# KILDST: Effective Knowledge-Integrated Learning for Dialogue State Tracking using Gazetteer and Speaker Information

**Hyungtak Choi[1], Hyeonmok Ko[1], Gurpreet Kaur[2], Lohith Ravuru[1], Kiranmayi Gandikota[2], Manisha Jhawar[2], Simma Dharani[2], and Pranamya Patil [2]**
[1]Samsung Research, Seoul, South Korea
[2]Samsung Research India Bangalore, Bangalore, India
{ht777.choi,felix.ko,k.gurpreet,loki.ravuru,k.gandikota m.jhawar,s.dharani,pran.patil}@samsung.com

## Abstract

Dialogue State Tracking (DST) is core research in dialogue systems and has received much attention. In addition, it is necessary to define a new problem that can deal with dialogue between users as a step toward the conversational AI that extracts and recommends information from the dialogue between users. So, we introduce a new task – DST from dialogue between users about scheduling an event (DST-USERS). The DST-USERS task is much more challenging since it requires the model to understand and track dialogue states in the dialogue between users and to understand who suggested the schedule and who agreed to the proposed schedule. To facilitate DST-USERS research, we develop dialogue datasets between users that plan a schedule. The annotated slot values which need to be extracted in the dialogue are date, time, and location. Previous approaches, such as Machine Reading Comprehension (MRC) and traditional DST techniques, have not achieved good results in our extensive evaluations. By adopting the knowledge-integrated learning method, we achieve exceptional results. The proposed model architecture combines gazetteer features and speaker information efficiently. Our evaluations of the dialogue datasets between users that plan a schedule show that our model outperforms the baseline model.

## 1 Introduction

DST is a task to determine the final dialogue states by continuously tracking the dialogue between the user and the system. It is a challenging and essential task because it can be applied to many real-world applications, such as voice assistant systems. Many approaches have been proposed to solve the DST problem (Lin et al., 2020; Kim et al., 2019). MinTL (Lin et al., 2020) framework adopts plug-and-play architecture to the pre-trained Seq2Seq model and can learn DST and NLU at the same time. SOM-DST (Kim et al., 2019) updates the dialogue state in two steps: state operation prediction such as ADD, UPDATE and DELETE operation and dialogue state updater. The approaches have the advantage of improving performance; however, the error gets propagated to the dialogue state tracking phase of the model if an error occurs in the prediction of state operation. Meanwhile, in the past few years, many innovative models in the field of MRC. Among them, the mainstream approach formalizes reading comprehension to the extent of extracting answers from a given text (Seo et al., 2016; Wang and Jiang, 2016; Xiong et al., 2017; Joshi et al., 2017; Dunn et al., 2017; Shen et al., 2017; Wang et al., 2017a,b; Tan et al., 2017; Devlin et al., 2018; Liu et al., 2019). There have been various attempts to apply this promising MRC technique to the DST field, and it has shown remarkable performance (Gao et al., 2019, 2020).

Despite much research on DST, there are still some challenging problems to be solved. Newly-coined words and unseen words pose problems (Bernier-Colborne and Langlais, 2020). Some slots, such as specific store names and movie names, are not general noun phrases and are more complicated to recognize (Ashwini and Choi, 2014; Jayarao et al., 2018). To overcome this problem, approaches for integrating external knowledge, such as gazetteer information or knowledge based on neural architectures, have been highlighted and studied again. One-hot vectors are typically used as inputs to the gazetteer encoder and are then concatenated with word representations. However, for some problems, simply integrating gazetteer information will not improve or reduce performance for some slots (Meng et al., 2021).

