An Experiment of Discourse and Sentiment Analysis for the Prediction of Empathy, Distress and Emotion

SURREY-CTS-NLP at WASSA 2022

Shenbin Qian, Constantin Orasan, Diptesh Kanojia,

Hadeel Saadany and Felix do Carmo

Centre for Translation Studies & Department of Computer Science

University of Surrey, UK

Task & Data Introduction

Track 1: Empathy Prediction (EMP), which is a regression task to predict both the empathy and distress score at the essay-level.

Track 2: Emotion Classification (EMO), which is to classify each essay into one of seven classes of emotion.

The main text data are responses from readers describing their feelings after they read some news articles. The target variable is the score which grades readers' empathy and distress level based on Batson's Empathic Concern and Personal Distress Scale (Batson et al., 1987).

Emotion labels are annotated by human annotators after machine learning algorithm pre-labelled these responses.

Size of dataset: 2130 examples in total, including the development dataset

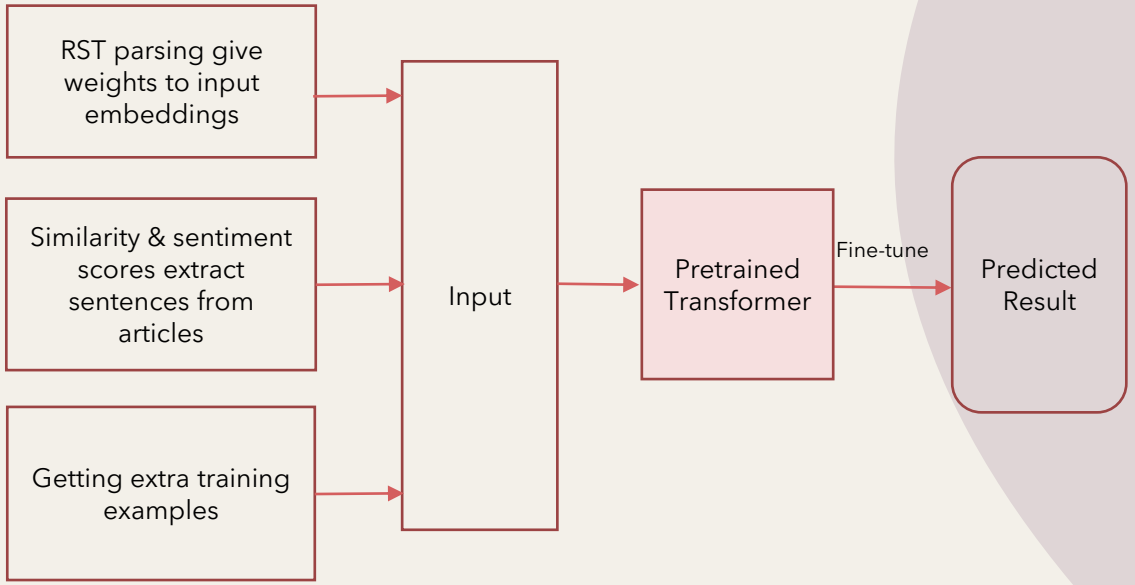
Objectives & Approaches

➢ **Main Objective**

- Solve the problem of small training size
- Get more data from the input perspective
- Explore different methods to solve the issue

