Summarizing Medical Conversations via Identifying Important Utterances

Yan Song[♠♡*], **Yuanhe Tian**[♥*], **Nan Wang**[♣], **Fei Xia**[♥]
[♠]The Chinese University of Hong Kong (Shenzhen)
[♡]Shenzhen Research Institute of Big Data
[♥]University of Washington [♣]Hunan University
[♠]songyan@cuhk.edu.cn [♥]{yhtian, fxia}@uw.edu
[♣]ncwang@hnu.edu.cn

Abstract

Summarization is an important natural language processing (NLP) task in identifying key information from text. For conversations, the summarization systems need to extract salient contents from spontaneous utterances by multiple speakers. In a special task-oriented scenario, namely medical conversations between patients and doctors, the symptoms, diagnoses, and treatments could be highly important because the nature of such conversation is to find a medical solution to the problem proposed by the patients. Especially consider that current online medical platforms provide millions of public available conversations between real patients and doctors, where the patients propose their medical problems and the registered doctors offer diagnosis and treatment, a conversation in most cases could be too long and the key information is hard to be located. Therefore, summarizations to the patients' problems and the doctors' treatments in the conversations can be highly useful, in terms of helping other patients with similar problems have a precise reference for potential medical solutions. In this paper, we focus on medical conversation summarization, using a dataset of medical conversations and corresponding summaries which were crawled from a well-known online healthcare service provider in China. We propose a hierarchical encoder-tagger model (HET) to generate summaries by identifying important utterances (with respect to problem proposing and solving) in the conversations. For the particular dataset used in this study, we show that high-quality summaries can be generated by extracting two types of utterances, namely, problem statements and treatment recommendations. Experimental results demonstrate that HET outperforms strong baselines and models from previous studies, and adding conversation-related features can further improve system performance.[1]

1 Introduction

Applying natural language processing (NLP) techniques to the medical field is a prevailing trend nowadays and has great potential in many applications, such as key information extraction in medical literature (Kim et al., 2011; Dernoncourt et al., 2017; Ševa et al., 2018), risk factor identification in electronic health records (Chang et al., 2015; Cormack et al., 2015; Cheng et al., 2016), and medical question answering (Pampari et al., 2018; Tian et al., 2019). As the demand for healthcare services increases greatly in the past decades,[2] it is urgent to improve the quality and efficiency of healthcare, reduce workload and mental stress of health providers and increase patient satisfaction. Recently, Internet-based healthcare platforms such as online doctor systems and doctor-patient cyber communities have been increasingly used by patients and health professionals with the hope that they would alleviate the ever-increasing demands for healthcare services and reduce the inaccessibility of services caused by geographical and socio-economic barriers. In such platforms, a patient can start a conversation to a registered doctor by typing their medical problems and then the doctor may ask the patients to specify his/her problem (e.g., symptoms, treatment has been taken, etc.). Since the conversation is asynchronous, it is possible that one speaker (either the

[*]Equal contribution.

[1]Our code, models, and the dataset are released at https://github.com/cuhksz-nlp/HET-MC.

[2]E.g., in China, the number of outpatient visits exceeded 7 billions and inpatient visits over 200 millions in a recent year.

Proceedings of the 28th International Conference on Computational Linguistics, pages 717–729
Barcelona, Spain (Online), December 8-13, 2020

	Role	Utterance	Translation	Tag
Conv.	P	胆碱能性荨麻疹怎么治疗	How to treat cholinergic urticaria and measles	PD
	D	这种情况多长时间了？用什么治疗过？	How long has this condition last? What treatment have you used?	OT
	P	好长时间了，之前治疗过，中西药都吃过就没治好	It has been a long time. I have taken both Chinese and Western medicine, but it is not working.	OT
	D	主要是避免诱因。胆碱能性荨麻疹要保持身体凉爽、避免出汗、避免精神紧张、进食热饮或酒精饮料等。	You need to avoid triggers of cholinergic urticaria. Keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc.	DT
	P	那怎么样能根治呢	How can it be cured?	OT
	D	目前医疗上，没有明确的根治方法。	At present, there is no clear cure for this disease.	OT
	D	内服药物之外还可以中药外洗.这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克，把这些药一起煎了进行外洗，一天二次。	In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day.	DT
SUM1		胆碱能性荨麻疹怎么治疗	How to treat cholinergic urticaria and measles	
SUM2	**A**	胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。内服药物之外还可以通过中药进行外洗。这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克,把这些药一起煎了进行外洗，一天二次。	You need to avoid triggers of cholinergic urticaria. keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc. In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day.	
	B	口服脱敏药物，同时避免诱因。胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。	You need to take desensitization drugs and avoid triggers. You should keep the body cool, avoid sweating, mental stress, hot drink, alcoholic beverages, etc.	

Figure 1: An example of a conversation and its different types of summaries. *P* and *D* stand for speaker roles, i.e., patient and doctor, and *PD*, *DT*, and *OT* in the last column refer to the utterance tags for problem description, diagnosis or treatment, and others, respectively. *SUM1* is a summary of the medical problem from the patient; *SUM2* is a summary of the diagnosis and treatment from the doctor. The English translation is not part of the corpus, which is added as a reference.

patient or the doctor) may type multiple lines (utterances) before the other speaker responds. Through this process, all key information regarding to a medical problem, as well as its diagnosis and medical recommendations, are recorded in the entire conversation. Once the platforms make all such conversations publicly available, other patients with similar medical problems can search relevant conversations and find potentially helpful solutions. However, when a conversation is too long or the key information is scattered in it, one could hardly find the essential contents or misread them in many cases. As a result, the summarization of the conversation, especially for problem statement and treatment recommendations, is an important task to help new patients locate useful information to address their medical concerns.

