Leveraging Active Learning for Few-Shot Relation Extraction using Large Language Models in the Biomedical Domain

Exposé for a Master Thesis

Dan Plischke
Supervisor: Ulf Leser

1 Motivation

The exponential growth of biomedical literature in recent years has led to a wealth of unstructured data containing invaluable insights into biological and chemical processes in living organisms [1]. The information present in this pool of unstructured data can be useful for disease understanding in the discovery and development phase of the drug development process [2]. The automatic information extraction process, for making this information easily accessible, harnesses approaches from the field of natural language processing (NLP). For example, Named Entity Recognition, Named Entity Linking, Coreference Resolution and Relation Extraction. Relation extraction is especially interesting in the biomedical field as interactions between entity types, such as proteins, genes, and chemical compounds, drive biological and chemical processes within living things. Understanding and leveraging existing knowledge about these processes, such as protein-protein interactions or compound-protein interactions is critical for efficiently developing novel drugs [3] [4].

Many approaches have been proposed for the relation extraction task in the biomedical domain, such as, prototypical networks [5], matching networks [6], knowledge graph-based systems [7] or embedding based systems [8]. In recent years however, research in the field of large language models (LLMs) has shown promising results in many NLP tasks [9], including relation extraction [10].The landscape of relevant relationship and entity types can shift during drug development projects. This leads to a shortage in data that these algorithms can be trained on to perform relation extraction on new entity pairs. Few-shot learning (FSL) is a category of approaches which can cope with limited amounts of training data to train a machine learning algorithm. In these low-resource settings, i.e., where limited labelled data is available, unlabeled data is often readily available. Due to this, the problem of selecting the best training samples to achieve the best generalization arises. This is where Active Learning (AL) approaches offer query strategies, in the so-called pool-based sampling setting, to select the most informative instances from a pool of unlabeled instances (unlabeled dataset) [11]. The union of few-shot learning using large language models and active learning for the relation extraction task in the biomedical domain has not been widely explored in literature yet. This is where this thesis will provide insights into the performance of few-shot learning for biomedical relation extraction enhanced by active learning.

2 Objective

The goal of this thesis is to define an evaluation approach for comparing the computational and qualitative performance of four distinct few shot learning methods on the biomedical relation extraction task. The concrete measures of computational and qualitative performance are outlined in the methodology section.

The beforementioned few-shot learning methods are the following:

Non-Weight Updating Methods

- **In-Context Learning:** tasks are reformulated into prompts using examples and/or task descriptions. Based on this, the model can generate responses or completions for a variety of tasks [9]. Using this approach, the model does not have to be running on a local machine of the user but can be me accessible through, e.g., an API. This approach will use the LLaMA model, which is described below.

Weight Updating Methods

- **Parameter-Efficient Fine-Tuning (**PEFT**):** PEFT is a collective term for fine-tuning methods which freeze some parameters of a pretrained LLM which are then not affected by further fine-tuning [13].
- **Meta Learning:** Meta-learning leverages knowledge acquired from previous tasks to enable faster adaptation and learning on new tasks with limited data. In the context of NLP, meta-learning involves training the model on a set of tasks to capture the shared patterns, structures, and representations across different NLP tasks. The meta-learner learns to effectively adapt its parameters to new tasks in a few-shot setting [14].

Few-shot learning methods can be used on a variety of LLMs. The choice of models is influenced by the FSL method as well as the models pretraining objective and dataset. Based on this, the following models are chosen for the evaluation in this thesis:

- **LLaMA:** A newly presented, open source LLM from Meta, claiming performance improvements over GPT-3 [15].
- **RoBERTa PM:** A BERT based model with pretraining objective enhancements, trained on the PubMed and PubMed Central corpus [16].

The active learning approaches that are going to be evaluated are the following:

- **Least Confidence:** Selects instances with the least prediction confidence. [17]
- **Contrastive Active Learning:** selects instances whose k-nearest neighbors exhibit the largest mean Kullback-Leibler divergence. [18]

To train and evaluate the previously defined models, two fully labeled datasets are chosen. This allows automatic retrieval of labels from relevant instances, identified by the active learning algorithms. This is described further in the methodology section. The datasets that are going to be used are the following:

- **DrugProt:** manually curated dataset of chemical and protein interactions from 5.000 abstracts and contains 24.526 relation triples [19].
- **BioRED:** contains relationships between genes, diseases, chemicals, variants, species, and cell lines and has been manually curated, resulting in 6.503 relation triples from 600 abstracts [20].

Based on the scope, defined in this section, these are the research questions to be answered:

- Can active learning techniques improve the qualitative performance of smaller models to be comparable with larger models on the biomedical relationship extraction task in the few-shot setting?
- How does the qualitative performance of few-shot learning methods, enhanced by active learning, compare to supervised learning methods on the biomedical relationship extraction task?

