Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3601–3609
Marseille, 20-25 June 2022

A Named Entity Recognition Corpus for Vietnamese Biomedical Texts to Support Tuberculosis Treatment

Uyen T.P. Phan[1,2], **Phuong N.V. Nguyen**[3], **Nhung T.H. Nguyen**[4]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[2] Vietnam National University, Ho Chi Minh city, Vietnam
[3]Pham Ngoc Thach University of Medicine, Vietnam
[4]Department of Computer Science, University of Manchester, UK
ptpuyen@fit.hcmus.edu.vn, nnvanphuong@pnt.edu.vn, nhung.nguyen@manchester.ac.uk

Abstract

Named Entity Recognition (NER) is an important task in information extraction. However, due to the lack of labelled corpora, biomedical NER has scarcely been studied in Vietnamese compared to English. To address this situation, we have constructed *VietBioNER*, a labelled NER corpus of Vietnamese academic biomedical text. The corpus focuses specifically on supporting tuberculosis treatment, and was constructed by collecting scientific papers and grey literature related to tuberculosis symptoms and diagnostics. We manually annotated a small set of the collected documents with five categories of named entities: Organisation, Location, Date and Time, Symptom and Disease, and Diagnostic Procedure. Inter-annotator agreement ranges from 70.59% and 95.89% F-score according to entity category. In this paper, we make available two splits of the corpus, corresponding to traditional supervised learning and few-shot learning settings. We also provide baseline results for both of these settings, in addition to a dictionary-based approach, as a means to stimulate further research into Vietnamese biomedical NER. Although supervised methods produce results that are far superior to the other two approaches, the fact that even one-shot learning can outperform the dictionary-based method provides evidence that further research into few-shot learning on this text type would be worthwhile.

Keywords: Named Entity Recognition, Vietnamese Biomedical Texts, Corpus Annotation

1. Introduction

According to WHO, prior to the coronavirus (COVID-19) pandemic, tuberculosis (TB) was the leading cause of death from a single infectious agent, ranking above HIV/AIDS (WHO, 2021). Meanwhile, Nguyen et al. (2020) reported that TB burden in Vietnam still remains high, despite government efforts to eradicate the disease. A system to support tuberculosis treatment for the Vietnamese language could therefore be extremely valuable in helping to improve the situation. Ideally, the system should be able to provide answers to the following questions:

- In which geographical areas is TB present?
- When have cases of TB been reported?
- Which methods are used to diagnose TB?
- Which symptoms/diseases are relevant to TB?
- Which organisations suggest/authorise a diagnostic procedure?

A key element in such a system would be a named entity recognition (NER) tool, which could automatically identify mentions of the important entities (e.g. locations, diseases and dates/times) that are needed to answer the above questions. Unfortunately, such a tool does not currently exist for the Vietnamese language, and developing one is very problematic, due to the lack of supporting resources. Approaches to NER can be classified into three main groups: rule-based, dictionary-based and machine learning-based. Rule-based systems can be developed without resources, but require a large amount of manual effort to construct, and tend to suffer from issues of low recall. Meanwhile, dictionary-based systems are reliant of the availability of wide coverage dictionaries; otherwise, they will fail when unseen entities are encountered. Machine-learning based systems can achieve higher levels of performance than the other two types of systems, and are able to recognise previously unseen entities, but are reliant on the availability of manually annotated training corpora.

There are many available annotated corpora for biomedical NER in English, e.g., NCBI Disease (Dogan et al., 2014), BioCreative V CDR (Li et al., 2016), i2b2 datasets (Uzuner et al., 2011; Sun et al., 2013; Stubbs and Uzuner, 2015), and CADEC (Karimi et al., 2015). These corpora have helped to shape state-of-the-art (SOTA) systems in NER for English biomedical text. However, no similar corpus exists for the Vietnamese language.

