Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen

Yixin Cao$^1$, Ruihao Shui$^{21}$, Liangming Pan$^{21}$, Min-Yen Kan$^1$, Zhiyuan Liu$^3$, Tat-Seng Chua$^1$

$^1$School of Computing, National University of Singapore, Singapore
$^2$NUS Graduate School for Integrative Sciences and Engineering
$^3$Department of CST, Tsinghua University, Beijing, China

ciaoyixin2011@gmail.com, {ruihaoshui,e0272310}@u.nus.edu
liuzy@tsinghua.edu.cn, {kanmy@comp.,dcscts}@nus.edu.sg

Abstract

The curse of knowledge can impede communication between experts and laymen. We propose a new task of expertise style transfer and contribute a manually annotated dataset with the goal of alleviating such cognitive biases. Solving this task not only simplifies the professional language, but also improves the accuracy and expertise level of laymen descriptions using simple words. This is a challenging task, unaddressed in previous work, as it requires the models to have expert intelligence in order to modify text with a deep understanding of domain knowledge and structures. We establish the benchmark performance of five state-of-the-art models for style transfer and text simplification. The results demonstrate a significant gap between machine and human performance. We also discuss the challenges of automatic evaluation, to provide insights into future research directions. The dataset is publicly available at https://srthu.github.io/expertise-style-transfer/.

1 Introduction

The curse of knowledge (Camerer et al., 1989) is a pervasive cognitive bias exhibited across all domains, leading to discrepancies between an expert’s advice and a layman’s understanding of it (Tan and Goonawardene, 2017). Take medical consultations as an example: patients often find it difficult to understand their doctors’ language. On the other hand, it is important for doctors to accurately disclose the exact illness conditions based on patients’ simple vocabulary. Misunderstanding may lead to failures in diagnosis and prompt treatment, or even death. How to automatically adjust the expertise level of texts is critical for effective communication.

In this paper, we propose a new task of text style transfer between expert language and layman language, namely Expertise Style Transfer, and contribute a manually annotated dataset in the medical domain for this task. We show four examples in Figure 1, where the upper sentence is for professionals and the lower one is for laymen. On one hand, expertise style transfer aims at improving the readability of a text by reducing the expertise level, such as explaining the complex terminology dyspnea in the first example with a simple phrase shortness of breath. On the other hand, it also aims to improve the expertise level based on context, so that layman’s expressions can be more accurate and professional. For example, in the second pair, causing further damage is not as accurate as ulcers, omitting the important mucous and disintegrative conditions of the sores.

There are two related tasks, but neither serve as suitable prior art. The first is text style transfer (ST), which generates texts with different attributes but with the same content. However, although existing approaches have achieved a great success regarding the attributes of sentiment (Li et al., 2018) and formality (Rao and Tetreault, 2018) among oth-
ers, expertise “styling” has not been explored yet. Another similar task is Text Simplification (TS), which rewrites a complex sentence with simple structures (Sulem et al., 2018b) while constrained by limited vocabulary (Paetzold and Specia, 2016). This task can be regarded as similar to our sub-task: reducing the expertise level from expert to layman language without considering the opposing direction. However, most existing TS datasets are derived from Wikipedia, and contain numerous noise (misaligned instances) and inadequacies (instances having non-simplified targets) (Xu et al., 2015; Surya et al., 2019); in which further detailed discussion can be found in Section 3.2.

In this paper, we construct a manually-annotated dataset for expertise style transfer in medical domain, named MSD, and conduct deep analysis by implementing state-of-the-art (SOTA) TS and ST models. The dataset is derived from human-written medical references, The Merck Manuals\(^1\), which include two parallel versions of texts, one tailored for consumers and the other for healthcare professionals. For automatic evaluation, we hire doctors to annotate the parallel sentences between the two versions (examples shown in Figure 1). Compared with both ST and TS datasets, MSD is more challenging from two aspects:

Knowledge Gap. Domain knowledge is the key factor that influences the expertise level of text, which is also a key difference from conventional styles. We identify two major types of knowledge gaps in MSD: terminology, e.g., *dyspnea* in the first example; and empirical evidence. As shown in the third pair, doctors prefer to use statistics (*About 1/1000*), while laymen do not (*quite small*).

