Understanding Medical Conversations with Scattered Keyword Attention and Weak Supervision from Responses

Xiaoming Shi,[1*] **Haifeng Hu,**[2] **Wanxiang Che,**[1†] **Zhongqian Sun,**[2] **Ting Liu,**[1] **Junzhou Huang**[2]
[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[2]Tencent AI Lab, Shenzhen, China
{xmshi, car, tliu}@ir.hit.edu.cn, {simbahu, sallensun, joehhuang}@tencent.com

Abstract

In this work, we consider the medical slot filling problem, i.e., the problem of converting medical queries into structured representations which is a challenging task. We analyze the effectiveness of two points: scattered keywords in user utterances and weak supervision with responses. We approach the medical slot filling as a multi-label classification problem with label-embedding attentive model to pay more attention to scattered medical keywords and learn the classification models by weak-supervision from responses. To evaluate the approaches, we annotate a medical slot filling data and collect a large scale unlabeled data. The experiments demonstrate that these two points are promising to improve the task.

Introduction

Recent advances in speech recognition and natural language processing have facilitated broad deployments of spoken dialogue systems (SDS) as natural interfaces for information access (Mori 1998; Smith and Hipp 1994; Zue and Glass 2000) , of which typical applications range from call center automation to virtual assistants for smart devices. A Critical component in SDS is spoken language understanding (SLU), which parses a natural language utterance into logical semantic representations that a computer can process effectively (Young et al. 2013). A typical pipeline of SLU includes: domain classification, intent detection, and slot filling (Tur and De Mori 2011). In this pipeline, slot filling has significant impact on the overall performance compared to intent detection (Li et al. 2017).

In the medical domain, dialogue systems that communicate with patients using natural language to obtain additional symptoms automatically attract more and more attention (Wei et al. 2018; Xu et al. 2019). In this work, we focus on slot filling on medical dialogues.

Slot filling is commonly treated as a structured prediction problem, where supervised learning algorithms, especially recurrent neural networks (Yao et al. 2013; Mesnil et al. 2015; Kurata et al. 2016; Jaech, Heck, and Ostendorf 2016; Hakkani-Tür et al. 2016) have demonstrated state-of-the-art performance. Traditional slot filling aims to label the explicit words in a given utterance according to a predefined domain ontology such that structured semantic representation (a.k.a slot-value pairs) can be extracted from the predicted labels (Mesnil et al. 2015; Kurata et al. 2016).

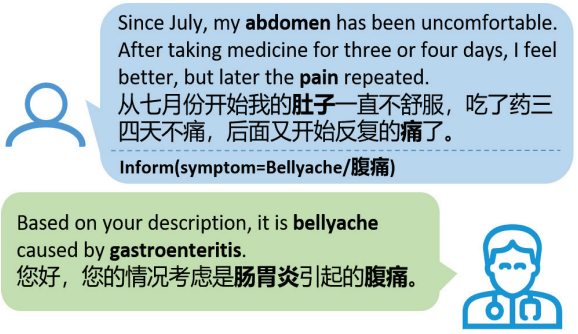

Figure 1: One turn of a medical conversation between a patient and a doctor.

Different from traditional slot filling tasks, slot filling for medical dialogues mainly faces two challenges: i) Medical dialogue data is unaligned where slot-values of the structured semantic representation do not explicitly occur in any specific spans; ii) Medical data annotation requires annotators with professional medical knowledge which leads to the high annotation cost.

Going deep into the first challenge, the unaligned issue mainly comes from two aspects: users' colloquial expression and scattered keyword. Concretely, patients' health reports are described in their colloquial language and vary from each other. What' more, patients always state their health status in a disorganized order that keywords for a single medical concept are dispersed. As an example in Figure 1, the medical concept *"Bellyache"* is described as the combination of *"abdomen"* and *"pain"* which is scattered in the original sentence. This unaligned issue between natural language queries and corresponding structured semantic representation has been tracked in (Barahona et al. 2016), treating

[*]Work done while Xiaoming Shi was Research Interns with Tencent AI Lab.
[†]Corresponding Author

the task as a multi-label classification task. Following this work, we treat the medical slot filling as a multi-label classification problem that aims to classify the sentences into pre-defined categories (i.e. slot-value pairs). Furthermore, to recognize those discontinuous ones better, we use the label-embedding attentive model (Wang et al. 2018) which makes the model more sensitive to medical keywords. To the best of our knowledge, we are the first to explore the scattered keyword issue for medical slot filling.

