A text style transfer system for reducing the physician–patient expertise gap

Bacco, Luca; Dell'Orletta, Felice; Lai, Huiyuan; Merone, Mario; Nissim, Malvina

Published in:
Expert systems with applications

DOI:
10.1016/j.eswa.2023.120874

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A text style transfer system for reducing the physician–patient expertise gap: An analysis with automatic and human evaluations

Luca Bacco [a,b,c], Felice Dell'Orletta [b], Huiyuan Lai [d], Mario Merone [a,*], Malvina Nissim [d]

[a] Unit of Computer Systems and Bioinformatics, Università Campus Bio-Medico di Roma, Department of Engineering, via Alvaro del portillo, 21, 00128 Rome, Italy
[b] ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", National Research Council, Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy
[c] R&D Lab, Webmonks S.r.l., Via del Triopio, 5, 00178 Rome, Italy
[d] CLCG, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

ARTICLE INFO

Keywords:
Healthcare
Natural language processing
Text style transfer
Text simplification
Semantic textual similarity

ABSTRACT

Physicians and patients often come from different backgrounds and have varying levels of education, which can result in communication difficulties in the healthcare process. To address this expertise gap, we present a "Text Style Transfer" system. Our system uses Semantic Textual Similarity techniques based on Sentence Transformers models to create pseudo-parallel datasets from a large, non-parallel corpus of lay and expert texts. This approach allowed us to train a denoising autoencoder model (BART), overcoming the limitations of previous systems. Our extensive analysis, which includes both automatic metrics and human evaluations from both lay (patients) and expert (physicians) individuals, shows that our system outperforms state-of-the-art models and is comparable to human-provided gold references in some cases.

1. Introduction

To communicate, humans use natural language in the form of speech or text. However, the information conveyed is not only represented by the content but also by the linguistic form in which it is presented, such as formal/informal attributes. These attributes, also known as *style*, can reflect the writer's intent (e.g., politeness) or reveal their characteristics (e.g., gender). For example, to be perceived as more professional, we might use a more formal vocabulary than what we use in our daily lives. On the other hand, if we are trying to explain a complex concept to someone unfamiliar with the subject, we might use simpler vocabulary and sentence structures.

The medical field faces a persistent challenge known as the *curse of knowledge* between doctors and patients (Camerer, Loewenstein, & Weber, 1989). This cognitive gap can result in misunderstandings and mistakes in treatment (Tan & Goonawardene, 2017; Tong et al., 2020). The language barrier given by differences in their background can lead doctors struggle to diagnose using a patient's language, and patients struggle to understand medical information, exacerbating this issue. Addressing the lack of *health literacy* (Apfel & Tsouros, 2013), which refers to the ability to effectively process and use health information, is crucial for improving communication and overall health outcomes (King, 2010; Tian, Xu, Mo, Dong, & Wong, 2020). This can also lower hospitalization rates and healthcare costs (Baker et al., 2002; Baker, Parker, Williams, & Clark, 1998; Mäenpää, Suominen, Asikainen, Maass, & Rostila, 2009). Reducing the knowledge gap between doctors and patients can enhance health equity and empower patients to take an active role in their care (Batterham, Hawkins, Collins, Buchbinder, & Osborne, 2016). This would also help decrease confusion and anxiety that results from misunderstandings of medical jargon, which importance is underscored by the growing use of health-related online resources such as mobile apps and social networks (Benigeri & Pluye, 2003; Tan & Goonawardene, 2017; White & Horvitz, 2009; Zielstorff, 2003). Among others, the World Health Organization acknowledges the need to address health literacy as part of its 2030 Agenda for promoting health.[1]

In this context, *Natural Language Processing* (NLP) presents the potential to play a crucial role. NLP has already demonstrated its ability to enhance patient care, by serving as support for physician diagnosis, training Computer-Aided Diagnosis (CAD) systems (Bacco et al., 2022; Long, Yang, & Liu, 2022; Wang et al., 2017), and identifying and improving flaws in the procedures and services offered by companies (Bacco, Cimino, Paulon, Merone, & Dell'Orletta, 2020; Zhang &

* Corresponding author.
E-mail addresses: l.bacco@unicampus.it (L. Bacco), felice.dellorletta@ilc.cnr.it (F. Dell'Orletta), h.lai@rug.nl (H. Lai), m.merone@unicampus.it (M. Merone), m.nissim@rug.nl (M. Nissim).

[1] WHO: *Shanghai Declaration on Promoting Health in the 2030 Agenda for Sustainable Development*. https://www.who.int/publications/i/item/WHO-NMH-PND-17.5.

https://doi.org/10.1016/j.eswa.2023.120874
Received 8 August 2022; Received in revised form 10 February 2023; Accepted 15 June 2023
Available online 24 June 2023

Jankowski, 2022). Both NLP and clinical informatics communities are actively working on a variety of tasks involving the automatic analysis of health-related texts (Gao et al., 2022). *Text Style Transfer* (TST) is one of them. TST is the process of changing the style of a text while preserving its content. By simplifying medical jargon, TST (also referred as *Text Simplification* in this case) can help patients better understand their health information and bridge the gap between doctors and patients. To help the community get over such issues, Cao et al. (2020) recently introduced the *Expertise Style Transfer* task for medical experts and lay people. The authors proposed a large non-parallel dataset for training models and a small parallel dataset annotated by experts for evaluation purposes. In this paper we present our work to handle such task. To improve the training of the models, we explored methods to collect parallel datasets from a large, non-parallel corpus. We used the Bidirectional and Auto-Regressive Transformers (BART) model (Lewis et al., 2020), which has been shown to be effective for style transfer tasks (Lai, Toral, & Nissim, 2021a, 2021b). We collected parallel datasets using the *Sentence Transformers* (Reimers & Gurevych, 2019) models, which are well-suited for the task of similarity search, and datasets from existing literature on (clinical) *Semantic Textual Similarity*. The collected parallel datasets and the style-transferred texts were evaluated using both automatic metrics and manual annotations from both lay and expert individuals, focusing on content preservation and the degree of style change.

Within this frame, our study advances the field by introducing a novel approach to reduce the expertise gap in communication between physicians and patients through a *Text Style Transfer* system. Despite being built upon previous works, our approach offers a fresh and innovative perspective on the *Expertise Style Transfer* task. In summary,

- this paper presents a novel approach to reduce the expertise gap in communication between physicians and patients by proposing a Text Style Transfer system;
- the system was developed by analyzing several *Semantic Textual Similarity* methods to collect parallel datasets
- and was extensively evaluated through human evaluations involving both lay people and experts in the field;
- the results showed that the system outperformed the state-of-the-art models based on both automatic and human evaluations.

The present study also highlights the feasibility of collecting pseudo-parallel data through *Semantic Textual Similarity* techniques, which leads to improved performance in a cost-effective manner. The combination of automated and human evaluations, coupled with the parallel data collection analysis, makes our study a valuable contribution to the current state-of-the-art, holding important theoretical implications and providing valuable insights into TST research in the medical field.

Our in-depth analysis of parallel data collection provides unique insights into the relationship between the quality of the collected datasets and the accuracy of the TST system. In particular, our human expert evaluations offer a rare perspective on the performance and quality of our methodology. Moreover, the expert annotations collected can be utilized in future studies to further improve and evaluate TST systems in the medical domain.

The present manuscript is structured as follows. First, we report the investigated literature and background information on Text Style Transfer and the expertise variants of TST (Section 2). In Section 3 we provide an overview on the datasets we exploited for both *Semantic Textual Similarity* and *Expertise Style Transfer*. In particular, we report some considerations about the *MSD* dataset proposed by Cao et al. (2020). After that, we present a detailed explanation of the methods used to address the Expertise Style Transfer task and the collection of pseudo-parallel datasets (Section 4), and the evaluation protocols to assess their quality (Section 5). Finally, in Section 6 we present a comprehensive discussion of the results from both automatic and human evaluations (and their correlations), focusing on content preservation, style strength, and fluency of the models' outputs and the collected parallel training sets, and including a qualitative analysis.

2. Related works

Style Transfer (Neural ST, in particular) is the task of reproducing some input content in a different style. Researchers investigated it for several media, from images and videos (Gatys, Ecker, & Bethge, 2015; Huang et al., 2017; Jing et al., 2020; Kim, Nam, Hong, & Park, 2022) to music (Cífka, Şimşekli, & Richard, 2020; Mukherjee & Mulimani, 2022). Some of the developed systems have already seen their application in industrial solutions.[2] *Text Style Transfer* (TST) shares the same principle as the other media: rewriting some textual input with a different attribute while minimizing the information loss. Researchers have investigated TST for various attributes such as formality (Lai et al., 2021b; Rao & Tetreault, 2018), politeness (Madaan et al., 2020; Niu & Bansal, 2018), and sentiment (Shen, Lei, Barzilay, & Jaakkola, 2017). Past works in TST have focused on these attributes as the related resources are easier to obtain. The sentiment Style Transfer, commonly known as polarity swap, has been questioned as a TST task (Lai, Toral, & Nissim, 2021a) as it does not preserve the original meaning of the source text, i.e., a positive sentence is changed to negative and vice-versa. However, in order to get a more comprehensive overview of TST tasks, we refer the reader to a few, recent reviews from Jin, Jin, Hu, Vechtomova, and Mihalcea (2021) and Toshevska and Gievska (2022).

Existing TST approaches can be grouped into three main categories, i.e., disentanglement, manipulation, and translation:

- *Disentanglement* methods attempt to learn separate representations for content and style (Fu, Tan, Peng, Zhao, & Yan, 2018; Hu, Yang, Liang, Salakhutdinov, & Xing, 2017; Shen et al., 2017), so that one can be manipulated without affecting the other. However, the success of disentanglement is difficult to assess, and some studies have shown that the latent representations may not actually be disentangled, being possible to recover information of style from the other (Elazar & Goldberg, 2018; Lample et al., 2019).
- *Manipulation* methods work by identifying specific words in the text that contribute to its style, such as professional language or clinical abbreviations (e.g., *qd*), and replacing them with synonyms or explanations (e.g., *once per day*) that are more appropriate for lay people (Li, Jia, He, & Liang, 2018; Shardlow & Nawaz, 2019). In the biomedical and clinical domain (Weng, Chung, & Szolovits, 2019; Zeng-Treitler, Goryachev, Kim, Keselman, & Rosendale, 2007), these methods often use *Consumer Health Vocabularies* (Manzini, Garrido-Aguirre, Fonollosa, & Perera-Lluna, 2022; Vydiswaran, Mei, Hanauer, & Zheng, 2014; Zielstorff, 2003, CHVs). Weng et al. (2019), in particular, used CHVs as a preliminary step to align embedding spaces and then used a translation-based technique to generate simplified sentences.
- *Translation* methods often use unsupervised training to learn style-specific translations (Lample et al., 2019) with back-translation or cycle reconstruction strategies. Back-translation (Sennrich, Haddow, & Birch, 2015) involves translating the source text to another language and back again. Prabhumoye, Tsvetkov, Salakhutdinov, and Black (2018) proposed it on the basis of the evidence shown by Rabinovich, Mirkin, Patel, Specia, and Wintner (2016) to reduce the style properties of the source text. Such strategy has already shown its efficiency (Lai, Toral, & Nissim, 2021a) Cycle reconstruction, instead, involves training a model to reconstruct the source text from the transferred output (Dai, Liang, Qiu, & Huang, 2019; Zhou et al., 2020). Parallel corpora can also be used for supervised training, but they can be expensive and time-consuming to collect.

[2] https://prisma-ai.com/; https://www.pikazoapp.com/; https://deepart.io/; https://groove2groove.telecom-paris.fr/

For the *Expertise Style Transfer* task at hand, Cao et al. (2020) evaluated models belonging to the three macro-categories of TST discussed earlier (see also Section 5 for an overview), while our approach falls into the latter category. In particular, we exploited the collection of pseudo-parallel corpora, built on the basis of a definition of similarity criterion between sentences, which has shown advantages over unsupervised training (Jin, Jin, Mueller, Matthews, & Santus, 2019), while being cost-effective if compared with the collection of human-annotated corpora.

In one related work, Luo et al. (2020) collected gold corpora from *MIMIC-III* database (Johnson et al., 2016), which was a time-intensive process requiring a certain degree of expertise. To overcome this issue, like us, Xu, Saxon, Sra, and Wang (2021) collected a large, pseudo-parallel corpus from the MSD training set. While sharing the same intent to collect pseudo-parallel corpora, there are some crucial differences. They used a language- and topic-agnostic LASER (Artetxe & Schwenk, 2019) framework to extract the embeddings and collected the largest number of training pairs above a fixed threshold on their similarity criterion. Our approach differs in the use of general and domain-specific monolingual Transformer-based models and in the investigation of the impact of different threshold ranges on the final TST system.

Disposing of parallel corpora can be an effective solution for these issues (Jin et al., 2019). When such data is not available, the automatic collection of pseudo-parallel data has proven to be effective, including in neural machine translation tasks (Imankulova, Sato, & Komachi, 2017, 2019). Style transfer, similarly to machine translation, is a rewriting task and shares similar modeling approaches. While machine translation deals with cross-language content, style transfer is typically within the same language. In some cases, it is approached from a multilingual perspective (Lai, Toral, & Nissim, 2022).

However, the collection of high-quality parallel datasets is a challenging task, especially in a specialized domain like healthcare, where human efforts and costs are significant. To address this issue, van den Bercken, Sips, and Lofi (2019) proposed using the BLEU score (Papineni, Roukos, Ward, & Zhu, 2002) to automatically collect a parallel dataset for a medical simplification task by utilizing texts from Wikipedia and Simple Wikipedia. However, this approach was found to be unsuitable for our use case due to the presence of many texts in both the expert and lay training corpora, and the significant differences between the expert and lay test samples. Another common technique for collecting parallel datasets is to train a classifier that can distinguish sentence pairs from two different corpora (Marie & Fujita, 2017; Zhu, Yang, & Xu, 2020). However, using a classifier for large corpora is often infeasible, especially when using Transformer architectures. Our approach addresses these limitations by employing bi-encoders. To the best of our knowledge, the use of bi-encoders in style transfer tasks, particularly in the technical domain of medicine, has not been explored previously.

