Self-Supervised Knowledge Assimilation for Expert-Layman Text Style Transfer

Wenda Xu, Michael Saxon, Misha Sra, William Yang Wang

University of California, Santa Barbara
Department of Computer Science
Santa Barbara, California, USA
{wendaxu,saxon}@ucsb.edu {sra,william}@cs.ucsb.edu

Abstract

Expert-layman text style transfer technologies have the potential to improve communication between members of scientific communities and the general public. High-quality information produced by experts is often filled with difficult jargon laypeople struggle to understand. This is a particularly notable issue in the medical domain, where layman are often confused by medical text online. At present, two bottlenecks interfere with the goal of building high-quality medical expert-layman style transfer systems: a dearth of pretrained medical-domain language models spanning both expert and layman terminologies and a lack of parallel corpora for training the transfer task itself. To mitigate the first issue, we propose a novel language model (LM) pretraining task, *Knowledge Base Assimilation*, to synthesize pretraining data from the edges of a graph of expert- and layman-style medical terminology terms into an LM during self-supervised learning. To mitigate the second issue, we build a large-scale parallel corpus in the medical expert-layman domain using a margin-based criterion. Our experiments show that transformer-based models pretrained on knowledge base assimilation and other well-established pretraining tasks fine-tuning on our new parallel corpus leads to considerable improvement against expert-layman transfer benchmarks, gaining an average relative improvement of our human evaluation, the Overall Success Rate (OSR), by 106%.

Introduction

Incompatible knowledge backgrounds between experts and laymen cause communication difficulties (Jerit 2009). These difficulties are particularly problematic in the medical domain when patients attempt to self-diagnose their ailments online (White and Horvitz 2010). Their search terms might be too vague, leading them to self-misdiagnose, followed by unnecessary treatment or tests, and potentially worse outcomes (Au 2019). Even if they find high-quality, correct online medical resources that match their condition, the incomprehensible medical jargon within can be confusing and frustrating (Benigeri and Pluye 2003). Misunderstandings from online medical information seeking have been shown to lead to increased health anxiety (White and Horvitz 2009). Expert-layman text style transfer technologies offer a potential method to resolve these problems. Accurate layman-to-expert style conversion of vague searches into precise terminology could improve the quality of retrieved documents. In turn, high-quality expert-to-layman translation of the retrieved documents would lead to more comfort, better understanding, and hopefully better overall outcomes.

Text style transfer is the task of transforming a passage of text from a source style (e.g., expert medical language) to a target style (e.g., layman language) while preserving the underlying meaning (Jin et al. 2021). Prior work has demonstrated impressive style transfer results across a variety of attributes, including sentiment (Li et al. 2018; Dai et al. 2019), formality (Rao and Tetreault 2018), and politeness (Sennrich, Haddow, and Birch 2016). However, the aforementioned style-transfer tasks are fundamentally surface-level transformations along fairly content-agnostic dimensions. Expert-layman style transfer is different in that it requires *domain-specific terminological correspondence knowledge*. For a given domain, the model must contain sets of mappings between specific expert and layman-style expressions for phenomena (e.g., "renal" and "relating to the kidneys"), and a system intended for one domain has no guaranteed applicability to another (e.g., linguistics to medicine).

In this paper, we tackle two core hurdles to building high-quality medical expert-layman style transfer systems: a lack of pretrained sequence-to-sequence language models containing medical domain-specific terminological correspondence knowledge, and a lack of parallel expert-layman medical corpora for fine-tuning. For the first hurdle, we introduce a novel language model (LM) pretraining task, **knowledge base assimilation** (KBA) to explicitly provide the LM with a learning signal across expert-layman phrasal realizations of concepts. We further augment our KBA training with the previously proposed *Mask*, *Switch*, and *Delete* self-supervised pretraining tasks (Devlin et al. 2019; Lample et al. 2018; Lewis et al. 2019) to build a robust medical LM containing terminological correspondence knowledge. To the best of our knowledge, this is the first work to investigate self-supervised representation learning in expert-layman text style transfer. To tackle the second hurdle, we produce a high-quality parallel extension of the non-parallel MSD medical text dataset (Cao et al. 2020) containing 11,512 expert- and layman-style medical sentences using a margin-based data mining criterion (Schwenk 2018). To the best of our knowledge, this is the first work to utilize supervised learning on a large data mined parallel corpus of expert-laymen sentences. We release our

code and parallel corpus for future research. [1]

Proposed Approach

A *knowledge domain* D contains many *concepts* C_i, which are sets of sentences $S_{i,j}$ of equivalent meaning. Sentences can be labeled with attributes, including whether they belong to the "expert" or "layman" style. We define the task of expert-layman style transfer as follows: given a sentence $S \in C_i$ with either the expert- or layman-style, generate a sentence S' in the other style that is also a member of C_i. This is not a simple task, as it requires the model to be aware of the underlying medical concept that links semantically equivalent but lexically unique medical phrases together.

