# Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots

Jia-Chen Gu[1], Tianda Li[2], Quan Liu[1,3], Zhen-Hua Ling[1], Zhiming Su[3], Si Wei[3], Xiaodan Zhu[2]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China

[2]ECE & Ingenuity Labs, Queen's University

[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

# Outline

- **Introduction**
- Speaker-Aware BERT
- Experiments
- Conclusion

# Introduction

- Multi-Turn Response Selection

  The task is to select the best-matched response from a set of candidates given the context of a conversation.

| | **Context** | |
|---|---|---|
| Utterance 1 | *Human*: How are you doing? | |
| Utterance 2 | *ChatBot*: I am going to hold a drum class in Shanghai. Anyone wants to join? | |
| Utterance 3 | *Human*: Interesting! Do you have coaches who can help me practice drum? | |
| Utterance 4 | *ChatBot*: Of course. | |
| Utterance 5 | *Human*: Can I have a free first lesson? | |
| | **Response Candidates** | |
| Response 1 | Sure. Have you ever played drum before? | **(Correct)** |
| Response 2 | What lessons do you want? | **(Wrong)** |

# Outline

- Introduction
- **Speaker-Aware BERT**
- Experiments
- Conclusion

# Speaker-Aware BERT

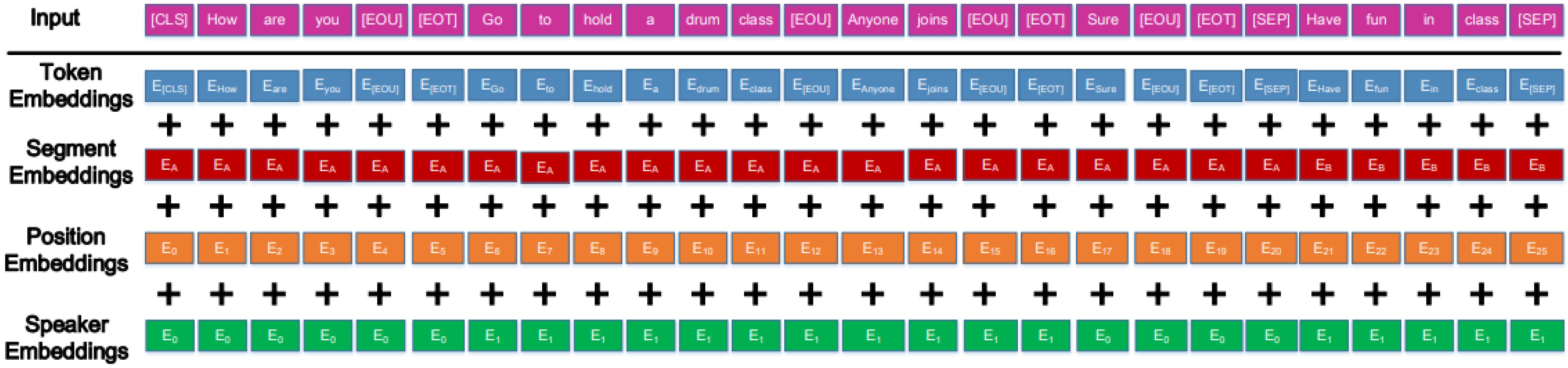
Three issues in present pre-trained language models (PLMs) for dialogue:

- Unable to model the speaker change in turn as a conversation progresses.
- Unable to process disentangled conversations with multiple speakers.
- Lack specific in-domain knowledge.

# Speaker-Aware BERT

Two strategies to model the speaker change information:

- Add the speaker embeddings to the token representation.
- Add the special segmentation tokens between the context utterances.



# Speaker-Aware BERT

A heuristic speaker-aware strategy to disentangle the dialogue:

- Define spoken-from speaker who is uttering an utterance
- Define spoken-to speaker who is receiving an utterance
- Select the utterances which have the same spoken-from or spoken-to speaker as the spoken-from speaker of the response
- The selected utterances are assigned with the spoken-from or spoken-to speaker embeddings respectively.

# Speaker-Aware BERT

Domain adaptation to incorporate in-domain knowledge into PLMs:

- Employ the training set of each dataset for domain adaptation without additional external knowledge.
- Continue post-training with: (1) a next sentence prediction (NSP) loss, and (2) a masked language model (MLM) loss.

