KNSE: A Knowledge-aware Natural Language Inference Framework for Dialogue Symptom Status Recognition

Wei Chen[1], **Shiqi Wei**[1], **Zhongyu Wei**[1,2]*, **Xuanjing Huang**[3]
[1]School of Data Science, Fudan University, China
[2]Research Institute of Intelligent and Complex Systems, Fudan University, China
[3]School of Computer Science, Fudan University, China
{chenwei18,sqwei19,zywei,xjhuang}@fudan.edu.cn

Abstract

Symptom diagnosis in medical conversations aims to correctly extract both symptom entities and their status from the doctor-patient dialogue. In this paper, we propose a novel framework called KNSE for symptom status recognition (SSR), where the SSR is formulated as a natural language inference (NLI) task. For each mentioned symptom in a dialogue window, we first generate knowledge about the symptom and hypothesis about status of the symptom, to form a *(premise, knowledge, hypothesis)* triplet. The BERT model is then used to encode the triplet, which is further processed by modules including utterance aggregation, self-attention, cross-attention, and GRU to predict the symptom status. Benefiting from the NLI formalization, the proposed framework can encode more informative prior knowledge to better localize and track symptom status, which can effectively improve the performance of symptom status recognition. Preliminary experiments on Chinese medical dialogue datasets show that KNSE outperforms previous competitive baselines and has advantages in cross-disease and cross-symptom scenarios.

1 Introduction

Dialogue symptom diagnosis is an important task in medical dialogue modeling, which is widely used in automatic construction of electronic medical records (EMRs) (Du et al., 2019; Lin et al., 2019; Gu et al., 2020; Zhang et al., 2020) and automatic diagnosis systems (Wei et al., 2018; Xu et al., 2019; Zhong et al., 2022; Chen et al., 2023b). Dialogue symptom diagnosis can be defined as two subtasks: symptom entity recognition (SER) and symptom status recognition (SSR). The former aims to identify symptom entities from doctor-patient dialogues, while the latter aims to further clarify the relationship between identified symptoms and patients.

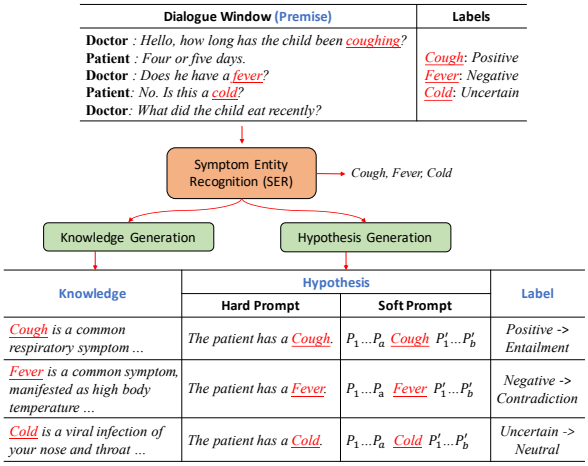

Figure 1: Data variations for dialogue symptom diagnosis in our KNSE framework. The Premise, Knowledge, Hypothesis, and Label (in blue font) are the converted data for NLI training.

Diagnosing symptoms from doctor-patient dialogues is challenging due to the common non-standard oral expressions in dialogues (Lin et al., 2019) and the fact that status information of a single symptom entity may be scattered across multiple utterances (Lou et al., 2021). Existing studies try to solve these issues by sequence labeling (Finley et al., 2018; Zhao et al., 2021), generative modeling (Du et al., 2019; He et al., 2021), semantic integration (Lin et al., 2019), context modeling (Zeng et al., 2022; Hu et al., 2022; Dai et al., 2022), etc. However, previous studies have the following limitations: 1) in most systems, symptoms and their status are jointly predicted, which makes it difficult to generalize to unseen symptoms; 2) symptom-related knowledge is rarely considered.

In this paper, we regard symptom status recognition (SSR) as a natural language inference (NLI) task, and propose a novel framework called KNSE for SSR task, as shown in Figure 1. For each symptom mentioned in a given dialogue window, we first generate knowledge about the symptom and

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10278–10286

hypothesis about the status of symptom, and construct a triplet of the form (premise, knowledge, hypothesis), where the premise is text of the dialog window. After encoding the concatenated text of the triples using BERT, KNSE further utilizes utterance aggregation, self-attention and cross-attention modules to extract knowledge and hypothesis related matching features, and adopts GRU module to track these matching features to generate symptom status features. Preliminary experiments on CMDD and its variant datasets demonstrate that KNSE outperforms previous competing baselines and has advantages in cross-disease and cross-symptom scenarios.

