Neural Network-Based Models with Commonsense Knowledge for Machine Reading Comprehension

Denis Smirnov
National Research University Higher School of Economics
Moscow, Russia
dmsmirnov@hse.ru

Abstract

State-of-the-art machine reading comprehension models are capable of producing answers for factual questions about a given piece of text. However, some type of questions requires commonsense knowledge which cannot be inferred from the given text passage. Thus, external semantic information could enhance the performance of these models. This PhD research proposal provides a brief overview of some existing machine reading comprehension datasets and models and outlines possible ways of their improvement.

1 Introduction

Machine reading comprehension (MRC) is one of the well-studied problems in artificial intelligence. This problem can be defined as a problem of creating an algorithm, which can understand the content of a given text in natural language, which is used by humans to communicate with each other. There is no formal way to define the quality of understanding. One of the most popular approaches to measure the understanding is the assess the ability to answer the questions about the given text, hence the problem of machine reading comprehension is closely related to the question answering problem and these concepts are often used as synonyms.

Question answering is a vital component of many real-world systems. More accurate answers to questions on the text, will improve the performance of intelligent assistants and search engines on the Internet or corporate knowledge bases.

There exist many datasets used to assess question answering models which contain texts and questions about its contents. This could be either multiple choice questions (Richardson, 2013), cloze-style questions, which require filling in the gap in the question definition (Hermann et al., 2015) or open questions, where the answer is a named entity from the context (Rajpurkar et al., 2018).

State-of-the-art question answering models perform fairly well for factual questions when the answer is clearly stated in the text but they fail to achieve comparable performance on questions which require common sense inference. This type of questions is often simple for humans but can be challenging for an algorithm because an answer cannot be derived without external knowledge about semantic relationships of entities described in the given text. Examples of such questions are demonstrated in the next section.

2 QA Datasets that Require Commonsense Knowledge

A well-known problem for commonsense evaluation is the Winograd Schema Challenge (WSC) (Levesque, 2011). The schema of the text passages and questions is based on co-reference resolution. The first part of the text mentions two entities, while the second part contains a pronoun or a possessive adjective which refer to any of the introduced entities. To illustrate the problem, consider the following question from the WSC dataset:

> Sam pulled up a chair to the piano, but it was broken, so he had to stand instead.
> What was broken?
> • The chair *(correct answer)*
> • The piano

The original dataset for the problem is very small, it contains only 150 schemas, as the new samples must be thoroughly handcrafted by humans. The most recent version consists of 285

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 90–94,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_014

schemas. This task remains difficult for models, state-of-the-art approach reports only 71.06% success rate for the problem (Prakash et al., 2019).

A more recent dataset for machine reading comprehension with commonsense knowledge is MCScript (Ostermann et al., 2018). It contains around 2100 scripts (narrative texts describing everyday activities) and approx. 14000 questions, written by crowdsourced workers Authors estimate that commonsense reasoning is required to answer 27.4% of questions.

More large-scale dataset was introduced by Zhang et al. (2018). Authors designed a multistage procedure to generate passage-question-answer triplets from CNN/Daily Mail dataset and Internet Archive which included performed automatic filtering of the triplets, leaving only those, which were unanswerable by the competitive MRC model, and further manual human filtering resulting in 120000 cloze-form questions. Authors analyzed a sample of resulting passages and questions and concluded, that automatic filtering allowed to exclude most of the questions, which could be answered with paraphrasing, while human filtering excluded ambiguous questions. 75% of sampled questions required commonsense reasoning or multisentence inference to obtain an answer.

A scalable approach for commonsense question generation and a new dataset, which consists from more than 12000 multiple-choice questions, was recently introduced by Talmor et al. (2019). In this dataset, questions are based on extracted subgraphs from ConceptNet. Given an extracted source concept and three target concepts, connected with the source by the same relation, crowdsourcers were asked to write three questions, that contain source concept and have only one of the target concepts as an answer. At the next stage, two more answer choices are added, to make the problem more challenging.

