On the interaction of automatic evaluation and task framing in headline style transfer

Lorenzo De Mattei[⋆◇†], **Michele Cafagna**[‡], **Huiyuan Lai**[†],
Felice Dell'Orletta[◇], **Malvina Nissim**[†], **Albert Gatt**[‡]
[⋆] Department of Computer Science, University of Pisa / Italy
[◇] ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa / Italy
[†] CLCG, University of Groningen / The Netherlands
[‡] LLT, University of Malta / Malta
lorenzo.dematteiei@di.unipi.it
{michele.cafagna,albert.gatt}@um.edu.mt
{h.lai,m.nissim}@rug.nl
felice.dellorletta@ilc.cnr.it

Abstract

An ongoing debate in the NLG community concerns the best way to evaluate systems, with human evaluation often being considered the most reliable method, compared to corpus-based metrics. However, tasks involving subtle textual differences, such as style transfer, tend to be hard for humans to perform. In this paper, we propose an evaluation method for this task based on purposely-trained classifiers, showing that it better reflects system differences than traditional metrics such as BLEU and ROUGE.

1 Introduction and Background

The evaluation of Natural Language Generation (NLG) systems is intrinsically complex. This is in part due to the virtually open-ended range of possible ways of expressing content, making it difficult to determine a 'gold standard' or 'ground truth'. As a result, there has been growing scepticism in the field surrounding the validity of corpus-based metrics, primarily because of their weak or highly variable correlations with human judgments (Reiter and Sripada, 2002; Reiter and Belz, 2009; Reiter, 2018; Celikyilmaz et al., 2020). Human evaluation is generally viewed as the most desirable method to assess generated text (Novikova et al., 2018; van der Lee et al., 2019). In their recent comprehensive survey on the evaluation of NLG systems, Celikyilmaz et al. (2020) stress that it is important that any used untrained automatic measure (such as BLEU, ROUGE, METEOR, etc) correlates well with human judgements.

At the same time, human evaluation also presents its challenges and there have been calls for the development of new, more reliable metrics (Novikova et al., 2017). Beyond the costs associated with using humans in the loop during development, it also appears that certain linguistic judgment tasks are hard for humans to perform reliably. For instance, human judges show relatively low agreement in the presence of syntactic variation (Cahill and Forst, 2009). By the same token, Dethlefs et al. (2014) observe at best moderate correlations between human raters on stylistic dimensions such as politeness, colloquialism and naturalness.

Closer to the concerns of the present work, it has recently been shown that humans find it difficult to identify subtle stylistic differences between texts. De Mattei et al. (2020b) presented three independent judges with headlines from two Italian newspapers with distinct ideological leanings and in-house editorial styles. When asked to classify the headlines according to which newspaper they thought they came from, all three annotators performed the task with low accuracy (ranging from 57% to 62%). Furthermore, agreement was very low (Krippendorff's $\alpha = 0.16$). Agreement was similarly low on classifying automatically generated headlines ($\alpha = 0.13$ or 0.14 for two different generation settings). These results suggest that human evaluation is not viable, or at least not sufficient, for this task.

In this work we focus on the same style-transfer task using headlines from newspapers in Italian, but address the question of whether a series of classifiers that monitor both style strength as well as content preservation, the core aspects of style transfer (Fu et al., 2018; Mir et al., 2019; Luo et al.,

Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pages 38–43,
Online (Dublin, Ireland), December 2020.

2019), can shed light on differences between models.

We also add some untrained automatic metrics for evaluation. As observed above, the fact that humans cannot perform this task reliably makes it impossible to choose such metrics based on good correlations with human judgement (Celikyilmaz et al., 2020). Therefore, relying on previous work, we compare the insights gained from our classifiers with those obtained from BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), since they are commonly used metrics to assess performance for content preservation and summarisation. Other common metrics such as METEOR (Banerjee and Lavie, 2005) and BLEURT (Sellam et al., 2020), which in principle would be desirable to use, are not applicable to our use case as they require resources not available for Italian.

More specifically, we train a classifier which, given a headline coming from one of two newspapers with distinct ideological leanings and in-house styles, can identify the provenance of the headline with high accuracy. We use this (the 'main' classifier) to evaluate the success of a model in regenerating a headline from one newspaper, in the style of the other. We add two further consistency checks, both of which aim at content assessment, and are carried out using additional classifiers trained for the purpose: (a) a model's output headline should still be compatible in content with the original headline; (b) the output headline should also be compatible in content with the article to which it pertains. A headline is deemed to be (re)generated successfully in a different style if both (a) and (b) are satisfied, and the main classifier's decision as to its provenance should be reversed, relative to its decision on the original headline.

