Learning to Retrieve Entity-Aware Knowledge and Generate Responses with Copy Mechanism for Task-Oriented Dialogue Systems

Chao-Hong Tan[1]*, Xiaoyu Yang[2]*, Zi'ou Zheng[2]* , Yufei Feng[2]*, Tianda Li[2]*,
Jia-Chen Gu[1], Quan Liu[1], Dan Liu[1], Zhen-Hua Ling[1]†, Xiaodan Zhu[2]†

*Equal contribution.
†Corresponding authors.

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China
[2]ECE & Ingenuity Labs, Queen's University, Kingston, Canada

Introduction

This challenge can be separated into three subtasks,
(1) knowledge-seeking turn detection,
(2) knowledge selection,
(3) knowledge-grounded response generation.

We use pre-trained language models, ELECTRA and RoBERTa, as our base encoder for different subtasks.
Subtask 1 and 2: coarse-grained information like domain and entity are used.
Subtask 3: latent variable at encoder and generate responses combined with copy mechanism.

Our proposed system ranks second under objective metrics and ranks fourth under human metrics.

Knowledge-seeking Turn Detection

Knowledge-aware ELECTRA:
(1) conduct an entity matching for each question, concatenate the domain label to the end of history
(2) add a one-bit knowledge flag to the end of the final hidden vector of the token <bos> to classify

Knowledge Selection

(1) **Retrieve and Rank Model:**

- Retrieve candidate knowledge base entities using text-matching based heuristics.
- Rank the candidate knowledge snippets by finetuning a Roberta model.

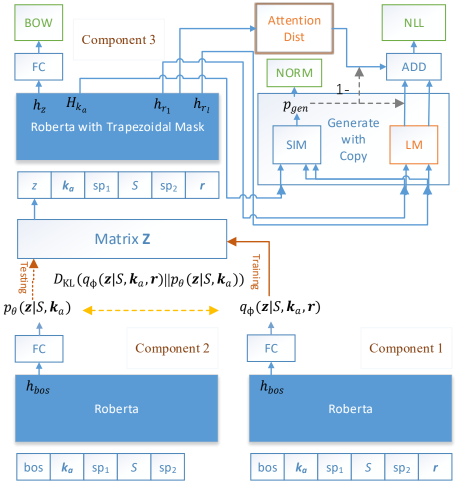
(2) **Three-step Model**:

- Predict domain, entity and document of the desired knowledge snippets with separate Roberta models.
- Augment the training data for better performance by leveraging the domain text in the knowledge base.

(3) **Ensemble Model**:

- Ensemble the Retrieve & Rank model with Three-step model together.

Knowledge-grounded Response Generation


BOW
FC
Component 3
Attention Dist
NLL
NORM
ADD
1-
h_z
H_{k_a}
h_{r_1}
h_{r_l}
p_{gen}
Roberta with Trapezoidal Mask
Generate with Copy
SIM
LM
z
k_a
sp₁
S
sp₂
r
Matrix Z
Testing
Training
$D_{KL}(q_\phi(\boldsymbol{z}|S,\boldsymbol{k}_a,\boldsymbol{r})||p_\theta(\boldsymbol{z}|S,\boldsymbol{k}_a))$
$p_\theta(\boldsymbol{z}|S,\boldsymbol{k}_a)$
$q_\phi(\boldsymbol{z}|S,\boldsymbol{k}_a,\boldsymbol{r})$
FC
Component 2
FC
Component 1
h_{bos}
Roberta
h_{bos}
Roberta
bos
k_a
sp₁
S
sp₂
bos
k_a
sp₁
S
sp₂
r

Model architecture. The trapezoidal mask is to prevent response information leakage from our bi-directional encoder.

(1) **Latent variable**: Capture the response information, work like shortcut connection, optimize under KLD
(2) **Knowledge Copy**: Combine auto-regressive generation with copy mechanism by weight adding.
(3) **Segmented Response Generation(SRG)**: Generate knowledge part and greedy part separately
(4) **Modified Beam Search(FFBS)**: Fix different first word of response then do normal beam search for each group
(5) **Post-processing Strategies**: Plus semantic similarity score and minus word similarity score between response candidates from FFBS and knowledge to NLL score.

