Learning to Generate Explanation from e-Hospital Services for Medical Suggestion

Wei-Lin Chen,[1] **An-Zi Yen,**[2] **Hen-Hsen Huang,**[3] **Hsin-Hsi Chen**[1]
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2]Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan
wlchen@nlg.csie.ntu.edu.tw, azyen@nycu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

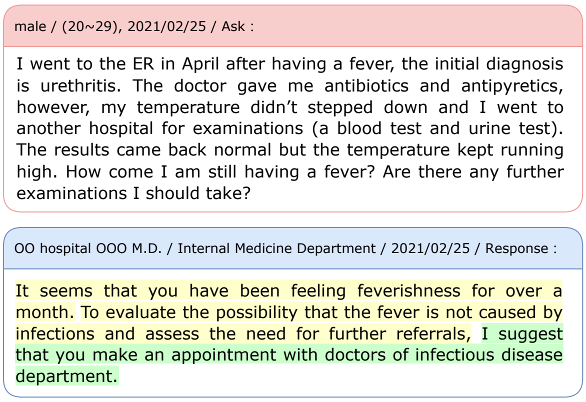
Abstract

Explaining the reasoning of neural models has attracted attention in recent years. Providing highly-accessible and comprehensible explanations in natural language is useful for humans to understand the model's prediction results. In this work, we present a pilot study to investigate explanation generation with a narrative and causal structure for the scenario of health consulting. Our model generates a medical suggestion regarding the patient's concern and provides an explanation as the outline of the reasoning. To align the generated explanation with the suggestion, we propose a novel discourse-aware mechanism with multi-task learning. Experimental results show that our model achieves promising performances in both quantitative and human evaluation.

1 Introduction

Neural models have shown remarkable success in various tasks, however, simply offering the predictions may not satisfy the requirement of end-users. Understanding how the decision has been reached by the model is essential in real-world applications. To provide a meaningful, human-comprehensible explanation, presenting it in natural language is a proper fashion. Note that simply present the explanation as a shopping list or fragments of text-highlight is not an ideal way. Humans prefer to read a text composed of a narrative structure, organized by the discourse relations elaborating the causality between the model input and output (Reiter, 2019). In this work, we propose a novel health consultancy model which can provide medical suggestions accompanied with natural language explanations learned from medical specialists to help humans make decisions in their daily life.

As many people are eager for help in addressing their health concerns, a model is necessitated to be capable of not only providing suggestions regarding their concerns, but also explaining the suggestions, alleviating their worries before visiting the doctors. Taking Figure 1 as an example, given a question asked by a patient, the physician answers it by explaining what disease might cause the symptoms mentioned in the question and suggests which medical specialty the patient should seek. In this example, the response has a clear narrative structure, where the explanation and suggestion are denoted by the yellow and green highlights, respectively. First, the patient's concern is addressed (e.g., having a fever for over a month). Then, before giving the suggestion, the physician explains the causality between the concerns and the suggestion. Recently, Explainable Artificial Intelligence (XAI), which aims at explaining how the decision is reached by the machine learning model (Ribeiro et al., 2016; Mullenbach et al., 2018; Pezeshkpour et al., 2019), has been gaining attentions in the research community, including works that provide textual explanations (Wu and Mooney, 2019; Rajani et al., 2019; Brahman et al., 2021). Generally, generating textual explanation can be regarded as a natural language generation task. Typically, most works collect explanations by asking annotators to write free-text sentences. Since human annotation is expensive, labor-intensive, and time-consuming, especially in domains where expertise is needed,



male / (20~29), 2021/02/25 / Ask :

I went to the ER in April after having a fever, the initial diagnosis is urethritis. The doctor gave me antibiotics and antipyretics, however, my temperature didn't stepped down and I went to another hospital for examinations (a blood test and urine test). The results came back normal but the temperature kept running high. How come I am still having a fever? Are there any further examinations I should take?

OO hospital OOO M.D. / Internal Medicine Department / 2021/02/25 / Response :

It seems that you have been feeling feverishness for over a month. To evaluate the possibility that the fever is not caused by infections and assess the need for further referrals, I suggest that you make an appointment with doctors of infectious disease department.

Figure 1: A Data Instance from the Health Consultancy Website

an alternative way is to construct synthesized explanations by designing rules to exploit information from other datasets as explanations (Li et al., 2018). Although previous works demonstrate that introducing an auxiliary generation task to explain the prediction enables performance improvement, two main issues remain to be tackled: **(1)** The explanations are annotation artifacts (Gururangan et al., 2018) since the annotators typically write vanilla and trivial description (Lei et al., 2020), resulting in a lack of linguistic variety (Parikh et al., 2020) that leads the model prone to overfit on annotator characteristics (Geva et al., 2019). **(2)** Whether the model faithfully explains the suggestion is still an open question (Jacovi and Goldberg, 2020). Producing an explanation just mimicking the way humans would say is impractical in fields involving high-stake scenarios.