Facing the second challenge, we leverage a large number of unlabeled data with their responses as weak supervision. And concretely, a large number of medical dialogues exist in web medical communities. Furthermore, it is common that doctors always retell the patient's symptoms with formal expressions in their responses which is easy to access by string matching with medical concepts in medical knowledge bases. As the example in the Figure 1, the slot-value *"bellyache"* is mentioned by the doctor. Thus, following the intuition that the medical concepts in doctors' responses are closely associated with patients' queries, we propose a novel methodology to learn medical slot filling with weak supervision from doctors' responses as the model pre-training before the fine-tuning on well-annotated data. To the best of our knowledge, we are the first to explore weak supervision from responses in the unlabeled dataset for medical slot filling.

To evaluate the approaches, we annotate a medical slot filling dataset which is well annotated by four medical experts. The evaluations on the dataset show that our model achieves state-of-the-art performance against strong baselines which indicate that the scattered keyword issue and weak supervision from responses are two promising problem-solving points.

The main contributions are listed as follows:

- We annotate a medical dialog data for medical slot filling and collect a large amount of high-quality unlabeled data with slot-values in responses.
- We explore the label-embedding attentive model to relieve the problem of scattered keywords which shows that the scattered keyword issue is a potential direction for improving medical slot filling performance.
- We explore approaches to make use of unlabeled data with responses as weak supervision which shows weak supervision from responses in unlabeled data is effective for improving medical slot filling performance.

The code is released[1].

Task Definition

The task of Medical Slot Filling (MSL) on unaligned colloquial medical dialogue can be defined as follows: given a medical database D with the itemized domain knowledge involved, this task aims at transforming a natural language medical query q, in which colloquial expressions exist, into the grounded formal representation with discrete logical forms, to perform correct query upon D. Figure 1 gives an example. The input is a patient's health expression which is stated in colloquial language and the output is grounded formal representation with corresponding medical concepts.

Basically, the MSF is a specific branch of the slot filling module of dialogue systems. However, since the current slot filling approaches, mainly based on sequence labeling modules, require the essential formal information elements explicitly existing in the original natural language query, it is only needed to extract the essential items and reform them into the required format. The MSF task, by contrast, has the following characteristics: i) is expressed in colloquial words which is totally different from formal slot-values; ii) key semantics are scattered; iii) the amount of well-annotated data is limited. In this situation, no explicit elements (the keywords directly equal to the concepts in the medical knowledge base) exist in the fuzzy queries, and consequently, the models need to understand the semantic of the colloquial query and transform it into formal medical concepts or their combination. At the same time, a large amount of unlabeled data is worth exploring.

Labeled Training Data	1,152
Labeled Validation Data	500
Labeled Test Data	1,000
Unlabeled Data	100,000
Avg. # of Tokens Per Utterance	12.39
Slot	Symptom
# of values	29
Avg. Frequency	387
Max. Frequency	979
Min. Frequency	185

Table 1: Statistics on the medical slot filling dataset.

Medical Slot Filling Dataset

To promote the studies on the MSF task, this paper first presents a dataset. Apparently, it is not a trivial work to collect and manually label the large amount of clinical queries and their appreciate answers based on user logs of dialog systems. Fortunately, such data are found being accumulated in the web Medical Community (MC) services, and thus it is possible to achieve the dataset required by the MSF task from the MC source.

Data Collection Set-up

This paper chooses the MC forums as the data source, such as DingXiangYuan [2], BaiduMuzhiDoctor [3], etc.. Our goal is to collect dataset includes natural language clinical questions, and medical concepts in their corresponding expert answers. This resulted in expression diversity and semantical richness of the collected data.

Data collection is performed in a three-step process. First, all turns are collected and we use rules to make coarse-grained screening (e.g., selecting with sentence length, fil-

[1] https://github.com/xmshi-trio/MSL.git.

[2] http://www.dxy.cn/

[3] https://muzhi.baidu.com/

tering with keywords) of these sentences. In this process, we want to ensure the high quality of the data. Secondly, pre-annotation is launched. Four annotators who major in medicine annotate a small part of the selected data, following an annotation guide. This setup allows the annotators to report errors (e.g., not following the task or confusing utterances and annotation guide) found in the collected turns. We summarize the found annotation issues, improve the annotation guide, and then reinterpret the guide to annotators. Finally, the annotation process is launched. From annotated data, we only keep data that is consistently labeled by four annotators.

For those unlabeled data, we match sentences with the medical concepts in medical knowledge base, and those medical concepts occurring in the responses are regarded as the medical concepts in responses.