Lexical & Structural Modification. Fu et al. (2019) has indicated that most ST models only perform lexical modification, while leaving structures unchanged. Actually, syntactic structures play a significant role in language styles, especially regarding complexity or simplicity (Carroll et al., 1999). As shown in the last example, a complex sentence can be expressed with several simple sentences by appropriately splitting content. However, available datasets rarely contain such cases.

Our main contributions can be summarized as:

- We propose a new task of expertise style transfer, which aims to facilitate communication between experts and laymen.
- We contribute a challenging dataset that requires knowledge-aware and structural modification techniques.
- We establish benchmark performance and discuss key challenges of datasets, models and evaluation metrics.

2 Related Work

2.1 Text Style Transfer

Existing ST work has achieved promising results on the styles of sentiment (Hu et al., 2017; Shen et al., 2017), formality (Rao and Tetreault, 2018), offensiveness (dos Santos et al., 2018), politeness (Sennrich et al., 2016), authorship (Xu et al., 2012), gender and ages (Prabhumoye et al., 2018; Lample et al., 2019), etc. Nevertheless, only a few of them focus on supervised methods due to the limited availability of parallel corpora. Jhatani et al. (2017) extract modern language based Shakespeare’s play from the educational site, while Rao and Tetreault (2018) and Li et al. (2018) utilize crowdsourcing techniques to rewrite sentences from Yahoo Answers, Yelp and Amazon reviews, which are then utilized for training neural machine translation (NMT) models and evaluation.

More practically, there is an enthusiasm for unsupervised methods without parallel data. There are three groups. The first group is Disentanglement methods that learn disentangled representations of style and content, and then directly manipulating these latent representations to control style-specific text generation. Shen et al. (2017) propose a cross-aligned autoencoder that learns a shared latent content space between true samples and generated samples through an adversarial classifier. Hu et al. (2017) utilize neural generative model, Variational Autoencoders (VAEs) (Kingma and Welling, 2013), to represent the content as continuous variables with standard Gaussian prior, and reconstruct style vector from the generated samples via an attribute discriminator. To improve the ability of style-specific generation, Fu et al. (2018) utilize multiple generators, which are then extended by a Wasserstein distance regularizer (Zhao et al., 2018). SHAPED (Zhang et al., 2018a) learns a shared and several private encoder–decoder frameworks to capture both common and distinguishing features. Some variants further investigate the auxiliary tasks to better preserve contents (John et al., 2019), or domain adaptation (Li et al., 2019).

\(^1\)https://en.wikipedia.org/wiki/The_Merck_Manuals
Another line of work argues that it is difficult to disentangle style from content. Thus, their main idea is to learn style-specific translations, which are trained using unaligned data based on back-translation (Zhang et al., 2019; Prabhumoye et al., 2018; Lample et al., 2019), pseudo parallel sentences according to semantic similarity (Jin et al., 2019), or cyclic reconstruction (Dai et al., 2019), marked with Translation methods.

The third group is Manipulation methods. Li et al. (2018) first identify the style words by their statistics, then replace them with similar retrieved sentences with a target style. Xu et al. (2018) jointly train the two steps with a neutralization module and a stylization module based on reinforcement learning. For better stylization, Zhang et al. (2018b) introduce a learned sentiment memory network, while John et al. (2019) utilize hierarchical reinforcement learning.

2.2 Text Simplification

Earlier work on text simplification define a sentence as simple, if it has more frequent words, shorter length and fewer syllables per word, etc. This motivates a variety of syntactic rule-based methods, such as reducing sentence length (Chandrasekar and Srinivas, 1997; Vickrey and Koller, 2008), lexical substitution (Glavas and Stajner, 2015; Paetzold and Specia, 2016) or sentence splitting (Woodsend and Lapata, 2011; Sulem et al., 2018b). Another line of work follows the success of machine translation (MT) (Klein et al., 2017), and regards TS as a monolingual translation from complex language to simple language (Zhu et al., 2010; Coster and Kauchak, 2011; Wubben et al., 2012). Zhang and Lapata (2017) incorporate reinforcement learning into the encoder–decoder framework to encourage three types of simplification rewards concerning language simplicity, relevance and fluency, while Shardlow and Nawaz (2019) improve the performance of MT models by introducing explanatory synonyms. To alleviate the heavy burden of parallel training corpora, Surya et al. (2019) propose an unsupervised model via adversarial learning between a shared encoder and separate decoders.

