24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Review of the Recent Techniques for Learning Commonsense Knowledge applied to the Winograd Schema Challenge

Aneta Koleva[1]

Abstract. The Winograd Schema Challenge (WSC) was proposed by Levesque et al. in 2011 as a new test in Artificial Intelligence (AI) and possibly as an alternative to the Turing test. WSC is a complex coreference resolution task which requires applying knowledge on commonsense reasoning. It is an easy task for humans but it still remains an unsolved challenge for computers. There are two categories of proposed approaches for tackling the WSC. The first encompasses techniques based on formalized Knowledge Representation and Reasoning (KRR), while the second entails Machine Learning (ML) approaches. In this paper we provide a review of the state-of-the-art approaches proposed from both categories and we outline their strengths and weaknesses. Additionally, we discuss a recent work which combines techniques from both categories as it seems to be a promising and innovative approach.

1 Introduction

A Winograd Schema (WS) [1] consists of three main parts. The first part represents a pair of twin sentences, where each sentence contains: (i) two noun phrases of the same semantic class and gender, (ii) one ambiguous pronoun that could refer to either of the antecedent noun phrases, and (iii) a special word such that when changed, the resolution of the pronoun is changed. The second part is a question which contains the special word, asking about the referent of the ambiguous pronoun. The third part contains the two possible answers corresponding to the noun phrases in the sentence. The original WSC dataset consists of 150 such problems[2] [3]. A typical example of a WS is the following one:

S: The trophy does not fit into the brown suitcase because it is too [small/large].
Q: What is too [small/large]?
A: The suitcase/the trophy

Here, the special word is one of the adjectives [*small/large*]. The ambiguous pronoun is *it* and the two antecedents are *trophy* and *suitcase*. In order to identify the correct referent of *it*, one needs to use commonsense knowledge about the size of objects. Having the special word in the sentences prevents any solver to rely on sentence structure and word order for finding the answer to the question. Consequently, applying similar methods as for solving the coreference resolution problem does not give better results then a randomly chosen answer. For a computer to be able to answer correctly a WSC problem, commonsense knowledge needs to be added. Therefore, the first challenge imposed by the WSC is how to obtain knowledge for commonsense reasoning. The second is how to formalize this knowledge such that all important information are preserved. Lastly, how to reason on top of formalized knowledge is the third main challenge when trying to solve the WSC.

Regarding the limitations of the WSC, the first and probably the most important limitation is the number of available sentences. Because the process of creating new WS is difficult, the number of available sentences is very small. For this reason, an additional dataset, called Pronoun Disambiguation Problems (PDPs), which contains examples found in literature, biographies and news or have been manually constructed by humans is also considered during evaluation of methods. A PDP may consist of more than one sentence and it can also have more than two possible answers, but unlike the WSs, these problems do not have twin sentences. Since the structure of the PDPs is not always consistent, resolving these problems is more difficult than resolving WSs. Another limitation is that no training data is provided for any of the datasets which makes it hard to approach the task as a machine learning problem.

The rest of the paper is structured as follows. In Sections 2 and 3 we describe the state-of-the-art approaches from the KRR and ML categories respectively. A combined approach is presented in Section 4. In the last section we conclude the paper.

2 Knowledge Representation and Reasoning

At the core of the KRR approaches lies the idea of formalizing background knowledge and formulating rules which can be used for resolving the missing pronoun. A very recent work by Sharma [5] achieves promising results among the KRR approaches. This work focuses on representing the WSC problem and the needed commonsense knowledge as semantic graphs. Then, it finds the correct answer by applying a method for graph matching, based on subgraph isomorphism. To this end, [5] defines a graph based representation for the WS sentence and for the needed knowledge. The definition for the commonsense knowledge assumes only the causal form. Additionally, [5] provides a reasoning algorithm based on ASP. The reported results from the conducted experiments are remarkably high. For the first experiment, they provided hand written graph representation of 240 WSC problems together with the corresponding background knowledge for each of the problems and their algorithm answered all 240 problems correctly. That is 84.2% from the entire

[1] TU Dresden, Germany, email: aneta.koleva@tu-dresden.de
[2] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html
[3] The number of WSC problems varies depending on the format in which they are available. In the .xml format there are 285 WSs, whereas in .html format there are 150 WSs.

WSC dataset. For the second and third experiments, they used the K-Parser[4] for the graph representation of the problems and the background knowledge. In the third experiment, they relied on automatic extraction for the background knowledge and it retrieved relevant knowledge for 120 problems. All of the 120 problems (42.1% of the WSC dataset) were answered correctly by the algorithm. The high number of correctly answered problems in the experiments and the possibility to retrieve an explanation for an answer is what makes this approach interesting. Furthermore, this method could set the path for future work in the direction of using graph representation of the WSC problems and graph matching for finding the correct answer. However, the method suffers from a couple of pitfalls at the moment . The first one is the limited definition of commonsense knowledge. Another observation is that the manual encoding of graphs for each WS is a very tedious and subjective process, which could lead to a biased representation of exactly what is missing for answering a problem.