In this paper, we conduct a new and challenging task that tracks the dialogue state between two users rather than the dialogue state tracking in the dialogue between users and systems, like in previous studies. Ultimately, this model can



*Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 60 - 66
December 15-18, 2022 ©2022 Association for Computational Linguistics

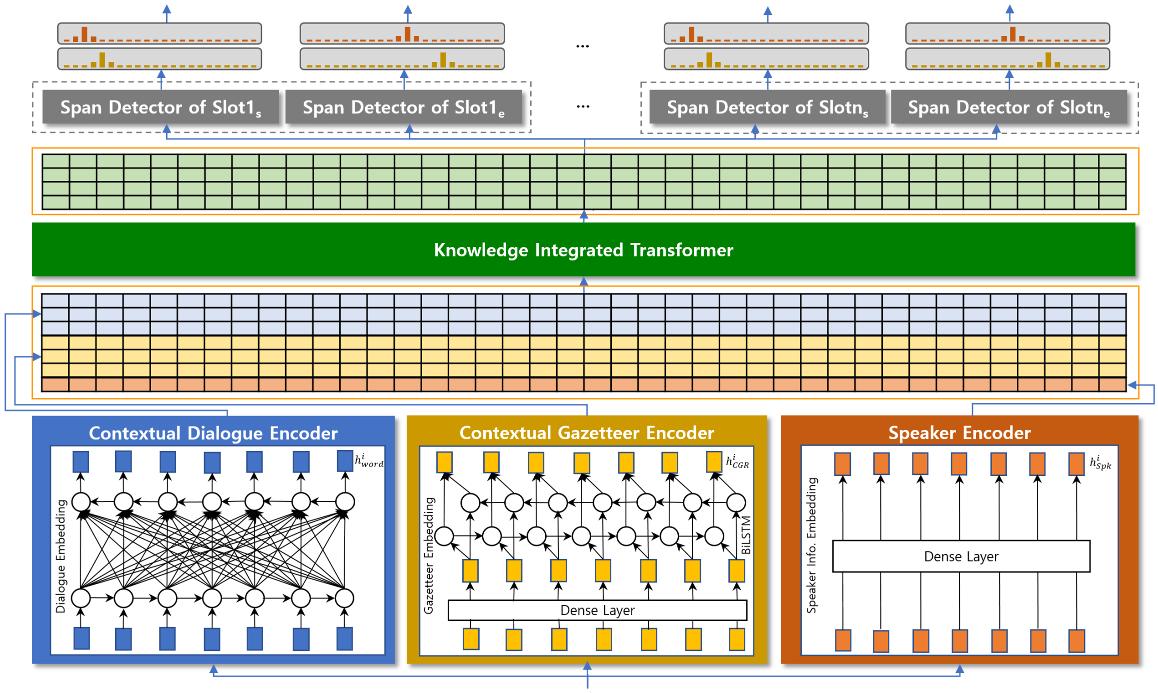


Figure 1: Illustration of our proposed KILDST architecture. The model consists of 3 different sub-encoders, 1) contextual dialogue encoder for encoding dialogue, 2) contextual gazetteer encoder for encoding domain knowledge, and 3) speaker encoder for encoding speaker information. Our proposed knowledge-integrated learning method efficiently consolidates information from multiple encoders and extracts the schedule using span detectors.

be applied to a recommendation service on a smartphone that extracts and recommends a schedule for WhatsApp, WeChat, Kakaotalk, and Telegram. Even if ambiguous slots are mentioned in the dialogue, the system cannot request explicit confirmation from the user, so it is more difficult to extract accurate information. Several studies use a gazetteer to improve performance (Song et al., 2020) to solve this problem. Especially, GEMNET (Meng et al., 2021) further enhances performance with effective gated gazetteer representations. We propose a novel architecture that effectively learns the gazetteer and speaker information to extract slots more accurately.
The contributions of our work are as follows:

- We propose a novel model based on the Transformer that efficiently consolidates the gazetteer knowledge and speaker information to improve the extraction performance of difficult words such as newly coined words and abbreviations used in dialogue between users.
- We propose a method of applying the GEMNET that efficiently utilizes the gazetteer knowledge in our integrated Transformer.
- We propose a method of understanding speaker information through speaker embeddings to know who suggested the schedule and who agreed.
- We evaluate our model for new dialogue datasets that contains a dialogue between users regarding scheduling an event and show that efficient use of gazetteer knowledge and speaker information improves performance.

