University of Milano-Bicocca
Department of Informatics, Systems and Communication
Master of Science in Computer Science

Bridging a GAP: Text Style Transfer from Journalistic to Conversational for enhanced social media dissemination of news

Supervisor
Gabriella Pasi

Dissertation by
Mattia Piazzalunga
851931

Academic Year 2023/2024

Summary

In 2023, the Digital News Report by the Reuters Institute for the Study of Journalism [13] provided clear evidence of the crisis overwhelming the news market. The main reason is the public's declining interest in staying informed, fueled by news agencies' difficulties in transitioning to a digital media environment; a shift driven by younger generations who rely almost entirely on digital media for accessing content. However, if properly leveraged, these platforms can prove to be a strategic resource: social networks, in particular, are used by 93.2% of users according to the Digital 2024 Global Overview Report [18], and 30% of people prefer staying informed through them, thanks to their more personalized and simplified approach to content sharing. In this landscape, research cannot remain indifferent: it must support the information market to help it embrace this media change. In the thesis titled "Bridging a GAP: Text Style Transfer from Journalistic to Conversational for Enhanced Social Media Dissemination of News", which emerged from an idea developed in collaboration with ANSA.it, Italy's leading news agency, the use of machine learning (ML) has been explored to reduce the high costs associated with transcribing news into a simplified format for social networks, allowing for easier consumption of content on these platforms and, by extension, supporting the news industry.

But how can all this be achieved? First of all, as highlighted, the process involves Natural Language Processing (NLP), a subfield of ML, in an attempt to modify the target style of the text, shifting from journalistic (standard) to conversational (social networks). An important definition of "style" is provided by McDonald et al. [11], who describe it as the way in which semantics are expressed, aiming, as emphasized by Hovy [6], to produce a stronger effect on the recipient of the discourse in text generation. To change the style of a text, however, it is necessary to engage both Natural Language Understanding (NLU), the area of NLP concerned with assigning meaning to sentences, and Natural Language Generation (NLG), which pertains to the generation of text. The relevant downstream task in the literature is, specifically, Text Style Transfer (TST), a process that requires the highest level of linguistic understanding: pragmatics. Pragmatics studies the use of language in real-world contexts, considering not only semantics but also how it varies based on factors unrelated to word structure. This makes the task of style transfer particularly complex and challenging, to the point that it requires the use of state-of-the-art (SOTA) neural architectures.

The representational power of the modern transformer architecture, introduced by Vaswani et al. [19], can be very useful in addressing the complexity of TST. It is a type of neural network that leverages the self-attention mechanism to assign meaning to a word in the text, given an arbitrarily large context, drawing on the intuition of Harris's distributional hypothesis [4]. Transformers have been proven to be particularly effective in the literature, achieving state-of-the-art results in numerous tasks. In the TST task, specifically, it is necessary to leverage the encoder-decoder architecture. The encoder can be thought of as the component performing the NLU task, extracting the informational essence of the journalistic text input and representing it in a latent space. The decoder, on the other hand, functions as a text generator, executing the NLG component. However, it is not possible to leverage the linguistic definition of style to work with ML; it is necessary to adopt a data-driven interpretation, as suggested by Mou et al. [12]. Specifically, given two corpora of text characterized by two

different styles, the invariance between the two represents the content, or the informational essence, while the variation between them constitutes the style; transformers must, therefore, learn to "play" with these two components.

