CHANGE-IT @ EVALITA 2020: Change Headlines, Adapt News, GEnerate

Lorenzo De Mattei
University of Pisa
CLCG, University of Groningen
ItaliaNLP Lab, ILC-CNR
Pisa, Italy
lorenzo.demattei@di.unipi.it

Michele Cafagna
Aptus.AI, Pisa, Italy
University of Malta, Malta
michele@aptus.ai

Felice Dell'Orletta
ItaliaNLP Lab, ILC-CNR
Pisa, Italy
felice.dellorletta@ilc.cnr.it

Malvina Nissim
CLCG, University of Groningen
The Netherlands
m.nissim@rug.nl

Albert Gatt
University of Malta
Malta
albert.gatt@um.edu.mt

Abstract

We propose a generation task for Italian – more specifically, a style transfer task for headlines of Italian newspapers. This is the first shared task on generation included in the EVALITA evaluation framework. Indeed, one of the reasons to have this task is to stimulate more research on generation within the Italian community. With this aim in mind, we release to the participating teams not only training data, but also a baseline sequence to sequence model that performs the task in order to help everyone get started, even when not accustomed to Natural Language Generation (NLG) approaches. Contextually, we explore the complex issue of automatic evaluation of generated text, which is receiving particular attention in the NLG community.

1 Task and Motivation

We propose a generation task for Italian in the context of the EVALITA 2020 campaign (Basile et al., 2020). More specifically, we design a *style transfer task for headlines of Italian newspapers*.

We believe it is the first time that a shared task on generation is offered in the context of EVALITA. Indeed, one of the reasons to have this task is to stimulate more research on generation within the Italian community. With this goal in mind, we release to the potential participating teams not only training data, but also a baseline sequence to sequence model that performs the task in order to help everyone get started, even when not accustomed to generation models, yet. This baseline model casts the style transfer problem as an extreme summarisation task, just showing how versatile the problem is in terms of possible approaches. Contextually, this task will help to further explore the complex issue of evaluation of generated text, which is receiving particular attention in the Natural Language Generation international community (Gatt and Krahmer, 2018; van der Lee et al., 2019).

Task The task is cast as a "headline translation" problem, and it is as follows. Given a collection of headlines from two Italian newspapers at opposite ends of the political spectrum, call them G and R, change all G-headlines to headlines into style R, and all R-headlines to headlines in style G.

In the context of this task we need to take care of two crucial aspects: data and evaluation. Details on data are provided in Section 2, and on evaluation in Section 3.

2 Data

We have collected news coming from two of the most important Italian newspapers situated at opposite ends of the political spectrum, namely *la Repubblica* (left) and *Il Giornale* (right), totalling approximately 152,000 article-headline pairs, with the two newspapers equally represented. Although the task only concerns headline change, the teams will receive both the headlines as well as their respective full articles.

Leveraging on an alignment procedure described below (see Cafagna et al. (2019) for fur-

cosine score	newspaper	alignment
0.96 (strict)	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 1: Example of alignments between *La Repubblica* and *Il Giornale*, extracted with different similarity scores. The second and the third examples would fall into the strict and the loose sets, respectively, according to the thresholds used to split the alignments. The first two headline pairs are well aligned, while the third pair has a very loose alignment.

ther details), we account for potential topic biases in the two newspapers, and we split the data set into strongly, weakly and not-aligned news. This information is useful in the creation of the datasets that we need to train our three evaluation classifiers (see Section 3). Additionally, it could help to better disentangle newspaper-specific style.

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 in approximately the same, short, temporal range 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 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 best alignments (*strict alignments*); the other one is looser (≥ 0.185, and < 0.5) — we define these latter as *weak alignments*. We consider the rest as basically not aligned.

Data splits We split the dataset into *strongly aligned news*, which are selected using the stricter threshold (∼20K aligned pairs, set `A*` in Figure 1a), and *weakly aligned and non-aligned news* (∼100K article-headline pairs equally distributed among the two newspapers, set `R` in Figure 1a).

The strictly aligned data is further split as shown in Figure 1a; this yields a total of four sets over the whole dataset (`A1`, `A2`, `A3`, and `R`). `A2` 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 system for the target task. These are shown in Figure 1b. Note that all sets also always contain the headlines' respective full articles, though these are not necessarily used.

