Common Sense Knowledge in Large Scale Neural Conversational Models

D.S. Tarasov(✉) and E.D. Izotova

Meanotek AI Research, Sibirsky Trakt 34, Kazan, Russian Federation
dtarasov@meanotek.io

Abstract. It was recently shown, that neural language models, trained on large scale conversational corpus such as OpenSubtitles have recently demonstrated ability to simulate conversation and answer questions, that require common-sense knowledge, suggesting the possibility that such networks actually learn a way to represent and use common-sense knowledge, extracted from dialog corpus. If this is really true, the possibility exists of using large scale conversational models for use in information retrieval (IR) tasks, including question answering, document retrieval and other problems that require measuring of semantic similarity. In this work we analyze behavior of a number of neural network architectures, trained on Russian conversations corpus, containing 20 million dialog turns. We found that small to medium neural networks do not really learn any noticeable common-sense knowledge, operating pure on the level of syntactic features, while large very deep networks shows do posses some common-sense knowledge.

Keywords: Deep highway networks · Conversational modeling · Information retrieval

1 Introduction

It was recently shown [1] that large-scale heterogenous dialog corpus can be used to train neural conversational model, that exhibits many interesting features, including capabilities to answer common-sense questions. For example, neural network model can tell that dog have four legs, and usual color of grass is green, even though these question/answer pairs do not explicitly exists in the dataset. This raises a question if such model can learn implicit ontology from conversations. If true, such models can be applied to the tasks outside of dialog modeling domain, such as information retrieval and question answering.

Unfortunately, this property has not received yet sufficient attention. Recent research on neural conversational models have been focused on incorporating longer context [2, 3], dealing with generic reply problem [4], incorporating attention and copying mechanism [5]. Attempts to connect neural conversational models to external knowledge bases were also made [6], however, we are not aware of any papers that investigated nature of knowledge that can be stored in neural network synaptic weights.

© Springer International Publishing AG 2018
B. Kryzhanovsky et al. (eds.), *Advances in Neural Computation, Machine Learning, and Cognitive Research*, Studies in Computational Intelligence 736,
DOI 10.1007/978-3-319-66604-4_6

In this work, we investigate the possibility of using large dialog corpus to train semantic similarity function. We train a number of neural network architectures, including recently proposed deep highway neural network model [7] on large number of dialog turns, extracted from both Russian part of OpenSubtitles database [8] and data collected from publicity available books in Russian, totaling 20 millions dialog turns. The training goal was to classify if sentence represent a valid response to previous utterance or not.

We found that smaller neural network models can learn general similarity function in sentence-space. This function performance is superior to simple neural bag of words models in selecting proper dialog responses and finding sentences, relevant to the query. However, these networks don't incorporate any meaningful knowledge about the world.

Large neural networks seem to incorporate some common-sense knowledge to semantic similarity function, as demonstrated by reranking possible answers to various common-sense and factoid questions.

2 Methods and Algorithms

2.1 Datasets

Russian part of OpenSubtitles database was downloaded from http://opus.lingfil.uu.se/. OpenSubtitles [8] is a large corpus of dialogs, consisting of movie subtitles. However, the data in this corpus is much smaller (about 10 M dialog turns after deduplication) then its English counterpart. OpenSubtitles is also very noise dataset, because it contains monologues, spoken by the same character, that are impossible to separate from dialogues, and also dialog boundaries are unclear.

To extend the available data for this work, we used Russian web-site lib.ru and mined publicly available fiction books for conversations of book characters. A heuristic parser was written to extract dialog turns from book texts. 10 M dialog turns was mined by this approach, resulting in total corpus size of 20 M dialog turns.

2.2 Neural Network Architectures

The structure of models, used for this work is shown on the Fig. 1. A number of specialized architectures were proposed for sentence matching task [9], including convolutional and LSTM models.

