REAL OR NOT?

Apachiței Maria Luisa, Munteanu Rareș Costin, Ursu Vlad, Vîrvarei Alexandru
under the guidance of Cușmuliuc Ciprian, PhD. Prof. Adrian Iftene

Computational Optimization
first year

1. Problem presentation

In recent years an increased research attention has been focused on evaluating the common sense of natural language statements. This push has been, among other things, driven by a desire of creating an autonomous intelligent machine capable of simulating human intelligence. Compared to humans, however, most machines have a weak common sense, being unable to recognize even trivial logical errors. One limitation of existing commonsense reasoning systems is that when verifying the soundness of a statement the returned output is usually binary, true or false, which means that benchmarking the system is difficult in the absence of a direct metric to quantitatively measure.

The goal of our research is to improve upon earlier prototypes by training a natural language neural network to generate a natural language reasoning explaining the decision. This decision record can be later used to analyze the correctness and accuracy of the model by evaluating the soundness of the returned reasoning and comparing it with a human generated analysis.

2. State-of-the-art

Researching related works, we have uncovered a few papers that were of particular importance. We have started by understanding the findings in the GPT-2 paper *Language Models are Unsupervised Multitask Learners* and *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* which allowed us to better understand the current capabilities of and limitations of state-of-the-art natural language neural networks in processing and understanding information encoded in natural language. From the uncovered strategies we have focused on fine-tuning all pretrained parameters of the down-stream tasks which allowed us to work with huge pre-trained models, saving us a lot of time and facilitating faster experiments and prototypes.

Furthermore, our approach was heavily influenced by the Winograd Schema Challenge, a dataset for common sense reasoning which employs Winograd Schema questions that require the

resolution of anaphora: the system must identify the antecedent of an ambiguous pronoun is a statement. A method of tackling the problems posed by the dataset is described in the A *Simple Method for Commonsense Reasoning* which studies the performance of unsupervised learning methods compared to deep neural networks and highlights some of the shortcomings of neural networks based on their dependence on labeled data in an area where such data is scarce. The improvements suggested by the paper allow a system to outperform previous state-of-the-art methods by a significant margin without using expensive annotated knowledge bases or hand-engineered features.

Another noteworthy work that helped us better understand the dynamic of the problem was *Event2Mind: Commonsense Inference on Events, Intents, and Reactions* which investigates commonsense inference. The paper studies a system where given an event described in a short free-form text a reason is generated about the likely intents and reactions of the event's participants. In their work the authors establish a baseline performance on the task, demonstrating that given the phrase-level inference dataset, neural encoderdecoder models can successfully compose phrasal embeddings for previously unseen events and reason about the mental states of their participants.

3. Solution

We propose a machine learning approach to the problem described above, combining state of the art generative model with data preprocessing methods and a way of modelling the problem so that we can fully leverage the power of the AI model.

3.1 Dataset

The starting point is the SemEval2020 NLP competition, particularly task4, subtask C, that states "Generate the reason why this statement is against common sense and we will use BELU to evaluate it.". The dataset for the task is publicly available on their GitHub and it structured as follows: One semantically wrong proposition followed by 3 explanations why it is wrong. The training set is formed of 10.000 such sentences, summing 30.000 of proposition-explanation pair samples.

3.2 Initial approach

The first thing we tried in the data preprocessing step is stemming, lemmatization and stop-words removal. We have tested several algorithms for these three methods, and performed empirical analysis after training the model. Due to the nature of the generative model we've used, we decided

to skip this part altogether, since the pre-trained word-embedding model is used to dealing with whole sentences, grammatically correct.

3.3 Solution

Generative model used is OpenAI's GPT-2, the former state-of-the-art model in generative problems, currently shadowed by it's successor, GPT-3. GPT-3 is not publicly available, and GPT-2 still remains the best public model. It comes in different shapes, ranging from 117 million parameters to 1.5 billion. In this task we have used the 345 million parameters model, but also experimented with the 117 million one. The capability of GPT-2 we are interested in is called "interactive generation", and based on an input sample, it generates the continuation of the sample until it reaches the End Of Sequence token.

