FLUTE Extension: Figurative Language Understanding through Textual Explanations

1st Alexandre Casarin
Politecnico di Torino
Turin, Italy
s306940@studenti.polito.it

2nd Federica Lupo
Politecnico di Torino
Turin, Italy
s302644@studenti.polito.it

***Abstract*—Figurative language understanding is an important and challenging task in the field of textual entailment, since it is difficult for machine learners to identify potential texts with non-literal meaning. This work[1], extends the previous study conducted by Chakrabarty et al. [1] on figurative language understanding through textual explanations. The aim is to present models and datasets used in figurative language and try to explore different approaches, such as model exploration, in order to improve the capacity of detecting figurative languages in pairs of text, correctly predicting if the sentences entail or contradict one another. The authors of the previous study fine-tuned a T5 model on the FLUTE dataset with explanations, generated through a model-in-the-loop framework based on GPT-3. In this project, we explore smaller models (*T5-small, T5-base, T5-large)* and observe the influence of changing training set on the performances. Results show that, even with smaller models and restrict computational capacity, it is still possible to obtain reasonable results, similar to consolidated articles in the field.**

***Index Terms*—Natural Language Processing, Figurative Language Understanding, textual entailment**

I. Introduction

Figurative language is present in basically every existing idiom in the world. It consists in textual expressions that not necessarily have their usual meaning, increasing the communication possibilities in a vocabulary. Since figurative language is still a recent topic in Natural Language Processing, there are some common issues faced by researchers in this area, such as lack of good quality datasets or performance benchmarks. Figurative Language Understanding through Textual Explanations [1] (FLUTE) is a new dataset in this field, containing 9000 examples of metaphor, sarcasm, simile or idiom generated with the help of GPT-3. Each example consists of a premise (literal sense phrase), a hypothesis, which is a phrase with the same sentence as the premise but containing the respective figurative language in a part of the text, a label, which tells us if the hypothesis contradicts or entails the premise, and an explanation for the given label. In our work, we train different models in the FLUTE dataset and test new features to understand what are their impacts in the quality of figurative language predictions and hopefully suggest possible improvements. The first extension studied is exactly the impact of model complexity, in which different versions of a T5 model are fine-tuned on the FLUTE dataset and their performances are compared.

T5 (Text-to-Text Transfer Transformer) is a model developed by *Google Research* [2] and is designed to perform various NLP tasks using a unified text-to-text framework. The T5 model can convert all NLP tasks into a text-to-text format, meaning both the input and output are text strings.

Model complexity strongly impacts the accuracy of the predictions. Larger model size generally means higher computational capacity, allowing a better understanding of the language nuances in the training phase and consequently more accurate predictions. On the other hand, experiments can get much more computational exhaustive, specially if there are hardware limitations. Other parameters in the model can also be modified in order to measure their importance in the performance. Examples of easy parameters to be tested are batch size, number of epochs and learning rate. Methods for these experiments are explained in the next section and their results are presented in section III.

After some analysis on the model used, we decided to focus on the dataset exploration to see the impact on the final evaluations. Two datasets have been used separately for the training:

- ***Fig-QA*** [2]: it consists of 10256 examples of human-written creative metaphors in English that are paired as a Winograd schema. It can be used to evaluate the commonsense reasoning of models.
- ***Multi-Genre Natural Language Inference Corpus (MNLI)*** [4]: it is a crowd-sourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). The premise sentences are gathered from ten different sources, including transcribed speech, fiction, and government reports.

In both cases some modification to the original structure have been made (further details in the next section).

II. Methods and experiments

To measure the performances in the experiments, automatic metrics have been used other than the simple label accuracy, as explained in the FLUTE project [1]. An *explanation score*

[1]Project available in this repository: https://github.com/federicalupo/model-in-the-loop-fig-lang.git

(between 0 and 100) is calculated as the mean between BERTScore and BLEURT. The standard accuracy of the model is reported as Acc@0, while Acc@50 and Acc@60 correspond to not only to the label accuracy, but they count the label accuracy given the explanation score greater than 50 and 60, respectively.

A. Extension 1 - Model Exploration

For the model exploration extension, the small, base and large versions of a T5 model have been fine-tuned on FLUTE, containing 60M, 220M and 770M parameters [3] respectively, as shown in table I. From the original dataset, 7035 samples were used for training the models, while 1500 samples were used for evaluating the performance of each setup (750 samples for sarcasm, 250 samples for idiom, metaphor, and simile). For all experiments, a single Tesla T4 GPU was used in Google Colab environment. Adam optimizer was initially used with constant learning rate 5e-5. Another experiment has been done later in the large model, with a different learning rate scheduler (see subsection 4).