To tackle this problem, we first pretrain a transformer language model on an ensemble of tasks, including our novel **knowledge base assimilation** (KBA) task and the previously demonstrated *Mask, Delete* and *Switch* **self-supervised learning** (SSL) tasks, to simultaneously model medical language in general while also capturing how specific concepts are phrased in each style. Then, we fine-tune this language model on **a new corpus of parallel expert-layman medical sentences** we extract using a margin-based criterion from the unaligned MSD dataset.

Dataset

We evaluate our proposed method and current SOTA models using the MSD dataset (Cao et al. 2020). To our knowledge, it is the only available dataset for the task of medical expert-layman text style transfer. MSD contains 245k medical training sentences which are each labeled with either the "expert" or "layman" style. Additionally, it contains a test set of 675 expert-layman sentence pairs of equivalent meaning. We extend the training set by producing 11,512 sentence pairs using a margin-based criterion (Schwenk 2018). There are 10810 medical concepts which are used by both expert and layman sentences. We use our edge refinement to create 1124 triples in our Terminology Bijective Graph and build 40,892 expert and 31,083 layman pseudo training sentences for KBA task. Specific statistics are listed in Table 1.

Pretraining Strategy

We use the standard transformer encoder-decoder (Vaswani et al. 2017) with 4 layers, 4 attention heads, and a hidden size of 256 as our language model. We perform a multi-task pretraining procedure where we train a single shared feedforward encoder-decoder framework across KBA and the three SSL tasks, *Mask, Switch* and *Delete*. For the KBA task, we construct 71,975 training sentences from MSD training set using a *Terminology Bijective Graph* described below. We construct training data for *Mask, Switch, Delete* tasks by separately applying their respective noise functions for each sentence in the MSD training dataset, as depicted in Figure 1. We optimize on all four pretraining tasks concurrently, by minimizing negative log-likelihoods:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KBA}} + \mathcal{L}_{\text{Mask}} + \mathcal{L}_{\text{Switch}} + \mathcal{L}_{\text{Delete}} \quad (1)$$

[1]Code available at https://github.com/xu1998hz/SSL_KBA_Expert_Layman_Style_Transfer.

Each mini-batch contains even distribution of each task's pretraining data. Details provided below.

Knowledge Base Assimilation

Knowledge Base Assimilation (KBA) resembles sequence-to-sequence knowledge distillation, except it utilizes a KB rather than a teacher model during training. In particular, we generate a sentence S' from a sentence S in either expert or layman style, where each term in S (a node in the KB) is replaced by a term with the same meaning but in the opposite style (a node opposite the original term along an edge in the KB) shown in Figure 1. The training task then becomes reconstructing S from S' by replacing all the terms, thereby training the LM to model edges in the Terminology Bijective Graph.

In particular, given a source sentence denoted by $S = \{w_1, m_2, w_3, w_4, m_5, ...w_n\}$, where w_i denotes a non-medical word and m_j denotes a medical phrase in the source style. The target style medical phrase which has the same meaning as m_j is denoted by m'_j. Both m_j and m'_j are connected by an edge in the Terminology Bijective Graph. The input of the KBA task is the sentences with the replaced medical phrases in the target style, $S' = \{w_1, m'_2, w_3, w_4, m'_5, ...w_n\}$. The model is required to reconstruct the original sentence from the replaced input sentence. The purpose of the KBA is to enable the model to pick out medical phrases which are misaligned with the sentence style and learn the mapping of concept pairs with identical meaning. We illustrate this process in Figure 1.

Terminology Bijective Graph

To perform KBA for expert-layman transfer in the medical domain, we require a knowledge base of expert term-layman term relation correspondence edges. To achieve this, we build a child knowledge base of the Unified Medical Language System (UMLS) (Bodenreider 2004) containing terms that appear in the aforementioned MSD dataset (Cao et al. 2020).

The UMLS is a standardized knowledge base maintained by the United States National Library of Medicine, containing a collection of Concept Unique Identifier (CUI) codes and corresponding descriptions. CUI codes provide fixed reference to "medical concepts" that are invariant to language or style (i.e., expert vs. layman). For example, CUI 'C0013404' corresponds to both 'dyspnea' and 'shortness of breath.' Cao et al. (2018) match every medical term in the MSD dataset

Number of MSD	Expert	Layman	Ratio
Training Sentences	130,349	114,674	0.88
Testing Sentences	657	657	1
KBA Synthesized Sentences	40892	31083	0.75
Data-Mined Sentences	11512	11512	1
Medical Concepts (CUI codes)	10810	10810	1
KB Triples	1124	1124	1

Table 1: MSD dataset statistics, and quantities for our derived KBA pretraining and fine-tuning data.