# Outline

- Introduction
- Speaker-Aware BERT
- **Experiments**
- Conclusion

# Experiments

- Datasets

| Datasets | | Train | Valid | Test |
|---|---|---|---|---|
| **Ubuntu V1** | pairs | 1M | 0.5M | 0.5M |
| | positive : negative | 1:1 | 1:9 | 1:9 |
| **Ubuntu V2** | pairs | 1M | 195k | 189k |
| | positive : negative | 1:1 | 1:9 | 1:9 |
| **Douban** | pairs | 1M | 50k | 10k |
| | positive : negative | 1:1 | 1:1 | 1:9 |
| **E-commerce** | pairs | 1M | 10k | 10k |
| | positive : negative | 1:1 | 1:1 | 1:9 |
| **DSTC8** | pairs | 11M | 1M | 1M |
| | positive : negative | 1: 99 | 1: 99 | 1: 99 |

# Experiments

- Overall Performance

| | Ubuntu Corpus V1 | | | | Ubuntu Corpus V2 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathbf{R_2@1}$ | $\mathbf{R_{10}@1}$ | $\mathbf{R_{10}@2}$ | $\mathbf{R_{10}@5}$ | $\mathbf{R_2@1}$ | $\mathbf{R_{10}@1}$ | $\mathbf{R_{10}@2}$ | $\mathbf{R_{10}@5}$ |
| TF-IDF (Lowe et al., 2015; Lowe et al., 2017) | 0.659 | 0.410 | 0.545 | 0.708 | 0.749 | 0.488 | 0.587 | 0.763 |
| RNN (Lowe et al., 2015; Lowe et al., 2017) | 0.768 | 0.403 | 0.547 | 0.819 | 0.777 | 0.379 | 0.561 | 0.836 |
| LSTM (Lowe et al., 2015; Lowe et al., 2017) | 0.878 | 0.604 | 0.745 | 0.926 | 0.869 | 0.552 | 0.721 | 0.924 |
| DL2R (Yan et al., 2016) | 0.899 | 0.626 | 0.783 | 0.944 | - | - | - | - |
| Match-LSTM (Wang and Jiang, 2016b) | 0.904 | 0.653 | 0.799 | 0.944 | - | - | - | - |
| MV-LSTM (Wan et al., 2016) | 0.906 | 0.653 | 0.804 | 0.946 | - | - | - | - |
| Multi-View (Zhou et al., 2016) | 0.908 | 0.662 | 0.801 | 0.951 | - | - | - | - |
| RNN-CNN (Baudis and Sedivý, 2016) | - | - | - | - | 0.911 | 0.672 | 0.809 | 0.956 |
| CompAgg (Wang and Jiang, 2016a) | 0.884 | 0.631 | 0.753 | 0.927 | 0.895 | 0.641 | 0.776 | 0.937 |
| BiMPM (Wang et al., 2017) | 0.897 | 0.665 | 0.786 | 0.938 | 0.877 | 0.611 | 0.747 | 0.921 |
| HRDE-LTC (Yoon et al., 2018) | 0.916 | 0.684 | 0.822 | 0.960 | 0.915 | 0.652 | 0.815 | 0.966 |
| SMN (Wu et al., 2017) | 0.926 | 0.726 | 0.847 | 0.961 | - | - | - | - |
| DUA (Zhang et al., 2018b) | - | 0.752 | 0.868 | 0.962 | - | - | - | - |
| DAM (Zhou et al., 2018b) | 0.938 | 0.767 | 0.874 | 0.969 | - | - | - | - |
| MRFN (Tao et al., 2019a) | 0.945 | 0.786 | 0.886 | 0.976 | - | - | - | - |
| IMN (Gu et al., 2019) | 0.946 | 0.794 | 0.889 | 0.974 | 0.945 | 0.771 | 0.886 | 0.979 |
| IOI (Tao et al., 2019b) | 0.947 | 0.796 | 0.894 | 0.974 | - | - | - | - |
| MSN (Yuan et al., 2019) | - | 0.800 | 0.899 | 0.978 | - | - | - | - |
| BERT | 0.950 | 0.808 | 0.897 | 0.975 | 0.950 | 0.781 | 0.890 | 0.980 |
| **SA-BERT** | **0.965** | **0.855** | **0.928** | **0.983** | **0.963** | **0.830** | **0.919** | **0.985** |

SA-BERT achieves new state-of-the-art performances on five public datasets.