To address the aforementioned two issues, our work is based on real-world data. We collect 86,399 question answering (QA) pairs [1] from an online health consultancy website called Taiwan e-Hospital,[2] which allows users to ask questions regarding their health conditions and physicians will respond to their concerns. The linguistic diversity from multiple users is greater than free-text produced by a few crowd-workers, and the explanation within an answer is more natural than handcrafted annotation. Then we propose pilot models to generate response consisting of explanation and suggestion according to the question. Note that our approach can be easily generalized since it is language and domain independent. The contributions of our work are summarized as follows:

1. We show a pilot study on health consulting with professional explanations.
2. We propose a novel discourse-aware mechanism that aligns the generated explanation with the suggestion.
3. Both qualitative and human evaluation show that our discourse-aware model achieves promising performances on suggestion and explanation generation.

2 Dataset

As mentioned in Section 1, we construct our dataset by crawling approximately 86k QA pairs from Taiwan e-Hospital website, where an answer is a free-text response written by the physician, containing the explanation and suggestion as Figure 1 shows. An ideal instance would be a triple of (q, s, e), where q is the question asked by the patient, and s and e are the suggestion and the explanation responded by the physician. However, the crawled raw text often carries greeting terms, salutations, and personal information, such as names of the patients and doctors, which are noise for our task and should be pruned. To gather the desired (q, s, e), we propose a rule-based keyword matching method to extract text snippets that belong to the suggestion and explanation, defined as follows.

- **Suggestion:** The suggested action regarding the patient's concerns, such as whether to seek medical attention, the department for making an appointment with, or the follow-up examination to undergo.
- **Explanation:** The text describing why a physician gives the suggestion. Generally, it includes medical knowledge to address the patient's concerns.

The details of our method are shown as follows.

Step1: We define a set of keyphrases $\mathcal{G}$ that belongs to greeting terms or salutations by regular expressions. Given a sequence of sentences $X = (x_1, x_2, ..., x_n)$ in the response, the i-th sentence x_i that contains a word $w \in \mathcal{G}$ will be filtered. Afterwards, a sequence of sentences $X^{'}$ is obtained, where $1 \leq |X^{'}| \leq n$.

Step2: We find that the sentences belonging to suggestions usually contain certain keywords such as "suggest", "recommend", and the name of the department to "seek". Hence, to identify whether a sentence $x \in X^{'}$ belongs to a suggestion, we manually collect a set of keywords $\mathcal{K}$ from several responses. Then, the sentences in $X^{'}$ that contain a word $w \in \mathcal{K}$ are regarded as the suggestion, and the remaining sentences are considered as the explanation.

Step3: In addition to preparing the (q, s, e) triples, we also construct binary labels $(0/1)$ from s, denoting whether the patient should receive medical assistance based on his/her health condition. Taking Figure 1 as an example, it would be labeled as 1 since the doctor suggests the user to make an

[1] The dataset for this paper is available at `https://github.com/ntunlplab/tw-eH`

[2] `https://sp1.hso.mohw.gov.tw/doctor/`

Setting	Input	Output
R_1	[user's question]	[doctor's response]
R_2	Suggest: [user's question] Explain: [user's question]	[doctor's suggestion] [doctor's explanation]
R_3	Suggest: [user's question] Explain: [user's question]	[doctor's suggestion] [doctor's explanation] [binary label for medical assistance]

Table 1: Three Input-Output Settings in the Experiments

Step1	Step2	Step3
0.82	0.80	0.81

Table 2: Evaluation Results of Regular Expressions in the Three Steps

appointment. In our expectation, this label, which indicates how serious the health risk the patient is facing, can play a useful auxiliary task. We compose a set of patterns to identify whether the physician suggests the patient to seek medical attention (e.g., mentioning a medical department in the suggestion).

The processed result from each step, namely R_1, R_2, and R_3, is utilized as references for our different proposed methods described in Section 3, i.e., *mT5* (3.1), *MTL mT5* (3.2), and *DMTL mT5* (3.3), respectively. The details of the dataset formats are shown in Table 1.

