Next utterance ranking based on context response similarity

Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin and Emmanuel Morin
Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes, 44322 Nantes Cedex 3, France
Email: (basma.boussaha, nicolas.hernandez, christine.jacquin, emmanuel.morin)@univ-nantes.fr

***Abstract*—Building dialogue systems that converse with humans in order to help them in their daily tasks is being a priority. Some systems converse by generating dialogues word by word whereas others retrieve the best utterance among a set of candidate responses. These retrieval systems rank the candidate responses by their relevance to the history of the conversation (context), the best response is then chosen. Approaches based on deep neural networks performed well on this task. In this work, we improve a state of the art approach based on an LSTM dual encoder and propose a new response retrieval dialogue system. Based on syntactic and semantic similarities between the context and the response extracted from word embeddings, our approach learns to match the context with the best response. Experimental results on the Ubuntu Dialogue Corpus show an important improvement of about 7%, 6% and 2% on Recall@(1, 2 and 5) compared to the best state of the art system.**

***Index Terms*—automatic assistance, dual encoder, LSTM, dialogue systems.**

I. INTRODUCTION

The exponential growth of Internet services and the number of users is making managing them a complex task. It is being crucial to develop machines that assist humans in doing their daily tasks. One important thing is to develop machines able to converse with humans and understand their needs. Using such machines could remarkably improve task performance and reduce human effort. The recent progress in computing power, the availability of large datasets and the new designed machine learning models helped developing natural language understanding and generating systems. These systems can understand and converse with humans by either generating responses word by word *(generative systems)* or by searching for the best answer among a set of predefined candidate responses *(response retrieval systems)*.

Conversations contain multiple exchanges, called utterances, between two or more people. In order to produce an adequate utterance to a given conversation, it is important to consider all the context utterances which make the task more difficult. Many recent works addressed the problem of predicting the next utterance in written conversations. Some of them select the best utterance by considering the one that matches all the context utterances. Whereas others reduce the hole conversation to its last utterance.

Given the example in Table I, a response retrieval system attempts to rank the candidate responses in order to find the next utterance of the given history of conversation. In this case the first response should be ranked before the second one. Note that it is important to consider all the context utterance and not only the last one since almost all the dialogue utterances share common words (written in bold) and common sens. Considering all these utterances yields to a better performance of the candidate responses ranking.

	Context
utterance 1	Hi, I can not longer access the graphical **login screen** on ubuntu 12.04
utterance 2	what exactly happen?
utterance 3	I can't remember the error message, would it have auto-logged to a file or should I reboot quick?
utterance 4	you mean it won't **automaticaly start** and what happen then?
utterance 5	it just stop at a text **screen**, but I can access the command line **login** via alt F1-6, and **start** x **manually** there. I think it might me **lightdm** that's break but I'm not sure
	Candidate responses
response 1	for me **lightdm** often won't start **automatically** either. It show me console tty1 instead and I have to **start lightdm manually**
response 2	what about sources.list ?

TABLE I
EXAMPLE OF A TECHNICAL CONVERSATION BETWEEN TWO USERS EXTRACTED FROM THE UBUNTU DIALOGUE CORPUS [1]. THE FIRST CANDIDATE RESPONSE IS GOOD WHEREAS THE SECOND ONE IS BAD.

The difficulty of the next utterance ranking task resides in the fact that the context and the response share common information that is, in most cases not implicit. According to [2], the challenges of this task are (1) how to identify important information (words, phrases, and sentences) in the context and to match this information with the other information in the response and (2) how to model the relationships between the context utterances. Existing works either use complex architectures to capture utterance level information and complex response matching mechanism or neglect utterance level information and consider the context as one long utterance (by concatenating all the utterances).

In this work, we improve an utterance ranking system based on dual encoder [1]. We encode the context and the candidate response into two separate vectors following the same process as them. The encoder is a shared recurrent neural network with Long Short-Term Memory (LSTM) cells [3] that learns a transformation of the context and the response into fixed size vectors. We compute a similarity vector as a cross product between these two vectors. Then the utterance ranking score is obtained by learning the transformation of this product vector using a fully connected layer and a sigmoid function. At the end, our model outputs a probability that a candidate utterance is the next utterance of the given context. We use this probability to rank all the candidate utterances. This new

ranking approach allows capturing the common semantic and syntactic features between the context and the utterance which is important to ditinguish between good utterances from bad ones. We evaluated our approach on a large dialogue corpus of Ubuntu chat and we followed [1] in the choice of Recall@k as an evaluation metric. Experimental results showed a significant improvements compared to the best state of the art system.