3 Related Work

Relation extraction has widely been explored using supervised learning approaches without large language models e.g., recurrent neural networks [21], CNNs [22] or graph-based models [23]. In recent years, LLMs have been trained in a supervised manner on the relation extraction task or the embeddings generated from them used to train other models [24]. In the few-shot setting, in-context learning [25], multi-task learning, prototypical networks [26] and matching networks [27] have been proposed. As to the best of the authors knowledge, Li et. al. [28] were the only ones proposing an active learning method for few-shot relation extraction in the biomedical field, using BERT as the base language model. This clearly indicates a lack of research in the biomedical field for few-shot relation extraction enhanced by active learning.

4 Methodology

One of the main goals of the methodology of this thesis is to automate the training and evaluation process of the previously mentioned models, few-shot learning and active learning methods. To achieve this, the SageMaker[1] service offered by Amazon Web Services, is going to be used to automatically submit training and evaluation jobs for the various model configurations. The models are going to be trained on the same instance size i.e., hardware requirements and automatically evaluated on this machine to provide comparable results. For the computational performance evaluation, the evaluation will be conducted multiple times to explore the variance induced by the underlying system. The concrete measures for the computational performance evaluation are video ram usage of the graphics processing unit (GPU), GPU usage and runtime. For the qualitative performance evaluation, F1 score and accuracy will be used.

The training and evaluation process for each composition of base-model, few-shot learning method, active learning method and dataset is depicted in Figure 1. First, the selected, labelled, dataset is split into a training and evaluation set. The training set is then fed into the active learning algorithm to select the most relevant training samples. If the active learning algorithm requires e.g., a model confidence score, a random subset is selected as the first training subset. For weight updating few-shot learning methods, the model is then trained on the selected subset. For non-weight updating few-shot learning, i.e., in-context learning, the model directly goes into the performance evaluation step. Once the performance has been determined, and it passes the stopping criterion, the next subset to train the model on is selected using the active learning method. This is done until it reaches the maximum number of shots previously defined, or the performance measure does not pass the stopping criterion. The stopping criterion is overall uncertainty which was proposed by Zhu et.al and uses the normalized prediction entropy as a performance measure and a manually defined threshold to stop the training process [29].

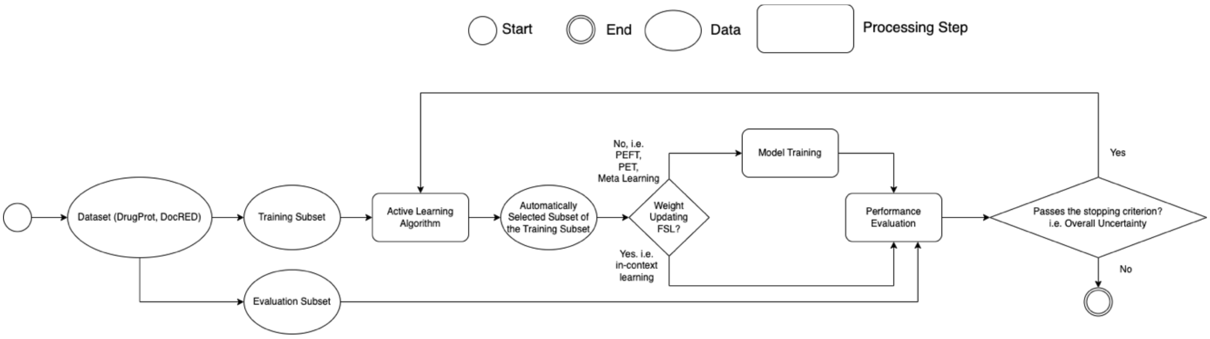
Start
End
Data
Processing Step
Dataset (DrugProt, DocRED)
Training Subset
Evaluation Subset
Active Learning Algorithm
Automatically Selected Subset of the Training Subset
Weight Updating FSL?
No, i.e. PEFT, PET, Meta Learning
Yes. i.e. in-context learning
Model Training
Performance Evaluation
Passes the stopping criterion? i.e. Overall Uncertainty
Yes
No

Figure 1: Model Training and Evaluation Approach

[1] https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html, Accessed 16 August 2023

5 Bibliography

[1] Cohen, A. M. and Hersh, W. R. (2005). “A survey of current work in biomedical text mining” in Briefings in Bioinformatics, 6(1), pp. 57–71.

[2] U.S. Food and Drug Administration (2018). “Step 1: Discovery and Development”. Available at: https://www.fda.gov/patients/drug-development-process/step-1-discovery-and-development. Accessed 20 July 2023.