The simplicity of language in the medical domain is particularly important. Terminologies are one of the main obstacles to understanding, and extracting their explanations could be helpful for TS (Shardlow and Nawaz, 2019). Deléger and Zweigenbaum (2008) detect paraphrases from comparable medical corpora of specialized and lay texts, and Kloehn et al. (2018) explore UMLS (Bodenreider, 2004) and WordNet (Miller, 2009) with word embedding techniques. Furthermore, Van den Bercken et al. (2019) directly align sentences from medical terminological articles in Wikipedia and Simple Wikipedia2, which confines the editors’ vocabulary to only 850 basic English words. Then, they refine these aligned sentences by experts towards automatic evaluation. However, the Wikipedia-based dataset is still noisy (with misaligned instances) and inadequate (instances having non-simplified targets) with respect to both model training and testing. Besides, it is usually ignored that the opposite direction of TS — improving the expertise levels of layman language for accuracy and professionalism — is also critical for better communication.

2.3 Discussion

To sum up, both tasks lack parallel data for training and evaluation. This prevents researchers from exploring more advanced models concerning the knowledge gap as well as linguistic modification of lexicons and structures. In this work, we define a more useful and challenging task of expertise style transfer with high-quality parallel sentences for evaluation. Besides, the two communities of ST and TS can shed lights to each other on sentence modification techniques.

3 Dataset Design

We describe our dataset construction that comprises three steps: data preprocessing, expert annotation and knowledge incorporation. We then give a detailed analysis.

3.1 Dataset Construction

The Merck Manuals, also known as the MSD Manuals, have been the world’s most trusted health reference for over 100 years. It covers a wide range of medical topics, and is written through a collaboration between hundreds of medical experts, supervised by independent editors. For each topic, it includes two versions: one tailored for consumers and the other for professionals.

**Step 1: Data Preprocessing.** Although the two versions of documents refer to the same topic, they
Pleural Effusion, Symptoms

**Expert**
Many cause dyspnea [C0013404], pleuritic chest pain [C0008033], or both.

**Laymen**
The most common symptoms, regardless of the type of fluid in the pleural space or its cause, are shortness of breath [C2707305;C3274920] and chest pain [C0008031;C2926613].

Table 1: Examples of parallel annotation in MSD, where the red fonts in brackets denote UMLS concepts.

---

**Step 2: Expert Annotation.** Given the aligned groups of sentences in professional and consumer MSD, we develop an annotation platform to facilitate expert annotations. We hire three doctors to select sentences from each version of group to annotate pairs of sentences that have the same meaning but are written in different styles. The hired doctors are formally medically trained, and are qualified to understand the semantics of the medical texts. To avoid subjective judgments in the annotations, they are not allowed to change the content. Particularly, the doctors are Chinese who also know English as a second language. Thus, we provide the English content accompanied with a Chinese translation as assistance, which helps to increase the annotation speed while ensuring quality. We also conduct verification on each pair of parallel sentences with the help of another doctor. Note that each pairing may contain multiple professional and consumer sentences; i.e., multiple alignment is possible, the alignments are not necessarily one-to-one. The strict procedure also discards many aligned groups, leading to 675 annotations for testing, with distribution of medical PCIO elements as shown in Figure 3.

**Step 3: Knowledge Incorporation.** To facilitate knowledge-aware analysis, we can utilize information extraction techniques (Cao et al., 2018a, 2019) to identify medical concepts in each sentence. Here, we use QuickUMLS (Soldaini and Goharian, 2016) to automatically link entity mentions to Unified Medical Language System (UMLS) (Bodenreider, 2004). Note that each mention may refer to multiple concepts, each for which we align to the highest ranked one. As shown in Table 1, the
mention dyspnea is linked to concept C0013404.

Through this three step process, we obtain a large set of (non-parallel) training sentences in each style, and a small set of parallel sentences for evaluation. The detailed statistics as compared with other datasets can be found in Table 2 and Table 3.