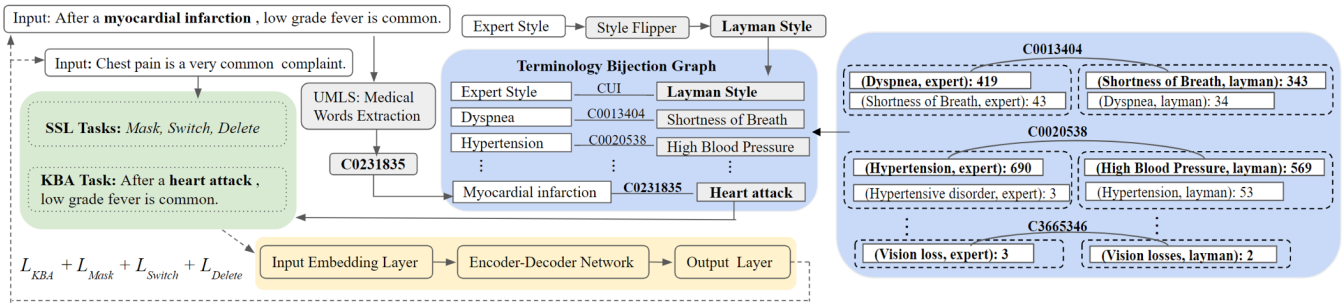

Figure 1: Multi-Task Self-Supervised Learning integrates KBA with *Mask, Switch, Delete* tasks. Four losses are optimized concurrently. The parameters of input embedding layer, encoder-decoder network and output layer are shared between pretraining and fine-tuning. On the right side, we demonstrate the construction of the Terminology Bijective Graph through edge refinement.

to its CUI code using QuickUMLS (Soldaini and Goharian 2016).

Knowing the CUI codes that appear in MSD a-priori, we construct a child knowledge base of UMLS, the *Terminology Bijective Graph*, exclusively containing pairs of terms with shared CUI in the expert and layman styles connected by a bidirectional "is in the other style" relation. To do this, we first collect 10,810 CUI codes which are both found in expert and layman sentences of MSD dataset. Based on those CUI codes, we form 10,810 medical phrasing subsets in each of expert and layman style. However, some phrasings for a CUI appear in both the expert- and layman-labeled medical phrasing subsets. Furthermore, sometimes two phrasings shared the same CUI code in opposite styles are just grammatical variations. We indicate both cases in the right side of Figure 1. To select the best candidate for each style in these cases, we apply a set of heuristics for *edge refinement*.

We first select our expert and layman terms from the candidates. For each CUI, we select one phrase to be the expert term and the other to be layman by a simple majority vote of the MSD style label for the sentences they appear in. For each medical phrasing subset of expert and layman style, the (term, style label) pair with the highest frequency is selected. Thus, each CUI code provides a connection in the Terminology Bijective Graph giving the correspondence between the most frequently used expert and layman phrasing of the underlying concept. Then, we apply Levenshtein distance (Miller, Vandome, and McBrewster 2009) with a threshold ($d = 4$) to exclude candidate phrasing pairs which are simply grammatical variations. This process is shown in the right side of Figure 1.

After the process of edge refinement we are left with a high-quality Terminology Bijective Graph containing 1,124 expert-layman edges with which we perform KBA. Although domain-specific terminological correspondence only requires one relation in the Terminology Bijective Graph, future works can extend KBA to assimilate more complex knowledge graph structures into SSL to novel domains.

Self-Supervised Learning Tasks

In Figure 1, we further augment the KBA edge modeling task with previously demonstrated self-supervised learning (SSL) tasks *Mask, Switch*, and *Delete* on the full MSD training corpus to build a more robust model of medical language. All three are sequence-to-sequence denoising autoencoder tasks, where an original sentence S is the reconstruction target given a perturbed input sentence S'. Details for each SSL task are provided below:

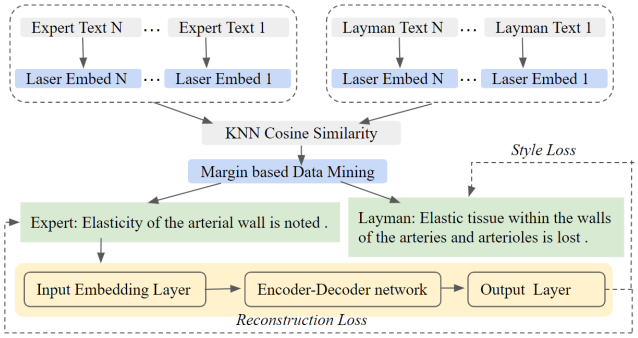

Figure 2: Data mining + Style Transfer Fine-tuning pipeline: Style and Reconstruction losses are optimized concurrently.

Mask The *Mask* task follows the BERT (Devlin et al. 2019) masking scheme. 15% of the word tokens in S are randomly replaced by <MASK> tokens to produce S'.

Switch The *Switch* task generates S' by shuffling the word order of a sentence. For each sentence S, we first select 15% of words in S at random to be shuffled. Then, the selected words are randomly reordered amongst themselves while preserving the order of the unselected words to generate S'. This is similar to Lample et al. (2018)'s noise function.

Delete In the *Delete* task, we randomly delete 15% of the word tokens.In contrast to the Mask Task, the *Delete* task requires the model to learn not only the contextual information for the deleted tokens but also learn the possible positions to insert words. This is similar to the token deletion pretraining task for BART (Lewis et al. 2019).

