TALP-UPC at ProbSum 2023: Fine-tuning and Data Augmentation Strategies for NER

Neil Torrero, Gerard Sant, Carlos Escolano
TALP Research Center, Universitat Politècnica de Catalunya,Barcelona
{neil.torrero, gerard.muniesa}@estudiantat.upc.edu, carlos.escolano@upc.edu

Abstract

This paper describes the submission of the TALP-UPC team to the Problem List Summarization task from the BioNLP 2023 workshop. This task consists of automatically extracting a list of health issues from the e-health medical record of a given patient. Our submission combines additional steps of data annotation with finetuning of BERT pre-trained language models. Our experiments focus on the impact of finetuning on different datasets as well as the addition of data augmentation techniques to delay overfitting.

1 Introduction

Healthcare is a vital sector in our society, playing a crucial role in both the prevention and treatment of diseases for the general population. In addition, healthcare workers often face large workloads due to the number of patients and emergencies that may occur during their shifts. Therefore, finding ways to automate their most repetitive tasks is worth exploring with AI system. With the addition of these systems, healthcare workers could reduce their cognitive load while at the same time providing better care and diagnosis to their patients. With these objectives in mind, this shared task aims to develop systems able to summarize medical health records. Healthcare workers collect daily notes on the treatment of patients, which are stored in electronic eHealth records. The summarizer must be able to extract the most relevant health issues from these records.

One of the main difficulties when creating these systems is the lack of available annotated data. Medical records are sensible material requiring a thorough revision in order to ensure the privacy of the patients, which implies some degree of human supervision. As a result of these steps, these datasets are usually expensive to create and include a limited number of examples. A popular approach for these tasks is resorting to recent NLP techniques such as Transformer (Vaswani et al., 2017) that allow working with large text context as well as pre-trained language models such as BERT (Devlin et al., 2019) that allow models to benefit from transfer learning from large corpora for several languages.

In this paper, we present our submission to the shared task of problem list summarization of the BioNLP 2023 workshop. The main contributions of our system are:

- Data preparation and annotation of the provided data to achieve consistent labeling between all corpora employed.
- Experimentation on the impact of BERT finetuning on different corpora and domains.
- Experimentation of different amounts of data augmentation and their impact on the results.

All code use for the this work is available at the following link: https://github.com/NeilTorrero/BioNLP.

2 Related Work

In this section, we introduce the general concepts of Named-Entity-Recognition (NER) as well as previous methods based on pre-trained language models and data augmentation to fine-tune them on limited amounts of data.

2.1 End-to-End NER

Named entities can be defined as nouns and complements that define an entity in a text. Such entities can be further classified as locations, people, or as in the case of our system, diseases name. The task of automatically extracting these entities is called Named-Entity-Recognition (NER). Systems usually approach this task as a sequence labeling task following the BIO tagging schema (Ramshaw and Marcus, 1995). In this schema, all tokens are

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 497–502
July 13, 2023 ©2023 Association for Computational Linguistics

tagged as either Beginning of an entity (B), Inside of an entity (I), or Outside of an entity (O). These tags can also be enriched to describe the entity's type. One drawback of this schema is that it produces significantly unbalanced sequences, as most of the tokens in a sentence do not belong to a named entity and are consequently tagged as O.

Several works have proposed methods to address NER using deep learning techniques. Hammerton (2003) proposed using LSTM (Hochreiter and Schmidhuber, 1997) to perform sequence labeling based on contextual token representations. Similarly, Collobert et al. (2011) proposed an approach based on CNN (LeCun et al., 1989). A common trait of these architectures is labeling each token independently without considering the sequence of labels already produced. To overcome this issue, Lample et al. (2016) proposed using LSTM+CRF, bridging contextual representations with a sequence decoding algorithm that considers the already decoded path.

2.2 Pre-trained Language Models

With the surge of pre-trained models such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). These architectures consist of a pre-training step on large unannotated corpora, allowing them to learn contextual representations that can be applied to other tasks, such as NER. The most common approach is fine-tuning, where the pre-trainied model's weights are frozen, and only the last linear layer, or classification head, is trained to ensure knowledge transfer without catastrophic forgetting of the model's original representation space. The main advantage of this approach is creating systems that benefit from transfer learning from pre-training on a large dataset when training on a smaller task dataset that can be several orders of magnitude smaller.

2.3 Data Augmentation

Another way of seeking better performance from a small corpus is data augmentation. This technique transforms the input data, creating different perspectives and allowing the model to learn new patterns from the same data. Another benefit of this technique is delaying over-fitting, as the different transformations prevent the model from memorizing the training data easily. Popular approaches to this method include randomly swapping or deleting tokens from the sentence, synonym substitution, and token masking (Lewis et al., 2020). These methods have in common that by modifying the sentence, systems have to use information from the context instead of individual tokens, leading to systems that better represent the context information and are less dependent on the positional information.

