**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

tutor
Gabriella Pasi

author
Mattia Piazzalunga

a. year
2023/2024

Access to accurate and quality information is a fundamental right.

**“The shaping of the mind or character”** [1]

Nonetheless, the news market is experiencing a crisis driven by the decline of public interest. [2]

**Problem**

The news agencies find it difficult to adapt to new digital media.

[1] Oxford English Dictionary, 2023. “Information”.
[2] Newmn et al., 2023. “Digital News Report”.

**36%** selectively avoid news. [2]

**- 63%** are interested in news compared to 2017. [2]

## Social networks are a resource

**93%** of internet users use socials. [3]

**New access points**

**-10%** of respondents inquire news via app/website compared to 2018. [2]

[2] Newmn et al., 2023. "Digital News Report".
[3] We Are Social et al., 2024. "Global Overview Report".

People prefer simplified and personalized access to content.

## Research can give a contribution

A first help is about lowering the costs associated with rewriting news in a conversational (social) style.

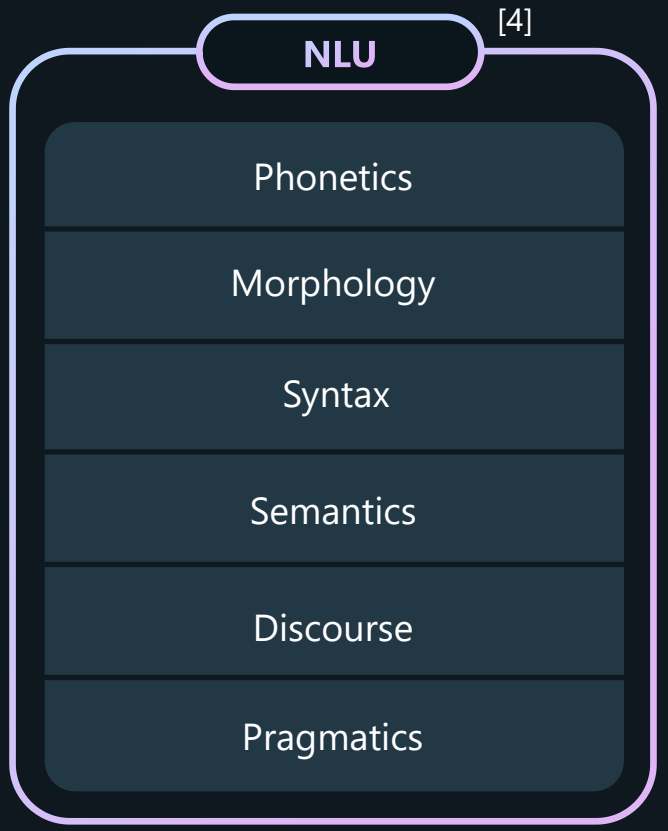
One can take advantage of the NLP to change the target style of the text.

**Collaboration**

Task idea born in collaboration with ANSA.it

**Open Science**

NLP is the combination of two sub-areas: Natural Language Understanding and Natural Language Generation.

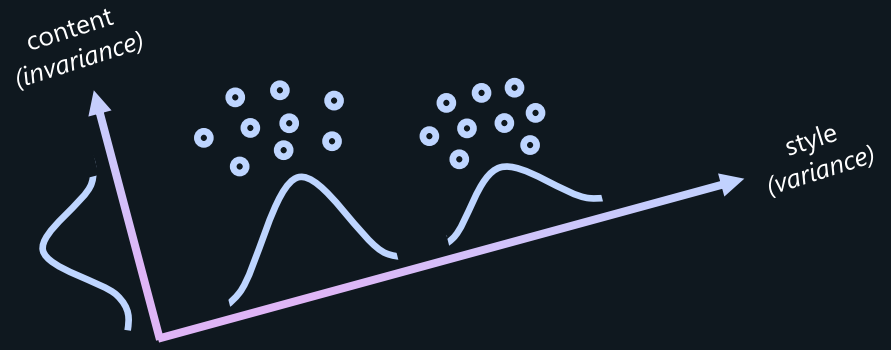



[4] Feldman, 1999. “NLP meets the Jabberwocky”.