Fine-tuning for Style Transfer

With our pretrained medical domain encoder-decoder language model, we turn to supervised learning to train an expert-layman medical style transfer model. This requires a parallel

corpus of medical sentences that share the same meaning, $(S_{i,\text{layman}}, S_{i,\text{expert}}) \in C_i$. We collect such a corpus by pairing sentences in the MSD dataset using a margin-based criterion.

Since the transformer already has considerable in-domain modeling capabilities from pretraining, fine-tuning converges very fast. During the fine-tuning stage, the Transformer is modeled by two losses: self-reconstruction and style transfer loss. We use a similar encoding strategy as Style Transformer (Dai et al. 2019), in which we encode a style embedding into the input. Two losses are optimized concurrently, by minimizing two negative log-likelihoods, Equation 2. Therefore, the model will preserve content while rewriting sentences into the target style, shown in Figure 2.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Self-Reconstruction}} + \mathcal{L}_{\text{Style}} \quad (2)$$

Data Mining with Margin Criterion

To collect training data for fine-tuning, we extract 11,512 paired sentences from 245k sentences (approximately 10%) of the MSD training set (Cao et al. 2020) using margin-based criterion (Schwenk 2018). This new parallel corpus is used for fine-tuning of the Transformer model, in Figure 2.

We first extract LASER embeddings (Artetxe and Schwenk 2018) of all sentences in both the expert and layman sets. The margin is defined as the ratio of cosine similarity of two sentence embeddings and the average cosine similarities of k-nearest neighbors in both forward and backward directions. x stands for one sentence in the source style set and y stands for one sentence in the target style set. $N_k(x)$ stands for k unique nearest neighbors of x in the target style set. Similarly for y, the k nearest neighbors are $N_k(y)$.

$$M(x,y) = \frac{\cos(x,y)}{\sum_{z \in N_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in N_k(y)} \frac{\cos(y,z)}{2k}} \quad (3)$$

We use the "*max-strategy*" from (Schwenk 2018) to calculate margin in both directions (expert to layman and layman to expert) for all sentences in both style sets. That allows us to build candidate pairs for both directions (expert$-$to$-$layman and layman$-$to$-$expert). Any sentence can occur, at most, once in the candidate pairs. Therefore, other candidate pairs of that sentence with smaller margin values will be excluded. We use a threshold on the margin score to select candidate sentence pairs that are mutual translations of each other. Discussion on mutual translations can be found in (Schwenk 2018). Margin criteria for our model is set as $k = 4$ and $threshold = 1.06$. Parallel corpus generation took 7.5 hours on a single Titan 1080 Ti GPU.

Experiments

We assess the performance of our strategy by pretraining our transformer encoder-decoder LM in four different conditions:

1. **Basic**, where the transformer model recieves only self-reconstruction loss in pretraining.
2. **SSL Only**, where only the three SSL tasks are used.
3. **KBA Only**, where only KBA pretraining is used.
4. **KBA+SSL**, where the model is simultaneously pretrained using KBA and the three SSL tasks.

For all four conditions, we fine-tune the resulting LM on the expert-layman transfer task using the paired dataset. Furthermore, we compare our model with three prior baselines.

Baseline Models

Following Cao et al. (2020), we choose baseline models that are commonly used and have publicly available code.

1. **DAR** (Li et al. 2018) reconstructs the input sequence after source-target word replacement using edit distance.
2. **Style Transformer** (Dai et al. 2019) uses cyclic reconstruction to preserve content while doing style transfer.
3. **Controlled Generation** (Hu et al. 2018) uses a variational encoder to reconstruct the content representation and an attribute discriminator to build the style vector.

None of these baseline models specifically deal with style transfer in the medical domain.

Training Details

We use the standard training settings for all models with Adam optimizer (Kingma and Ba 2015) and early stops applied. Max sequence length, learning rate and drop out rate are set to $100, 1e-4$ and 0.5 respectively. The three baselines, being unsupervised, are only trained on the non-parallel MSD corpus; they cannot be fine-tuned on our parallel corpus without modification.

Our model architecture follows Dai et al. (2019), with 4 layers, 4 attention heads per layer, and hidden size 256. We add one style token into the input sequence with 256 hidden units after the embedding layer. The positional encoding is applied to the entire input sequence except style embedding. For different SSL task combinations, the pretraining took 6 hours on average and fine-tuning took 1.5 hours on a single Titan 1080 Ti GPU. We use clinical-BERT's (Huang, Altosaar, and Ranganath 2020) tokenization for all models.

Finally, we augment our expected "best" condition of **KBA + SSL** pretraining by making **KBA + SSL Large**, identical to the other transformer models but for a hidden size of 512. We pretrain and fine-tune this model identically to the others.

Human Evaluation

We hired crowdworkers on Amazon Mechanical Turk[2] (AMT) to rate the output of all systems. We collected a random subset of 500 MSD test set sentences to evaluate the performance of all our models. For each source sentence and its style-transferred output, workers were asked to rate the output on three aspects: **content preservation** (the extent to which the two sentences match), **style transfer strength** (extent to which the desired change in style takes place), and **grammar fluency**. Crowdworkers answer questions on each aspect of a set of translations on a 5 point Likert scale.