Supervised machine learning approaches are suitable when large amounts of training data are available, which is the case for most of the corpora cited above. However, few-shot learning (Hofer et al., 2018; Yang and Katiyar, 2020; Huang et al., 2020) may be applied when only a small amount of annotated data is available. This latter approach is considered more practical than the first one, especially in the case of low-resourced languages/domains, given that the production of large annotated corpora can be expensive and

time-consuming.

According to the importance of annotated corpora for developing practical NER tools, this paper introduces a novel Vietnamese NER corpus for biomedical texts, called *VietBioNER*. The corpus was created by collecting and digitising scientific papers and theses related to tuberculosis. We then randomly selected a set of 1706 sentences and manually annotated five categories of entities: Organisation, Location, Date and Time (DateTime), Symptom and Disease (Symtom_and_Disease), and Diagnostic Procedure (DiagnosticProcedure). The average Inter-Annotator Agreement (IAA) rate between the two annotators is 80.69%, indicating that the annotations are reliable and thus suitable for training NER models.

We additionally report the baseline performance of dictionary-based, supervised learning and few-shot learning approaches. For the dictionary-based approach, we compiled a list of dictionaries for locations, organisations, diagnostic procedures, symptoms and diseases. For supervised learning approaches, we employed Bi-LSTM (Lample et al., 2016), BERT (Devlin et al., 2019), and PhoBERT (Nguyen and Nguyen, 2020). For few-shot learning, we used NNShot and StructShot (Yang and Katiyar, 2020).

Our experimental results show that the best performance is obtained using supervised learning approaches, followed by few-shot learning and dictionary-based ones. Among the supervised learning approaches, PhoBERT was able to produce the best results, thanks to more accurate word embedding representations than those used in the BiLSTM and BERT approaches,

To the best of our knowledge, VietBioNER is the first publicly available Vietnamese corpus for NER that focuses on academic text in the biomedical domain. It is our hope that the corpus and benchmarks will act as a stimulus for further research into Vietnamese biomedical NER systems. The corpus and the experimental results will be made available at https://github.com/ptpuyen1511/VietBioNER.

2. Related Work

The majority of currently available annotated biomedical corpora are in English. Among them, NCBI Disease, i2b2/n2c2, and CADEC (CSIRO Adverse Drug Event Corpus) have been commonly used to build state-of-the-art (SOTA) systems.

NCBI Disease (Dogan et al., 2014) consists of 793 PubMed abstracts, annotated with 6892 disease mentions belonging to 790 unique disease concepts.

i2b2/n2c2 are a series of Shared Task Challenges for Clinical Data NLP, with different tasks being set each year. The NER challenges i2b2-2010 (Uzuner et al., 2011) and i2b2-2012 (Sun et al., 2013) both included de-identified discharge summaries provided by Partners HealthCare and MIMIC-II. The i2b2-2010 dataset consists of discharge summaries, annotated with three entity types: Medical problems, Treatments, and Tests. Meanwhile, the i2b2-2012 dataset has 310 summaries with two types of annotations: Clinically relevant events and Temporal relations. The i2b2-2014 (Stubbs and Uzuner, 2015) includes 1304 longitudinal medical records annotated with seven entity types: Name, Profession, Location, Age, Date, Contact, and IDs. The IAA of i2b2-2014 is approximately 89% for entity-based evaluation and 93% for token-based evaluation. The n2c2-2018 (Henry et al., 2019) consists of 505 de-identified discharge summaries drawn from the MIMIC-III clinical care database (Johnson et al., 2016). The n2c2-2018 was annotated with Adverse Drug Events (ADEs) and has nine annotation types: Drug, Strength, Form, Dosage, Frequency, Route, Duration, Reason, and ADE.

CADEC (Karimi et al., 2015) consists of 12533 posts from a medical forum called AskaPatient. The CADEC dataset has 9111 annotations with five entity types: ADR, Disease, Drug, Symptom, and Finding. The dataset was shown to be reliable, with an IAA ranges from 78% to 95%.