Furthermore, Xu et al. (2021) focused mainly on human evaluation and compared their outputs only with inputs using (self-)BLEU, ignoring reference sentences in their analysis (ref-BLEU). The interpretation of high self-BLEU scores is not trivial: a score close to 100% between input and output only means that the model has learned to reproduce the input without making any changes to the style. Moreover, it has been established that surface-based metrics like BLEU are not ideal for TST tasks, as they exhibit low correlation with human judgments (Briakou, Agrawal, Tetreault, & Carpuat, 2021; Lai, Mao, Toral, & Nissim, 2022). For these reasons, we evaluated our outputs and those of the models presented by Cao et al. (2020) using several other metrics, referred to both input and target sentences. Computing these metrics for the gold source and target texts allowed us to highlight the degree of content changes in the test set, as suggested in previous works (Basu, Vasu, Yasunaga, Kim, & Yang, 2021; Cao et al., 2020; Vásquez-Rodríguez, Shardlow, Przybyła, & Ananiadou, 2021), and confirmed through our human evaluations. This issue may stem from the loss of contextual information when working at a sentence level. As a result, a few studies have taken a paragraph-level approach to the medical-style transfer task from the perspective of Plain Language Summarization (Devaraj, Marshall, Wallace, & Li, 2021; Guo, Qiu, Wang, & Cohen, 2021, PLS), also in languages other than English (Grabar & Cardon, 2018).

Our study makes a unique and significant contribution to the field of Text Style Transfer by presenting an extensive examination of the collection of parallel data and offering unique insights specific to its application. Furthermore, our human expert evaluations set our work apart from previous studies, providing a valuable and rare perspective on the performance and quality of our system. Despite building upon previous works, our approach offers a fresh and innovative perspective on the task of Text Style Transfer. The combination of automated and human evaluations, coupled with the in-depth analysis of parallel data collection, makes our study a valuable addition to the current state-of-the-art. Table 1 provides a summary of previous studies on TST in the medical and clinical domain. The table highlights the main characteristics of each study, including the differences with the current study being discussed.

3. Datasets

In our work, we utilized and combined three datasets for both similarity and style transfer tasks.

3.1. ClinicalSTS2019 (CSTS)

Wang, Fu et al. (2020) collected a total of 2054 sentence pairs annotated by two clinical experts for the *Clinical Semantic Textual Similarity* track in the *n2c2/OHNLP* challenge of 2019. The training set is an extension of the dataset presented in the previous year's challenge (Wang et al., 2018). Authors asked experts to independently annotate each pair based on their semantic equivalence on a scale from 0 to 5, where 0 indicates completely dissimilar sentences and 5 indicates a perfect semantic match. For more information and data examples, see Table 4 and Wang, Afzal et al. (2020).

3.2. Medical question pairs (MQP)

McCreery, Katariya, Kannan, Chablani, and Amatriain (2020) collected a dataset of *COVID-19* related questions.[3] It contains 1524 pairs of questions, where each pair consists of one positive and one negative example. The positive examples are rephrased by doctors to maintain the content while restructuring the original question as much as possible. The negative examples are rephrased in a manner that the answer of that question would be incorrect or irrelevant while maintaining the same structure and keywords. The pairs are labeled as similar or dissimilar (1 or 0) based on their semantic equivalence.

3.3. MSD

The MSD dataset contains a large collection of medical sentences in both expert and layman styles (~130k and ~115k sentences, respectively), and a smaller annotated set of parallel texts (675 pairs) for evaluation purposes. Cao et al. (2020) collected such a dataset from the *Merck Manuals* (also known as the *MSD Manuals*) website,[4] which is one of most world-widely trusted reference in health. The authors also provided medical entities and concepts from the Unified Medical Language System (UMLS, Bodenreider, 2004) for each sample using the QuickUMLS (Soldaini & Goharian, 2016) tool.

The empirical analysis we conducted on the parallel test set revealed various problematic patterns. Table 2 reports an overview of some examples. These problematic patterns can compromise the evaluation

Table 1
Summary table of TST papers in medical and clinical domain, i.e., *Expertise ST*, *Medical ST*, and *Plain Language Summarization* (PLS). The former may be seen as an instance of the second one. The latter see the TST task along with the summarization of some paragraphs (abstracts from the medical literature).

Authors	Task	Level	Data	Dataset	Language	Metrics	Characteristics
Cao et al. (2020)	Expertise ST	sentence	non-parallel	MSD	English	Style Acc., self-/ref-BLEU, perplexity, human (lay)	(i) collection of a new dataset; (ii) use of non-parallel corpus; (iii) only human-lay evaluation; (iv) missing analysis of automatic and human metrics correlation
Xu et al. (2021)	Expertise ST	sentence	pseudo-parallel	MSD	English	Style Acc., self-BLEU, perplexity, human (lay)	(i) language- and topic-agnostic framework (LASER); (ii) empirical choice of similarity threshold; (iii) no ref-metrics for content preservation assessment; (iv) only human-lay evaluation; (v) missing analysis of automatic and human metrics correlation
Zeng-Treitler et al. (2007)	Medical ST	word	non-parallel	Unknown	English	Readability, human (exp)	(i) synonym replacement; (ii) explanation insertion; (iii) human evaluation with only one individual; (iv) missing analysis of automatic and human metrics correlation
Weng et al. (2019)	Medical ST	word/sentence	non-parallel	MIMIC III	English	Precision at k, human	(i) use of non-parallel corpus; (ii) human evaluation with both experts and lay people; (iii) missing analysis of automatic and human metrics correlation; (iv) two-steps word/sentence translation method using CHVs
van den Bercken et al. (2019)	Medical ST	sentence	pseudo-parallel	Wikipedia	English	ref-BLEU, SARI, human (lay)	(i) collection of a new dataset; (ii) BLEU-based pseudo-parallel data collection; (iii) only human-lay evaluation; (iv) missing analysis of automatic and human metrics correlation
Luo et al. (2020)	Medical ST	sentence	parallel	MedLane	English	several	(i) collection of a new dataset; (ii) costly human annotation; (iii) no human evaluation
Grabar and Cardon (2018)	PLS	paragraph	parallel	CLEAR	French	/	(i) collection of a new, multi-source dataset
Devaraj et al. (2021)	PLS	paragraph	parallel	CDSR	English	Readability, ROUGE, BLEU, SARI	(i) only qualitative human evaluation; (ii) missing analysis of automatic and human metrics correlation
Guo et al. (2021)	PLS	paragraph	parallel	CDSR	English	Readability, ROUGE, human (lay)	(i) only human-lay evaluation; (ii) missing analysis of automatic and human metrics correlation

Table 2
The analysis of the *MSD* test dataset has revealed some problematic pairs. Most of them belong to one of the following patterns: (i) duplicate texts for both styles, (ii) poor fluency, (iii) missing information, (iv) different gold target references for the same source text, (v) acronyms, and (vi) different meanings between source and target texts. The truncated texts are indicated with "[...]" to accommodate them in the table.

	Expert	Layman
(i)	The change in LDL levels may partly explain why atherosclerosis and thus coronary artery disease become more common among women after menopause. [...]	The change in LDL levels may partly explain why atherosclerosis and thus coronary artery disease become more common among women after menopause. [...]
(ii)	Treatment of underlying disorder	Treatment of cause
(iii)	The most common causative organisms of occult bacteremia are Streptococcus pneumoniae and Haemophilus influenzae. [...]	Children **under 3 years old** who develop a fever **(particularly if their temperature is 102.2° F [39 °C] or higher)** sometimes have bacteria in their bloodstream (bacteremia). [...]
(iv)	Clinical evaluation Clinical evaluation Clinical evaluation.	Physical examination A doctor's evaluation A doctor's examination.
(v)	**IV** fluids.	Fluids given by vein
(vi)	[...] **It occurs predominantly in men** practicing receptive anal intercourse and can occur in women who participate in anal sex.	**It occurs mainly in women**. Anal sex with an infected partner may result in gonorrhea of the rectum.

of the models, and a more thorough evaluation is needed to better understand the difficulties of the medical style transfer task.

Regarding the training dataset, we discovered that there were overlapping texts in both styles, particularly in instances of fixed word patterns. We regarded these instances as irrelevant and filtered them out by removing sentences that were short (less than 10 tokens) or displayed specific patterns using simple regular expressions. This preprocessing stage reduced the number of samples to approximately 110k for the expert style and 97k for the layman style.

4. Text style transfer system

From a mathematical standpoint, the aim of a TST system is to model the probability $p(y|x)$ where $x(c, a)$ is the source sentence and $y(c, b)$ is the target sentence, with the same content c but different attributes (styles) a and b. If the system can also model the reverse direction $p(x|y)$, it is referred to as bidirectional (Jin et al., 2021). If the transformation is from a more complex source text to a simpler one, such as from expert to layman style, it is also referred to as *Text Simplification*.

For our system, we exploited the collected pseudo-parallel training sets (Section 4.1) to fine-tune a Bidirectional and Auto-Regressive Transformers (BART) model (Lewis et al., 2020). *BART* is a denoising autoencoder for pre-training sequence-to-sequence model. Given a source sentence $\boldsymbol{x} = \{x_1, \dots, x_n\}$ and a target sentence $\boldsymbol{y} = \{y_1, \dots, y_m\}$, its loss function is the cross-entropy between the decoder's output and the target sentence:

$$L_{(\phi)} = -\Sigma_i \log(p(y_i|y_{1:i-1}, \boldsymbol{x}; \phi)) \tag{1}$$

The entire system pipeline is depicted in Fig. 1 and consists of four steps.

i Initializing the *Sentence Transformers* with the pre-trained *BERT*-based models' weights.
ii Fine-tuning the Sentence Transformers with the datasets described in Section 3.
iii Using the bi-encoders to perform a similarity search on the expert and layman corpora from the MSD training data.
iv Fine-tuning the *BART* model for the Text Style Transfer task using pseudo-parallel data collected by setting a similarity threshold.

The resulting model can then be used during inference to simplify medical texts for a lay audience.

4.1. Pseudo-parallel data collection

Our supervised approach starts with collecting pseudo-parallel data from the expert and layman corpora. To do this, we used the *Sentence-Transformers* (Reimers & Gurevych, 2019) architecture. This architecture consists of a bi-encoder, a siamese network that trains one Transformer encoder to produce semantically meaningful embeddings. This results in outputs of semantically similar sentences being closer to each other in the vector space compared to dissimilar sentences.

The bi-encoder architecture also makes it computationally efficient to conduct large-scale semantic search, which is what our task required. The bi-encoder reduces the complexity of retrieving representations for each paired combination in the dataset to obtaining one embedding for each sentence and computing the (cosine) similarity between paired embeddings through the *FAISS* (Johnson, Douze, & Jégou, 2019) library, which is optimized for GPU usage. This allowed us to efficiently find the nearest layman equivalent for each expert sample.

4.1.1. Pre-trained models

We evaluated several Transformers encoders to obtain sentence embeddings for the training dataset. We first compared two general-topic and domain-specific encoders, i.e., *BERT* (Devlin, Chang, Lee, & Toutanova, 2019) and *(Bio-)ClinicalBert* (Alsentzer et al., 2019). Both encoders have the same architecture, which eliminates the influence of different architectural models. *(Bio-)ClinicalBert* was initialized using *BERT* and then pre-trained on large medical and clinical domain data. As expected, this led to better similarity performance, as shown in Table 3. Therefore, we used *(Bio-)ClinicalBert* for our training strategy.

4.1.2. Training strategies

We implemented the *Multiple Negatives Ranking (MNR) loss* (Henderson et al., 2017) as the loss for the *contrastive (representation) learning* (Hadsell, Chopra, & LeCun, 2006). It pushes the model to create closer representations in the vector space for similar sentences and more distant for dissimilar ones, based on some distance/similarity metric. At each step, the training process aims to minimize the following equation:

$$L_{MNR} = -\frac{1}{K}\sum_{i=1}^{K}[s(x_i, y_i) - log\sum_{j=1}^{K} e^{s(x_i, y_j)}] \tag{2}$$

in which (x_i, y_i) indicates any ith anchor-positive (premise and hypothesis) pair and (x_i, y_j) indicates any anchor-negative pair in the batch of size K; $s(.,.)$, instead, indicates the score based on the defined metric (cosine similarity in our case).

We trained our models on *MQP* and *ClinicalSTS* training datasets. In some cases, we used only positive (pos) pairs (for *ClinicalSTS*, we considered a semantic equivalence score of 4 or higher). For the *MSD* training dataset, which did not have any content equivalence labels, we followed the unsupervised *SimCSE* (Similarity Contrastive Sentence Embedding) framework as proposed in Gao, Yao, and Chen (2021). In this framework, we used the Transformer encoder and anchor-positive pairs consisting of the same input sentence. The randomness of the *dropout* (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) masks in the encoder's layers generated two different noisy representations of the same sentence, and the model learned to generate closer embeddings from the noisy representations while distancing anchor-negative pairs in the batch.

4.1.3. STS evaluation

To be consistent with past literature, we evaluated the models using two common metrics for semantic textual similarity, Pearson and Spearman correlations between the similarity scores $\boldsymbol{x} = \{x_1, \dots, x_n\}$ produced by the sentence embeddings and the *CSTS* official test set labels $\boldsymbol{y} = \{y_1, \dots, y_n\}$. Eq. (3) reports the formulas of these metrics,

$$pearson = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \; ; \quad spearman = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

where $\overline{x}$ and $\overline{y}$ indicate the mean of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, n is the number of elements, and d_i the pairwise distance of the ranks of the ith elements (x_i and y_i). In particular, we defined the score x_i between two sentences $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ as the cosine similarity of their embeddings, as reported in Eq. (4).

$$cos(\boldsymbol{a}_i, \boldsymbol{b}_i) = \frac{\boldsymbol{a}_i \cdot \boldsymbol{b}_i}{\|\boldsymbol{a}_i\| \cdot \|\boldsymbol{b}_i\|} \tag{4}$$

In addition, we also assessed the models' performance by calculating the average cosine similarity between expert-layman pairs ($\boldsymbol{a}_i$ and $\boldsymbol{b}_i$) in the *MSD* test set as

$$similarity = \frac{1}{N}\sum_{i}^{N} cos(\boldsymbol{a}_i, \boldsymbol{b}_i) \tag{5}$$

where N is the number of pairs. Table 3 reports the results of these evaluations.