Evaluation(On Test Set)

Model	Knowledge Detection Precision/Recall/F1	Knowledge Selection MRR@5/Recall@1/Recall@5	Knowledge Grounded Generation BLEU-1/2/3/4	 METEOR	 ROUGE-1/2/L
Baseline	0.9933 / 0.9021 / 0.9455	0.7263 / 0.6201 / 0.8772	0.3031 / 0.1732 / 0.1005 / 0.0655	0.2983	0.3386 / 0.1364 / 0.3039
Ours	0.9933 / 0.9677 / 0.9803	0.9195 / 0.8975 / 0.9460	0.3779 / 0.2532 / 0.1731 / 0.1175	0.3931	0.4204 / 0.2113 / 0.3765

The evaluation results (objective/human metrics) on test dataset of our all three subtasks. Compared with the baseline, our model achieves huge improvement in all three subtasks. In addition, human metrics show that the performance of our model is close to human.

Model	Accuracy	Appropriateness	Average
Baseline	3.7155	3.9386	3.8271
Ours	4.3793	4.2755	4.3274
Ground-Truth	4.5930	4.4513	4.5221

Analysis(On Validation Set)

Subtask-1 Model	Precision	Recall	F1-score
Baseline(GPT2)	0.957	0.966	0.961
Xlnet	0.944	0.983	0.963
Roberta	0.947	0.993	0.970
ELECTRA	0.950	0.994	0.972
Our	**0.996**	**0.999**	**0.998**

Subtask-2 Model	MRR@5	Recall@1	Recall@5
Roberta (baseline)	0.8691	0.8058	0.9428
Retrieve & Rank	**0.9747**	0.9622	**0.9880**
Three-step w/ data aug.	0.9739	0.9660	0.9828
Three-step w/o data aug.	0.9692	0.9607	0.9783
Our Ensemble Model	0.9743	**0.9678**	0.9813

Subtask-3 Model	BLEU-4	METEOR	ROUGE-L
Ours w/ PP	0.1450	0.4526	0.4256
Ours w/o PP	**0.1512**	0.4534	0.4317
–FFBS	0.1490	**0.4536**	**0.4342**
–SRG	0.1473	0.4485	0.4311
–Copy	0.1424	0.4322	0.4194
–Z	0.1327	0.4210	0.4096

(1) **Subtask-1**: Compared with the original ELECTRA, the F1-score of our model has **2.6%** improvement.
(2) **Subtask-2**: Retrieve & Rank model has a better recall@5 while Three-step model with data augmentation provides better recall@1, especially better in domain and entity prediction. The ablation test on data augmentation also shows that the data augmentation technique is helpful at the domain and entity level.
(3) **Subtask-3**: Latent z contributes about **1%** to on each metrics, while knowledge-copy mechanism wins about **0.5%** on BLEU-4, about **1.6%** on METEOR, about **1.2%** in ROUGE-L. Note that the use of post-processing could increase human evaluation scores, while the objective scores may reduce.

Generation Example

Context	By the way, what hours is the **station** where the **train** departs open?
Knowledge(Body)	The hours of operation are **Monday-Saturday 05:00 - 23:00**, and on **Sundays 07:00 - 22:55**.
Ground-Truth	**Monday through Saturday 5 AM to 11:00 PM**, **Sundays 7 AM to 11:55 PM**. Can I help you with any more information?
Ours w/ PP	Hours of operation at the **train station** are **Monday-Saturday 05:00-23:00** and on **Sundays 07:00 - 22:55**. Anything else I can do for you?
Ours w/o PP	The station opens at **5:00 am Monday-Saturday** and **7:00 on Sundays**. Anything else I can do for you?
Ours w/o Copy	The station opens at **5 am Monday-Saturday** and **7 am on Sunday**. Anything else I can do for you?

Conclusion

(1) This paper describes our overall system that is evaluated in Track 1 of DSTC 9.
(2) Pre-trained language models are used as our base encoder, and task-specific components are applied to improve performance.
(3) In the released evaluation results, we rank second under objective metrics and rank fourth under human metrics.

Contact

Email:
chtan@mail.ustc.edu.cn
{xiaoyu.yang, ziou.zheng, feng.yufei, tianda.li}@queensu.ca
Github:
github.com/lxchtan/DSTC9-Track1