PhoNER_COVID19 (Truong et al., 2021) is a recently released annotated corpus in the Vietnamese language that includes biomedical entities. The corpus consists of 34984 entities over 10027 sentences. The corpus is constituted by articles from popular Vietnamese online news sites. Ten entity types including Patient ID, Person Name, Age, Gender, Occupation, Location, Organisation, Transportation, Date, and Symptom & Disease were annotated in the corpus.

In contrast to PhoNER_COVID19, VietBioNER consists of academic-related text, including both scientific and grey literature. Given that machine learning based tools are sensitive to the type of text on which they are trained, VietBioNER constitutes an important resource that can complement PhoNER_COVID19 to facilitate the recognition of biomedical entities in a range of different text types.

3. Corpus Annotation

3.1. Document Sources

We manually selected 220 documents, consisting of both scientific articles and theses related to tuberculosis. The scientific articles were collected from 5 medical journals: the Medical Journal of Ho Chi Minh City, the Journal of Practical Medicine, the National Journal of Tuberculosis and Lung Disease, the Can Tho University Journal of Medicine and Pharmacy, and the Vietnam Medical Journal. The collected theses were written by resident doctors, speciality doctors, masters, associate doctorates and doctorates from different medical universities. The number of collected documents from different sources is reported in Table 1.

It should be noted that most of the collected documents took the form of hard copies. We therefore scanned them and applied VietOCR[1] in order to digitise them.

[1] http://vietocr.sourceforge.net/

Table 1: The number of collected documents from different sources in VietBioNER.

Document Source	#Doc
Medical Journal of Ho Chi Minh City	40
Journal of Practical Medicine	34
National Journal of Tuberculosis and Lung Disease	26
Can Tho University Journal of Medicine and Pharmacy	6
Vietnam Medical Journal	4
Theses	110

Table 2: Agreement between annotators for each entity type (F-score).

Entity Type	IAA (%)
DiagnosticProcedure	70.59
DateTime	76.19
Symptom_and_Disease	81.96
Organisation	95.00
Location	95.89
All	80.69

3.2. Manual Annotation

One of our main aims in creating VietBioNER is to facilitate the automatic extraction of information that can support domain experts in carrying out TB treatment in Vietnam. According to the types of information that are important for this activity, we focus on five categories of entities: Organisation, Location, Date and Time (DateTime), Symptom and Disease (Symptom_and_Disease), and Diagnostic Procedure (DiagnosticProcedure).
Below, we provide a more detailed description of each of these categories[2].

3.2.1. Organisation

Names of specific organisations are tagged in this category. These include names of government or non-government organisations, and names of offices or unions, such as, "Bộ Y tế" (Ministry of Health), "Khoa Phổi Thận" (Lung and Kidney Department), and "Bệnh viện Nhân dân Gia Định" (Gia Dinh People's Hospital).

3.2.2. Location

Mentions of geographic locations, i.e., any identifiable point or area in the planet, ranging from continents, major bodies of water (e.g., oceans, rivers, lakes), named landforms, countries, states, cities and towns, are marked up as Location entities. This entity type includes instances of proper names and their abbreviations, except when used in the context as a political entity. Some examples for Location entities are "phía Bắc Việt Nam" (North Vietnam), "quận Tân Bình" (Tan Binh district), and "Khu vực Đông Nam Á" (Southeast Asia).

3.2.3. DateTime

Date or specific periods of time (having a specific beginning and ending) are marked in this category. Some examples of this entity type are: seasons of the year; dates, months and years; specific time, etc. For examples "mùa hè" (summer), "tháng Ba" (March), "từ 1990-1992" (from 1990-1992), and "25 - 26/10".

3.2.4. Symptom and Disease

We include all diseases, illness, inflammations and disorders/conditions occurring in humans. We also annotated words/phrases that describe signs of the disease process, rather than the disease/illness itself. This includes alterred physical appearance or behaviours. Examples include "lao đa kháng thuốc" (multidrug-resistant tuberculosis), "ung thư" (cancer), "đái tháo đường" (diabetes), "ho có đờm" (coughing up phlegm), "đau ngực" (chest pain) , and "khó thở" (shortness of breath).