### 3.2 Dataset Analysis

Let us compare our MSD dataset against both publicly available ST and TS datasets. SimpWiki (Van den Bercken et al., 2019) is a TS dataset derived from the linked articles between Simple Wikipedia and Normal Wikipedia. It focuses on the medical domain and extracts parallel sentences automatically by computing their BLEU scores. GYAFC (Rao and Tetreault, 2018) is the largest ST dataset on formality in the domains of Entertainment & Music (E&M) and Family & Relationships (F&R) from Yahoo Answers. It contains more than 50,000 training sentences (non-parallel) for each domain, and over 1,000 parallel sentences for testing, obtained by rewriting informal answers via Amazon Mechanical Turk. Yelp and Amazon (Li et al., 2018) are sentiment ST datasets by rewriting reviews based on crowdsourcing. They both contain over 270k training sentences (non-parallel) and 500 parallel sentences for evaluation. Authorship (Xu et al., 2012) aims at transferring styles between modern English and Shakespearean English. It contains 18,395 sentences for training (non-parallel) and 1,462 sentence pairs for testing.

#### Dataset Statistics

Table 2 presents the statistics of expertise and layman sentences in our dataset as well as SimpWiki. We split the sentences using NLTK, and compute the ratio of layman to expert in each metric to denote the gap between the two styles (a lower value implies a smaller gap but that for #Sentence). Three standard readability indices are used to evaluate the simplicity levels: FleshKincaid (Kincaid et al., 1975), Gunning (Gunning, 1968) and Coleman (Coleman and Liau, 1975). The lower the indices are, the simpler the sentence is. Note that SimpWiki does not provide a train/test split, and thus we randomly sample 350 sentence pairs for evaluation. We follow the same strategy in our experiments.

Compared with SimpWiki, we can see that: (1) MSD evaluates the structure modifications. As the layman language usually requires more simple sentences to express the same meaning as in the expert language, each expert sentence in MSD Test refers to 1.13 layman sentences on average, while the number in SimpWiki is only 0.99. (2) MSD is more distinct between the two styles, which is critical for style transfer. This is markedly demonstrated by the larger difference between their (concepts) vocabulary sizes (0.62/0.81 vs. 0.85 in ratio of layman to expert), and between the readability indices (0.81/0.81 vs. 0.84 on average). (3) we have more complex professional sentences in expert language (14.57/14.07 vs. 13.55 in the three readability indices on average) but comparatively simple sentences in laymen language (11.89/11.45 vs. 11.40). This is intuitive because both versions of Wikipedia are written by crowdsourcing editors, and MSD is written by experts in medical domain.

#### Quality of Parallel Sentences

One of the main concerns in ST is the limitations of parallel sentences towards automatic evaluation. On one hand, assuming that the parallel sentences have the same meaning, many datasets find the aligned sentences to have higher string overlap (as measured by BLEU). On the other hand, the two sentences should have different styles, and may vary a lot in expressions: and thus leading to a lower BLEU. Hence how to build a testing dataset that considers both criteria is critical. We analyze
the quality of testing sentence pairs in each dataset.

| Datasets   | BLEU  | ED   | Task |
|------------|-------|------|------|
| GYAFTeam   | 16.22 | 28.53| ST   |
| GYAFVar    | 16.95 | 29.35| ST   |
| Yelp       | 24.76 | 22.20| ST   |
| Amazon     | 44.52 | 19.75| ST   |
| Authorship | 19.43 | 36.70| ST   |
| SimpWiki   | 49.98 | 64.16| TS   |
| MSD        | 14.01 | 139.73| ST & TS |

Table 3: BLEU (4-gram) and edit distance (ED) scores between parallel sentences. Concept words are masked for ED computation (Fu et al., 2019). Higher BLEUs imply two more similar sentences, while higher edit distances imply more heterogeneous structures.

Table 3 presents the BLEU and edit distance (ED for short) scores. Note that each pair of parallel sentences is verified to convey the same meaning during annotation. We see that: (1) MSD has the lowest BLEU and highest ED. This implies that MSD is very challenging that requires both lexical and structural modifications. (2) TS datasets reflect more structural differences (with higher ED values) as compared to ST datasets. This means that TS datasets concerning the nature of language complexity (simplicity) are more complex to transfer.