Style is the way in which semantics are expressed [6], aiming to produce a stronger effect on the recipient of the discourse in text generation. [5]

## Text Style Transfer is the key

**Data-driven definition of style.** "given two corpora differentiated by their style, the invariance between them is the content, or the informative essence, while the variance is the style". [7]




[5] Hovy, 1987. "Generating natural language under pragmatic constraints".
[6] McDonald, 1985. "A computational theory of prose style for natural language generation".
[7] Mou et al., 2020. "Stylized Text Generation: Approaches and Applications".

There is no such downstream TST task in the literature.

**NEW TASK**

| INPUT | OUTPUT | CONDITIONS |
| --- | --- | --- |
| - $x_s = x_1^s ... x_n^s$ the source sentence with the style attribute $a_s$;<br>- $a_s$ = "journalistic";<br>- $a_t$ = "conversational" the desired style attribute. | - $x_t = x_1^t ... x_n^t$ the target sentence, with style attribute associated $a_t$. | - $a_t \neq a_s$ and therefore it is a stylistic transfer;<br>- $x_s$ and $x_t$ must share the content. |

**Journalistic to Conversational TST (J2C-TST)**

This task requires text pairs on which to train models
<journalistic_style, conversational_style>

## CORPUS IT

**Name: J2C_news_IT;**

**# text pairs:** 1.478;

**Language:** Italian;

**Creation methodology:** Created by aligning Facebook posts and original ANSA.it articles and removing the harmful noise.

Dataset created in collaboration with ANSA.it

**Certified by ANSA.it**

## CORPUS EN

**Name: J2C_news_EN;**

**# text pairs:** 5.352;

**Language:** English;

**Creation methodology:**

1. A 2016 dataset [8] of Facebook posts associated, via links, with the original post was exploited;
2. Leveraging publicly available data from 9 news outlets, alignments were created, removing noise;
4. The pairs were retained if the text embeddings of the two parts, generated by a Reimers model [9], had a cosine similarity ≥ 0,7.

**Data from 9 news agencies.**

[8] https://data.world/martinchek/2012-2016-facebook-posts.
[9] Reimers et al., 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks".

The corpora products have can also be useful in other tasks. An example is given here by exploiting J2C_news_EN.

## "Characterization of the styles involved"

**Composition**

| | # unique words | avg. text length |
|---|---|---|
| **jour** | 79.692 | 537 |
| **conv** | 16.979 | 27 |

**Subjectivity** [10]

| | # objective | # subjective |
|---|---|---|
| **jour** | 4.866 | 486 |
| **conv** | 4.646 | 706 |

[*] **10%** news becomes informal when switching to conversational style.

[-] **15y/o** school age level for understanding the text

**Ease of reading** [12] [-]

| | Flesch Reading Ease |
|---|---|
| **jour** | 60,44 |
| **conv** | 62,14 |

**Formality** [11] [*]

| | # formal | # informal |
|---|---|---|
| **jour** | 5.125 | 227 |
| **conv** | 4.679 | 673 |

**Sentiment** [13]

| | # positive | # neutral | # negative |
|---|---|---|---|
| **jour** | 683 | 3.576 | 1.094 |
| **conv** | 694 | 3.023 | 1.635 |

[10] GroNLP/mdebertav3-subjectivity-English.
[11] Babakov, et al., 2023. "A study on content preserving style transfer".
[12] Flesh, 1979. "How to write plain English : a book for lawyers and consumers".
[13] Loureiro, et al., 2022. "TimeLMs: Diachronic Language Models from Twitter".