The remainder of this work is as follows: section II investigates works around conversational systems. Section III describes the problem and the architecture of our system. In section IV we present the dataset on which we evaluated our system, results and comparison with state of the art systems. Finally we conclude in section V with some perspectives of future work.

II. RELATED WORK

Recently many works were interested in constructing task-oriented conversational systems. We distinguish two categories of dialogue systems: generative and response retrieval systems [4]. Most of the generative systems are based on the *sequence-to-sequence* architecture of [5] in order to generate dialogues word by word [6], [7], [8]. Despite the capacity of these systems to generate customized responses for each conversation context, they tend to generate short and general responses [9]. Thus, they prefer generating for example *"I don't know"* and *"Good !"* in most of the times. This is due essentially to the lack of diversity in their objective function [10]. In technical assistance, a dialogue system is supposed to generate accurate and customized responses to help the user solving his specific problem. In the other hand, response ranking systems are able to provide more accurate and syntactically correct utterances in case they have been already seen. This category of dialogue systems is in the center of our interest in this work.

[1] built an utterance ranking system based on *dual encoder*. Their main idea is to use word embeddings to present the text input (the context and the candidate utterance). Then they encode separately the context and the utterance embeddings into two fixed size vectors. These vectors contain a compressed information of the hole context and response independently of their initial length (in number of words). A dot product is computed between a learned parameter matrix and these two vectors. The product is transformed into a probability used as a ranking score between the candidate utterances. Moreover, two variants of the dual encoder with RNNs and LSTMs cells were implemented and evaluated in the same work. An extension of this study was realized by [11] in which an ensemble system was deployed. It regroups 11 LSTMs, 7 Bi-LSTMs and 10 CNNs trained with different hyper-parameters.

Inspired by the human brain, [12] incorporated domain knowledge into their system in order to improve context and response modeling. They introduced for the first time a new cell called *r-LSTM* which has an extra gate called *Recall Gate*. As indicated by its name, this cell helps in memorizing information about domain knowledge in addition to encoding the context and the response with the same process of [1] explained previously. [2] designed a response ranking system which considers this time the context utterances separately. From each utterance in the context, they extracted two information: the word level and utterance level similarities. These information are encoded using a succession of convolution and pooling and then accumulated using GRU units [13]. At the end a probability is computed using softmax on the weights of the accumulating GRU.

Unlike all these works, [14], [15] only considered the last utterance of the context. [15] used the conversation topic as an extra information to improve the quality of the selected response. The conversation topic words were extracted from both the last utterance of the context and the response using the state-of-the-art topic Twitter LDA model for short texts [16]. The context, the response and their topic words, were embedded into a vector space using Neural Tensor Networks [17], [18]. The response ranking score is computed as in the works presented previously using a softmax function.

In this work, we adopt the first neural system that addressed the next utterance ranking task: the dual encoder [1]. We build an utterance ranking system that captures context response similarities between the context and the response. Unlike [11] and [12], our approach is simple and can be easily adapted to other domains since it does not require domain related information. Our approach is trained in end-to-end fashion without any need to an extra module trained offline to generate extra information. Moreover, the problem of reproducibility of [2] as explained in section IV did not help us using their system. These were the essential motivations of our choice to improve the initial system.

III. OUR MODEL

A. PORBLEM FORMALIZATION

Given a conversation context C between two users as a succession of n utterances u_i such as $C = \{u_1, u_2, u_3, ...u_n\}$. The problem consists of selecting the next utterance u_{n+1} called the response among a set of m candidate utterances $u_{n+1} \in \{r_1, r_2, r_3, ..., r_m\}$. We define the problem as a ranking task in which we want to order candidate responses by their increasing suitability to the conversation context. The utterance with the highest score is then chosen as the next utterance.