Below is an example question from CommonsenseQA:

Where would I not want a fox?
- hen house *(correct answer)*
- england
- mountains
- english hunt
- california

Authors also performed multiple experimental evaluations and showed, that current state-of-the-art models are far away from human performance on this dataset.

3 Existing Approaches

Modern approaches to question answering problem mostly rely on deep neural networks. More specifically, they often use recurrent neural networks (RNNs), a special type of networks, which process input sequentially. In such networks, the result of the processing of previous input affects the consecutive outputs. One limitation of this architecture is that the state is updated on each timestep, so it is hard to keep track of long-range dependencies. This happens because of the vanishing gradient problem. To address this issue a modification of recurrent layer called Long Short-Term Memory (LSTM) was introduced (Hochreiter and Schmidhuber, 1997). In this type of layer, there is an additional path to carry data flow (carry flow) through time steps, which is capable of capturing long-range dependencies. Another possible improvement of RNN architecture, frequently used in NLP models, is the simultaneous processing of input sequence in s forward and backward direction, which is done in bi-directional RNNs (Schuster and Paliwal, 1997). The same trick can be applied to LSTM (BiLSTM) (Graves and Schmidhuber, 2005).

A standard component of deep neural networks in the whole natural language processing domain are embeddings. They are used to transform words into low-dimensional dense real-valued vector representation which can be easily used as an input for any type of neural network. There are several standard embeddings pre-trained on large text corpora, like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or FastText (Bojanowski et al., 2017), and most of the question answering models use one of the available implementations.

The rest of this section describes several architectures of the most important approaches for question answering problem. The architectures under consideration both use and do not use the external semantic information.

3.1 RNN-Based Models

BiDAF (Bi-Directional Attention Flow) was proposed by Seo et al. (2017). In this architecture, embeddings are calculated using the input data (pair context and question) both at the word level

and at the character level (using the convolutional network char-CNN). Embeddings vectors are fed as an input to BiLSTM, the context and the question use layers that are not interconnected. Then, the attention mechanism is applied to the activations of these layers (it is proposed to use it in two directions from the context to the question and vice versa) and the result passes through one common two-layer BiLSTM and the final layer forms the answer.

One notable model for machine reading comprehension is DAANet (Xiao et al., 2018). This architecture does not use any external semantic information. However, the dual learning objective of this model deserves attention. Instead of training the model with a single task to answer the question, the authors proposed a way to simultaneously train the network to generate a question using the answer and generate the answer using the question.

3.2 Pre-Trained Language Models

Nowadays, the best results in many datasets are achieved by universal deep pre-trained language models that are fine-tuned for a specific task.

The common limitation of the models, which use Word2Vec, GloVe, FasText or similar embeddings is the static nature of word vectors obtained by such embeddings. The word vector values are the same, regardless of the context, and thus, these embeddings are not capable of capturing polysemy. At the same time, embeddings, obtained with language models, overcome this issue, they produce different vectors for words in different contexts.

One of the best models in this category is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The main innovation of this model is the training method, in which, unlike the usual approach to learning language models (when the objective is to predict the next word), the network learned to predict a randomly chosen masked word in a phrase and thus learned the representation of the surrounding context of the word.

Another deep language model also based on transformer architecture (Vaswani et al., 2017) is GPT-2 (Radford et al., 2019). While BERT was originally trained on BooksCorpus and English Wikipedia, GPT-2 was trained on a more diverse set of Internet texts. It also has much more trainable parameters (1.5B in the largest unreleased version vs. 340M in the BERT-large). Authors claim that GPT-2 achieves state-of-the-art results on several NLP tasks, including the Winograd Schema Challenge.

3.3 Adding Commonsense to Models

When it comes to enrichment of models with commonsense knowledge, semantic networks are the number one choice. One of the largest semantic networks is ConceptNet (Speer et al., 2017), where one can find a vocabularies of concepts in multiple languages, which are interconnected by 34 relations, forming a graph with 34 million edges.

