
**Taesun Whang\*, Dongyub Lee\*, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee**

* Equal Contribution.
Multi-turn Response Selection

- Selecting the optimal response given a user and dialog context in multi-turn dialog systems.

[Dialog Context]

User: Good morning! What can I do for you?

Agent: I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?

Agent: With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.

User: How much does a seven-day tour by bus cost?

[Response Candidates]

1. Two thousand dollars.

2. Does that include hotels and meals?

...
Recent success of PrLM based models

- BERT-VFT (Whang et al., 2020)
- Obtained state-of-the-art results.
- Tend to make predictions based on relatedness of history and candidates.
- Limits in adapting the sequential nature of multi-turn dialog.

- BERT-SS-DA (Lu et al., 2020)

Lu et al., Improving Contextual Language Models for Response Retrieval in Multi-Turn Conversation. SIGIR 2020.
Recent success of PrLM based models (Adversarial Experiments)

[Dialog Context]

Hello, is there anything I can help you with?

Hi, I want to get some suggestions about next semester's course selections.

Great, what is your major?

I'm interested in computer engineering.

What level of programming are you capable of?

I have some programming experience in C++ and Matlab after taking …

That works. Are there any suggestions of advanced classes using Python?

Nice try of it.

Next term, I will learn Python, there are other topics that I like also.

[Response Candidates]

I'd recommend that your take EECS280 and EECS203 as soon as you can. They are important for your computer science major.

(a) Ground Truth (BERT score : 0.813)

That works. Are there any suggestions of advanced classes using Python?

(b) Adversarial Example (BERT score : 0.993)
Multi-turn Response Selection (Challenges)

- Domain adaption based on an additional training on a target corpus is extremely time-consuming and computationally costly.
- Formulating response selection as a dialog-response binary classification task is insufficient to represent intra- and inter-utterance interactions.
- Existing models tend to select the optimal response depending on how semantically similar it is to a given dialog.

Utterance Manipulation Strategies (UMS)
Contributions

• Show that existing response selection models are more likely to predict a semantically relevant response with its dialog rather than the next utterance.

• Propose highly effective self-supervised learning methods, utterance manipulation strategies (UMS), which aid the model towards maintaining dialog coherence.

• State-of-the-art performance on multiple public benchmarks (i.e., Ubuntu, Douban, and E-commerce).
Proposed Method (Overview)

[Dialog Context]
- Good morning! What can I do for you?
- I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?
- With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.
- How much does a seven-day tour by bus cost?
- Two thousand dollars.
- Does that include hotels and meals?
- Oh, yes, and admission tickets for places of interest as well.

[Response]
- That sounds reasonable.

(a) Response Selection

[Utterance Insertion]

(a) I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?
(b) With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.
(d) Two thousand dollars.
(e) Does that include hotels and meals?
(c) How much does a seven-day tour by bus cost?

[Utterance Deletion]

(a) I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?
(b) With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.
(c) I'd like to taste some local dishes. What would you recommend?
(d) How much does a seven-day tour by bus cost?
(e) Does that include hotels and meals?
(d) Two thousand dollars.

[Utterance Search]

(a) I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?
(b) With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.
(c) How much does a seven-day tour by bus cost?
(d) Two thousand dollars.
(e) Does that include hotels and meals?
(f) Oh, yes, and admission tickets for places of interest as well.

(b) Utterance Manipulation Strategies
Proposed Method (Language Models for Response Selection)

- Pre-trained Language Models: BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020)
- Domain-specific Post-training (Additional training on a target corpus with PrLM objectives)
- Training Response Selection Models: BERT (Whang et al., 2020)

\[
X = ([CLS] u_1 u_2 \ldots u_{n_u} [SEP] r [SEP])
\]

\[
g(U, r) = \sigma(w^T x_{[CLS]} + b)
\]

\[
\mathcal{L}oss = - \sum_{(U, r, y) \in \mathcal{D}} y \log(g(U, r)) + (1 - y) \log(1 - g(U, r))
\]

Clark et al., ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR 2020.
Proposed Method (UMS – Utterance Insertion)

[Utterance Insertion]

(a) I’m thinking of traveling to California in May. Could you recommend some tourist programs for that?
(b) With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.
(d) Two thousand dollars.
(e) Does that include hotels and meals?
(c) How much does a seven-day tour by bus cost?