3.2.5. DiagnosticProcedure

We label procedures, methods and techniques used to determine the composition, quality or concentration of a specimen, which are carried out in a clinical laboratory, for examples, "sinh thiết màng phổi bằng kim" (pleural needle biopsy), and "GeneXpert".

After finalising the guidelines, we randomly selected 63 segments, consisting of the following sections from the collected documents: summary, introduction, problem statement and objective. The selected segments were then passed to two annotators, who are senior students of the Faculty of Public Health, Pham Ngoc Thach University of Medicine. Both annotators have medical knowledge and were supported by detailed and formal annotation guidelines. Annotation was carried out using brat–a web-based annotation tool[3]. The tool allows annotators to easily create new annotations, as well as modify or delete existing annotations.

4. Corpus Statistics

To calculate the Inter-Annotator Agreement (IAA), we randomly selected 7 segments for double annotation, i.e., the two annotators labelled the same documents without discussing about their choices. The agreement between annotators is reported in Table 2, in terms of F-score.
As shown in the table, the entity type with the lowest IAA is DiagnosticProcedure. This can be a challenging type of annotate, due to difficulties in determining appropriate boundaries. For example, given the entity "nhuộm soi AFB mô màng phổi" (staining for AFB in pleural tissue), one annotator tagged the whole span as DiagnosticProcedure, while the other tagged a shorter

[2]We will publish the full guidelines with our corpus.

[3]https://brat.nlplab.org/

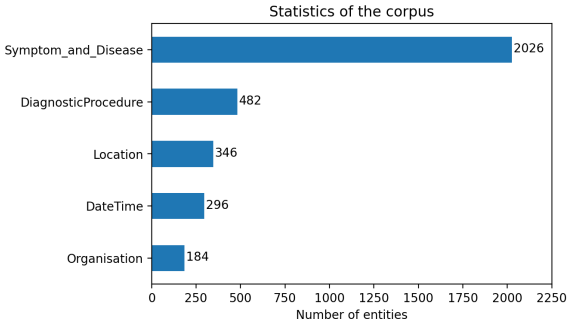

Figure 1: Distribution of each entity type in VietBioNER.

span of "nhuộm soi AFB" (staining for AFB). We observed similar difficulties in the case of DateTime entities. Nevertheless, an average IAA of 80.69% F-score was achieved over all annotation categories, which indicates that the quality of the corpus annotation is sufficiently reliable.

The final annotated corpus contains 1706 sentences with an average length of 31 tokens. About 74% of sentences in the corpus contained annotated entities. The corpus is annotated with 3334 entities, whose distribution among the 5 categories is detailed in Fig. 1. It can be observed that Symptom/Disease entities are by far the most dominant entity type in the corpus, followed by Diagnostic Procedure and Location.

4.1. Data Visualisation

The BERT-based models that we use for our supervised approaches are reliant on accurate word embeddings that can accurately represent and distinguish the meanings of different entities. A good set of word embeddings would therefore be expected to assign similar representations to entities belonging to the same semantic category. Our experiments use two different language models for word embeddings and, in order to provide an indication of the effectiveness of these models in producing embeddings for our corpus, we visualise the entity embeddings for the entire corpus using two language models, i.e., Multilingual BERT (Devlin et al., 2019) and PhoBERT (Nguyen and Nguyen, 2020)[4] to extract word embeddings. We then calculated entity embeddings by taking average of word embeddings in an entity. We finally employed t-SNE (van der Maaten and Hinton, 2008) to reduce the dimensionality of the embeddings, and plotted them in a 2D space. The resulting visualisations as displayed in Fig. 2.

From Fig. 2, we can see that for both models, the embeddings of entities belonging to the same category are roughly clustered together. However with Multilingual BERT, entity embeddings are sometimes intertwined each other while with PhoBERT they are clustered more neatly. This clustering behaviour is to be expected, since PhoBERT was specifically fine-tuned on Vietnamese texts, while Multilingual BERT was pretrained on 104 different languages.