TABLE I
T5 MODEL SIZE CHARACTERISTICS

T5-Model Size	**Features**		
	Parameters	***# Layers***	***# Heads***
Small	60M	6	8
Base	220M	12	12
Large	770M	24	16
3B	3B	24	32

1) T5-small: T5-small was trained in 10 epochs with a batch-size of 16. Since this is the smallest model used in this work, we expect its performance in the four different types of figurative language to be the lowest among all models tested, but also the fastest to be trained.

2) T5-base: T5-base was trained in the same conditions presented in T5-small. We expect this model to perform better in terms of accuracy than T5-small, but to have a bigger training time.

3) T5-large: Since we have hardware constraints provided by the Google Colab account, e.g. only a single GPU, T5-large was the biggest model able to be trained in these conditions. Despite that, in order to avoid memory problems, we also had to use batch-size equal 2 and train for only 6 epochs (in previous testing, convergence was reached after around 5 epochs). We expect T5-large to have the biggest performance in terms of Acc@0, Acc@50 and Acc@60, but to be significantly slower than the previous models. It is important to highlight that reducing the batch-size leads to memory usage decrease and faster training time, at the cost of an expected slight drop in the performance.

4) Polynomial learning rate: Since T5-large is the biggest model used in this work, then with biggest expectations of accuracy, we used the same configuration of the last test in this model, despite the usage of a polynomial learning rate instead of a constant scheduler. This study permits a better understanding of how this important parameter affects the final results of the model. Figure 1 shows the difference between a constant scheduler (in blue) and a polynomial one, which decreases the learning rate at each epoch, since the model should be more close to the minimum loss point towards the end of the training phase.

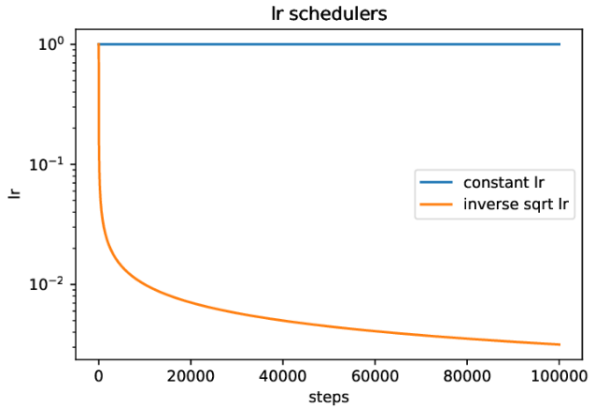

Fig. 1. Learning rate schedulers comparison

B. Extension 2 - Dataset Exploration

As anticipated in the previous section we decided to fine-tuned the model by considering two different datasets after adapting them to the type of input required by the model:

- *Fig-QA*: in the original dataset, given a startphrase, and two possible literal meanings, the binary label indicates if the startphrase is associated to the first literal meaning (label 0), or to the second one (label 1). In order to adapt this dataset to our purpose, we considered the first literal meaning as premise and the startphrase as hypothesis, then if the label proposed was 0, we consider an entailment (label "Entails") between the premise and the hypothesis, otherwise a contradiction (label "Contradicts").
- *MNLI*: this dataset presents a column for premises, one for hypothesis and a label indicating an entailment (label 0), contradiction (label 2) or neutral (label 1) relation between the two. In the dataset used only the labels related to entailment and contradiction have been considered.

In this extension only Acc@0 has been considered as metric since details on the explanations were missing. We tested the ability of the model to distinguish between contradiction or entailment labels.

III. RESULTS

A. Extension 1 - Model Exploration

Tables II, III and IV show the accuracy performances, in the automatic metrics already described before, for the small, base and large T5 models, respectively. As expected, T5-large outperformed the other models in every metric, but even with reduced batch-size and number of epochs, it still took twice the time to be trained on FLUTE if compared to the base model.

On the other hand, standard accuracy for small and base models were slightly above 50%, which is almost a random choice. Our T5-large results are similar to those presented in [1], when trained in T5-3b model, with even better results for metaphor data. This result is important to demonstrate that good results can be achieved in figurative language recognition even with limited computational resources. In table IV, that now can be used as a benchmark, we observe that sarcasm had better accuracy performances if compared to the others figurative language data, followed by idiom. If we compare the training set size for each data type, sarcasm contains around 4000 samples (considering entailment and contradiction pairs), while idiom has 2000 samples, and the others have 1500 samples [1]. Therefore, accuracy performances in the test phase looks highly correlated to the amount of data provided in the training phase, as expected. The models were evaluated on test sets for each category, composed by 750 randomly selected samples from sarcasm dataset, 250 for the other categories. Finally, table V shows that a polynomial learning rate performs worse than the constant rate. Some hypothesis that can explain this results are:

- The initial learning rate is the same as in the constant approach. Increasing this value for the polynomial scheduler could lead to faster convergence.
- Figurative language understanding is complex and samples contain very heterogeneous data. A constant learning rate might be more adequate, since the model keeps learning well, despite of how much it has already been trained.