- Find where the selected utterance should be inserted.
- [INS] tokens are positioned before each utterance and after the last utterance.
- \( u_t \) is the target utterance and \([\text{INS}]_t\) is the target insertion token.

\[
X_{\text{INS}} = [[\text{CLS}][\text{INS}]_1 u_1[\text{INS}]_2 u_2 \ldots u_{t-1} [\text{INS}]_t u_{t+1} \ldots u_k [\text{INS}]_k [\text{SEP}] u_t [\text{SEP}]]
\]
I'd like to taste some local dishes. What would you recommend?

Two thousand dollars.

With pleasure. We arrange two kinds of tourist programs for California, a seven-day tour by bus and a five-day flying journey.

I'm thinking of traveling to California in May. Could you recommend some tourist programs for that?

How much does a seven-day tour by bus cost?

Two thousand dollars.

**Proposed Method (UMS – Utterance Deletion)**

**[Utterance Deletion]**

- Find an unrelated utterance to the dialog.
- The unrelated utterance is sampled from the random dialog.
- [DEL] tokens are positioned before each utterance.
- \( u^{\text{rand}} \) is the utterance from the random dialog and \([\text{DEL}]_t\) is the target deletion token.

\[
X_{\text{DEL}} = [[\text{CLS}] [\text{DEL}]_1 u_1 [\text{DEL}]_2 u_2 \ldots [\text{DEL}]_t u^{\text{rand}} [\text{DEL}]_{t+1} u_t \ldots [\text{DEL}]_{k+1} u_k [\text{SEP}]]
\]
Proposed Method (UMS – Utterance Search)

[Utterance Search]

- Find the previous utterance of the last utterance from the jumbled utterances.
- Shuffle utterances except for the last utterance.
- [SRCH] tokens are positioned before each utterance.
- $u'_t (u_{k-1})$ is the previous utterance of the last utterance $u_k$ and $[SRCH]_t$ is the target search token.

$$X_{SRCH} = [[CLS] \ [SRCH]_1 u'_1 [SRCH]_2 u'_2 \ ...
[SRCH]_t u'_t \ ... u'_{k-1} [SEP] u_k [SEP]]$$
Proposed Method (Multi-task Learning Setup)

• The output representations of special tokens ([INS], [DEL], and [SRCH]) are used to classify whether each token is in a correct position to be inserted, deleted, and searched.

• Target tokens for each task ([INS]_t, [DEL]_t, and [SRCH]_t) are labeled as 1, otherwise 0.

\[
p(y_{\text{TASK}} = 1 | X_{\text{TASK}}) = \sigma \left( w^T x_{\text{TASK}} + b \right), \quad \text{where } \text{TASK} \in \{\text{INS, DEL, SRCH}\}
\]

• Binary cross-entropy loss for all auxiliary tasks to optimize the model.

• Response Selection loss and UMS losses are summed with the same ratio.
Experimental Setup

• Dataset
  • Ubuntu : Ubuntu internet relay chats (troubleshooting the Ubuntu OS).
  • Douban : Chinese open-domain dialogs (web-crawled from Douban Group).
  • E-Commerce : Chinese customer consultation dialogs (Taobao).
  • Kakao : Korean open-domain (Twitter and Reddit) constructed by Kakao Corporation.

• Evaluation Metrics
  • $R_n@k$ ($k=\{1, 2, 5\}$), P@1, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ubuntu</th>
<th>Douban</th>
<th>E-Commerce</th>
<th>Kakao</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Val</td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td># pairs</td>
<td>1M</td>
<td>500K</td>
<td>500K</td>
<td></td>
</tr>
<tr>
<td>pos:neg</td>
<td>1:1</td>
<td>1:9</td>
<td>1:9</td>
<td></td>
</tr>
<tr>
<td># avg turns</td>
<td>10.13</td>
<td>10.11</td>
<td>10.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Baselines

• Single-turn Matching Models
  • CNN, LSTM, BiLSTM
  • MV-LSTM, Match-LSTM

• Multi-turn Matching Models
  • Multi-View, DL2R
  • SMN, DUA, DAM, IoI
  • MSN

• BERT-based Models
  • Vanilla BERT, BERT-SS-DA, SA-BERT
## Quantitative Results (Ubuntu, Douban, and E-Commerce Corpus)