B. SYSTEM ARCHITECTURE

Inspired by the system of [1], we propose an improved architecture of the LSTM dual encoder trained in end-to-end fashion. The idea consists of representing the context C and the response R using word embeddings. Then these word embeddings $e_1, e_2, .., e_j$ are given in chronological order to a recurrent neural network with LSTM cells called encoder. The hidden layer of this network is updated each time a word embedding is fed. The aim of the encoder is to provide a fixed length vector for each input text which has a variable size. This process is described in figure 1 with the dark frame, it is the same as the one deployed in [1]. At the end we get the hidden layer of the encoder C' and R' which represent in this case the hole context and the response respectively.

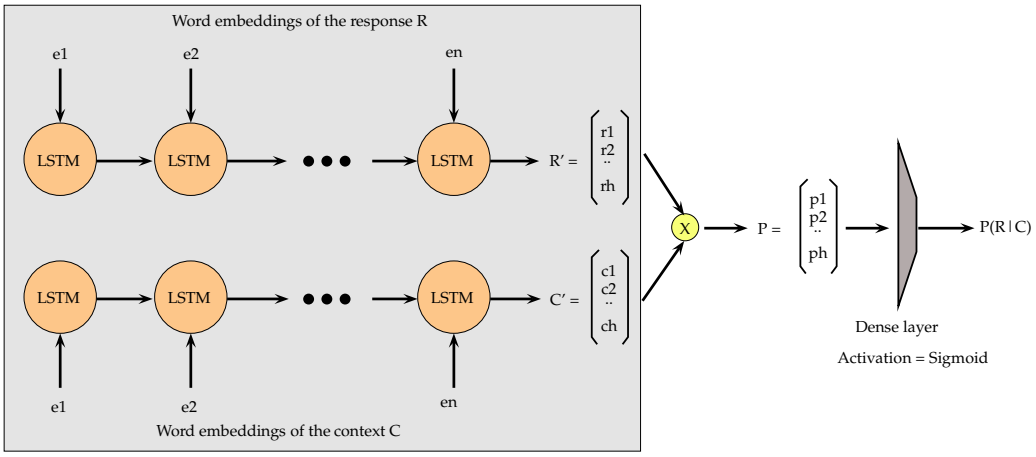

Fig. 1. Architecture of our system based on LSTM dual encoder

Unlike the approach of [1], in which the candidate utterance score is computed as the dot product between C', R' and a matrix of learned parameters M, we compute the score differently. In our approach the score is computed as a cross product P between C' and R' which reflects the similarity between the context and the candidate response. The similarity vector P is fed into a fully connected layer and transformed into probability score using a sigmoid function. This architecture is motivated by the fact that the context and the response share common concepts (common words, sens, etc). These concepts are first captured with word embeddings and then using the encoders and the similarity, we capture semantic and syntactic similarities.

The advantages of our system compared to the state of the art ones are: (1) we do not require any external module to provide extra information such us context and response topics or related knowledge unlike [12] and [15]; (2) the architecture is trained in end-to-end where the classification error is back-propagated through the network to improve the training process from the embedding layer to the probability prediction; (3) we designed an utterance ranking system that is domain independent. It means we can adapt this same architecture from one assistance domain to another, for example from Ubuntu to Visa and immigration assistances by simply changing the dataset.

IV. EVALUATION AND RESULTS

A. Ubuntu Dialogue Corpus

[1] collected a large corpus of Ubuntu dialogues called the Ubuntu Dialogue Corpus (UDC). The corpus contains around one million written conversations between two users who exchanged dialogues at lest three times. These conversations are issued from the chat logs of the channel *#Ubuntu* on the Freenode Internet Relay Chat (IRC)[1]. Conversations on this source are multi users, some heuristics were applied on these conversations in order to extract two-user discussions. Some of these heuristics are based on user name mention to identify recipient and on a time frame of a fixed duration in order to limit the conversation length and limit extracted discussions to one subject [1].

These conversations are written in English, they address different technical problems related to Ubuntu. The first version V1 of the corpus raised some problems related to the distribution of the data separation through the time and the sampling procedure for the context length in the validation and test sets, etc[2]. These bugs were addressed in the second version V2 of the corpus and hence the results on these two versions are not directly comparable. Table II summarizes statistics on the V2 of the Ubuntu Dialogue Corpus.