## STYLES

| JOURNALISTIC | CONVERSATIONAL |
| --- | --- |
| Designed to offer in-depth coverage of events, providing detailed information that helps the reader understand the full context and develop critical and informed thinking. | It aim to engage the reader, often with a more personal tone, trying to elicit responses and comments. Its strength lies in brevity and the ability to focus on key information without details that might distract the reader. |

## + EXTRA

In a down market, do publishers follow the guidelines to increase engagement?

**54**/5352 articles follow SEO guidelines that recommend articles between 2100 and 2400 words. [14]

**12**/5352 Facebook posts stay under 50 characters that guarantee more interactions. [15]

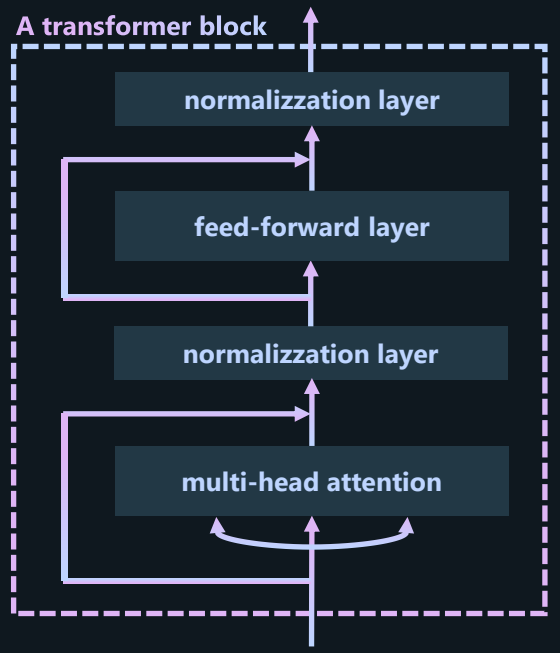
**0** /5352 text pairs follow both guidelines.

[14] https://blog.hubspot.com/marketing/how-long-should-your-blog-posts-be-faq.
[15] https://buzzsumo.com/blog/ultimate-guide-facebook-engagement-2017/.

Working at the pragmatic level, the Text Style Transfer Task can prove to be really complex

**Representational power of the modern transformer architecture can be very useful** [16]

It uses self-attention [16] to build contextual word representations by leveraging and managing distant information in sentences.



