Automatically Extracting Information in Medical Dialogue: Expert System And Attention for Labelling

Xinshi Wang
Rensselaer Polytechnic Institute
wangx47@rpi.edu

Daniel Tang
University of Luxembourg
xunzhu.tang@uni.lu

Abstract

Medical dialogue information extraction is becoming an increasingly significant problem in modern medical care. It is difficult to extract key information from electronic medical records (EMRs) due to their large numbers. Previously, researchers proposed attention-based models for retrieving features from EMRs, but their limitations were reflected in their inability to recognize different categories in medical dialogues. In this paper, we propose a novel model, Expert System and Attention for Labelling (ESAL). We use mixture of experts and pre-trained BERT to retrieve the semantics of different categories, enabling the model to fuse the differences between them. In our experiment, ESAL was applied to a public dataset and the experimental results indicated that ESAL significantly improved the performance of Medical Information Classification. The code is available at here. [1]

Introduction

Increasingly, hospitals are prioritizing Medical Dialogue Information Extraction (MDIE) due to the adoption of Electronic Health Records (EHR). Using MDIE, detailed medical information can be extracted from doctor-patient conversations. MDIE can be viewed as a multi-label classification problem made up of different classes and their status labels. Specifically, the data set we used in this paper includes symptoms, surgeries, tests, and other information.

Medical dialogue information extraction has received an increasing amount of attention from scholars, and various approaches have been developed. Finley et al. converted Doctor-patient dialogues to electronic medical records, but no solution is proposed. As a result, 186 symptom codes and their corresponding statuses were defined as a new task. By proposing two novel models, Du et al. were able to solve this problem. The first model was a span-attribute tagging model, and the second was a sequence-to-sequence model. Even though they covered a wide range of symptoms in their data set, they didn't consider other critical medical information. As a means of incorporating more medical information, Zhang et al. proposed a novel dataset that includes four main categories, namely symptoms, surgeries, tests, and other information. Furthermore, they predefined several specific items with corresponding statuses. In addition, they proposed a novel method of annotation, the sliding window technique, so that the dialogues included within the document could contain the proper amount of information. Meanwhile, they developed a Medical Information Extractor (MIE) for multi-turn dialogues. A matching mechanism was used to match dialogues between predefined category-item representations and status representations. The utterance's category-item information is exploited to match its most suitable status in a window to aggregate its category-item and corresponding status information.

With the help of mixture of experts (Jacobs et al. (1991);Ma et al. (2018);Zhuang et al. (2020)), we propose a model called Expert System and Attention for Labeling (ESAL) that extracts various representations of dialogue to address the different categories within the dialogue. To get category-specific representations, we first use BERT (Devlin et al. (2018)) to extract contextual representations of the dialogue and feed them to the category-specific BiLSTM (Schuster and Paliwal (1997)) expert. After that, we calculate the attention value between the encoded candidate representation and the encoded dialogue representation in order to obtain the candidates. In a similar manner, we calculate the status using the same attention mechanism.

To summarize, this paper makes the following contributions:

- This paper proposes an expert system attention for labelling model for extracting medical dialogue information. Each specified category can be captured in terms of the utterance representation.
- In this study, we introduce an expert system that effectively strengthens the understanding of doctor-patient dialogue. To facilitate understanding, we also introduce a learnable embedding layer.
- On a widely used medical dialogue dataset, we perform extensive experiments. On window-level evaluation, our model scores 70.00, while on dialogue-level evaluation, it scores 72.17. On the benchmark dataset, it outperforms the state-of-the-art approaches by a significant margin, demonstrating its effectiveness.

[1] https://anonymous.4open.science/r/Expert-System-and-Attention-for-Labelling-8B14/

Related Work

Medical Dialogue Information Extraction

Medical Dialogue Information Extraction has attracted increasing scholar attention due to the growing priority of building Electronic Health Records in hospitals. (Wachter and Goldsmith Xu) The first public doctor-patient dialogue dataset was proposed by Zhang et al.. The dataset was obtained from from a popular Chinese online medical consultation website, Chunyu-Doctor [2]. They converted the dataset into dialogue forms and labeled the utterances with 4 predefined categories, symptom, test, surgeries, and other information, as well as their corresponding status. They formulated the information extraction task as a multi-label classification problem and proposed a novel deep matching model called MIE. MIE first encoded the dialogue and the predefined candidates through a LSTM network (Hochreiter and Schmidhuber (1997)). Then, MIE treats both the encoded category-item and status representation as queries to calculate the attention values toward the original utterances in the deep matching module. After obtaining the desired attention values, the category-item information of the utterance is exploited to match its status in a aggregate module. Since this dataset currently is the largest dataset with well-designed labels, our model also uses this dataset as a benchmark.