Data Statistics

Table 1 gives the basic statistics upon the dataset and the medical concepts concluded to in our dataset. There are totally 1152 labeled training data, 500 for the validation and another 1000 for the test. Besides, there are other 100,000 unlabeled data with medical concepts referred to in their responses is provided. According to Table 1, the average count of tokens in each utterance is 12.39 which shows that the sentences are long. In this task, we mainly annotate the slot `symptom`. There are totally 29 slot-values included in the dataset. And the average frequency of values is 387, 979 for maximum frequency and 185 for minimum frequency.

Approach

In this section, the label-embedding attentive model and weak supervision from responses as illustrated by Figure 2 and Figure 3, will be introduced. As discussed in the previous sections, formal medical concepts in doctors' responses can help understanding patients' queries, moreover, it is even much more challenging to using unlabeled data to improve the performance of the task.

Hence, in this part, we propose to perform weak supervised learning for model pre-training upon the rough query-answer pairs as a pre-training process for the slot filling model, without any human-annotated structure semantic representations.

The model is composed of *Label-embedding Attentive Model*, *Query Encoding and Classification*, and *Weak Supervision Learning based Model Pre-training* modules, which will be detailed in the parts below.

Label-embedding Attentive Model

Attention mechanism is widely used recently (Wang, Che, and Liu 2016). As the diagram in Figure 2 shows, the input word embedding sequence and the candidate slot-value pair representation directly interact through the label attention. This component recognizes which words are highly related to the slot filling task and aims to select keywords for the slot filling task.

Let $\mathcal{X} = \{x_1, \cdots, x_n\}$ be the input word embedding sequence, $\mathcal{S} = \{s_1, \cdots, s_m\}$ be the candidate symptom em-

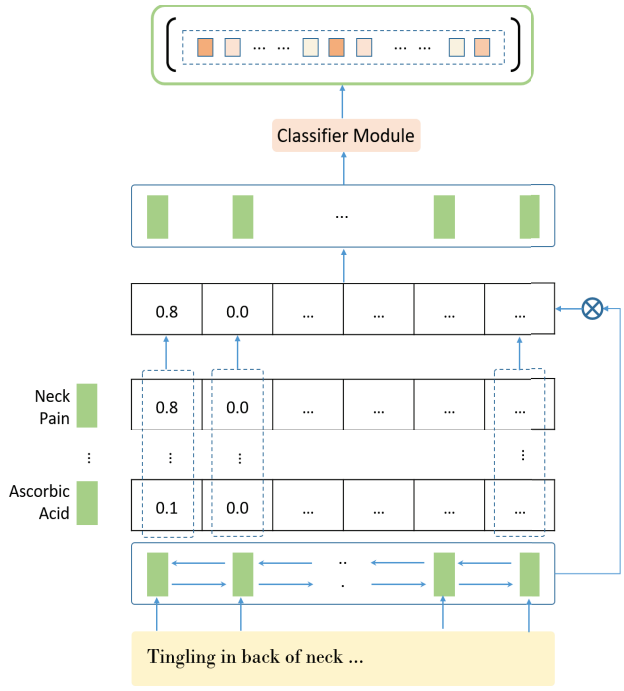

Figure 2: The illustration of the label-embedding attentive model.

beddings. These two representations are then forced to interact in order to learn a similarity metric which discriminates whether a word is relative with the slot-value pairs, and then the maximum value is taken:

$$a_i = max_j(x_i \cdot s_j) \tag{1}$$

where $\cdot$ denotes dot product, x_i denotes the ith sentence in $\mathcal{X}$, and s_j denotes the jth elements in $\mathcal{S}$. Then the word sequence is represented with the dot product a_i and x_i:

$$\mathcal{X}' = \{x_1 \cdot a_1, \cdots, x_n \cdot a_n\} \tag{2}$$

Query Encoding and Classification

The encoder aims to transfer the natural language input requests into real-valued vectors.

We use several text classification encoder as query encoders, including TextCNN(Kim 2014), RCNN(Lai et al. 2015), TextRNN(Liu, Qiu, and Huang 2016), DRNN(Wang 2018), RegionEmbedding(Qiao et al. 2018), and Star-Transformer encoder(Vaswani et al. 2017; Guo et al. 2019).