We can observe that in this new configuration, with the polynomial learning rate, sarcasm is still the class with higher performance, which confirms the hypothesis of training dataset size being essential to achieve good results in figurative language understanding.

TABLE II
T5-SMALL RESULTS

Figurative Type	Accuracy types		
	Acc @0	***Acc @50***	***Acc@ 60***
Sarcasm	59.9	30.0	4.0
Simile	47.2	20.0	2.4
Metaphor	53.6	13.7	1.2
Idiom	50.8	20.4	2.8

TABLE III
T5-BASE RESULTS

Figurative Type	Accuracy types		
	Acc @0	***Acc @50***	***Acc@ 60***
Sarcasm	55.6	43.5	13.5
Simile	51.2	37.6	10.8
Metaphor	51.2	28.2	4.8
Idiom	52.0	43.2	21.2

TABLE IV
T5-LARGE RESULTS (BATCH-SIZE = 2 / EPOCHS = 6)

Figurative Type	Accuracy types		
	Acc @0	***Acc @50***	***Acc@ 60***
Sarcasm	**93.6**	**87.7**	**52.4**
Simile	**62.8**	**56.0**	**27.6**
Metaphor	**82.3**	**65.7**	**33.5**
Idiom	**84.8**	**82.0**	**60.8**

TABLE V
T5-LARGE RESULTS (BATCH-SIZE = 2 / EPOCHS = 6 / POLYNOMIAL LR)

Figurative Type	Accuracy types		
	Acc @0	***Acc @50***	***Acc@ 60***
Sarcasm	81.8	76.4	43.4
Simile	50.0	44.4	22.4
Metaphor	53.6	42.7	15.3
Idiom	59.2	56.8	44.4

B. Extension 2 - Dataset Exploration

Diving into the results for the second extension proposed (table VI), we can see a big drop for the sarcasm category when we pass from the FLUTE fine-tuning to the other training set. Using the Fig-QA as training set, leads to an increase of 10.8% for the simile category in comparison to using the FLUTE training set. While using the MNLI dataset increases of 1.6% the accuracy related to the metaphor category. In table VII, a detail of the dimension of the dataset used is highlighted. It is important to highlight that the results were obtained by training the T5-large model, with batch-size 2 and 6 epochs, since it was the best model obtained in the first analysis. Overall, for three categories out of four, we can see similar performances, except for the sarcasm one in which we have a big drop, motivated also by the difficulties embedded in this category. Understanding sarcasm requires having some context and a deep knowledge of the language, that overcomes shallow patterns.

Finally, it is important to mention that the FLUTE dataset is a dataset created after different analysis and collaboration between human intervention and automatic processes, this leads to more precise information, helpful in boosting the model performances.

IV. CONCLUSION

In this project, starting from the work of Chakrabarty et al. on FLUTE: Figurative Language Understanding through

TABLE VI
T5-LARGE RESULTS PER TRAINING SET

Figurative Type	Accuracy types Acc@0		
	FLUTE	***Fig-QA***	***MNLI***
Sarcasm	**93.6**	64	69.5
Simile	62.8	**73.6**	60.8
Metaphor	82.3	78.2	**83.9**
Idiom	**84.8**	79.2	82.8

TABLE VII
DATASET DIMENSIONS

Dataset	***FLUTE***	***Fig-QA***	***MNLI***
Training	7035	7739	5353
Validation	499	967	669
Test	1500	1500	1500

Textual Explanations, we proposed two in-depth analysis.

First of all, we focused on the fine-tuning of different variations of the T5 model: *small, base, large*, by proving that using a larger model increases the performances, both in the case of the simple accuracy that in the case of the metrics considering explanations, as expected at the cost of computational power.

Then, we tried the Fig-QA and the MNLI dataset to fine-tune the large model, in two cases out of four, fine-tuning on the two dataset led to an increase of the performances, the downside of the two datasets is the less attention paid to create the dataset, as we can see from the previous work with FLUTE dataset.

One of the limitations of dealing with figurative language is that it has the potential to be used in a harmful way, especially against minority and historically disadvantaged groups. All potentially offensive examples have been removed, aware that the removal practise is still subject to personal judgment.

As future steps, further studies can be conducted in other languages, since the main language used in this work is English. Moreover, multiple datasets could be merged together by exploiting the potential of the Large Language Models (as GPT) to create new explanations, even if this would require an important amount of comptutational capacity.

REFERENCES

[1] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[2] Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language, 2022.

[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. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.