ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414073

KNOWLEDGE-BASED CHAT DETECTION WITH FALSE MENTION DISCRIMINATION

Wei Liu, Peijie Huang[*]*, Dongzhu Liang, Zihao Zhou*

College of Mathematics and Informatics, South China Agricultural University, China

liuliulz09@stu.scau.edu.cn, pjhuang@scau.edu.cn, {liang_dz, zz}@stu.scau.edu.cn

ABSTRACT

Chat detection is critical for recently emerged personal intelligent assistants (PIA), which can be seen as a hybrid of domain-specific task-oriented spoken dialogue systems and open-domain non-task-oriented ones. Recent advances have attempted to utilize external domain knowledge to enhance utterance semantics understanding and can contribute to chat detection. However, it also inevitably introduces false mention (i.e., token spans being misidentified as entity mentions) in `Chat` utterances, causing performance to degrade. To deal with this issue, this paper proposes a new model for knowledge-based chat detection with false mention discrimination (FMD-KChat). A two-stage pipeline is adopted, which contains an additional neural network-based classifier in the first stage for distinguishing the false mentions and a feature fusion gate in the chat detection stage for combining the contextual representation with the external knowledge feature based on the false mention discrimination probability. Experiments on the SMP-ECDT benchmark corpus show the well performance of the proposed model.

Index Terms— chat detection, personal intelligent assistants, false mention discrimination, knowledge-based model

1. INTRODUCTION

Recently emerged personal intelligent assistants (PIA) on smartphones and home electronics (e.g., Siri and Alexa) typically perform various tasks (e.g., Web search, weather checking, and alarm setting) while being able to have chats with users [1]. They can be seen as a novel hybrid of domain-specific task-oriented spoken dialogue systems (SDS) [2] and open-domain non-task-oriented SDS [3]. To realize such hybrid SDS, we have to determine whether or not a user is going to have a chat with the system. For example, if a user says "What is your hobby?" it is considered that she is going to have a chat with the system. On the other hand, if she says "Set an alarm at 8 o'clock," she is probably trying to operate her smartphone. This task is referred as chat detection and is treated as a binary classification problem [1].

Recent advances on nature language understanding (NLU) are overwhelmingly contributed by deep learning techniques [4–6], which have taken the state-of-the-art of NLU to a new level. The core architecture of the chat detection models that using a binary neural classifier is shown in Figure 1(a). However, fully data-driven neural classification models [7–11] tend to have limitation on inadequate expression of domain entity mentions (i.e., token spans representing entities in a user utterance) for lack of domain knowledge and result in insufficient semantics learning of entity mentions.

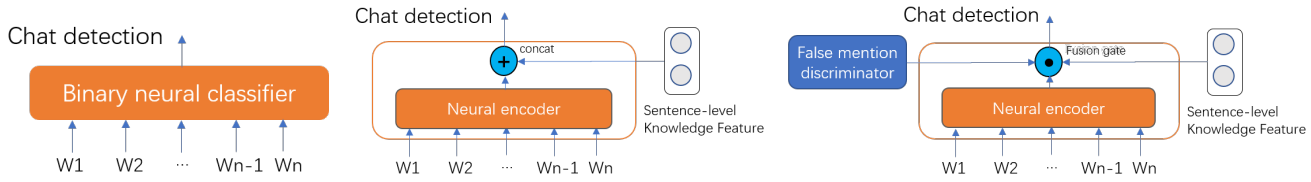
ID	Utterance	Label
S1	Launch **Benben Reading**.	`NonChat`
S2	I don't want to say. **I am in a bad mood**.	`Chat`

Table 1: *Example of utterances with true or false mention.* ***Bold*** *for entity mentions.*

Several studies have been proposed to integrate knowledge information to learn a knowledge-based sentence representation on text classification [12], speech understanding [13], and utterance domain classification (UDC) [14, 15]. In this paper, we follow this line of research and propose a base model of knowledge-based chat detection (KChat). The core architecture of KChat is shown in Figure 1(b), in which the knowledge feature is concatenated with the contextual representation in the binary neural classifier. The knowledge feature used to enhance utterance semantics understanding can derive from mention-type pairs retrieved from external knowledge bases (KBs). S1 in Table I shows an examples of retrieved entity mention in `NonChat` utterance that will benefit the chat detection model. However, some false mentions are also inevitably introduced in `Chat` utterances. S2 in Table 1 shows the false mention problem, where the music type mention "I am in a bad mood" in the utterance is mistakenly retrieved from KB. As a result, the performance of knowledge-based chat detection model will be degraded by penetrating false knowledge into neural classifier. Previous work [15] proposes a knowledge-gated mechanism to control the weights for external knowledge. However, it only uses the representation of external knowledge to realize gate mechanism, which does not truly consider how to discriminate whether the external knowledge in an utterance should be introduced and is lack of interpretability.