[3] https://huggingface.co/datasets/medical_questions_pairs

[4] https://www.msdmanuals.com/

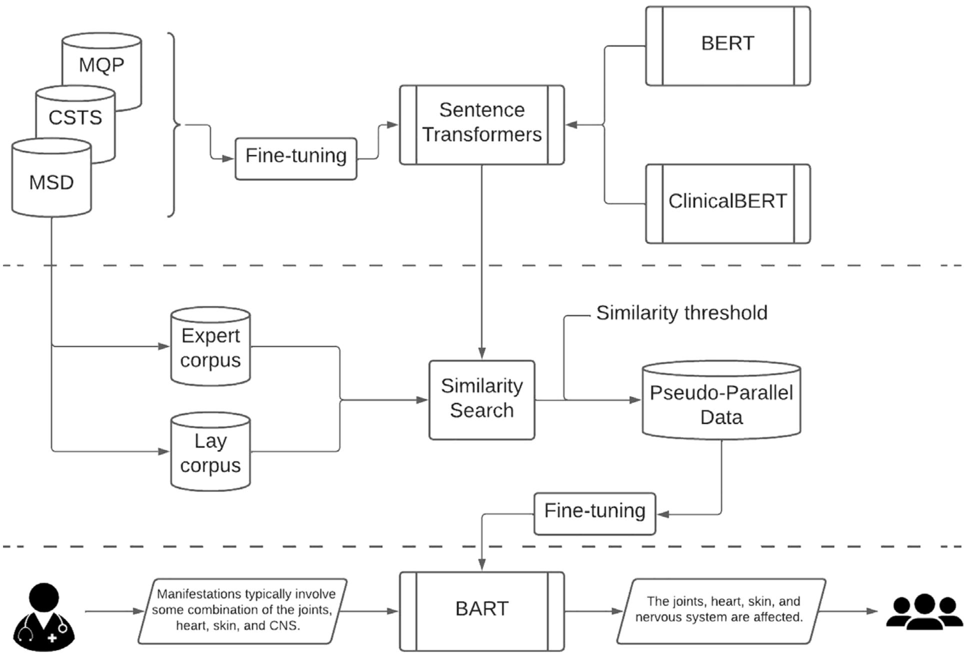

Fig. 1. Our approach consists of the following pipeline: (i) retrieving pre-trained Transformers as a bi-encoder and (ii) fine-tuning them with *Semantic Textual Similarity* datasets or *MSD* training set; then, (iii) using the fine-tuned bi-encoder to perform a similarity search on the expert and layman corpora derived from the MSD training set. By setting a similarity threshold to collect pseudo-parallel data, (iv) fine-tuning the style transfer model using the collected pseudo-parallel data. In the end, the fine-tuned model is used during inference time to simplify medical texts from physicians to patients.

Table 3
The table reports the performance of the Semantic Textual Similarity models, each of which is identified by the first column. The models were fine-tuned starting from a pre-trained model on a specific training set, i.e., mqp (medical question pairs), csts (clinicalSTS2019), and msd. The superscript (pos) indicates that only the positive pairs were included in the training process. The first two rows report basic pre-trained models without further training. The performance of the models was evaluated in terms of Pearson and Spearman correlation coefficients computed on the clinicalSTS2019 dataset, and on the average cosine similarity computed on the parallel samples of the msd test set.

Id	Model	Training set	Pearson (%)	Spearman (%)	Similarity (%)
bert	Bert-base-uncased	/	21.64	25.03	87.70
cb	Bio-ClinicalBert	/	30.07	31.84	**93.99**
cb_mqp1	Bio-ClinicalBert	mqp$^{(pos)}$	68.26	71.29	69.72
cb_mqp	Bio-ClinicalBert	mqp	80.27	77.41	67.12
cb_csts1	Bio-ClinicalBert	csts$^{(pos)}$	43.91	47.44	**79.28**
cb_csts	Bio-ClinicalBert	csts	61.61	56.08	66.33
cb_mqp_csts1	cb_mqp	csts$^{(pos)}$	**81.12**	**78.29**	69.93
cb_mqp_csts	cb_mqp	csts	66.51	62.33	65.17
cb_msd	Bio-ClinicalBert	msd	53.22	53.67	47.93

As previously mentioned, *(Bio-)ClinicalBert* outperformed the other pre-trained model, *Bert* (**cb** and **bert** in the table), for all the metrics. This was expected since the former passed through a pre-training (domain adaptation) phase in the biomedical and clinical domains. The **cb_mqp_csts1** model achieved the highest correlation scores. It is a *(Bio-)ClinicalBert* we first fine-tuned on the *MQP* dataset and then on the positive samples of the *CSTS* dataset. Interestingly, the second fine-tuning step only slightly improved the performance (see **cb_mqp**). Apart from the pre-trained models, for what concerns the evaluation on the *MSD* test set, the model fine-tuned only on the *CSTS* positive samples (**cb_csts1**) achieved the highest averaged cosine similarity. The model fine-tuned using only the *MSD* data, instead, performed poorly on the averaged cosine similarity. Such a result may indicate that the training strategy employed for this model was not suitable for the test set at hand, which presents a high degree of aggressiveness in the changes between source and related target texts.

4.1.4. Datasets creation

As the last step, we collected the pseudo-parallel datasets. We conducted the next analyses with a selected set of the implemented models. Based on their performances, we retrieved the pairs collected using **cb_mqp_csts1** and **cb_csts1** models. To analyze the impact of the fine-tuning strategies, as well as the pre-training domain adaptation step, we included the **cb_msd**, **cb** and **bert** models, too. Furthermore, we also computed the similarity search between the lay corpus and a "corrected" expert one, for which we switched expert terms with their lay-related terms. To do so, following a similar approach of Xu et al. (2021), we first collect all the *Concept Unique Identifier* (CUI) codes in the *MSD* training set, as well as the number of occurrences of the terms appearing in the texts for each style. Then, we switched each expert term with the most represented one in the lay texts that share the same CUI(s). From now on in the paper, we refer to this dataset as **cb_msd_swap**.

To analyze the impact collected training sets may have on the final task at different similarity thresholds, we retrieved several datasets at different threshold ranges based on the quantiles they separate in the entire training set. We thus selected the following ranges between the following quantiles: {99%, 95%, 90%, 85%, 80%, 75%, 70%, 50%}. To minimize the impact of the training set size, we used the same number of samples for each interval (with the exception of the ones above

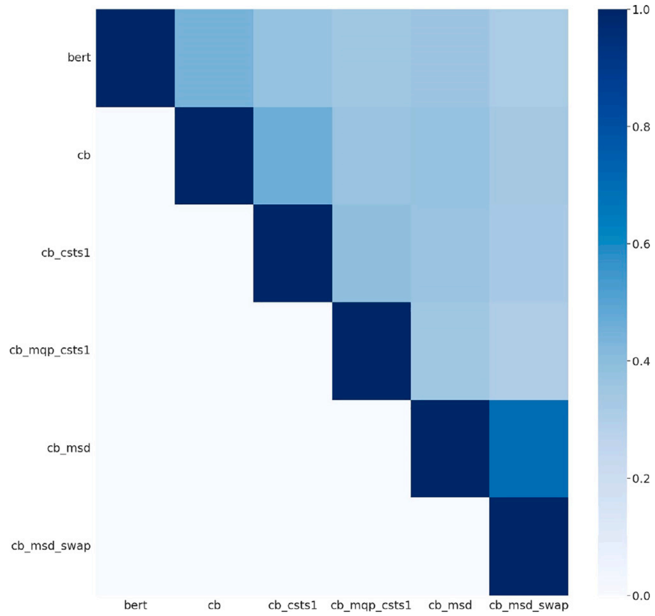

Fig. 2. Average overlaps between collected datasets.

99% and between 99% and 95%, which contained a smaller number of samples). We then evaluated the overlap between the datasets collected by the several models by averaging their overlap at each quantile. As shown in Fig. 2, datasets were more or less dissimilar, on average. The pre-trained models are more similar to each other than to the fine-tuned ones. Also, the two datasets collected with the **cb_msd** model look mostly overlap.

5. Automatic and human evaluation

We used both automatic and human evaluation methods to comprehensively assess the quality of the style transfer systems and to understand the strengths and limitations of the models. Each metric refers to one of the text's aspects, i.e., the style strength (degree of style transfer) of the target text, the content preservation between source and target texts, and the fluency of the generated text. In particular, we compared our results with some baseline models, i.e., an *unsupervised BART* model (having the same architecture as in our system) and all the models provided by Cao et al. (2020), which include two text simplification models and three style transfer models. We fine-tuned the *unsupervised BART* model without parallel data by employing an iterative back-translation approach (Hoang, Koehn, Haffari, & Cohn, 2018). Two models for the two transfer directions get trained (almost) simultaneously. Specifically, each model generates synthetic parallel data for the other. In this way, the models get trained in a pseudo-supervised fashion, each in one direction.

The baselines provided by Cao et al. (2020) include:

- *OpenNMT+PT* (Shardlow & Nawaz, 2019), an OpenNMT-based (Klein, Kim, Deng, Senellart, & Rush, 2017) supervised model that replaces complex words with their simple synonym based on a phrase table;
- *UNTS* (Surya, Mishra, Laha, Jain, & Sankaranarayanan, 2019), an unsupervised neural model consisting of a shared encoder and a pair of attentional decoders; it is trained with discrimination-based losses and denoising;
- *ControlledGen* (Hu et al., 2017), a neural generative model combining variational auto-encoders and style attribute discriminators for the effective imposition of semantic structures;
- *DeleteAndRetrieve* (Li et al., 2018), an editing-based method that first deletes style-related words, then retrieves new phrases associated with the target attribute and uses a neural model to combine them as the final output;
- *StyleTransformer* (Dai et al., 2019), a Transformer-based model that uses cycle reconstruction to learn content and style representation without parallel data.

5.1. Automatic evaluation

Following previous works (Lai, Toral, & Nissim, 2021a; Lai et al., 2021b; Luo et al., 2019; Sancheti, Krishna, Srinivasan, & Natarajan, 2020), we used the following strategies. To assess the **content** aspect, we computed BLEU (Papineni et al., 2002) and BERTScore (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020) between the generated sentence and the human source and reference. BLEU counts the n-gram matches in the candidate text with the reference one, this can be roughly formulated as

$$\text{BLEU-}n = \frac{\sum_{C\in\{Candidates\}}\sum_{\text{n-gram}\in C} Count_{match}(\text{n-gram})}{\sum_{C\in\{Candidates\}}\sum_{\text{n-gram}\in C} Count(\text{n-gram})} \quad (6)$$

where C represents the candidate text and $match$ means that a n-gram appears in both the candidate and either the source (*self-BLEU*) or the reference (*ref-BLEU*). In particular, we used the *overall* BLEU by averaging the scores obtained with $n = \{1, 2, 3, 4\}$. BERTScore uses greedy matching to maximize the matching similarity score for each token in the candidate sentence with each token in either the source (*self-BERTScore*) or the reference (*ref-BERTScore*), and combines recall (R) and precision (P) to compute an F_1 measure. This can be formulated as

$$R = \frac{1}{|x|}\sum_{x_i\in x}\max_{\hat{x}_j\in\hat{x}}\mathbf{x}_i^{\top}\hat{\mathbf{x}}_j\ , \quad P = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j\in\hat{x}}\max_{x_i\in x}\mathbf{x}_i^{\top}\hat{\mathbf{x}}_j\ , \quad F_1 = 2\frac{R\cdot P}{R+P} \quad (7)$$

where $\hat{x}$ and x represent the candidate and reference, respectively. We also included two learnable metrics, BLEURT (Sellam, Das, & Parikh, 2020) and COMET (Rei, Stewart, Farinha, & Lavie, 2020), as they have shown promising correlation results with human judgments in the evaluation of machine translation, as well as style transfer tasks as formality (Lai, Mao et al., 2022).

For what concerns the evaluation of the **style** of texts, we used a TextCNN-based (Kim, 2014) classifier (trained on the entire training set) to evaluate the target style accuracy of the transferred texts.

Regarding the **fluency**, we assessed the perplexity in an analogous way as in Cao et al. (2020), computing a (pseudo-)perplexity with a masked language model. As in Salazar, Liang, Nguyen, and Kirchhoff (2020), for each text W^i, for each of its tokens w_t, we first computed the conditional log probability $P_{MLM}(w_t|W^i_{\backslash t})$ obtained by the model giving in input the sentence $W^i_{\backslash t}$, that is the same as W^i but with the tth token masked. Then, we computed the pseudo-likelihood PLL^i of each text by summing the contribution for each token. Finally, we added the scores of all the corpus S of sentences together, normalized the result with respect to the total number of tokens N in the corpus, and exponentiated the result to obtain a measure of pseudo-perplexity $pPPL$. The process is summed up in the following equation:

$$\begin{aligned} pPPL(S) &= exp(-\frac{1}{N}\sum_{i=1}^{|S|} PLL(W^i)) = \\ &= exp(-\frac{1}{N}\sum_{i=1}^{|S|}\sum_{t=1}^{|W^i|} log P_{mlm}(w^i_t|W^i_{\backslash t})) \end{aligned} \quad (8)$$

To compute such scores, we used *(Bio-)ClinicalBert* as well as its versions fine-tuned on the training sets for lay or expert styles. To balance the data sizes and remove any influence given by the different corpus dimensions, we reduced the number of experts' texts during fine-tuning.

