Comparing Cross Language Relevance vs Deep Neural Network Approaches to Corpus-based End-to-end Dialogue Systems*

Seyed Hossein Alavi and **Anton Leuski** and **David Traum**
Institute for Creative Technologies
University of Southern California
{seyedhoa@ict.usc.edu and leuski@ict.usc.edu and traum@ict.usc.edu}

Abstract

We compare two models for corpus-based selection of dialogue responses: one based on cross-language relevance and a cross-language LSTM model. Each model is tested on multiple corpora, collected from two different types of dialogue source material. Results show that while the LSTM model performs adequately on a very large corpus (millions of utterances), its performance is dominated by the cross-language relevance model for a more moderate-sized corpus (ten thousands of utterances).

1 Introduction

End-to-end neural network models of conversational dialogue have become increasingly popular for conversational tasks (e.g., (Ritter et al., 2011; Serban et al., 2015; Zhao et al., 2017)). These models eschew traditional modeling approaches that include internal hand-crafted domain models and representations of dialogue context and multimodal input signals, and separate components for understanding natural language (converting to the internal representation language), updating dialogue state, state-based response generation, and natural language generation (e.g., (Traum and Larsson, 2003; Raux et al., 2005; Nasihati Gilani et al., 2018)). Instead, these models learn to respond directly from a corpus, either by generating new responses or selecting a response from the corpus training data, using dual encoding and hidden layers to learn appropriate dialogue continuations. However, there are still a number of questions remaining about how well such models really work for real applications, and how much data is needed to achieve acceptable performance. Other machine learning approaches have been shown to be useful, with much smaller datasets.

In this paper, we compare two different kinds of end-to-end system, a neural network model based on (Lowe et al., 2015) and an older kind of end-to-end dialogue model, based on cross-language retrieval (Leuski et al., 2006), implemented in the publicly available NPCEditor (Leuski and Traum, 2011), and previously used for systems that have been displayed in museums (Traum et al., 2012, 2015). We compare these models on two different datasets: the Ubuntu Corpus (Lowe et al., 2015), and one derived from one of the museum system datasets (Traum et al., 2015).

2 Datasets and models

We utilized a number of datasets in our experiments to compare NPCEditor with a deep neural network model. The *Ubuntu Dialogue corpus* (Lowe et al., 2015) was constructed from Linux support message boards, where people posted problems and solutions. It contains 1 million multi-turn dialogues, with a total of over 7 million utterances and 100 million words. The training set has 50% relevant and 50% irrelevant pairs of $<$ context, response $>$. In the development set, for a given context it has 1 relevant response and 9 distractors (irrelevant responses).

We constructed three other datasets out of the data made available from the system described in (Traum et al., 2015). Pinchas_10 consists of 33350 samples for the training set, 50% of which are negative samples and the rest are positive. In the development and test sets, for each question, there is a relevant response and 9 randomly selected non-relevant responses. (Similar to the dev and test sets in the Ubuntu corpus)

Pinchas_1444 is constructed to investigate how the models would perform on a task inspired by a

*This work was supported by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

real problem (Traum et al., 2015) in which we may have more than one thousand possible responses gathered from interviews. The training set is created similar to Pinchas_10. Nonetheless, for the development and test sets, instead of 10 distractors, we used the whole set of possible responses. Another important difference between Pinchas_1444 and Pinchas_10 is that in this new set there might be more than one relevant response for a given question. Given that very few of the 1444 responses are appropriate for any given question, showing an even number of positive and negative examples might inappropriately prefer recall over precision. In a second version, Pinchas_1444_v2, we increased the negative samples in the training set from 50% to 90%.

The first model we test is NPCEditor (Leuski and Traum, 2011), which was used for the system in (Traum et al., 2015). At the core of NPCEditor is a statistical regression approach based on cross-lingual language model proposed by Lavrenko for cross-lingual information retrieval (Lavrenko, 2004). Leuski and Traum successfully adopted his approach to question answering and applied it in many different applications (Leuski and Traum, 2008, 2011).