Due to the nature of medical conversation, i.e., a task that seeks solutions to provide medical recommendations for particular health problems, it is possible to perform the task by identifying important utterances in such conversations. In this study, *important utterances* refer to the utterances that contain key information for the medical problem or for the treatment. Therefore, our focus is different from existing studies on utterances in conversations, where they pay more attention to assessment of utterances with respect to the functionalities of utterances in the conversations, such as analyzing automatically generated utterances regarding their suitability within particular conversational contexts (Inaba and Takahashi, 2016; Lison and Bibauw, 2017), evaluation of human conversational performance on readability, sensibility, and social involvement (Dascalu et al., 2010), and identification of segments of utterance that are produced with more emphases for certain interactional purposes (Takeuchi et al., 2007). Little research has been done to identify important utterances that contribute to a specific outcome of a conversation, which in this study refers to the content about patient's problem and recommendation treatment in the conversation.

To conduct the medical conversation summarization task, in this paper, we propose a new benchmark

Data	Total #		Avg. # per Case						Avg. Length		
	Case	Utt.	Utt.	PD	DT	OT	P	D	Utt.	SUM1	SUM2
All	44,983	855,403	19.0	1.3	4.5	13.3	9.7	9.3	16.4	22.8	113.0
Train	35,987	684,611	19.0	1.3	4.5	13.3	9.8	9.3	16.4	22.9	112.8
Test	8,996	170,792	19.0	1.3	4.5	13.2	9.7	9.3	16.5	22.8	114.0

(a) Statistics of the cleaned corpus. *"P"* and *"D"* are the number of utterances by patients and doctors, respectively. SUM2 is the concatenation of SUM2-A and SUM2-B. *Avg. length* is the average number of Chinese characters in an utterance, SUM1, or SUM2.

	A Only	BOTH
Train	33,306 (92.5%)	2,681 (7.5%)
Test	8,299 (92.3%)	697 (7.7%)

(b) The number (percentage) of conversations with SUM2-A only and with both SUM2-A and SUM2-B.

Table 1: The statistics of the corpus. Table 1a illustrates the overall statistics, and Table 1b reports the number of conversations with SUM2-A only or with both SUM2-A and SUM2-B.

dataset in Chinese, which has over 40K cases covering nearly 2K disease types. In each case, there is a medical conversation between a patient and a doctor, and two summaries: one for problem statement and the other for treatment recommendations. Figure 1 shows an example conversation with the two types summaries: *"SUM1"* for problem statement and *"SUM2"* for treatment recommendations. SUM2 has two types, i.e., type A and B, which will be explained in the next section.

Besides, we propose a *hierarchical encoder-tagger* (HET) model for extractive summarization to tag each utterance in a medical conversation with regard to whether an utterance is a problem statement or a treatment recommendation. We further enhance the model with end-to-end memory networks (Sukhbaatar et al., 2015) to incorporate the information in relevant utterances in the conversation. We use BERT (Devlin et al., 2019) as the token-level encoder and try several utterance-level encoders and taggers. Experimental results show that HET outperforms strong baselines as well as models from previous studies on this dataset. Analyses are also conducted to better understand our findings from the results.

2 A Corpus of Medical Conversations

Medical conversation is a type of task-oriented conversation. Different from ordinary conversations in which topics are often fluid, in task-specific conversations, participants interact to accomplish a projected set of goals and sub-goals (Litman and Allen, 1987; Drew and Heritage, 1992). Specifically, for conversations in the medical domain from online medical platforms, the projected goal is for the doctor to diagnose and offer treatment recommendation for the patient's problem (Drew and Heritage, 1992; Robinson, 2012; Wang et al., 2020). Particularly in China, many platforms make such medical conversations publicly available so that new patients with similar problem can search relevant conversations and find helpful information from them. Therefore, summarization of the patient's problem and doctor's recommendations in a conversation could be highly important because such summaries can help the new patients locate the key information, especially when a conversation is too long. To conduct such summarization, a straightforward solution is to identity the important utterances that contain key information for problem statements or treatment recommendations. However, limited corpus can be found to train such summarization model, especially for Chinese. Therefore, we develop a corpus in Chinese for medical conversation summarization and illustrate the details in the following text.

The Raw Data The original data are crawled from one of the most well-known online health provider platforms[3] in China, under a section called "*Frequently Inquired Health Problems*."[4] In these conversations, patients consult registered doctors[5] about some health problems; doctors help them to determine the nature of the problems, provide treatment recommendations, and/or advise them to seek further medical attention from other health facilities. Instead of isolated question-answer segments or part of the conversations, this data contain full conversations between patients and doctors, covering the entire interaction process. In addition to dialogues, each conversation contains meta information such as the type of disease and the corresponding hospital department, as well as the speakership of the utterances in conversation.[6] Many

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

[4] https://www.chunyuyisheng.com/pc/qalist/

[5] Normally, these doctors are manually verified by the website.

[6] While the patient is allowed to interact with the doctor using typed messages, audio messages, uploaded photos, or emojis,

(but not all) conversations include a summary added by doctors after the conversations are conducted. The summary has two parts: *SUM1* describes the medical problem that the patient has; *SUM2* summarizes the doctor's diagnosis or treatment recommendations. *SUM2* is of two types: Type *A* (we denote it as *SUM2-A*) is the concatenation of a few utterances in the conversation, whereas Type *B* (we denote it as *SUM2-B*) is a more concise summary written by the doctor and may contain text that does not appear in the conversation. In all, we crawled 109,850 conversations from 23 hospital departments or sub-divisions, and the conversations cover 1,839 disease types, which forms our raw corpus. Among them, only half of them contains both SUM1 and SUM2. This again emphasizes the necessity of this summarization task, because if we can automatically generate the missing summaries for problem statement and treatment recommendations, new patients may have more references when they search conversations that are relevant to their problem.

Data Processing To facilitate the task of conversation summarization, we process the raw corpus by only reserving the conversations that have both SUM1 and SUM2, and further clean the resulted data by removing duplicates and those conversations containing only one utterance. The cleaned data contain both input and output for the summarization task. Particularly, SUM1 and SUM2-A are the concatenation of selected utterances in the conversation that provide key information for problem statement and treatment recommendations. Therefore, the important utterances identified in a conversation are those likely to appear in the summary. In detail, following Nallapati et al. (2017) and Chen and Bansal (2018), we use ROUGE scores to measure the overlap between an utterance and a summary, and label the utterances accordingly; that is, we break the summary into segments,[7] and then for each segment, find the closest utterance in the conversation according to ROUGE-1 score. If the score is greater than a threshold, we label that utterance as "PD" if the summary is *SUM1* and "DT" if it is *SUM2*. For all other utterances, we label them with "OT". We call those resulting "PD", "DT", and "OT" as *silver-standard labels*.