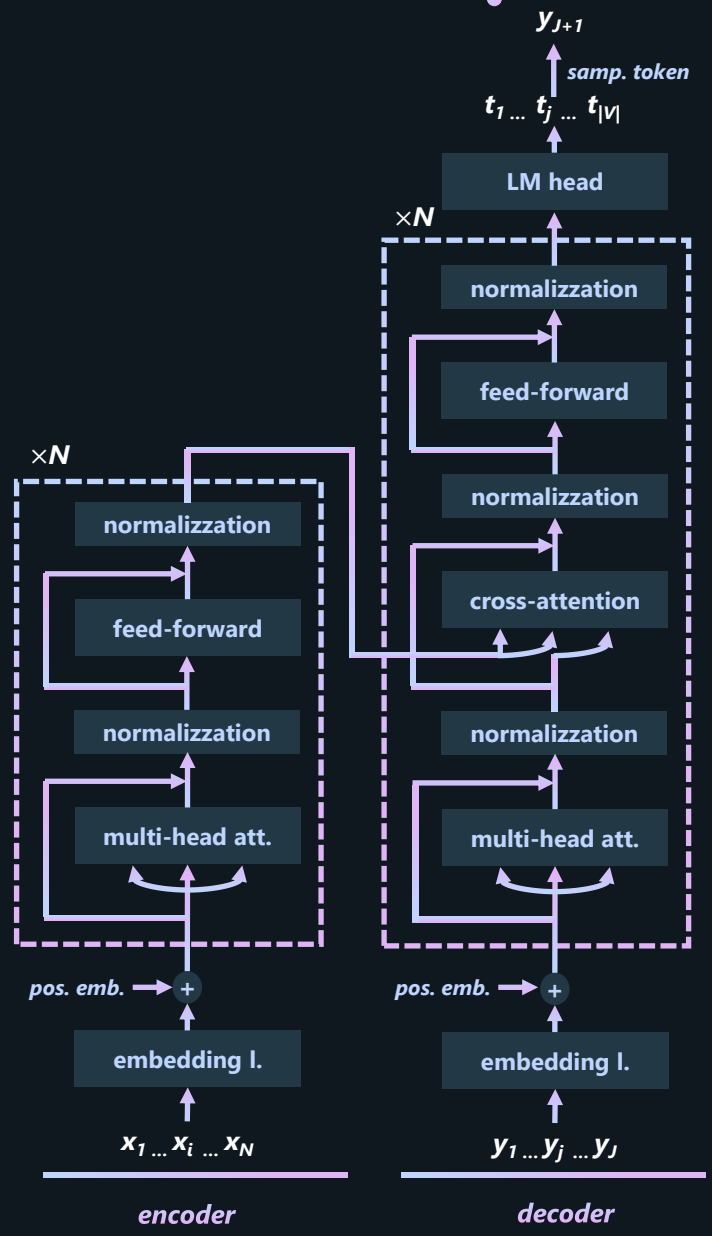

[16] Vaswani et al., 2017. "Attention is all you need".

In the TST task, specifically, it is necessary to leverage the **encoder-decoder architecture** [17].

- The encoder can be thought of as the component performing the NLU task.

- The decoder, on the other hand, functions as a text generator (NLG component).

The decoding block of the transformer is enriched with a special layer called "cross-attention" [16], which allows each position of the decoder to attent all positions of the input sequence.




[16] Vaswani et al., 2017. "Attention is all you need".
[17] Cho et al., 2014. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation".

Training a model from scratch is too resource-intensive and, in any case, the datasets are too small. Need to capture general language knowledge for a complex task like J2C-TST.

**FOUNDATIONAL MODELS**

| **T5 v1.1 (base)** [19] | **mT5 (base)** [18] |
|---|---|
| Language: english; | Language: +100; |
| Trained on: C4; | Trained on: mC4; |
| Type: T2T; | Type: T2T; |
| Parameters: 220 million | Parameters: 500 million |
| Extras: improved version of T5 | Extras: multilanguage version of T5 |

[18] Conneau et al., 2020. "Unsupervised Cross-lingual Representation Learning at Scale".
[19] Raffel et al., 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer".

To achieve high performance in a downstream task , the foundational model must be adapted, optimizing its weights. However, the training regime is very computationally expensive, with a strong environmental and economic impact. [22]

## LoRA (PEFT) is the key

**-98,59%** training parameters for T5;

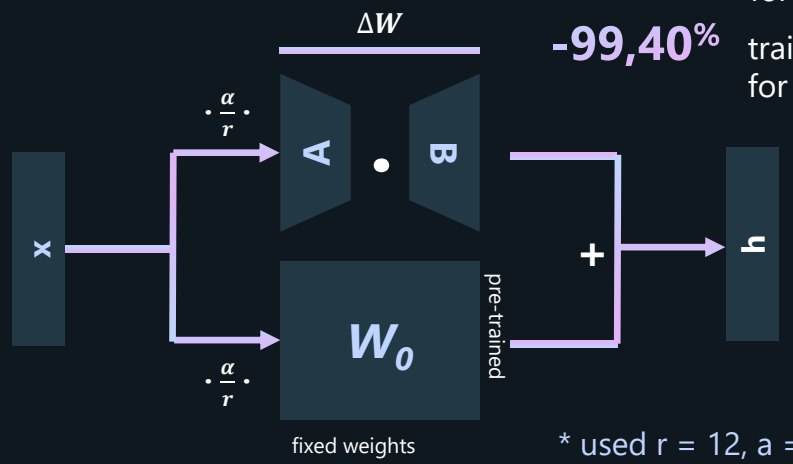
**-99,40%** training parameters for mT5.