4 Experiments

We reimplement five SOTA models from prior TS and ST studies on both MSD and SimpWiki datasets. A further ablation study gives a detailed analysis of the knowledge and structure impacts, and highlights the challenges of existing metrics.

4.1 Baselines

We choose the following methods to establish benchmark performance on the two datasets on expertise style transfer, because they: (1) achieve SOTA performance in their fields; (2) are typical methods (as grouped in Section 2); and (3) release codes for reimplementation.

The TS models selected are: (1) Supervised model OpenNMT+PT that incorporates a phrase table into OpenNMT (Klein et al., 2017), which provides guidance for replacing complex words with their simple synonym (Shardlow and Nawaz, 2019); and (2) Unsupervised model UNTS that utilizes adversarial learning (Surya et al., 2019).

The models for ST task selected are: (1) Disentanglement method ControlledGen (Hu et al., 2017) that utilizes VAEs to learn content representations following a Gaussian prior, and reconstructs a style vector via a discriminator; (2) Manipulation method DeleteAndRetrieve (Li et al., 2018) that first identifies style words with a statistical method, then replaces them with target style words derived from given corpus; and (3) Translation method StyleTransformer (Dai et al., 2019) that uses cyclic reconstruction to learn content and style vectors without parallel data.

4.2 Training Details

We use the pre-trained OpenNMT+PT model released by the authors. Other models are trained using MSD and SimpWiki training data. We leave 20% of the training data for validation. The training settings follow the standard best practice; where all models are trained using Adam (Kingma and Ba, 2015) with mini-batch size 32, and the hyper-parameters are tuned on the validation set. We set the shared parameters the same for baseline models: the maximum sequence length is 100, the word embeddings are initialized with 300-dimensional GloVe (Pennington et al., 2014), learning rate is set to 0.001, and adaptive learning rate decay is applied. We adopt early stopping and dropout rate is set to 0.5 for both encoder and decoder.

4.3 Evaluation Metrics

Following Dai et al. (2019), we make an automatic evaluation on three aspects:

Style Accuracy (marked as Acc) aims to measure how accurate the model controls sentence style. We train two classifiers on the training set of each dataset using fasttext (Joulin et al., 2017).

Fluency (marked as PPL) is usually measured by the perplexity of the transferred sentence. We fine-tune the state-of-the-art pretrained language model, Bert (Devlin et al., 2019), on the training set of each dataset for each style.

Content Similarity measures how much content is preserved during style transfer. We calculate 4-gram BLEU (Papineni et al., 2002) between model outputs and inputs (marked as self-BLEU), and between outputs and gold human references (marked as ref-BLEU).

Automatic metrics for content similarity are arguably unreliable, since the original inputs usually achieve the highest scores (Fu et al., 2019). We only report TS models for expertise to laymen language, since they do not claim the opposite direction.

---

5 We only report TS models for expertise to laymen language, since they do not claim the opposite direction.

6 https://github.com/senisioi/NeuralTextSimplification/
Thus also conduct human evaluation. To evaluate over the entire test set, only layman annotators are involved, but we ensure that the layman style sentences are accompanied as references to assist understanding. Each annotator is asked to rate the model output given both input and gold references. The rating ranges from 1 to 5, where higher values indicate that more semantic content is preserved.

**Text Simplification Measurement.** The above metrics may not perform well regarding language simplicity (Sulem et al., 2018a). So, we also utilize a TS evaluation metrics: SARI (Xu et al., 2016). It compares the n-grams of the outputs against those of the input and human references, and considers the added, deleted and kept words by the system.