[4]We used the default hyper-parameters provided by https://huggingface.co/models

Table 3: Number of entities in each benchmark setting. In the case of few-shot learning, we calculated the average and performed ceiling rounding over 5 different support sets.

Entity Type	Supervised Learning			Few-shot Learning (Average over 5 sets)		
	Train	Valid	Test	1-shot	5-shot	10-shot
Symptom_and_Disease	838	378	810	5	20	35
DiagnosticProcedure	191	89	202	3	8	13
Location	162	78	106	4	8	17
DateTime	141	58	97	2	7	13
Organisation	77	36	71	2	6	11

5. Benchmark Settings

We provide two sets of benchmark results for models trained on our corpus using two different settings. The first set of results is obtained using a standard supervised learning setting, which typically includes three disjoint subsets: training, validation, and testing. The second set of results is obtained using few-shot learning setting, in which only a small set of labelled instances is used for model training.

5.1. Standard Supervised Learning

We constructed the training, validation, and test sets with an approximate ratio of 7:3:7. Specifically, the training set consists of 706 sentences, the validation set consists of 300 sentences, and the test set has 700 sentences.

5.2. Few-shot Learning

Unlike traditional supervised learning, few-shot learning consists of two phases: meta training and inference (testing). The meta-training phase uses a larger or more richly labelled corpus, which is usually referred as the corpus in a source domain. Meanwhile, the inference phase uses the base model obtained from the meta-training phase on a specific corpus in the target domain (i.e., a corpus with limited labelled instances).

For inference, we build 1-shot, 5-shot, and 10-shot learning sets from the training set mentioned in Section 5.1. For each n-shot learning set, we generated 5 random support sets using the Greedy Sampling algorithm (Yang and Katiyar, 2020). n denotes the number of entities in each category that are selected for inclusion in each support set. Consequently, each n-shot support set will have a maximum of $(n \times num\_entity\_categories)$ sentences. We use the same test set as in the supervised setting.

The distributions of all entity types in the two different settings are reported in Table 3.

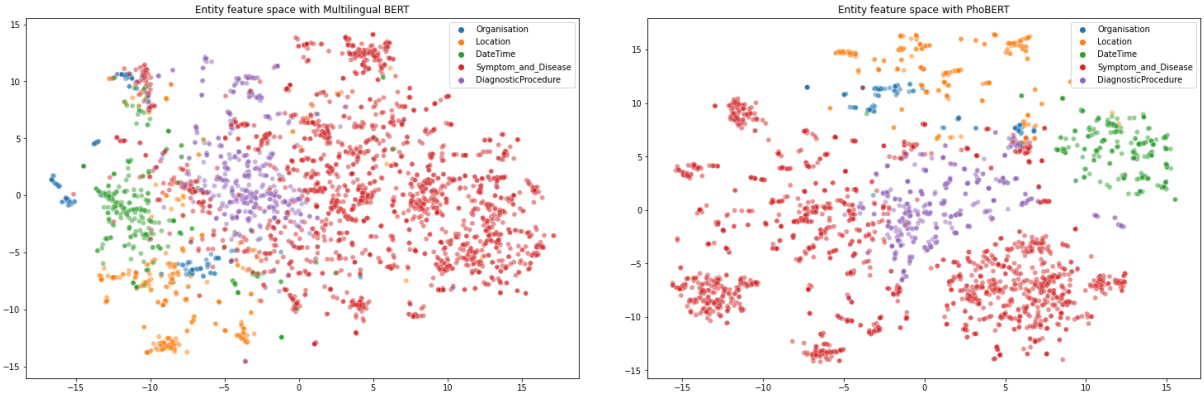

Figure 2: Visualisations of entity embeddings. The left figure shows entity embeddings obtained from Multilingual BERT while the right figure shows entity embeddings obtained from PhoBERT.