## 2 Proposed Model

We propose an effective Knowledge-Integrated Learning method for Dialogue State Tracking using Gazetteer and Speaker Information (KILDST). As shown in Figure 1, the proposed model consists of 3 different sub-encoders, 1) a contextual dialogue encoder for encoding dialogue, 2) a contextual gazetteer encoder for encoding domain knowledge, and 3) a speaker encoder for encoding speaker information. Our proposed knowledge-integrated learning method efficiently consolidates information from multiple encoders and extracts the slots related to a schedule, such as a date, time, and location using span detectors.

### 2.1 Contextual Dialogue Encoder

A contextual dialogue encoder encodes the input dialogue based on a pre-trained BERT (Devlin et al., 2018). We use syllable units for tokens because this outperforms other token units in our experiments using Korean datasets.

$$h_{word} = BERT(D) \quad (1)$$

In the above formula, D refers to an index list represented by the dialogue text tokenized in syllable units in the form of "$[CLS]$ user A's dialogue sentence $[TURN]$ user B's dialogue sentence $[TURN]$". "$[TURN]$" is a special token for distinguishing the dialogue turn between users.

### 2.2 Contextual Gazetteer Encoder

Gazetteer information can be provided directly as an input feature, but more is needed and sparse. We use linear projection to obtain a dense representation that captures interactions between multiple matches per syllable unit. We encode Contextual Gazetteer Representations (CGR) with gazetteer information. BiLSTM is then applied to contextualize this representation (Meng et al., 2021).

$$\begin{aligned} h^i_{forward} &= LSTM(h^{i-1}_{forward}, h^i_{gaz}) \\ h^i_{backward} &= LSTM(h^{i+1}_{backward}, h^i_{gaz}) \\ h^i_{cgr} &= [h^i_{forward}, h^i_{backward}] \end{aligned} \quad (2)$$

### 2.3 Speaker Encoder

Our task differs from general dialogue system tasks between the user and the system. Since the dialogue is between two users and not between a user and a system, it is crucial to learn each dialogue information so that user A and user B can be distinguished. Therefore, it performs better when the model is accompanied by additionally providing a one-hot encoder to represent the speaker id for a given utterance.

### 2.4 Knowledge-Integrated Transformer

We integrate contextual dialogue embedding, contextual gazetteer embedding, and speaker embedding. Especially, We integrate the Mixture of Experts (MoE) mechanism (Pavlitskaya et al., 2020; Meng et al., 2021) at the knowledge-integrated Transformer to utilize both the dialogue and gazetteer information efficiently. We add gating networks to create a weighted linear combination of words and gazetteer representations. Training the gating network prevents the overuse or underuse of features.

$$\begin{aligned} w_e &= \sigma(\theta[h_{word}, h_{cgr}]), \\ h_{moe} &= w_e \cdot h_{word} + (1 - w_e) \cdot h_{cgr} \end{aligned} \quad (3)$$

$$h_{integrated} = Transformer([h_{moe}, h_{spk}]) \quad (4)$$

$h_{word}$, $h_{cgr}$ and $h_{spk}$ are the output of respective sub-modules. They are used to train the gating network. $\sigma$ is a Sigmoid activation function and $\theta$ is a trainable parameter. [.,.] represents a concatenation. We learn gating weights $w_e$. The model can learn how to dynamically calculate each syllable unit's hidden information $h_{integrated}$. After obtaining $h_{integrated}$, we feed it to the integrated Transformer encoder to learn the integrated embedding vector.