A core element in our setup is testing our evaluation classifiers/strategies in different scenarios that arise from different ways of framing the style transfer task, and different degrees of data availability. Indeed, we frame the task either as a translation problem, where a headline is rewritten in the target style or as a summarisation problem, where the target headline is generated starting from the source article, using a summarisation model trained on target style. The two settings differ in their needs in terms of training data as well as in their ability to perform the two core aspects of style transfer (style strength and content preservation).

We observe how evaluation is affected by the different settings, and how this should be taken into account when deciding what the best model is.

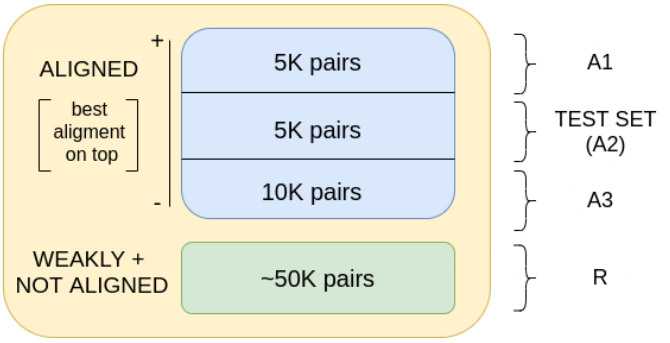
Data and code used for this paper are available at https://github.com/michelecafagna26/CHANGE-IT. The data and task settings also lend themselves well as material for a shared task, and they have indeed been used, with the summarisation system described here as baseline, in the context of the EVALITA 2020 campaign for Italian NLP (De Mattei et al., 2020a).

2 Task and Data

Our style transfer task can be seen as a "headline translation" problem. Given a collection of headlines from two newspapers at opposite ends of the political spectrum, the task is to change all rightwing headlines to headlines with a leftwing style, and all leftwing headlines to headlines with a rightwing style, while preserving content. We focus on Italian in this contribution, but the methodology we propose is obviously applicable to any language for which data is available.

Collection We used a corpus of around 152K article-headline pairs from two wide circulation Italian newspapers at opposite ends of the political spectrum namely *la Repubblica* (left-wing) and *Il Giornale* (right-wing) provided by De Mattei et al. (2020b). The data is balanced across the two sources. Though we are concerned with headlines, full articles are used in two ways: (a) *alignment*; and (b) the consistency check classifiers (see Section 4 for details). For the former, we leverage the alignment procedure proposed by Cafagna et al. (2019) and we split our dataset into strongly aligned, weakly aligned and non-aligned news. The purpose of alignment is to control for potential topic biases in the two newspapers so as to better disentangle newspaper-specific style. Additionally, this information is useful in the creation of our datasets, specifically as it addresses the need for parallel data for our evaluation classifiers and the translation-based model (see below).

Alignment We compute the tf-idf vectors of all the articles of both newspapers and create subsets of relevant news filtering by date, i.e. considering only news which were published approximately within the same, short time interval for the two sources. On the tf-idf vectors we then compute cosine similarities for all news in the resulting subset, rank them, and retain only the alignments that are



(a) Data splits

training sets		
EVAL	main	R+A3+A1
	HH	A1 + random pairs
	AH	R+A3+A1
TASK	SUM	R+A3
	S2S1	A3 (10K)
	S2S2	A3+A1 (15K)
	S2S3	A1 (5K)

(b) Training sets

Figure 1: Data splits and their use in the different training sets

above a certain threshold. The threshold is chosen taking into consideration a trade-off between number of documents and quality of alignment. We choose two different thresholds: one is stricter (> 0.5) and we use it to select the best alignments; the other one is looser (> 0.185, and $<= 0.5$).

Data splitting We split the dataset into *strongly aligned news*, which are selected using the stricter threshold (∼20K aligned pairs), and *weakly aligned and non-aligned news* (∼100K article-headline pairs equally distributed among the two newspapers). The aligned data is further split as shown in Figure 1a. SA is left aside and used as test set for the final style transfer task. The remaining three sets are used for training the evaluation classifiers and the models for the target task in various combinations. These are described in Figure 1b and in connection with the systems' descriptions.[1]

3 Systems

Our focus is on the interaction of different evaluation settings and approaches to the task. Accordingly, we develop two different frameworks with different takes on the same problem: (a) as a true translation task, where given a headline in one style, the model learns to generate a new headline in the target style; (b) as a summarisation task, where headlines are viewed as an extreme case of summarisation and generated from the article. We exploit article-headline generators trained on opposite sources to do the transfer. This approach does not in principle require parallel data for training.