3 Machine Learning

The proposals in this category rely on machine learning and deep learning techniques. They exploit the vast amount of available online data for learning commonsense knowledge. The model described by Raffel et. al [3], called T5 (Text-To-Text Transfer Transformer), is currently the best performing solution for the WSC as well as for other NLP problems. T5 is implemented as an encoder-decoder Transformer with self-attention and two layer stacks, one for the encoder and one for the decoder. It is called text-to-text because both the input and the predicted output for all tasks are in textual format. As a first step, the model is pre-trained in an unsupervised setting on a large corpora ($\sim$750GB) of cleaned text which has been scraped from Common Crawl[5]. The idea behind this is that the model learns general knowledge before fine-tuning it for a specific task. The objective of the unsupervised training is to first randomly sample and mask 15% of the tokens in the input text. Each of the masked tokens is replaced by a unique sentinel token. After that, the model is trained to predict and output the masked tokens. In the next step, the pre-trained model is trained in a supervised manner on a specific task. For solving the WSC, the input would be a WS sentence with a highlighted ambiguous pronoun and the predicted output should be the correct noun phrase. The input in the training phase is a WS, the ambiguous pronoun form the WS, the correct possible answer and a label *True*. Because this eliminates half of the WSC dataset, an additional dataset, which was provided by [4] of around 1000 WSs, was also used. The model was then tested on the PDP dataset. Although the smallest implementation of T5 already beat the previous best model, it was the largest implementation of T5 with 11 billion parameters which is the new state-of-the-art by achieving 93,8% accuracy. While these are very good results for the WSC, compared to the previous work, this method relies on learning with no inferential reasoning involved. Thanks to available computational power and large corpora of cleaned text, this method achieves very good results. Nevertheless, the training process of the model makes this method very expensive.

4 A Combined Approach

Prakash et al. [2], to the best of our knowledge, are the first ones to combine methods from KRR with methods from ML such as language models. They propose a framework in which four different steps are followed. In the first step a knowledge hunting module sends queries, which contain the verb phrases from the WS, to a search engine in order to extract text similar to the text in the WS. The second step defines an alignment function between the WS and the retrieved text snippets. The goal of this step is to find a mapping by matching the possible answers and the verb from the WS, to the entities from the retrieved text. In the third step, a pre-trained language model assigns probabilities to the sentences with substituted pronouns. These probabilities are then used in the last step as truth values for each of the possible answers. Finally, the results from the alignment and the probabilities from the language models are combined in a Probabilistic Soft Logic (PSL) framework. PSL is based on first order logic, but the predicates in PSL are values in the interval $[0, 1]$. Prakash et al. [2] formalize a weighted rule in PSL which assigns a truth value to each of the possible answers, based on the output from the alignment and the probabilities predicted by the language models. The answer with the higher truth value is considered to be the correct one for the WS. For the evaluation, the authors used pre-trained language models from previous approaches and report an achieved accuracy of 71.06%. Moreover, for all the language models the addition of knowledge from the knowledge hunting module improved their accuracy. This is an interesting approach because it paves the path for exploring other possible combinations of the results from the powerful language models with additional knowledge and reasoning procedures.

5 Conclusion

In conclusion, it appears that the task of learning commonsense knowledge, so far, cannot be solved by applying techniques from KRR or from ML alone. Despite the vast amounts of available data and computational power, it is still difficult to close the gap between human (100%) and computer (93.8%) performance. Research into how to learn commonsense knowledge by employing techniques from both categories could lead to better results.

ACKNOWLEDGEMENTS

This work is funded by Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR 248 (see `https://perspicuous-computing.science`).

REFERENCES

[1] Hector J. Levesque, Ernest Davis, and Leora Morgenstern, 'The winograd schema challenge', in *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*, (2012).

[2] Ashok Prakash, Arpit Sharma, Arindam Mitra, and Chitta Baral, 'Combining knowledge hunting and neural language models to solve the winograd schema challenge', in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6110–6119, (2019).

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, 'Exploring the limits of transfer learning with a unified text-to-text transformer', *arXiv e-prints*, (2019).

[4] Altaf Rahman and Vincent Ng, 'Resolving complex cases of definite pronouns: The winograd schema challenge', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pp. 777–789, (2012).

[5] Arpit Sharma, 'Using answer set programming for commonsense reasoning in the winograd schema challenge', *TPLP*, **19**(5-6), 1021–1037, (2019).

[4] kparser.org
[5] commoncrawl.org