Expert System

Expert system is composed of many separate networks, each of which learns to handle a subset of the complete set of training cases (Jacobs et al. (1991)). The ensemble of individual experts has proven to be able to improve performance (Caruana (1993);Hinton et al. (2015)). Then, Eigen, Ranzato, and Sutskever and Shazeer et al. made the mixtures of experts system into a basic building block. Expert system has been applied to various fields, such as multi-domain fake news detection (Nan et al. (2021)) and recommendation systems (Ma et al. (2018)).

Approach

In this section, we will elaborate the architecture of ESAL. The architecture is shown in figure 1. ESAL is composed of 4 different stages: 1). Embedding layer 2). Expert information extraction Layer 3). Self-Attention labelling Layer 4). Output Layer.

Embedding layer

For each doctor-patient dialogue, we first tokenize its content with Bert Tokenizer (Devlin et al. (2018)). We then add special tokens for classification (*i.e.*,$[cls]$) as well as separation (*i.e.*,$[sep]$) to obtain a list of tokens $X = [[cls], token_1, token_2, ..., token_n, [sep]]$. We then feed the list of tokens into BERT to obtain word embedding $V = BERT(X)$. Similarly, we perform the same operation on the candidates for matching to obtain the embedding $U = BERT(Q)$ for query Q.

[2] https://www.chunyuyisheng.com

Expert information extraction layer

With the advantage of Mixture-of-Experts, we employ multiple experts (i.e., network) to extract category-specific and status-specific representations of the utterance. We select the bidirectional long short-term memory network (BiLSTM) (Schuster and Paliwal (1997)) with attention mechanism (Vaswani et al. (2017)) as our individual network. BiLSTM has been widely used to extract contextual text features.

The equation below denotes the process for encoding each dialogue, where i indicates the output from the i^{th} category expert.

$$H_C[i], H_S = BiLSTM(V), BiLSTM(U) \quad (1)$$

$H_C[i]$ consists of the contextual representation of embedding V specific to category i.

For candidates in the form of $\{Category : Item - Status\}$, we denote the Cartesian product between item and status given the category as Q_C. We then feed Q_c to the corresponding category expert to obtain the embedding and apply self-attention to the embedding to obtain a single vector C_C that compresses the information of the entire sequence in a weighted way. The procedures above can be described with the following equation, where $\sigma = \frac{exp(i)}{\sum_{i=1}^{n} \exp(i)}$ denotes the softmax operation.

$$\begin{aligned} U_C[i], U_S &= BiLSTM(Q_C), BiLSTM(Q_S) \\ A_C[i], A_S &= WU_C[i] + b, WU_S + b \\ P_C[i], P_S &= \sigma(A_C[i]), \sigma(A_S) \\ C_C, C_S &= \sum_{i}^{n}(P_C[i]U_C[i]), \sum_{i}^{n}(P_S U_S) \end{aligned} \quad (2)$$

Self-Attention Labeling Layer

We employ self-attention to capture the most relevant candidate features from the utterance representation, where the candidate representation is treated as a query to calculate the attention value Q_C towards the category specific utterance representation. Similarly, the candidate status representation is treated as another query to calculate the attention value toward the original utterances to obtain the most relevant status features from utterance representation.

$$\begin{aligned} &P_C[i], P_S[i] = \sigma(C_C[i]H_C[i]), \sigma(C_S[i]H_S) \\ &Q_C[i], Q_S[i] = \sum_{j}^{n}(P_C[i,j]H_C[i,j]), \sum_{j}^{n}(P_S[i,j]H_S[i,j]) \end{aligned} \quad (3)$$

