An Approach to the Frugal Use of Human Annotators to Scale up Auto-coding for Text Classification Tasks

❖ Li'An Chen (li'an.chen@anu.edu.au)

❖ Hanna Suominen (hanna.suominen@anu.edu.au)

Australian National University

Acknowledgement We are grateful for the support from the anonymous paper reviewers, Prof. Inger Mewburn, Dr. Will Grant, Dr. Lindsay Hogan, Chenchen Xu, Emsi Burning Glass Inc, PostAc®, and ANU CV Discovery Translation Fund2.0.

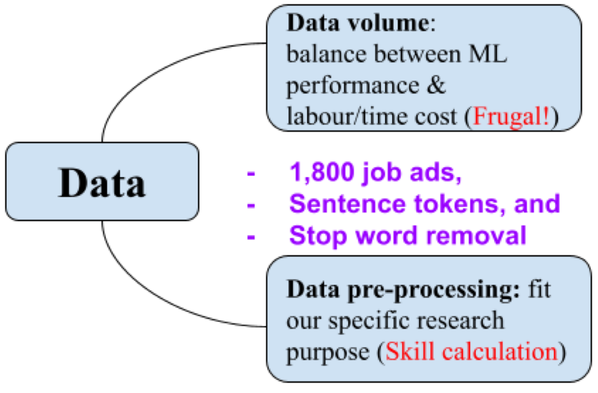
Introduction

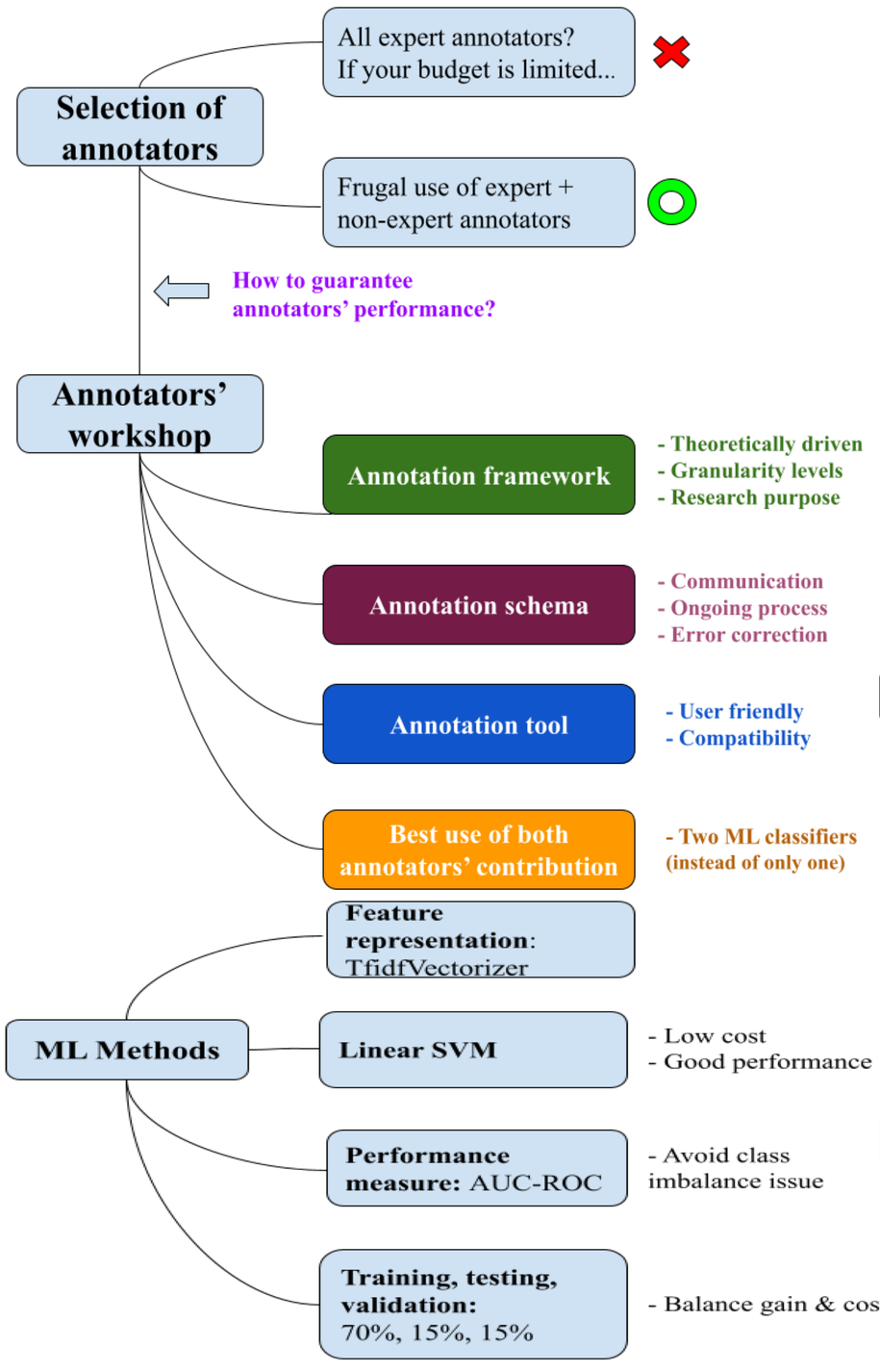
Human annotation for establishing the training data is often a very costly process in natural language processing (NLP) tasks which has led to frugal NLP approaches becoming an important research topic. Many research teams struggle to complete projects with limited funding, labor, and computational resources. Driven by the Move-Step analytic framework theorized in the applied linguistics field, our study offers a rigorous approach to the frugal use of two human annotators to scale up auto-coding for text classification tasks.

Research questions

- For automatic text classification tasks, how could we design human annotators' workshop frugally and at the same time maintain good performance of the machine?
- How could we design the human annotators' workshop to enable easy identification and fixation of problems in the human annotation schema?
- If multiple human annotators were involved, which annotator's labelled data should be adopted for training?

Methods





Primary findings

- The frugal use of an expert annotator and a non-expert annotator generated an averaged Cohen's Kappa of 0.76.
- The total time investment of our frugal approach to human annotation was 376 hours (the time consumed by two human annotators).
- The frugal use of only two human annotators plus a limited amount of labelled data resulted in an averaged area under the receiver operating characteristic (ROC) curve (AUC) score of 0.80.
- Differentiation of coarse-grained and fine-grained labels allowed for enhanced interpretability of the ML performance. It also allowed for strategically hybrid use of multiple human annotators' labels to optimize the ML performance.

Discussion & Conclusion

- Frugal use of human annotators can generate good inter-rater agreement & ML performance, but rigorous design of the annotating process is a must.
- 'Neutering' might not apply well to all NLP tasks, as the benefits of having two granularities in our study show, particularly when interpretability is concerned.
- Our study does not guarantee generalizability. Instead, we recommend researchers prioritize the annotation methods' compatibility with specific research purposes.

Future directions

- It would be interesting in the future to compare the performance given by 1) crowdsourcing, 2) pure expert annotators, and 3) expert + non-expert annotators (rigorous process design involved).