To deal with the above challenge, we propose a knowledge-based chat detection model with false mention discrimination (FMD-KChat). The overall principle can be seen in Figure

[*] Corresponding author.



(a) Existing framework (b) KChat framework (c) FMD-KChat framework

Fig. 1: (a) The core architecture of the chat detection models using a binary neural classifier. (b) The base model of knowledge-based chat detection (KChat). (c) Our idea of knowledge-based chat detection with false mention discrimination (FMD-KChat).

1(c). A false mention discrimination (FMD) is proposed to discriminate whether the external knowledge in an utterance should be utilized or not. Note that our FMD's supervised labels can be extracted from chat detection task without extra labeling effort. Moreover, we propose a two-staged pipeline to integrate the false mention discrimination into KChat. An additional neural network-based classifier is applied to distinguish the false mentions from the retrieved ones. Then in the chat detection stage, a feature fusion gate is used to combine the contextual representation with the external knowledge feature based on the false mention discrimination probability instead of incorporating external knowledge directly. Experimental results on public benchmark corpus SMP-ECDT show that the proposed FMD-KChat model achieves significant improvements over the compared models in chat detection, and larger margin of accuracy improvement over the base KChat in the detection of `Chat` utterances with false mention.

2. THE PROPOSED APPROACH

2.1. KChat Model

Knowledge Retrieval. Knowledge retrieval is a process that retrieving external knowledge from knowledge bases (KBs). In detail, we follow the previous work [14, 15], obtaining external knowledge (i.e. entity types [16], a set of words which are semantic categories to which entities belong) in utterances from CN-Probase [17], which is a large-scale Chinese taxonomy for entity types retrieval. Besides, for some missing entities in KBs, we follow previous work [14, 15], adopting some other reliable sources (e.g. Baidu Baike, QQ music) as supplement to improve the coverage of external knowledge.

To be more specific, for every utterance $\boldsymbol{u}$, we rely on KBs to retrieve a type set $\mathcal{T}$ with respect to the entity mention set $\mathcal{M}$ for the utterance. Formally, given an utterance $\boldsymbol{u} = [w_1, w_2, ..., w_n]$, we match each word w_i to a named entity e_i in KBs, if the word w_i has an corresponding entity e_i in KBs. After that, we obtain the corresponding entity types $\boldsymbol{t_i}$ from KBs based on the named entity e_i. According to every $\boldsymbol{t_i}$ in an utterance, we add all entity types inside them into $\mathcal{T}$, getting the entity type set for the utterance.

Knowledge Incorporation. After getting the $\mathcal{T} = \{\tau_1, \tau_2, \tau_3, ..., \tau_m\}$, where τ_j is the entity type and m is the number of retrieved entity types, we vectorize every τ_j by using a BERT encoder, which has been pre-trained by a large amount of data, and so it can well encode the external knowledge as embeddings with abundant prior knowledge. Notice that since entity types are a set of words, they can be encoded by BERT directly. Furthermore, since there may be multiple elements in $\mathcal{T}$, we sum these embeddings encoded by BERT and scale them to control the size of the value. Finally, we concatenate external knowledge vector $\boldsymbol{d}$ with the output of utterance encoder $\theta(w)$. The process can be shown by:

$$\boldsymbol{\psi_j} = BERT(\tau_j) \tag{1}$$

$$\boldsymbol{d} = \frac{\sum_{j=1}^{m} \boldsymbol{\psi_j}}{m} \tag{2}$$

$$\boldsymbol{o} = \text{concat}[\theta(w), \boldsymbol{d}] \tag{3}$$

where the utterance encoder is an attentive BiLSTM model [9]. We leverage the BiLSTM model to learn the hidden state for each time step which contains sequential information for the utterance and implements the soft attention mechanism to give some key words larger weights. Here we denote $\boldsymbol{\phi^E(w_i)}$ as the word embedding of w_i, which is got through BERT, and $\boldsymbol{w_a^T}$ and $\boldsymbol{W_s}$ are trainable parameters. The utterance encoder can be given by, where:

$$\boldsymbol{h_i^w} = \textbf{BiLSTM}^w(\boldsymbol{\phi^E(w_i)}, \boldsymbol{h_{i-1}^w}) \tag{4}$$

$$\theta(w) = \sum_{i=1}^{n} \alpha_i \boldsymbol{h_i} \propto \exp(\boldsymbol{w_a^T} tanh(\boldsymbol{W_s h_i})) \tag{5}$$