From the pool of previous deep neural net models, such as (Hochreiter and Schmidhuber, 1997), (Olabiyi et al., 2018), (Shao et al., 2017), (Zhou et al., 2018), (Zhang et al., 2018), (Devlin et al., 2018), (Mehri and Carenini, 2017), we chose the Dual encoder model first introduced by (Lowe et al., 2015). We trained the model with the same parameters that (Lowe et al., 2015) did.

3 Experiments and Evaluation

We conduct a series of experiments to compare the NPCEditor and the Dual-Encoder model. Following (Lowe et al., 2015), we use R@k as the evaluation metric, which is the percentage of times that the expected response is retrieved in the top-k responses. R@1 is equivalent to accuracy. We first test the Dual-Encoder model on both the Ubuntu corpus (to compare with the model in (Lowe et al., 2015), as a sanity check on the implementation), and on the Pinchas_10 dataset, which has a test-set parallel in structure to Ubuntu. Next we compare the NPCEditor and the Dual-Encoder model on the Pinchas_10 dataset. Then we compare the performance of the NPCEditor and Dual-Encoder model on Pinchas_1444_v1 and Pinchas_1444_v2 datasets.

Dataset	Pinchas_10		Ubuntu
Model	**NPCEditor**	**DE**	**DE**
1 in 10 R@1	**0.78**	0.64	0.60
1 in 10 R@2	**0.84**	0.83	0.74
1 in 10 R@5	0.92	**0.97**	0.92

Table 1: Results from the experiment 1 and 2 using various R@k measures.

Pinchas_1444	v2		v1
Model	**NPCEditor**	**DE**	**DE**
1 in 1444 R@1	**0.7663**	0.1238	0.0625
1 in 1444 R@2	**0.8175**	0.1939	0.1305
1 in 1444 R@5	**0.8758**	0.3089	0.2392
1 in 1444 R@10	**0.9106**	0.4217	0.3441

Table 2: Results from experiment 3 and 4.

4 Results

Experiment 1 showed that the Pinchas data appears easier than the Ubuntu data - with a much smaller training set size, the Dual-Encoder model was able to improve on R@k in the Pinchas_10 dataset compared to the Ubuntu dataset. Experiment 2 showed that given the amount of available training data (10s of thousands of examples), the NPCEditor significantly out-performs the Dual-Encoder model in R@1 on this data set. Experiment 3 showed that the results are even more striking for a more real-world example, where the system's task is to pick the best response out of a set of over 1000 available. Here, the Dual-Encoder model does not perform well enough to engage in a meaningful dialogue, while the NPCEditor performs similarly to results reported in (Traum et al., 2015), which led to much-reported user engagement. The improved performance of the Pinchas_1444_v2 training set, with a much higher proportion of negative examples, does perhaps point to a direction for improvement. Future work should perhaps look at the even higher distribution of negative to positive examples.

These results do show that despite the recent popularity of deep learning models, there is still a place for more traditional machine learning algorithms, that can operate well on more moderate-sized data sets for problems of interest. It may also be the case that different types of dialogue have different optimal models. For example, (Gandhe and Traum, 2010) show very different upper bounds for retrieval approaches to dialogue in different domains/datasets.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sudeep Gandhe and David Traum. 2010. I've said it before, and i'll say it again: an empirical investigation of the upper bound of the selection approach to dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 245–248. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Victor Lavrenko. 2004. *A Generative Theory of Relevance*. Ph.D. thesis, University of Massachusetts at Amherst.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *Proceedings of the 26th Army Science Conference*, Orlando, Florida, USA.

Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *IJCNLP*.

Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal dialogue management for multiparty interaction with infants. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 5–13. ACM.

Oluwatobi Olabiyi, Alan Salimov, Anish Khazane, and Erik T. Mueller. 2018. Multi-turn dialogue response generation in an adversarial learning framework. *CoRR*, abs/1805.11752.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. *Proceeding of the International Speech Communication Association*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating long and diverse responses with neural conversation models. *CoRR*, abs/1701.03185.

David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and grace: Direct interaction with museum visitors. In *International conference on intelligent virtual agents*, pages 245–251. Springer.

David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015. Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 199–208, Prague, Czech Republic. Association for Computational Linguistics.

David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *CoRR*, abs/1806.09102.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.