Overall, our model consists of two encoder layers that compute representations of source sentences, one or more processing layers stacked on top of each other and output layer, consisting of a single unit that outputs the probability of response being appropriate to context. In this work we tested two types of encoders LSTM-based encoder along with simpler fully connected encoder.

Neural bag of words (NboW) model is a fixed length representation xf obtained by summing up word vectors in the text and normalizing result (by multiplying by 1/lxfl). This model was used as a baseline in [9].

2.3 Word Vectors

Real-valued embedding vectors for words were obtained by unsupervised training of Recurrent Neural Network Language Model (RNNLM) [10] over entire Russian Wikipedia. Text was preprocessed by replacing all numbers with #number token and all occurrences of rare words were replaced by corresponding word shapes.

3 Results and Discussion

3.1 Reply Selection Accuracies

Table 1 reports response selection accuracies for three different models on the test set, consisting of 10,000 contexts. For each context 4 random responses were given to classifier to rank along with "correct" (actual response from dataset).

Table 1. Model accuracies in selecting right context/response pairs

Model	Accuracy
Random baseline	19.7
Neural bag of word encoder with 1 fully connected processing layer	21.2
Fully connected encoder with 1 fully connected processing layer	37.8
Fully connected encoder with 4 highway processing layers	41.1
LSTM encoder with 4 highway processing layers	39.3

Two findings are particularly surprising. First, NboW model did not achieved any significant improvements over random baseline, in contrast with results reported in [9] for matching English twitter responses. This result might be due to the fact that our corpus is much larger (about 10 times) and much more noisy. Second, LSTM encoder actually performs worse than simple fully connected encoder, and it is also much slower. This is interesting, because fully-connected encoders with zero-padded sentences are not commonly evaluated for such tasks, because they are assumed to be bad models, because of their potential to overfit the data. However, with a special case of conversation, where most responses are small in size, and given a lot of data, apparently fully-connected encoders could be usable option.

Another interesting point here is that we observed that small model with 1 processing layer also scored 29.8 on the task of matching English sentences using pre-trained word vectors for English language, without training the network itself on English data. This result indicates that small models actually learn some language-independent generic similarity function that operate on word vectors and not involve deeper understanding of the content.

3.2 Factoid Answer Selection from Alternatives

To evaluate model capability for question answering, we designed a test set of 300 question-answers pairs, using search engine snippets as candidate answers. The task was

to select snippet, containing the correct answer (all snippets were first evaluated by human, to asses if they contain necessary answers). Top 10 snippets were selected for evaluation for each question. Table 2 summarizes results of all models.

Table 2. Accuracies on factoid question answering

Model	Percent of correct answers
First snippet baseline	30.3%
Fully connected encoder with 1 fully connected processing layer	36.2%
Fully connected encoder with 4 highway processing layers	34.7%
LSTM encoder with 4 highway processing layers	32.0%

For this task, a model with one processing layer demonstrated best results. Overall, improvements over were small, probably because search engine snippets already represent strong baseline. Manual inspection of ranking results revealed, that improvements were due to models capacity to distinguish between snippets that contained answers and snippets that were just copies of the questions (see Table 3).

Table 3. Example ranking of candidate snippets for the question "сколько звезд на небе" (How many stars are there in the sky?)

Answer text	Answer ranking
В ясную погоду на небе видно около 3000 звезд ("*on clear whether, about 3000 stars can be seen at the sky*")	0.76
Сколько же звезд на небе? ("*How many stars are in the sky*?")	0.68
На этой странице вы узнаете, сколько звезд на самом деле видно на небе ("*On this page you will learn how many stars can be seen at the sky*")	0.55

We therefore conclude, that model can use sentence structure to decide if it can be viewed as appropriate answer or not.

3.3 Common Sense Questions

Finally, to test models capacity to understand the world, we prepared a set of 100 common-sense questions, like "what is the color of the sky?", "what pizza is?". Like in previous setup, we evaluate model capability to choose correct answer out of 5 options. Results are summarized in Table 4.