Based on the observation that during adaptation to a specific task, foundational models can effectively learn through weight matrix projections into lower-dimensional spaces [20], LoRA [21] works like this:

Forward pass become:

$$h = W_0x + \frac{\alpha}{r}\Delta Wx = W_0x + \frac{\alpha}{r}BAx$$

with $W_0 \in R^{dxk}$, $B \in R^{dxr}$, $A \in R^{rxk}$ and $r \ll \min(d, k)$.




* used r = 12, a = 32 and dropout of 10%.

[20] Aghajanyan, et al., 2021. "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning".
[21] Hu et al., 2021. "LoRA: Low-Rank Adaptation of Large Language Models".
[22] Strubbel et al., 2019. "Energy and Policy Considerations for Deep Learning in NLP".

## EVALUATION [24]

### TRANSFERRED STYLE STRENGTH

Useful to assess whether the style after transfer is the target style.

A binary classifier can be useful for this purpose.

### CONTENT PRESERVATION

To assess how much of the main content is preserved.

Three metrics are used:
- BLEU [26] (standard) based on n-grams precision with GTs;
- METOR, an improved version of BLEU that consider recall [25];
- BERT-score [28] leverages the power of contextual embeddings for more accurate evaluation.

### FLUENCY

Analyzed by calculating "perplexity", a measure of how well a SOTA pre-trained model predicts the data.

Two models are used:
- GPT-2 [27] can be used to evaluate the english outputs;
- GePpeTto [23], an italian variant of GPT-2, employed for italian outpus.

In-vivo evaluation is expensive and not comparable, we are using an **intrinsic evaluation**

[23] De Mattei et al., 2020. "GePpeTto "Carves Italian into a Language Model". [24] Jin et al., 2022. "Deep Learning for Text Style Transfer: A Survey". [25] Lavie et al., 2007. "Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments". [26] Papineni et al., 2002. "BLEU: a method for automatic evaluation of machine translation". [27] Radford et al., 2019. "Language Models are Unsupervised Multitask Learners". [28] Zhang et al., 2020. "BERTScore: Evaluating Text Generation with BERT".

# The classifiers

## RoBERTa finetuned

| | precision | recall | F1-score |
|---|---|---|---|
| **jour** | 100% | 99% | 99% |
| **conv** | 99% | 100% | 99% |
| **accuracy** | | | 99% |

A classifier based on RoBERTa [30], a bidirectional encoder, was created on the English dataset.

Some hyperparameters and/or techniques: linear learning rate scheduler and an early -stopping with patience = 5.

## XLM-RoBERTa finetuned

| | precision | recall | F1-score |
|---|---|---|---|
| **jour** | 99% | 99% | 99% |
| **conv** | 99% | 99% | 99% |
| **accuracy** | | | 99% |

A classifier based on XLM-RoBERTa [29], a multilingual version of RoBERTa, was created on the Italian dataset.

[29] Conneau et al., 2020. "Unsupervised Cross-lingual Representation Learning at Scale".
[30] Liu et al., 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach".

At first, the fine-tuned models did not show high accuracy with respect to the analyzed parameters. It is necessary to change the generation, replacing greedy decoding.

## NEW PARAMETERS

*Seq2SeqTrainer class customization request. [31]

**Maximum output sequence length.** The generation length has been limited (50 tokens for T5, 64 for mT5).

**Sampling.** This parameter enables sampling-based generation. When set to true, the model selects tokens randomly.

**Top-p.** Set to 0,85: the model will only "select" from tokens that cumulatively represent 85% of the total probability.