### 4.4 Overall Performance

Table 4 presents the overall performance. Since each pair of parallel sentences has been verified during annotation, we did not report human scores to avoid repeated evaluations. We can see that: (1) Parallel sentences in MSD have higher quality than SimpWiki, because our gold references are more fluent (4.29 vs. 7.65 in perplexity on average) and more discriminable (91% vs. 60% on average style accuracy). (2) The transfer for L2E is more difficult (except in content similarity) than that for E2L: 39.55% vs. 42.50% in Acc on average, 11.50 vs. 10.33 in PPL on average and 2.80 vs. 2.63 in human ratings on average. This is because the increase in expertise levels requires more contexts and knowledge, and is harder than simplification. (3) TS models perform similarly with ST models. Besides, supervised model OpenNMT+PT outperforms the unsupervised UNTS in fluency and content similarity due to the additional supervision signals. On the other hand, UNTS achieves higher Acc since it utilizes more non-parallel training data. (4) The style accuracy is the reverse to content similarity, making it more challenging to propose a comprehensive evaluation metric that can balance the two opposite directions. In terms of content similarity, even if both self-BLEU and ref-BLEU show a strong correlation with human ratings (over 0.98 Pearson coefficient with p-value < 0.0001), the higher scores of ControlledGen cannot demonstrate its superior performance, as it actually makes little modifications to styles. Instead, DeleteAndRetrieve, presents a strong ability to control styles (70% on average in Acc on MSD), but hardly preserves the contents. Style Transformer performs more stably.

Next, we discuss key factors of MSD. We take the E2L as the exemplar for discussion, as we have observed similar results for the opposing direction.

### 4.5 Impact of Concepts

Figure 4a shows the performance curves of BLEU and style accuracy. We choose the concept range to ensure they contain similar number of sentences. Along with the increasing number of concepts, we can see a downward BLEU trend. This is because it becomes more difficult to preserve content when the sentence is more professional. As for style accuracy, DeleteAndRetrieve achieves the peak around [8,12) concepts, while the performance of other models drops gradually. Clearly, a lower number of concepts benefit the model for better understanding the sentences due to their correlated semantics, but a larger number of concepts requires knowledge-aware text understanding.

### 4.6 Impact of Structures

Figure 4b presents the performance curves regarding the structure differences, where the edit distance is computed as mentioned in Section 3.2. Higher score denotes more heterogeneous structures. We see a similar trend with the curves of concepts. That is, existing models perform well.
in simple cases (fewer concepts and less structural differences), but becomes worse if the language is complex. We doubt that the encoder in each model is able to understand the domain-specific language sufficient well without considering knowledge. We thus propose a simple variant of ControlledGen by introducing terminology definitions, and observe some interesting findings in Section 4.10.

4.7 Performance on Medical PCIO
The style of medical PCIO elements (e.g., symptoms) are slightly different. We separately evaluate each model and present the results in Figure 4c. Style accuracy remains similar among these medical PCIO elements, but there are significant differences among the models in their performance for preserving content. Specifically, models perform well for those sentences about treatment, but perform poorly for evaluation, because this type of sentences usually involve many rare terms, challenging understanding.

4.8 Performance using Simplification Metrics

| Dataset          | SimpWiki | MSD |
|------------------|----------|-----|
|                  | E2L      | L2E | E2L      | L2E  |
| OpenNMT+PT       | .2695    | -   | .2204    | -    |
| UNTS             | .3115    | -   | .3313    | -    |
| ControlledGen    | .3187    | .2856 | .2170    | .1636 |
| DeleteAndRetrieve| .1983    | .1684 | .3378    | .3345 |
| StyleTransformer | .3189    | .2933 | .3541    | .3411 |

Table 5: Performance using SARI.

Table 5 presents the performance based on the TS evaluation metric, SARI. We utilize the Python package\(^7\) and follow the settings in the original paper. Surprisingly, SARI on MSD presents a relatively comprehensive evaluation that is consistent with the above analysis as well as our intuition. ControlledGen and OpenNMT+PT are ranked lower since they tend to simply repeat the input. DeleteAndRetrieve and UNTS are ranked in the middle due to the accurate style transfer but poor content preservation. StyleTransformer is ranked highest as it performs stably in Table 4 and Figure 4a, 4b, 4c. This inspires us to further investigate automatic evaluation metrics based on TS studies, which is our ongoing work. Even so, we still recommend necessary human evaluation in the current stage.

4.9 Case Study

Table 6 presents two examples of transferred sentences. In the first example, both OpenNMT+PT and UNTS make lexical changes: replacing progresses with goes. DeleteAndRetrieve transfers style successfully but also changes the content slightly. The other two output the original expert sentence, that is the reason why they achieve higher BLEU (also PPL) but fails in Acc. Manipulation method (i.e., DeleteAndRetrieve) is more progressive in changing the style, but disentanglement method, ControlledGen, prefers to stay the same.