Table 4. Accuracies on multiple-choice common-sense questions

Model	Result
Random baseline	19.5%
Fully connected encoder with 1 fully connected processing layer	20.3%
Fully connected encoder with 4 highway processing layers	26.5%
LSTM encoder with 4 highway processing layers	19.8%

Only deep model with fully-connected encoder demonstrated some understanding of common sense questions above random baseline and even here results are generally poor. Table 5 shows example rankings of answers to a typical question by best model.

Table 5. Example ranking of candidate answers for common sense questions

Что такое собака? *What dog is?*	Что такое автомобиль? *What automobile is?*	где живет человек? *Where does human live?*	Какого цвета земля? *What is the color of the ground?*
0.71 животное *(animal)*	0.84 мотор *(motor)*	0.844 нора (burrow)	0.87 зеленая (green)
0.49 растение *(plant)*	0.70 механизм *(mechanism)*	0.841 дом (house)	0.86 желтая (yellow)
0.48 концепция *(concept)*	0.67 животное *(animal)*	0.52 лес (forest)	0.83 бурая (brown)
0.45 планета *(planet)*	0.65 фонарь *(lamp)*	0.48 лужа (water pool)	0.82 черная (black)
0.43 механизм *(mechanism)*	0.59 квадрат *(square)*	0.42 джон (John)	0.79 белая (white)
0.38 фонарь *(lamp)*	0.54 дом *(house)*	0.41 мотор (motor)	0.70 дом (house)
0.36 дом *(house)*	0.46 планета *(planet)*	0.40 фонарь (lamp)	0.62 красная (red)
0.32 квадрат *(square)*	0.46 концепция *(concept)*	0.35 квадрат (square)	0.46 фонарь (lamp)
0.30 африка *(Africa)*	0.45 африка *(Africa)*	0.17 море (sea)	0.45 синяя (blue)
0.19 джон *(John)*	0.39 растение *(plant)*		0.16 джон (John)

Manual examination rankings revealed, that questions that concern relationships of two and more entities are more difficult to answer, compared to the questions related to the single entity (Table 5)

4 Conclusions

We found that large neural dialog models can learn some common-sense knowledge, although to the limited extent. There is, however, a room for improvement, because we found that even our large model did not significantly overfit the training set, and there is also a possibility for collecting more training data.

Another interesting finding is that our models learned to understand sentence structure of question/answer pairs and can select answers those structure is more likely to contain answers to the question.

Finally, we observed that simple encoder, based on fully-connected layer with padded input outperforms LSTM-based encoders both in computing speed and response

selection accuracy. Further analysis is needed to understand the significance of this finding.

Subsequent work should probably include analysis of even larger models, and detailed analysis of what happens in encoding layers, to better understand how these models really operate and what they can do. Also, testing sets need to be expanded in both size and extend of coverage of various common-sense topics.

References

1. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint, arXiv:1506.05869 (2015)
2. Yao, K., Zweig, G., Peng, B.: Attention with intention for a neural network conversation model. arXiv preprint, arXiv:1510.08565 (2015)
3. Chen, X., et al.: Topic aware neural response generation. arXiv preprint, arXiv:1606.08340 (2016)
4. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint, arXiv:1510.03055 (2015)
5. Mihail, E., Manning, D.: A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. arXiv preprint, arXiv:1701.04024 (2017)
6. Ahn, S., et al.: A neural knowledge language model. arXiv preprint, arXiv:1608.00318 (2016)
7. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint, arXiv:1505.00387 (2015)
8. Tiedemann, J.: News from OPUS—a collection of multi-lingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, pp. 237–248. John Benjamins Publishing Company, Amsterdam (2009)
9. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in Neural Information Processing Systems, pp. 2042–2050 (2014)
10. Mikolov T., Karafiat M., Burget L., Cernocky J., Khudanpur S.: Recurrent neural network based language model. In: INTERSPEECH, pp. 1045–1048 (2010)