[3] Mohs, R.C. and Greig, N.H. (2017). “Drug discovery and development: Role of basic biological research”. Alzheimer’s & Dementia: Translational Research & Clinical Interventions, 3(4), pp.651–657.

[4] Perera, N., Dehmer, M. and Emmert-Streib, F. (2020). “Named Entity Recognition and Relation Detection for Biomedical Information Extraction”, Frontiers in Cell and Developmental Biology, 8.

[5] Liu T, Ke Z, Li Y and Silamu W. (2023). „Knowledge-enhanced prototypical network with class cluster loss for few-shot relation classification”. PLoS One.18(6):e0286915.

[6] Wei Z, Guo W, Zhang Y, Zhang J and Zhao J. (2023). “Bidirectional matching and aggregation network for few-shot relation extraction”. PeerJ Comput Sci. ;9:e1272.

[7] S. Yin, W. Zhao, X. Jiang and T. He (2020). "Knowledge-aware Few-shot Learning Framework for Biomedical Event Trigger Identification". 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), pages. 375-380.

[8] Sänger, M. and Leser, U. (2021). Large-scale entity representation learning for biomedical relationship extraction, Bioinformatics, Volume 37, Issue 2, pages 236–242.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A et al. (2020). “Language Models are Few-Shot Learner”, in Larochelle, H. et al. (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 1877–1901.

[10] Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang (2023). “How to Unleash the Power of Large Language Models for Few-shot Relation Extraction?”, in Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP). Toronto, Canada (Hybrid): Association for Computational Linguistics, pp. 190–200.

[11] Zhang, Z., Strubell, E., and Hovy, E. (2022). “A Survey of Active Learning for Natural Language Processing”. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 6166–6190). Association for Computational Linguistics.

[13] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. and Chen, W. (2021). “LoRA: Low-Rank Adaptation of Large Language Models”. CoRR, abs/2106.09685.

[14] Lee, H.y., Li, S.W. and Vu, T. (2022). “Meta Learning for Natural Language Processing: A Survey”. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 666–684).

[15] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et.al. (2023). “LLaMA: Open and Efficient Foundation Language Models”.

[16] Lewis, P., Ott, M., Du, J., and Stoyanov, V. (2020). “Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art”. In Proceedings of the 3rd Clinical Natural Language Processing Workshop (pp. 146–157). Association for Computational Linguistics.

[17] Lewis, D.D. and Gale, W.A. (1994). “A Sequential Algorithm for Training Text Classifiers”. In: Croft, B.W., van Rijsbergen, C.J. (eds) SIGIR ’94. Springer, London.

[18] Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y. and Slonim, Y. (2020). “Active Learning for BERT: An Empirical Study”. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7949–7962.

[19] Krallinger, M., Rabal, O., Miranda-Escalada, A. and Valencia, A. (2021). “DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions (1.2)” [Data set]. Zenodo.

[20] Luo, L., Lai, P.T., Wei, C.H., Arighi, C. and Lu, Z. (2022). “BioRED: A rich biomedical relation extraction dataset”. Briefings in Bioinformatics, Volume 23, Issue 5, September 2022.

[21] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). “Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification”. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 207–212). Association for Computational Linguistics.

[22] Wang, L., Cao, Z., Melo, G., and Liu, Z. (2016). “Relation Classification via Multi-Level Attention CNNs”. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1298–1307). Association for Computational Linguistics.

[23] Baldini Soares, L., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). “Matching the Blanks: Distributional Similarity for Relation Learning”. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2895–2905). Association for Computational Linguistics.

[24] Su, P., Peng, Y., and Vijay-Shanker, K. (2021). “Improving BERT Model Using Contrastive Learning for Biomedical Relation Extraction”. In Proceedings of the 20th

Workshop on Biomedical Language Processing (pp. 1–10). Association for Computational Linguistics.

[25] Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J. and Kurohashi, S. (2023). "GPT-RE: In-context Learning for Relation Extraction using Large Language Models".

[26] Wen, W., Liu, Y., Ouyang, C., Lin, Q. and Chung, T. (2021). "Enhanced Prototypical Network for Few-Shot Relation Extraction". Information Processing & Management, vol. 58, no. 4, July 2021, p. 102596.

[27] Liu, F., Lin, H., Han, X., Cao, B. and Sun, L. (2022). "Pre-training to Match for Unified Low-shot Relation Extraction". In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5785–5795). Association for Computational Linguistics.

[28] Q. Li, H. Xu, H. Wang and B. Tang (2022). "S3 AAL: Support Set Selection based on Adversarial Active Learning for Medical Few-Shot Relation Extraction". 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

[29] Zhu, J., Wang, H. and Hovy E. (2008). "Multi-Criteria-Based Strategy to Stop Active Learning for Data Annotation". In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1129–1136.