Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210325

Few-Shot Tuning Framework for Automated Terms of Service Generation

Ha Thanh NGUYEN [a,1], Kiyoaki SHIRAI [a] and Le Minh NGUYEN [a]
[a] *Japan Advanced Institute of Science and Technology*

Abstract. In this paper, we introduce BART2S a novel framework based on BART pretrained models to generate terms of service in high quality. The framework contains two parts: a generator finetuned with multiple tasks and a discriminator finetuned to distinguish the fair and unfair terms. Besides the novelty in design and the implementation contributions, the proposed framework can support drafting terms of service, a growing need in the digital age. Our proposed approach allows the system to reach a balance between automation and the will expression of the service provider. Through experiments, we demonstrate the effectiveness of the method and discuss potential future directions.

Keywords. few-shot tuning, terms of service, generation

1. INTRODUCTION

Natural language generation comes along with the development history of NLP. The first generation of these systems are simple rule-based systems, typically represented by ELIZA (1). These systems have a complex set of rules but can only generate language in a very limited context. Later systems with knowledge bases and statistical methods can perform this task better in more problems such as weather forecasts (2), storytelling (3), and dialogue system (4). With breakthroughs in hardware and architecture in recent years, transformer-based systems like GPT-2 (5), GPT-3 (6) and BART (7) are recently received great attention from both the industry and the research community.

These models have been very successful with destructive or teasing applications such as fake news, fake images, fake videos, but that does not guarantee they can generate high-quality content like terms of service. Compared to a naive copy-paste mechanism, a generative model generates not only the learned examples but also the synthetic sample from them. As a result, it gives editors more flexibility in drafting the documents. However, this problem contains two main difficulties. First, it requires a balance between automation and the will of the editor. Second, the meaning of the content that the system generates should be of high quality and fairness.

Finding a solution for terms of service generation problem, this paper proposes BART2S, a novel generator-discriminator framework for *terms of service generation*.

[1] Corresponding Author: Nguyen Ha Thanh, Japan Advanced Institute of Science and Technology; E-mail: nguyenhathanh@jaist.ac.jp

The generator is designed to solve the problem called *title-based generation* in order to both express the will of the editor and reduce their drafting effort. The editor provides a title and the framework completes the content according to the patterns it learned from the data. To implement this paradigm, we pretrain the generator with multiple sequence-to-sequence tasks. The discriminator is trained on the same vocabulary to classify the generated terms as fair or not. Our experiments show interesting results and prove the effectiveness of the approach.

2. BART2S Framework

2.1. Title Based Generation

Each content in terms of service has a title reflecting it. We recognize this feature as an opportunity for editors to participate in editing with minimal effort. The title is usually a sentence that describes the topic of the content and can even reflect the editor's point of view. Therefore, we use the title as information for the editor to guide the system. With a title as input and content as output, we propose the title-based generation problem. This problem can be considered as a conditional generation problem that the generated content must reflect the topic mentioned in the title.

The problem brings a challenge in signal recovery, usually, the information in the title is often much more concise than its content. To be able to fulfill the ideas from a short sentence or even a word of the title, the model needs to understand the patterns of idea development in a particular domain. For meaningful generated content, the pretraining stage needs to be done with the appropriate tasks and the appropriate data domain. We propose three tasks that help to train the generator for the desired goal: writing the next sentence, writing content from the title, and paraphrasing. In addition, we propose using a pretrained discriminator to evaluate and adjust the output of the generator.

2.2. Pretraining Encoder

The first task is the next sentence generation. We prepare the noised input similar to how BART (7) is trained. Given a noised sentence, the model needs to generate the sentence right after it. Our goal in training the model on this task is for the model to learn how to use words in the field of law. We assume that this skill can bring a better generation for title based generation problem. In addition, the title that the editor input can be an incomplete sentence. Trained by this task, the model is able to complete the idea from the input.

The second task that the model needs to learn is to generate content from the title. This task is directly related to the title-based generation problem. The title of a paragraph is usually a summary of it or the topic it covers. Generating content from the title demonstrates the model's ability to understand the title as well as find out the content representing that topic. This task also serves as a model guide in generating the desired output as the content from the title as input.