2 Related Work

Medical Dialogue Dataset Extracting structured information from medical dialogues has received increasing attention, and various human-annotated medical dialogue datasets are constructed to support this research. Du et al. annotated a Chinese dialogue corpus, where 186 symptoms are defined and each symptom is associated with a three-value status (Positive, Negative, Unknown). Lin et al. constructed a chinese medical dialogue datasets called CMDD, where each symptom is annotated in the dialogue with BIO format, with its corresponding status and standardized name. Zhang et al. created CHYU dataset containing 1,120 dialogues, where more medical entity categories and their status are annotated, including symptoms, tests, operations, etc. Chen et al. created IMCS-21, a more extensive manually annotated medical conversation dataset, including medical entities and status, dialog intentions, and medical reports.

Medical Information Extraction Several methods have been proposed for extracting structured information from medical dialogues. Finley et al. first introduced a linear processing pipeline system to automatically extract EMRs from oral medical dialogues. Du et al. developed a span-attribute tagging model and a Seq2Seq model to infer symptom entities and their status from medical dialogues. Lin et al. utilized attention mechanism and symptom graph to integrate semantic information across sentences. Zeng et al.; Hu et al.; Hu et al. proposed context modeling approaches to learn the joint representation of context and symptoms. The closest study to our method for dialogue symptom diagnosis is the machine reading comprehension (MRC) framework proposed by Zhao et al., in which the author adopted a similar sentence pair classification method to identify the status of each symptom. Our method extends this approach by introducing additional symptom knowledge and adopting a more complex network structure compared to a simple BERT (Devlin et al., 2019) encoder.

3 Method

3.1 Task Formulation

Given a dialogue window $X = \{U_1, U_2, ..., U_n\}$, where U_i represents a patient (or doctor) utterance, and n is the window size. The set of all symptoms is denote as T, and the set of all symptom status is denoted as S.

For each dialogue window X, dialogue symptom diagnosis task aims to extracting all mentioned symptoms and corresponding status, i.e., $y = \{(t_i, s_i)\}_{i=1}^{k}$, where $t_i \in T$, $s_i \in S$.

3.2 Symptom Entity Recognition

Since symptom entity recognition (SER) is not the focus of this study, we simply utilize the BERT-CRF (Devlin et al., 2019) model to extract the text span of symptom entities, and adopt a SVM (Hearst et al., 1998) classifier to standardize the identified symptom entities. The BERT-CRF and SVM models achieved 95% F1 scores and 98% accuracy on the test set respectively, indicating the SER task is relatively simple and will not bring too much noise to the next step, which is consistent with the conclusion in (Zhao et al., 2021). Therefore, improving the performance of symptom status recognition (SSR) is currently the most pressing obstacle. We will focus on the major contributions of the proposed KNSE framework in subsequent chapters.

3.3 KNSE Framework

3.3.1 Symptom Hypothesis Generation

We regard SSR task as a natural language inference (NLI) (Chen et al., 2017) problem, where the concatenated text of dialogue window is regarded as the *premise*, and the statement of symptom status is regarded as the *hypothesis*. We set the hypothesis that **the patient has the given symptom** (Figure 1), and consider two ways to generate the hypothesis, i.e., hard prompt and soft prompt.

Hard Prompt The hard prompt based template is set to "*The patient has a {}.*", where the content in curly brackets is filled with the given symptom.

Soft Prompt The soft prompt template is set to "$P_1 ... P_a$ {} $P_1^{'} ... P_b^{'}$", where a and b prompt tokens are added before and after the given symptom respectively, and the embeddings of these prompt tokens are trainable. Note that a and b are hyperparameters.

3.3.2 Symptom Knowledge Generation

Knowledge about symptoms may help better localize positive symptoms, we utilize large language models (LLMs) to obtain symptom knowledge for convenience. We first construct the following question template, "*Please briefly describe the {} symptom*", where the content in curly brackets will be filled with a specific symptom. Then we feed the question to ChatGPT (OpenAI, 2022), and cache the answer as the symptom knowledge. It is worth noting that the acquisition of symptom knowledge does not rely on LLMs, as it can be obtained through other sources, such as relevant entries on Wikipedia, professional medical websites, etc.