To assign the correct candidates to each dialogue window, we need to match every $Q_C[i]$ with the corresponding $Q_S[i]$. The category-item pair information and the status information does not necessarily appear in the same dialogue window, so we need to take the interactions between utterances among multiple dialogue windows into consideration. The process can be described with following equation, where *concat* denotes the concatenate operation:

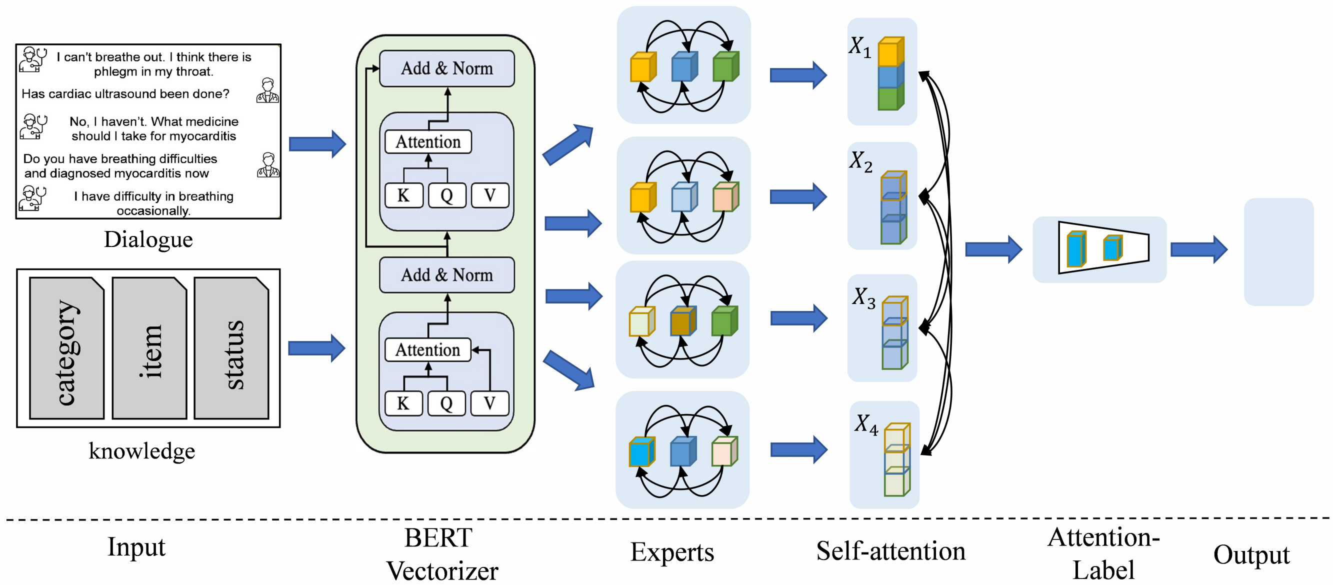

Figure 1: Model Architecture

$$
\begin{aligned}
C[i] &= \sigma(Q_C[i] W Q_S[i]) \\
\hat{Q}_S[i] &= \sum_{j=1}^{n} (C[i,j] Q_S[i]) \\
F[i] &= concat(C[i], \hat{Q}_S[i])
\end{aligned}
\tag{4}
$$

The output of the equation above gives the candidate information assigned to the doctor patient dialogue U.

Output Layer

We use the output from the Self-Attention Labeling Layer, (*i.e.*F[i]) to generate the output for our model. Using a feed-forward network, we can project the utterance's representation F[i] onto the 355 corresponding candidate positions, and then apply a softmax function to select the final prediction label. The process can be described with the following equations, where f denotes the feedforward network and $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ denotes the sigmoid function:

$$
\begin{aligned}
s[i] &= f(F[i]) \\
y &= h_\theta(max(s[i]))
\end{aligned}
\tag{5}
$$

Loss Function

We adopt the cross entropy loss as our loss function. The function is defined as the following equation:

$$
L = \frac{1}{I \times J} \sum_i \sum_j -y_j^i \ln(\hat{y_j^i}) + (1 - y_j^i) \ln(1 - \hat{y_j^i})
\tag{6}
$$

The y_j^i denotes label of j^{th} candidate from the i^{th} label.I denotes the number of samples and J denotes the number of candidates. $\hat{y_j^i}$ denotes the ground truth value of label y_j^i.