5.2. Human evaluation

The human evaluation was conducted with two different protocols to capture both the lay people and professional physician's perspective, thus we elaborated two different protocols. The lay people were asked to judge only the style perception of the samples, while the physicians were asked to judge both the style and the content preservation. For the evaluation of the pseudo-parallel datasets collected with the *Semantic Textual Similarity* models, only the evaluation by the physicians was conducted. The goal was to ensure that the content preservation was evaluated by experts in the field, as lay people without the right field expertise were deemed unreliable in assessing it.

Due to the high cost of hiring professional healthcare personnel, we carried out the annotations in only one direction, from expert to layman. The reason for this choice is that text simplification tasks have been more widely studied in the past and are considered more important for real-world applications. Due to cost constraints, we selected only one of our TST models for evaluation. We made this decision based on the results of our automatic evaluation (Section 6.1). We examined the results of the pseudo-parallel datasets collected with respect to similarity quantiles and focused on models trained on datasets collected at the 85% quantile. This was because the parallel sets at the 85% quantile score were closer to the results obtained on the gold test set and the TST models trained on them generally showed good balance between content preservation and style evaluation. Among all the TST models trained on the pseudo-parallel sets (Section 4.1), we chose the one trained on the **cb_mqp_csts1** set (at 85% quantile) because it achieved higher content preservation scores on average. We also compared the outputs of this model and the model trained on **cb_msd_swap** (at 85% quantile), which performed similarly. We found that the former tended, in some cases (especially when the input sentence was relatively short), to generate more accurate explanations of medical terms (see Section 6.4). Regarding perplexity metrics, the trends were not clearly separable, so they did not influence our final decision.

To compare our system with state-of-the-art models, we selected the **Style Transformer** as our competitor. This decision was made for two reasons. Firstly, among the previously proposed models in the literature for *Expertise Style Transfer*, the Style Transformer showed more consistent results in terms of the balance between content preservation and style strength, as noted in Cao et al. (2020). Secondly, its architecture and training approach are similar to our unsupervised model, allowing us to demonstrate the improvement of our methodology compared to unsupervised approaches. In particular, we excluded our unsupervised model from the human evaluation due to its high content preservation scores, which resulted in outputs that were mostly repetitions of the inputs. Moreover, we included the gold lay references in the comparison to evaluate the models' proximity and accuracy compared to the gold references. The annotators were not aware of which system generated which text.

During the evaluation process, we excluded samples with source texts that were less than 5 or more than 32 tokens, as well as samples where either of the models produced an output that was identical to the source text. This ensured a fair comparison between the models and reduced the annotators' workload, eliminating trivial examples. We also used different annotation protocols for lay people and experts and asked both groups to judge the same texts.

5.2.1. Annotation protocol for layman

We hired ten lay individuals who were proficient in English but lacked a background in medicine or related fields. We asked them to select the easier text to understand from each pair, which consisted of the source text and one of the system outputs (or the reference text). To reduce bias, the pairs were shuffled before being presented to the annotators.

Each subject annotated 30 samples, consisting of 3 pairs for each system. There was an overlap of 10 samples with another subject to evaluate the agreement between annotators, resulting in 250 annotations. We measured the agreement between annotators using Cohen's Kappa (K^{lay}) (Cohen, 1960) and evaluated the style transfer as the ratio of texts judged to be easier to understand than the related source text (Sty^{lay}).

5.2.2. Annotation protocol for expert

We hired four physicians proficient in English from the Department of Orthopaedic Surgery of University Campus Bio-Medico of Rome, Italy. The expert evaluation consisted of two sets of annotations. The first set evaluated the content preservation of the collected pseudo-parallel data, while the second set evaluated the style transfer of the outputs of the style transfer systems.

They were asked to assess the content preservation for both sets and the style strength for the outputs, based on the quality of the changes made to the original style. The experts were asked to judge the texts independently, taking into account the specific guidelines provided. Regarding content preservation, they were instructed to assess content preservation based on the guidelines followed in previous literature (Wang, Afzal et al., 2020) To assess style strength, they were asked to consider terminology and empirical evidence knowledge gaps, as highlighted in Cao et al. (2020), while ignoring fluency issues and content as well. The questions and answers included in the protocols are summarized in Table 4, along with the scores associated with each answer.

For evaluating the pseudo-parallel data, a total of 350 samples were presented to the annotators, consisting of one expert sentence and its lay counterpart. These samples were randomly selected from one of the quantile-dependent sets collected with **cb_mqp_csts1**. We excluded samples from the 99% quantile, which contained pairs of identical texts. This protocol was designed to analyze how the quality perceived by physicians changes across different quantile ranges.

The evaluation of the outputs of the three systems was done by presenting 250 samples to the annotators. Each sample consisted of one source text and three rephrased texts. To decrease the cognitive effort required, the source text and all outputs were presented on the same page to the annotators. This approach may have introduced bias as the annotators were not asked to perform a *relative rating* (Briakou, Agrawal, Zhang, Tetreault, & Carpuat, 2021) between systems. We thus evaluated the content preservation (Cnt) and the style strength (Sty) by looking at the average scores and their ranking comparisons.

To assess the consistency between the annotators, we presented a subset of 100 training samples and 50 outputs samples to both the physicians involved in the annotation phase of training sets and outputs, respectively. We utilized the quadratic weighted version of the *Cohen's Kappa* (K_w) (Cohen, 1968), which just require the distribution of the distance between two annotations to be ordinal (Vanbelle, 2016). By considering y_i and y_j as the annotations made by annotators i and j, respectively, and computing the weights using the following equation

$$w_{i,j} = \frac{(y_i - y_j)^2}{(N-1)^2} \tag{9}$$

where N is the number of choices, we can measure the agreement while taking into account the severity of the disagreement between annotators. Trivially, a disagreement between annotators evaluating a pair of texts as *completely equivalent* and *unimportant details differ* is weighted less than a disagreement between *completely equivalent* and *completely dissimilar* (refer to Table 4).

6. Results and discussion

In this section, we present and analyze all of the results, including both automatic and human evaluations. The annotations provided us with the opportunity to examine the correlation between automatic scores and human judgments. The input from the annotators also facilitated a qualitative analysis, where we highlighted the key aspects of the style transfer task and the results obtained from the models.

Table 4
Questions and answers, included in the expert protocols, for the evaluations of content preservation and style strength. On the left of each answer, the associated score is reported.

	Content preservation
Q:	To what extent is the rewritten text still conveying the same content as the source text?
0:	The two texts are *completely dissimilar*.
1:	The two texts are not equivalent, but are on the *same topic*.
2:	The two texts are not equivalent, but share *some details*.
3:	The two texts are roughly equivalent, but some *important information differs/missing*.
4:	The two texts are mostly equivalent, but some *unimportant details differ*.
5:	The two texts are *completely equivalent*, as they mean the same thing.
	Style strength
Q:	To what extent the process of rewriting for a lay audience can be considered a good attempt?
0:	The rewriting process is not a good attempt, performing *no changes* from the source text.
1:	The rewriting process is not a good attempt, performing *some changes* that are *not good* for the scope.
2:	The rewriting process made some *minimal good changes* but the rewritten text still mostly targets an expert audience.
3:	The rewriting process made quite *substantial changes*, although there are some elements for an expert audience.
4:	The rewritten text *really* targets a lay audience.

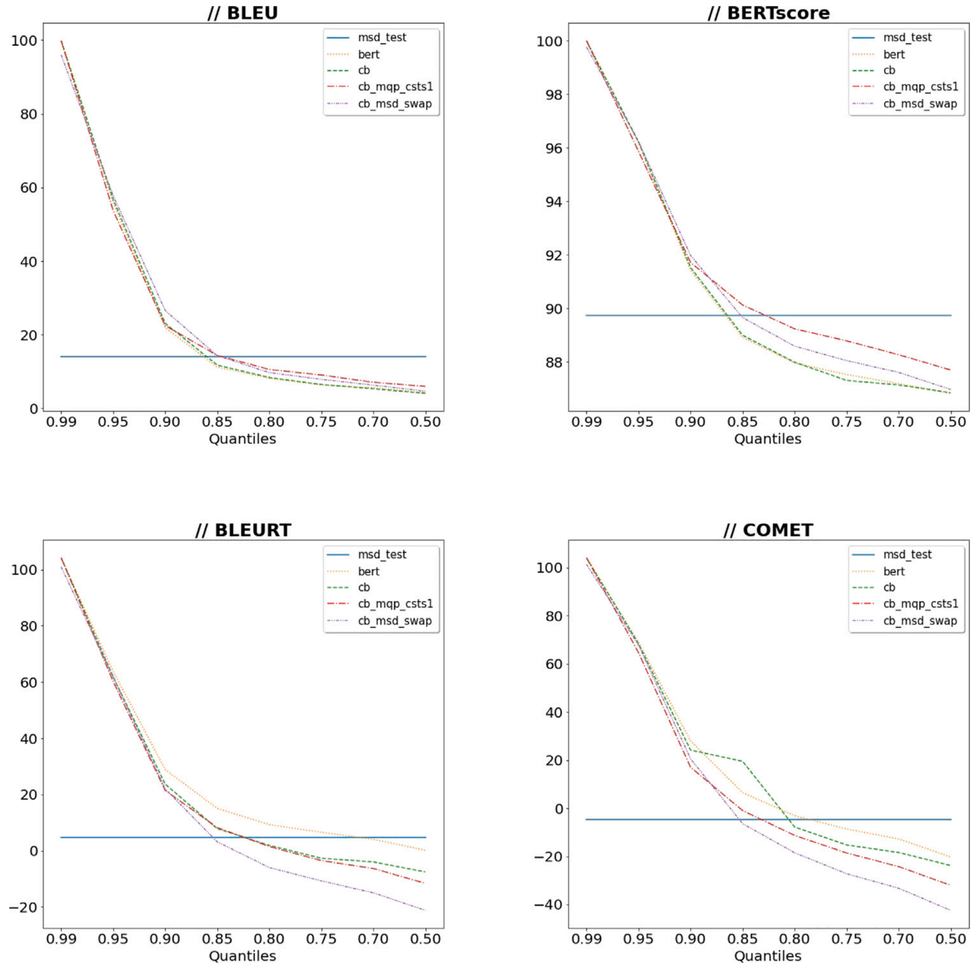

Fig. 3. Automatic content preservation metrics (in terms of %) for the collected parallel sets, indicated with the // symbol over the quantile ranges. The most relevant models are reported. The blue solid horizontal line indicates the score computed between the source and the gold reference.

6.1. Automatic evaluation

With regards to the automatic evaluation, we provide plots to understand the impact of different training sets on TST performance. Fig. 3 displays the content preservation metrics for the collected parallel training sets, indicated with the symbol "//". The same metrics were assessed for the TST system outputs in relation to both the source and target, denoted with the "self-" and "ref-" prefixes, respectively, and are shown in Figs. 4 and 5. Fig. 6 reports the automatic metrics assessing the style strength and perplexity of the outputs. We present the (pseudo-)perplexities using the original *(Bio-)ClinicalBert* and its fine-tuned versions on the lay or expert corpora. For clarity, we only show the most relevant models in the plots, and each subfigure presents the scores of the *MSD* test set as the baseline(s) and the best-performing state-of-the-art models in terms of content preservation (*ControlledGen*), stability across content preservation and style strength (*StyleTransformer*), and our unsupervised *BART* model as competitors. The results for all systems can be found in Tables A.1 and A.2 in the appendix.

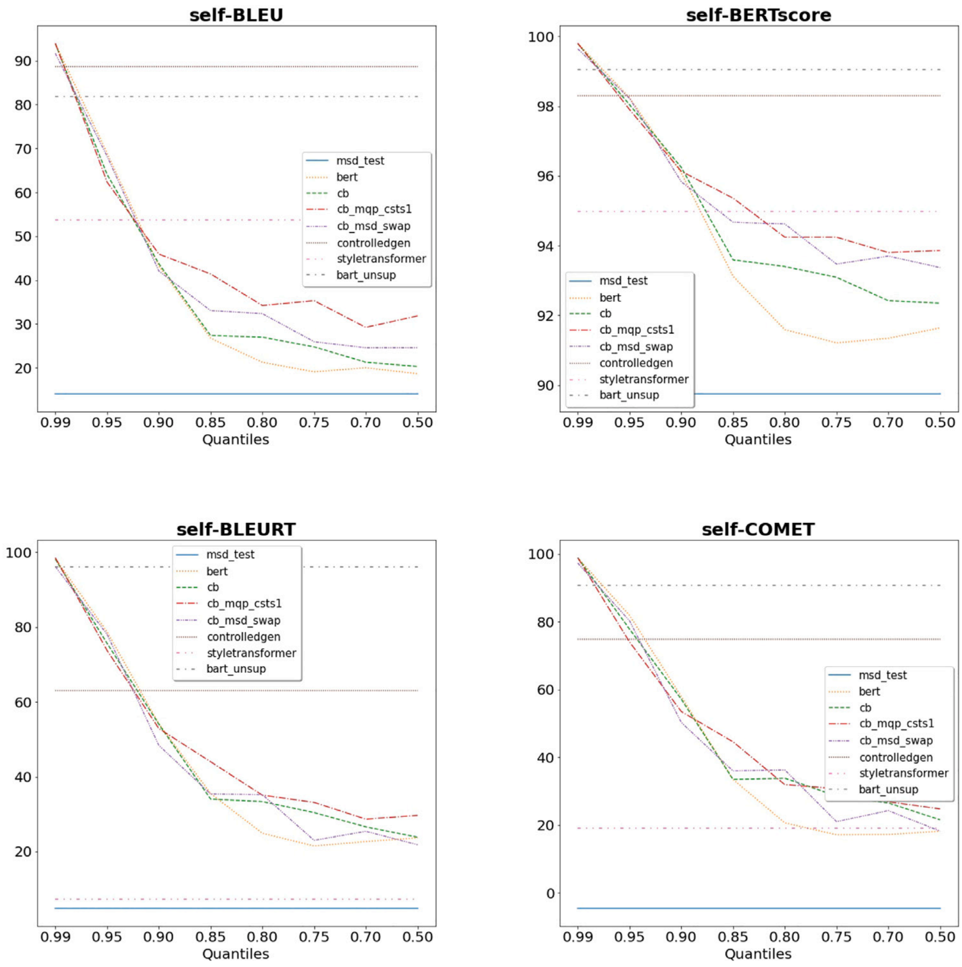

Fig. 4. Automatic content preservation metrics (in terms of %) over the quantile ranges for the most relevant models, computed with respect to the source (*self-*). The blue solid horizontal line indicates the score computed between the source and the gold reference, while the other horizontal lines refer to the competitors.