3.3.3 Natural Language Inference

For a given triplet (P, H, K), i.e., the premise P, the generated hypothesis H and knowledge K, the natural language inference (NLI) module aims to predict whether the hypothesis is true (entailment), false (contradiction), or undetermined (neutral), given the premise and the knowledge. Inspired by recent studies in multi-turn dialogue modeling (Zhang et al., 2018; Chen et al., 2022b,a), we propose a similar matching model, which consists of modules including encoder, utterance aggregation, self attention, cross attention and GRU (Dey and Salem, 2017).

Encoder We first adopt BERT (Devlin et al., 2019) to encode the triplet. We concatenate the triplet with special token [SEP] and feed them into BERT to obtain their respective hidden vectors H_P, H_H and H_K, whose dimension is the corresponding length multiplied by the hidden vector dimension d:

$$H_P, H_H, H_K = \text{Encoder}(P, H, K)$$

Utterance Aggregation The hidden vector of hypothesis H_H is then treated as a query to apply an attention mechanism to the hidden vector of premise H_P, which can aggregate the hypothesis related information from each utterance into a single vector. The process of aggregation works as:

$$\begin{aligned} a[i][j] &= \max_k(H_P[i][j]WH_H[k]^T) \\ p[i] &= \text{softmax}(a[i]) \\ C_{hyp}[i] &= \sum_j p[i][j]H_P[i][j], \end{aligned} \tag{1}$$

where $W \in \mathcal{R}^{d \times d}$ is trainable weights, $[i][j]$ represents the j-th word in the i-th utterance of premise, $[k]$ represents the k-th word in the hypothesis, and $C_{hyp}[i]$ is the aggregated hypothesis-related information of the i-th utterance of the premise. The process will assign high values to the words related to the hypothesis, and thus extract the most relevant information within an utterance.

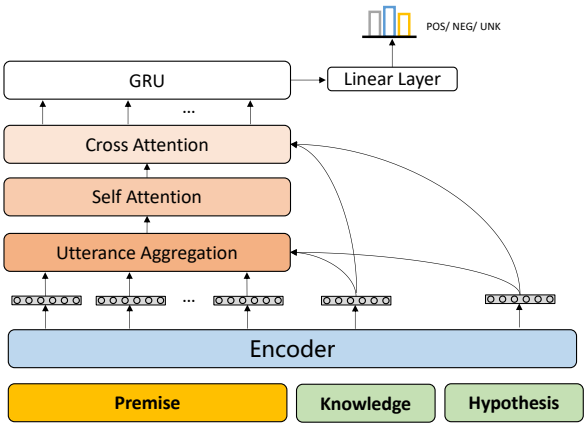

Figure 2: The structure of natural language inference module.

Self Attention We use self-attention to enhance contextual utterance representation as follows:

$$C_{hyp}^{sa} = \text{SA}(C_{hyp}, C_{hyp}, C_{hyp}), \tag{2}$$

where $\text{SA}(\cdot)$ represents a combination of self-attention, residual connection, and layer normalization modules.

Cross Attention We adopt cross-attention to enhance hypothesis information fusion as follows:

$$C_{hyp}^{ca} = \text{CA}(C_{hyp}^{sa}, H_H, H_H), \tag{3}$$

where $\text{CA}(\cdot)$ represents a combination of cross-attention, residual connection, and layer normalization modules.

GRU We have presented C_{hyp}^{ca}, which is regarded as the matching features between hypothesis and premise. In the same way, we can obtain

C^{ca}_{knw}, i.e., the matching feature between knowledge and premise, by using knowledge as the key in utterance aggregation module. Afterwards, we concatenate C^{ca}_{hyp} and C^{ca}_{knw} and update the matching features using a bidirectional GRU (Dey and Salem, 2017) as follows:

$$\hat{h} = \text{GRU}([C^{ca}_P; C^{ca}_K]), \tag{4}$$

where $\hat{h}$ is our final representation of symptom status, which is recursively updated from sentence-level matching features. We employ a linear layer to map $\hat{h}$ to the probability distribution of symptom status and train KNSE with the cross entropy objective.

4 Experiments

4.1 Dataset

We conduct extensive experiments on Chinese Medical Diagnosis Dataset (CMDD) (Lin et al., 2019) to demonstrate the effectiveness of our framework on dialogue symptom diagnosis task. We convert CMDD to a sliding window format on all dialogues, with a windows size of 5, following the settings of previous studies (Hu et al., 2022).

