Towards Simple but Efficient Next Utterance Ranking

Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin and Emmanuel Morin

LS2N, UMR CNRS 6004, Université de Nantes, France

firstname.lastname@ls2n.fr

1. Overview

- The next utterance ranking task consists of selecting a response for a given context of conversation.
- This category of dialogue systems can produce longer, meaningful and more syntactically correct responses compared to generative dialogue systems.
- Retrieval-based dialogue systems have been successfully applied in industry such as the Microsoft's social bot *XiaoIce* and *AliMe*.

2. Motivations

We believe that the existing retrieval-based dialogue systems suffer from the following drawbacks.

- The complexity of their architectures.
- Their domain dependency.

3. Approach

We propose a **simple**, **domain independent** and **efficient** retrieval-based dialogue system that performs *as good as* the previous complex systems and sometimes *better* while being **conceptually** simpler.

1. Encode the context and the response with a **shared** LSTM and compute their cross product: sequence-level similarity.
2. In parallel, compute a dot product between the embedding matrices and encode it with another LSTM: word-level similarity.
3. Concatenate both vectors and transform them into a probability using a FFNN: response ranking score.

4. Datasets & Metrics

- Ubuntu Dialogue Corpus (UDC): a domain specific corpus of Ubuntu related English chats.
- Douban Conversation Corpus: an open domain corpus extracted from Douban a Chinese social network.
- We used the Recall@k, Precision@1, MRR and MAP as evaluation metrics.

5. Multi-Level Retrieval-Based Dialog System

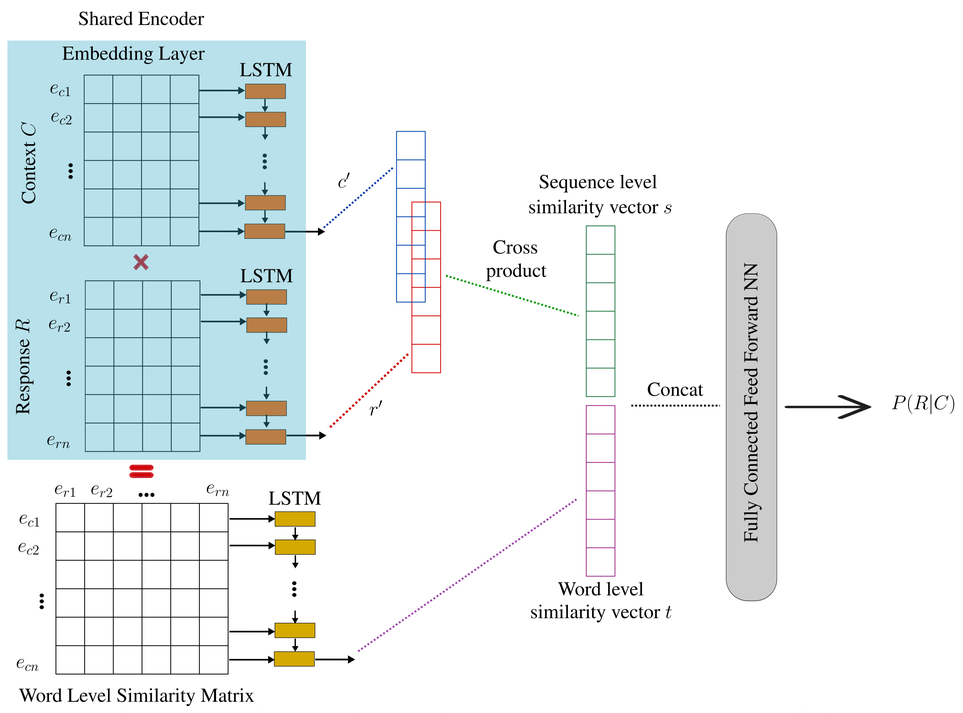


Figure 1: Architecture of our multi-level context response matching dialogue system.

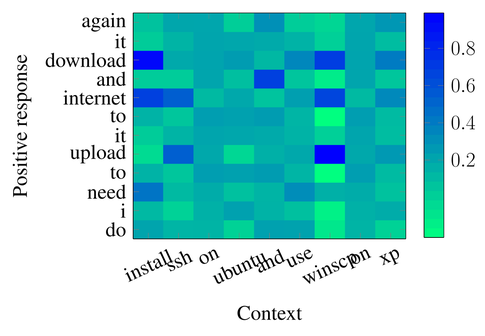
6. Experiments

System	Ubuntu Dialogue Corpus (V1)				Douban Conversation Corpus					
	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5	P@1	MAP	MRR
TF-IDF	0.659	0.410	0.545	0.708	0.096	0.172	0.405	0.180	0.331	0.359
LSTM	0.901	0.638	0.784	0.949	0.187	0.343	0.720	0.320	0.485	0.527
BiLSTM	0.895	0.630	0.780	0.944	0.184	0.330	0.716	0.313	0.479	0.514
DL2R	0.899	0.626	0.783	0.944	0.193	0.342	0.705	0.330	0.488	0.527
Multi-View	0.908	0.662	0.801	0.951	0.202	0.350	0.729	0.342	0.505	0.543
$SMN_{dynamic}$	0.926	0.726	0.847	0.961	0.233	0.396	0.724	0.397	0.529	0.569
DAM	**0.938**	**0.767**	**0.874**	**0.969**	0.254	0.410	0.757	**0.427**	**0.550**	**0.601**
Our system	0.935	0.763	0.870	0.968	**0.255**	**0.414**	**0.758**	0.418	0.548	0.594
Only sequence similarity	0.917	0.685	0.825	0.957	0.209	0.357	0.702	0.358	0.500	0.543
Only word similarity	0.926	0.744	0.853	0.956	0.223	0.370	0.719	0.373	0.513	0.556

- Our system outperforms several complex baselines with a good margin.
- Our system neither matches each context turn with the candidate response nor uses complex **cross-** and **self**-attention in addition to matching and accumulation mechanisms but achieves almost the same performance as the DAM.

7. Visualization

- We visualize the Word Level Similarity Matrix.
- Important (key) words in the context and the response were recognized by our system and were given higher scores.



8. System Ablation

- Having only one similarity level results in lower scores.
- Even with only one similarity level, our system outperforms several baselines.
- Having only word similarity information yields to a better performance compared to only having sequence similarity.
- Both similarity levels are complementary.

9. Error Analysis

We performed a human evaluation on 200 randomly selected test samples from UDC (V1).

Error class	Percentage
Functionally equivalent	31%
Semantically equivalent	20%
Out of context	35.5%
Very general responses	13.5%

Almost 50% of errors were due to the negative sampling approach (random selection of negative responses) used in the construction of the corpus.

10. Conclusion

- We proposed an end-to-end domain independent retrieval-based dialog system that matches the context with the correct response on **two levels**.
- We showed that a simple end-to-end architecture is more efficient than complex state-of-art systems.
- The visualization and the system ablation studies shows and confirms the importance of extracting both similarity levels.

11. Perspectives

- We plan to improve our system while keeping the simplicity of the approach.
- We encourage the scientists to care about the complexity of their architectures.
- We want to enrich the text with discursive information such as dialogue acts and rhetorical relations.
- The results of the error analysis shows important information about the impact of the dataset on the performance of the system.

Code and data

https://github.com/basma-b/multi_level_chatbot

Contact Information

- Web: https://basma-b.github.io/
- Email: basma.boussaha@univ-nantes.fr

LS2N LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES

UNIVERSITÉ DE NANTES

EIEL ing 2019