**Beam search hypotheses.** The beams number has been set to 10.

**Temperature**. The temperature has been set to 0,8, which moderates the level of randomness.

**Length penalty.** A value of 0.85 penalizes long sequences.

**Repetition penalty.** This reduces the likelihood of repeating tokens already generated (1.5).

[31] Wolf et al., 2020. “Transformers: State-of-the-Art Natural Language Processing”

# First results

**BASE models**

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 base** | 27,27 | 43,00 | 89,77% | 50,64 | 100% | 251.116.800 | 3.538.944 |
| **mT5 base** | 32,84 | 45,16 | 77,44% | 56,24 | 100% | 585.940.224 | 3.538.944 |

**SMALL models**

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 small** | 24,53 | 36,45 | 88,84% | 65,59 | 100% | 78.337.408 | 1.376.256 |
| **mT5 small** | 30,48 | 42,85 | 76,38% | 55,28 | 100% | 301.553.024 | 1.376.256 |

Although there is no universal standard for evaluating the scores obtained, as they heavily depend on the specific task, considering the demands of style transfer and the complexity of the task:

- **BLEU** ∈ [20, 30] is acceptable, while BLEU ∈ [30, 40] is good;
- **METEOR** ∈ [20, 40] is acceptable, while METEOR ∈ [40, 60] is good;
- **BERT-score** ∈ [50, 70] is acceptable, while BERT-score ∈ [70, 90] is good;
- **Perplexity** ∈ [100, 50] is acceptable, while Perplexity ∈ [50, 20] is good.

Although no human evaluation has been conducted, at least for now, one might still wonder whether they are truly functioning as expected.

**T5-v1_1-base-news-style-J2C-EN-v1**

**Predicted:** «ly pardoned Robert Downey Jr. on Thursday for a nearly 20- year-old felony drug conviction that led to the Oscar-nominated actor's imprisonment for roughly a year.»

**Ground Truth:** «California's governor has pardoned Robert Downey Jr. for a drug conviction that sent the 'Iron Man' actor to prison.»

**mT5-base-news-style-J2C-IT-v1**

**Predicted:** «Un grave incidente stradale si è verificato nella tarda mattinata sull'autostrada A23, nel tratto tra Gemona e Carnia, in direzione Tarvisio. Nell'incidente una ventina di persone sarebbero rimaste ferite, alcune in modo grave.»

**Ground Truth:** «Un grave incidente stradale si è verificato nella tarda mattinata sull'autostrada A23, nel tratto tra Gemona e Carnia, in direzione Austria. Nell'incidente una ventina di persone sarebbero rimaste ferite, alcune in modo grave. Lo scontro ha coinvolto numerosi veicoli. #ANSA.»

ON THESIS
**106, 114**

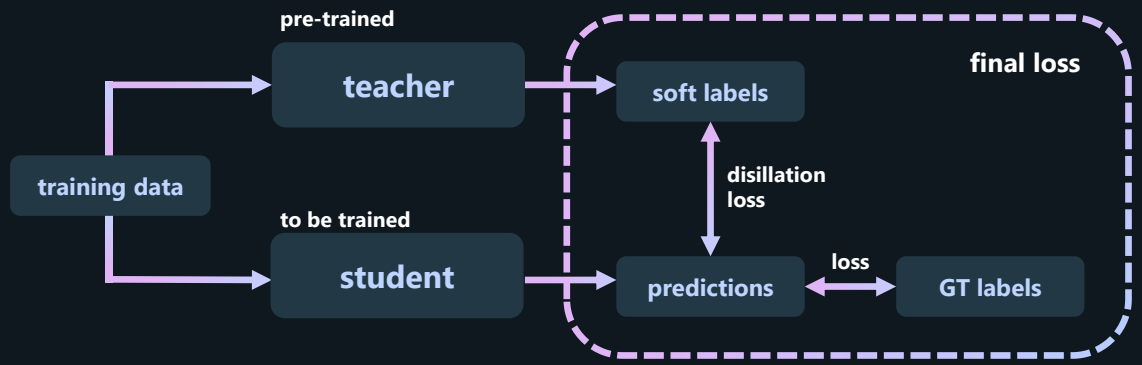
Not all parameters of a model are essential [33]. Instead of creating a small model from 0, we can start from the base.