# utterances (total)	7,100,000
# turns (total)	5,139,574
# words (total)	100,000,000
Min. # turns per dialogue	3
Avg. # turns per dialogue	4.94
Avg. # word per turn	10.34
# train samples	1,000,000
# test samples	18,920
# validation samples	19,560

TABLE II
STATISTICS OF THE V2 OF THE UBUNTU DIALOGUE CORPUS

The corpus contains 1 million of dialogues for training, 19,560 and 18,920 dialogues for validation and test respectively. Each sample in the train is a triplet *(context, response, label)*. The label is set to "1" if the response is the next utterance of the given the context, else it is set to "0". In the validation and test sets, each sample is composed of a context and 10 candidate responses where one is the *ground-truth response* and 9 are bad responses. The bad responses were randomly sampled from the corpus. The task on this corpus consists of ranking the good response (the ground-truth response) on top of the bad candidate responses. We chose to work on this corpus for mainly two reasons. First,

[1]For the period 2004-2015 available on https://irclogs.ubuntu.com/

[2]More details on https://github.com/rkadlec/ubuntu-ranking-dataset-creator

its dialogue nature contrary to monologue datasets. Second, many utterance ranking systems were evaluated on this dataset, which provide a comparison environment for our approach.

B. Evaluation results

Evaluation of conversational systems is an open research domain, in which there are no standard evaluation metrics [19], [20]. We followed [1], [12], [15], [2] in using *Recall@k* as an evaluation metric for our utterance ranking system. As attested by [19], Recall@k is suitable for the response ranking task. This metric measures the capacity of the system to rank the the good response among the k top best responses retrieved by our system.

While, the approaches of [12], [15], [2] were evaluated on the V1 of the Ubuntu Dialogue Corpus, their evaluation was not performed on the latest version of the corpus. Moreover, they are not comparable between each other because the test set used in each approach is not the same. For these reasons the direct comparison between these approaches is not straightforward. Therefore, we compared our approach to the system of [1]. Even if the results reported on their first paper were obtained on V1 of UDC, they released updated results on the V2, in their second paper [4] and on their project web-page on GitHub[3].

Method	Recall@1	Recall@2	Recall@5
TF-IDF [4]	48,8 %	58,7 %	76,3 %
RNN Dual Encoder [4]	37,9 %	56,1 %	83,6 %
LSTM Dual Encoder [4]	55,2 %	72,1 %	92,4 %
BiLSTM Dual Encoder* [11]	54,2 %	71,6 %	91,9 %
Similarity LSTM Dual Encoder	**62,2** %	**78,0** %	**94,9** %
Similarity BiLSTM Dual Encoder	**62,3** %	**78,2** %	**95,1** %

TABLE III
EVALUATION RESULTS USING RETRIEVAL METRICS RECALL@K. ***Note*** *: WE EVALUATED THE APPROACH OF [11] ON THE V2 INITIALLY EVALUATED ON V1 AND WE REPORTED RESULTS IN THE TABLE

In the table III, the first three rows report the system results of [4] on the V2 of UDC. The BiLSTM in row 4 is the system of [11] that we evaluated on the V2. Our system outperforms with a good margin all these state of the art systems on all Recall@k metrics. Recall@1 is a hard metric which evaluates the capacity if the system to rank the best response on the rank 1/10 which is not easy to perform by an utterance ranking system.

As explained in section III, the difference of the way we compute the ranking score as a similarity between the context and the response, we improve significantly the results. Thus we gain around 7% on Recall@1. Moreover, using Bidirectional LSTMs (BiLSTM) in our approach brings more gain on all Recall@k metrics. With the BiLSTM we encode each input in two different directions using two LSTMs into two vectors. Then we concatenate the vectors to representative vector for each of the context and the response.

[3]https://github.com/npow/ubottu/tree/master/

C. Prediction analysis

We analyzed the predictions made by our system on the test set in order to understand the cases of good and bad predictions. Table IV contains an example that has been successfully classified regarding Recall@k. Our system ranked the best response on top of the candidate utterances. Even though the candidate responses do not share common words with the context, our model was able to recognize the best response and assigned it the highest probability.