Let "Encoder" and r be the query encoder and query representation vector. The intermediate representations are obtained by putting label-embedding attentive model output through these encoders:

$$r = \text{Encoder}(\mathcal{X}') \tag{3}$$

The classifier is designed to map the query representation into probability distribution of all candidate classes. If $\phi_{dim}(x) = \sigma(Wx + b)$ is a layer which maps input vector x to a vector of size dim, the input to the final multi-label

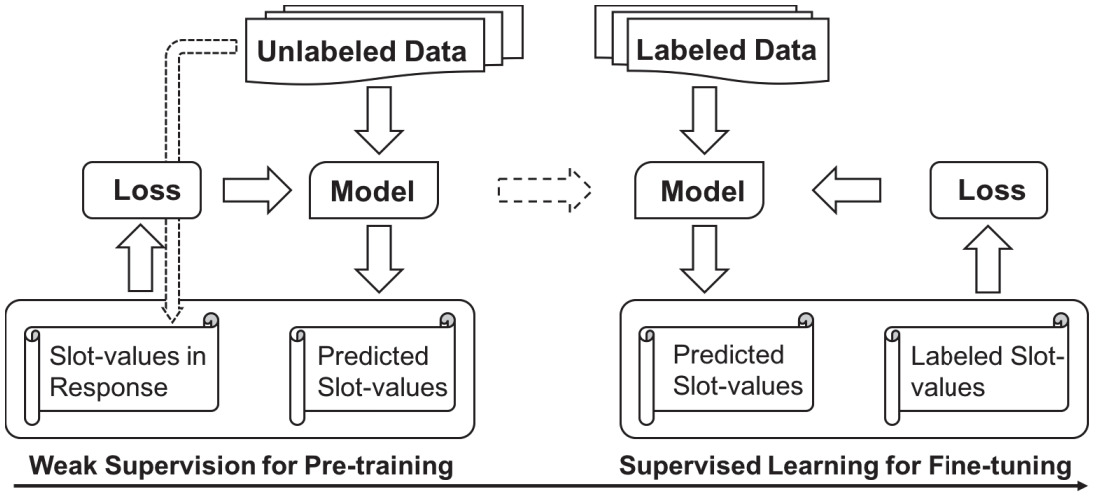

Figure 3: The architecture of Weak Supervision for Pre-training and Supervised Learning for Fine-tuning. When a model is being pre-training, medical concepts extracted from unlabeled data's responses are used as weak supervision of medical queries.

activation function $\mathrm{sigmoid}$ (which represents the decision) is given by:

$$y = \phi_c(r) \tag{4}$$

where c is the total amount of candidate labels, and y is the output probability distribution.

Two Steps for Model Training

We pre-train classifiers with weak supervision of slot-values in unlabeled dataset's responses. And then the learned model is fine-tuned in well-annotated dataset. The module is illustrated in Figure 3.

Weak Supervision for Pre-training This approach comes from the intuition that doctors' replies contain the relative medical concept with patients' health condition or retell patients' symptoms with formal medical concepts. Thus, it is possible to use doctors' replies as patients' queries' weak supervision, while doctors' replies may contain relative medical concepts, not only the medical concept describing patients' issues. What's more, the weak supervision method makes full of unlabeled data which helps reducing labeling costs. To do this, the classifier models are pre-trained on unlabeled data with the weak supervision from responses. Although pre-training with relevance data may result in the model not learning accurate labels, it can help the model eliminate most negative labels. After this, the pre-trained model will be fine-tuned on well-annotated data.

Let the parameters of classifier model be θ, and the loss function $\mathrm{BCEWithLogitsLoss}$ is represented as Loss. Given the input i medical query x_i, the model predicted ith probability distribution y_i of x_i where dimensions represent candidate slot-value pairs, and the target ith probability distribution y_i' which is extracted from responses. Then the learning object is

$$loss = \mathrm{Loss}(p(y_i|x_i, \theta), y_i') \tag{5}$$

The goal of the training is minimizing the loss to learn a best θ.

Supervised Learning for Fine-tuning The classifier models has already learned relative slot-value pair information on unlabeled data (i.e. been pre-trained), then the model will be trained on well-labeled data which is all fine-tuning means. The pre-training step with unlabeled data with weak supervision helps the model eliminate most negative labels, and the fine-tuning step aims to make the models more accurate on precise classes based on the relative class information from the pre-training step.

Let the parameters of classifier model be θ, the parameters of pre-trained model be θ' and the loss function $\mathrm{BCEWithLogitsLoss}$ be Loss. Given the input i medical query x_i, the model predicted ith probability distribution y_i of x_i where dimensions represent candidate slot-value pairs, and the target ith probability distribution y_i' which is annotated by experts. Then the learning object is

$$loss = \mathrm{Loss}(p(y_i|x_i, \theta), y_i') \tag{6}$$

where the parameters θ is initialized with θ'. The goal of the training is minimizing the loss to learn a best θ.

