from the other newspaper, we expect that the model, if it has learnt appropriately, still tries to come up with a *la Repubblica*-specific title. We did the same with *Il Giornale* model, running it over *la Repubblica* test set.

Evaluating Generation Models through Classification For classification, we trained two classifiers: one on the topic-biased train-D (settings1-3), the other on the mixed train-M (settings2-4). At classification stage, we assess the performance of the generators using the respective classifier for each setting over the following headlines:

i. a validation set which comes from the same distribution of each training set;

ii. gold headlines in the test set;

iii. generated headlines in the test set:

- in *settings1-2* we test in-newspaper generated headlines;

	Ann 1	Ann 2	Ann 3	Agreement
gold	0.58	0.62	0.57	0.16
setting 1	0.57	0.59	0.54	0.14
setting 2	0.57	0.60	0.56	0.13

Table 4: Annotators' accuracy and agreement on sampled aligned test sets. The agreement is computed as Krippendorff's Alpha Reliability.

- in *settings3-4* we test cross-newspaper generated headlines.

In each case, we assess the influence of topic bias and similarity between training and test set by testing both the model trained on train-D and that trained on train-M.

5.3. Expectations

The experiments were designed and run with the following expectations for the classification models:

E1 reasonable classification performance (above 50% baseline) on the generated headlines in all settings, indicating that the generators are able to capture newspaper-specific traits and reproduce them in the generated headlines. We expect in any case the performance to be lower than on gold headlines in the same setting;

E2 better classification performance on the generated headlines in setting2 than in setting1, as the test set is strict-aligned, thus topic-unbiased, while train-D (setting1) is highly topic-biased;

E3 worse classification performance on gold headlines of the test sets than those of the validation sets as the latter come from the same distribution as the training sets, while the test set is strict-aligned; this is especially true for setting1, where we expect a larger gap between validation and test; the gap should be smaller in setting2, since the training set is closer to the test set;

E4 good performance on the cross-generated headlines (settings3-4), showing that a newspaper's style is preserved in headlines even when generated from articles of a different newspaper, though lower than the classification performance of the in-newspaper generation (settings1-3). The smaller the difference between setting1 and setting3 (and setting2 and setting4), the better the model captures newspaper-specific stylistic features.

5.4. Results

We discuss the classifiers' results in relation to our expectations. Before doing so, we run a few more human-based evaluations, which we report on first.

In order to further assess human ability to distinguish headlines from the two newspapers in the same settings of the classifiers (rather than a random split as briefly reported in Section 5.1. above), we asked three annotators to label 200 gold headlines each picked randomly from the aligned test set (100 from *la Repubblica*, 100 from *Il Giornale*). Also we asked the annotators to label 200 headlines generated automatically in setting 1 and setting 2. This evaluation is therefore directly comparable with the automatic evaluation over the gold data and the generated headlines in the corresponding settings. All annotators are familiar with at least one of the two newspapers.[2]

Test set	Rep F1	Gio F1	AVG F1
train-D (settings1-3)			
validation	0.819	0.815	0.817
gold	0.755	0.703	0.729
in-generated (setting1)	0.701	0.630	0.666
cross-generated (setting3)	0.682	0.548	0.615
train-M (settings2-4)			
validation	0.810	0.809	0.810
gold	0.782	0.770	0.776
in-generated (setting2)	0.690	0.653	0.672
cross-generated (setting4)	0.646	0.567	0.607
human evaluation on sample from test set			
gold (avg)	0.543	0.620	0.582
in-generated (setting1) (avg)	0.600	0.527	0.563
in-generated (setting2) (avg)	0.607	0.530	0.569

Table 5: Results for the different experiments.

The results reported in Table 4 show that human annotators definitely do not perform well at distinguishing the gold headlines, not much above the 50% baseline. Similar scores are observed in the assessment of the automatically generated headlines for both settings. Also the level of agreement (computed as the Krippendorff's Alpha Reliability) is very low for both gold and generated headlines, further indicating that human evaluations are not reliable for this task. To provide a few concrete examples, in Table 6 we show some gold and generated headlines together with their human and automatic evaluation.

Table 5 reports the results for all settings and the average of the performance of the three human evaluators for comparison.