To validate the results of our regular expressions in Steps 1, 2, and 3, we randomly sample 100 instances from R_1, R_2, and R_3, respectively. And by checking the correctness of these instances with human evaluation described as follows, we can assess the quality of the regular expressions. For Steps 1 and 2, an instance is considered incorrect if it contains sentences that should be filtered or missing sentences that should be retained. That is, the regular expression admits or filters the wrong sentences. Otherwise, it is considered as correct. For Step 3, the correctness of an instance is determined by checking if the binary label l is the same as whether the physician suggests the patient to seek medical attention or not. The results are measured by $\frac{\#\text{ correct instances}}{100}$ and presented in Table 2. And Table 3 shows the statistics of the top-10 departments ranked by the number of QA pairs, where Exp. and Sug. indicate Explanation and Suggestion, respectively.

		Avg # sent.	
Department	# QA pairs	Exp.	Sug.
Gynecology & Obstetrics	11,676	7.28	1.45
Gastroenterology	7,497	6.36	1.56
Dermatology	7,487	5.93	1.38
Urology	6,895	8.55	1.42
Orthopedics	6,870	6.60	1.47
General Surgery	5,966	7.10	1.51
Ophthalmology	5,086	9.86	1.44
Otorhinolaryngology	5,010	7.54	1.60
Psychiatry	4,489	13.78	1.72
Dentistry	3,998	6.84	1.38

Table 3: Statistics of the Top-10 Departments with Average Numbers of Explanation (Exp.) and Suggestion (Sug.) Sentences

3 Methodology

In this section, we introduce our models for the task of suggestion and explanation generation. As the references of suggestions and explanations are collected by handcrafted rules without human annotation, the models are weakly supervised. Given a question q, our goal is to learn a generator $\mathbf{gen}(\cdot)$ to generate a textual response with a narrative structure consisting of a suggestion s and an explanation e. Three $\mathbf{gen}(\cdot)$ models are described as follows.

3.1 The mT5 Model

We adopt the pre-trained multilingual T5-base model (Xue et al., 2021), which casts natural language processing problems in an unified "text-to-text" form with great flexibility. The input and output (label) data are the patient's question and the corresponding response R_1, i.e., the result of Step 1 in Section 2. Since R_1 does not explicitly extract s and e from the given responses, the generated results of *mT5* do not distinguish the suggestions and the explanations. The pre-trained mT5-base model

	Full Response			Suggestion			Explanation		
Method	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5	20.335	6.844	17.135	–	–	–	–	–	–
MTL mT5	21.470	7.429	19.383	22.559	7.615	21.764	20.691	7.296	17.679
DMTL mT5	**22.176**	**7.619**	**20.096**	**22.717**	**7.840**	**21.893**	**21.789**	**7.461**	**18.811**

Table 4: Results of Suggestion and Explanation Generation, Reported in ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L)

also serves as the backbone for the following two $\mathbf{gen}(\cdot)$s.

3.2 The mT5 Model with Multitask Learning

Since *mT5* does not always generate suggestions and explanations as expected, i.e., the generated response would sometimes contain only suggestion or only explanation, we implement a multi-task learning mT5-base model, *MTL mT5*, to address this issue. To train the *MTL mT5*, we use R_2 as the output data. For input data, we add a prefix text, "Suggest:" or "Explain:", to specify which task the model should perform. Concretely, given q and a response (s, e) in R_2, the two formats of (input, output) data are ("Suggest: q", s) and ("Explain: q", e). In this way, the *MTL mT5* model can generate suggestions and explanations explicitly.

3.3 Discourse-aware MTL mT5 (DMTL mT5)

With the multi-task setting, given an input question q, both the suggestion s and the corresponding explanation e are generated. Ideally, the user can assess the need to seek medical attention based on the model's generated suggestion accompanied with the explanation. However, if the explanation cannot support the suggestion, i.e., they are not related, the user would be confused, so that s/he may decrease the degree of confidence to the system.

To mitigate this problem, we use the dataset R_3, which contains the binary labels indicating whether the patient needs to receive medical assistance. We propose a discourse-aware mechanism into the *MTL mT5* model by introducing a new objective function with a weighted parameter to focus on generating the correct binary label l, where $l \in \{0, 1\}$. We view l as a concise summary of the suggestion, and assume that the model would generate explanations aligned with suggestions in order to predict this binary label l. Specifically, considering a sequence of n words with a binary label $Y = (y_1, y_2, ..., y_n, l)$ output by *DMTL mT5* and the reference sequence $\hat{Y} = (\hat{y}_1, ..., \hat{y}_n, \hat{l})$, the weighted cross-entropy loss ψ of the task of explanation generation is computed as follows:

$$\psi = \left(\sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) \right) + \alpha \times \left(\mathcal{L}(l, \hat{l}) \right)$$

where $\mathcal{L}$ denotes the cross entropy loss, and α is a hyper-parameter for the weighted loss function. We set $\alpha = 1.1$ by tuning with the validation set.