2.2. False Mention Discrimination

In KChat architecture, the mention-type pairs knowledge are incorporated for the `NonChat` utterances to help identifying some specific instructions. However, those mentions will be also inevitably introduced to `Chat` utterances, doing harm to the chat detection task. In FMD, we define mentions in `Chat` utterances as false mentions and mentions in `NonChat` utterances as true mentions. The goal of FMD is to discriminate whether the mentions in an utterance are proper for it. Intuitively, those utterances that have true mentions are more likely to have similar context of mentions with each other. For example, utterance "Play Jackie Chan's action movie" and utterance "Play Chris Evans's science fiction movie" are all `NonChat` utterances and have true mentions for them (i.e.

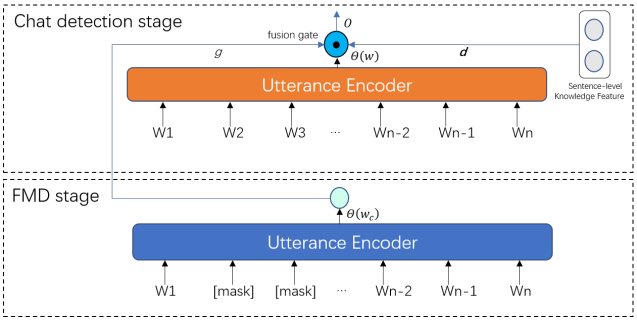


Fig. 2: The FMD-KChat architecture

Jackie Chan and Chris Evans), and the context of their mentions are very similar. However, in the utterance "Who is the wife of Jackie Chan?", the user only wants to have a chat with PIA, so the PIA system should not integrate the external knowledge respected to the mention (i.e. Jackie Chan) into chat detection model, and the context of the mention is different from above two utterances. Therefore, we take the context of mentions in an utterance as input to discriminate whether the mentions in it are true mentions. In detail, we mask all mentions in an utterance and define the FMD task as a binary classification task, labeling the context of true mentions as 1 and the context of false mentions as 0.

2.3. FMD-KChat

In terms of the mechanism that integrating FMD into KChat, we hope to get two strong modules to finish two tasks, FMD and chat detection. However, these two modules have mutual effect with each other since one module takes the other's output as input. Specifically, during training, the wrong output of KChat will wrongly indicate FMD to update its parameters when the output of FMD is correct. And the wrong output of FMD will also degrade the performance of KChat. So, we can find it is hard for these two modules to converge to optimal together. Therefore, a two-staged pipeline is designed to integrate FMD into chat detection model explicitly, which can contribute to train two strong modules. In two-staged mechanism, we divide the model into two stages, the first of which is FMD stage and the second is the chat detection stage. Firstly, we train the FMD model and the KChat model separately. Then we link these two models as a pipeline. The completed FMD-KChat architecture is shown in Figure 2.

FMD Stage. In the FMD stage, we leverage the aforementioned attentive BiLSTM model, whose input is the context of mentions in an utterance, to discriminate the false mentions and judge whether the utterance should incorporate external knowledge. And the output in FMD stage will be sent to the chat detection stage, which will be taken as an input of fusion gate. The process of FMD stage is given by:

$$g = sigmoid(\hat{\theta}(w_c)) \tag{6}$$

where $\hat{\theta}(w_c)$ denotes the context of mentions in an utterance, which has been encoded, and g is the output of FMD stage.

Chat Detection Stage. In chat detection stage we take the utterance as input and encode it by attentive BiLSTM. The next step is the difference between KChat and FMD-KChat. In KChat, we just concatenate the external knowledge feature with contextual representation. But in FMD-KChat, we add the output of FMD stage as an input to form a fusion gate, helping the contextual representation concatenate with the reliable external knowledge. The fusion gate is given by:

$$\hat{\boldsymbol{d}} = g\boldsymbol{d} \tag{7}$$

$$o = concat[\theta(w), \hat{\boldsymbol{d}}\,] \tag{8}$$

where $\boldsymbol{d}$ denotes the sentence-level knowledge feature mentioned above, $\hat{\boldsymbol{d}}$ denotes the knowledge feature after multiplied by g and $\theta(w)$ is the contextual representation of the utterance, and o is the output of the fusion gate.

3. EXPERIMENTS

3.1. Experiment Settings

Dataset. We evaluate our models on the public benchmark corpus of SMP-ECDT [18, 19] provided by iFLYTEK Corporation. The dataset consists of the two top categories *chat* and *task-oriented*. In our evaluation, we take utterances in the top category *chat* as `Chat` utterances and take utterances in the top category *task-oriented* as our `Nonchat` utterances. The dataset contains 3736 training data (3076 `Chat` and 660 `Nonchat`) and 4528 test data items (2594 `Chat` and 1934 `Nonchat`), which are all single-turn short utterances.