Experiment

Experimental Setting

Baselines We run several encoders on the dataset for the task: TextCNN(Kim 2014), RCNN(Lai et al. 2015), TextRNN(Liu, Qiu, and Huang 2016), DRNN(Wang 2018), RegionEmbedding(Qiao et al. 2018)and Star-Transformer (Vaswani et al. 2017; Guo et al. 2019). On the base of these encoders, we add label-embedding attentive model and weak supervision and test their performance on the dataset.

Evaluation We focus on five key evaluation metrics:

- Precision: the fraction of relevant instances among the retrieved instances.

Model	Precision	Recall	Micro F1	Macro F1	Turn Accuracy
TextCNN	87.91	59.11	70.69	56.95	45.50
TextCNN+A	87.50	62.73	73.07	62.40	49.50
TextCNN+A+WS	87.76	71.23	**78.76**	**73.76**	**54.30**
TextRCNN	83.11	64.08	72.36	65.86	49.70
TextRCNN+A	79.70	71.23	75.23	68.03	51.90
TextRCNN+A+WS	79.56	74.17	**76.77**	**72.02**	**55.30**
TextRNN	88.89	56.02	68.73	53.15	45.30
TextRNN+A	88.73	62.27	73.19	60.47	50.40
TextRNN+A+WS	90.08	64.98	**75.50**	**68.42**	**50.80**
DRNN	83.43	67.85	74.83	65.17	52.50
DRNN+A	82.11	70.86	76.07	67.42	51.90
DRNN+A+WS	82.94	79.44	**81.15**	**76.95**	**58.30**
RegionEmbedding	80.65	52.71	63.75	57.86	35.90
RegionEmbedding+A	85.17	57.53	68.67	61.63	40.60
RegionEmbedding+A+WS	86.85	64.16	**73.80**	**68.35**	**46.30**
Star-Transformer	81.92	67.55	74.04	69.76	48.90
Star-Transformer+A	79.66	74.32	76.90	72.34	**53.20**
Star-Transformer+A+WS	81.71	74.70	**78.05**	**73.48**	52.80

Table 2: The micro F1 score, macro F1 score and turn accuracy results of the six encoders: TextCNN(Kim 2014), TextRCNN(Lai et al. 2015), TextRNN(Liu, Qiu, and Huang 2016), AttentiveConvNet(Yin and Schütze 2018), DRNN(Wang 2018), RegionEmbedding(Qiao et al. 2018), Star-Transformer encoder(Vaswani et al. 2017; Guo et al. 2019). 'A' represents label-embedding attentive model and 'WS' represents weak supervision.

- Recall: the fraction of relevant instances that have been retrieved over the total amount of relevant instances.
- Micro and Macro F1 score: the F1 score in both micro-average and macro-average manners are used to measure the diffierence between predicted and gold slot-values.
- Turn Accuracy: the proportion of dialogue turns where all the slot-value pairs are correctly identified.

Settings When training models, we use mini-batch with batch-size=16 and the Adam optimizer with default parameters (a fixed learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$) (Kingma and Ba 2014). The number of iterations per batch is set to be 15 in the experiments. To avoid overfitting, we shuffle batches in the dataset when process the dataset. And the dropout rate is set to 0.5 to avoid overfitting. We use GloVe (Pennington, Socher, and Manning 2014) as the pre-trained embedding and the dimension of word is set as 300.

Experimental Results and Analysis

Main Results The main results on the dataset are listed in Table 2. We use 10, 000 unlabeled data for pre-training and all labeled data for fine-tuning. In the table, 'A' represents adding label-embedding attentive model and 'WS' represents adding weak supervision.

Comparing raw classifiers to those with label-embedding attentive, we can see that models with label-embedding attentive model perform better than raw classifiers both in Micro F1 and Macro F1 for improvement of 3.12% and 3.92% respectively. This shows that label-embedding attentive model can significantly improve the performance of the model. What's more, the label-embedding attentive gets more improvement on Macro F1 than Micro F1. This indicates that the attention can help the model improve more on low-source labels, for that Macro F1 calculates the average F1 score on all labels which cares labels as equal. Besides, the label-embedding attentive model helps model get 5.27% improvement on recall which tells that keyword contributes to recognize colloquial expressions unseen in training data when testing. All these results tell that paying more attention to scattered medical keywords is an effective way to improve the medical slot filling.