# Experiments

- Overall Performance

| | Douban Conversation Corpus | | | | | | E-commerce Corpus | | |
|---|---|---|---|---|---|---|---|---|---|
| | **MAP** | **MRR** | **P@1** | **$R_{10}$@1** | **$R_{10}$@2** | **$R_{10}$@5** | **$R_{10}$@1** | **$R_{10}$@2** | **$R_{10}$@5** |
| TF-IDF (Lowe et al., 2015) | 0.331 | 0.359 | 0.180 | 0.096 | 0.172 | 0.405 | 0.159 | 0.256 | 0.477 |
| RNN (Lowe et al., 2015) | 0.390 | 0.422 | 0.208 | 0.118 | 0.223 | 0.589 | 0.325 | 0.463 | 0.775 |
| LSTM (Lowe et al., 2015) | 0.485 | 0.527 | 0.320 | 0.187 | 0.343 | 0.720 | 0.365 | 0.536 | 0.828 |
| Multi-View (Zhou et al., 2016) | 0.505 | 0.543 | 0.342 | 0.202 | 0.350 | 0.729 | 0.421 | 0.601 | 0.861 |
| DL2R (Yan et al., 2016) | 0.488 | 0.527 | 0.330 | 0.193 | 0.342 | 0.705 | 0.399 | 0.571 | 0.842 |
| MV-LSTM (Wan et al., 2016) | 0.498 | 0.538 | 0.348 | 0.202 | 0.351 | 0.710 | 0.412 | 0.591 | 0.857 |
| Match-LSTM (Wang and Jiang, 2016b) | 0.500 | 0.537 | 0.345 | 0.202 | 0.348 | 0.720 | 0.410 | 0.590 | 0.858 |
| SMN (Wu et al., 2017) | 0.529 | 0.569 | 0.397 | 0.233 | 0.396 | 0.724 | 0.453 | 0.654 | 0.886 |
| DUA (Zhang et al., 2018b) | 0.551 | 0.599 | 0.421 | 0.243 | 0.421 | 0.780 | 0.501 | 0.700 | 0.921 |
| DAM (Zhou et al., 2018b) | 0.550 | 0.601 | 0.427 | 0.254 | 0.410 | 0.757 | - | - | - |
| MRFN (Tao et al., 2019a) | 0.571 | 0.617 | 0.448 | 0.276 | 0.435 | 0.783 | - | - | - |
| IMN (Gu et al., 2019) | 0.570 | 0.615 | 0.433 | 0.262 | 0.452 | 0.789 | 0.621 | 0.797 | 0.964 |
| IOI (Tao et al., 2019b) | 0.573 | 0.621 | 0.444 | 0.269 | 0.451 | 0.786 | 0.563 | 0.768 | 0.950 |
| MSN (Yuan et al., 2019) | 0.587 | 0.632 | 0.470 | 0.295 | 0.452 | 0.788 | 0.606 | 0.770 | 0.937 |
| BERT | 0.591 | 0.633 | 0.454 | 0.280 | 0.470 | 0.828 | 0.610 | 0.814 | 0.973 |
| **SA-BERT** | **0.619** | **0.659** | **0.496** | **0.313** | **0.481** | **0.847** | **0.704** | **0.879** | **0.985** |

SA-BERT achieves new state-of-the-art performances on five public datasets.

# Experiments

- Overall Performance

| DSTC 8-Track 2-Subtask 2 | | | |
|---|---|---|---|
| Set | Model | **MRR** | **$R_{100}$@1** |
| Valid | IMN (Gu et al., 2019) | 0.443 | 0.322 |
| | IMN (Gu et al., 2019) + SDS | 0.504 | 0.375 |
| | BERT | 0.335 | 0.258 |
| | BERT + SDS | 0.560 | 0.440 |
| | SA-BERT - SDS | 0.344 | 0.265 |
| | **SA-BERT** | **0.594** | **0.477** |
| | SA-BERT (Ensemble) | 0.611 | 0.496 |
| Test | SA-BERT (Ensemble) | 0.621 | 0.506 |

SA-BERT achieves new state-of-the-art performances on five public datasets.

Ablation tests by ablating the speaker-aware disentanglement strategy (SDS).

# Experiments

- Analysis

| Pre-train | Speaker embeddings | $\mathbf{R}_2$@1 | $\mathbf{R}_{10}$@1 | $\mathbf{R}_{10}$@2 | $\mathbf{R}_{10}$@5 |
|---|---|---|---|---|---|
| No | No | 0.950 | 0.781 | 0.890 | 0.980 |
| No | Yes | 0.950 | 0.786 | 0.890 | 0.981 |
| Yes | No | 0.961 | 0.825 | 0.915 | 0.984 |
| **Yes** | **Yes** | **0.963** | **0.830** | **0.919** | **0.985** |

Ablation tests on the Ubuntu V2 dataset by ablating the speaker embeddings.

# Experiments

- Analysis

| Adaptation corpus | $\mathbf{R}_2$@1 | $\mathbf{R}_{10}$@1 | $\mathbf{R}_{10}$@2 | $\mathbf{R}_{10}$@5 |
|---|---|---|---|---|
| None | 0.950 | 0.786 | 0.890 | 0.981 |
| DSTC8 | 0.954 | 0.803 | 0.902 | 0.981 |
| Ubuntu V1 | 0.961 | 0.824 | 0.914 | 0.985 |
| **Ubuntu V2** | **0.963** | **0.830** | **0.919** | **0.985** |

Results on the test set of Ubuntu V2 dataset, by domain adaptation with different corpora and fine-tuning all on the training set of Ubuntu V2 dataset.

The more similar to the task this adaptation corpus is, the more improvement it can help to achieve.

# Outline

- Introduction
- Speaker-Aware BERT
- Experiments
- **Conclusion**

# Conclusion

- Speaker change is an important and intrinsic property of multi-turn dialogues, which should be modeled in PLMs.
- In addition to general knowledge, specific in-domain knowledge is also important for response selection in retrieval-based chatbots.

Jia-Chen Gu

Tianda Li

Quan Liu

Zhen-Hua Ling

Zhiming Su

Si Wei

Xiaodan Zhu

# Thanks!

Code and Model: https://github.com/JasonForJoy/SA-BERT