3 Methodology

Inspecting the MIMIC dataset, we can observe that most of the example's ground truths follow the same pattern. We have a list of diagnoses split by *;* without connectors or additional information for each example. In order to generate a summary that follows these ground truths, we propose a system based on BIO tagging using two additional corpora and a Fine-tuning step based on BERT. Then we process the inputs of the patient's treatment notes to extract the keywords and transform them into the expected test format.

3.1 Data processing

Figure 1 shows an example of the data preprocessing and annotation used in our experiments. The first thing we noticed when working with the MIMIC dataset was the presence of special characters such as *[** Know Last Patient Name **]* to anonymize sentences that include sensitive data. In order to work with English grammatical sentences, those tags were removed from the sentences.

Sentence: Mr. [**Known lastname 4385**] is a 37 yo male with lower GI bleed in setting of active UC flare. Remove Anonymization: Mr. Known lastname is a 37 yo male with lower GI bleed in setting of active UC flare.
Tokenized: 'mr', '.', 'known', 'lastname', 'is', 'a', '37', 'yo', 'male', 'with', 'lower', 'gi', 'bleed', 'in', 'setting', 'of', 'active', 'uc', 'flare', '.'
Annotation: 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'I', 'I', 'O', 'O', 'O', 'O', 'B', 'I', 'O'

Figure 1: Example of annotation from the MIMIC corpus

Input data is divided into three columns: Assessment, Subjective Sections, and Objective Sections. Upon reviewing the various examples and comparing them to the ground truth summary, we concluded that the first two columns contained the majority of the information. In contrast, the latter column primarily consisted of measurements and symptom checklists indicating whether the patient exhibited them. Therefore, we decided to skip this

column, significantly reducing the input context length. Other modifications tested were training on each section individually or concatenated together into a single input. We also performed experiments using individual and concatenated sections to increase the number of examples.

After cleaning, to make the MIMIC notes work for token classification, each of the examples had their inputs divided into tokens and added a label for each of them following the same format as the others with the BIO schema. Sentences were manually labeled to match the list of terms in the ground truth summary by selecting the words in the different columns. A tool was developed [1] to create and revise annotations graphically, which sped up the process significantly.

3.2 Model Fine-tuning

The following step of our submission is the training of the NER classifier. For all our experiments, we used BERT as the base of our system and fine-tuned it for our task.

Data-wise, two different strategies were tested. First, we tried to fine-tune the model on all the data available for the task. In this experiment, we tried to maximize the data available during training. Second, we tried a two-step fine-tuning strategy. The BERT model is trained in the first step without using the MIMIC corpus. The rationale behind this step is to have a model that can perform the task without focusing on the evaluation's domain. In the second step, we fine-tune the MIMIC data only to enforce a final model that adapts better to the specific details of the evaluation.

As the MIMIC corpus is less than 800 sentences long, we added data augmentation to this second fine-tuning step to delay over-fitting. We randomly replace tokens with BERT's *[MASK]* token for each sentence. We tested various probabilities, ranging from 5% to 40% of the input tokens being replaced. The labels of the masked tokens remained unchanged. By applying this step, the model must attend to the context of the masked tokens to classify, reducing the dependency on known words or positional information over the input sequence.

3.3 Post-processing

After fine-tuning the model, we observed a series of recurrent issues in the outputs resulting from the format used in both medical notes and labels. We noticed that medical notes usually included the same problem several times, while the label did not include repetitions. Another problem was the differences in using acronyms between notes and labels. As a result, some diseases correctly identified in the text were considered wrong by the metric as the strings did not match. We applied several post-processing steps to mitigate these issues. We removed duplicated terms for all outputs, considering acronyms and partial forms of multi-word entities. Finally, we concatenated and separated all entities using the *;* delimiter.

[1] Available here

4 Experimental Framework

4.1 Datasets

In addition to the notes provided by MIMIC III, we also employed two additional datasets from the biomedical domain: *NCBI Disease* (Doğan et al., 2014) and *BC5CDR* (Wei et al., 2016). These contain disease names and concept annotations from PubMed abstracts and articles- Both corpora are formatted for NER tasks with BIO tagging.

Dataset	Train	Val	Test
BC5CDR	5228	5330	5865
NCBI	5433	924	941
MIMIC	608	76	76

Table 2: Size and Distribution of datasets

Table 2 shows the number of sentences on all three corpora. It is worth noting that by incorporating *NCBI* and *BC5CDR* datasets, the available data increases by an order of magnitude, from 608 examples in the MIMIC dataset to a total training size of 11269.