Taking a look at the performance of classifiers with the label-embedding attentive model and weak supervision pre-training, we can find that the weak supervision pre-training approach is obviously very effective for improvement of 3.36% and 6.78% in Micro F1 and Macro F1 respectively. Unlabeled data contains more of sparse labels of the labeled data which promotes the improvement of the Macro F1 score more than the Micro F1 score. At the same time, the recall value gets further improvement. The reason may be that unlabeled data contains more different colloquial expressions not occurring in labeled data which leads to the model can identify examples that have not been seen in the training data when testing. All these results indicate that weak supervision from responses benefits the medical slot filling.

Adding Bert Pre-training step in word embedding methods are always conducted on language models to get a well-initialized word embedding from the aspect of transfer learning, while the pre-training step in our methods focus on the same task with weak supervised data to get well-initialized model parameters.

(Devlin et al. 2018) shows rather surprising results in a large number of tasks. In this experiment, we aim to analyze the influences of pre-trained word embeddings on the pro-

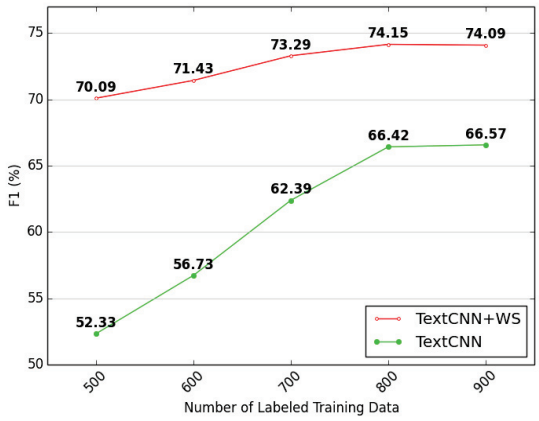

Figure 4: The F1 score on different training data amount when fine-tuning. 'WS' means adding weak supervision as pre-training.

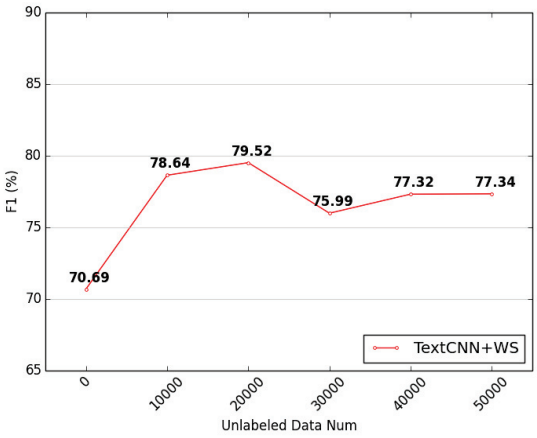

Figure 5: The F1 score on different unlabeled data amount for pre-training. 'WS' means adding weak supervision as pre-training.

	Micro F1	Macro F1	TA
TextCNN	70.69	56.95	45.50
TextCNN+WS	72.59	58.65	**53.30**
TextCNN+B	71.22	59.23	47.10
TextCNN+B+WS	**75.56**	**67.17**	52.80

Table 3: The results of TextCNN with Bert as fixed feature. 'WS' means weak supervision, 'B' means adding Bert as fixed feature, and 'TA' represents turn accuracy.

	Micro F1	Macro F1	Turn Accuracy
TextCNN+Q	53.64	54.32	27.50
TextCNN+R	72.59	58.65	53.30

Table 4: The results of TextCNN with different pre-training weak supervision. 'Q' means that weak supervision comes from medical concepts occurring in queries, while 'R' means that weak supervision comes from medical concepts occurring in responses.

posed weak supervision methods. In this work, Bert is used as fixed feature to get a better performance. TextCNN is selected as the testbed. Table 3 summarizes the results with Bert as fixed feature.

The results show that 1)Bert shows promising results on the medical slot filling task; 2) based on Bert, the proposed weak supervision method can make a great improvement of the model.

Training Data Amount Analysis We try to analyze the influences of the training data amount when unlabeled data is fixed. As shown in Figure 4, the red line shows the performance of the TextCNN classifier with weak supervision as pre-training, and the green line shows the performance of the raw TextCNN classifier on 500, 600, 700, 800, 900 training data amount respectively.

The trend of the lines shows that 1) when the labeled data amount is small, weak supervision leads to more contributions; 2) weak supervision with unlabeled data always helps the model to get better performance; 3) a small amount of well-annotated data with weak supervision can get a strong performance whose requirement for labeled data is relatively small. The results indicate that weak supervision from responses is potential for MSL which is consistent with our expectations to reduce the requirement of well-labeled data.