Format The data is distributed in the form of *one CSV file* with the following fields:

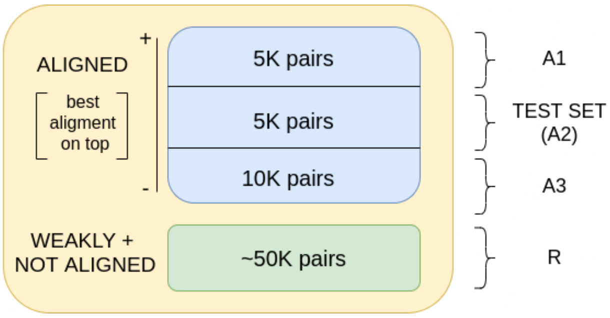
```
id, headline, article, label [R,G]
```

3 Evaluation

Human evaluation is generally viewed as the most desirable method to assess generated text (Novikova et al., 2018; van der Lee et al., 2019). However, human evaluation is not always a viable option, due to resources, but also due to the fact that humans might not be capable of reliably assessing the task at hand. Related to the current challenge, De Mattei et al. (2020a) have shown that people find it difficult to identify subtle stylistic differences between texts.

Automatic, reliable metrics should therefore also be sought (Novikova et al., 2017). For our task, we propose a fully automatic strategy 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 'sanity 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 1b.

In what follows we describe these classifiers in



(a) Overall data splits

EVALUATION		
train & test	main	R+A3+A1
	HH	A1 + random pairs
	AH	R+A3+A1
TASK		
train		R+A3
test		A2

(b) Training/test sets

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

more detail. When discussing baseline results, we will show how the contribution of each classifier is crucial towards a comprehensive evaluation.

Main classifier The main classifier uses a pre-trained BERT (Devlin et al., 2019) 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 2). 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 2).

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 2).

		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 2: Performance of the evaluation 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*).

4 Baseline System

We developed a baseline system using a summarisation approach, 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, as done in (De Mattei et al., 2020b). The advantage of this approach is that in principle it doesn't require parallel data for training.

Specifically, we use two pointer-generator networks (See et al., 2017), which include a *pointing mechanism* able to copy words from the

Il Giornale → La Repubblica		
E in Sicilia è scattata l’allerta rossa *[en: And in Sicily it’s now red alert]*	⟶	Migranti, la Protezione civile continua dimenticata *[en: Migrants, the Civil Protection Department goes on forgotten]*
Nozze gay, toghe contro i sindaci: ”Le trascrizioni sono illegittime” *[en: Gay marriages, gowns against mayors: “Transcriptions are not valid”]*	⟶	Il Consiglio di Stato boccia le nozze gay all’estero *[en: The State Council rejects gay marriages abroad]*
La Repubblica → Il Giornale		
Castelnuovo, lo sdegno di cittadini e associazioni: ”Attacco all’integrazione che funziona” *[en: Castelnuovo, the indignation of citizens and associations: “Attack to the integration that works”]*	⟶	I migranti non sono più rifugiati *[en: Migrants are not refugees anymore]*
Da Renzi a Di Maio, ecco il reddito dichiarato dai politici italiani. Fedeli il ministro con l’imponibile più alto *[en: From Renzi to Di Maio: here it’s the income declared by the Italian politicians. Fedeli is the minister with the highest taxable income]*	⟶	Grillo e Giggino italiani conquistano l’elenco dei redditi italiani *[en: Grillo and Giggino Italians conquer the list of Italian incomes]*

Table 3: Examples of headlines generated by the baseline system.

source as well as pick them from a fixed vocabulary, thereby allowing better handling of out-of-vocabulary 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.

The results of the baseline system, measured as performance of each classifier as well as the overall compliancy score, are reported in Table 4.

	HH	**AH**	**Main**	**compl.**
rep2gio	.649	.876	.799	.449
gio2rep	.639	.871	.435	.240
avg	**.644**	**.874**	.616	.345

Table 4: Baseline performance on test data.

5 Outlook

This shared task proposal was intended to stimulate research in NLG, with a specific focus on style transfer and automatic evaluation, in the Italian community. Over ten teams expressed their interest in participating in the shared task officially, but eventually there were no submitted runs. We do hope that the materials developed in the context of this challenge will nevertheless be of use to promote research in a field that is still under-researched in the Italian NLP landscape. All materials are available: `https://github.com/michelecafagna26/CHANGE-IT`.

References

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

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*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020a. Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020b. Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Albert Gatt and Emiel Krahmer. 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.

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*, pages 2241–2252, Copenhagen, Denmark, September. 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, June. 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.

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, October–November. Association for Computational Linguistics.