<table>
<thead>
<tr>
<th>Models</th>
<th>Ubuntu $R_{10@1}$</th>
<th>Ubuntu $R_{10@2}$</th>
<th>Ubuntu $R_{10@5}$</th>
<th>MAP</th>
<th>MRR</th>
<th>Douban $P@1$</th>
<th>Douban $R_{10@1}$</th>
<th>Douban $R_{10@2}$</th>
<th>Douban $R_{10@5}$</th>
<th>E-commerce $R_{10@1}$</th>
<th>E-commerce $R_{10@2}$</th>
<th>E-commerce $R_{10@5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN (Kadlec, Schmid, and Kleindienst 2015)</td>
<td>0.549</td>
<td>0.684</td>
<td>0.896</td>
<td>0.417</td>
<td>0.440</td>
<td>0.226</td>
<td>0.121</td>
<td>0.252</td>
<td>0.647</td>
<td>0.328</td>
<td>0.515</td>
<td>0.792</td>
</tr>
<tr>
<td>LSTM (Kadlec, Schmid, and Kleindienst 2015)</td>
<td>0.638</td>
<td>0.784</td>
<td>0.949</td>
<td>0.485</td>
<td>0.537</td>
<td>0.320</td>
<td>0.187</td>
<td>0.343</td>
<td>0.720</td>
<td>0.365</td>
<td>0.536</td>
<td>0.828</td>
</tr>
<tr>
<td>BiLSTM (Kadlec, Schmid, and Kleindienst 2015)</td>
<td>0.630</td>
<td>0.780</td>
<td>0.944</td>
<td>0.479</td>
<td>0.514</td>
<td>0.313</td>
<td>0.184</td>
<td>0.330</td>
<td>0.716</td>
<td>0.365</td>
<td>0.536</td>
<td>0.825</td>
</tr>
<tr>
<td>MV-LSTM (Wan et al. 2016)</td>
<td>0.653</td>
<td>0.804</td>
<td>0.946</td>
<td>0.498</td>
<td>0.538</td>
<td>0.348</td>
<td>0.202</td>
<td>0.351</td>
<td>0.710</td>
<td>0.412</td>
<td>0.591</td>
<td>0.857</td>
</tr>
<tr>
<td>Match-LSTM(Wang and Jiang 2016)</td>
<td>0.653</td>
<td>0.799</td>
<td>0.944</td>
<td>0.500</td>
<td>0.537</td>
<td>0.345</td>
<td>0.202</td>
<td>0.348</td>
<td>0.720</td>
<td>0.410</td>
<td>0.590</td>
<td>0.858</td>
</tr>
<tr>
<td>Multi-View (Zhou et al. 2016)</td>
<td>0.662</td>
<td>0.801</td>
<td>0.951</td>
<td>0.505</td>
<td>0.543</td>
<td>0.342</td>
<td>0.202</td>
<td>0.350</td>
<td>0.729</td>
<td>0.421</td>
<td>0.601</td>
<td>0.861</td>
</tr>
<tr>
<td>DL2R (Yan, Song, and Wu 2016)</td>
<td>0.626</td>
<td>0.783</td>
<td>0.944</td>
<td>0.488</td>
<td>0.527</td>
<td>0.330</td>
<td>0.193</td>
<td>0.342</td>
<td>0.705</td>
<td>0.399</td>
<td>0.571</td>
<td>0.842</td>
</tr>
<tr>
<td>SMN (Wu et al. 2017)</td>
<td>0.726</td>
<td>0.847</td>
<td>0.961</td>
<td>0.529</td>
<td>0.569</td>
<td>0.397</td>
<td>0.233</td>
<td>0.396</td>
<td>0.724</td>
<td>0.453</td>
<td>0.654</td>
<td>0.886</td>
</tr>
<tr>
<td>DUA (Zhang et al. 2018)</td>
<td>0.752</td>
<td>0.868</td>
<td>0.962</td>
<td>0.551</td>
<td>0.599</td>
<td>0.421</td>
<td>0.243</td>
<td>0.421</td>
<td>0.780</td>
<td>0.501</td>
<td>0.700</td>
<td>0.921</td>
</tr>
<tr>
<td>DAM (Zhou et al. 2018)</td>
<td>0.767</td>
<td>0.874</td>
<td>0.969</td>
<td>0.550</td>
<td>0.601</td>
<td>0.427</td>
<td>0.254</td>
<td>0.410</td>
<td>0.757</td>
<td>0.526</td>
<td>0.727</td>
<td>0.933</td>
</tr>
<tr>
<td>IoI (Tao et al. 2019b)</td>
<td>0.796</td>
<td>0.894</td>
<td>0.974</td>
<td>0.573</td>