Note that the main purpose of the binary label l is to provide loss signals encouraging the generated explanations to align with the suggestions. For inference, l is not exposed to end-users, that is, we conduct post-processing to trimmed l from the generated explanation.

4 Experiments and Discussions

We conduct both quantitative and qualitative evaluations to compare the generated suggestions and explanations of our proposed methods. The dataset are randomly split into train, validation, and test sets by the ratio 8:1:1 (69,119, 8,640, 8,640), where every instance is a QA pair. And we adopt teacher-forcing strategy (Williams and Zipser, 1989) with the cross-entropy loss as the objective function for optimizing all models, i.e., *mT5*, *MTL mT5* and *DMTL mT5*. The results reported in this section are conducted on the test set.

4.1 Quantitative Evaluation

The ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) between the generated responses and the reference responses are shown in Table 4, denoted as R-1, R-2, R-L, respectively. Since *mT5* does not generate s and e individually, we combine the generated s and e from the multi-task learning models, i.e., *MTL mT5* and *Discourse-aware MTL mT5*, as one full response to compare across three models. As shown in Table 4, the multi-task learning models outperform the *mT5* model. It confirms

Method	Relevan.	Suggest.	Explan.
mT5	3.64	3.09	2.00
MTL mT5	3.75	3.52	2.21
DMTL mT5	**3.87**	**3.56**	**2.42**

Table 5: Results of Human Evaluation

that explicitly learning how to generate suggestions and explanations is a proper fashion. Furthermore, the *Discourse-aware MTL mT5* outperforms other models in all ROUGE metrics. It shows that introducing weighted loss benefits the explanation generation as well as the suggestion generation.

We also measure the statistical significance level with the sampling-based bootstrap test, following the guidelines of (Dror et al., 2018). We compare the *DMTL mT5* with *mT5* and *MTL mT5* on the full response, and *DMTL mT5* significantly outperforms the other models at $p < 0.05$. To further measure the qualities of the generated suggestions and explanations, we also conduct qualitative human evaluation in Section 4.2.

4.2 Human Evaluation

For human evaluation, we invite a group of physicians and randomly sample 100 instances from the test set, where each instance is assigned to two physicians to assess the following three aspects:

1. **Relevance**: whether the generated response is related to the patient's question.

2. **Correctness of suggestion**: whether the generated suggestion is correct.

3. **Correctness of explanation**: whether the generated explanation can explain the generated suggestion and help patients understand the reason why such a suggestion is given.

Note that for the multi-tasking methods, we present the generated suggestion and explanation of each instance jointly as one response to the physicians. Each aspect is ranging from zero (does not meet the given aspect) to five (totally meets the given aspect).

The evaluation results are reported in Table 5, where Relevan., Suggest., and Explan. correspond to the three aspects described above, respectively. The *Discourse-aware MTL mT5* achieves the highest scores in all aspects, suggesting that the discourse-aware mechanism enables the model to generate explanations more aligned with suggestions, and makes the response more relevant to the question. Compared to the single task learning *mT5*, *MTL mT5* and *Discourse-aware MTL mT5* obtain higher scores on "Correctness of Suggestion" and "Correctness of Explanation", indicating that multi-task learning makes the model more attend on learning information benefiting both tasks. Overall, our proposed models achieve promising performances on generating suggestions by learning from the QA pairs only. However, the scores obtained on "Correctness of Explanation" are lower than half of the full score. This might indicate that generating correct explanations is still challenging due to the lack of medical knowledge.

5 Conclusion

This paper proposes a discourse-aware generative model based on multi-task learning to generate narrative structured responses consisting of suggestions and explanations to the questions. Experimental results show that our model with the discourse-aware mechanism outperforms baseline models on both quantitative and qualitative evaluations. However, based on the human evaluation results, there is still ample room for improvement on providing medical explanations. As the correctness of explanation is still relatively lower than our expectation. On the other hand, without integrating explicit medical knowledge, there exists potential risks of producing unfaithful results. In the future, we plan to incorporate external domain knowledge, e.g., medical knowledge base, into the model to generate enriched and faithful explanations that are not only relevant to suggestions, but also contain correct information.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 110-2221-E-002-128-MY3 and MOST 110-2634-F-002-050-.

References

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592–12601.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online. Association for Computational Linguistics.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Ehud Reiter. 2019. Natural language generation challenges for explainable ai. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.