6. Experiments

6.1. Experimental Methods

We carried out a number of different experiments using VietBioNER, corresponding to the two benchmark learning settings, in addition to a simple dictionary-based method. For the supervised learning experiments, we employed Bi-LSTM (Lample et al., 2016) and BERT (Devlin et al., 2019). For few-shot learning, we used NNShot and StructShot (Yang and Katiyar, 2020).

6.1.1. Dictionary-based method

We used simple left-right maximum matching (LRMM) to match entities against a dictionary. A different dictionary was compiled for each entity category. For Location and Organisation, the dictionaries were automatically constructed by crawling data on Wikipedia, while the dictionary for Diagnostic Procedure entities was manually compiled by medical experts. For Symptom_and_Disease entities, the dictionary was constructed using a combination of automatic and manual methods. Finally, the validity of DateTime entities was checked using regular expressions.

It should be noted that the training and validation parts of the corpus were not used in this experiment; only the test set was used, in order to validate the performance of the dictionaries in recognising relevant entities.

6.1.2. Supervised learning

We applied two popular supervised learning-based approaches to our corpus as follows.

- *Bi-LSTM:* We employed the Bi-LSTM NER by Lample et al. (2016), one of the SOTA approaches prior to the emergence of BERT.

- *BERT:* We used BERT-based NER model with two pre-trained language models, Multilingual BERT (Devlin et al., 2019) and PhoBERT (Nguyen and Nguyen, 2020) with the default hyper-parameters provided by Hugging Face[5].

In this experiment, we used the standard supervised learning benchmark setting, i.e., the training, validation and test sets.

6.1.3. Few-shot learning

Among existing approaches to few-shot learning for NER, we experimented with NNShot and StructShot (Yang and Katiyar, 2020).

For the meta training phase, we used PhoNER_COVID19 (Truong et al., 2021) as a source domain to generate the base model for the NER task. In the inference phase, we used the support sets in the few-shot learning benchmark to predict the label of instances in the test set, i.e., the target domain.

6.2. Experimental Results

It is noted that in our experiments, we used the IO scheme instead of the common BIO one since the IO scheme was reported to produce better results (Yang and Katiyar, 2020). All experimental results obtained on the test set are reported in Table 4.

It can be observed that the supervised learning approach produced the best results among the three experimental approaches. Better performance using this setting is to be expected, compared to few-shot learning, because the supervised models were trained using a larger amount of labelled data.

Although the dictionary-based method used domain-specific dictionaries, it is very difficult to ensure that sure dictionaries provide comprehensive coverage of concepts in the domain, at least without considerable time and effort; many domain-specific dictionaries for English have been under constant development for many years. Perhaps unsurprisingly, therefore, many of the entities in the test set were not present in the dictionaries, and this method achieved the lowest recall in comparison to the other methods, although precision was somewhat higher than recall. The low performance is likely to be caused by the many different ways

[5]https://huggingface.co/models

Table 4: Results of applying different NER methods to the test set of VietBioNER. In the case of few-shot learning, we report the average and standard deviation of scores from 5 different support sets.

Method		P (%)	R (%)	F_1 (%)
Dictionary-based Method				
LRMM		51.24	17.73	26.34
Few-shot Learning Methods				
1-shot	NNShot	27.37 ± 5.6	41.44 ± 10.9	32.31 ± 5.9
	StructShot	31.46 ± 5.5	39.72 ± 10.8	34.61 ± 6.5
5-shot	NNShot	28.96 ± 3.1	44.53 ± 5.9	35.00 ± 3.5
	StructShot	31.75 ± 3.8	39.38 ± 3.7	34.89 ± 1.6
10-shot	NNShot	30.32 ± 1.7	49.43 ± 3.2	37.57 ± 2.1
	StructShot	32.17 ± 2.8	43.44 ± 1.3	36.89 ± 1.9
Supervised Learning Methods				
Bi-LSTM		79.00	77.84	78.42
Multilingual BERT		75.43	80.74	77.99
PhoBERT		**77.49**	**81.83**	**79.60**

in which entity names such as diseases, symptoms and diagnostic procedures can be described in text; it is virtually impossible to ensure that all such variations are comprehensively covered by dictionaries.