6.1.1. *Automatic evaluation of pseudo-parallel data*

The results in Fig. 3 indicate that the different automatic metrics share similar trends, although their values may vary. The *// BERTscore* has a limited range compared to the other metrics, while, as expected, *// BLEU* does not show negative values like *// BLEURT* and *// COMET*. The pseudo-parallel datasets collected with different models exhibit similar trends across all metrics, with similar values in the beginning, which suggests that the sets overlap more at high quantiles regardless of the model used for collection. After the 90% quantile, the differences in the collected datasets are captured by the metrics, except for *// BLEU*. As previously discussed in Section 5.2, the sets are generally closer to the test set around the 85% quantile, regardless of the metric used for evaluation.

6.1.2. *Automatic evaluation of systems' outputs*

In Figs. 4 and 5, it is evident that the characteristics of the parallel training sets have an impact on the performance of the style transfer systems. The metrics used to measure content preservation in the outputs show a similar trend, with the metrics denoted with the *self*-prefix showing a closer range of values compared to those denoted with the *ref*- prefix. Specifically, the *ref*-metrics have a lower starting point and a gentler slope, with the exception of the *ref-BERTscore*, which shows a modest increase at first. However, its variations are still limited. At lower quantiles, the different metrics present different rankings for the models, with the model trained on the set collected using the *cb_mqp_csts1* model showing higher ranking positions for both the *self*- and *ref*-metrics.

However, the low content scores of the gold references highlight the difficulty of the task. The models generally achieved higher *self*-scores, which may indicate either their superiority over the human references or that the *self*-metrics are not appropriate for this task.

The trend observed in the first plot of Fig. 6 is reversed when considering the style strength accuracy metric calculated using the TextCNN style classifier. This difference in behavior can be attributed to the parallel dataset used to train the models. At high quantiles, the parallel texts are considered very similar or even identical, as demonstrated by the close to 100% *// BLEU* scores. However, at the lower quantiles, the texts are too dissimilar from one another. At one extreme, the model was trained to mainly reproduce the input, while at the other extreme, the model was trained to generate outputs that are too dissimilar from the source, but closer to the desired style. It is worth noting that, with the early quantiles, the content preservation performance of our model outperformed the state-of-the-art systems. However, in the following quantiles, our model outperformed the style strength of state-of-the-art systems, excluding the score achieved by the *DeleteAndRetrieve* system, which was found to be the worst in terms of content preservation, as indicated in Appendix.

The results of the different models on the style transfer task suggest that there is a trade-off between style strength and content preservation. Models trained on non-fine-tuned models, such as **cb** and **bert**, showed higher style strength, but worse content preservation (especially in *self*-metrics) compared to other models. This may indicate that non-specialized *Semantic Textual Similarity* models tend to collect less related text pairs to train TST models, that thus produce outputs that are less related to the input, resulting in a higher variability.

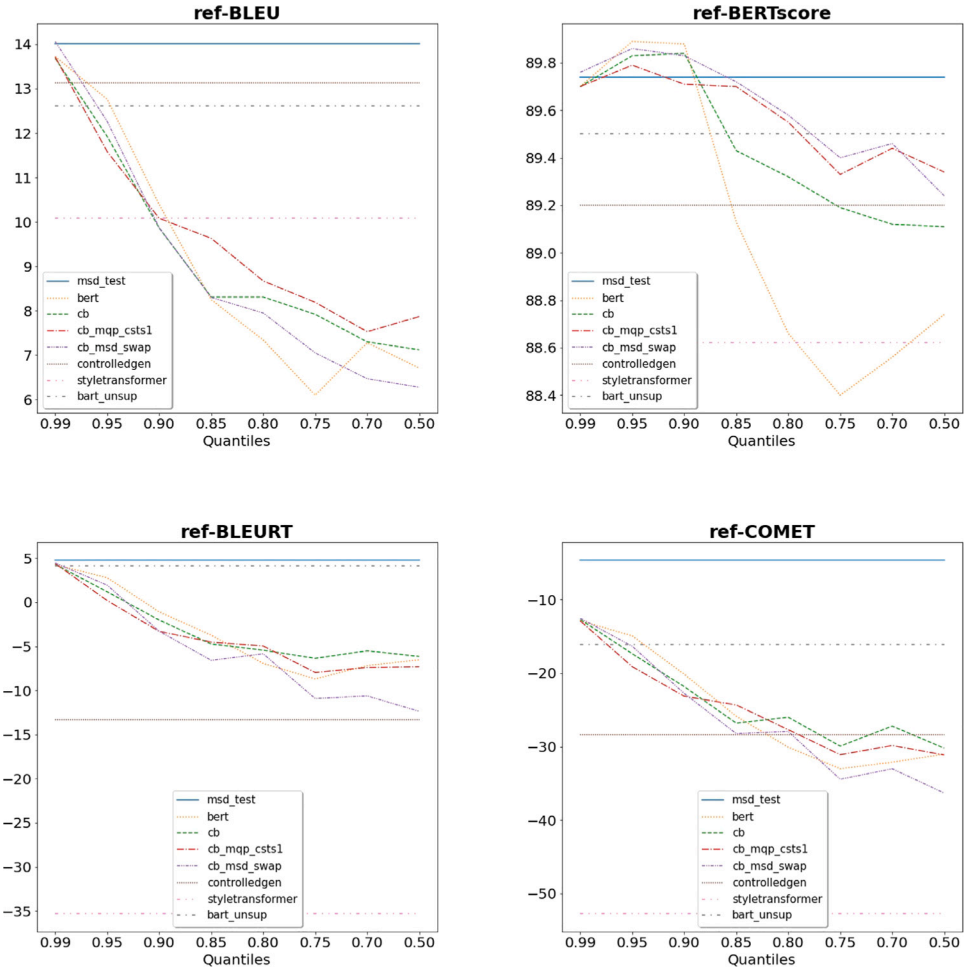

Fig. 5. Automatic content preservation metrics (in terms of %) over the quantile ranges for the most relevant models, computed with respect to the gold reference (*ref-*). The blue solid horizontal line indicates the score computed between the source and the gold reference, while the other horizontal lines refer to the system competitors.

The *ControlledGen* and unsupervised *BART* models, on the other hand, showed strong content preservation, but poor style strength. This trade-off between content preservation and style strength explains why the *StyleTransformer* was chosen as the model as our competitor during human evaluations. Unfortunately, none of the models were able to reach the level of style strength seen in the test set. However, the high accuracy reached by the style classifier on the test set confirms the ability of this model to distinguish between expert and lay styles.

The plots in Fig. 6 display the perplexities calculated using *(Bio-)ClinicalBert* ($pPPL$) and its fine-tuned models on the lay and expert corpora ($pPPL_{lay}$ and $pPPL_{exp}$, respectively). These metrics show a trend that is similar to the trend observed for the content preservation metrics. As the quantile ranges increase, the perplexity decreases, reflecting the increased variability in the related training set. This behavior is more pronounced for the perplexity model that was fine-tuned on the lay corpus. Interestingly, the $pPPL_{lay}$ scores are higher than the $pPPL_{exp}$ scores for each model. This suggests that the lay outputs generated by the models are more similar to the expert training corpus than to the lay corpus. The lay test corpus is the exception for it (while the expert test corpus shares this behavior as was expected), reflecting that the test set was extracted from the same corpus from which the training dataset was collected. At lower quantiles, the perplexity metrics for models trained on sets collected using fine-tuned models are similar to those for models trained on sets collected using non-fine-tuned models, and vice versa. This can be attributed to the increased variability of the training sets obtained using non-fine-tuned models. This behavior can be attributed to the variability of the training sets. The use of non-fine-tuned models to collect parallel sets results in pairs of more dissimilar sentences, which increases the variability of what the model has seen during training. In general, our models have lower perplexities than the state-of-the-art models and our unsupervised BART baseline, regardless of the quantile range, demonstrating their effectiveness.

6.2. Human evaluation

The results of the human evaluation of our model (based on the training set collected with **cb_mqp_csts1** at the 85% quantile), compared with the state-of-the-art (*StyleTransformer*) and gold references, are reported in Table 5.

6.2.1. Human evaluation agreements

Before analyzing the results, the quality of the annotation process was assessed by measuring the agreement between annotators. The lay annotators evaluated the style as a binary task, choosing the easier-to-understand text among the source and a system output. The average Cohen’s Kappa (K^{lay}) was .32 (±.15), which can be considered fair agreement as suggested by previous literature (Landis & Koch, 1977). However, the high standard deviation suggests that some pairs are easier to annotate than others. The agreement on the individual systems was also evaluated. It is worth noting that, despite the fact that the annotators agreed the most on the outputs generated by *StyleTransformer* (ST), this model achieved the lowest score for lay people in terms of style-related performance. This means that the lay annotators

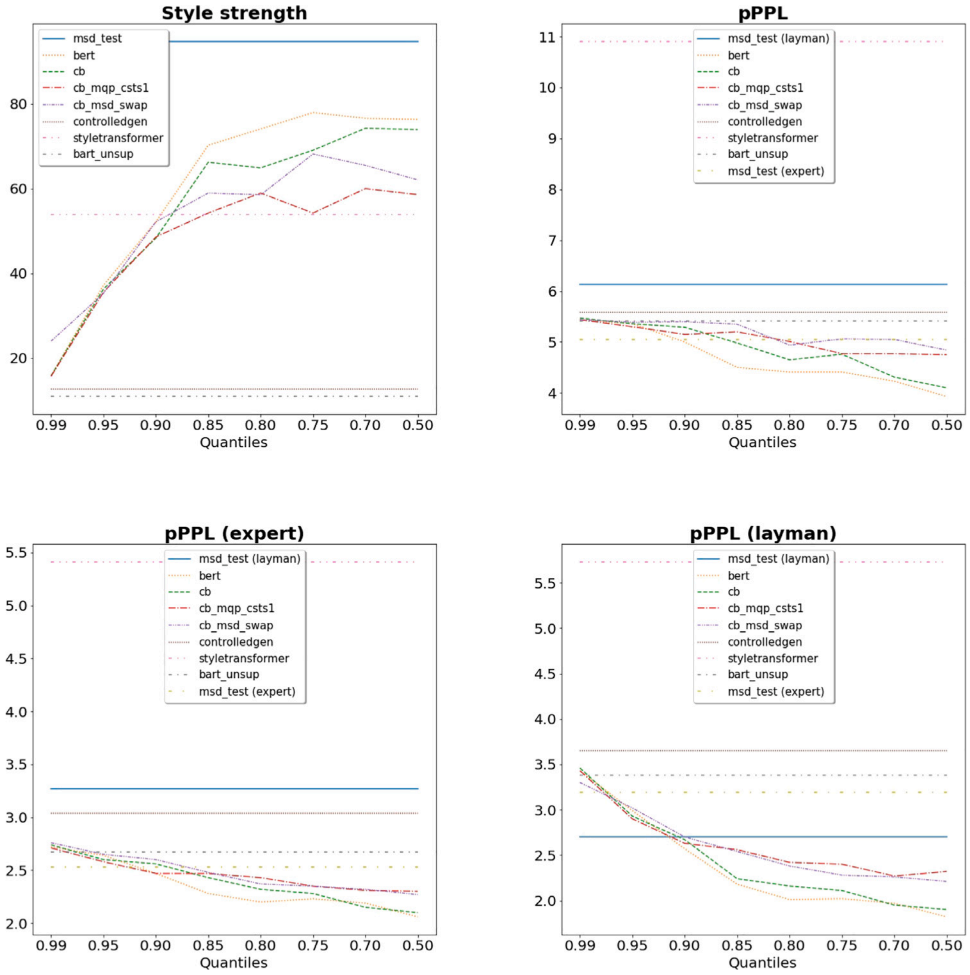

Fig. 6. Automatic style strength metric (in terms of accuracy percentage) and (pseudo)perplexity metrics. The latter were computed using a *(Bio-)ClinicalBert* masked language model ($pPPL$) and its fine-tuned versions on expert and lay corpora ($pPPL_{exp}$ and $pPPL_{lay}$, respectively).

mostly agreed its outputs was not easier to understand compared to the source. This aspect was discussed in more detail in the qualitative analysis Section 6.4. On the other hand, our model achieved higher results, comparable to the gold reference, indicating that it was able to effectively make changes in the direction of simplification.

Unlike lay annotators, we decided to have experts to judge using a scale. Thus, the original Cohen's Kappa was not appropriate for measuring their agreement. We thus applied its quadratic weighted version (as described in Section 5.2), instead. Plus, we asked the physicians to judge not only the style but the content preservation too. We assessed the agreement score for both content (K_w^{cnt}) and style (K_w^{sty}), separately. However, since the weighted Cohen's Kappa interpretation is debated, with its results influenced by the weight scale (Vanbelle, 2016), we also assessed the Spearman correlation indices (ρ^{cnt} and ρ^{sty}). The expert annotators showed moderate agreement, with a Kappa score of .42 for content preservation and .50 for style strength. The Spearman scores support these results (Shrout, 1998). We also looked at the agreement scores for each system individually. The annotators were more in agreement on the content preservation of the system outputs than on the gold reference (Ref). This indicates that the reference may have undergone more changes from the source text, leaving room for more interpretation by the annotators. For the style analysis, the results from the expert annotators were similar to those obtained from lay people. Regarding the style analysis, the expert annotators showed outcomes analogous to the ones obtained with lay people.