The CMDD dataset contains 2,067 dialogues and 87,005 windows, including 52,952/16,935/17,118 dialogue windows in train/develop/test sets, respectively, covering 160 symptoms, and 3 possible status (Positive, Negative, and Unknown) for each symptom.

4.2 Baselines

Five baseline models are used for comparison, including Plain-Classifier (Zhou et al., 2016), BERT-MTL (Devlin et al., 2019), MIE-multi (Zhang et al., 2020), MRC (Zhao et al., 2021) and CANE (Hu et al., 2022). The Plain-Classifier, MIE-multi and CANE models regard the task as multi-label classification problem and jointly predict the symptoms and their status; while BERT-MTL and MRC models adopt a pipeline approach, i.e., first predict the mentioned symptoms, and then predict the status of each symptom.

We also compare several variants of KNSE. Encoder only represents after encoding the triplet, the hidden vector of [CLS] is directly used to predict the symptom status. Hard prompt indicates using fixed, non-trainable prompt to generate the hypothesis. KNSE w/o. knowledge means not using knowledge.

4.3 Experimental Settings

We use BERT (Devlin et al., 2019) as our encoder. We set the maximum length of each utterance to 50 to ensure that the length of the dialogue window does not exceed 256, and we set the maximum length of symptom knowledge to 64. We use the AdamW (Loshchilov and Hutter, 2017) as the optimizer, and its betas, weight decay and other parameters follow the settings in RoBERTa (Liu et al., 2019). We set the batch size as 64, the learning rate as 1e-5. We adopt soft prompt and the hyperparameter a and b are set to 10 and 5 respectively. We train a total of 20 epochs and choose the model that performs best on the develop set.

4.4 Evaluation Metrics

We report the micro-averaged Precision, Recall and F1 score in the multi-label classification (Zhang and Zhou, 2013) scenario to measure the overall performance of the system, where the label space is $|T|*|S|$, and only if both the symptom and its status are correct can they be considered as real positive cases. Both the window-level and the dialogue-level results are reported in the paper, see details in (Hu et al., 2022).

4.5 Main Results

Table 1 shows the experimental results on the CMDD dataset. It can be seen that KNSE outperforms all baselines in both window-level and dialogue-level evaluation metrics. This illustrates the effectiveness of the KNSE framework. It is worth noting that since each symptom is identified independently, KNSE does not take advantage of the co-occurrence of some symptoms and their status like MIE-multi and CANE. The results in window-level are relatively higher than the results in dialogue-level. This is because the latter is stricter than the former, which has been verified in previous studies (Hu et al., 2022).

It is more interesting to analyze the effectiveness of KNSE components. From the results of the ablation experiments: The variant KNSE encoder only underperforms, suggesting that the inductive bias introduced by these additional modules in addition to the encoder are effective for symptom status representation learning; Using hard prompt tokens instead of soft prompt tokens will slightly reduce the model performance, we guess that it may be because tunable soft prompts can help the model learn to pay attention to important words in the dia-

Model	Window-Level			Dialogue-Level		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Plain-Classifier (Zhou et al., 2016)	79.80	75.90	76.84	67.81	67.81	65.58
BERT-MTL (Devlin et al., 2019)	80.20	77.18	77.25	70.28	70.12	68.21
MIE-multi (Zhang et al., 2020)	81.63	80.45	80.23	74.12	72.38	72.51
MRC (Zhao et al., 2021)	80.05	78.45	79.24	73.56	74.92	74.24
CANE (Hu et al., 2022)	82.54	81.36	81.33	75.78	75.79	75.20
KNSE (Ours)	**84.17**	**82.86**	**83.57**	**77.32**	**76.59**	**76.83**
Ablation Study						
encoder only	82.12	80.74	80.84	74.81	74.59	74.52
hard prompt	83.78	82.52	83.14	76.92	76.21	76.39
w/o. knowledge	83.05	82.40	82.93	75.89	76.03	75.89

Table 1: Experimental results on CMDD dataset.

Model	Dataset	POS	NEG	UNK
MRC(Zhao et al.)	CMDD	87.1	73.8	72.4
	CD-CMDD	81.8	64.3	63.2
	CS-CMDD	74.2	58.7	56.6
KNSE (Ours)	CMDD	89.6	76.5	74.2
	CD-CMDD	86.5	72.7	71.4
	CS-CMDD	83.4	69.8	66.5

Table 2: Experimental results on cross-domain variants of CMDD dataset, where POS, NEG and UNK are the abbreviations of Positive, Negative and Unknown respectively.

logue window; Introducing symptom knowledge is effective, since intuitively, knowledge can help us better identify positive symptoms.