In comparison to the dictionary-based method, the few-shot learning methods achieved significantly higher recall. However, the trade-off is that precision is significantly lower. In terms of F-score, however, even the use of 1-shot learning produces better results than the dictionary-based method. This result illustrates that, even with only a very small amount of training data, machine leaning methods can outperform the dictionary-based method, whose dictionaries required a large amount of manual effort to create.

When we increased the number of training instances in the few-shot learning experiments to 5 and 10, performance is generally increased, with a lower standard deviation than the results obtained using 1-shot, indicating that performance becomes more stable when a small number of additional training instances are used. However, the performance gap between few-shot and supervised learning is still large, which illustrates the need for further research into few-shot learning, especially for a low-resourced language like Vietnamese.

In terms of the supervised learning approaches, the PhoBERT-based method achieved the best performance. This result can be explained by the fact that the PhoBERT model was pre-trained on a large amount of Vietnamese data, hence producing better word representations, as previously discussed and illustrated in Fig 2.

To investigate the PhoBERT results in more detail, we report the performance of the model on the validation set for each entity category in Table 5. The table shows that the model performed well in recognising Symptom_and_Disease, Location and DateTime entities. Meanwhile, performance for DiagnosticProcedure entities is noticeably lower than for other categories. We hypothesise that a probable reason for these

Table 5: Performance for each named entity category using PhoBERT on the validation set.

Entity Type	P (%)	R (%)	F_1 (%)
DiagnosticProcedure	50.46	61.80	55.56
Organisation	69.44	69.44	69.44
DateTime	70.69	70.69	70.69
Location	76.47	83.33	79.75
Symptom_and_Disease	81.33	80.69	81.01

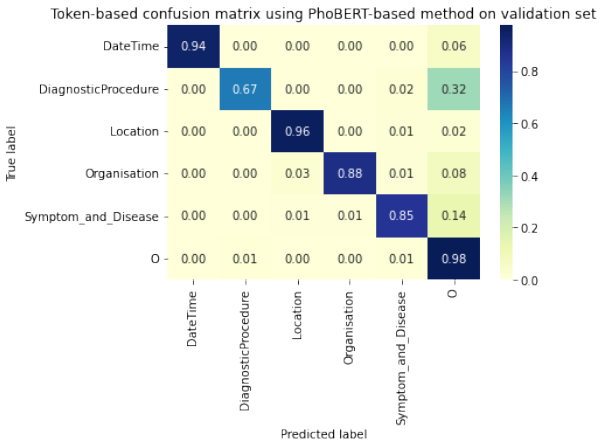

Figure 3: Token-based confusion matrix on the predictions of the validation set by PhoBERT.

results is that the vocabulary used to describe Symptom_and_Disease, Location and DateTime entities is also found in the corpus on which PhoBERT was fine-tuned. In contrast, this is not the case for the DiagnosticProcedure category, whose entities generally constitute highly specific tuberculosis diagnostic procedures. Since PhoBERT is a language model for the general domain, it cannot provide representations for all such entities.

To further analyse the results, we depict a token-based confusion matrix on the predictions of the validation set in Fig. 3. From the confusion matrix, it can be noted that many of the tokens that should have been labelled as DiagnosticProcedure tokens were incorrectly predicted as non-entity tokens, i.e., tag 'O' in the matrix (the confusion is 32%). As explained above, since PhoBERT's vocabulary does not include all of the words used in this type of entity, such type of confusion in understandable. As follow-up work, we plan to develop methods to improve the performance of detecting DiagnosticProcedure entities.

7. Conclusion

In this paper, we have described the construction of VietBioNER, a novel Vietnamese NER corpus for the biomedical domain. The corpus was annotated with 5 entity types: Symptom_and_Disease, Diagnostic Pro-

cedure, Location, Organisation, and Date Time. Although the corpus was constructed to support TB treatment, we believe that models trained on the corpus would be capable of recognising information about other diseases in the biomedical and clinical domains.