It is essential to emphasize at this point that the work presented introduces a new downstream task for TST, not yet present in the literature, which can be named *Journalistic2Conversational TST* (J2C-TST). Given the novelty of this task, it was necessary to create two supporting datasets. The first, for the Italian language, was developed in collaboration with ANSA, consisting of 1.478 pairs of texts <journalistic style, conversational style>, taken respectively from the official website and the associated Facebook post. The second one, in English, is derived from a 2016 corpus [1] and includes 5,352 pairs of texts from 9 English news agencies, whose alignment, after a data-cleaning phase, was verified by analyzing the cosine similarity between the contextual embeddings of the texts in each pair, generated using an encoder proposed by Reimers et al. [17]. The definition of the task and the datasets represent the first significant contribution of the work undertaken during the thesis. In addition to supporting the J2C-TST task, the two corpora open new avenues for studying the two writing styles. For instance, in the thesis, the stylistic differences between journalistic and conversational styles were examined using three state-of-the-art models [2]. The results show that, with the same level of formality, subjectivity, and sentiment, the conversational style tends to be associated with shorter texts aimed at engaging the reader, focusing on essential information, and encouraging clicks to the news outlet's official website.

Starting from the two presented corpora, it was possible to train models for the transfer task. However, it is important to emphasize that not only were the pairs in the datasets limited in number, preventing the model from capturing general linguistic details useful for the task, but in general, training a model from scratch is very resource-intensive. To address this problem, a common practice is to use a foundational model, pre-trained on a vast amount of data in an unsupervised manner, and then adapt it to the specific task through fine-tuning. Pre-training allows the model to acquire a general knowledge base that is useful for various NLP tasks, especially for complex tasks such as stylistic transfer. After a careful review of the existing literature, the generalist models identified as most suitable for the task were T5 version 1.1 [16], for the english language, and mT5 [21], its multilingual version, for the Italian corpus.

What does fine-tuning mean? Fine-tuning involves adapting a pre-trained model to a specific task, allowing it to acquire targeted knowledge by updating its internal weights. However, given the vast number of parameters present in models like T5-base and mT5-base, this process can still be extremely computationally expensive. To address this issue, the dissertation adopts PEFT (Parameter-Efficient Fine-Tuning) using the LoRA method, introduced by Hu et al. [7]. LoRA is based on an observation by Aghajanyan et al. [1], which suggests that during adaptation to a specific task, foundational models can effectively learn through weight matrix projections into lower-dimensional spaces. So, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, its update is constrained by representing it with a low-rank decomposition $W_0 + \Delta W = W_0 + BA$ where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$

[1] `https://data.world/martinchek/2012-2016-facebook-posts`

[2] `https://huggingface.co/s-nlp/roberta-base-formality-ranker` - formality, `https://huggingface.co/GroNLP/mdebertav3-subjectivity-english` - subjectivity and `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest` - sentiment .

and the rank $r << \min(d, k)$. During training, W_0 is frozen and does not receive gradient updates, unlike A and B. Thanks to the use of LoRA, the number of trainable parameters has been significantly reduced: from 251.116.800 to 3.538.944 for T5 (1.41%) and from 585.940.224 to 3.538.994 for mT5 (0.60%). This has made the computational process much more efficient and in line with the concept of "green AI".

At this point, it is essential to emphasize that, in addition to the base models, their "small" versions were also trained in the dissertation to ensure the creation of resource-efficient neural networks, even in inference, while at the same time enabling the use of models in contexts with limited computational capacity. The goal is to reduce costs and time without compromising the accuracy of the style transfer. The TST models were evaluated according to three criteria, as defined by Jin et al. [8]:

- Transferred Style Strength. This criterion assesses whether the transferred style effectively corresponds to the conversational style. Since there is no suitable classifier for this task in the literature, two models were produced using the J2C datasets. For English, RoBERTa [10] was fine-tuned, while for Italian, XML-RoBERTa [2] was adapted. Both classifiers achieved accuracy and F1 scores above 99%, demonstrating, on the one hand, the presence of discriminating characteristics in the text of the two classes and, on the other hand, the classifiers' effectiveness in evaluating J2C-TST.
- Content Preservation. The effectiveness of the models in preserving content during style transfer was measured using standard metrics such as BLEU [14], METEOR [9], and BERT-score [22]. BLEU is the traditional metric for content preservation in machine translation, METEOR represents an improved version, and BERT-score leverages the power of contextual embeddings for more accurate evaluation.
- Fluency. The fluency of the generated texts was analyzed by calculating "perplexity", a measure of how well a model predicts the data, using SOTA pre-trained models specific to each language. In particular, GPT-2 [15] was used to evaluate the English version, while for the Italian version was employed GePpeTto [3], an Italian variant of GPT-2. Thanks to their exposure to large amounts of data, these models can learn syntactic structures, grammatical rules, semantic context, and statistical regularities of the language, thereby effectively assessing the fluency of the texts.