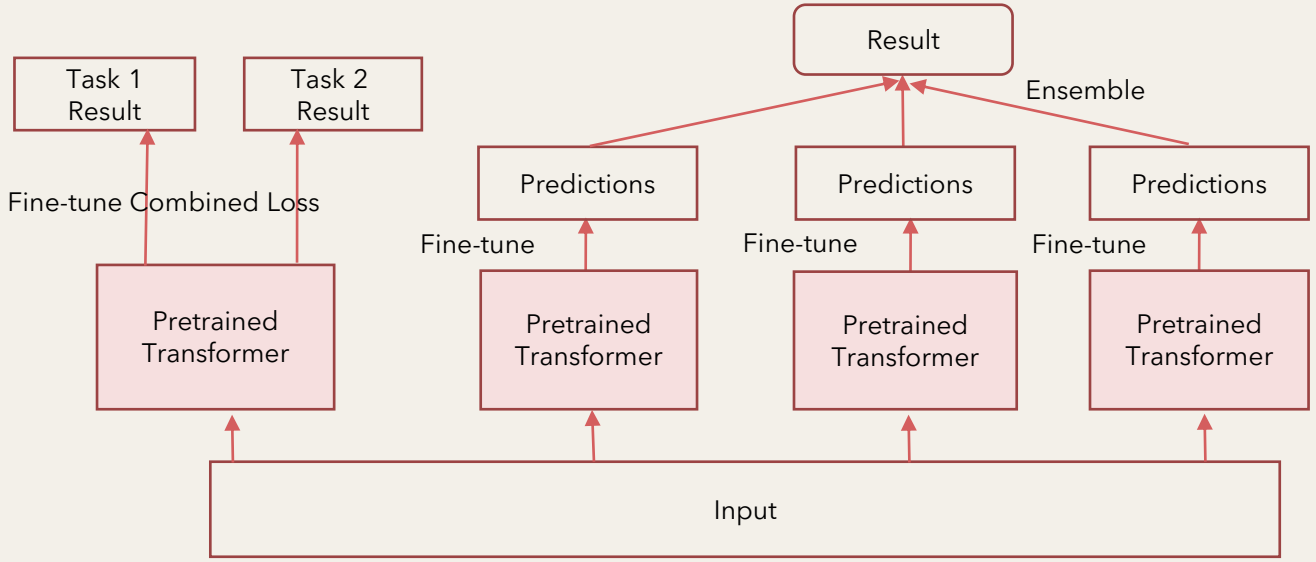
➢ **Data Perspective**

- Incorporating discourse information in the input text
- Extracting features from news articles since essays are responses to articles
- Data augmentation for emotion classification as they are more available



Objectives & Approaches

- **Method Perspective**
- Ensemble learning: more models to tackle less data
- Multi-task learning: correlated tasks to learn information from each other

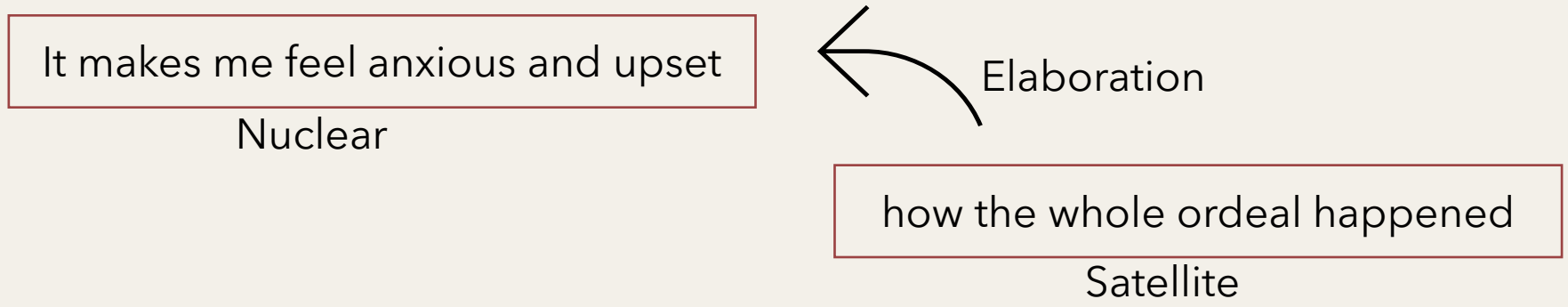


Method Description—RST parsing

Rhetorical structure theory (Mann and Thompson, 1987) parsing is a discourse analysis theory which aims to build a discourse-relation tree for a sequence of text units.

Examples:

Sentence: It makes me feel anxious and upset how the whole ordeal happened.



Method Description—RST parsing

Assumption: Nuclei should be given more weights on the text embeddings while satellites less weights during the training process to make the input more reasonable, since nuclei are those that are more emotional, carrying the intention of the writer, whereas satellites are only a rephrasing of the events in the corresponding news article.

In experiments, we used the text-level discourse rhetorical structure (DRS) parser by Zhang et al. (2021), and tested with different weights to the embeddings of nuclei and satellites for the essays. We found that giving 0.3 to the nuclei and 0.7 to the entire essays for fine-tuning a RoBERTa base model (Liu et al., 2019) leads to our best performance.

Method Description–Similarity & Sentiment Score

- **Motivation**

Similarity and sentiment scores are used with the purpose to get more texts i.e. extra contextual information as input for fine-tuning a transformer model, since the essays and articles can be regarded as one unified discourse.