	Context		Candidate responses
u1	how do i remove the chat option from the envelope icon at the top of the screen i already delete empathy	**0.98**	**i tried that but it's still there**
		0.25	thank the internet wasnt working because of this
u2	i wouldn't think so	0.22	thank so much
		0.14	sorry not mean for you

TABLE IV
AN EXTRACTED EXAMPLE FROM THE TEST SET. OUR SYSTEM SUCCESSFULLY RANKED THE BEST RESPONSE ON TOP OF THE CANDIDATE UTTERANCES. NOTE THAT IN THIS EXAMPLE WE REDUCED THE NUMBER OF CANDIDATE RESPONSES TO 4 FOR FORMATTING REASONS.

We are interested in error analysis and understanding as possible the reasons of bad predictions made by our system. We randomly chose a test sample on which the system was not able to retrieve the best response as show in table V. In this case, the expected response is *thank you*, whereas our system predicted *it's only annoying when the cursor drag really slowly* to be the best response. Note that the other candidate responses obtained a higher score compared to the ground-truth response.

	Context		Candidate responses
u1	http://www.howtogeek.com how to add screensavers to ubuntu 12.04 see also	0.99	it's only annoying when the cursor drag really slowly
	http://askubuntu.com questions how can i change or install	0.87	apt-get install hwinfo
u2	ok it won't become an issue on system upgrade	0.85	ok what is that ok just figure it out you just help me out haha
u3	then you probably just need to log out back in to restart indicator messages	**0.27**	**thank you**

TABLE V
AN EXTRACTED EXAMPLE FROM THE TEST SET. OUR SYSTEM FAILED IN RANKING THE GROUND TRUTH RESPONSE *thank you*. NOTE THAT IN THIS EXAMPLE WE REDUCED THE NUMBER OF CANDIDATE RESPONSES TO 4 FOR FORMATTING REASONS.

We estimate that this is mainly due to the generalization capacity of our model in the case of complex contexts or unseen data. Although we think that the candidate responses randomly sampled could be potential responses such as the third response in our example. It is as general as *thank you* and we suppose that the corpus plays an important role in building such response retrieval systems. More details about the corpus are given in the next section.

D. Further analysis

We believe that the Ubuntu Dialogue Corpus contains some bias which makes building utterance ranking systems harder for at least three reasons. First, negative responses are randomly sampled from the hole corpus without any human judgment. Some negative responses could be potential responses for the given context such as *"Thank you", "yes !", "will try it", etc*. Second, the conversations of these corpus were originally between more than two persons and then some heuristics were used in order to reduce the multi-user conversation to a two users. Thus the coherence of the conversation when removing some important information risks to be lost. Third, the nature of these conversations is chat, which is very noisy compared to email, FAQ and forum datasets which contain less abbreviations, typos, etc.

Despite all these disadvantages, we believe that it is important to build dialogue models on this kind of datasets since chat is a large and available source of conversations as well as emails and forums. As an alternative solution, we can recruit labelers to judge the candidate responses as positive or negative and in these case we tolerate the possibility to have multiple good utterances for the same context. In this case, the Recall@k would not be the best metric to evaluate next utterance ranking systems. Precision@1, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) will be more appropriate in the case of presence of multiple good utterances.

E. Parameter tuning

Word embeddings were initialized with Glove [21] pre-trained on Common Crawl Corpus[4] then fine tuned during training[5]. The only preprocessing performed on the dataset is tokenization, lemmatization and stemming available as options when downloading the corpus. The System parameters were updated using Stochastic Gradient Decent (SGD) with Adam algorithm [22]. The model was trained on a single Titan X GPU.

Initial learning rate was set to 0.001 and Adam parameters β_1 et β_2 were set to 0.9 and 0.999 respectively. As regularization strategy we used *early-stopping* and to train the model we used mini batch of size 256. The size of word embeddings and the size of the hidden layer of LSTM and BiLSTM were set to 300. We limited the size of both the context and the response to 160 words. We implemented our system with Keras [23] with Tensorflow[24] in backend. We release the code that reproduces our results on https://github.com/basma-b/dual_encoder_udc. These hyper-parameters were obtained with a grid search on the development set.

V. CONCLUSION AND PERSPECTIVES

We proposed in this work an utterance ranking system based on dual encoder by improving a state of the art response retrieval system. Experimental results show that our approach brings significant improvements compared to the state of the art systems. Our new approach based on semantic and syntactic similarities between the context and response allows to better distinguish between good and bad responses.