At first, the fine-tuned models did not show high accuracy with respect to the analyzed parameters. To improve the outputs, it was necessary to significantly customize the generation phase. By default, models generally use "greedy decoding" to produce text, a technique that generates sentences word by word, or rather token by token, always choosing the most probable option. Although this approach is fast, it tends to produce suboptimal results. To address this limitation, "beam search" was initially employed, which explores multiple paths simultaneously, leading to the first significant improvements. However, to achieve the best results, the Seq2SeqTrainer module from the Huggingface transformers library [20], which supported the use of state-of-the-art transformer architectures throughout the dissertation experiments, had to be directly modified. This customization allowed the incorporation of new generation parameters during the evaluation phase, enabling the production of high-quality responses by balancing creativity and coherence, while effectively controlling repetitions and ex-

cessive lengths. In addition to "beam search", the following were introduced: "maximum generation length", "sampling" to explore not only the most probable options combined with "nucleus sampling" to reduce the influence of less probable choices, and "temperature" to moderate randomness. Moreover, "length penalty" was used to penalize sequences that were too long, given the nature of conversational text, and "repetition penalty" to avoid unwanted repetitions.

In an attempt to reduce the accuracy gap between base models and small models, independently fine-tuned, the knowledge distillation technique proposed by Hinton et al. [5] was tested. Instead of performing a regular fine-tuning, the base models (teacher), already optimized for the J2C-TST task, guided the learning of the small models (student). To implement the distillation, it was necessary to modify the Seq2SeqTrainer class once again, customizing the loss function for training. The new loss function was created by combining two components: the "distillation loss", which measures the difference between the student's predictions and the teacher's "soft" predictions (non-binary and richer), and the traditional loss function based on the ground truth. Although this technique has indeed improved the performance of the T5-small model compared to independent fine-tuning, the results for mT5-small have remained unchanged. This outcome is not easily justifiable, but the multilingual pretraining of mT5-base could lead to less specialized representations for the italian language, likely making it more difficult for the student to extract relevant information.

A final experiment was conducted in an attempt to further improve performance. A data augmentation technique, specifically backtranslation, was employed. This technique generates new variants of a text by translating it into another language (German, in this case) and then translating it back into the original language, aiming to enrich the training datasets and potentially boost performance. The Italian dataset was augmented with 75% artificial pairs, while the English dataset, due to the greater availability of data, was enriched by 25%. However, the evaluation of the results on the test set showed no improvements, likely due to possible inconsistencies generated by the translations, which failed to capture linguistic nuances.

To summarize. The thesis, alternating between detailed theoretical and practical notions, has defined a new TST task in the literature, the J2C-TST, useful in supporting the free information market. Additionally, it presented two datasets, one in italian and one in english, to support the task, studying them and making them available for further analysis. Moreover, in defining accurate and resource-efficient J2C-TST models, alongside careful hyperparameter tuning and modifications to existing libraries, the validity of using SOTA compression and data augmentation techniques for the task was explored. Finally, six models are presented: two style classifiers (for Transferred Style Strength evaluation), two models for the Italian J2C-TST, and two for the English task. The J2C-TST models results are available in the table [3].