Experiments

In this section, we will conduct experiments on the MIE dataset (Zhang et al. (2020)). We will firstly describe the dataset and evaluation metrics. Then we will present results with a case study of the experiment.

Dataset Description

We evaluate our model on a public dataset MIE (Zhang et al. (2020)). An example of a dialogue window is illustrated in Table 1 below.

Table 1: Dialogue Window

Role	Dialogue
Patient:	Doctor, is it premature beat?
Doctor:	Yes, Do you feel short breath?
Patient:	No. Should I do radio frequency ablation?
Doctor:	You should. Any discomfort in chest?
Patient:	I always have bouts of pain.

The annotation of the sliding window dialogue is composed of several labels in the form of $\{Category : Item - Status\}$. An example of the annotated label is given in table 2.

$Category$ contains four main categories (Symptom, Surgery, Test, and Other Info). $Item$ stands for the frequent items with respect to each category. There are 45, 4, 16, and 6 items, respectively. The $status$ is defined as doctor-pos, doctor-neg, patient-pos, patient-neg, or unknown. There are in total 1,120 dialogues, resulting in 18,212 windows. The data is divided into train/develop/test sets of size 800/160/160 for dialogues and 12,931/2,587/2,694 for windows respectively. In total, there are 46,151 annotated labels,

Models \ Levels	Window-Level									Dialogue-Level								
Model	Category			Item			Full			Category			Item			Full		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Plain-Classifier	67.21	63.78	64.92	60.89	49.20	53.81	53.13	49.46	50.69	93.57	89.49	90.96	83.42	73.76	77.29	61.34	52.65	56.08
MIE-Classifier-Single	80.51	76.39	77.53	76.58	64.63	68.30	8.20	61.60	62.87	97.14	91.82	93.23	91.77	75.36	80.96	71.86	56.67	61.78
MIE-Classifier-Multi	80.72	77.76	78.33	76.84	68.07	70.35	67.87	64.71	64.57	96.61	92.86	93.45	90.68	82.41	84.65	68.86	62.50	63.99
MIE-Single	78.62	73.55	74.92	76.67	65.51	68.88	69.40	64.47	65.18	96.93	90.16	92.01	94.27	79.81	84.72	75.37	63.17	67.27
MIE-Multi	80.42	76.23	77.77	77.21	66.04	69.75	70.24	64.96	66.40	98.86	91.52	92.69	95.31	82.53	86.83	76.83	64.07	69.28
ESAL	92.42	89.66	90.26	89.46	83.38	84.85	72.08	70.93	70.00	96.51	95.05	94.74	92.52	90.88	90.50	73.68	73.10	72.17

averaging 2.53 labels in each window, 41.21 labels in each dialogue.

Table 2: Dialogue Annotation

Category	Item(Status)
Symptom:	Premature beat (doctor-pos)
Test:	Electrocardiogram (patient-pos)
Symptom:	Cardiopalmus (patient-neg)
Symptom:	Dyspnea (patient-neg)
Surgery:	Radiofrequency ablation (doctor-pos)
Symptom:	Chest pain (patient-pos)

Evaluation Metrics

We use the precision, recall, and F1 score to evaluate our results. We also follow the evaluation metrics Zhang et al. employed to further analyze the model behavior.

Window-level: The results of each segmented window are evaluated and reported by the micro-average of all windows in the test set.

Dialogue-level: We merge the results with the same category and item of all the windows in the same dialogue. For category-item pair with multiple status assigned, we replace the unknown status with any other status occurred and replace the negative status with positive status if occurred.

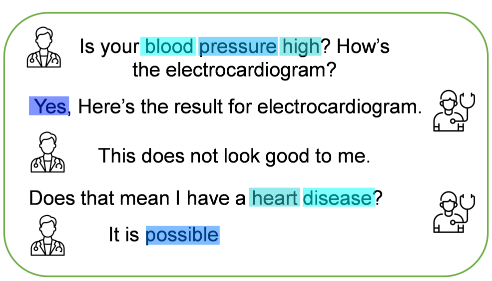

Figure 2: Symptom Expert Attention Heat Map

Main Results

The experimental results are show in Table 1. From the table, we can make the following observations.