To ensure the layman crowdworkers understood both sentences, we include supplementary medical definitions for all

[2]Each crowdworker, from the English-speaking locales of {US, CA, UK, AU, NZ, IE} was paid $0.80 per task and averaged 5.5 minutes of completion time with an average compensation of $8.73/hr.

medical terms in each sentence. Crowd-workers were able to access those definitions with a mouse-over of the underlined medical words in the interface (See Appendix Figure 3 and Appendix "Implementation of Human Evaluation Interface").

Due to the knowledge gap between expert and layman sentences, the understanding comparison between the transferred sentence and the source sentence is the most direct way to assess the strength of expert-layman text style transfer. The comparison is quantified by the number of times that crowd-workers had to check supplementary definitions of medical words in the sentence. Fewer checks in the transferred sentence imply an easier-to-understand sentence compared to the source sentence, and vice versa. In the expert to layman (E2L) direction, higher understanding score means the transferred sentence is easier for laymen annotators to understand, while in the direction of layman to expert (L2E), a higher score indicates the transferred sentence is harder to understand.

Following Li et al. (2018), we report six success rates. We consider a transferred sentence "successful" in one evaluation criteria if it is rated 4 or 5 by AMT workers.

Content Success Rate (CSR)—the percentage of sentences that receive 4 or 5 rating in the content criterion.

Understanding Success Rate (USR)—the percentage of sentences that receive 4 or 5 rating for "understand."

Grammar Success Rate (GSR)—the percentage of sentences that receive 4 or 5 rating for the grammar criterion.

Understanding + Content Success Rate (UCSR)—the percentage of sentences that receive 4 or 5 rating for both content and "understand" criteria.

Understanding + Grammar Success Rate (UGSR)—the percentage of sentences that receive 4 or 5 rating for both grammar and "understand" criteria.

Overall Success Rate (OSR)—the percentage of sentences that receive 4 or 5 rating in all three criteria.

We use CSR, USR and GSR to directly assess the model's performance in content preservation, style transfer strength and grammar fluency respectively. We further define a concept, *effective style transfer*: An effective style transfer happens only when model can preserve the sentence meaning or fluency during Style Transfer. We include UCSR, UGSR, OSR to indicate the percentage of sentences that can achieve effective style transfer and OSR can also reflect the overall performance of models on three criteria.

Automated Evaluation

Following previous work by (Dai et al. 2019; Cao et al. 2020), we compute three automatic evaluation metrics (see Table 4). We train a style classifier on the MSD training set using FastText (Joulin et al. 2016) to estimate the style accuracy of the transferred sentence. The **style classifier score** indicates the percentage of the transferred sentences that labeled as corresponding style. We also use NLTK (Bird, Klein, and Loper 2009) to calculate **4-gram BLEU** (Papineni et al. 2002) scores between the transferred sentence and the original sentence. We use KenLM (Heafield 2011) to train a 5-gram language model on the MSD training set. We use **PPL** to measure the fluency of the transferred sentence.

Results

From our human evaluation results in Table 2, our model trained on the SSL task only achieves the highest CSR score. Our KBA+SSL and KBA+SSL Large variants achieve the highest USR, indicating their most progressive style transfer strengths. Although, ControlledGen (CtrlGen) (Hu et al. 2018) achieves the highest GSR, it mostly copies input to the output with limited style transfer, as indicated by its low USR and OSR. All baseline models only achieve limited effective style transfer, as indicated by their low UCSR, UGSR and OSR. In contrast, our KBA+SSL Large has relative improvements over the best performing baseline model DeleteAndRetrieve (DAR) (Li et al. 2018) by 39% in UCSR, 59% in UGSR and 75.6% in OSR.To better understand the characteristics of the models, we provide a case study of an expert-to-layman input example and a layman-to-expert example in Table 5.

Effects of KBA and SSL Tasks Table 2 shows that even the **basic transformer** model without specialized SSL or KBA pretraining achieves competitive USR compared to the baseline models on fine-tuning alone. However, its outputs tend to lose the original sentence meaning or fluency while it adapts to our parallel corpus, indicated by its low CSR and GSR. Moreover, its effective style transfer is limited, indicated by its low UCSR, UGSR and OSR.

Adding **KBA** training (KBA only condition) improves USR by learning terminological mappings between the styles. Surprisingly, as the model learns to reconstruct sentences from the replaced target style medical words both content and grammar scores are improved. Those improvements over all three criteria leads to the enhance of effective style transfer, indicating by UCSR, UGSR and OSR.

Using the auxiliary learning of the **SSL** tasks significantly improves CSR and GSR. This finding verifies our assumptions that context-aware learning gives rise to better apparent content understanding and syntactic awareness. Interestingly, this multitask context-aware learning also demonstrates robust performance in effective style transfer, leading to steep increases of all UCSR, UGSR and OSR.