## We can use Knowledge Distillation [32][34]

Compared with other compression techniques, it showed high fidelity with the original model. [35]

The base models (teacher), already optimized for task, guided the learning of the small models (student). 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);
- the traditional loss function based on the ground truth.

* Seq2SeqTrainer class customization request. [31]




$$L_{total} = \alpha_{CE} * L_{CE} + \alpha_{distill} * T^2 * KL(softmax(\frac{logits_{teacher}}{T})||\log\_softmax(\frac{logits_{student}}{T}))$$

[31] Wolf, 2020. "Transformers: State-of-the-Art Natural Language Processing. [32] Bucila et al., 2006. "Model compression". [33] Lecun, 1998. "Gradient-based learning applied to document recognition". [34] Hinton et al., 2015. "Distilling the Knowledge in a Neural Network". [35] Xu et al., 2021. "Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression".

# Second results

**BASE models**

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 base** | 27,27 | 43,00 | 89,77% | 50,64 | 100% | 251.116.800 | 3.538.944 |
| **mT5 base** | 32,84 | 45,16 | 77,44% | 56,24 | 100% | 585.940.224 | 3.538.944 |

**SMALL models**

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 small** | 24,53 | 36,45 | 88,84% | 65,59 | 100% | 78.337.408 | 1.376.256 |
| **mT5 small** | 30,48 | 42,85 | 76,38% | 55,28 | 100% | 301.553.024 | 1.376.256 |

**SMALL models distilled**

NEW

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 small** | 25,01 | 38,70 | 89,24% | 57,66 | 100% | 78.337.408 | 1.376.256 |
| **mT5 small** | 30,29 | 42,16 | 76,00% | 53,65 | 100% | 301.553.024 | 1.376.256 |

In an attempt to improve the accuracy of previously defined models. Due to the scarcity of data in the J2C-TST datasets, data augmentation (DA) techniques can be adopted.

## Backtranslation is a possibility [35]

Backtranslation consists of the following steps:
1 - The original text from the dataset to be augmented is taken in a source language;
2 - It is translated into another language (pivot language);
3 - Finally, the text is back-translated from the pivot language to the source language.

German is a good pivot language for backtranslation:
- There are high-quality machine translation models;
- It has a different syntactic and grammatical structure compared to languages like italian or English;

**+75% J2C_news_it**

**IT → DE → IT**

**+25% J2C_news_EN**

**EN → DE → EN**

* The integration of the translation functionality was implemented using Google Translate APIs through a third-party library.

[36] Sennrich, et al., 2016. "Neural Machine Translation of Rare Words with Subword Units".

# Third results

**BASE models**

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 base** | 27,27 | 43,00 | 89,77% | 50,64 | 100% | 251.116.800 | 3.538.944 |
| **mT5 base** | 32,84 | 45,16 | 77,44% | 56,24 | 100% | 585.940.224 | 3.538.944 |

**BASE models**

NEW

| | BLEU | METEOR | BERT-score | Perplexity | % conversational | # parameters | # adapted parameters |
|---|---|---|---|---|---|---|---|
| **T5 base** | 25,15 | 40,89 | 89,43% | 49,84 | 97,65% | 251.116.800 | 3.538.944 |
| **mT5 base** | 32,16 | 44,56 | 77,30% | 56,02 | 97,65% | 585.940.224 | 3.538.944 |

**Considerations**

The backtranslation technique did not yield the expected improvements:
- It can be hypothesized that the quality of the translations was insufficient;
- The augmentation percentages adopted may also have played a role

## FUTURE RESEARCHES

**New benchmarks.** One could expand datasets, extending collaborations to English agencies and also exploiting posts from different platforms.

**Extrinsic evaluations.** While complex and not very comparable, it would be useful to conduct an evaluation through end users and experts in the field.

**Enhancing output quality.** You may consider evaluating the use of new foundational models, such as BART [37], use new DA techniques, or test different pivot languages and different rates for backtranslation.