The third task is about paraphrase generation. The skills required in this task enable the model to understand the text and represent it in a different way. This task is useful to train the model not only for the flexibility of the model but also for the coherence of the generated content. In essence, the content is a paraphrase of the title with ancillary information. Our assumption is that learning to paraphrase will help the model to generate the content from the title better.

2.3. Pretraining Discriminator

The discriminator proposed in this paper is used for regulating generated content. It is pretrained to distinguish between fair and unfair terms. The discriminator is fed by the input having the same format as the output of the generator. Let $C = [wc_1, wc_2, ..., wc_n]$ be the content and $L = [0, 1]$ be its corresponding fairness label. The discriminator is trained to map the content with their fairness label. This component makes our approach different from other systems; it enables us to build a system toward a constructive goal of generating high-quality content.

2.4. Cross-model Few-shot Tuning

The models can be represented as differentiable functions $G(x, \theta_g)$ and $D(x, \theta_d)$ with x, θ_g, and θ_d are the input, generator's parameters and discriminator's parameters, respectively. Tuning the generated output by the generator, we minimize $log(1 - D \circ G(x, \theta_g))$ using gradient descent process. In the backpropagation, for the loss to be able to pass through the two models, we replace *ArgMax* function at the last layer of the generator with *SoftArgMax* function represented in Equation 1.

$$\text{SoftArgMax}(x) = \sum_i \frac{e^{\beta x_i}}{\sum_j e^{\beta x_j}} i \tag{1}$$

where $x = [x_1, x_2, x_3, ..., x_n]$ and $\beta \geq 1$.

The discriminator's weights are frozen during the tuning process, which guarantees that this component is an independent observation. It is only based on the knowledge learned during the pretraining process to make the assessment. The loss reduces when the generator adjusts the generated content to make it fairer. This ability creates a bold difference in our framework compared to naive copy/paste systems and other non-regulated systems.

3. Experiments

3.1. Experimental Setup

Generator. Task 1's data is formed from the content of terms of service documents we collect on the Internet. Data for Task 3 is extracted from MSRP dataset (8), we only keep the sentence pairs with positive labels for paraphrasing. Data for Task 2 and data for evaluation are crawled from *Law Insider*[2], an online corpus that contains contract terms with their title. For each input, we use Token Masking, Token Deletion, and Token Infilling to transform the input in the same way that BART is pretrained. Sentence Permutation and

[2] https://lawinsider.com

Document Rotation are not applicable in this case. With this transformation, the model must learn the output based on the incomplete input. After being processed as above, our training data has 5,323 samples for Task 1; 901 samples for Task 2; and 3,728 samples for Task 3.

The terms of service generation problem is a multiple ground truth problem. In fact, there are many terms with the same name but different content. Therefore, we designed the test set to fit that characteristic. The ground truth of each title includes 1,000 most popular corresponding content according to statistics of Law Insider. Accordingly, the BLEU score is the measure we use to evaluate the performance of the model according to different training strategies.

Discriminator. For the discriminator, we use labelled samples provided by *ToS;DR project*[3] to finetune and evaluate the model. There are in total 4,152 samples in which 2,308 samples are fair terms with positive labels and 1,844 samples are unfair terms with negative labels. For both components, we use the configuration of *BART large* to initialize the models.

BART2S Framework. After finetuning the generator and the discriminator, we verify the BART2S framework presented in Section 2. The generator generates temporary output, and this output is assessed by the discriminator. The generator's weights are updated until the output meets the condition of fairness verified by the discriminator.

We design human-based metrics to evaluate and compare the tuned model with the BART2S framework and other candidates with the same configuration. In terms of ToS generation, we consider 4 aspects of good content as *Grammar*, *Readability*, *Relevance*, and *Fairness*. The content needs to be written in human language with good grammar, readable, and relevant to the title. Most importantly, the model needs to generate a fair term.

Among the release model of BART (7), we only consider models with BART Large configuration. Besides, for each classification and generation tuning task, we choose one candidate with the best performance reported by the authors. Finally, the three candidates to compare with the outputs of BART2S are as follows:

- BART Large w/o Ft: BART Large without finetuning on any task.
- BART Large MNLI: BART Large finetuned on MNLI dataset (9).
- BART Large CNN: BART Large finetuned on CNN-DM dataset (10).