For the translation approach (S2S), we train a supervised BiLSTM sequence-to-sequence model with attention from OpenNMT (Klein et al., 2017) to map the headline from left-wing to right-wing, and viceversa. Since the model needs parallel data, we exploit the aligned headlines for training. We experiment with three differently composed training sets, varying not only in size, but also in the strength of the alignment, as shown in Figure 1b.

For the summarisation approach (SUM), we use two pointer-generator networks (See et al., 2017), which include a *pointing mechanism* able to copy words from the source as well as pick them from a fixed vocabulary, thereby allowing better handling of out-of-vocabulary words. ability to reproduce novel words. One model is trained on the *la Repubblica* portion of the training set, the other on *Il Giornale*. In a style transfer setting we use these models as follows: Given a headline from *Il Giornale*, for example, the model trained on *la Repubblica* can be run over the corresponding article from *Il Giornale* to generate a headline in the style of *la Repubblica*, and vice versa. To train the models we use subset R, but we also include the lower end of the aligned pairs (A3), see Figure 1b.

4 Evaluation

Our fully automatic strategy is based on a series of classifiers to assess style strength and content preservation. For style, we train a single classifier (*main*). For content, we train two classifiers that perform two 'consitency checks': one ensures that the two headlines (original and transformed) are still compatible (*HH classifier*); the other ensures that the headline is still compatible with the original article (*AH classifier*). See also Figure 1a.

In what follows we describe these classifiers in more detail. When discussing results, we will show how the contribution of each classifier is crucial towards a comprehensive evaluation.

[1] Note that all sets also always contain the headlines' respective full articles, though these are not necessarily used.

Main classifier The main classifier uses a pre-trained BERT encoder with a linear classifier on top fine-tuned with a batch size of 256 and sequences truncated at 32 tokens for 6 epochs with learning rate 1e-05. Given a headline, this classifier can distinguish the two sources with an f-score of approximately 80% (see Table 1). Since style transfer is deemed successful if the original style is lost in favour of the target style, we use this classifier to assess how many times a style transfer system manages to reverse the main classifier's decisions.

HH classifier This classifier checks compatibility between the original and the generated headline. We use the same architecture as for the main classifier with a slightly different configuration: max. sequence length of 64 tokens, batch size of 128 for 2 epochs (early-stopped), with learning rate 1e-05. Being trained on strictly aligned data as positive instances (`A1`), with a corresponding amount of random pairs as negative instances, it should learn whether two headlines describe the same content or not. Performance on gold data is .96 (Table 1).

AH classifier This classifier performs yet another content-related check. It takes a headline and its corresponding article, and tells whether the headline is appropriate for the article.The classifier is trained on article-headline pairs from both the strongly aligned and the weakly and non-aligned instances (`R+A3+A1`, Figure 1b). At test time, the generated headline is checked for compatibility against the source article. We use the same base model as for the main and HH classifiers with batch size of 8, same learning rate and 6 epochs. Performance on gold data is >.97 (Table 1).

		prec	rec	f-score
main	**rep**	0.77	0.83	0.80
	gio	0.84	0.78	0.81
HH	**match**	0.98	0.95	0.96
	no match	0.95	0.98	0.96
AH	**match**	0.96	0.99	0.98
	no match	0.99	0.96	0.97

Table 1: Performance of the classifiers on gold data.

Overall compliancy We calculate a compliancy score which assesses the proportion of times the following three outcomes are successful (i) the *HH classifier* predicts 'match'; (ii) the *AH classifier* predicts 'match'; (iii) the *main classifier*'s decision is *reversed*. As upperbound, we find the compatibility score for gold at 74.3% for transfer from *La Repubblica* to *Il Giornale* (*rep2gio*), and 78.1% for the opposite direction (*gio2rep*).

5 Results and Discussion

Table 2 reports results of our evaluation methods both for the summarization system (`SUM`) and for the style transfer systems (S2S) in the different training set scenarios.