6.2.2. Human evaluation of systems' outputs

Moving to the proper evaluation analysis, our model ($Ours$) performed better on average compared to the state-of-the-art model, further highlighting the improvements brought by our approach. While our model's content preservation scores were even higher than the reference, its style scores were still lower. This suggests that our model prioritizes maintaining the meaning of the input over making significant changes to the text. While this is not ideal, it is preferable to avoid losing information, even if it means making only minor changes that may not be as noticeable to a layperson. In this sense, it can hardly compete with the abstraction level of the gold references. The heatmaps in Fig. 7 show that it outperforms the reference in content preservation, but is not as good as the reference in changing the style. while both our model and the reference outperformed the *StyleTransformer* in both content preservation and style, as well as overall. Additionally, despite often making only minimal changes, the outputs from our model were still rated as easier to understand than the source texts by lay people, which is a significant improvement compared to the state-of-the-art. In the lay evaluation, in particular, the outputs from our model were found to be very similar to the gold references.

6.2.3. Experts' evaluation of pseudo-parallel data

In addition, we performed an expert evaluation to assess the content preservation quality of the parallel sets at different quantiles. To determine the agreement between annotators, we used 100 samples and calculated the quadratic weighted version of Cohen's Kappa and the Spearman correlation. The annotators showed moderate to substantial

Table 5
Evaluation results for the gold reference (Ref), the StyleTransformer (ST), and our model ($Ours$), as well as the three systems together (All). The first block regards the agreement between annotators assessed with a given number of samples (#). For lay annotators, the agreement is assessed with Cohen's Kappa (K^{lay}), while for the experts it is measured with the quadratic weighted version (K_w) and the Spearman correlation index (ρ), for both content preservation (cnt) and style strength (sty). The second block reports the human evaluation results (in terms of percentages) of the different systems for lay and expert annotations. For the former case, the style is evaluated as the ratio between the number of texts judged easier to understand than the related source text (Sty^{lay}). For the latter, both content and style scores are normalized with the range of the related scale. The third block is dedicated to the automatic (self-)metrics computed with respect to the source text and the style strength. The best results for each metric are shown in **bold**.

System	#	Agreement					Human evaluation			Automatic evaluation				
		K^{lay}	K_w^{cnt}	K_w^{sty}	ρ^{cnt}	ρ^{sty}	Sty^{lay}	Cnt	Sty	$BLEU$	$BERT$	$BLEURT$	$COMET$	SS
Ref	50	.26 ± .36	.24	.21	.41	**.31**	**69.00**	65.12 ± 29.23	**71.60 ± 27.59**	14.01	89.74	4.77	−4.62	**94.67**
ST	50	**.31 ± .45**	**.63**	**.34**	.62	**.31**	28.50	62.00 ± 25.35	31.35 ± 17.15	**53.66**	94.98	7.38	19.02	53.93
Ours	50	.16 ± .27	.57	.20	**.66**	.19	66.50	**79.48 ± 20.42**	48.80 ± 27.95	41.40	**95.36**	**43.99**	**44.56**	54.22
All	150	.32 ± .15	.42	.50	.50	.52	–	–	–	–	–	–	–	–

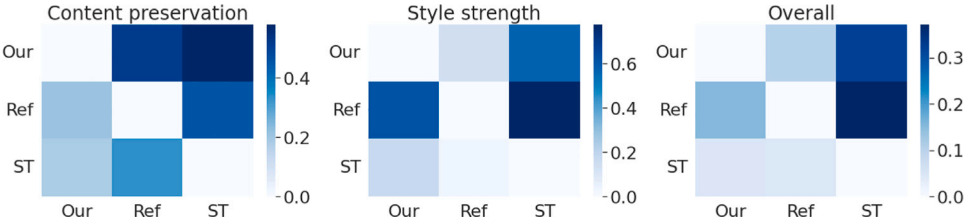

Fig. 7. Ranking comparison regarding human evaluations for content preservation, style strength, and a combination of the two (overall). The darker the color of the cell (i,j), the more times the ith system (on the y-axis) was ranked better than the jth model (on the x-axis) on the same sample. Note that the sum between the cell (i,j) and the cell (j,i) is lower than 1 because of cases of draws. For the same reason, the diagonal is represented by all zeros.

agreement (K_w^{cnt} = .60, ρ^{cnt} = .64). The results of the evaluation are shown in Fig. 8, which displays the normalized average and standard deviation of the content preservation scores for the expert annotators. The average content preservation score generally decreases as the quantile threshold decreases, but the trend becomes less clear for lower quantiles, especially between 70% and 50%. This is also reflected in the automatic metrics, which show similar values for the lower quantile-related parallel sets (Fig. 3). The same pattern is observed in the *self*- and *ref*-metrics for the style transfer task (Figs. 4 and 5). These results suggest that decreasing the threshold below a certain value leads to parallel sets of similar low quality.

The results of our evaluation indicate valuable insights, even though direct comparison with the outputs and pseudo-parallel sets annotation tasks is not feasible due to varying setups and annotators. The higher level of agreement between annotators implies that they may be more in agreement about the quality of content preservation in parallel corpora automatically collected compared to those generated by a model or even those annotated by human experts (gold references). These results suggest that our pipeline for collecting parallel corpora for the style transfer task in the medical domain can be applied effectively based on the similarity threshold, and could potentially be used as a pre-annotation phase to minimize the annotators' workload and limit the changes between source and target texts.

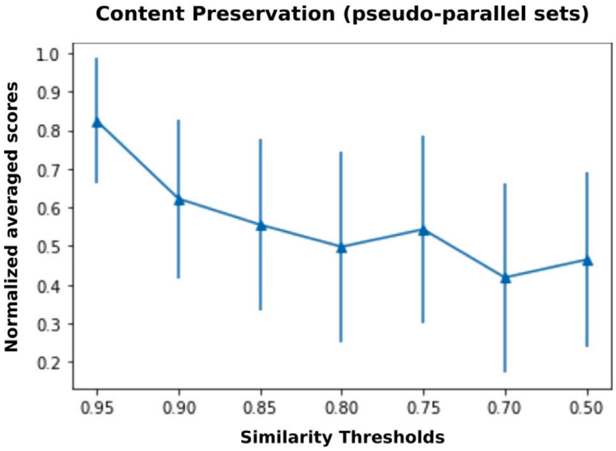

Fig. 8. Human evaluation results in content preservation for the collected pseudo-parallel datasets over the quantile thresholds. The results are reported in terms of the normalized average and standard deviation scores.

6.3. Comparing automatic and human evaluations

The comparison of the automatic evaluation scores with the human evaluation scores in Table 5 reveals some interesting findings. Although the comparison of BLEU scores between the *StyleTransformer* and our model are not directly comparable, the automatic *self*-metrics for content preservation tend to have a similar behavior as the human evaluation scores. To further explore the agreement between the automatic and human evaluation metrics, we analyzed the correlation between them for both content preservation and style strength. Table 6 shows the Spearman correlations for *self*- and *ref*-content metrics (ρ^{self} and ρ^{ref}, respectively), as well as for style (ρ^{ss}). The results show that the correlation between the *self*-metrics and human judgments is higher compared to the correlation between the *ref*-metrics and human judgments. This is consistent with past literature (Lai, Mao et al., 2022). Overall, BLEURT and COMET are the metrics that show the highest correlation with human judgments, both in the *self*- and *ref*-setting. It is also worth noting that the reference texts (Ref) have a lower correlation with *self*-metrics compared to the two models, which highlights the aggressive differences between the reference texts and the associated source texts. Furthermore, the *StyleTransformer* model (ST) shows higher correlation scores, suggesting that there is a stronger

Table 6
Spearman correlation scores between expert human judgments and automatic metrics for the gold reference (Ref), the StyleTransformer (ST), and our model ($Ours$), as well as the three systems together (All). The # column reports the number of samples used to assess the correlation scores. For content preservation scores, we reported correlation involving both self- (ρ^{self}) and ref- (ρ^{ref}) metrics. The last column instead assesses the correlation (ρ^{ss}) between the style annotations and the outputs of our trained style classifier. The best content-related correlations for each system are shown in **bold**.

System	#	Humans-BLEU		Humans-BERT		Humans-BLEURT		Humans-COMET		Humans-SS
		ρ^{self}	ρ^{ref}	ρ^{self}	ρ^{ref}	ρ^{self}	ρ^{ref}	ρ^{self}	ρ^{ref}	ρ^{ss}
Ref	250	**.45**	–	.42	–	.41	–	.43	–	.10
ST	250	.59	.21	**.68**	.22	.67	**.43**	.67	.42	- .03
Ours	250	.64	**.34**	.60	.21	.62	.30	**.65**	.30	.20
All	750	.39	.26	.45	.27	**.60**	**.46**	.58	.44	.34

correlation between automatic metrics and human evaluations when judging a less-performing system.

When examining the results for the style aspect, a noticeable feature is the low correlation score for the *StyleTransformer*. This is likely due to the model's strategy of replacing complex terms with simpler but often unrelated words, which are evaluated as simplifications by the classifier that is less influenced by the outputs' meaning and fluency than humans. Additionally, the correlation score for reference texts, which the style classifier was able to identify well, is notably low. This highlights the difficulty for humans in assessing the style strength, separating it from the structure and semantics. These findings are in line with recent studies in the field (De Mattei, Cafagna, Dell'Orletta, & Nissim, 2020).

6.4. Qualitative analysis

Our manual inspection was conducted on a significant number of examples and incorporated the feedback from the expert annotators. The three models we focused on were: our *cb_mqp_csts1* model at 85% quantile, the *StyleTransformer* system, and a model based on the *cb_msd_swap* dataset (at 85% quantile) which had similar performance to our model. We did not take into consideration the models' outputs that were not mere repetitions of the input. The results of our analysis allowed us to draw some qualitative conclusions. Firstly, we observed that each model employed different strategies to simplify the text. In instances where our models could not substitute a complex term, they attempted to provide an explanation:

> **Source:** Pulmonary arteries are affected, sometimes causing pulmonary hypertension.
>
> **StyleTransformer:** Pulmonary arteries are affected, sometimes causing intravenously recurring.
>
> **cb_mqp_csts1 (85%):** Pulmonary hypertension is a condition in which blood pressure in the lungs is too high.

This behavior is particularly noticeable for shorter or incomplete sentences:

> **Source:** IV fluids.
>
> **StyleTransformer:** common fluids.
>
> **cb_mqp_csts1 (85%):** IV fluids are given intravenously.
>
> **cb_msd_swap (85%):** Blood and urine are given intravenously.

Of course, not all of the found examples are good attempts, as sometimes the models produced irrelevant or meaningless explanations:

> **Source:** Biopsy.
>
> **cb_mqp_csts1 (85%):** Biopsy is the most common type of bleeding disorder.
>
> **cb_msd_swap (85%):** Biopsy is given intravenously.

Despite that, these examples indicate that our models tend to provide explanations when unable to substitute significant terms, and at times, exhibit good domain knowledge. Additionally, we discovered some domain knowledge related to gender in our models, which was interesting.

> **Source:** Most *patients* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.
>
> **StyleTransformer:** Most *people* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.
>
> **cb_mqp_csts1 (85%):** Most *women* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.

Another trend that we observed, is that all models tend to replace the word *patient* with a less domain-specific term such as *people* (or *woman*, as in the previous example). This may indicate that the lay corpus does not use the word "patient" in the same way as the expert corpus. However, being a common word, this change was often not considered by annotators as valid. Additionally, we noticed that our models were better at handling certain abbreviations, such as *hr* and *yr*, while the *StyleTransformer* model struggled with these:

> **Source:** Jaundice usually peaks within 1 to 2 *wk*. Recovery phase: During this 2- to 4-wk period, jaundice fades.
>
> **StyleTransformer:** Jaundice usually peaks within 1 to 2 *relieving*.
>
> **cb_mqp_csts1 (85%):** Jaundice usually peaks within 1 to 2 *weeks*.
>
> **cb_msd_swap (85%):** Jaundice usually disappears within 1 to 2 *weeks*.

The above example shows that the models tend to simplify the structure of the source text when it is complex. Specifically, if the source text consists of multiple sentences, the models often truncate the output, removing either the left or right context. We believe it to be due to the one-sentence nature of the training corpus and the limited input token lengths of the models.

In particular, when the *StyleTransformer* model encounters stylistic change, its output often appears messy. This is partially confirmed by higher perplexities and the annotators' results and feedback. The lay annotators frequently commented on the presence of poorly structured sentences in the model outputs. Despite the model's attempt to simplify the input, it often made the sentences difficult to comprehend, leading the annotators to prefer the original expert text. This posed a challenge for the annotators when the source text was already hard to understand and the meaning of the model's output was clearer, despite being messy.

Some annotators found shorter sentences easier to understand, while others preferred longer texts for their added context. Choosing between the two was often complicated by differences in meaning caused by extra information in one of the texts, too. Moreover, minimal changes, such as capitalization of common terms or the substitution of common words (e.g., *patients*), were often randomly judged by lay annotators, while experts typically considered these situation as *no good changes* to *no changes* when no other alterations were made.

7. Conclusion

Our study has demonstrated the effectiveness of our Text Style Transfer system in improving the communication between physicians and patients, alleviating the issues arising from the *curse of knowledge*. By leveraging a pre-trained denoising autoencoder (BART) model

Table A.1
Results of the automatic evaluations of our models with respect to the collected parallel training sets. Both sets and models were evaluated at various quantile thresholds. For the // metrics, with the **bold** font we indicate the values closer to scores obtained on the test set. For the others, we used it to indicate the best scores obtained. In particular, the values in red indicates the state-of-the-art scores. For the test set, the perplexity scores of both lay and expert corpora are reported in this order.