Pre-training Unlabeled Data Amount Analysis We try to analyze the influences of the unlabeled data amount for pre-training when labeled data is fixed. We select 0, 10000, 20000, 30000, 40000 and 50000 unlabeled data for this experiment, and 0 unlabeled data means no pre-training. As shown in Figure 5, the red line shows the performance of the TextCNN classifier with weak supervision as pre-training.

The line chart shows that 1) weak supervision with unlabeled data always helps the model to get better performance; 2) excessive unlabeled data for pre-training will not always help further improve performance.

Weak Supervision Source Analysis We aim to analyze the different weak supervision sources. We try to employ those medical concepts occurring in queries as the weak supervision source. TextCNN is selected as the testbed. Table 4 summarizes the results of different weak supervision sources. From the table, we can find that using medical concepts in responses is significantly better than using those in queries. The main reason is that medical concepts in queries are in a colloquial language which is hard to extract medical concepts by matching, thus using concepts in queries contain too many errors which may lead to low performance. And doctors always retell patients' health condition with formal concepts that are easy to access which is more suitable for weak supervision source.

Case Study for Label-embedding Attentive Model Figure 6 shows three examples for the label-embedding attentive model where gray value is proportional to the proba-

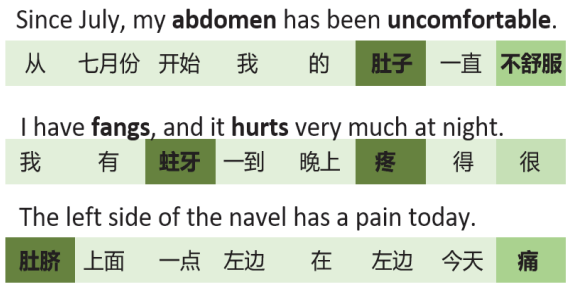

Figure 6: Three cases of the label-embedding attentive model.

bility value. In the three sentences, a single medical concept is expressed with several keywords that are dispersed in different parts of the sentences. The cases show that the label-embedding attentive model can help models get more attention to these scattered keywords which can help relieve the scattered keyword issue to some extent and is in favor of following classifiers.

Related Work

Our work is closely relative with SLU from unaligned data, and medical concept normalization from user generate texts.

SLU from Unaligned Data

Sequence labeling discriminative models such as CRFs and sequence neural networks (Yao et al. 2013; Mesnil et al. 2015; Kurata et al. 2016; Jaech, Heck, and Ostendorf 2016; Hakkani-Tür et al. 2016) have been widely used for SLU. Traditional SLU is defined as sequence labeling problem which needs word-level semantic annotations. However, utterances in real life are always unaligned. This issue has attracted attentions recently years. (Zhou and He 2011) proposed learning CRFs from unaligned with manually tuned lexical or syntactic features. (Barahona et al. 2016) presented a two-step multi-label classifier for semantic decoding in SDS. (Zhao and Feng 2018) used a generative neural network model for slot filling based on a sequence-to-sequence model together with a pointer network to solve the problem of data sparsity and out-of-vocabulary (OOV) caused by unknown slot values.

The unaligned issue also exits in medical conversations. The gap between colloquial language and formal medical concept makes medical dialogue data are unaligned where slot-values such as symptom expressions are scattered in sentences and different from formal medical concept which is hard to annotate the data in word-level with high quality. Based on the above analysis, we follow the approach in (Barahona et al. 2016) and employ multi-label classification models for conducting medical slot filling.

Medical Concept Normalization from User Generate Texts

Medical concept normalization for user-generated texts aims at mapping a health condition phrase described in colloquial language to a medical concept in standard ontologies (Zhao et al. 2019).

Traditional approaches used for medical concept normalization include lexicon-based string matching, heuristic string matching, and rule-based text mapping to a set of pre-defined variants of terms. Recently, deep learning has been widely used for this task (Limsopatham and Collier 2015; 2016; Lee et al. 2017; Tutubalina et al. 2018). (Limsopatham and Collier 2015) treated the mapping as a machine translation (MT) approach where a social medical phrase is translated into a formal medical concept. Recently, (Limsopatham and Collier 2016; Lee et al. 2017; Tutubalina et al. 2018) use multiple deep learning architectures such as CNN and RNN with input word embeddings trained on various clinical domain-specific knowledge sources to map a health condition described in the colloquial language to a medical concept defined in standard clinical terminologies.

They treat the task as a multi-class classification problem which is highly similar to our task. The difference between the two tasks is that medical concept normalization aims to parse a medical concept phrase, while the input of our task is a natural language sentence. Besides, there is always only one medical concept in the medical concept normalization task instead of multi-targets in medical slot filling.