4.6 Cross-domain Scenarios

We further explore the model performance in cross-disease and cross-symptom scenarios. Specifically, we redivide the CMDD dataset, where CMDD-CD and CMDD-CS represent the division of training sets, develop sets and test sets according to disease and symptom, respectively. In CMDD-CD dataset, the diseases in the test set are not seen in the training set, but there may be some overlapping symptoms. In CMDD-CS dataset, all symptoms in the test set in do not appear in the training set.

We assume that the symptoms are already known, and adopt the MRC and KNSE models to predict the status of each symptom. We report F1 score for each category of symptom status in Table 2. The experimental results show that the dataset divided by symptoms is more difficult on the SSR task than the dataset divided by diseases, which is intuitive. Besides, it can be seen that the performance degradation of KNSE on CD-CMDD and CS-CMDD datasets (about 3~8%) is much lower than that of MRC (about 9~16%), suggesting that compared with MRC, KNSE has a stronger ability to recognize the status of symptoms that have not been seen in the training set.

5 Conclusion

In this paper, we investigate the problem of symptom diagnosis in doctor-patient dialogues. We proposed a knowledge-aware framework by formalizing the symptom status recognition problem as a natural language inference task. Our framework is able to encode more informative prior knowledge to better localize and track symptom status, which can effectively improve the performance of symptom status recognition. We develop several competitive baselines for comparison and conduct extensive experiments on the CMDD dataset. The experimental results demonstrate the effectiveness of our framework, especially in cross-disease and cross-symptom scenarios.

Limitations

This study has potential limitations. Firstly, we only test our method on one dataset. We plan to apply our model to more datasets in future versions. Secondly, ablation experiments are not sufficient. We will conduct comprehensive ablation experiments to demonstrate the contribution of different components.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 6217020551) and Science and Technology Commission of Shanghai Municipality Grant (No.21QA1400600).

References

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, et al. 2022a. Dialogved: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864.

Wei Chen, Yeyun Gong, Can Xu, Huang Hu, Bolun Yao, Zhongyu Wei, Zhihao Fan, Xiaowu Hu, Bartuer Zhou, Biao Cheng, et al. 2022b. Contextual fine-to-coarse distillation for coarse-grained response selection in open-domain conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4865–4877.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023a. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Wei Chen, Cheng Zhong, Jiajie Peng, and Zhongyu Wei. 2023b. Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1):btac744.

Jianhua Dai, Chao Jiang, Ruoyao Peng, Daojian Zeng, and Yangding Li. 2022. Chinese medical dialogue information extraction via contrastive multi-utterance inference. *Briefings in Bioinformatics*, 23(4):bbac284.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15.

Qizheng Gu, Cong Nie, Ruixiang Zou, Wei Chen, Chaojun Zheng, Dongqing Zhu, Xiaojun Mao, Zhongyu Wei, and Dong Tian. 2020. Automatic generation of electromyogram diagnosis report. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1645–1650. IEEE.

Yifei He, Yan Li, and Senbao Hou. 2021. Document-aware information extractor for chinese medical dialogue. In *2021 3rd International Workshop on Artificial Intelligence and Education (WAIE)*, pages 65–68. IEEE.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Gangqiang Hu, Shengfei Lyu, Xingyu Wu, Jinlong Li, and Huanhuan Chen. 2022. Contextual-aware information extractor with adaptive objective for chinese medical dialogues. *Transactions on Asian and Low-Resource Language Information Processing*.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. 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 5033–5042, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dongfang Lou, Zhilin Liao, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. Mlbinet: A cross-sentence collective event detection network. *arXiv preprint arXiv:2105.09458*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 201–207.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

Daojian Zeng, Ruoyao Peng, Chao Jiang, Yangding Li, and Jianhua Dai. 2022. Csdm: A context-sensitive deep matching model for medical dialogue information extraction. *Information Sciences*.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

Xiongjun Zhao, Yingjie Cheng, Weiming Xiang, Xiang Wang, Lin Han, Jiandong Shang, and Shaoliang Peng. 2021. A knowledge-aware machine reading comprehension framework for dialogue symptom diagnosis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1185–1190. IEEE.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section Limitations

☑ A2. Did you discuss any potential risks of your work?
section Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 4.1

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 4.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 4.5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.