Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

Data-Efficient Algorithms and Neural Natural Language Processing: Applications in the Healthcare Domain

Heereen Shim
Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Leuven, Belgium
heereen.shim@kuleuven.be

Abstract

Recently proposed pre-trained language models can be easily fine-tuned to a wide range of downstream tasks. However, fine-tuning requires a large training set. This PhD project introduces novel natural language processing (NLP) use cases in the healthcare domain where obtaining a large training dataset is difficult and expensive. To this end, we propose data-efficient algorithms to fine-tune NLP models in low-resource settings and validate their effectiveness. We expect the outcomes of this PhD project could contribute to the NLP research and low-resource application domains.

1 Introduction

A pre-trained language model [Devlin *et al.*, 2019] can be easily fine-tuned to various downstream tasks. However, fine-tuning still requires a large task-specific labelled dataset and some application domains have limited access to large-scale data. The main goal of this PhD project is to fine-tune NLP models with the limited data for low-resource application domains, such as healthcare. In this project, we aim to support healthcare professionals and encourage healthcare recipients to engage in the care processes. Specifically, we introduce novel NLP use cases in the three phases of a sleep coaching programme, from assessment to coaching and monitoring. To this end, we focus on the following research questions:

RQ1. How can we fine-tune a model when only a small-sized training set is available?

RQ2. How can we fine-tune a model when only a small subset of data is labelled? Also, how can we use manual labelling resources efficiently in terms of performance gain per labelling effort?

RQ3. Can we exploit other resources (e.g., structured information, prior knowledge, etc) to improve the performance?

Throughout the PhD project, these three research questions are addressed within the 3 use cases as illustrated in Table 1.

Use-cases	RQ1	RQ2	RQ3
Assessment	✓	✓	✗
Coaching	✓	✓	
Monitoring	✓	✗	✗

Table 1: Use-cases and research questions (RQ).✓and ✗indicate a completed task and a work-in-progress task, respectively.

2 Methodologies

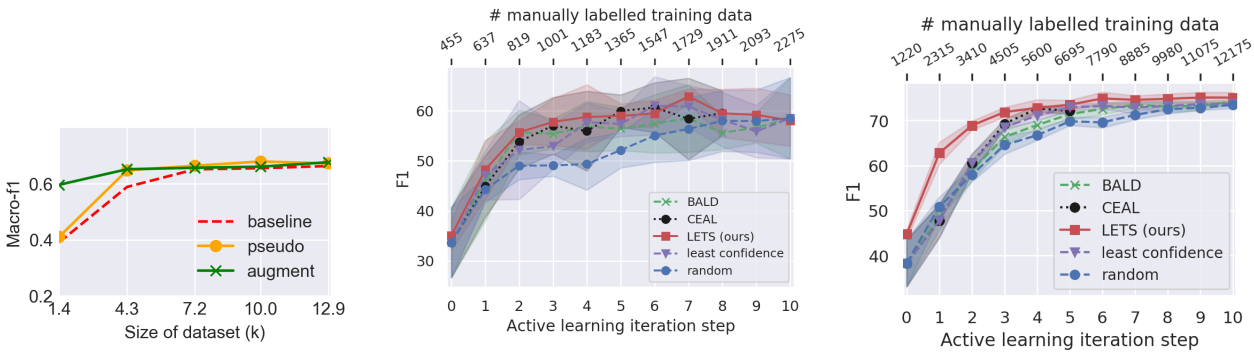
Data Augmentation and Semi-Supervised Learning. The first use case is understanding the participants' complaints based on free text and parsing them into pre-defined sleep issue categories. Within this use case, we consider a limited data resource setting when we have a small-sized labelled training set with a large but unlabelled training set. To mitigate this, we proposed a method which is a combination of data augmentation and semi-supervised learning. The data augmentation technique increases the size of the labelled data set and semi-supervised learning is an iterative learning framework that annotates unlabelled data by using the trained model's predictions.