Data Properties Table 1 shows the statistics of the processed dataset, where Table 1a reports the overall statistics of all data and the train/test splits (we use 80% for training and 20% for testing), and Table 1b illustrates the number of conversations where SUM2 is SUM2-A only or has both SUM2-A and SUM2-B. A few points are worth mentioning. First, on average, each dialogue includes 19.0 UTTERANCE (and about half of them are by the doctor), but only 4.5 of them are tagged with label "DT", which demonstrates that more than half of doctors' utterances are not included in the summary. Such utterances can be greetings, symptom inquiry, etc. Second, the conversations between the patient and the doctor are asynchronous: either party can type some messages, walk away, and later come back to continue the discussion. This property makes the corpus different from other benchmark corpora (such as AMI (McCowan et al., 2005)) consisting of dialogues during in-person meetings. Third, for SUM2, all conversations have SUM2-A, and only a small portion (around 7.5% in the training and testing sets) have both SUM2-A and SUM2-B. Therefore, for the conversations with both SUM-2A and SUM-2B, we use their concatenation to compute the average length reported in Table 1a. Forth, while this paper focuses on summarization, the corpus can be used for other NLP tasks such as question answering and dialogue analysis.

3 Summarization via Tagging

To model conversation, a common approach is to use a two-level hierarchical sequential model (Serban et al., 2016), in which a conversation may be modeled as a sequence of utterances, and each utterance is modeled as a sequence of words or characters. Using such hierarchical models, conventional studies mainly focused on conversation generation (Sordoni et al., 2015; Serban et al., 2016; Serban et al., 2017), where a decoder is employed to generate responses conditioning upon the vectors encoded from the hierarchical modeling of previous utterances.

For our dataset, there is a big overlap between utterances and the summaries; for instance, as shown in Table 1b, SUM2 in the majority of the conversations (92.5% in training and 92.3% in the test set) are

in this dataset we only include typed messages for our research.

[7]The summary in this corpus often uses a full-width comma (U+FF0C) as a delimiter, and we use this delimiter to break a summary into segments.

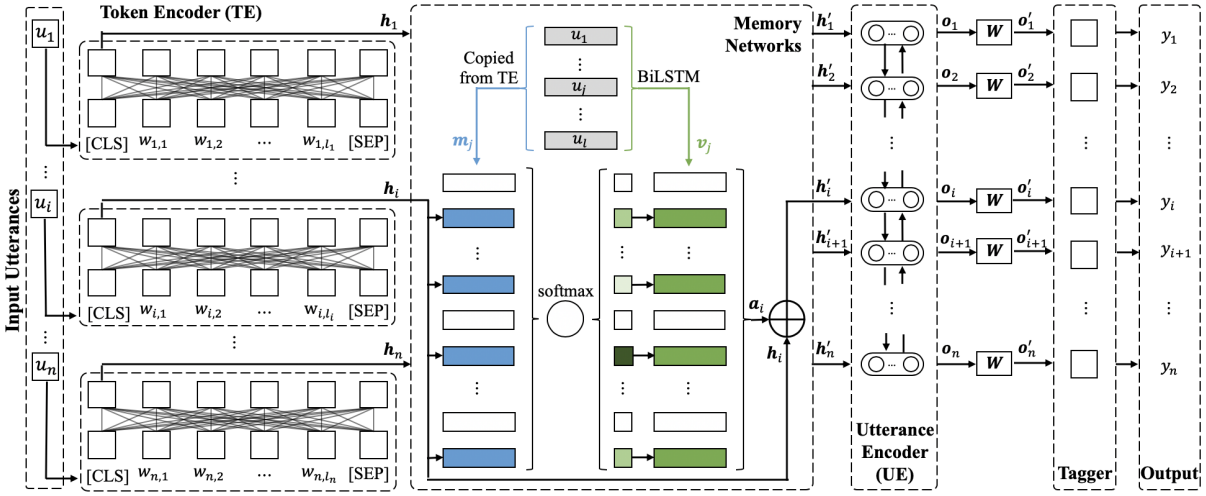

Figure 2: An illustration of the hierarchical encoder-tagger model (HET) with memory networks for identifying important utterances in conversations. Here, u_i is the i^{th} utterance in the conversation.

of type SUM2-A only and the rest contain both SUM2-A and SUM2-B, where SUM2-A is generated by concatenating several utterances in the conversation. To take advantage of such a property, we treat summarization as a tagging task; that is, we generate the summaries by first labeling the utterances with the PD, DT, OT tags and then concatenating the labeled utterances to form summaries.

We define the input utterance sequence as $\mathcal{U} = u_1, u_2, \cdots, u_i \cdots, u_n$ with each u_i presented as a sequence of basic tokens (e.g., word or character) $u_i = w_{i,1}, w_{i,2}, \cdots, w_{i,l_i}$. To model the input, our model follows the typical hierarchical structure in which the tokens and utterances are encoded with by separate encoders and hierarchically stacked. Then a tagger is attached at the utterance-level to predict PD/DT/OT labels. Afterwards, we concatenate the utterances labeled by PD and DT to generate the summary of medical problems and doctor's diagnosis, respectively. To further enhance our model, we adopt memory networks (Sukhbaatar et al., 2015) to incorporate the information from relevant utterances in the conversation. Therefore, our model is a *hierarchical encoder-tagger* (HET) with the memory module applied between the token-level and utterance-level encoders, which is illustrated in Figure 2. Also, it is worth noting that our method can generate the two types of summaries simultaneously, since they directly come from the predicted PD/DT/OT labels. In the following texts, we firstly introduce the memory module and then elaborate the whole hierarchical tagging process with the memories.