	BLEU	**METEOR**	**BERT-score**	**Fluency**	**% Conversational**
T5-v1_1-base-news-style-j2c-EN-v1	27.27	43.00	89.77%	50.64	100%
T5-v1_1-small-news-style-j2c-EN-v1	25.01	38.70	89.24%	57.66	100%
mT5-base-news-style-j2c-IT-v1	32.84	45.16	77.54%	56.24	100%
mT5-small-news-style-j2c-IT-v1	30.48	42.85	76.38%	55.28	100%

[3] BLEU $\in [20, 30]$ is acceptable, while BLEU $\in [30, 40]$ is good; METEOR $\in [20, 40]$ is acceptable, while METEOR $\in [40, 60]$ is good; BERT-score $\in [50, 70]$ is acceptable, while BERT-score $\in [70, 90]$ is good; Perplexity $\in [100, 50]$ is acceptable, while Perplexity $\in [50, 20]$ is good.

The results obtained, despite the differences detailed in the dissertation (which include aspects related to model architecture, training dataset, language, and parameters), demonstrate the effectiveness of the developed solutions. However, this represents only a starting point for further analyses aimed at improving the J2C task, as well as related areas such as machine translation.

As Brad Smith, president of Microsoft, stated: "Our future success depends on our ability to embrace artificial intelligence and use it for good". Contributing to the field of open information through AI, in my opinion, represents an excellent starting point. Therefore, I will extend these studies during my PhD at the University of Milan Bicocca.

The thesis was written adhering to the principles of Open Science, a set of practices that promote transparency and open sharing of the scientific process. For this reason, not only the results were provided, but also the models, code, and procedures necessary to replicate the experiments, thereby fostering greater transparency, scientific progress, and reducing duplications. This approach aligns with the philosophy of "Green AI", which aims for more sustainable development of artificial intelligence.

References

[1] Armen Aghajanyan et al. 'Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning'. In: ACL, 2021.

[2] Alexis Conneau et al. 'Unsupervised Cross-lingual Representation Learning at Scale'. In: ACL, 2020.

[3] Lorenzo De Mattei et al. 'GePpeTto Carves Italian into a Language Model'. In: *CoRR* (2020).

[4] Zellig S Harris. 'Distributional structure'. In: *Word* (1954).

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015.

[6] Eduard Hovy. 'Generating natural language under pragmatic constraints'. In: *Journal of Pragmatics* (1987).

[7] Edward J. Hu et al. 'LoRA: Low-Rank Adaptation of Large Language Models'. In: *CoRR* (2021).

[8] Di Jin et al. 'Deep Learning for Text Style Transfer: A Survey'. In: *Computational Linguistics* (2022).

[9] Alon Lavie et al. 'Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments'. In: ACL, 2007.

[10] Yinhan Liu et al. 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. In: *CoRR* (2019).

[11] David D. McDonald et al. 'A computational theory of prose style for natural language generation'. In: ACL, 1985.

[12] Lili Mou et al. 'Stylized Text Generation: Approaches and Applications'. In: ACL, July 2020.

[13] N Newman et al. *Digital news report 2023*. Tech. rep. 2023.

[14] Kishore Papineni et al. 'BLEU: a method for automatic evaluation of machine translation'. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL, 2002.

[15] Alec Radford et al. 'Language Models are Unsupervised Multitask Learners'. In: 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[16] Colin Raffel et al. 'Exploring the limits of transfer learning with a unified text-to-text transformer'. In: *J. Mach. Learn. Res.* (2020).

[17] Nils Reimers et al. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. In: *CoRR* (2019).

[18] We Are Social and Meltwater. *Digital 2024 Global Overview Report*. 2024.

[19] Ashish Vaswani et al. 'Attention is all you need'. In: Curran Associates Inc., 2017.

[20] Thomas Wolf et al. 'Transformers: State-of-the-Art Natural Language Processing'. In: ACL.

[21] Canwen Xu et al. 'Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression'. In: ACL, 2021.

[22] Tianyi Zhang et al. 'BERTScore: Evaluating Text Generation with BERT'. In: 2020.