Regarding **E1**, we indeed observe that for gold headlines the performance of the classifier is higher than for generated headlines, although for all the generated headlines the classifier performance is significantly higher than a random baseline. This suggests that the generators are able to intercept stylistic features and to generate text accordingly.

Also **E2** is confirmed by empirical results. For both generated and gold headlines of the test set we observe better performances when the classifier is trained on train-M, which is more similar than train-D to the test set, in terms of con-

[2]We did seek a collaboration with expert title creators for one of the two newspapers, as they are likely to have a different perception of the headlines, but received a negative response. We discuss this further in Section 6. in the context of future work.

example	generated	newspaper	human pred	machine pred
Usa - Cuba , Obama : " Bienvenido a Cuba " . E l' Avana accoglie tre giorni *en: [Usa - Cuba, Obama: "Bienvenido a Cuba". And Havana welcomes three days]*	Yes	rep	rep	gio
La verita su Twitter : " Macchina del fango " . Ma il Pdl è insorto *en: [The truth on Twitter: "Mud Machine". The PDL has arisen]*	Yes	gio	gio	gio
De Benedetti : " Riforma Popolari , tutta la storia di Pulcinella " . Il Pd : " Ne parlavano tutti " *en: [De Benedetti: "Populars reform, the whole story of Pulcinella", PD: "Everyone was talking about it"]*	Yes	rep	gio	rep
Rai verso le nomine per le reti : ecco i nomi *en: [Rai towards the nominations for the channels: here are the names]*	No	gio	gio	rep
Nasa , la Terra ha sette " sorelle " : scoperto un nuovo sistema planetario *en: [Nasa, the Earth has six "sisters": a new planetary system is discovered]*	No	rep	rep	rep
Vaccino antinfluenzale : ecco i cinque miti da sfatare *en: [Flu vaccine: here are the five myths to dispel]*	No	gio	rep	gio

Table 6: Examples of human and automatic evaluation of gold and generated headlines. The examples are randomly picked from any setting.

model	example
gold (rep)	Erdogan - Netanyahu , accuse durissime : " Israele come Hitler " , " No , tu sei un dittatore e stragista " (Erdogan - Netanyahu , very serious accusations : " Israel like Hitler " , " No , you are a dictator and mass killer ")
rep_D	Erdogan - Israele , la replica : " Israele e il Paese piu fascista "
rep_M	Israele , Netanyahu : " Israele e il Paese piu sionista , Hitler fascista fra i curdi "
gio_D2rep	Erdogan : " Premier razzista del mondo " Il piano di accuse per i curdi
gio_M2rep	Erdogan : " Il Paese piu sionista , razzista del mondo " . La replica araba
gold (gio)	Ecco le cellule hackerate per sconfiggere il cancro (Here are the hacked cells to defeat cancer)
gio_D	Il Mit di Boston : " Hackerare e riprogrammare le cellule per combattere il cancro "
gio_M	Hackerare le cellule per il cancro ' : ' riprogrammare il Dna '
rep_D2gio	Boston , ecco il codice genetico per combattere i tumori . " E ora un linguaggio "
rep_M2gio	Il Mit di un codice del Dna : cosi possibile hackerare le cellule sane e riprogrammarle

Table 7: Examples of generated headlines in the different settings for *la Repubblica* and *Il Giornale*

trolling for topic, (settings1-3). We also see a gap between validation and test performance in all settings, but smaller when the classifier is trained on train-M (**E3**).

Lastly, there is a drop in performance between in-generated and cross-generated headlines for both setting1-3 and setting2-4, although the performance on cross generated headline is still higher than the random baseline (matching **E4**). This goes to show that when a model trained on *la Repubblica* is asked to generate a headline starting from an *Il Giornale* article, it will do so preserving the style it has learnt from *la Repubblica*, in spite of having generated from the other newspaper's text.

As final evidence, we trained a newspaper-agnostic generator by mixing half of *La Repubblica* and half of *Il Giornale* from train-M (weakly aligned, closer to the test set than train-D), with a resulting size comparable to the other training sets (65k). By design, this model cannot learn any newspaper-specific style, and we therefore expect it to be unable to produce any newspaper-specific traits in generation. The measurable consequence of this is that the classifier should indeed not be able to distinguish them. A resulting average F1 score of 0.47, when compared to the scores in Table 5, is further proof that our models are indeed learning newspaper-specific style for headline generation.