Label Augmentation and Active Learning. The second use case is understanding each participant's experience by analysing their reviews. We aim to build a system that can extract fine-grained sentiment values towards multiple aspects. In this use case, we consider how to efficiently label and unlabelled training set. To this end, we proposed a label-efficient training scheme that integrates three elements: (i) a task-specific pre-training to exploit unlabelled task-specific corpus data; (ii) label augmentation to maximise the utility of labelled data; and (iii) active learning to strategically prioritise unlabelled data points to be labelled.

Synthetic Data Generation and Multi-Task Learning. The third use case is monitoring sleep activity and assessing subjective sleep quality. This use case focuses on extracting temporal information of sleep-related events and the corresponding subjective information, which is difficult to be captured by sensors or a structured questionnaire. The main technical challenge is enhancing the numeracy of a pre-trained language model with the lack of training data. To address this, we proposed a synthetic data generation algorithm and a novel multi-task model.

3 Results and Next Steps

Data Augmentation and Semi-Supervised Learning. In our first study [Shim *et al.*, 2020], we investigated the ef-



(a) Results on various dataset sizes (b) Results on the custom dataset (c) Results on the benchmark dataset

Figure 1: Experimental results from the previous studies: (a) [Shim *et al.*, 2020], (b) and (c) [Shim *et al.*, 2021a].

fect of data augmentation and semi-supervised learning. We showed that data augmentation can improve performance, especially with a small training dataset (Figure 1a). Results also show that data augmentation can improve the performance of minority label classes which is critical when the data is imbalanced. More details can be found in the paper [Shim *et al.*, 2020]. In the future study, we will investigate integrating knowledge, such as demographic information or the prior knowledge of class distribution into a language model.

Label Augmentation and Active Learning. In our second study [Shim *et al.*, 2021a], we proposed a novel label-efficient training scheme to not only effectively reduce manual labelling efforts but also maximise the utility of data. Results show that the proposed method can reduce manual labelling efforts 2-3 times and increase generalisability. More importantly, we validate the effectiveness of the proposed method on the custom (Figure 1b) and benchmark datasets (Figure 1c).

Synthetic Data Generation and Multi-Task Learning. In our third study [Shim *et al.*, 2021b], we evaluated the effectiveness of utilising synthetic data and multi-task learning. Experimental results show that using synthetic data can improve the performance when the augmentation factor is 3. Also, we found that training a model for a target task (i.e., normalised time value extraction) with an auxiliary task (i.e., answer span detection) can improve the performance significantly when the synthetic data is used. In the future study, we will extend this work to the temporal reasoning that requires reasoning over the text.

4 Conclusion

This extended abstract summarises the research outcomes of the PhD project on low-resource NLP for healthcare applications. We briefly introduced the proposed methods, including data augmentation techniques and learning strategies, and the results. The main contribution of this PhD project is the proposed methods are reusable for other applications. Therefore, we expect that low-resource application domains beyond healthcare can benefit from the research outcomes. Also, we believe this project will bring a broader impact by supporting healthcare professionals and empowering healthcare recipients.

Acknowledgments

I thank my supervisors for their valuable feedback and support throughout the project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766139. This article reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains.

References

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[Shim *et al.*, 2020] Heereen Shim, Stijn Luca, Dietwig Lowet, and Bart Vanrumste. Data augmentation and semi-supervised learning for deep neural networks-based text classifier. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1119–1126, 2020.

[Shim *et al.*, 2021a] Heereen Shim, Dietwig Lowet, Stijn Luca, and Bart Vanrumste. Lets: A label-efficient training scheme for aspect-based sentiment analysis by using a pre-trained language model. *IEEE Access*, 9:115563–115578, 2021.

[Shim *et al.*, 2021b] Heereen Shim, Dietwig Lowet, Stijn Luca, and Bart Vanrumste. Synthetic data generation and multi-task learning for extracting temporal information from health-related narrative text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 260–273, Online, November 2021. Association for Computational Linguistics.