The top panel in Table 2 shows the results for systems where training data is weakly aligned or unaligned. The summarisation system `SUM` does better at content preservation (HH and AH) than `S2S1`. However, its scores on the *main* classifier are worse in both transfer directions, as well as on average. The average compliancy score is higher for `S2S1`. In summary, for data which is not strongly aligned, our methods suggest that style transfer is better when conceived as a translation task. BLEU is higher for `SUM`, but the overall extremely low scores across the board suggest that it might not be a very informative metric for this setup, although commonly used to assess content preservation in style transfer (Rao and Tetreault, 2018). Our HH and AH classifiers appear more indicative in this respect, and ROUGE scores seem to correlate a bit more with them, when compared to BLEU. It remains to be investigated whether BLEU, ROUGE, and our content-checking classifiers do in fact measure something similar or not.

With better-aligned data (bottom panel), the picture is more nuanced. Here, the main comparison is between two systems trained on strongly aligned data, one of which (`S2S2`) has additional, weakly aligned data. The overall compliancy score suggests that this improves style transfer (and this system is also the top performing one over all, also outperforming `S2S1` and `SUM`). As for content preservation (AH and HH scores), `S2S3` is marginally better on average for HH, but not for AH, where the two systems are tied.

Overall, the results of the classification-based evaluation also highlight a difference between a summarisation-based system (`SUM`), which tends to be better at content preservation, compared to a translation-based style transfer setup (especially `S2S2`) which transfers style better. Clearly, a corpus-based metric such as BLEU fails to capture these distinctions, but here does not appear informative even just for assessing content preservation.

		HH	AH	Main	Compl.	BLEU	ROUGE
without top aligned data							
SUM	**rep2gio**	.649	.876	.799	.449	.020	.145
	gio2rep	.639	.871	.435	.240	.026	.156
	avg	**.644**	**.874**	.616	.345	**.023**	**.151**
S2S1	**rep2gio**	.632	.842	.815	.436	.011	.136
	gio2rep	.444	.846	.864	.321	.012	.130
	avg	.538	.844	**.840**	**.379**	.012	.133
with top aligned data							
S2S2	**rep2gio**	.860	.845	.845	.549	.018	.159
	gio2rep	.612	.846	.847	.442	.016	.151
	avg	.736	**.846**	**.849**	**.496**	**.017**	**.155**
S2S3	**rep2gio**	.728	.844	.845	.520	.012	.139
	gio2rep	.760	.848	.649	.420	.013	.156
	avg	**.744**	**.846**	.747	.470	.013	.148

Table 2: Performance on test data.

One aspect that will require further investigation, since we do not have a clear explanation for it as of now, is the performance difference between the two translation directions. Indeed, transforming a *La Repubblica* headline into a *Il Giornale* headline appears more difficult than transforming headlines in the opposite directions, under most settings.

6 Conclusions

This paper addressed the issue of how to evaluate style transfer. We explicitly compared systems in terms of the extent to which they preserve content, and their success at transferring style. The latter is known to be hard for humans to evaluate (Dethlefs et al., 2014; De Mattei et al., 2020b). Our aim was primarily to see to what extent different evaluation strategies based on purposely trained classifiers could distinguish between models, insofar as they perform better at either of these tasks and in different training scenarios.

Our findings suggest that our proposed combination of classifiers focused on both content and style transfer can potentially help to distinguish models in terms of their strengths. Interestingly, a commonly used metric such as BLEU does not seem to be informative in our experiments, not even for the content preservation aspects.

To the extent that stylistic distinctions remain hard for humans to evaluate in setups such as the one used here, a classification-based approach with consistency checks for content preservation is a promising way forward, especially to support development in a relatively cheap and effective way.

Future work will have to determine how the various metrics we have used relate to each other (especially our classifiers and BLEU/ROUGE), and whether human judgement can be successfully brought back, and in case in what form, at some stage of the evaluation process.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. ACL.

Michele Cafagna, Lorenzo De Mattei, and Malvina Nissim. 2019. Embeddings shifts as proxies for different word use in italian newspapers. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), Bari, Italy*.

Aoife Cahill and Martin Forst. 2009. Human evaluation of a German surface realisation ranker. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 112–120, Athens, Greece. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv 2006.14799*.

Lorenzo De Mattei, Michele Cafagana, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020a. CHANGE-IT @ EVALITA 2020: Change Headlines, Adapt News, GEnerate. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020b. Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France. European Language Resources Association.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 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, Gothenburg, Sweden. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirtieth Conference on Innovative Applications of Artificial Intelligence (IAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 97–104, Harriman, New York, USA. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 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, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.