Training Details. We conduct all experiments with word embeddings provided by BERT [20] (BERT-Base, Chinese[1]). To explore the proper hyper-parameters, we performed 10-fold cross validation. The earlystop strategy is employed to terminate training when the loss on validation data does not decrease. To avoid overfitting, dropout [21] is used. For optimizer, we use Adam [22] along with learning rate 0.001. The hidden layer size of BiLSTM is set to 50 and the batch size is set to 64. The loss functions of both FMD and chat detection are binary cross entropy. All the results are the mean of 10 independent experiments. For the training of FMD, since the `Chat` utterances which have mentions in training dataset usually take a relatively small proportion and the other `Chat` utterances have very similar contextual features as them. We add `Chat` utterances without mention as a complementary training data here, and give a label of 0 to them.

Compared Methods. We compare our proposed FMD-KChat with suitable baselines: (1) **BERT Fine tune**: A pre-trained language model [20], and it is demonstrated to be effective in ATIS intent classification [23]. (2) **BiLSTM**: A classic baseline that widely used for spoken language understanding [6]. (3) **BiLSTM Hard ATT**: A standard BiLSTM

[1]https://github.com/google-research/bert#pre-trained-models

Models	F1 ↑	EER ↓	AUC ↑
BERT Fine tune	0.796	0.137	0.939
BiLSTM	0.798	0.150	0.916
BiLSTM Hard ATT	0.801	0.133	0.940
BiLSTM Soft ATT	0.810	0.130	0.940
KChat	0.826	0.118	0.946
KChat Gate	0.829	0.122	0.946
FMD-KChat	**0.838**	**0.116**	**0.947**

Table 2: *Chat detection performance. The notation ↑ means higher values are better, and ↓ means lower values are better.*

with hard attention mechanism [24] is employed for chat detection. (4) **BiLSTM Soft ATT**: A standard BiLSTM with soft attention mechanism, which has been successfully applied in utterance classification task [9, 10]. (5) **KChat**: a base model of knowledge-based chat detection, which follows previous works [12–15] to concatenate the knowledge feature with the contextual representation in the attentive BiLSTM. (6) **KChat Gate**: a base model of knowledge-gated chat detection, which has been proposed to control the information of external knowledge to flow into the UDC model [15].

Metrics. We consider various metrics for evaluating: F1, EER, AUC and accuracy, where equal error rate (EER) is a commonly used out-of-domain (OOD) detection metric [25], which is the error rate when the confidence threshold is located where false acceptance rate (FAR) is equivalent to false rejection rate (FRR).

3.2. Result and Analysis

Table 2 shows the result on test data of our proposed model and competing approaches. As is showed Table 2, for the metrics of F1, our model significantly outperforms the compared methods. Compared to KChat and KChat gate, performance gain of FMD-KChat reaches 1.2% and 0.9%. For ERR, FMD-KChat also outperforms KChat and KChat gate. And for AUC, our FMD-KChat outperforms slightly better than above two models. Besides, compared to the data-driven models, both of the knowledge-based models achieve larger margin of performance improvement.

Figure 3 shows the ROC curves obtained from these models. KChat model and KChat gate surpasses all the fully data-driven baselines, and FMD-KChat further improves the performance of chat detection. Note that the performance improvement of the FMD-KChat model over the two models is consistent but relatively small. This is partly because when the FPR is too large or too small, FMD has little effect on chat detection.

We further investigate performances on false mention discrimination. The comparison of FMD-KChat and KChat on the detection of `Chat` utterances with false mentions is shown in Table 3. We use BiLSTM Soft ATT model as reference. It can be found that KChat fails to detect `Chat` utterance with false mention, and gets an accuracy of only 0.416. Compared to KChat, FMD-KChat achieves significantly improvement on the detection of `Chat` utterances with false mentions. The accuracy of FMD-KChat which outperforms BiLSTM Soft ATT demonstrates that the proposed FMD-KChat has successfully avoided the impact on the inevitably introduced false mention in knowledge-based models.

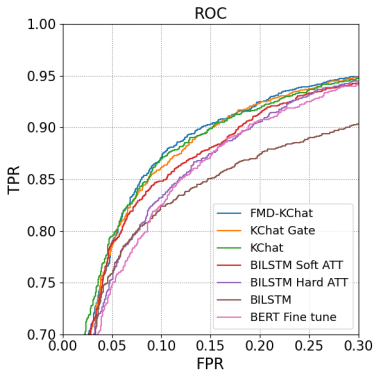


Fig. 3: ROC curves for different models. The upper-left corner of the ROC curves is zoomed in to facilitate a clearer view.

