End-to-end Gated Self-attentive Memory Network for Dialog Response Selection

Shuo Sun[1*] , Yik-Cheung Tam[2*] , Jie Cao[3*], Canxiang Yan[2] , Zuohui Fu[4], Cheng Niu[2] , Jie Zhou[2]

[1] Center for Language and Speech Processing, Johns Hopkins University
[2] WeChat AI - Pattern Recognition Center, Tencent Inc.
[3] School of Computing, University of Utah
[4] Department of Computer Science, Rutgers University

* Equal contribution. Work done when Shuo Sun, Jie Cao and Zuohui Fu were interns at WeChat AI - Pattern Recognition Center, Tencent inc.

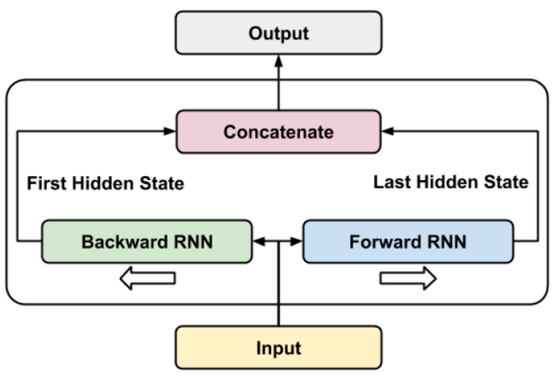
INTRODUCTION

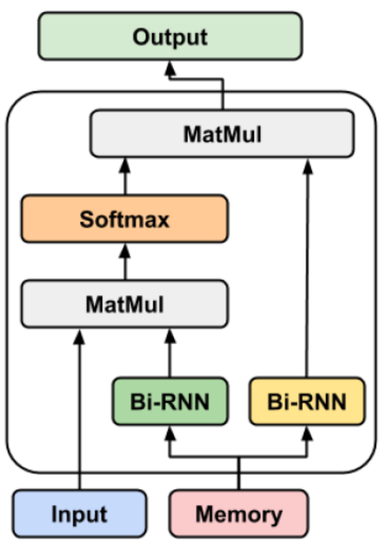
- Gated Self-attentive Memory Network encodes dialog history and external domain knowledge in an end-to-end trainable manner.
- Novelty is that each utterance in the memory is enhanced with self-attention building the connection between dialog history and external domain.
- Our approach ranks at **the second place for both student advising and Ubuntu subtasks integrated with external domain knowledge.**

NETWORK MODULES

Bi-RNN based Utterance Encoding

- Bi-RNN is used as sentence encoders.
- Output is concatenation of the first hidden state of backward RNN and last hidden state of forward RNN.



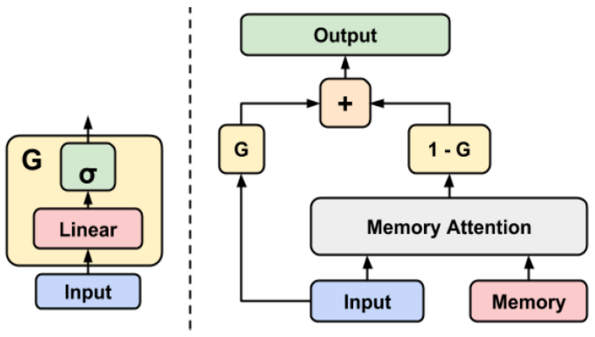


Memory Attention

- Attention mechanism is used to "retrieve" relevant memory vectors.
- Output is a weighted sum of the memory vectors.

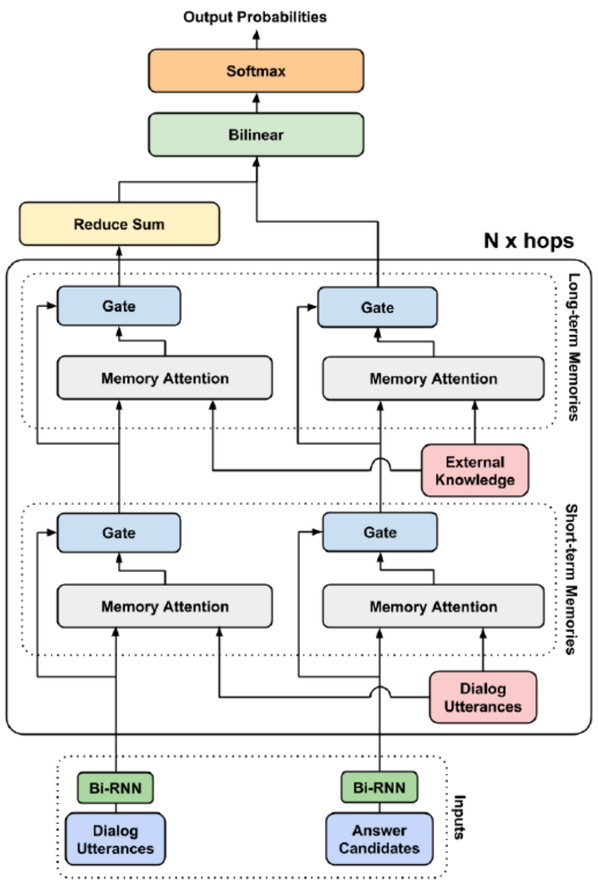
Gated Memory Attention

- Input and the retrieved memories are combined via a gating mechanism.
- Regulates the degree of enhancement.
- Prevents information overload.