3.1 Utterance Memories

As discussed above, we regard our summarization task as an utterance tagging process. Similar to other tagging tasks in which contextual information is highly helpful in determining the output tags (Song and Xia, 2012; Marcheggiani and Titov, 2017; Higashiyama et al., 2019; Tian et al., 2020a; Tian et al., 2020b), for each utterance u_i in the conversation, relevant utterances in each conversation also provide useful information (Zhang et al., 2018) to determine whether a particular utterance is important. To exploit the information from relevant utterances, we adopt end-to-end memory networks (Sukhbaatar et al., 2015), which (as well as the variants) have been demonstrated to be useful in many tasks (Miller et al., 2016; Tian et al., 2020c), to learn from them to facilitate important utterance tagging. In doing so, we first map all utterances $[u_1, \cdots, u_j, \cdots, u_n]$ in the conversation into their memory vectors and value vectors. The memory vectors (denoted by $\mathbf{m}_j$ for u_j) are directly copied from the utterance representation obtained from the token encoder; the value vectors (denoted by $\mathbf{v}_j$ for u_j) are obtained by a BiLSTM encoder. Specifically, memory vectors $\mathbf{m}_j$ are used to compute the similarity with the input utterance; while $\mathbf{v}_j$ carries u_j's encoding information for generating final memory output. Then for each utterance u_i with its representation $\mathbf{h}_i$, we use it to address relevant utterance through the memory, which is formalized as

$$p_{i,j} = \frac{\delta_{i,j} \cdot exp(\mathbf{h}_i \cdot \mathbf{m}_j)}{\sum_{j=1}^{l_j} \delta_{i,j} \cdot exp(\mathbf{h}_i \cdot \mathbf{m}_j)} \quad (1)$$

Here, $\delta_{i,j} \in \{0, 1\}$ is a binary activator which equals 1 if the speaker of u_j is identical with that of u_i and equals 0 otherwise; $\mathbf{m}_j = \mathbf{h}_j$ because the memory vectors are copied from the utterance representation obtained from the token encoder (TE); $p_{i,j}$ is the weight measuring the relevance between u_j and u_i. Afterwards, the value vectors $\mathbf{v}_j$ of u_j are weighted with $p_{i,j}$ and summed by

$$\mathbf{a}_i = \sum_{j=1}^{l} p_{i,j} \mathbf{v}_j. \tag{2}$$

where $\mathbf{a}_i$ is the vector to represent the information from relevant utterances via a weighted sum operation.

3.2 The Hierarchical Encoder-tagging with Memories

To obtain the representation of each input utterance u_i, we apply BERT (Devlin et al., 2019) as our token-level encoder (TE), and use the encoded hidden vector of "[CLS]"[8] as $\mathbf{h}_i$ to represent the utterance u_i. Once $\mathbf{a}_i$ is obtained from the memory module, we concatenate it with $\mathbf{h}_i$ and get the resulting utterance representation for the utterance level encoding[9] by

$$\mathbf{h}'_i = \mathbf{h}_i \oplus \mathbf{a}_i \tag{3}$$

Then, an utterance-level encoder (UE) is applied to model the utterance representations in a sequential way. For example, if we use LSTM for UE, the utterance-level encoding is formulated as

$$\mathbf{o}_i = LSTM(\mathbf{o}_{i-1}, h'_i) \tag{4}$$

where the $\mathbf{o_i}$ is the step-wise state for utterances and h_i is used as the input to the UE at each time step. Note that, in addition to LSTM, there are many other ways for UE, e.g., BiLSTM. Herein we use LSTM as an example of the UE for the sake of simplicity.

On the top of the encoder, there is the tagger layer performing the identification task, where a trainable matrix $\mathbf{W}$ and bias vector $\mathbf{b}$ is used to align $\mathbf{o}_i$ to the output space:

$$\mathbf{o}'_i = \mathbf{W} \cdot \mathbf{o}_i + \mathbf{b} \tag{5}$$

Afterwards, a softmax or conditional random field (CRF) (Lafferty et al., 2001) algorithm is applied to $\mathbf{o}'_i$ to obtain the output tags. Finally, we concatenation all utterance with the label PT and DT to generate the summary of patient's problem (SUM1) and doctor's diagnoses (SUM2), respectively.

4 Experiments

4.1 Settings

We experiment our HET model with and without the memory on our corpus. For model implementation, at the token-level encoder (TE), we use the Chinese version of BERT[10] and ZEN (Diao et al., 2019)[11] with their default settings, where for both BERT and ZEN, we use 12 layers of multi-head attentions with the dimension of hidden vectors set to 768; for the utterance level, we firstly run experiments with no encoder; then following previous studies such as (Kalchbrenner and Blunsom, 2013; Zhao et al., 2017; Kumar et al., 2018), we experiment with two recurrent neural network models (namely, LSTM and BiLSTM) to encode the utterance sequence for each conversation, where the dimension of hidden states is set to 300 for LSTM and 150 for BiLSTM encoder.

In the memory module, the embedding matrix and BiLSTM encoder for obtaining the value vectors $\mathbf{v}_j$ for u_j are applied directly to the Chinese characters in the utterance. All parameters in the embedding matrix and the BiLSTM encoder in the memory module are initialized randomly, with the dimension of embedding and hidden states set to 768 and 384, respectively (which allows the dimension of $\mathbf{v}_i$ to match that of the hidden vector of BERT and ZEN).

For the tagger, we run two types of them, i.e., softmax and CRF, in order to test whether there is a strong

[8] "[CLS]" is the sentence initial symbol added by BERT.

[9] There are different naming schemes for the hierarchy of conversation encoders. To clarify, *TE* in this paper is called *utterance encoder* in other studies, while our *UE* is also called *context-* or *conversation-level encoder* elsewhere.

[10] We use the Chinese base model from `https://s3.amazonaws.com/models.huggingface.co/`.

[11] We obtain the pre-trained ZEN model from `https://github.com/sinovation/ZEN`.