By adding context-aware learning into KBA (KBA+SSL condition), we observe improvements across all criteria, demonstrating the importance of context-aware learning to the final pretraining scheme. By adding KBA to SSL, we observe consistent improvements of USR, UCSR and UGSR. This finding suggests that the shared representation of context-aware understandings and terminology mappings can improve style transfer strength and this improvement is "effective" in considering content and fluency. However, since the learning of KBA is directly enforced through our Terminology Bijective Graph, KBA-generated representations only have limited sentence context, leading to drops in CSR and GSR in the composed setting, compared to the more context-aware SSL tasks. Therefore, investigating a more sophisticated multi-task pretraining scheme to fully incorporate the power of KBA and context-aware SSLs is a

Model	Human Evaluation					
	CSR	USR	GSR	UCSR	UGSR	OSR
Style Transformer (Dai et al. 2019)	0.70	0.28	0.62	0.18	0.06	0.11
DeleteAndRetrieve (Li et al. 2018)	0.70	0.32	0.47	0.23	0.15	0.12
ControlledGen (Hu et al. 2018)	0.85	0.20	**0.74**	0.10	0.15	0.09
Basic Tr.	0.68	0.30	0.47	0.16	0.09	0.07
KBA Pretraining Only	0.73	0.32	0.60	0.19	0.13	0.11
SSL Pretraining Only	**0.87**	0.30	0.73	0.26	0.22	0.20
KBA+SSL Pretraining	0.83	**0.37**	0.65	0.30	0.22	0.20
KBA+SSL Pretraining Large(512)	0.86	**0.37**	0.67	**0.32**	**0.24**	**0.22**

Table 2: Human Evaluation Table: Our supervised models are below the middle line, the unsupervised baselines are above.

Data	CSR	USR	GSR	UCSR	UGSR	OSR
40%	**0.85**	0.30	0.71	0.25	0.19	0.17
80%	0.85	**0.32**	**0.72**	0.26	0.20	0.18
100%	0.83	0.37	0.65	**0.30**	**0.22**	**0.20**

Table 3: Human evaluation of our KBA+SSL (256) model output using different percentages of our parallel corpus.

good future research direction. We include further discussion in the Appendix Section "Additive Effects of SSL Tasks" to demonstrate how each individual SSL task contributes.

Fine-tuning Data Quantity Effects We repeat our fine-tuning experiment for KBA+SSL (256) on 40% and 80% subsets of our parallel corpus. We find that USR, UCSR, UGSR, and OSR drop compared to using the complete set of parallel sentences. Surprisingly, we found that both CSR and GSR improve when fewer training samples are used. The model might focus less on transferring into target sentence style but more on reconstructing the original sentence in these restricted data conditions. Compared to the three baseline models, both 40% and 80% of parallel data mined sentences outperform baseline models on UCSR, UGSR and OSR. This finding demonstrates the importance of the KBA+SSL tasks and mild parallel data dependency of the pipeline.

Effect of Embedding Size Table 2 shows that training a larger LM on KBA+SSL improves CSR and GSR, leading to a 9.6% relative improvement in OSR. Although USR stays the same, KBA+SSL Large can further enhance effective style transfer, indicating by 6% relative improvement in UCSR and 5% in UGSR. Increased size giving better results is consistent with previous work (Devlin et al. 2019; Lewis et al. 2019; Wang et al. 2019).

Correlation to the Automatic Evaluation

To investigate the quality of autmoated metrics we compute a system-level correlation between BLEU score and human judgement CSR, between style accuracy and our human judgement USR and between PPL and human judgement GSR. Similar to previous results (Li et al. 2018; Cao et al. 2020), we find that the BLEU score has moderate correlation to human evaluation CSR with Pearson correlation (PCC) 0.64 ($p = 0.086$). However, BLEU score is not a reliable indicator in expert layman style transfer, as it tends to penalize the semantically-correct phrases when they differ from the surface form of the reference (Zhang et al. 2019). We find that Style Accuracy has the moderate correlation to human evaluation USR with PCC 0.46 ($p = 0.257$). Similar to previous finding (Cao et al. 2020), we observed that Style Classifier can be easily fooled and achieving high accuracy by adding random target style words, e.g. "patient", into layman sentences and "people" into expert sentences. However, none of those transforms is valid because they don't improve layman or expert's understandings of the original sentences. Although, our KBA+SSL Large performs the best under PPL evaluation, we find a weak correlation between PPL and our human judgement GSR, with PCC -0.38 ($p = 0.352$). Overall, we conclude that three automatic evaluation metrics can be useful for model developments as they exist some correlation to the human evaluation. However, human evaluation is non-replaceable at the current stage. We include further discussion and one concrete example in the Appendix Section "Case Study of Automatic Evaluation" and Table 7.