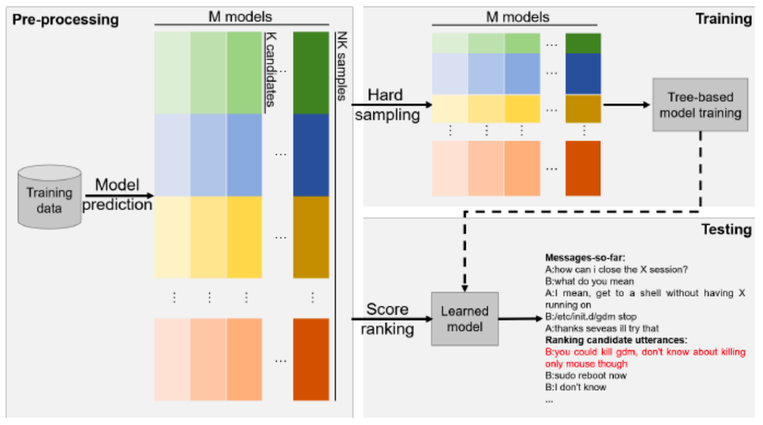
GATED SELF-ATTENTIVE MEMORY NETWORK

- Short-term memories and long-term memories are stacked sequentially.
- Process is repeated in a multi-hop fashion.
- Weights are not shared across hops.
- Memory enhanced dialog utterances are summed into a single vector.



ENSEMBLE LEARNING

- XGBoost is employed to train a second stage meta classifier.
- Scores from many GSMN models with various hyperparameters are used as features.
- Training instances with mean prediction scores outside the range [0.001,0.95] are filtered out.



EXPERIMENTAL RESULTS

Model	R@1	R@10	R@50	MRR
Dual-LSTM baseline	0.062	0.296	0.728	N/A
HGRU baseline	0.164	0.632	0.922	0.299
SMN w/ 1 hop	0.218	0.642	0.956	0.337
2 hops	0.198	0.620	0.938	0.320
3 hops	0.206	0.648	0.942	0.333
GSMN w/ 1 hop	0.220	0.632	0.954	0.343
2 hops	0.214	0.644	**0.960**	0.338
3 hops	0.214	0.628	0.956	0.335
1 hop + EK	0.220	0.644	0.956	0.343
2 hops + EK	**0.224**	**0.654**	0.944	**0.354**

Baseline and GSMN results on the Flex advising dev set. EK denotes external knowledge.

Model	R@1	R@10	R@50	MRR
Dual-LSTM baseline	0.083	0.360	0.804	N/A
SMN w/ 1 hop	0.326	0.671	0.952	0.445
SMN w/ 2 hop	0.337	0.686	0.956	0.455
GSMN w/ 1 hop	0.379	0.733	0.973	0.497
2 hops	0.389	**0.755**	0.972	0.508
3 hops	**0.398**	0.761	**0.976**	**0.515**

Baseline and GSMN results on the Ubuntu dev set. (Single model)

EFFECTIVENESS OF SELF-ATTENTION AT UTTERANCE LEVEL

Model	R@1			R@5			R@10			MRR		
	LU	AU	Gain	LU	AU	Gain	LU	AU	Gain	LU	AU	Gain
1 hop	0.194	0.220	▲0.026	0.592	0.643	▲0.051	0.910	0.954	▲0.044	0.317	0.343	▲0.026
2 hops	0.208	0.214	▲0.006	0.596	0.644	▲0.048	0.920	0.960	▲0.040	0.326	0.338	▲0.012
3 hops	0.208	0.214	▲0.006	0.616	0.628	▲0.012	0.928	0.956	▲0.028	0.329	0.335	▲0.006
1 hop + EK	0.172	0.220	▲0.048	0.578	0.644	▲0.066	0.910	0.956	▲0.046	0.294	0.343	▲0.049
2 hops + EK	0.186	0.224	▲0.038	0.602	0.654	▲0.052	0.932	0.944	▲0.012	0.319	0.354	▲0.035

LU: restricted the inputs to only the last utterance of a dialog history.
AU: use almost all utterances of a dialogue history (last 6 utterances)

- Significant performance gains across all evaluation metrics and model settings were observed.
- This confirms that modeling co-references among dialog utterances via self-attention was effective.

OFFICIAL EVALUATION RESULTS

Measure	Ubuntu	Advising - Case 1	Advising - Case 2
Recall@1	0.475	0.494	0.18
Recall@10	0.814	0.85	0.562
Recall@50	0.978	0.98	0.94
MRR	0.595	0.6078	0.3069

Subtask 1

Measure	Ubuntu	Advising - Case 1	Advising - Case 2
Recall@1	0.504	0.538	0.178
Recall@10	0.827	0.864*	0.608
Recall@50	0.98	0.986	0.944
MRR	0.6172	0.6455*	0.3149

Subtask 5
*** Denotes the 1st-place evaluation results.**

Discussion

- Bi-RNNs in GSMN might have heavily memorized the language styles in training data and failed to adapt to the novel styles in test-case2 .
- Our models could not handle cases where subject and direct object of sentence were reversed.
- For example, *"EECS281 is taken by most students in second semester"* instead of *"Most students take EECS281 in second semester".*

CONCLUSION

- Gated Self-attentive Memory Network effectively integrates external knowledge and dialog history in an end-to-end fashion.
- For future work, we believe that improving the encoding power of dialog history and external domain knowledge as well as their interaction will be crucial for further performance improvement.