MINOR PROJECT SYNOPSIS

Title: Commonsense Validation Sen Making

Department of Computer Engineering

Faculty of Engineering and Technology

JAMIA MILLIA ISLAMIA UNIVERSITY

Project By – Debal Hussain Abbas (18BCS046)

Mauwaz Ahmed Farooqui (18BCS048)

ABSTRACT

Background:

Introducing common sense to natural language understanding systems has received increasing research attention. It remains a fundamental question on how to evaluate whether a system has the sense-making capability. Existing benchmarks measure common sense knowledge indirectly or without reasoning. In this project, we release a benchmark to directly test whether a system can differentiate natural language statements that make sense from those that do not make sense. The results are evaluated based on the accuracy score.

Methodology:

We are using a transformer model to classify the most sensible sentence out of the given two sentences.

A web application has been designed to provide a suitable interface to the project. The user inputs two statements on the website and the sentence which is sensible is returned as output.

INTRODUCTION

The project is the implementation of the first part of the paper (Sense-Making) titled –

"Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation" authored by Wang Cunxiang, Liang Shuailong, Jin Yili, Wang Yilong, Zhu Xiaodan and Zhang Yue

Let us consider the following example –

> Task: Which statement of the two is against common sense?
>
> Statement1: He put a turkey into the fridge.
>
> Statement2: He put an elephant into the fridge.
>
> Sensible Statement: He put a turkey into the fridge

Natural Language Understanding (NLU) has received increasing research attention in recent years. With language models trained on large corpora, algorithms show better performance than humans on some benchmarks. Compared to humans, however, most end-to-end trained systems are rather weak on common sense. For example, it is straightforward for a human to understand that someone can put a turkey into a fridge but he can never put an elephant into a fridge with basic commonsense reasoning, but it can be non-trivial for a system to tell the difference. Existing datasets test common sense indirectly through tasks that require extra knowledge. They verify whether a system is equipped with common sense by testing whether it can give a correct answer where the input does not contain such knowledge.

The paper mentions two tasks - first task is to choose from two natural language statements with similar wordings which one makes sense and which one does not make sense; The second task is to find the key reason why a given statement does not make sense. This project implements the first task which is the validation of sensible statements.

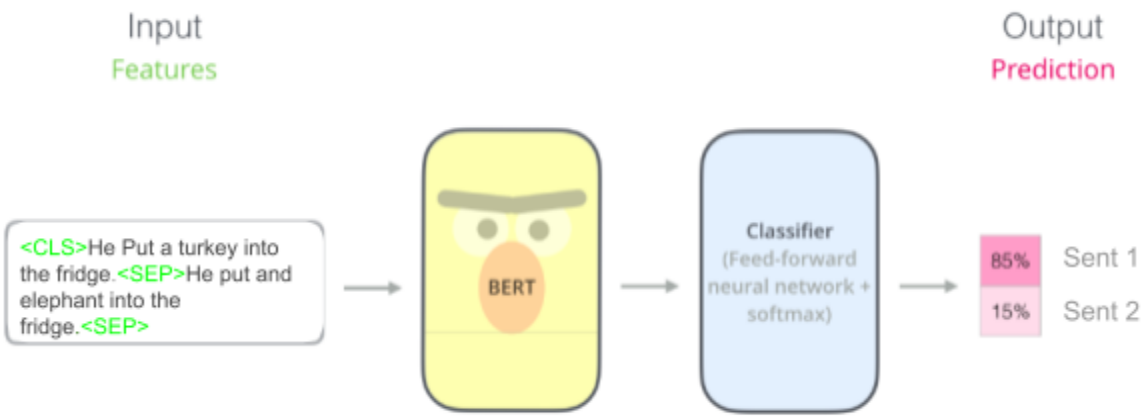
The dataset contains 2021 instances for each subtask, manually labeled by 7 annotators. The dataset is distributed into Train, Test and Validation subsets. Each instance of the data consists of 2 sentences – sentence0 and sentence1, one of which makes sense while the other does not, the context of both remaining the same. Alongside the pair of sentences is a number which is either 0 or 1 indicating which of the above statements makes sense. Each instance of data can be identified by a unique id.

Human performance on the benchmark is 99.1% for the Sen-Making task. In this pilot study, we evaluate contextualized representations trained over large-scale language modeling tasks on our benchmark. Results show that there is still a large gap behind human performance despite that the models are trained over 100 million natural language sentences.

METHODOLOGY

We choose state-of-the-art language models trained over large texts as our baselines, assuming that common sense knowledge is encoded over texts.

We will add a single feed forward layer on top of the transformer language model to having a two neuron and softmax as activation function . It will output **0** if the first sentence makes sense and **1** if the second sentence makes sense.



DEVELOPMENT ENVIRONMENT

The web application has been developed using the Flask framework for the backend and HTML/CSS for the frontend.

Data exploration has been done using Numpy and Pandas. ELMo has been used for contextualized word embeddings.

Transformers and pytorch library has been used for creating the Deep Learning Sequence Model for Classification of the text.

REFERENCES/CITATION

You can use the following if you want to use our dataset or cite our work:

Cunxiang Wang, Shuailong Liang, Yili Jin, Yi-long Wang, Xiaodan Zhu, and Yue Zhang. 2020.SemEval-2020 task 4: Commonsense Validation and Explanation. In Proceedings of The 14th International Workshop on Semantic Evaluation. Association for Computational Linguistics.

```
@inproceedings{wang-etal-2020-semeval,

    title = "{S}em{E}val-2020 Task 4: Commonsense Validation and
Explanation",

    author = "Wang, Cunxiang and Liang, Shuailong and

      Jin, Yili and Wang, Yilong and Zhu, Xiaodan and Zhang, Yue",

    booktitle = "Proceedings of The 14th International Workshop on
Semantic Evaluation",

    year = "2020",

    publisher = "Association for Computational Linguistics",

}


@inproceedings{wang-etal-2019-make,
```

```
   title = "Does it Make Sense? And Why? A Pilot Study for Sense Making
and Explanation",

   author = "Wang, Cunxiang and Liang, Shuailong and

             Zhang, Yue and Li, Xiaonan and Gao, Tian",

   booktitle = "Proceedings of the 57th Annual Meeting of the Association
for Computational Linguistics",

   month = jul,

   year = "2019",

   address = "Florence, Itay",

   publisher = "Association for Computational Linguistics",

   url = "https://www.aclweb.org/anthology/P19-1393",
```

```
   pages = "4020--4026",

   abstract = "Introducing common sense to natural language understanding
systems has received increasing research attention. It remains a
fundamental question on how to evaluate whether a system has the
sense-making capability. Existing benchmarks measure common sense
knowledge indirectly or without reasoning. In this paper, we release a
benchmark to directly test whether a system can differentiate natural
language statements that make sense from those that do not make sense. In
addition, a system is asked to identify the most crucial reason why a
statement does not make sense. We evaluate models trained over large-scale
language modeling tasks as well as human performance, showing that there
are different challenges for system sense-making.",

}
```