Model	Style Acc	BLEU	PPL
Style Transformer	0.43	61.2	207
DeleteAndRetrieve	0.64	30.6	141
ControlledGen	0.14	**80.4**	171
Basic Tr.	**0.66**	19.8	338
KBA Only	0.62	37.1	240
SSL Only	0.41	59.2	163
SSL+KBA	0.63	39.2	159
SSL+KBA Large (512)	0.61	40.2	**127**

Table 4: Automatic evaluation results using Style Accuracy, BLEU and PPL

Case Study

In the (E2L) of Table 4, both Style Transformer (Style Tr) and ControlledGen make lexical substitutions to the target style

Model Name	Expert Input → Generated Layman Sentences
Expert Input	Fluid accumulation in the lungs may cause **dyspnea** and **crackles on auscultation** .
Style Transformer	Fluid accumulation in the lungs may cause **attention** and **literally on 4.4** .
DeleteAndRetrieve	Fluid may cause **various symptoms (such as a head injury)** .
ControlledGen	Fluid accumulation in the lungs may cause dyspnea and crackles on **pupils** .
Basic Tr.	Fluid accumulation in the lungs may cause **shortness of liquids during the pregnancy** .
KBA+SSL	The fluid accumulation in the lungs may cause **difficulty breathing (dyspnea)** and **airway narrowing (auscultation)**
KBA+SSL (Lg)	Fluid may be surgically in the lungs and may cause **shortness of breath** .
Gold Reference	If fluid accumulates in the lungs , people may become **short of breath** .

Table 5: Examples of baseline and our model outputs. All model's and expected modifications on medical concepts are bold.

words. However, those changes cause complete deviation from the sentence meanings. In most cases, ControlledGen stays the same as input, which is the reason that it achieves the high CSR and GSR. DeleteAndRetrieve and Basic Tr. are the most progressive baseline models in changing sentence style which seems to be the reason why both of them achieve competitive USRs. However, neither of them achieve effective style transfer, as the transferred sentences are both disfluent and deviating from original meanings, indicated by their low CSR, GSR, UCSR and UGSR. In our two best performing systems, KBA+SSL and KBA+SSL Large, both models are able to accurately translate "dyspnea" to either "difficulty breathing" or "shortness of breath." Moreover, our KBA+SSL is able to deduce the reason of "crackles on auscultation" as "airway narrowing" (See Appendix Table 8 for more examples).

Related Work

Text Style Transfer Due to the limited availability of parallel corpora, most prior work relies unsupervised learning. One approach disentangles style and content representations to generate target-style text sequences by directly manipulating latent representations (Shen et al. 2017; Hu et al. 2018; John et al. 2019). Another approach synthesizes parallel expert-layman sentence pairs through back translation (Prabhumoye et al. 2018; Zhang et al. 2018; Lample et al. 2019) or cyclic reconstruction (Dai et al. 2019; Huang et al. 2020) to enable supervised learning.

Jin et al. (2020) iteratively harvest pseudo-parallel sentences for supervised learning, but this small-scale data mining cannot generate adequately large parallel corpora. Malmi, Severyn, and Rothe (2020) replace source words to the target words using two pretrained masked language models. Similar to our KBA task, Li et al. (2018) reconstruct sentences after source-target word replacement using edit distance. However, simple pretraining scheme and inaccurate edit distance measures restrain their applications to our task. As a result, for this defined task (Cao et al. 2020), domain-specific terminological correspondence knowledge and large parallel corpus generation are our main focuses.

BiText Data Mining Margin-based criterion has demonstrated good performance in low resources setting (Chaudhary et al. 2019; Koehn et al. 2019), LASER embedding (Artetxe and Schwenk 2018) and margin-based bi-text mining (Schwenk et al. 2020). Margin-based embedding has been widely studied in bilingual and multilingual sentence representations (Bouamor and Sajjad 2018; Grégoire and Langlais 2017). But, there is limited prior works that applies this idea to monolingual text style transfer and uses semantic embedding to create pseudo parallel supervisions.

Self-Supervised learning SSL aims to train a neural network with automatically generated data (Peters et al. 2018; Devlin et al. 2019). There are two existing approaches for pretrained language models, feature-based learning (Peters et al. 2018) and fine-tuning (Devlin et al. 2019). To tackle medical domain-specific terminologies, there are many variants (Peng, Yan, and Lu 2019; Alsentzer et al. 2019; Beltagy, Lo, and Cohan 2019; Lee et al. 2019). But, all these works either pretrain on an encoder or a decoder. Therefore, they are not good fits for seq-to-seq generation. pretrained encoder-decoder framework (Song et al. 2019; Lewis et al. 2019) has been successful in many downstream sequence generation tasks like text summary, machine translation and question answering. SSL on text style transfer remains under-studied.

Conclusion

We built a large-scale parallel corpus extending the MSD dataset using margin-based criterion. We introduced a novel pretraining task, knowledge base assimilation, which combined with established SSL tasks produces a high-quality LM to fine-tune with the parallel corpus. This model outperforms unsupervised baselines considerably on human evaluations. We hope that future work will explore a more sophisticated pretraining scheme to fully incorporate KBA and context-aware SSLs and assimilate more complex knowledge graph structures into LMs by extending KBA to novel domains.

Ethics Statement

As we restricted our training to the publicly available descriptive data, systems trained on it have no risk of personal information leaks beyond those generally associated with any web search. The biggest ethical risk to this work is that erroneous transfers could be misleading to users leading to negative clinical outcomes, if and when used in applications. Extensive human evaluations on the system outputs are required at the current stage before pushing this model into the clinical applications.

