Neural Text Rewriting

Lai, Huiyuan

DOI:
10.33612/diss.910519765

CHAPTER 13

Conclusions

Deep neural networks have yielded great breakthroughs in natural language generation in recent years, especially with the development of pretrained language models. However, the vast majority of NLP research has focused on literal language, while research on modelling text attributes, or "style" has received less attention. The focus of this thesis is neural text rewriting, investigating the problem of *how to model the "style" of the given text, automatically generating a new text*. To this end, we first provided a thorough overview of neural text rewriting, and then proposed a set of solutions for both modelling approaches and evaluation practices. Specifically, we posed nine research questions, aiming to address several challenges in this field. In the following, we conclude the thesis with our answers to these questions and suggest directions for future work.

RQ1 Can reward learning be used to successfully augment PLMs by explicitly optimising core aspects of style transfer even when small amounts of parallel data are available?

The answer to this question is yes. In Chapter 4, we proposed a framework for supervised formality style transfer, which employs reward learning to explicitly optimise models on two core aspects of the task: style strength and content preservation. We showed that augmenting PLMs with rewards is a successful strategy for this task, especially towards the content aspect. Particularly, finetuning PLMs with 10% of parallel data is more successful than training on 100% of data from scratch, thereby reducing the need for parallel data to achieve competitive performance.

RQ2 Can large amounts of existing generic resources be exploited for style transfer when no task-specific parallel data is available?

To answer this question in Chapter 5, we presented an unsupervised approach based on a general pretrained language model for two rewriting tasks, namely formality transfer and polarity swap. We found that although these two tasks are usually conflated into a single "*style transfer*" label, formality transfer can be seen much more as rewriting than polarity swap due to actual content preservation, which does not exactly apply in polarity swap. For this reason, formality transfer can be conceived akin to the more general task of paraphrasing. Therefore, we proposed to strengthen the model's rewrite ability by training the model on generic paraphrase data for formality transfer, and on synthetic pairs created using general lexical resources for polarity swap. Besides achieving state-of-the-art results, we reflected on the nature of the two tasks and highlighted their differences by showing how they respond to our approach.

RQ3 If parallel data is available in one language only, can parameter-efficient adaptation strategies make it possible to perform style transfer in and with multiple languages?

In Chapter 6, we proposed an adaptation framework to model style transfer in a multilingual fashion exploiting parallel data available for one language only (English). Specifically, we introduced two language- and task-specific adaptation strategies using specific modules with few parameters. The former addresses the problem of some languages being not well represented in the pretrained language model, and the latter supports the transfer of the task knowledge to other target languages. We found that the adaptation strategies with task-specific parallel data from a different language are effective, yielding competitive results and outperforming more classic iterative back translation approaches with style-labelled data. Our approach may serve as a blueprint for other tasks and languages.

RQ4 Taking human judgments as the compass to navigate the evaluation of style transfer, how do automatic metrics fare in comparison to human judgements in different evaluation dimensions?

In Chapter 7, we considered various automatic metrics on the three evaluation aspects of formality transfer and assessed them against human judge-

ments. For style strength, a style regressor performs worse than classifiers when evaluating high-quality systems. Regarding content preservation: (i) if using the source sentence, learnable metrics have much higher correlation scores than those of surface-based metrics; (ii) when human references are available, most metrics are reliable for measuring performances at the system level, surface-based metrics however correlate poorly with human judgment at the sentence level. For fluency, perplexity can be used for evaluating the informal-to-formal direction, while it is clearly less reliable for the opposite direction. Finally, our observations on evaluating against source sentences and human references hopefully stimulate further debate about the nature of evaluation on style transfer tasks and (possibly) other rewriting tasks.

RQ5 Given the need to employ (and train) separate metrics for each evaluation dimension, can LLMs be reliably used instead as single, multidimensional style transfer evaluators in a zero-shot setting?

The answer to this question seems to be yes. In Chapter 8, we investigated the potential of ChatGPT as an evaluator for text style transfer from multiple evaluation aspects: style strength, content preservation, and fluency. By prompting ChatGPT with specific evaluation instructions, we found that it achieves competitive correlations with human judgements compared to existing automatic metrics. This is particularly the case for the most difficult setting evaluated (dataset-level), on which ChatGPT outperforms the other metrics in almost all cases. Our findings on the possibility of using LLMs as multidimensional evaluators for text style transfer and maybe other rewriting tasks, can prompt further discussion in the community about the need to use (and train) separate metrics.

RQ6 Does cross-language and cross-figurative knowledge transfer benefit multiple figurative language detection tasks?

To answer this question in Chapter 9, we introduced a multilingual multifigurative language understanding benchmark that focuses on sentence-level figurative language detection, covering three common figures of speech and seven languages. Based on prompt learning, we proposed a

framework to unify the interrelated detection tasks across multiple figures of speech and languages, while having no task- or language-specific modules. Our results show that the unified model benefits from cross-lingual and cross-figurative knowledge transfer in sentence-level detection. Therefore, our answer to this question is yes. This also provides a new avenue for joint modelling of interrelated tasks, which can be further explored in future work.