### 2.5 Span Detector

The span detector model, which uses a composite embedding vector, predicts the position information of all the slots related to the scheduled event. For all the slots of the scheduled events, the span detector is considered to take as the token level representation $[t_1, \cdots, t_n]$, the output of the knowledge-integrated Transformer. Each token representation $t_i$ is projected linearly through a common layer whose output values correspond to start and end positions. Softmax is then applied to the position values to produce a probability distribution for all tokens. Finally, we extract the span value with the highest probability distributions for each target slot and provide that as the output. The formula of the learning method of this model is as follows.

$$\begin{aligned} P_s &= W_s \cdot h_{integrated} + b_s \\ P_e &= W_e \cdot h_{integrated} + b_e \\ P_{joint} &= [P_s, P_e] \\ L_s &= CCE(P_s, y_s) \\ L_e &= CCE(P_e, y_e) \\ L_{joint} &= JE(P_{joint}, y_{joint}) \\ L &= L_s + L_e + L_{joint} \end{aligned} \quad (5)$$

In Equation (4) above, $h_{integrated}$ means integrated token on the knowledge-integrated Transformer, $W_s$ and $W_e$ mean the weight matrix, and $b_s$ and $b_e$ mean the bias. $P_s$ and $P_e$ mean the probability distribution of start and end positions for each token in the dialogue input. $y_s$ and $y_e$ denote the position labels for the correct answer range. Along with modeling start position and end position probabilities separately using Categorical Cross Entropy (CCE) loss, we use Jaccard Expectation (JE) loss to optimize start and end positions jointly instead of CCE loss. Because it showed better results when using JE loss.

| Model Type | 3K dataset JGA | 3K dataset Slot Acc. | 3K dataset Slot F1 | 10K dataset JGA | 10K dataset Slot Acc. | 10K dataset Slot F1 |
|---|---|---|---|---|---|---|
| SOM-DST (Kim et al., 2019)* | 48.00 | - | 83.73 | 62.40 | - | 89.33 |
| DSTRC (baseline) (Gao et al., 2019)* | 50.70 | 91.56 | 84.97 | 68.80 | 94.47 | 90.52 |
| BERT-SpanDetector | 60.00 | 94.12 | 90.38 | 73.00 | 95.71 | 92.47 |
| BERT-SpanDetector +Gaz. | 63.67 | 94.41 | 91.18 | 73.30 | 95.60 | 92.29 |
| BERT-SpanDetector +Gaz. +Transformer | 66.67 | 95.29 | 92.26 | 73.60 | 95.80 | 92.46 |
| BERT-SpanDetector +Gaz. +Transformer +Spk | 68.00 | 95.12 | 91.84 | 73.37 | 95.80 | 94.41 |
| BERT-SpanDetector +Context Gaz. +Transformer | 68.33 | 95.67 | 92.29 | 73.80 | 95.78 | 92.64 |
| BERT-SpanDetector +Context Gaz. +Transformer +MoE | 69.44 | 95.33 | 92.29 | 75.40 | 96.10 | 93.09 |
| BERT-SpanDetector +Context Gaz. +Transformer +MoE +Spk (**KILDST**) | **70.16** | **95.35** | **92.57** | **77.80** | **96.37** | **93.61** |

Table 1: Results on the 3K and 10K datasets for all models . MoE is a mixture of experts. Spk is speaker embedding.

# 3 Experiments and Results

## 3.1 Datasets and Experimental Setup

One of the essential goals of our work is to collect and create dialogue datasets between users that plan specific appointments. We collected and utilized their actual dialogue datasets in Korean with the consent of users to use the provided datasets for research. Also, to collect dialogue data of various ages, gender, and relationship combinations between users who have dialogue, we use a crowdsourcing platform. We created a dataset by providing chat rooms where real users can chat under certain predefined conditions. We provided the purpose of the dialogue, the relationship between users, and guidance information to these user chat rooms. If we set a specific profile of the chat room we want to collect, only users who meet the conditions can attend the chat room. We can collect various combinations and types of dialogue datasets by introducing a real-chat simulator between users in a constrained environment. We evaluate our models on two datasets, such as the 3K and 10K datasets. The 3K dataset has 3,000 dialogues, and 33,585 dialogue turns. 10K dataset has 10,000 dialogues, and 109,971 dialogue turns. The dataset was experimented with and evaluated by the ratio of train set 9 and test set 1. All models are implemented using Tensorflow 2.5.

