Taesun Whang
Natural Language Processing & Artificial Intelligence
(+82) 10-4905-1211 | taesunwhang@gmail.com | taesunwhang

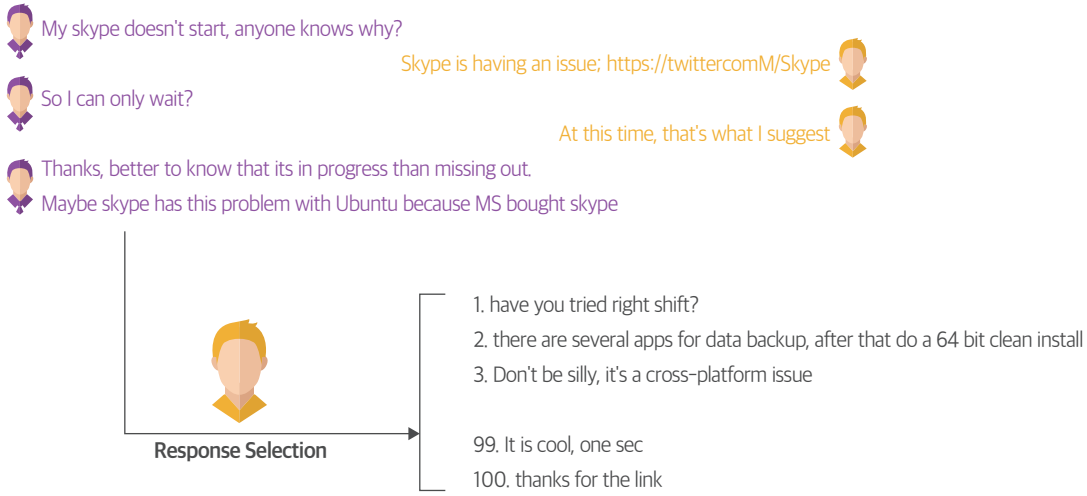
Multi-turn Response Selection

Utterance Manipulation Strategies for Multi-turn Response Selection

- **Whang, T.***, Lee, D.*, Oh, D., Lee, C., Han, K., Lee, D. H., & Lee, S. (2021). Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In Proceedings of 35th AAAI Conference on Artificial Intelligence (pp. 14041-14049).
- Github: https://github.com/taesunwhang/UMS-ResSel

01. 개요

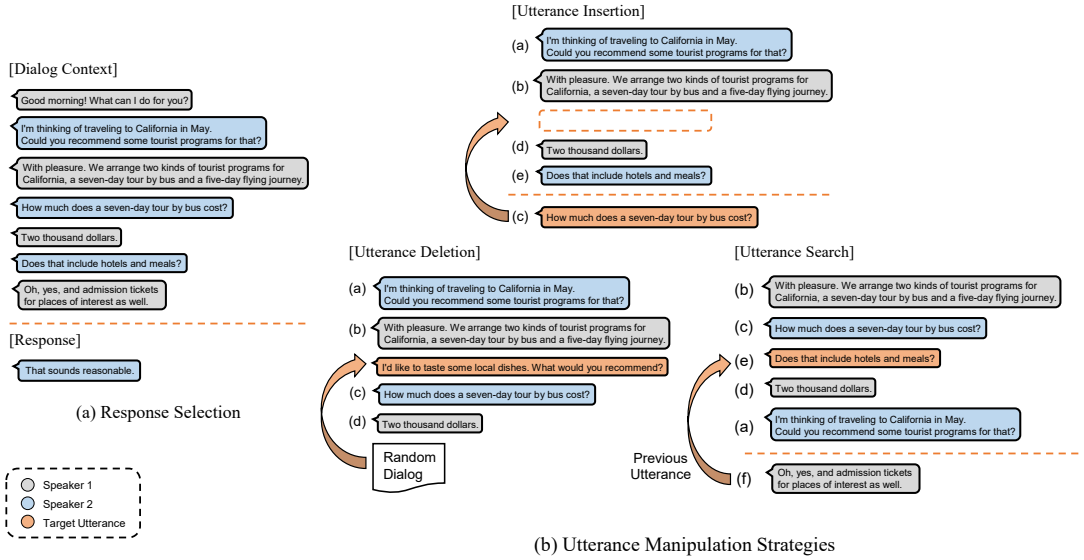
- Response Selection은 검색 기반 대화 시스템에서 시스템의 자연스럽고 정확한 응답을 위해 주어진 응답 후보(candidates) 중에서 대화에 이어지는 응답을 찾아내는 것입니다.
- 대화 문맥 정보를 바탕으로 주어진 대화에 가장 밀접하고 관련성이 높은 응답을 사용자에게 제공해 주는 것을 목표로 하며, 챗봇을 위한 대화 시스템 분야에서 생성 기반 모델과 더불어 많은 연구가 진행되고 있습니다. 기존의 연구들로는 RNN 기반의 모델에서부터 Attention 기반의 대화-응답 matching 모델 등이 있으며, 최근에는 Transformer Encoder를 기반으로 하는 연구가 좋은 성능을 보여 주었습니다.
- 아래 예시는 검색 기반 대화 시스템에서 대화에 이어지는 다음 발화를 예측하는 예시입니다. 본 예시에서는 Ubuntu 운영체제에서 문제가 발생한 질문자와 해결책을 제시하는 답변자의 대화를 예로 들었으며, 질문자의 마지막 발화에 어울리는 답변자의 예상 답변을 선택하는 것이 Response Selection task의 목적입니다.