RQ7 Can a single model handle the generation between different figures of speech when literal-figurative pairs are only available for individual figures of speech?

In Chapter 10, we introduced the novel task of multi-figurative language generation and proposed a novel framework for this task. Specifically, we leverage paraphrasing data in further pretraining to enhance both form strength and context preservation; we also designed a mechanism for injecting the target figurative information into the encoder, thus guiding the encoder to represent the source sentence. Our framework can achieve generation between different figures of speech even without parallel figurative-figurative pairs. This confirms, as also observed in our detection experiments, that our unified model does benefit from cross-figurative knowledge transfer in the context of figurative language generation.

RQ8 Is our approach applicable to new rewriting tasks without task-specific parallel training data?

In Chapter 11, we proposed the new task of responsibility perspective transfer and experimented with two approaches: unsupervised and prompt learning. Most models score at least 6/10 on average on human judgements on content preservation, and achieve the perspective change where the perceived level of responsibility on the perpetrator increases substantially compared to the source. On the other hand, there is still much room for improvement: none of the models comes close to having the same level of blame as the target sentences do. Overall, our human and automatic evaluations suggest that prompting a LLM (GPT-3) performs the best, with a relatively high level of responsibility placed on the perpetrator and a good degree of content preser-

vation. This shows that it becomes feasible to leverage large language models to perform more rewriting tasks.

RQ9 Can meaning representations be included in the pretraining of a language model together with natural language for meaning-to-text generation?

The answer to this question is yes. In Chapter 12, using DRS-based meaning representations as an additional language aside natural languages, we introduced a novel multilingual pretrained language-meaning modelling framework for meaning-to-text generation. Indeed, exploiting parallel data and DRS language neutrality allows to boost performance in lesser-resourced languages. To reduce the gap between the pretraining and finetuning objectives, we proposed a supervised denoising training that exploits more explicitly the relationship between a DRS and each corresponding text as well as between the parallel texts in the different languages. Our experiments show that our approach outperforms several baselines on the task of multilingual DRS-to-text generation.

Outlook

This thesis has presented a set of solutions and findings for neural text rewriting. Furthermore, one higher-level contribution of this thesis is that we provided a deeper understanding of the nature of different rewriting tasks, as well as a unified view of the joint modelling of interrelated tasks. We expect this to inspire future research by providing other researchers with a different perspective to rethink research directions in this area. In the following, we outline some possible extensions of this thesis that we believe are critical and valuable for neural text rewriting.

Data The development of NLP over the past few decades is closely related to data, especially in the era of pretrained language models. We believe that data-centric research will continue to play an important role in the next decade. In NLP, constructing a standard benchmark dataset is often the first and crucial step to tackle a new task, which can enable the development of various models, the assessment of their performances, and comparisons

among them. Although NLP researchers have built various benchmarks for a wide range of text rewriting tasks, there are several potential directions for building more diverse rewriting benchmarks: (i) as most works focus only on English, it is important to construct multilingual datasets that have wide language coverage; (ii) although there have been many successful works on sentence-level text rewriting, document-level rewriting, a more challenging task with many promising applications such as AI-driven writing assistance, has received very little attention in the NLP research community (Lin et al., 2021), especially in attribute-controlled text rewriting; (iii) we focus on modelling text surface properties such as formality and sentiment, as well as more complex figures of speech, more diverse ones such as personalized style is a promising direction for future work. For the data construction aspect, since data annotation is a time-consuming and expensive process, data augmentation (Feng et al., 2021) and LLM-aided annotation (He et al., 2023) will therefore be important research directions in the future. All in all, from our perspective, data-centricity is a promising trajectory for future work.

Models This thesis mainly employs state-of-the-art PLMs to address several challenges facing text rewriting. In recent years, the PLMs scale has continued to increase, reaching the order of hundreds of billions of parameters, leading NLP and even the entire artificial intelligence field into a new era. In text rewriting, Wahle et al. (2022) show that GPT-3-generated paraphrases are rated as high-quality by human experts as the original scientific text provided in input, and we also showed that LLMs have good potential in both modelling and evaluation in Chapters 8 and 11. In this context, leveraging LLMs for text rewriting could be a promising direction to extend the practical impact of this thesis. For example, one potential direction is modular learning, as we discussed in Chapter 6, which allows a subset of modules to be specialized for the target task (or style). This not only mitigates catastrophic forgetting and training costs of LLMs, but also enables the deployment of LLMs in a lightweight manner.

Applications Text rewriting can serve a wide range of purposes, primarily dividing into two categories: aiding in various downstream NLP tasks and supporting the development of application products. Previous works have

shown that text rewriting can help with many NLP tasks, such as text sumarization (Cao et al., 2017) and question answering (Xu et al., 2016). For further research, relevant to this thesis, controlling the style and tone of generated text (e.g. a summary) is a promising direction. In product applications, for instance, recent work showed that metaphors with varying degrees of perceived warmth can shape users' expectations of conversational agents, and thus lead to different effects on many aspects such as willingness to use and cooperation (Khadpe et al., 2020). Therefore, future work can investigate how to generate text with an appropriate style and use of figurative language on the basis of the context to improve user satisfaction and engagement of conversational agents.