## 3.2 Evaluation Metrics and Results

We use joint goal accuracy, slot accuracy, and slot F1 score to evaluate our model. **Joint Goal Accuracy** is an accuracy that checks whether all slot values predicted at a dialogue exactly match the ground truth values. The comparison of the results of our model with the state-of-the-art model using MRC techniques on the datasets is presented in Table 1. Our KILDST model achieves higher scores in all evaluation metrics in the test dataset. It shows that the joint goal accuracy increases in our experiments. Whenever we add new features such as speaker encoder, contextual gazetteer encoder, and a mixture of experts gating modules (Meng et al., 2021), the joint goal accuracy increases in our experiments. The composite output vector consists of speaker embedding, contextual gazetteer embedding, and dialogue text embedding. Our best model extracts most slots better than the state-of-the-art model and also all other models in the evaluation dataset.

| Slot Type | JGA | Slot Acc. | Slot F1 |
|---|---|---|---|
| Overall Slots | **77.80** | **96.37** | **93.61** |
| YEAR | 99.10 | 99.10 | 99.10 |
| MONTH | 99.10 | 99.10 | 99.10 |
| WEEK | 98.60 | 98.60 | 98.60 |
| DATE | 93.40 | 93.40 | 93.40 |
| AMPM | 93.20 | 93.20 | 93.20 |
| HOUR | 95.70 | 95.70 | 95.70 |
| MINUTE | 99.40 | 99.40 | 99.40 |
| LOCATION | 92.50 | 92.50 | 92.50 |

Table 2: The best model overall slots result on the 10K

| Hyperparameter | Search Range | Optimal Value |
|---|---|---|
| Batch size | [16,32] | 32 |
| Epochs | [50,75,100] | 100 |
| Learning rate | - | 0.0001 |
| Optimizer | [Adam,AdamW] | AdamW |
| Dropout rate | - | 0.1 |
| BERT Max length | - | 512 |
| BERT Hidden size | - | 256 |
| BiLSTM Input word size | - | 512 |
| BiLSTM Hidden size | - | 256 |
| Integrated Transformer #layers | [1,2] | 2 |

Table 3: Hyperparameters for KILDST model

## 4 Ablation Analysis

### 4.1 Effect of Gazetteer

Integrating additional gazetteer information into the model improves the performance of all models. In particular, models trained with 3,000 datasets benefit from gazetteer knowledge more than models trained with 10,000 datasets. Experiments using a gazetteer in 3,000 datasets have improved JGA performance by 3.67 %. In the case of fine-turned models with insufficient training data, using the gazetteer knowledge has a more significant effect on improving performance. On the other hand, when training the models with a large number of training datasets, such as the model trained through 10,000 datasets, it is interpreted that it gets a relatively small benefit because much gazetteer information is already included in the training datasets. In addition, the results of the model using the contextual gazetteer encoder by the situation encoded by BiLSTM have improved the JGA performance by 1.66 %in the experiments tested with 3,000 data sets than the simple one-hot gazetteer embedding. The results indicate the high efficiency of the CGR, which clearly shows the effect of the integration of the gazetteer.

### 4.2 Effect of Mixture of Experts

The KILDST model is trained and fused with the BERT dialogue encoder and contextual gazetteer encoder. To assess the MoE component's impact, we concatenate the output vector of the contextual dialogue encoder and contextual dialogue encoder without MoE. By applying MoE, there was a 1.11% JGA performance improvement effect in the model experiment with 3,000 datasets and a 1.6% JGA performance improvement effect in the model experiment with 10,000 datasets. Table 1 shows more improvement in JGA performance than other accuracies when applying MoE.

### 4.3 Effect of Integrated Transformer

In the case of our existing BERT-SpanDetector model, the output vectors of the contextual gazetteer representation and the contextual dialogue representation are concatenated and transferred to the span detector without additional training. However, our proposed KILDST approach, which once again trains through the Transformer Encoder, effectively integrates two output vectors, generates an updated embedding vector, and transfers it to the span detector. Additional training using an integrated transformer has improved JGA performance by 3% in model experiments using 3,000 datasets.