**A text-to-video approach to news.** In some markets, and especially among younger people, there is a preference for video formats for news. T2V approaches could be explored.

**A text-to-audio approach to news.** Among people under 35, a significant percentage is inclined to choose listening, thanks to the widespread use of smartphones and headphones. T2A approaches could be explored.

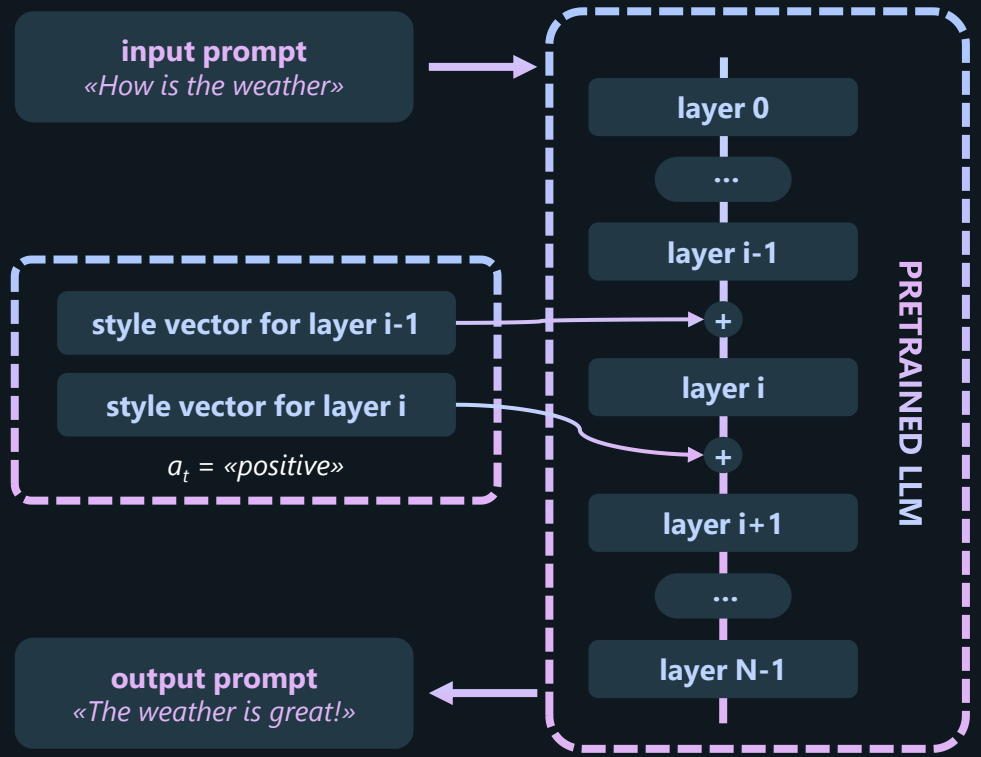
**Context-aware TST, a multimodal approach.** TST models could be modified to incorporate a multimodal context. For example, the style could be inferred directly from a visual context.

[37] Lewis et al., 2020. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension".

Training a deep neural network from scratch is costly in every respect: money, time, and environmental impact. Even fine-tuning large models is often prohibitive, either due to a lack of data or the computational resources required.

For an efficient TST that also does not require parallel pairs for training is based on **activation engineering**. By embedding various documents in the target style embedding in a foundational model, the average activation values of one or more layers can be recorded, thus constructing one or more style embeddings.

These vectors are then added, as a difference from a generic vector, to the activation values of the generalist model in an attempt to influence the style of the generated content




[38] Konen et al., 2024. "Style Vectors for Steering Generative Large Language Model".

> **“ Our future success hinges on our ability to embrace AI and use it for good**

Brad Smith - Microsoft President

**University of Milano-Bicocca**

Department of Informatics, Systems and Communication
Master of Science in Computer Science

# Thank you

tutor
Gabriella Pasi

author
Mattia Piazzalunga

a. year
2023/2024