To adapt our problem to the capability of GPT-2, we have arranged the training set in pairs of proposition-answer samples followed by a new token, <|endoftext|> and a newline (eg. Elephants eat helicopters. Helicopters are not food. <|endoftext|>) . GPT-2's End Of Sequence token is not necessarily at the end of a sentence, it can appear in the middle of a sentence, so we need a way to separate samples. Our novel token is added in the byte-pair encoding of the model, and then learned by it, so when it will generate new samples it will delimit them by that token.

We needed about 5000 iterations of fine-tuning until the loss function was not decreasing anymore and we could conclude that the model is well-enough tuned for our problem. Interesting part about GPT-2 in this case is that it has also learned to generate illogical sentences, since it has encountered plenty in the fine-tuning process.

To get an explanation for our input illogical sentence, we just need to feed the illogical sentence to the GPT-2 model, followed by a full stop to mark the end of sentence. Since it was fine tuned on pairs looking like "<Illogical sentence.Logical explanation>" , it has learned that after the illogical sentence and a full stop, the logical explanation is followed, so it will generate it.

3.4 Limitations

We noticed, based on a thorough empirical analysis, that the obtained model has some persistent limitations, regardless of the input. If it is fed a logical proposition, then it will most likely output nonsense words, sometimes not even grammatically correct. The second corner case is when a sentence contains two semantical error, such as "Children fly on the plate". First error is constituted by "Children fly" and the second by "fly on the plate". In these cases, the generative model falls short.

4. Evaluation

In order to evaluate the model, we took in consideration multiple solution, such as BLEU score, which is an *automatic testing tool*, and then we have decided on the *manual evaluation*.

4.1 Automated testing

BLEU (**bilingual evaluation understudy**) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics.
Scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness are not taken into account.

BLEU's output is always a number between 0 and 1. This value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts. Few human translations will attain a score of 1, since this would indicate that the candidate is identical to one of the reference translations. For this reason, it is not necessary to attain a score of 1. Because there are more opportunities to match, adding additional reference translations will increase the BLEU score.

Inevitably, BLEU algorithm because it is an *automatic testing tool*, has different problems.

One of them is the fact that it only outputs 0(for bad explication) or 1(for good explication). From these types of output we can’t extract the motive that the explanation is bad. Is it semantically wrong or grammatically wrong?

Another problem with BLEU is the fact that if we give as output the input sentence, the score will be high, so the model theoretically is very good.

4.2. Manual Testing

To solve this problem, we used a manual testing that implies 100 sentences given to the model and interpret the results by hand.

The test data was provided by the *SemEval* competition[1], among with a score scale[2]. The score has the following form:

Score	Description
0	The reason is not grammatically correct, or not comprehensible at all, or not related to the statement at all.
1	The reason is just the negation of the statement or a simple paraphrase. Obviously, a better explanation can be made.
2	The reason is relevant and appropriate, though it may contain a few grammatical errors or unnecessary parts. Or like case 1, but it's hard to write a proper reason.
3	The reason is appropriate and is a solid explanation of why the statement does not make sense.

5. Results

5.1. Examples of results

- *For case 0*, we entered sentence: "*I got ready for bed so I drank a cup of coffee.”*. The model returned: "*coffee = liquid not solid.”*. There is not necessary a further explanation why this response has an output value of 0.

[1] https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation-and-Explanation/blob/master/ALL%20data/Test%20Data/subtaskC_test_data.csv

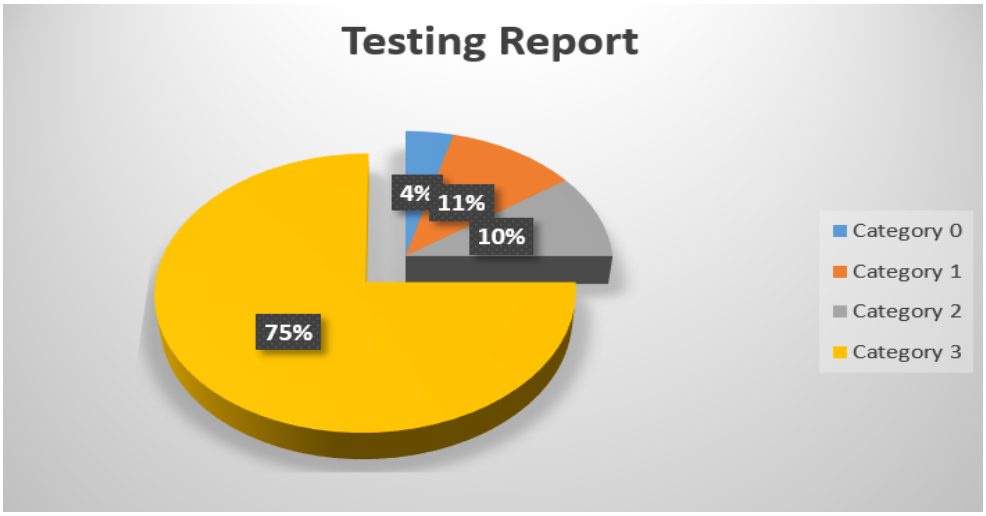
[2] https://www.aclweb.org/anthology/2020.semeval-1.39.pdf