The second example shows structural modifications. We can see that the supervised OpenNMT+PT simply deletes the complex terminol...
Prostate cancer usually progresses slowly and rarely causes symptoms until advanced.

Prostate cancer usually goes slowly and rarely causes symptoms until advanced.

Prostate cancer usually goes slowly and rarely causes symptoms until advanced.

Prostate cancer usually progresses slowly and rarely causes symptoms until advanced.

Prostate cancer usually progresses slowly and rarely causes symptoms until advanced.

Cystic lung disease and recurrent spontaneous pneumothorax may occur. These disorders can cause pain and shortness of breath.

Cystic lung disease can cause pain and shortness of breath. Lung disease and roughly something pneumothorax may occur.

Cystic lung disease and recurrent spontaneous pneumothorax may occur. These disorders can cause pain and shortness of breath.

ear skin disease in the lungs and the lungs may occur in other disorders and may cause chest pain and shortness of breath.

Cystic lung disease and exposed spontaneous pneumothorax may occur.

Air-filled sacs (cysts) may develop in the lungs. The cysts may rupture, bringing air into the space that surrounds the lungs (pneumothorax). These disorders can cause pain and shortness of breath.

Table 6: Examples of model outputs. Red/blue words with underlines highlight model/expected modifications.

| Expertise input  | OpenNMT+PT       | UNTS       | ControlledGen | DeleteAndRetrieve | StyleTransformer | Laymen Gold |
|------------------|------------------|------------|---------------|-------------------|------------------|-------------|
| Prostate cancer  | Prostate cancer  | Prostate   | Prostate      | Prostate          | Prostate         | Prostate    |
| goes slowly      | goes slowly      | goes slowly| goes slowly   | goes slowly       | goes slowly      | goes slowly |
| rarely causes    | rarely causes    | rarely     | rarely causes | rarely causes     | rarely causes    | rarely causes|
| symptoms        | symptoms         | symptoms   | symptoms      | symptoms          | symptoms         | symptoms    |
| until advanced.  | until advanced.  | until      | until advanced.| until advanced.   | until advanced.  | until advanced.|

4.10 Discussion

We have two observations from the aspects of model and evaluation. For models, there is a huge gap between all of the above models and human references. MSD is indeed challenging to conduct language modifications considering both knowledge and structures. Most of the time, these models basically output the original sentences without any modifications, or simply cut off the complex words randomly. Besides, all of the above models still perform much worse than human, which motivates research into better models.

5 Conclusion

We proposed a practical task of expertise style transfer and constructed a high-quality dataset, MSD. It is of high quality and also challenging due to the presence of knowledge gap and the need of structural modifications. We established benchmark performance of five SOTA models. The results shown a significant gap between machine and human performance. Our further discussion analyzed the challenges of existing metrics.

In the future, we are interested in injecting knowledge into text representation learning (Cao et al., 2017, 2018b) for deeply understanding expert language, and will help to generate knowledge-enhanced questions (Pan et al., 2019) for laymen.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.
References

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In WWW.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research.

Colin Camerer, George Loewenstein, and Martin Weber. 1989. The curse of knowledge in economic settings: An experimental analysis. Journal of political Economy.

Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018a. Neural collective entity linking. In COLING.

Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Chengjiang Li, Xu Chen, and Tiansi Dong. 2018b. Joint representation learning of cross-lingual words and entities via attentive distant supervision. In EMNLP.

Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In EMNLP.

Zheting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In ICML.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Re-thinking text attribute transfer: A lexical analysis. ArXiv.

Zhixin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In AAAI.

Goran Glavas and Sanja Stajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In ACL.

Robert Gunning. 1968. Technique of clear writing.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In AAAI.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In ICLR.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. ICLR.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In ACL.

Nicholas Kloechn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P Yuan, and Debra Revere. 2018. Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in english and spanish. JMIR.

Guillaume Lamplde, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In ICLR.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and William B. Dolan. 2019. Domain adaptive text style transfer. EMNLP.