UE	Tagger	$\mathcal{M}$	PD (SUM1)						DT (SUM2)					
			P	R	F	R-1	R-2	R-L	P	R	F	R-1	R-2	R-L
None	softmax	×	**90.01**	81.94	85.78	**87.18**	**83.85**	**87.18**	86.09	**90.26**	88.12	79.62	70.58	79.59
		√	89.11	**84.03**	**86.49**	87.06	83.55	87.06	**87.63**	89.64	**88.62**	**80.03**	**70.98**	**80.00**
	CRF	×	89.76	**83.91**	**86.74**	**87.58**	**84.78**	**87.58**	86.31	89.67	87.96	79.99	70.73	79.95
		√	**90.85**	82.83	86.65	87.37	84.58	87.37	**87.96**	**90.12**	**89.02**	**80.17**	**71.29**	**80.13**
LSTM	softmax	×	91.58	**85.95**	88.68	90.80	86.78	90.80	87.67	**89.14**	88.40	80.41	71.29	80.38
		√	**92.35**	85.67	**88.88**	**90.85**	**86.83**	**90.85**	**88.89**	88.98	**88.93**	**80.75**	**71.76**	**80.71**
	CRF	×	90.78	**86.39**	**88.53**	90.05	86.08	90.05	86.95	**89.92**	88.41	80.08	71.20	80.05
		√	**91.26**	85.70	88.39	**90.14**	**86.17**	**90.14**	**88.52**	89.69	**89.10**	**80.54**	**71.60**	**80.50**
BiLSTM	softmax	×	91.56	**85.68**	**88.52**	91.02	87.12	91.02	88.68	89.77	89.22	80.74	71.61	80.70
		√	**92.02**	85.24	88.50	**91.06**	**87.15**	**91.06**	**88.73**	**90.44**	**89.57**	**80.94**	**72.40**	**80.91**
	CRF	×	91.65	85.24	88.33	90.98	87.28	90.98	87.53	89.28	88.40	80.44	71.36	80.41
		√	**92.03**	**85.25**	**88.51**	**91.01**	**87.38**	**91.01**	**88.88**	**89.88**	**89.37**	**80.87**	**72.07**	**80.84**

(a) Results from BERT

UE	Tagger	$\mathcal{M}$	PD (SUM1)						DT (SUM2)					
			P	R	F	R-1	R-2	R-L	P	R	F	R-1	R-2	R-L
None	softmax	×	88.24	**84.78**	**86.48**	**87.34**	**83.97**	**87.34**	**90.58**	88.37	89.46	80.11	71.64	80.08
		√	**90.66**	82.54	86.41	87.32	83.94	87.32	89.84	**89.21**	**89.52**	**80.42**	**71.99**	**80.39**
	CRF	×	89.83	83.96	86.80	**87.73**	**84.32**	**87.73**	89.52	89.13	89.32	80.18	71.60	80.15
		√	**91.07**	**82.97**	**86.83**	87.59	84.18	87.59	**89.94**	**89.22**	**89.58**	**80.46**	**71.99**	**80.43**
LSTM	softmax	×	**92.55**	**85.90**	**89.11**	**91.07**	**87.08**	**91.07**	**90.94**	88.92	89.92	80.96	72.73	80.93
		√	91.86	85.74	88.70	90.91	86.93	90.91	90.92	**89.35**	**90.13**	**81.28**	**73.17**	**81.25**
	CRF	×	90.89	**86.66**	**88.73**	90.15	86.12	90.15	**89.96**	88.55	89.25	80.72	72.25	80.68
		√	**91.39**	85.81	88.51	**90.21**	**86.24**	**90.21**	89.92	**89.94**	**89.93**	**81.10**	**72.75**	**81.07**
BiLSTM	softmax	×	91.75	**85.93**	**88.75**	91.13	87.19	91.13	90.64	89.62	90.13	81.80	73.62	81.77
		√	**91.76**	85.27	88.39	**91.26**	**87.36**	**91.26**	**90.73**	**89.88**	**90.30**	**81.96**	**73.89**	**81.93**
	CRF	×	91.88	**85.40**	88.53	91.16	87.28	91.16	90.45	**89.82**	**90.13**	81.51	73.32	81.48
		√	**93.25**	84.81	**88.83**	**91.20**	**87.38**	**91.20**	**93.14**	87.04	89.99	**81.90**	**73.88**	**81.87**

(b) Results from ZEN

Table 2: The results of HET using BERT and ZEN as the token encoder with and without the memory module ($\mathcal{M}$). We also try different combinations of utterance encoders (UE) (i.e., none, LSTM, and BiLSTM) and taggers (i.e., softmax and CRF). *PD* and *DT* are the two tags for important utterances; P, R, and F are the precision, recall, and F scores of the predicted labels when compared with the silver-standard PD/DT/OT labels; *R-1*, *R-2*, and *R-L* are ROUGE-1, ROUGE-2, and ROUGE-L scores of the generated summaries when compared with gold references in the corpus (i.e., the SUM1 and SUM2).

dependency between the importance labels of adjacent utterances. We use cross-entropy and negative log-likelihood as loss functions for softmax and CRF, respectively.

For evaluation, we use F scores for the tagging results[12] and use ROUGE-1, ROUGE-2, and ROUGE-L scores[13] to evaluate the generated summaries using SUM1 and SUM2 in the dataset as the gold standard. If the SUM2 of a conversation includes both SUM2-A and SUM2-B, we treat the concatenation of SUM2-A and SUM2-B as the gold standard for SUM2 in all the experiments, except the results in Table 4.

4.2 Basic HETs

The first experiment is to explore how the HET models perform under different settings on the proposed dataset, where models with and without the memory module and configured with different token encoders (BERT and ZEN), UEs (no UE, LSTM, and biLSTM), and taggers (softmax and CRF) are tested. Table 2(a) and 2(b) show the results of utterance tagging (in terms of precision, recall, and F scores) and summarization (in terms of ROUGE-1, ROUGE-2, and ROUGE-L) for both problem statement (SUM1)

[12]We use the code in the *sklearn* framework `https://scikit-learn.org/stable/modules/classes.html`.