Models	Accuracy
KChat	0.416
FMD-KChat	**0.549**
BiLSTM Soft ATT	0.541

Table 3: *Performance on the detection of `Chat` utterances with false mention.*

4. CONCLUSIONS

In this paper, we have proposed a knowledge-based chat detection model with false mention discrimination (FMD-KChat), which to the best of our knowledge is the first knowledge-based neural model that incorporates with false mention discrimination into chat detection. In FMD-KChat, we design a two-staged pipeline, which contains an additional neural network-based classifier in the first stage for distinguishing the false mentions and a feature fusion gate in the chat detection stage for combining the contextual representation with the discriminated external knowledge feature. Experimental results on SMP-ECDT benchmark corpus demonstrate the proposed model achieves significant improvements over the compared models in chat detection, and larger margin of accuracy improvement over the base KChat in the `Chat` utterances with false mention.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 71472068) and the Innovation Training Project for College Students of Guangdong Province (No. 201910564164).

References

[1] S. Akasaki and N. Kaji, "Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems," in *ACL 2017*, pp. 1308–1319.

[2] J. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 393–422, 2007.

[3] R. Wallace, "The anatomy of alice," in *Parsing the Turing Test*, pp. 181–210. Springer, 2009.

[4] G. Tür, L. Deng, D. Hakkani-Tür, and X. He, "Towards deeper understanding: Deep convex networks for semantic utterance classification," in *ICASSP 2012*, pp. 5045–5048.

[5] P. Zhou, W. Shi, J. Tian, et al., "Attention-based bidirectional long short-term memory networks for relation classification," in *ACL 2016*, pp. 207–212.

[6] N. Vu, P. Gupta, H. Adel, et al., "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *ICASSP 2016*, pp. 6060–6064.

[7] P. Xu and R. Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," in *ICASSP 2014*, pp. 136–140.

[8] S. V. Ravuri and A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," in *INTERSPEECH 2015*, pp. 135–139.

[9] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *INTERSPEECH 2016*, pp. 685–689.

[10] Y. Kim, D. Kim, A. Kumar, et al., "Efficient large-scale neural domain classification with personalized attention," in *ACL 2018*, pp. 2214–2224.

[11] J. Lee, R. Sarikaya, and Y. Kim, "Locale-agnostic universal domain classification model in spoken language understanding," in *NAACL-HLT 2019*, pp. 9–15.

[12] J. Wang, Z. Wang, D. Zhang, et al., "Combining knowledge with deep convolutional neural networks for short text classification," in *IJCAI 2017*, pp. 2915–2921.

[13] M. Graja, M. Jaoua, and L. Belguith, "Statistical framework with knowledge base integration for robust speech understanding of the tunisian dialect," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2311–2321, 2015.

[14] Y. He, P. Huang, Z. Du, et al., "Distant supervision based utterance domain classification with domain-specific NER," *J. Chin. Inf. Process.*, vol. 34, no. 5, pp. 10–18, 2020.

[15] Z. Du, P. Huang, Y. He, et al., "A knowledge-gated mechanism for utterance domain classification," in *NLPCC 2019*, pp. 142–154.

[16] Q. Wang, Z. Mao, B. Wang, et al., "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, 2017.

[17] B. Xu, Y. Xu, J. Liang, et al., "CN-DBpedia: A never-ending chinese knowledge extraction system," in *IEA/AIE 2018*, pp. 428–438.

[18] W. Zhang, Z. Chen, W. Che**,** et al., "The first evaluation of chinese human-computer dialogue technology," *CoRR*, vol. abs/1709.10217, 2017.

[19] Z. Zhao, W. Zhang, W. Che, et al., "An evaluation of chinese human-computer dialogue technology," *Data Intell.*, vol. 1, no. 2, pp. 187–200, 2019.

[20] J. Devlin, M. Chang, K. Lee, et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, pp. 4171–4186.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*.

[23] L. Qin, W. Che, Y. Li, et al., "A stack-propagation framework with token-level intent detection for spoken language understanding," in *EMNLP-IJCNLP 2019*, pp. 2078–2087.

[24] S. Shankar, S. Garg, and S. Sarawagi, "Surprisingly easy hard-attention for sequence to sequence learning," in *EMNLP-IJCNLP 2018*, pp. 640–645.

[25] I. Lane , T. Kawahara , T. Matsui, et al., "Out-of-domain utterance detection using classification confidences of multiple topics," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 150–161, 2007.