As a future work, we plan to re-evaluate the other state of the art approaches [15], [2], [12] on the same test set of the second version of the Ubuntu Dialogue Corpus. Our next goal is to improve the context representation by modeling dialogue utterances separately and then using attention mechanism [25], [26] instead of simply concatenating them. We plan to evaluate the impact of text preprocessing on the performances such as removing stop words and replacing urls, numbers, etc with specific tags. Moreover, an evaluation of our approach on larger datasets and on other corpora of other languages such as *Baidu Tieba* [15] and *Douban* [2] is planned.

[4]http://commoncrawl.org/the-data/

[5]Note that we trained word embeddings on the training set without but no improvement was observed.

REFERENCES

[1] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'15)*, Prague, Czech Republic, September 2015, pp. 285–294. [Online]. Available: http://aclweb.org/anthology/W15-4640

[2] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, Vancouver, Canada, 2017, pp. 496–505. [Online]. Available: http://www.aclweb.org/anthology/P17-1046

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] R. T. Lowe, N. Pow, I. V. Serban, L. Charlin, C.-W. Liu, and J. Pineau, "Training end-to-end dialogue systems with the ubuntu dialogue corpus," *Dialogue & Discourse*, vol. 8, no. 1, pp. 31–65, 2017.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 2014 conference on Advances in Neural Information Processing Systems (NIPS'14)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Montreal, Canada, 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[6] O. Vinyals and Q. Le, "A neural conversational model," in *Workshop on Deep Learning at the 31 st International Conference on Machine Learning (ICML'15)*, Lille, France, 2015.

[7] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, Phoenix, AZ, USA, 2016, pp. 3776–3783. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016387.3016435

[8] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*, Melbourne, Australia, 2015, pp. 553–562.

[9] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, and R. Kurzweil, "Generating high-quality and informative conversation responses with sequence-to-sequence models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, Copenhagen, Denmark, September 2017, pp. 2210–2219. [Online]. Available: https://www.aclweb.org/anthology/D17-1235

[10] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'16)*, San Diego, CA, USA, June 2016, pp. 110–119. [Online]. Available: http://www.aclweb.org/anthology/N16-1014

[11] R. Kadlec, M. Schmid, and J. Kleindienst, "Improved deep learning baselines for ubuntu corpus dialogs," in *Workshop on Machine Learning for Spoken Language Understanding and Interaction at the 29th Annual Conference on Neural Information Processing Systems (NIPS'15)*, Montreal, Canada, 2015.

[12] Z. Xu, B. Liu, B. Wang, C. Sun, and X. Wang, "Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'17)*, Anchorage, AK, USA, May 2017, pp. 3506–3513.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Workshop on Deep Learning and Representation Learning at the 28th Annual conference on Advances in Neural Information Processing Systems (NIPS'14)*, Montreal, Canada, 2014.

[14] H. Wang, Z. Lu, H. Li, and E. Chen, "A dataset for research on short-text conversations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, Seattle, WA, USA, 2013, pp. 935–945. [Online]. Available: http://www.aclweb.org/anthology/D13-1096

[15] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Response selection with topic clues for retrieval-based chatbots," *arXiv preprint arXiv:1605.00090*, 2016.

[16] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 338–349. [Online]. Available: http://dl.acm.org/citation.cfm?id=1996889.1996934

[17] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proceedings of the 26th international conference on Advances in Neural Information Processing Systems (NIPS'13)*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Lake Tahoe, NV, UDA: Curran Associates, Inc., 2013, pp. 926–934.

[18] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. Buenos Aires, Argentina: AAAI Press, 2015, pp. 1305–1311. [Online]. Available: http://dl.acm.org/citation.cfm?id=2832415.2832431

[19] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, 2017, pp. 1116–1126. [Online]. Available: http://www.aclweb.org/anthology/P17-1103

[20] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, Austin, Texas, November 2016, pp. 2122–2132. [Online]. Available: https://aclweb.org/anthology/D16-1230

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, Doha, Qatar, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations (ICLR'15)*, San Diego, CA, USA, 2015.

[23] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[25] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.