- *For case 1*, we entered the sentence: "*You die twice.*". The model returned: "*In some stories, a cat can die for nine times.*". The explanation is alright for other input like: "*A cat can die one time*", but for our case a better explanation can be made.
- *For case 2*, we entered the sentence: "*Birds like to stay in the same location all year long.*". The model returned: "*Birds go extinct when they only joined a location once in a lifetime.*". The explanation has some scientific facts behind, but it is not the answer that we are looking for. For this example, perhaps a better explanation could not be made.
- *For case 3*, we entered the sentence: "*It might rain later today, be sure to grab your breakfast before you leave.*". The model returned: "*Breakfast is a meal, it will not keep you dry from rain.*". The explanation is valid, even the input sentence is semantically hard to understand for the model.

5.2. General evalution of results

5.3. Statistics of results

After manually testing 100 samples, our model had an accuracy of 75% to generate really good explanations and only 4% of the time the output was completely irrelevant.



Given the fact that the average results have a larger proportion in comparison with the worst results and the best ones have a 75% chance of being obtained, we can appreciate the quality of the implemented system.

Even if the manual testing tend to be subjective, it is the best approach for our model, given the fact that we were able to understand deeply the context of the sentence and tag the results according to its complexity.

6. Comparison with other solutions

Until the recent explosive advancements in machine learning and the massive growth in computational power, this kind of problems were approached either in a fully deterministic way, where programmers tried to think of every corner case that may appear and address it, either in a more mathematical way, for instance modelling the problem using graphs, where the vertexes represent words in a sentence and edges the connections between them. Then, based on the structure of the graph, another graph constituting the logical explanation could be generated, based on deterministic rules as well.

For the SemEval2020 competition, all the teams that placed on top for this task have used the GPT-2 generative model, due to the fact that it has proven itself extremely powerful in all kind of generative tasks, and especially on language generation. The problem has been modelled in diverse ways, and the word embedding model has been either augmented in diverse ways, either changed altogether.

7. Future work

It is clear that the approaches similar to those described above constitute the current basis for the research in this direction, and the experimentation is usually done mostly at the word embedder level. GPT-2 is powerful enough to adapt to any word embedder it uses, so that is where the progress must be made.

We have yet to test the capabilities of the state-of-the-art generative model together with some state-of-the-art word-embedding algorithms. These fall in two categories: pre-trained and not pre-trained. The pre-trained ones are usually trained on huge databases of books or internet posts, and the relationships between words are well made. But for some datasets, such as ours, these relationships may differ than in most use cases, so we can either train the word-embedding together with the generative-model, either train it on other dataset, possibly smaller, that resembles the word styling of the dataset we use for the main task.

8. Conclusions

The problem of Commonsense Validation and Explanation has become increasingly popular among software developers and software analysts due to its increased applicability in various fields such as artificial intelligence, computer vision, etc.

Our solution, although it gave promising results, represents only a beginning, a small step in solving this problem, which has a very high degree of complexity.

With multiple new directions of development studied or to be discovered, our solution can be improved and present a real solution on commonsense validation and explanation.

Also, understanding natural language at a high level requires a long study, as proposed and our solution, which in time could solve real problems, such as: fake news on social networks, whose impact can affect humanity and many other things.

9. Bibliography

[1] https://arxiv.org/pdf/1906.00363.pdf

[2] https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[3] https://arxiv.org/pdf/1810.04805.pdf

[4] https://arxiv.org/pdf/1806.02847.pdf

[5] https://arxiv.org/pdf/1805.06939.pdf

[6] https://en.wikipedia.org/wiki/BLEU

[7] https://www.aclweb.org/anthology/2020.semeval-1.39.pdf