4.2 Hyperparameters

Hyperp.	NCBI+BC5CDR	MIMIC
Batch size	8	8
Epochs	1	3
Learning rate	4.7e-5	5.5e-5
Weight decay	0.125	0.004

Table 3: Fine-tuning hyperparameters

For both stages of training, we rely on the pre-trained BERT base uncased model available through HuggingFace's Transformers library (Wolf

	Private Test			Public Test		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Base	0,265	0,299	0,247	0,329	0,223	0,238
Two-Step Fine-tuning (TSF)	0,289	0,294	0,257	0,304	**0,258**	0,252
TSF + post-processing	0,348	0,302	0,292	0,330	0,246	0,257
TFS + Data Augmentation 10%	0,357	0,304	0,296	0,333	0,246	0,258
TFS + Data Augmentation 25%	0,361	0,305	0,299	0,331	0,239	0,255
TFS + Data Augmentation 40%	0,337	0,291	0,280	0,329	0,248	**0,259**
TSF all data	0,351	**0,327**	0,304	0,316	0,252	0,253
TFS all data + post-processing	**0,393**	0,310	**0,316**	**0,344**	0,237	0,256

Table 1: Summarization results as Precision, Recall and F1 score. On the left, the private test set held out from the training corpus. On the right, the public test set from the shared task.

et al., 2020). As the model undergoes two fine-tuning steps, the parameters differ between the first step, which uses the NCBI and BC5CDR datasets for biomedical NER, and the second step, which trains with the MIMIC dataset.

Table 3 shows the hyperparameters used for both finetuning steps. Due to the small size of the MIMIC corpus, both the learning rate and the number of epochs were increased, while weight decay was decreased. All other hyperparameters are left with the standard values provided by the library.

5 Results

This section will discuss the main results from the different experiments performed during this shared task. Table 1 overviews the best-performing models. Results are computed over private text extracted from the provided MIMIC training data and the public test set used on the CodaLab [2] competition. Metrics are reported using the ROUGE-L metric (Lin, 2004) as precision, recall, and F1 score, using the script provided by the organization.

We performed experiments on fine-tuning strategies. Results show that the two-step fine-tuning approach (TSF) outperforms fine-tuning on all available data by more than 1% on private and public test sets. This performance increment is even more significant when the post-processing step is applied. We observe a 4% F1 score improvement on the private test set, much larger than the 0,05% observed on the public test set. An explanation for this difference may be that the public test set is cleaner than the provided training data. These steps are applied to all our experiments.

[2] https://codalab.lisn.upsaclay.fr/competitions/12388results

Another critical factor in our experiments was data scarcity. To mitigate it, we evaluated different levels of data augmentation by randomly masking different percentages of the input tokens. We observe slight improvements when adding more data augmentation up to 40% of tokens. These results are consistent for both private and public test sets, being this the best model on the public test set and our final submission to the competition.

Finally, once we had the final hyperparameters for all models, we decided to run a final experiment, including the development data for the training to increase the amount of data available. Although this configuration performed best on our private test set, it yielded slightly worse results on the public test set compared to the original model that used less data.

Overall, these results showcase the importance of fine-tuning domain-specific data and using data augmentation to maximize the performance of our systems.

6 Conclusions

In this paper, we have presented the UPC submission to the BioNLP 2023 Problem List Summarization. Our results demonstrate that modeling this task as a NER problem and combining pre-train model fine-tuning and data augmentation is an effective approach to solving this task, even when limited training data is available. Experimental results show how these two techniques can outperform fine-tuning on additional out-of-domain data or adding small amounts of domain data during the fine-tuning process. In future work, we plan to explore leveraging additional non-annotated data from the biomedical domain.

Limitations

During the manual labeling process of the notes, we searched for keywords to select for the Named-Entity-Recognition and compared them to those appearing on the summary. For the system to work effectively, the list of problems and diagnoses should appear in its entirety in the original text columns. However, this was not the case. Most of the examples had summaries with additional diagnoses and problems with words that could not be extracted from the original text. In addition to these cases, having an inconsistent use of acronyms leads to different versions of the same term in the original text and the summary.

Finally, when comparing both lists of tokens, the items found in both segments had a distinct order of appearance, having examples score low for only counting one of them as a match in Rouge-L, where the list of words is the same but in a different order.

Ethics Statement

Any work on biomedical data presents a series of ethical considerations. First, the data employed includes sensible data from actual patients and, therefore, a possible breach of their privacy. Thorough corpus curation and deidentification are required to ensure that no information could be related to the patients. Secondly, decisions in the medical domain may have consequences for the patient's health. Automating any task in this domain without proper human supervision may lead to wrong treatments or diagnoses and the associated risks to those decisions.

In the particular case of problem list summarization, our proposed system only produces outputs already present in the provided human-generated data. Errors resulting from false diseases hallucinated by the model can be discarded. However, errors due to missing or partially annotated entities are likely, especially in cases with the provided notes are not entirely consistent. For this reason, this system should be employed under the supervision of a human expert in the medical field and not as a standalone automatization tool.

Acknowledgements

This work was funded by Spanish State Research Agency (AEI) project PID2019-107579RB-I00 (AEI/10.13039/501100011033) and the "European Union NextGenerationEU/PRTR" under the project ROB-IN (PLEC2021-007859)

References

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 172–175.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Wayne Hubbard, and Lawrence Jackel. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.