On both the window-level and dialogue level evaluation, Our model outperforms other models in most metrics. On window-level Full evaluation, our method has the performance improved by 5.4% compared to the MIE-multi in F1 score. On dialogue-level full evaluation, our method achieves an improvement of 4.17% in F1 score. These results demonstrate that the ESAL is performing better compared to the previous state-of-the-art model.

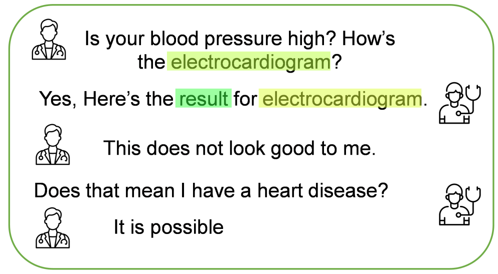

Figure 3: Test Exper Attention Heat Map

On Window-level evaluation, our model outperforms other models significantly in Category and Item evaluation. For Category evaluation, Our model has a performance improvement of 16.90% in F1 score. For Item evaluation, our model has a improvement of 21.65% in F1 score. Also, the improvement on Precision and Recall are significant. These results demonstrate that ESAL is able to extract a better domain-specific representation of the utterance.

Case Analysis

In this section, we perform an analysis on a specific case to verify the effectiveness of the mixture of experts. We did a data visualization on the attention value from Symptom expert and Test expert on the same utterance in graph 2 and graph 3. Brighter Color suggests a higher attention value. The label for the utterance is {Symptom: high blood pressure- doctor-pos, Symptom: heart disease-unkown, Test: electrocardiogram-pos}. As we can see from graph 2, the highest attention value comes from "Yes", which suggests that our Symptom Expert captures the status information correctly. It also captures the status information for heart disease. Similarly, the test expert has captured the item and status. These two outputs gave category specific attention value on different items, thus proved the effectiveness of our model in capturing category-specific representations.

Conclusion

In this paper, we proposes an expert system attention for labelling model for extracting medical dialogue information, which utilizes two techniques: mixture of experts and an embedding layer. Experimental results on a public available dataset have shown that ESAL has the ability to capture category specific utterance representations and has better understanding of doctor-patient dialogue compared to previous models. For future work, We plan to investigate the interaction between doctor and patient to handle the pronoun ambiguity.

References

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, 41–48. Citeseer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, N.; Chen, K.; Kannan, A.; Tran, L.; Chen, Y.; and Shafran, I. 2019. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Finley, G.; Edwards, E.; Robinson, A.; Brenndoerfer, M.; Sadoughi, N.; Fone, J.; Axtmann, N.; Miller, M.; and Suendermann-Oeft, D. 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*, 11–15.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2(7).

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation* 3(1):79–87.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11):2673–2681.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Wachter, R., and Goldsmith, J. 2018. To combat physician burnout and improve care, fix the electronic health record. *Harvard Business Review*.

Xu, R. 2018. The burnout crisis in american medicine. *The atlantic*.

Zhang, Y.; Jiang, Z.; Zhang, T.; Liu, S.; Cao, J.; Liu, K.; Liu, S.; and Zhao, J. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6460–6469.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1):43–76.

References

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, 41–48. Citeseer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, N.; Chen, K.; Kannan, A.; Tran, L.; Chen, Y.; and Shafran, I. 2019. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Finley, G.; Edwards, E.; Robinson, A.; Brenndoerfer, M.; Sadoughi, N.; Fone, J.; Axtmann, N.; Miller, M.; and Suendermann-Oeft, D. 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*, 11–15.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2(7).

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation* 3(1):79–87.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11):2673–2681.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Wachter, R., and Goldsmith, J. 2018. To combat physician burnout and improve care, fix the electronic health record. *Harvard Business Review*.

Xu, R. 2018. The burnout crisis in american medicine. *The atlantic*.

Zhang, Y.; Jiang, Z.; Zhang, T.; Liu, S.; Cao, J.; Liu, K.; Liu, S.; and Zhao, J. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6460–6469.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1):43–76.