<td>0.621</td>
<td>0.444</td>
<td>0.269</td>
<td>0.451</td>
<td>0.786</td>
<td>0.563</td>
<td>0.768</td>
<td>0.950</td>
</tr>
<tr>
<td>MSN (Yuan et al. 2019)</td>
<td>0.800</td>
<td>0.899</td>
<td>0.978</td>
<td>0.587</td>
<td>0.632</td>
<td>0.470</td>
<td>0.295</td>
<td>0.452</td>
<td>0.788</td>
<td>0.606</td>
<td>0.770</td>
<td>0.937</td>
</tr>
<tr>
<td>BERT (Gu et al. 2020)</td>
<td>0.808</td>
<td>0.897</td>
<td>0.975</td>
<td>0.591</td>
<td>0.633</td>
<td>0.454</td>
<td>0.280</td>
<td>0.470</td>
<td>0.828</td>
<td>0.610</td>
<td>0.814</td>
<td>0.973</td>
</tr>
<tr>
<td>BERT-SS-DA (Lu et al. 2020)</td>
<td>0.813</td>
<td>0.901</td>
<td>0.977</td>
<td>0.602</td>
<td>0.643</td>
<td>0.458</td>
<td>0.280</td>
<td>0.491</td>
<td>0.843</td>
<td>0.648</td>
<td>0.843</td>
<td>0.980</td>
</tr>
<tr>
<td>SA-BERT (Gu et al. 2020)</td>
<td>0.855</td>
<td>0.928</td>
<td>0.983</td>
<td>0.619</td>
<td>0.659</td>
<td>0.496</td>
<td>0.313</td>
<td>0.481</td>
<td>0.847</td>
<td>0.704</td>
<td>0.879</td>
<td>0.985</td>
</tr>
<tr>
<td>BERT (ours)</td>
<td>0.820</td>
<td>0.906</td>
<td>0.978</td>
<td>0.597</td>
<td>0.634</td>
<td>0.448</td>
<td>0.279</td>
<td>0.489</td>
<td>0.823</td>
<td>0.641</td>
<td>0.824</td>
<td>0.973</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>0.826</td>
<td>0.908</td>
<td>0.978</td>
<td>0.602</td>
<td>0.642</td>
<td>0.465</td>
<td>0.287</td>
<td>0.483</td>
<td>0.839</td>
<td>0.609</td>
<td>0.804</td>
<td>0.965</td>
</tr>
<tr>
<td>UMSBERT</td>
<td>0.843</td>
<td>0.920</td>
<td>0.982</td>
<td>0.597</td>
<td>0.639</td>
<td>0.466</td>
<td>0.285</td>
<td>0.471</td>
<td>0.829</td>
<td>0.674</td>
<td>0.861</td>
<td>0.980</td>
</tr>
<tr>
<td>UMS ELECTRA</td>
<td>0.854</td>
<td>0.929</td>
<td>0.984</td>
<td>0.608</td>
<td>0.650</td>
<td>0.472</td>
<td>0.291</td>
<td>0.488</td>
<td>0.845</td>
<td>0.648</td>
<td>0.831</td>
<td>0.974</td>
</tr>
<tr>
<td>BERT+</td>
<td>0.862</td>
<td>0.935</td>
<td>0.987</td>
<td>0.609</td>
<td>0.645</td>
<td>0.463</td>
<td>0.290</td>
<td>0.505</td>
<td>0.838</td>
<td>0.725</td>
<td>0.890</td>
<td>0.984</td>
</tr>
<tr>
<td>ELECTRA+</td>
<td>0.861</td>
<td>0.932</td>
<td>0.985</td>
<td>0.612</td>
<td>0.655</td>
<td>0.480</td>
<td>0.301</td>
<td>0.499</td>
<td>0.836</td>
<td>0.673</td>
<td>0.835</td>
<td>0.974</td>
</tr>
<tr>
<td>UMSBERT+</td>
<td>0.875†</td>
<td>0.942†</td>
<td>0.988†</td>
<td>0.625</td>
<td>0.664</td>
<td>0.499</td>
<td>0.318</td>
<td>0.482</td>
<td>0.858</td>
<td>0.762</td>
<td>0.905</td>
<td>0.986</td>
</tr>
<tr>
<td>UMS ELECTRA+</td>
<td>0.875</td>
<td>0.941</td>
<td>0.988</td>
<td>0.623</td>
<td>0.663</td>
<td>0.492</td>
<td>0.307</td>
<td>0.501</td>
<td>0.851</td>
<td>0.707</td>
<td>0.853</td>
<td>0.974</td>
</tr>
</tbody>
</table>
Quantitative Results (Ubuntu, Douban, and E-Commerce Corpus)