02. 연구내용

- 최신 언어 모델 기반한 Response Selection 모델은 대화와 응답 후보군을 입력받으면, 후보 문장의 적정성 여부를 binary classification한 결과를 내놓습니다. 본 연구에서는 의미적 유사도를 기반으로 점수를 내는 언어 모델의 특성상, 응답으로 적절하지 않은 문장에 정답보다 더 높은 점수를 부여하는 경향성을 보이는 기존 방식의 한계를 지적했습니다. 한가지 예로, LM에 기반한 최신의 응답 선택 모델이 대화에 이어지는 응답보다, 대화의 맥락에 적절하지 않지만 의미적 유사도가 높은 문장 (대화 내 존재하는 문장)에 더 높은 점수를 부여하는 것을 확인하였습니다.
- 본 연구에서는 기존의 한계를 극복하고자 UMS (Utterance Manipulation Strategies)를 제안했습니다. 이 기법은 대화에서 특정 발화가 어느 위치에 삽입돼야 하는지(insertion), 현재 대화 흐름에서 어떤 발화가 올바르지 않은지(deletion), 특정 발화의 바로 이전 발화의 위치가 어딘지(search)를 배우는 3가지 task로 구성됩니다. Self-supervised learning을

통해 사람이 따로 라벨링 작업을 할 필요가 없고, 기존의 응답 선택 모델을 따로 조정할 필요 없이 fine-tuning단계에서 joint-training을 진행합니다.



03. 실험 및 결과

- 본 연구에서는 Multi-turn Response Selection에서 주요 benchmark 데이터셋으로 활용되는 Ubuntu Corpus V1 (English), Douban Corpus (Chinese), E-commerce Corpus (Chinese), 그리고 새로 구축한 Kakao Corpus (Korean)에 대해서 UMS 방법에 대한 성능 평가를 진행 하였습니다. 기존 RNN, Attention, Transformer 기반의 모델에서부터 pre-trained LM 기반의 모델들까지 baseline으로 설정하고 성능을 비교하였습니다.
- pre-trained LM의 경우 BERT와 ELECTRA 두 가지 언어 모델에대해서 실험을 진행하였고, 기존 baseline 모델들과 공정한 비교를 위해 각 언어모델을 task-specific corpus에 post-training을 진행한 모델 (BERT+, ELECTRA+)에 대해서도 평가를 진행하였습니다.
- 실험 결과, UMS 방법을 적용하였을 때 모든 benchmark 데이터셋에서 주목할 만한 성능 향상을 보여주었습니다. 특히 UMS_{BERT+} 모델이 기존 baseline과 비교했을 때 state-of-the-art (SOTA)의 성능을 보여주었습니다.

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

- 추가로 adversarial 실험을 설계하여 UMS 방법을 통해 학습한 모델이 response selection task에 robust한 것을 검증하고자 하였습니다. adversarial 실험은 대화 내 임의의 한 발화를 response candidate 안에 포함시켜 모델의 candidate ranking 결과를 평가하였습니다. 실험 결과, UMS를 통해 학습한 모델이 그렇지 않은 모델보다 adversarial test set에서 더 좋은 성능을 보여주었습니다.