TST Model	STS Model	Quantile	// BLEU	self-BLEU	ref-BLEU	// BERT	self-BERT	ref-BERT	// BLEURT	self-BLEURT	ref-BLEURT	// COMET	self-COMET	ref-COMET	SS	pPPL	pPPL (lay)	pPPL (exp)
Test set	–	–	14.01	14.01	14.01	89.74	89.74	89.74	4.77	4.77	4.77	−4.62	−4.62	−4.62	94.67	6.14/5.05	2.70/3.19	3.27/2.53
BART (unsup)	–	–	–	81.78	12.62	–	99.04	89.50	–	96.04	4.14	–	90.75	−16.07	10.96	5.41	3.38	2.67
BART	bert	0.99	99.84	**93.94**	**13.72**	100.00	**99.80**	**89.70**	104.21	**98.26**	**4.26**	104.07	**98.79**	**−12.86**	15.85	5.43	3.43	2.72
		0.95	55.87	68.94	12.76	96.17	98.24	89.89	63.79	78.67	2.77	68.93	81.86	−14.93	37.19	5.36	2.99	2.64
		0.90	21.92	43.27	10.39	91.42	96.08	89.88	28.82	54.14	−1.09	27.97	57.90	−20.14	52.15	5.00	2.57	2.47
		0.85	**11.22**	26.75	8.25	**88.91**	93.12	89.13	15.04	35.50	−3.74	6.47	33.46	−25.88	70.22	4.50	2.18	2.28
		0.80	8.20	21.27	7.34	87.95	91.59	88.66	9.27	24.88	−6.94	**−3.08**	20.65	−30.08	74.07	4.41	2.01	2.2
		0.75	6.36	19.09	6.10	87.52	91.21	88.40	6.51	21.49	−8.69	−8.64	17.13	−33.00	**77.93**	4.41	2.02	2.23
		0.70	5.60	20.02	7.28	87.18	91.34	88.56	**4.01**	22.66	−7.18	−12.79	17.20	−32.12	76.59	4.23	1.97	2.19
		0.50	4.23	18.66	6.71	86.82	91.64	88.74	0.12	23.68	−6.53	−20.27	18.21	−31.02	76.30	**3.93**	**1.82**	**2.06**
BART	cb	0.99	99.88	**93.85**	**13.68**	99.99	**99.79**	**89.70**	104.21	**98.09**	**4.29**	104.03	**98.85**	**−12.69**	15.85	5.47	3.46	2.74
		0.95	56.77	64.08	11.92	96.22	98.04	89.83	61.81	75.60	1.18	68.26	77.80	−17.36	36.30	5.36	2.93	2.60
		0.90	23.09	43.73	9.86	91.52	96.24	89.84	23.65	54.02	−2.02	24.15	57.13	−21.83	48.30	5.29	2.67	2.56
		0.85	**11.75**	27.38	8.31	**88.99**	93.59	89.43	7.80	34.02	−4.74	19.48	33.46	−26.81	66.22	4.98	2.24	2.43
		0.80	8.38	26.98	8.31	87.98	93.40	89.32	**1.85**	33.35	−5.44	**−7.81**	33.84	−26.01	64.89	4.65	2.16	2.32
		0.75	6.49	24.78	7.92	87.30	93.09	89.19	−2.73	30.41	−6.36	−15.29	28.59	−29.93	69.04	4.76	2.11	2.28
		0.70	5.31	21.28	7.30	87.13	92.42	89.12	−4.03	26.59	−5.51	−18.41	26.53	−27.21	**74.22**	4.31	1.95	2.15
		0.50	4.10	20.30	7.12	86.84	92.35	89.11	−7.60	23.83	−6.15	−23.82	21.56	−30.21	73.93	**4.10**	**1.90**	**2.10**
BART	cb_csts1	0.99	99.87	**93.95**	**13.72**	100.00	**99.80**	89.70	104.21	**98.15**	**4.26**	104.05	**98.82**	**−12.81**	15.85	5.49	3.47	2.75
		0.95	54.57	66.77	12.40	95.76	98.08	**89.89**	59.39	75.53	1.38	63.62	77.70	−16.64	35.56	5.39	2.98	2.64
		0.90	21.52	41.00	9.83	90.92	95.28	89.72	19.22	47.11	−2.35	14.09	47.09	−23.02	56.74	5.10	2.53	2.48
		0.85	**12.12**	30.31	8.50	**89.10**	93.83	89.50	**5.44**	32.65	−5.00	**−5.98**	33.17	−26.65	63.41	4.92	2.32	2.38
		0.80	9.02	30.89	7.90	88.34	94.20	89.58	0.17	35.33	−4.11	−15.09	34.54	−26.45	61.48	4.96	2.36	2.38
		0.75	7.66	34.14	8.20	87.90	94.28	89.47	−4.21	35.80	−5.01	−20.30	34.12	−27.26	58.81	4.88	2.35	2.36
		0.70	6.45	29.24	7.95	87.59	93.78	89.56	−6.30	32.14	−5.11	−24.76	29.44	−27.41	62.52	4.60	2.22	2.25
		0.50	5.02	29.01	7.62	87.11	93.49	89.36	−10.12	29.32	−7.19	−30.48	2 4.49	−29.44	**65.63**	**4.57**	**2.16**	**2.25**
BART	cb_mqp_csts1	0.99	99.86	**93.97**	**13.71**	100.00	**99.80**	89.70	104.22	**98.48**	**4.37**	104.06	**98.81**	**−12.89**	15.70	5.44	3.43	2.71
		0.95	53.70	62.35	11.58	95.86	97.90	**89.79**	60.17	73.71	0.19	64.64	74.00	−19.12	35.56	5.30	2.90	2.58
		0.90	22.48	45.94	10.08	91.71	96.13	89.71	21.41	52.87	−3.31	17.01	53.49	−23.14	48.59	5.15	2.63	2.47
		0.85	**14.39**	41.40	9.63	**90.12**	95.36	89.70	8.19	43.99	−4.52	**−0.97**	44.56	−24.33	54.22	5.20	2.56	2.47
		0.80	10.57	34.23	8.67	89.23	94.24	89.55	**1.44**	35.06	−4.97	−11.32	31.93	−27.70	58.96	5.01	2.42	2.43
		0.75	9.06	35.31	8.19	88.78	94.24	89.33	−3.56	33.13	−7.96	−18.61	30.75	−31.09	54.22	4.77	2.40	2.35
		0.70	7.09	29.22	7.53	88.26	93.8	89.44	−6.37	28.66	−7.40	−24.27	26.88	−29.85	**60.00**	4.77	2.27	2.31
		0.50	5.97	31.86	7.87	87.69	93.86	89.34	−11.65	29.66	−7.31	−32.05	24.78	−31.13	58.57	**4.75**	**2.32**	**2.30**
BART	cb_msd	0.99	99.86	**93.80**	**13.75**	100.00	**99.79**	89.70	104.22	**97.97**	**4.23**	104.09	**98.70**	**−12.90**	16.15	5.52	3.47	2.78
		0.95	58.51	67.87	12.50	96.33	98.22	**89.87**	62.04	78.72	2.08	69.00	80.77	−15.20	35.26	5.35	2.95	2.62
		0.90	27.24	47.88	10.16	92.13	96.58	89.75	22.64	55.73	−2.91	22.19	58.80	−22.43	44.59	5.27	2.73	2.55
		0.85	**14.11**	35.38	8.60	**89.76**	95.00	89.68	**3.93**	39.08	−5.17	**−4.69**	39.50	−26.45	56.74	5.13	2.51	2.48
		0.80	9.89	30.67	7.79	88.71	94.41	89.53	−5.42	33.48	−7.99	−17.40	33.51	−30.28	59.26	5.26	2.44	2.41
		0.75	7.90	29.98	6.87	88.10	94.39	89.41	−10.69	29.32	−9.91	−25.89	31.58	−32.46	52.89	5.03	2.44	2.39
		0.70	6.45	29.26	7.44	87.67	93.90	89.46	−14.34	28.48	−8.68	−31.65	29.17	−30.13	**61.63**	4.96	2.33	2.37
		0.50	4.71	24.18	6.44	87.00	93.42	89.31	−21.07	23.04	−10.62	−41.42	20.34	−35.33	60.30	**4.66**	**2.19**	**2.21**
BART	cb_msd_swap	0.99	96.05	**91.66**	**14.06**	99.76	**99.64**	89.76	100.94	**96.11**	**4.44**	101.33	**97.27**	**−12.52**	24.00	5.44	3.30	2.76
		0.95	57.78	68.24	12.26	96.23	98.22	**89.86**	61.69	78.02	1.93	67.74	80.25	−16.34	35.26	5.39	3.02	2.65
		0.90	26.58	42.08	9.87	91.97	95.83	89.83	21.94	48.37	−3.29	20.52	50.27	−22.76	52.15	5.40	2.70	2.60
		0.85	**14.31**	33.06	8.30	**89.65**	94.67	89.72	**3.11**	35.37	−6.59	**−6.42**	35.96	−28.23	58.96	5.35	2.54	2.48
		0.80	9.70	32.36	7.95	88.58	94.62	89.58	−6.04	35.19	−5.85	−18.53	36.26	−27.94	58.52	4.94	2.38	2.37
		0.75	7.89	25.95	7.05	88.04	93.47	89.40	−10.77	23.00	−10.92	−27.23	21.02	−34.44	**68.15**	5.06	2.28	2.35
		0.70	6.30	24.58	6.47	87.60	93.70	89.46	−15.02	25.40	−10.62	−33.26	24.27	−33.02	65.48	5.05	2.26	2.32
		0.50	4.65	24.59	6.28	86.96	93.37	89.24	−21.27	21.82	−12.37	−42.46	18.23	−36.32	62.07	**4.84**	**2.21**	**2.27**

Table A.2
Results of the automatic evaluations of the state of the art models are reported, as in Table A.1: with the **bold** font we indicated the best scores obtained and the red color the state-of-the-art scores. For the test set, the perplexity scores of both lay and expert corpora are reported in this order.

TST Model	STS Model	Quantile	// BLEU	self-BLEU	ref-BLEU	// BERT	self-BERT	ref-BERT	// BLEURT	self-BLEURT	ref-BLEURT	// COMET	self-COMET	ref-COMET	SS	pPPL	pPPL (lay)	pPPL (exp)
Test set	–	–	14.01	14.01	14.01	89.74	89.74	89.74	4.77	4.77	4.77	−4.62	−4.62	−4.62	94.67	6.14/5.05	2.70/3.19	3.27/2.53
OpenNMT+PT	–	–	–	59.89	9.92	–	97.16	89.13	–	60.75	**−10.96**	–	56.43	−33.62	21.04	5.49	**3.24**	**2.64**
UNTS	–	–	–	20.49	3.94	–	87.87	83.25	–	−47.99	−74.96	–	46.61	−96.89	37.48	31.78	18.46	14.89
ControlledGen	–	–	–	**88.61**	**13.13**	–	**98.29**	**89.20**	–	**63.10**	−13.35	–	**74.81**	**−28.38**	12.74	5.58	3.65	3.04
DeleteAndRetrieve	–	–	–	6.66	2.95	–	85.05	83.97	–	−78.48	−83.77	–	−91.43	−110.99	**79.56**	**5.46**	4.44	4.51
StyleTransformer	–	–	–	53.66	10.09	–	94.98	88.62	–	7.38	−35.33	–	19.02	−52.77	53.93	10.9	5.73	5.41

trained with pseudo-parallel data cost-effectively collected through Semantic Textual Similarity techniques, we achieved significantly better results than existing methods in terms of content preservation, style strength, and perplexity. Our human evaluations, in particular, show comparable performance to gold target texts, thus proving applicable for efficiently improving patient–physician communication, enhancing the overall health outcomes while reducing healthcare costs. Our results reinforce the value of our cost-effective methodology for building improved TST systems and provide solid evidence for the significant contributions our work has made to the advancement of this research field.

Furthermore, the comprehensive human evaluation phase involving experts and lay people, integrated with a qualitative analysis, shed light on the characteristics and issues of datasets, models, and evaluation metrics. Such results open the way to new challenges for future developments. For example, given the appropriate training corpora, our model could be trained at different expertise levels, bridging the gap for individuals based on their background. Also, our strategy could be implemented as a cost-effective preliminary step to minimize the workload of annotators involved in the collection of parallel datasets.

Moreover, the annotations we collected with the experts can be used in future studies to develop and evaluate Semantic Textual Similarity and Text Style Transfer systems in the medical field.

CRediT authorship contribution statement

Luca Bacco: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Felice Dell'Orletta:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Huiyuan Lai:** Methodology, Software, Writing – original draft. **Mario Merone:** Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Supervision. **Malvina Nissim:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to express their sincere gratitude to the physicians of the Department of Orthopaedic Surgery at the University Campus Bio-Medico of Rome, Italy, for their professional contribution to the development of this work. Special thanks are due to L. Ambrosio (MD), G. Papalia (MD), F. Russo (MD), and G. Vadalà (MD) for their invaluable assistance. The authors would also like to acknowledge the efforts and feedback of all the experts and lay annotators, too, who played a crucial role in the study. Their contribution is greatly appreciated.

Appendix. Comprehensive results

See Tables A.1 and A.2.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., et al. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop* (pp. 72–78). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-1909, URL https://aclanthology.org/W19-1909.

Apfel, F., & Tsouros, A. D. (2013). *Health literacy: the solid facts* (pp. 3–26). Copenhagen: World Health Organization.

Artetxe, M., & Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. http://dx.doi.org/10.1162/tacl_a_00288.

Bacco, L., Cimino, A., Paulon, L., Merone, M., & Dell'Orletta, F. (2020). A machine learning approach for sentiment analysis for Italian reviews in healthcare. *Computational Linguistics CLiC-It 2020*, *630*(699), 16.

Bacco, L., Russo, F., Ambrosio, L., D'Antoni, F., Vollero, L., Vadalà, G., et al. (2022). Natural language processing in low back pain and spine diseases: A systematic review. *Frontiers in Surgery*, *9*, http://dx.doi.org/10.3389/fsurg.2022.957085, URL https://www.frontiersin.org/articles/10.3389/fsurg.2022.957085.