For completeness, and to give an idea of the generated headlines we obtain using the various models, we report a few examples in Table 7. This shows two examples of headlines (one from *la Repubblica* and one from *Il Giornale*) with the automatically generated headlines versions in the different settings.

6. Conclusions

We trained a few pointer network models under different training settings that learnt to generate headlines according to a given newspaper's style, controlling for topic biases. We also trained a few classifiers that are able to distinguish the source of a given headline with high accuracy. Using such classification models as evaluators we were able to verify that the generators we have trained are indeed style-aware. This was confirmed through an additional experiment which showed that if the headlines are generated by a model trained in a newspaper-agnostic fashion, the classifier is indeed not able to distinguish them.

This whole evaluation procedure is done in a completely automated fashion. This is an advantage not only in terms of saving human effort, but especially because our experiments suggest that humans cannot perform this task reliably enough. An aspect to concentrate on in future work concerns the nature of the human judges who perform the evaluation. It would be desirable to collaborate with journalists, possibly title-creating experts from the specific newspapers we work with. Such experts should be better able than lay

people to spot and judge whether the generated style is appropriate for their own newspaper. At earlier stages of this work we did seek collaboration with one of the two papers we worked, but received a negative response. We still find this would be a valuable avenue to explore, and we plan to do it in the future. In any case, coupling generation and classification appears to be a successful evaluation methodology which we believe can be applied more generally, especially in absence of reliable human judgement.

Lastly, the data we have used for our experiments is part of a larger corpus that we have collected and that contains news articles from a large proportion of all Italian newspapers. The corpus is enriched with information about geographical provenance of the newspapers, density and amount of circulation, in addition to political positioning. The very same approach that we have described in this paper could therefore be applied to more data and to other dimensions of stylistic variation.

7. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baroni, M. and Bernardini, S. (2005). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Belz, A. and Kow, E. (2010). Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 7–15. Association for Computational Linguistics.

Cafagna, M., Mattei, L. D., Bacciu, D., and Nissim, M. (2019a). Suitable doesn't mean attractive. human-based evaluation of automatically generated headlines. In *CLiC-it*.

Cafagna, M., Mattei, L. D., and Nissim, M. (2019b). Embeddings shifts as proxies for different word use in italian newspapers. In *CLiC-it*.

Cahill, A. and Forst, M. (2009). Human evaluation of a german surface realisation ranker. In *Empirical methods in natural language generation*, pages 201–221. Springer.

Cimino, A., De Mattei, L., and Dell'Orletta, F. (2018). Multi-task learning in deep neural networks at evalita 2018. In *EVALITA@ CLiC-it*.

Dethlefs, N., Cuayáhuitl, H., Hastie, H., Rieser, V., and Lemon, O. (2014). Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 702–711.

Ficler, J. and Goldberg, Y. (2017). Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633*.

Flekova, L., Preoţiuc-Pietro, D., and Ungar, L. (2016). Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany, August. Association for Computational Linguistics.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

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

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Controllable text generation. *arXiv preprint arXiv:1703.00955*, 4.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. (2018). Disentangled representation learning for text style transfer. *arXiv preprint arXiv:1808.04339*.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Malmasi, S., Tetreault, J., and Dras, M. (2015). Oracle and human baselines for native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado, June. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., and Zweig, G. (2013). word2vec. *URL https://code. google. com/p/word2vec*.

Mikolov, T. (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80.

Potash, P., Romanov, A., and Rumshisky, A. (2015). Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Em-*

pirical Methods in Natural Language Processing, pages 1919–1924.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Reddy, S. and Knight, K. (2016). Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tian, Y., Hu, Z., and Yu, Z. (2018). Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Tikhonov, A. and Yamshchikov, I. P. (2018). Guess who? multilingual approach for the automated generation of author-stylized poetry. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 787–794. IEEE.

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., and Plank, B. (2018). Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia, July. Association for Computational Linguistics.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG'19)*, Tokyo, Japan. Association for Computational Linguistics.