Acknowledgements

This work was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No.1650114 and by the National Science Foundation award No.2048122. We would also like to thank the Robert N. Noyce Trust for their generous gift to the University of California via the Noyce Initiative. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the sponsors

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. B. A. 2019. Publicly Available Clinical BERT Embeddings. arXiv:1904.03323.

Artetxe, M.; and Schwenk, H. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *CoRR*, abs/1812.10464.

Au, A. 2019. Why Physician Guidance Matters: A Night of Neuralgia, Meningitis, and WebMD. *The Annals of Family Medicine*, 17(5): 462–464.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676.

Benigeri, M.; and Pluye, P. 2003. Shortcomings of health information on the Internet. *Health promotion international*, 18(4): 381–386.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue): 267–270.

Bouamor, H.; and Sajjad, H. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In Rapp, R.; Zweigenbaum, P.; and Sharoff, S., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA). ISBN 979-10-95546-07-8.

Cao, Y.; Hou, L.; Li, J.; and Liu, Z. 2018. Neural Collective Entity Linking. arXiv:1811.08603.

Cao, Y.; Shui, R.; Pan, L.; Kan, M.-Y.; Liu, Z.; and 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*, 1061–1071. Online: Association for Computational Linguistics.

Chaudhary, V.; Tang, Y.; Guzmán, F.; Schwenk, H.; and Koehn, P. 2019. Low-Resource Corpus Filtering using Multilingual Sentence Embeddings. arXiv:1906.08885.

Dai, N.; Liang, J.; Qiu, X.; and Huang, X. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. arXiv:1905.05621.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Grégoire, F.; and Langlais, P. 2017. BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 46–50. Vancouver, Canada: Association for Computational Linguistics.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *In Proc. of the Sixth Workshop on Statistical Machine Translation*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2018. Toward Controlled Generation of Text. arXiv:1703.00955.

Huang, K.; Altosaar, J.; and Ranganath, R. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342.

Huang, Y.; Zhu, W.; Xiong, D.; Zhang, Y.; Hu, C.; and Xu, F. 2020. Cycle-Consistent Adversarial Autoencoders for Unsupervised Text Style Transfer. *CoRR*, abs/2010.00735.

Jerit, J. 2009. Understanding the Knowledge Gap: The Role of Experts and Journalists. *The Journal of Politics*, 71: 442 – 456.

Jin, D.; Jin, Z.; Hu, Z.; Vechtomova, O.; and Mihalcea, R. 2021. Deep Learning for Text Style Transfer: A Survey. arXiv:2011.00416.

Jin, Z.; Jin, D.; Mueller, J.; Matthews, N.; and Santus, E. 2020. IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and Translation. arXiv:1901.11333.

John, V.; Mou, L.; Bahuleyan, H.; and Vechtomova, O. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 424–434. Florence, Italy: Association for Computational Linguistics.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. arXiv:1612.03651.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koehn, P.; Guzmán, F.; Chaudhary, V.; and Pino, J. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3:*

Shared Task Papers, Day 2), 54–72. Florence, Italy: Association for Computational Linguistics.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv:1711.00043.

Lample, G.; Subramanian, S.; Smith, E.; Denoyer, L.; Ranzato, M.; and Boureau, Y.-L. 2019. Multiple-Attribute Text Rewriting. In *International Conference on Learning Representations*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.

Li, J.; Jia, R.; He, H.; and 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)*, 1865–1874. New Orleans, Louisiana: Association for Computational Linguistics.

Malmi, E.; Severyn, A.; and Rothe, S. 2020. Unsupervised Text Style Transfer with Padded Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8671–8680. Online: Association for Computational Linguistics.

Miller, F. P.; Vandome, A. F.; and McBrewster, J. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press. ISBN 6130216904.

Papineni, K.; Roukos, S.; Ward, T.; and 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*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, Y.; Yan, S.; and Lu, Z. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv:1906.05474.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style Transfer Through Back-Translation. arXiv:1804.09000.

Rao, S.; and Tetreault, J. 2018. Dear Sir or Madam, May I introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. arXiv:1803.06535.

Schwenk, H. 2018. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 228–234. Melbourne, Australia: Association for Computational Linguistics.

Schwenk, H.; Wenzek, G.; Edunov, S.; Grave, E.; and Joulin, A. 2020. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. arXiv:1911.04944.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. 35–40.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. arXiv:1705.09655.

Soldaini, L.; and Goharian, N. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. arXiv:1905.02450.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, S.; Che, W.; Liu, Q.; Qin, P.; Liu, T.; and Wang, W. Y. 2019. Multi-Task Self-Supervised Learning for Disfluency Detection. *CoRR*, abs/1908.05378.

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

White, R. W.; and Horvitz, E. 2010. Web to world: predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. In *AMIA Annual Symposium Proceedings*, volume 2010, 882. American Medical Informatics Association.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. *CoRR*, abs/1904.09675.

Zhang, Z.; Ren, S.; Liu, S.; Wang, J.; Chen, P.; Li, M.; Zhou, M.; and Chen, E. 2018. Style Transfer as Unsupervised Machine Translation. *CoRR*, abs/1808.07894.