- Two different PrLMs (BERT and ELECTRA)
- Domain-specific post-training (denoted as BERT+ and ELECTRA+)
- ELECTRA vs UMS\textsubscript{ELECTRA}
  - $R_{10}@1$: + 2.8% (Ubuntu), + 3.9% (E-Commerce)
  - P@1: + 0.7% (Douban)
- BERT+ vs UMS\textsubscript{BERT+}
  - $R_{10}@1$: + 1.3% (Ubuntu), + 3.7% (E-Commerce)
  - P@1: + 3.3% (Douban)

<table>
<thead>
<tr>
<th>Models</th>
<th>Ubuntu</th>
<th>Douban</th>
<th>E-commerce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{10}@1$</td>
<td>$R_{10}@2$</td>
<td>$R_{10}@5$</td>
</tr>
<tr>
<td>BERT (ours)</td>
<td>0.820</td>
<td>0.906</td>
<td>0.978</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>0.826</td>
<td>0.908</td>
<td>0.978</td>
</tr>
<tr>
<td>UMS\textsubscript{BERT}</td>
<td>0.843</td>
<td>0.920</td>
<td>0.982</td>
</tr>
<tr>
<td>UMS\textsubscript{ELECTRA}</td>
<td>0.854</td>
<td>0.929</td>
<td>0.984</td>
</tr>
<tr>
<td>BERT+</td>
<td>0.862</td>
<td>0.935</td>
<td>0.987</td>
</tr>
<tr>
<td>ELECTRA+</td>
<td>0.861</td>
<td>0.932</td>
<td>0.985</td>
</tr>
<tr>
<td>UMS\textsubscript{BERT+}</td>
<td>0.875$^\dagger$</td>
<td>0.942$^\dagger$</td>
<td>0.988$^\dagger$</td>
</tr>
<tr>
<td>UMS\textsubscript{ELECTRA+}</td>
<td>0.875</td>
<td>0.941</td>
<td>0.988</td>
</tr>
</tbody>
</table>
Quantitative Results (Kakao Corpus)

- BERT vs UMS$_{BERT}$
- UMS$_{BERT}$ improves performance compared to the baseline for both Web and Clean.
- Absolute improvement of 5.1% (Web) and 6.8% (Clean) in P@1.

<table>
<thead>
<tr>
<th>Test Split</th>
<th>Approach</th>
<th>MAP</th>
<th>MRR</th>
<th>P@1</th>
<th>$R_{10}@1$</th>
<th>$R_{10}@2$</th>
<th>$R_{10}@5$</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web</td>
<td>BERT</td>
<td>0.671</td>
<td>0.720</td>
<td>0.555</td>
<td>0.391</td>
<td>0.599</td>
<td>0.890</td>
<td>+ 5.1%</td>
</tr>
<tr>
<td></td>
<td>UMS$_{BERT}$</td>
<td>0.699</td>
<td>0.751</td>
<td>0.606</td>
<td>0.428</td>
<td>0.623</td>
<td>0.911</td>
<td></td>
</tr>
<tr>
<td>Clean</td>
<td>BERT</td>
<td>0.726</td>
<td>0.792</td>
<td>0.648</td>
<td>0.395</td>
<td>0.612</td>
<td>0.888</td>
<td>+ 6.8%</td>
</tr>
<tr>
<td></td>
<td>UMS$_{BERT}$</td>
<td>0.761</td>
<td>0.834</td>
<td>0.716</td>
<td>0.431</td>
<td>0.663</td>
<td>0.903</td>
<td></td>
</tr>
</tbody>
</table>
Adversarial Experiment