[13]The code is from `https://github.com/google-research/google-research/tree/master/rouge`.

Models	PD (SUM1)						DT (SUM2)					
	P	R	F	R-1	R-2	R-L	P	R	F	R-1	R-2	R-L
Nallapati et al. (2017)	76.69	81.74	79.13	89.03	84.68	89.03	78.70	72.66	75.56	52.10	37.90	52.07
Wang et al. (2019)	76.86	81.41	79.07	89.25	84.92	89.25	77.77	71.66	74.59	53.34	37.41	47.73
Ours (BERT)	**92.02**	85.24	**88.50**	91.06	87.15	91.06	88.73	**90.44**	89.57	80.94	72.40	80.91
Ours (ZEN)	91.76	**85.27**	88.39	**91.26**	**87.36**	**91.26**	**90.73**	89.88	**90.30**	**81.96**	**73.89**	**81.93**

Table 3: Experimental results of our runs of models from previous studies as well as our best HET (with BiLSTM UE, softmax tagger, and the memory module).

and treatment recommendation (SUM2) when BERT and ZEN token encoders are used.

Some observations are stated in order below. First, the overall results demonstrate that the method of generating summaries via tagging works well on our dataset. In most cases, models that perform well on tagging (F scores) also perform well on summarization (ROUGE scores). Second, for both BERT and ZEN encoders, the HET model works well with different combinations of UEs and taggers, which illustrates the validity of our approach. Among different settings, the one using BiLSTM UE outperforms others, suggesting that the sequential organization of utterances play an important role in identifying important utterances in conversations. Third, compared with models without the memory module, models with memories achieve greater improvements on the doctor diagnoses (SUM2). However, the effect of memories is not as good for the problem description (SUM1). One possible explanation could be that the information of other utterances is more useful for determining whether an utterance can be tagged for SUM2 than that for SUM1; the memory module can appropriately model such information and thus including the memory module in HET is more helpful on SUM2 than that for SUM1.

4.3 Comparison with Previous Studies

On our dataset, we compare our approach with two previous extractive summarization models. The first one is SummaRuNNer (Nallapati et al., 2017) and the other is a contextualized extractive method (CEM) proposed by Wang et al. (2019).[14] Since these models are originally designed for document summarization, which cannot generate summaries for patient's problem and doctor's diagnosis simultaneously, in our experiments, we directly concatenation all utterances to form a document as the input (i.e., the conversation utterances are regarded as document sentences) and train the models for SUM1 and SUM2 separately . For both models, we apply the Chinese character embeddings from Tencent Embedding[15] (Song et al., 2018) and select the top ranked 7%, and 24%[16] of the utterances (sentences) as the summarization of patient's problem and doctor's diagnosis, respectively. Table 3 shows the best results of the two reference models as well as our model using BERT and ZEN with the best setting (i.e., BiLSTM UE, softmax tagger, as well as the memory module), where our approach outperforms both referential systems on both SUM1 and SUM2, where the model with ZEN obtains the best results.

4.4 SUM2-A vs. SUM2-B as Gold Standard

As shown in Table 1b, 7.7% of conversations in the test set contain both SUM2-A and SUM2-B. So far, for those conversations, we have used the concatenation of SUM2-A and SUM2-B as the gold standard (see Table 2-3). Table 4(a) shows the performance of the four systems (i.e., Ref-1 from Nallapati et al. (2017) and Ref-2 from Wang et al. (2019)) and our model using BERT and ZEN as TE under the best setting (e.g., BiLSTM UE, softmax tagger, with the memory module)) on the entire test set, but with SUM2-A as the gold standard. Not surprisingly, for all three models, their performances with SUM2-A as the gold standard are higher than the ones with concatenation of SUM2-A and SUM2-B as the gold standard (see the last three columns in Table 3).

[14]We use their released code at `https://github.com/hpzhao/SummaRuNNer` and `https://github.com/hongwang600/Summarization`.

[15]We use the official release at `https://ai.tencent.com/ailab/nlp/embedding.html`.

[16]The two numbers are selected according to the percentage of utterances labeled by PD and DT in uttereances in each conversation in the training set (see Table 1).

Model	SUM2-A		
	R-1	R-2	R-L
Ref-1	53.27	38.98	53.24
Ref-2	54.29	38.39	48.77
Ours (BERT)	82.24	73.83	83.22
Ours (ZEN)	**83.25**	**75.32**	**83.23**

(a) Results on the entire test set with SUM2-A as the gold standard.

Model	SUM2-A			SUM2-B		
	R-1	R-2	R-L	R-1	R-2	R-L
Ref-1	57.79	42.98	57.77	17.21	12.01	17.20
Ref-2	58.96	42.69	54.58	16.73	11.33	15.75
Ours (BERT)	78.99	69.81	78.99	**21.08**	16.18	**21.08**
Ours (ZEN)	**79.62**	**70.90**	**79.62**	21.07	**16.24**	21.07

(b) Results on the 697 test conversations that have both SUM2-A and SUM2-B, with either SUM2-A or SUM2-B as the gold standard.

Table 4: ROUGE scores of two reference models (i.e., Ref-1 (Nallapati et al., 2017) and Ref-2 (Wang et al., 2019)) and our best model (with BiLSTM UE, softmax tagger, and the memory module), where different part (i.e., SUM2-A or SUM2-B) of SUM2 is regarded as the gold standard.

Table 4(b) reports the results on the 697 conversations in the test set that have both SUM2-A and SUM-2B, with either SUM2-A or SUM-2B as the gold standard. For all three systems, ROUGE scores with SUM2-B as gold standard are much lower than the ones with SUM2-A, indicating that generating summaries that are similar to manually crafted summaries is still a challenge task.