## 5 Conclusions

This paper presents a novel model architecture that effectively learns the gazetteer knowledge and speaker information for DST. Our model consists of 3 different sub-encoders a contextual dialogue encoder for encoding dialogue, a contextual gazetteer encoder for encoding domain knowledge, and a speaker encoder for encoding speaker information. The knowledge-integrated learning method we proposed outperforms other models by integrating the information of various encoders more efficiently and accurately extracting slots. In addition, we have collected and created new dialogue datasets between users that plan a schedule. Our evaluation of this dialogue dataset shows improvement over the state-of-the-art model by better extracting the schedule from a dialogue containing difficult words such as newly coined words and abbreviations. Our model can be applied to a messenger app such as WhatsApp, WeChat, Kakaotalk, and Telegram as a recommendation system, extracting a schedule from dialogue and recommending a schedule to the user. In the future, we plan to expand our proposed model architecture to a model that extracts another meaningful event from a dialogue among more users.

## Acknowledgement

The authors would like to thank Chanwoo Kim, Munjo Kim, Jonggu Kim, Siddhartha Mukherjee, Raghavendra Hanumantasetty Ramasetty, and Aniruddha Uttam Tammewar for their helpful suggestions and discussion during this project.

## References

Sandeep Ashwini and Jinho D Choi. 2014. Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

Gabriel Bernier-Colborne and Philippe Langlais. 2020. Hardeval: Focusing on challenging tokens to assess robustness of ner. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1704–1711.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Shuyang Gao, Sanchit Agarwal, Tagyoung Chung, Di Jin, and Dilek Hakkani-Tur. 2020. From machine reading comprehension to dialogue state tracking: Bridging the gap. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of Special Interest Group on Discourse and Dialogue*.

Pratik Jayarao, Chirag Jain, and Aman Srivastava. 2018. Exploring the importance of context and embeddings in neural ner models for task-oriented dialogue systems. *arXiv preprint arXiv:1812.02370*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *CoRR*, abs/1911.03906.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

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

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Svetlana Pavlitskaya, Christian Hubschneider, Michael Weber, Ruby Moritz, Fabian Huger, Peter Schlicht, and Marius Zollner. 2020. Using mixture of expert models to gain insights into semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 342–343.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055.

Chan Hee Song, Dawn Lawrie, Tim Finin, and James Mayfield. 2020. Improving neural named entity recognition with gazetteers. *arXiv preprint arXiv:2003.03072*.

Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *arXiv preprint arXiv:1706.04815*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017a. Evidence aggregation for answer re-ranking in open-domain question answering. *arXiv preprint arXiv:1711.05116*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017b. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

## A Appendices

| Korean conversation between 2 users | Korean conversation translated into English |
|---|---|
| User1: 예슬아 언제 시간됨? | User1: Jennie, When are you free? |
| User2: 왜? | User2: Why? |
| User1: 용진이한테 빌린 보드게임 돌려줘야하는데 그런김에 같이 모이게 | User1: I have to return the board game I borrowed from Yongjin, so we might as well meet up sometime |
| User2: ㅇㅋ 나 이번주는 안되고 다음주 일요일에 될듯? 저번처럼 스타벅스? | User2: Okay, I am not free this week, but next Sunday is good. Starbucks like last time? |
| User1: ㅇㅇ넴 저녁먹게 6시에 보죠 | User1: Let's meet up at 6 o'clock and have dinner |
| User2: 그때 다른 약속있거든... 그래서 한시간 늦게보자 | User2: I have an appointment at that time, but let's make it an hour late |
| User1: 난 괜찮음 | User1: Sounds good |

Table 4: Korean conversation dataset between 2 users

| Korean ground truth | Ground truth translated into English |
|---|---|
| • schedule.week: 다음주 | • schedule.week: next week |
| • schedule.date: 일요일 | • schedule.date: Sunday |
| • schedule.hour: 7시 | • schedule.hour: 7 |
| • schedule.ampm: 저녁 | • schedule.ampm: dinner |
| • schedule.location: 스타벅스 | • schedule.location: Starbucks |

Table 5: Ground truth of Korean conversation dataset