These models are used as an end-to-end generator without the discriminator part as proposed in the BART2S framework. We create a collection of 30 short titles with an average length of 23 characters, feed them in the models and invite 10 evaluators to assess the generated content with the 4 metrics mentioned above. For each metric, we use a binary evaluation, the evaluators only need to check whether the content is acceptable or not. To avoid biases in the assessments, we only provide the evaluators with the title and the corresponding generated content. The evaluators do not know about the models and the process of generating the content.

The final score of each model in each aspect is calculated as in Formula 2.

$$score_a(\mathrm{M}) = \frac{1}{n}\sum_{i=1}^{n}\frac{p_a^i}{s} \tag{2}$$

[3] https://tosdr.org/

In which, $score_a(\mathrm{M})$ is the evaluation score of model M in aspect a, s is the total of sentences, n is the number of evaluators, p_a^i is the number of sentences evaluated as possitive by i^{th} evaluator in the aspect a.

3.2. Experimental Results

Approach	BLEU Score
All tasks	**60.07**
W/o Task 1	59.85
W/o Task 2	57.29
W/o Task 3	56.34

Table 1. Performance of generator trained with different approaches.

Generator. Table 1 summarizes our experimental results on training the generator with different settings. The model training with all tasks achieved the best performance. The surprising thing about the experimental results was that the model trained without Task 2 was not the model with the worst performance. From that result, we assume that Task 2 can be learned indirectly through the next sentence generation task and paraphrasing task. This once again confirms the idea of using multi-task learning for this problem is appropriate.

System	Grammar	Readability	Relevance	Fairness
BART Large w/o Ft	0.34	0.31	0.41	0.43
BART Large MNLI	0.32	0.33	0.37	0.37
BART Large CNN	0.69	0.73	0.72	0.86
BART2S	**0.80**	**0.82**	**0.87**	**0.94**

Table 2. Evaluation results on grammar, readability, relevance, and fairness of each system. The underlined line indicates our proposed system.

Discriminator. With the early stopping setting, the discriminator training process ends when the loss value on the validation set stops to decrease after 10 epochs. The accuracy on the training set is 66.6% and the accuracy on the validation set is 66%. These values reflect the difficulty of the fairness classification problem. It's not straightforward to detect an unfair term provided only its content. However these values are significantly greater than a random guess, which proves that there are latent patterns that support the model to do the task.

BART2S Framework. We feed the models with the short titles as described in Section 3.1. The max length of the generated content for every model is set to 512 subwords. With the given 30 titles, BART2S needs at most 2 epochs to tune the generator for generating desired content. Since we do not provide any ground truth of the data, BART2S is solely based on pretrained knowledge to adjust the outputs. Table 2 presents the evaluation results on grammar, readability, relevance, and fairness aspect of BART2S Framework and the controls. The BART2S framework leads all evaluation aspects, followed by the BART Large CNN model. BART Large w/o Finetuning model and the BART Large MNLI perform worst in the ranking.

4. Conclusions

This paper proposed BART2S, a regulated generative framework for generating terms of services automatically using the generative models. To ensure a balance between automation and expression of will, the framework is based on the title-based generation problem. The framework contains two sub modules as a generator and a discriminator. We use a custom pretrained model trained on 3 different tasks as the generator and a pretrained classification model with the same configurations as the discriminator. We also propose a novel tuning process to adjust the fairness of the generated content. The experimental results show that our approach is appropriate and the framework can produce high-quality results. Although this framework was proposed in the terms of service generation problem, the idea is general and can be applied to many different fields. Replacing the fairness discriminator with another set of constraints could be a potential direction.

References

[1] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966;9(1):36-45.

[2] Angeli G, Liang P, Klein D. A simple domain-independent probabilistic approach to generation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010. p. 502-12.

[3] Holtzman A, Buys J, Forbes M, Bosselut A, Golub D, Choi Y. Learning to write with cooperative discriminators. arXiv preprint arXiv:180506087. 2018.

[4] Wolf T, Sanh V, Chaumond J, Delangue C. Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:190108149. 2019.

[5] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.

[6] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:200514165. 2020.

[7] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:191013461. 2019.

[8] Dolan WB, Quirk C, Brockett C. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics; 2004. p. 350-6.

[9] Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:170405426. 2017.

[10] Hermann KM, Kočiskỳ T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. arXiv preprint arXiv:150603340. 2015.