- Investigate whether language models for response selection are trained properly.
- Randomly extract an utterance from the dialog context and replace it with one of negative responses.
- $R_{10}@1$ score decreases by 58% (baselines) and 48% (UMS) on average.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Model</th>
<th>Original $R_{10}@1$</th>
<th>Original MRR</th>
<th>Adversarial $R_{10}@1$</th>
<th>Adversarial MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td>BERT</td>
<td>0.820</td>
<td>0.887</td>
<td>0.199</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>BERT+</td>
<td><strong>0.862</strong></td>
<td><strong>0.915</strong></td>
<td>0.203</td>
<td>0.573</td>
</tr>
<tr>
<td></td>
<td>ELECTRA</td>
<td>0.826</td>
<td>0.890</td>
<td>0.304</td>
<td>0.614</td>
</tr>
<tr>
<td></td>
<td>ELECTRA+</td>
<td>0.861</td>
<td>0.914</td>
<td><strong>0.329</strong></td>
<td><strong>0.636</strong></td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td>0.842</td>
<td>0.902</td>
<td>0.259</td>
<td>0.596</td>
</tr>
<tr>
<td>UMS</td>
<td>BERT</td>
<td>0.843</td>
<td>0.902</td>
<td>0.310</td>
<td>0.622</td>
</tr>
<tr>
<td></td>
<td>BERT+</td>
<td><strong>0.875</strong></td>
<td><strong>0.923</strong></td>
<td>0.363</td>
<td>0.656</td>
</tr>
<tr>
<td></td>
<td>ELECTRA</td>
<td>0.854</td>
<td>0.910</td>
<td>0.397</td>
<td>0.668</td>
</tr>
<tr>
<td></td>
<td>ELECTRA+</td>
<td><strong>0.875</strong></td>
<td><strong>0.922</strong></td>
<td><strong>0.437</strong></td>
<td><strong>0.692</strong></td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td>0.862</td>
<td>0.914</td>
<td>0.377</td>
<td>0.660</td>
</tr>
</tbody>
</table>

*Lower is better*
Ablation Study

- One auxiliary task (i.e., 3 > 2 ≈ 4)
- Two auxiliary tasks (i.e., 5 ≈ 7 > 6)
- Overall, DEL > INS ≈ SRCH
- Improvement of 2.8% w.r.t. $R_{10@1}$

<table>
<thead>
<tr>
<th>Auxiliary Tasks</th>
<th>$R_{10@1}$</th>
<th>$R_{10@2}$</th>
<th>$R_{10@5}$</th>
<th>MRR</th>
<th>$R_{10@1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 None</td>
<td>0.826</td>
<td>0.908</td>
<td>0.978</td>
<td>0.890</td>
<td></td>
</tr>
<tr>
<td>2 INS</td>
<td>0.836</td>
<td>0.917</td>
<td>0.980</td>
<td>0.897</td>
<td>+ 1.0%</td>
</tr>
<tr>
<td>3 DEL</td>
<td>0.848</td>
<td>0.924</td>
<td>0.983</td>
<td>0.905</td>
<td>+ 2.2%</td>
</tr>
<tr>
<td>4 SRCH</td>
<td>0.834</td>
<td>0.915</td>
<td>0.981</td>
<td>0.896</td>
<td>+ 0.8%</td>
</tr>
<tr>
<td>5 INS + DEL</td>
<td>0.853</td>
<td>0.927</td>
<td>0.984</td>
<td>0.909</td>
<td>+ 2.7%</td>
</tr>
<tr>
<td>6 INS + SRCH</td>
<td>0.841</td>
<td>0.920</td>
<td>0.982</td>
<td>0.901</td>
<td>+ 1.5%</td>
</tr>
<tr>
<td>7 DEL + SRCH</td>
<td>0.852</td>
<td>0.927</td>
<td>0.983</td>
<td>0.908</td>
<td>+ 2.6%</td>
</tr>
<tr>
<td>8 INS + DEL + SRCH</td>
<td>0.854</td>
<td>0.929</td>
<td>0.984</td>
<td>0.910</td>
<td>+ 2.8%</td>
</tr>
</tbody>
</table>
Visualization

(a) Insertion  (b) Deletion  (c) Search
Conclusion

- Pointed out the limitations of existing works based on PrLMs, such as BERT in retrieval-based multi-turn dialog systems.
- Proposed highly effective utterance manipulation strategies (UMS) for multi-turn response selection.
- UMS are fully applied in self-supervised manner and can be easily incorporated into existing models.
- New state-of-the-art results on multiple public benchmark datasets.
Thank you

Our code is publicly available at