Baker, D. W., Gazmararian, J. A., Williams, M. V., Scott, T., Parker, R. M., Green, D., et al. (2002). Functional health literacy and the risk of hospital admission among medicare managed care enrollees. *American Journal of Public Health*, *92*(8), 1278–1283.

Baker, D. W., Parker, R. M., Williams, M. V., & Clark, W. S. (1998). Health literacy and the risk of hospital admission. *Journal of General Internal Medicine*, *13*(12), 791–798.

Basu, C., Vasu, R., Yasunaga, M., Kim, S., & Yang, Q. (2021). Automatic medical text simplification: Challenges of data quality and curation.

Batterham, R., Hawkins, M., Collins, P. A., Buchbinder, R., & Osborne, R. H. (2016). Health literacy: applying current concepts to improve health services and reduce health inequalities. *Public Health*, *132*, 3–12.

Benigeri, M., & Pluye, P. (2003). Shortcomings of health information on the internet. *Health Promotion International*, *18*(4), 381–386.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267–D270.

Briakou, E., Agrawal, S., Tetreault, J., & Carpuat, M. (2021). Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1321–1336). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.100, URL https://aclanthology.org/2021.emnlp-main.100.

Briakou, E., Agrawal, S., Zhang, K., Tetreault, J., & Carpuat, M. (2021). A review of human evaluation for style transfer. In *Proceedings of the 1st workshop on natural language generation, evaluation, and metrics* (pp. 58–67). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.gem-1.6, URL https://aclanthology.org/2021.gem-1.6.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232–1254.

Cao, Y., Shui, R., Pan, L., Kan, M.-Y., Liu, Z., & Chua, T.-S. (2020). Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1061–1071). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.100, URL https://aclanthology.org/2020.acl-main.100.

Cífka, O., Şimşekli, U., & Richard, G. (2020). Groove2Groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2638–2650. http://dx.doi.org/10.1109/TASLP.2020.3019642.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. http://dx.doi.org/10.1177/001316446002000104, arXiv:10.1177/001316446002000104.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Dai, N., Liang, J., Qiu, X., & Huang, X. (2019). Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5997–6007). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1601, URL https://aclanthology.org/P19-1601.

De Mattei, L., Cafagna, M., Dell'Orletta, F., & Nissim, M. (2020). 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* (pp. 6709–6717). Marseille, France: European Language Resources Association, URL https://aclanthology.org/2020.lrec-1.828.

Devaraj, A., Marshall, I., Wallace, B., & Li, J. J. (2021). Paragraph-level simplification of medical texts. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4972–4984). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.395, URL https://aclanthology.org/2021.naacl-main.395.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

Elazar, Y., & Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 11–21). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1002, URL https://aclanthology.org/D18-1002.

Fu, Z., Tan, X., Peng, N., Zhao, D., & Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*. (1).

Gao, Y., Dligach, D., Christensen, L., Tesch, S., Laffin, R., Xu, D., et al. (2022). A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, *29*(10), 1797–1806. http://dx.doi.org/10.1093/jamia/ocac127.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894–6910). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.552, URL https://aclanthology.org/2021.emnlp-main.552.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Grabar, N., & Cardon, R. (2018). CLEAR–simple corpus for medical french. In *Proceedings of the 1st workshop on automatic text adaptation* (pp. 3–9).

Guo, Y., Qiu, W., Wang, Y., & Cohen, T. (2021). Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35* (1), (pp. 160–168).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2* (pp. 1735–1742). http://dx.doi.org/10.1109/CVPR.2006.100.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., et al. (2017). Efficient natural language response suggestion for smart reply. arXiv, arXiv:1705.00652.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Hoang, V. C. D., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 18–24). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W18-2703, URL https://aclanthology.org/W18-2703.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup, & Y. W. Teh (Eds.), *Proceedings of machine learning research*: *vol. 70*, *Proceedings of the 34th international conference on machine learning* (pp. 1587–1596). PMLR, URL https://proceedings.mlr.press/v70/hu17e.html.

Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., et al. (2017). Real-time neural style transfer for videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 783–791).

Imankulova, A., Sato, T., & Komachi, M. (2017). Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th workshop on asian translation (WAT2017)* (pp. 70–78). Taipei, Taiwan: Asian Federation of Natural Language Processing, URL https://aclanthology.org/W17-5704.

Imankulova, A., Sato, T., & Komachi, M. (2019). Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(2), http://dx.doi.org/10.1145/3341726.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., & Mihalcea, R. (2021). Deep learning for text style transfer: A survey. *Computational Linguistics*, 1–51.

Jin, Z., Jin, D., Mueller, J., Matthews, N., & Santus, E. (2019). IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3097–3109). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1306, URL https://aclanthology.org/D19-1306.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2020). Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, *26*, 3365–3385. http://dx.doi.org/10.1109/TVCG.2019.2921336.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, A freely accessible critical care database. *Scientific Data*, *3*(1), 1–9.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1181, URL https://aclanthology.org/D14-1181.

Kim, Y. H., Nam, S. H., Hong, S. B., & Park, K. R. (2022). GRA-GAN: Generative adversarial network for image style transfer of gender, race, and age. *Expert Systems with Applications*, *198*, Article 116792. http://dx.doi.org/10.1016/j.eswa.2022.116792, URL https://www.sciencedirect.com/science/article/pii/S0957417422002512.

King, A. (2010). Poor health literacy: A 'hidden'risk factor. *Nature Reviews Cardiology*, *7*(9), 473–474.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, system demonstrations* (pp. 67–72). Vancouver, Canada: Association for Computational Linguistics, URL https://aclanthology.org/P17-4012.

Lai, H., Mao, J., Toral, A., & Nissim, M. (2022). Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd workshop on human evaluation of NLP systems* (pp. 102–115). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.humeval-1.9, URL https://aclanthology.org/2022.humeval-1.9.

Lai, H., Toral, A., & Nissim, M. (2021a). Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4241–4254). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.349, URL https://aclanthology.org/2021.emnlp-main.349.

Lai, H., Toral, A., & Nissim, M. (2021b). Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 2: Short Papers)* (pp. 484–494). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-short.62, URL https://aclanthology.org/2021.acl-short.62.

Lai, H., Toral, A., & Nissim, M. (2022). Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 262–271). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-short.29, URL https://aclanthology.org/2022.acl-short.29.

Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M., & Boureau, Y.-L. (2019). Multiple-attribute text rewriting. In *International conference on learning representations*. URL https://openreview.net/forum?id=H1g2NhC5KQ.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.703, URL https://aclanthology.org/2020.acl-main.703.

Li, J., Jia, R., He, H., & Liang, P. (2018). Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long Papers)* (pp. 1865–1874). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1169, URL https://aclanthology.org/N18-1169.

Long, R., Yang, D., & Liu, Y. (2022). DiseaseNet: A novel disease diagnosis deep framework via fusing medical record summarization. *IAENG International Journal of Computer Science*, *49*(3).

Luo, F., Li, P., Zhou, J., Yang, P., Chang, B., Sui, Z., et al. (2019). A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 5116–5122). URL https://www.ijcai.org/proceedings/2019/0711.pdf.

Luo, J., Zheng, Z., Ye, H., Ye, M., Wang, Y., You, Q., et al. (2020). A benchmark dataset for understandable medical language translation. arXiv preprint arXiv:2012.02420.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., et al. (2020). Politeness transfer: A tag and generate approach. arXiv preprint arXiv:2004.14257.

Mäenpää, T., Suominen, T., Asikainen, P., Maass, M., & Rostila, I. (2009). The outcomes of regional healthcare information systems in health care: A review of the research literature. *International Journal of Medical Informatics*, *78*(11), 757–771. http://dx.doi.org/10.1016/j.ijmedinf.2009.07.001, URL https://www.sciencedirect.com/science/article/pii/S1386505609001051.

Manzini, E., Garrido-Aguirre, J., Fonollosa, J., & Perera-Lluna, A. (2022). Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. *Expert Systems with Applications*, *204*, Article 117446. http://dx.doi.org/10.1016/j.eswa.2022.117446, URL https://www.sciencedirect.com/science/article/pii/S0957417422007813.

Marie, B., & Fujita, A. (2017). Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 392–398).

McCreery, C. H., Katariya, N., Kannan, A., Chablani, M., & Amatriain, X. (2020). Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3458–3465).

Mukherjee, S., & Mulimani, M. (2022). ComposeInStyle: Music composition with and without style transfer. *Expert Systems with Applications*, *191*, Article 116195. http://dx.doi.org/10.1016/j.eswa.2021.116195, URL https://www.sciencedirect.com/science/article/pii/S0957417421015128.

Niu, T., & Bansal, M. (2018). Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, *6*, 373–389.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/1073083.1073135, URL https://aclanthology.org/P02-1040.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. W. (2018). Style transfer through back-translation. *CoRR*, arXiv:1804.09000, arXiv:1804.09000, URL http://arxiv.org/abs/1804.09000.

Rabinovich, E., Mirkin, S., Patel, R. N., Specia, L., & Wintner, S. (2016). Personalized machine translation: Preserving original author traits. arXiv preprint arXiv:1610.05461.

Rao, S., & 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.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 2685–2702). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.213, URL https://aclanthology.org/2020.emnlp-main.213.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2699–2712). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.240, URL https://aclanthology.org/2020.acl-main.240.

Sancheti, A., Krishna, K., Srinivasan, B. V., & Natarajan, A. (2020). Reinforced rewards framework for text style transfer. In *Advances in information retrieval* (pp. 545–560). URL https://arxiv.org/pdf/2005.05256.pdf.

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7881–7892). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.704, URL https://aclanthology.org/2020.acl-main.704.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709.

Shardlow, M., & Nawaz, R. (2019). Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 380–389). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1037, URL https://aclanthology.org/P19-1037.

Shen, T., Lei, T., Barzilay, R., & Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. *Advances in Neural Information Processing Systems*, *30*.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317.

Soldaini, L., & Goharian, N. (2016). Quickumls: A fast, unsupervised approach for medical concept extraction. In *MedIR workshop, Sigir* (pp. 1–4).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958, URL http://jmlr.org/papers/v15/srivastava14a.html.

Surya, S., Mishra, A., Laha, A., Jain, P., & Sankaranarayanan, K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2058–2068). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1198, URL https://aclanthology.org/P19-1198.

Tan, S. S.-L., & Goonawardene, N. (2017). Internet health information seeking and the patient-physician relationship: A systematic review. *Journal of Medical Internet Research*, *19*(1), Article e5729.

Tian, C. Y., Xu, R. H., Mo, P. K.-H., Dong, D., & Wong, E. L.-Y. (2020). Generic health literacy measurements for adults: A scoping review. *International Journal of Environmental Research and Public Health*, *17*(21), http://dx.doi.org/10.3390/ijerph17217768, URL https://www.mdpi.com/1660-4601/17/21/7768.

Tong, A., Levey, A. S., Eckardt, K.-U., Anumudu, S., Arce, C. M., Baumgart, A., et al. (2020). Patient and caregiver perspectives on terms used to describe kidney health. *Clinical Journal of the American Society of Nephrology*, *15*(7), 937–948. http://dx.doi.org/10.2215/CJN.00900120, URL https://cjasn.asnjournals.org/content/15/7/937.

Toshevska, M., & Gievska, S. (2022). A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, *3*(5), 669–684. http://dx.doi.org/10.1109/TAI.2021.3115992.

van den Bercken, L., Sips, R.-J., & Lofi, C. (2019). Evaluating neural text simplification in the medical domain. In *The world wide web conference* (pp. 3286–3292). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3308558.3313630.

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, *81*(2), 399–410.

Vásquez-Rodríguez, L., Shardlow, M., Przybyła, P., & Ananiadou, S. (2021). Investigating text simplification evaluation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 876–882).

Vydiswaran, V. V., Mei, Q., Hanauer, D. A., & Zheng, K. (2014). Mining consumer health vocabulary from community-generated text. In *AMIA annual symposium proceedings, Vol. 2014* (p. 1150). American Medical Informatics Association.

Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., et al. (2020). MedSTS: A resource for clinical semantic textual similarity. *Language Resources and Evaluation*, *54*(1), 57–72.

Wang, Y., Afzal, N., Liu, S., Rastegar-Mojarad, M., Wang, L., Shen, F., et al. (2018). Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. *Proceedings of the BioCreative/OHNLP Challenge*, *2018*.

Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., Liu, H., et al. (2020). The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. *JMIR Medical Informatics*, *8*(11), Article e23375.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Weng, W.-H., Chung, Y.-A., & Szolovits, P. (2019). Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3121–3131).

White, R. W., & Horvitz, E. (2009). Experiences with web search on medical concerns and self diagnosis. In *AMIA annual symposium proceedings, Vol. 2009* (p. 696). American Medical Informatics Association.

Xu, W., Saxon, M., Sra, M., & Wang, W. Y. (2021). Self-supervised knowledge assimilation for expert-layman text style transfer. arXiv preprint arXiv:2110.02950.

Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., & Rosendale, D. (2007). Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA annual symposium proceedings, Vol. 2007* (p. 846). American Medical Informatics Association.

Zhang, N., & Jankowski, M. (2022). Hierarchical BERT for medical document understanding. arXiv preprint arXiv:2204.09600.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhou, C., Chen, L., Liu, J., Xiao, X., Su, J., Guo, S., et al. (2020). Exploring contextual word-level style relevance for unsupervised style transfer. arXiv preprint arXiv:2005.02049.

Zhu, S., Yang, Y., & Xu, C. (2020). Extracting parallel sentences from nonparallel corpora using parallel hierarchical attention network. *Computational Intelligence and Neuroscience*, *2020*.

Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, *36*(4), 326–333. http://dx.doi.org/10.1016/j.jbi.2003.09.015, URL https://www.sciencedirect.com/science/article/pii/S1532046403000960, Building Nursing Knowledge through Informatics: From Concept Representation to Data Mining.