- **Overview**

Cosine similarity is an effective way to extract sentences that are highly related to the essays from the articles based on the distance of their embeddings.

Sentiment score is usually used to grade the sentiment of a sentence between –1 to 1 (negative to positive). We used it to extract sentences in the articles that contain more extreme sentiments so that more emotional information can be kept together with the essays.

Method Description–Similarity & Sentiment Score

- **Cosine Similarity**

In experiments, we tried to get sentence embeddings from Sentence BERT (Reimers and Gurevych, 2019) and Universal Sentence Encoder (Cer et al., 2018) for the calculation of cosine similarity. We also tested different similarity thresholds to filter sentences in the article.

- **Sentiment Score**

For the calculation of sentiment scores, we used a simple rule-based sentiment analysis tool, VADER (Hutto and Gilbert, 2014) to extract sentiment-intensive features.

Method Description–Data Augmentation

- **Motivation:**

As the original training data is small in size and relatively skewed in distribution, data augmentation is something that we could do to get more training examples, especially for the emotion classification task.

The GoEmotions dataset (Demszky et al., 2020) is a manually annotated high-quality dataset with 27 emotion categories, which we only need those 7 categories.

As texts in the GoEmotions dataset might have different writing styles and sequence lengths compared with our essays, we selected those texts that are longer than 25 words and make sure that more joy and surprise examples are included to compensate the skewed distribution.

Method Description–Multi-task Learning

- **Motivation**

Both Batson's Empathy Theory and the high Pearson correlation score (0.45) of empathy and distress suggests that multi-task learning might help us train one model for both sub-tasks.

In this paper, a weighted loss considering the homoscedastic uncertainty (Kendall et al., 2017) of our two sub-tasks was applied to a RoBERTa model (Liu et al., 2019) to predict both empathy and distress scores.

Method Description–Ensemble Learning

- **Motivation**

Combining the results of several different models would be another way to compensate the small training size.

Majority voting is to use the predictive result that most models predict as the final prediction.

To get a better result, we fine-tuned the RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020) and DeBERTa (He et al., 2020) base models for majority voting.

Results for Empathy Prediction

	RST Parsing	Similarity & Sentiment Score	Multi-task Learning	Simple Fine-tuning
Empathy	0.431	0.501[1]	0.480	0.504
Distress	0.465	0.535	0.458	0.530

The Pearson correlation scores produced by the model using RST parser are not as high as expected, but results using extracted article sentences by cosine similarity and sentiment score are pretty high, especially the distress score. Just fine-tuning a RoBERTa base model also achieves high scores.

Multi-task learning is also not bad at predicting the empathy score, but we might still need to design a better loss function to train the model.

1 Only this result is based on fine-tuning a RoBERTa large model, not the base model

Results for Emotion Prediction

	GoEmotions	Ensemble Learning	Simple Fine-tuning
Accuracy	0.634	0.619	0.646
F1 score	0.548	0.534	0.571
Precision	0.576	0.564	0.595
Recall	0.532	0.520	0.559

The F1 score for the GoEmotions result is higher than the one for ensemble learning, which implicitly suggests that getting more training data is more important than using larger and more models, especially when training datasets are particularly small.

However, just fine-tuning a RoBERTa base model appears to have a slightly better result than data augmentation for this task.

Conclusion

- **Summary**

We tried different ways to improve model performance from the perspective of discourse and sentiment analysis, data augmentation and method optimisation like RST parsing, sentiment score and ensemble learning.

We proposed a reliable method to analyse and extract information from both the news articles and the essays to compensate the small training size for empathy and distress prediction. For emotion classification, simple fine-tuning gave us the best result though we also tried data augmentation and ensemble learning.

Conclusion

- **Discussions and Future Directions**

For empathy prediction, RST parsing or even other methods for discourse analysis is still something we can try to get useful information from the articles.

Demographic and personality information were also included to train a tabular model in training and we achieved our best performance. So adding these additional information is also something worthy trying in future research if they were given.

For emotion classification, how to get and sample extra data to compensate the skewed distribution or experimenting with feature extraction techniques on existing information in the training data like the news articles could be possible ways to improve model performance.

References

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint*.

Kevin Clark, Minh-Thang Luong, Google Brain, Quoc V Le Google Brain, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint*.

C J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216– 225.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. pages 394 Association for Computational Linguistics.