Conclusion

In this paper, we study medical slot filling and propose to use label-embedding attentive model and weak supervision from responses. Experiments show that our method significantly improves the model's performance. Comparison analysis gives empirically guarantee for our method. In the future, we will further improve the data and try more ways to improve the performance of medical slot filling from the aspect of weak supervision from responses.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions. This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61976072, 61632011 and 61772153.

References

Barahona, L. M. R.; Gasic, M.; Mrkšić, N.; Su, P.-H.; Ultes, S.; Wen, T.-H.; and Young, S. 2016. Exploiting sentence and context representations in deep neural models for spoken language understanding. *arXiv preprint arXiv:1610.04120*.

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*.

Guo, Q.; Qiu, X.; Liu, P.; Shao, Y.; Xue, X.; and Zhang, Z. 2019. Star-transformer. 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)*, 1315–1325.

Hakkani-Tür, D.; Tür, G.; Celikyilmaz, A.; Chen, Y.-N.; Gao, J.; Deng, L.; and Wang, Y.-Y. 2016. Multi-domain

joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, 715–719.

Jaech, A.; Heck, L.; and Ostendorf, M. 2016. Domain adaptation of recurrent neural networks for natural language understanding. *arXiv preprint arXiv:1604.00117*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurata, G.; Xiang, B.; Zhou, B.; and Yu, M. 2016. Leveraging sentence-level information with encoder lstm for semantic slot filling. *arXiv preprint arXiv:1601.01530*.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Lee, K.; Hasan, S. A.; Farri, O.; Choudhary, A.; and Agrawal, A. 2017. Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 462–469. IEEE.

Li, X.; Chen, Y.-N.; Li, L.; Gao, J.; and Celikyilmaz, A. 2017. Investigation of language understanding impact for reinforcement learning based dialogue systems. *arXiv preprint arXiv:1703.07055*.

Limsopatham, N., and Collier, N. 2015. Adapting phrase-based machine translation to normalise medical terms in social media messages. *arXiv preprint arXiv:1508.02285*.

Limsopatham, N., and Collier, N. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1014–1023.

Liu, P.; Qiu, X.; and Huang, X. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Mesnil, G.; Dauphin, Y.; Yao, K.; Bengio, Y.; Deng, L.; Hakkani-Tur, D.; He, X.; Heck, L.; Tur, G.; Yu, D.; et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM* 23(3):530–539.

Mori, R. D. 1998. Spoken dialogues with computers, ser. signal processing and its applications. *Academic Press*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Qiao, C.; Huang, B.; Niu, G.; Li, D.; Dong, D.; He, W.; Yu, D.; and Wu, H. 2018. A new method of region embedding for text classification. In *ICLR*.

Smith, R., and Hipp, D. 1994. Spoken natural language dialog systems: A practical approach. *Oxford University Press*.

Tur, G., and De Mori, R. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Tutubalina, E.; Miftahutdinov, Z.; Nikolenko, S.; and Malykh, V. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics* 84:93–102.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*.

Wang, S.; Che, W.; and Liu, T. 2016. A neural attention model for disfluency detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 278–287.

Wang, B. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2311–2320.

Wei, Z.; Liu, Q.; Peng, B.; Tou, H.; Chen, T.; Huang, X.; Wong, K.-F.; and Dai, X. 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)*, 201–207.

Xu, L.; Zhou, Q.; Gong, K.; Liang, X.; Tang, J.; and Lin, L. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *AAAI*.

Yao, K.; Zweig, G.; Hwang, M.-Y.; Shi, Y.; and Yu, D. 2013. Recurrent neural networks for language understanding. In *Interspeech*, 2524–2528.

Yin, W., and Schütze, H. 2018. Attentive convolution: Equipping cnns with rnn-style attention mechanisms. *Transactions of the Association for Computational Linguistics* 6:687–702.

Young, S.; Gašić, M.; Thomson, B.; and Williams, J. D. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. of the IEEE* 101(5):1160–1179.

Zhao, L., and Feng, Z. 2018. Improving slot filling in spoken language understanding with joint pointer and attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 426–431.

Zhao, S.; Liu, T.; Zhao, S.; and Wang, F. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 817–824.

Zhou, D., and He, Y. 2011. Learning conditional random fields from unaligned data for natural language understanding. In *European Conference on Information Retrieval*, 283–288. Springer.

Zue, V., and Glass, J. 2000. Conversational interfaces: Advances and challenges. *Proc IEEE, vol. 88, no. 8, pp. 1166–1180*.