4.5 HETs with Meta-Information

In addition to the utterances, each conversation in the dataset has three major types of meta information; namely, speaker role (patient or doctor) (SR), hospital department (HD), and disease name (DN). We experiment with adding such meta-information on top of our model using BERT and ZEN as TE under the best setting. To incorporate the meta-information, we use a single-layer neural network to transfer them into vectorized representation, and concatenate them to their correspondent encoder layers. Specifically, SR is added to TE; HD and DN are added to UE.[17]

Table 5 reports the performance of our HET models with different combination of the meta-information, where the results without using any meta-information are shown in the first row (which is identical to the last row in Table 3). Compared to the baselines, models with meta-information achieve better performance in most cases. Specifically, adding SR results in higher improvements compared with HD and DN. One possible explanation could be that the utterances from the patient and the doctor could be more important in generating problem statement (SUM1) and treatment recommendation (SUM2), respectively. Therefore, adding SR would help our model to focus more on the utterances for the patients and the doctors when it is predicting PD and DT labels for SUM1 and SUM2, respectively.

5 Related Work

5.1 Extractive Summarization

As a direct research line related to our work, extractive summarization aims to extract important sentences in the input and use them to form a summary. Most previous studies focused on document summarization (Nallapati et al., 2017; Narayan et al., 2018; Wang et al., 2019; Zhang et al., 2019; Xiao and Carenini, 2019; Luo et al., 2019) while some focused on summarization of meeting transcripts (Riedhammer et al., 2010; Singla et al., 2017), where their problem settings and data preparation are different from ours. Specifically, compared with summarization for documents, our task of conversation summarization is more challenging because utterances in the conversation are less formally written and there are speaker role changes during the entire conversation; compared with summarization for meeting transcripts, where the summary is similar to a short meeting-log, our task requires to generate more informative summaries to facilitate the needs of providing useful information to potential patients from the online platform. General extractive approaches for summarization always face challenges of redundancy when they use extracted sentences to generate an informative and readable summary within a length, in which additional modeling is required to address it even though with powerful neural models, e.g., BiLSTM (Nallapati et al., 2017), transformers (Zhang et al., 2019), and attentions (Xiao and Carenini, 2019). On the contrary,

[17]More specifically, we concatenate the SR representation with $\mathbf{h}_i$ obtained from Eq. 3 and concatenate the DR and DN representation with $\mathbf{o}_i$ from Eq. 4.

Meta-info			PD (SUM1)						DT (SUM2)					
SR	HD	DN	P	R	F	R-1	R-2	R-L	P	R	F	R-1	R-2	R-L
×	×	×	92.02	85.24	88.50	91.06	87.15	91.06	88.73	90.44	89.57	80.94	72.40	80.91
√	×	×	95.45	83.74	89.21	92.40	88.72	92.40	**91.60**	89.12	**90.34**	82.17	74.15	82.14
×	√	×	93.39	84.67	88.81	91.71	87.84	91.71	88.62	89.58	89.10	81.18	72.25	81.16
×	×	√	92.54	85.73	89.01	91.17	87.24	91.17	89.40	88.75	89.08	81.24	72.33	81.20
√	√	√	**95.75**	**83.94**	**89.46**	**92.67**	**88.88**	**92.67**	90.92	**89.30**	90.10	**82.47**	**74.43**	**82.44**

(a) Results from BERT

Meta-info			PD (SUM1)						DT (SUM2)					
SR	HD	DN	P	R	F	R-1	R-2	R-L	P	R	F	R-1	R-2	R-L
×	×	×	91.76	85.27	88.39	91.26	87.36	91.26	90.73	89.88	90.30	81.96	73.89	81.93
√	×	×	95.61	83.83	89.33	92.49	88.77	92.49	**92.52**	89.42	**90.94**	83.10	75.31	83.07
×	√	×	93.42	84.64	88.82	91.95	88.12	91.95	91.83	89.23	90.51	82.12	74.30	82.10
×	×	√	92.98	**85.29**	88.97	91.48	87.65	91.48	91.43	90.08	90.75	82.11	74.40	82.08
√	√	√	**95.95**	84.04	**89.60**	**92.80**	**88.97**	**92.80**	91.11	**90.66**	90.88	**83.31**	**75.48**	**83.29**

(b) Results from ZEN

Table 5: Results of our models using BERT and ZEN TE under the best setting (with BiLSTM UE, softmax tagger, and the memory module). "SR", "HD", and "DN" stand for the meta-information of speaker roles, hospital departments, and disease names, respectively.

in our work, this challenge may not be an issue because the redundancy in the original input is limited and directly concatenating selected utterances with their same order in the original conversation does not lead to unreadable summmaries in most cases. Therefore, to have a good performance in conversation summarization in the medical domain, task-specific designs of summarization model are expected.

5.2 Utterance Modeling in Conversations

Studies on dialogue systems have drawn much attention recently, where many of them have been done on utterance modeling in human-human conversations (Wang et al., 2018a; Liu et al., 2019). In these studies, one stream of utterance modeling focuses on dialogue act classifications, which aims to attribute one of predefined acts to each utterance in conversations (Lee and Dernoncourt, 2016; Liu et al., 2017; Kumar et al., 2018; Wang et al., 2018b; Raheja and Tetreault, 2019). Another stream focuses on assessment of utterances in terms of their quality in various aspects, such as sentiment analysis (Inaba and Takahashi, 2016; Lison and Bibauw, 2017; Misra et al., 2019). Our study on extractive summarizations for conversation can be regarded as in the line of the latter stream in evaluating utterances for human-human conversations, where little research has been done for utterances based on their importance to the pragmatic outcomes (i.e., summaries for problem statement and treatment recommendations in our study) of the conversations.

6 Conclusion and Future Work

In this paper, we proposed a new task of medical conversation summarization, which is performed by identifying important utterances in the conversation between patients and doctors. Based on the real data from a Chinese online medical service provider, a hierarchical encoder-tagger model (HET), which is enhanced by the memory module, was proposed to tag each utterance in a conversation with problem statement or treatment recommendation. The labeled utterances are then concatenated to form summaries. The experimental results demonstrate the validity of our approach to medical conversation summarization via identifying important utterances on the proposed dataset. For future work, we plan to perform further key information extraction on the conversation summaries from similar medical problems, so that we can obtain relevant information such as symptoms and treatment recommendations to a particular medical problem and help new patients to locate more precise references that are covered in many cases.