A few attempts to add external semantic information are found in the works of Wang et al. (2018) and González et al. (2018). which use information from ConceptNet. Wang et al. (2018) propose the model, which is similar to BiDAF with the only exception that the embeddings of words and features from context, question and answer are also considered, then they pass through separate BiLSTM layers and their activations are aggregated with attention. In the features of the model, which are counted for context, there is a place for a vector of 10 values, encoding the relation of a word and any of the words in a question or answer (the fact that there is an edge in ConceptNet). If there are several such relations, one is chosen randomly.

An approach of González et al. (2018) is the replacement of standard embeddings with NumberBatch semantic vectors, trained using connections from ConceptNet. As far as embeddings are a standard component of deep neural networks in natural language processing domain, this approach of embeddings replacement allows enriching a great variety of neural network-based models with external semantic information, contained in word representation.

The missing knowledge can be extracted not only from knowledge graphs, but also from the text repositories. One possible technique is the knowledge hunting. In case of co-reference resolution problem like in WSC, knowledge hunting would consist in finding the similar piece of text, which does not have ambiguity in referencing mentioned entities. This method have been successfully applied to WSC by Prakash et al. (2019). Authors combined a two-stage knowledge hunting

procedure with the outputs of a neural language model using a probabilistic soft logic, and it currently achieves state-of-the art results in this challenge.

4 Discussion

We have seen so far, that even enormous pre-trained deep language models cannot pick up the ability to reason about the text, even when trained on large and diverse corpora, so the improvement of methods of extraction and representation of commonsense knowledge in neural networks is one of the directions of my PhD research.

The idea from DAANet (Xiao et al., 2018) could be used in a modified version of such model, which can simultaneously produce a query to a semantic network (i.e question about the text) and an answer to a given target question. This approach could be a possible solution to the problem of defining relevant external information for question answering algorithm.

The improvement of extraction and representation of commonsense knowledge is only one aspect of the work. So far we have explored the datasets and models for English language, which has enormous amount of labeled and unlabeled resources. Other languages have much less available resources for training. Another difficulty arises, when the same models are being applied to more agglutinative languages, which require morphological disambiguation, like Russian. The adaptation of existing models to Russian is another direction of my work.

Deep pre-trained language models can be trained in multilingual setting, and, for example, BERT nominally supports Russian. However, the monolingual model outperforms multilingual version (Devlin et al., 2019). It has been shown, that multilingual model can be a good initialization for finetuning of monolingual version of the model (Kuratov and Arkhipov, 2019). Exploration of the possibilities and limitations of transfer learning between languages for question answering.

Finally, the lack of resources for evaluation of models for Russian language, encourages me to collect my own dataset for machine reading comprehension.

Acknowledgments

I would like to express my appreciation to Dr. Dmitry Ilvovsky, my research supervisor, for his valuable and constructive suggestions during the planning and development of this work.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL* 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota.

José-Ángel González, Lluís-F. Hurtado, Encarna Segarra, and Ferran Pla. 2018. ELiRF-UPV at SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*. volume 4, pages 2047–2052 vol. 4.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 1693–1701.

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

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR* abs/1905.07213.

Hector J. Levesque. 2011. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar.

Ashok Prakash, Arpit Sharma, Arindam Mitra, and Chitta Baral. 2019. Combining knowledge hunting and neural language models to solve the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners .

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia.

Matthew Richardson. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Emprical Methods in Natural Language Processing (EMNLP 2013)*.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*. pages 4444–4451.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 5998–6008.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018. Yuanfudao at SemEval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana.

Hang Xiao, Feng Wang, Jianfeng Yan, and Jingyao Zheng. 2018. Dual ask-answer network for machine reading comprehension. *CoRR* abs/1809.01997.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *CoRR* abs/1810.12885.