Acknowledgements

This work is supported by The Chinese University of Hong Kong (Shenzhen) under University Development Fund UDF01001809.

References

Nai-Wen Chang, Hong-Jie Dai, Jitendra Jonnagaddala, Chih-Wei Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2015. A Context-aware Approach for Progression Tracking of Medical Concepts in Electronic Medical Records. *Journal of Biomedical Informatics*, 58(S):S150–S157, Dec.

Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of ACL*.

Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the SIAM International Conference on Data Mining*, pages 432–440.

James Cormack, Chinmoy Nath, David Milward, Kalpana Raja, and Siddhartha R. Jonnalagadda. 2015. Agile Text Mining for the 2014 I2B2/UTHealth Cardiac Risk Factors Challenge. *Journal of Biomedical Informatics*, 58(S):S120–S127, Dec.

Mihai Dascalu, Stefan Trausan-Matu, and Philippe Dessus. 2010. Utterances Assessment in Chat Conversations. *Research in Computing Science*, 46:323–334, March.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural Networks for Joint Sentence Classification in Medical Paper Abstracts. In *Proceedings of EACL*, pages 694–700, Valencia, Spain, April.

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.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2019. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. *ArXiv*, abs/1911.00720.

Paul Drew and John Heritage, editors. 1992. *Talk at Work: Interaction in Institutional Settings*. Cambridge University Press, Cambridge.

Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. 2019. Incorporating Word Attention into Character-Based Word Segmentation. 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 2699–2709, Minneapolis, Minnesota, June.

Michimasa Inaba and Kenichi Takahashi. 2016. Neural Utterance Ranking Model for Conversational Dialogue Systems. In *Proceedings of SIGDIAL*, pages 393–403, Los Angeles, September.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. *arXiv preprint*, abs/1306.3584.

Su Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic Classification of Sentences to Support Evidence Based Medicine. *BMC bioinformatics*, 12 Suppl 2:S5, 03.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF. In *Proceedings of AAAI*, New Orleans, Louisiana, USA.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289, San Francisco, CA, USA.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Proceedings of NAACL*, pages 515–520, San Diego, California, June.

Pierre Lison and Serge Bibauw. 2017. Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models. In *Proceedings of SIGDIAL*, pages 384–394, Saarbrucken, Germany, August.

Diane J. Litman and James F. Allen. 1987. A Plan Recognition Model for Subdialogues in Conversations. *Cognitive Science*, 11(2):163–200.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using Context Information for Dialog Act Classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark, September.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R. Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, Wai Leng Chow, and Nancy F. Chen. 2019. Fast Prototyping a Dialogue Comprehension System for Nurse-Patient Conversations on Symptom Monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31, Minneapolis, Minnesota, June.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading Like HER: Human Reading Inspired Extractive Summarization. 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 3033–3043, Hong Kong, China, November.

Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark, September.

Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The AMI Meeting Corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Amita Misra, Mansurul Bhuiyan, Jalal Mahmud, and Saurabh Tripathy. 2019. Using Structured Representation and Data: A Hybrid Model for Negation and Sentiment in Customer Service Conversations. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 46–56, Minneapolis, USA, June.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*, pages 3075–3081.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking Sentences for Extractive Summarization with Reinforcement Learning. 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 1747–1759, New Orleans, Louisiana, June.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium, October-November.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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 3727–3733, Minneapolis, Minnesota, June.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long Story Short–global Unsupervised Models for Keyphrase Based Meeting Summarization. *Speech Communication*, 52(10):801–815.

J.D. Robinson. 2012. Overall Structural Organization. *The Handbook of Conversation Analysis*, pages 257–280.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-end Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI*, pages 3776–3783.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of AAAI*, pages 3295–3301.

Jurica Ševa, Martin Wackerbauer, and Ulf Leser. 2018. Identifying Key Sentences for Precision Oncology Using Semi-Supervised Learning. In *Proceedings of BioNLP*, pages 35–46.

Karan Singla, Evgeny Stepanov, Ali Orkan Bayer, Giuseppe Carenini, and Giuseppe Riccardi. 2017. Automatic Community Creation for Abstractive Spoken Conversations Summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 43–47, Copenhagen, Denmark, September.

Yan Song and Fei Xia. 2012. Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *LREC*, pages 3853–3860.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 175–180.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of CIKM*, pages 553–562.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end Memory Networks. In *Advances in neural information processing systems*, pages 2440–2448.

Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, and Shourya Roy. 2007. Automatic Identification of Important Segments and Expressions for Mining of Business-Oriented Conversations at Contact Centers. In *EMNLP*, pages 458–467, Prague, Czech Republic, June.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. ChiMed: A Chinese Medical Corpus for Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy, August.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online, July.

Yuanhe Tian, Yan Song, and Fei Xia. 2020b. Supertagging Combinatory Categorial Grammar with Attentive Graph Convolutional Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020c. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Nan Wang, Yan Song, and Fei Xia. 2018a. Coding Structures and Actions with the COSTA Scheme in Medical Conversations. In *Proceedings of the BioNLP 2018 workshop*, pages 76–86, Melbourne, Australia, July.

Nan Wang, Yan Song, and Fei Xia. 2018b. Constructing a Chinese Medical Conversation Corpus Annotated with Conversational Structures and Actions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.

Hong Wang, Xin Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Self-Supervised Learning for Contextualized Extractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2221–2227, Florence, Italy, July.

Nan Wang, Yan Song, and Fei Xia. 2020. Studying Challenges in Medical Conversation with Structured Annotation. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 12–21, Online, July.

Wen Xiao and Giuseppe Carenini. 2019. Extractive Summarization of Long Documents by Combining Global and Local Context. 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 3011–3021, Hong Kong, China, November.

Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding Conversation Context for Neural Keyphrase Extraction from Microblog Posts. 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 1676–1686, New Orleans, Louisiana, June.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy, July.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative Encoder-Decoder Models for Task-Oriented Spoken Dialog Systems with Chatting Capability. In *Proceedings of SIGDIAL*, pages 27–36.