In addition to constructing the data set and creating two different benchmark settings, we also reported performance of baseline systems using three different approaches to NER. Experimental results showed that the supervised learning method performs better than dictionary-based and few-shot learning methods. Specifically, the model using BERT-base model gives better results than Bi-LSTM. While the results achieved using few-shot learning method were lower than those obtained using supervised learning, the former method could still outperform the dictionary-based baseline. Since few-shot learning is more practical for real life applications, as it requires very little annotated data, our initial baseline results provide motivation to further investigate this type of learning approach.

Although the size of the corpus is relatively small compared with other corpora in English, VietBioNER still constitutes an important resource, given that, to our knowledge, it is the first corpus of Vietnamese academic text that has been annotated with biomedical entities. As such, it may be used in a variety of ways, e.g. as a benchmark resource for use by the biomedical text mining community or to develop downstream applications such as semantic search engines and text classification tools.

In the future, we plan to label additional entity types such as Drug, Laboratory Procedure, Therapeutic or Preventive Procedure, as well relations between entities. We also plan to investigate whether the application of other few-shot learning methods using our corpus could result in improved model performance.

8. Acknowledgements

We would like to thank Paul Thompson and the anonymous reviewers for their useful comments. In addition, we would like to acknowledge the annotators from the Faculty of Public Health, Pham Ngoc Thach University of Medicine. This research was partially funded by the University of Science, VNU-HCM, Vietnam under grant number CNTT2020-04.

9. Bibliographical References

Al-Hegami, A. S., Othman, A. M. F., and Bagash, F. (2017). A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set.

Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., and Dai, L. (2021). ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. *Complexity*, 2021:1–6, 03.

Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., and Xu, H. (2015). A Study of Active Learning Methods for Named Entity Recognition in Clinical Text. *J. of Biomedical Informatics*, 58(C):11–18, December.

Chen, D., Che, N., Le, J., and Pan, Q. (2019). A co-training based entity recognition approach for cross-disease clinical documents. *Concurr. Comput. Pract. Exp.*, 31(23).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *CoRR*, abs/1703.03400.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 07.

Hakala, K. and Pyysalo, S. (2019). Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China, November. Association for Computational Linguistics.

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, O. (2019). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 10.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November.

Hofer, M., Kormilitzin, A., Goldberg, P., and Nevado-Holgado, A. J. (2018). Few-shot Learning for Named Entity Recognition in Medical Text. *CoRR*, abs/1811.05468.

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., and Han, J. (2020). Few-Shot Named Entity Recognition: A Comprehensive Study. *CoRR*, abs/2012.14978.

Ju, Z., Wang, J., and Zhu, F. (2011). Named Entity Recognition from Biomedical Text Using SVM. In *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese Neural Networks for One-shot Image Recognition. In *ICML deep learning workshop*, volume 2.

Lample, G., Ballesteros, M., Subramanian, S.,

Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep.

Munkhdalai, T. and Yu, H. (2017). Meta Networks. *CoRR*, abs/1703.00837.

Nguyen, D. Q. and Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. *CoRR*, abs/2003.00744.

Nguyen, H. T. M., Ngo, Q. T., Vu, L. X., Tran, V., and Nguyen, H. T. (2019). VLSP Shared Task: Named Entity Recognition. *Journal of Computer Science and Cybernetics*, 34:283–294.

Nguyen, H. V., Tiemersma, E. W., Nguyen, H. B., Cobelens, F. G. J., Finlay, A., Glaziou, P., Dao, C. H., Mirtskhulava, V., Nguyen, H. V., Pham, H. T. T., Khieu, N. T. T., de Haas, P., Do, N. H., Nguyen, P. D., Cung, C. V., and Nguyen, N. V. (2020). The second national tuberculosis prevalence survey in vietnam. *PLOS ONE*, 15(4):1–15, 04.

Nichol, A., Achiam, J., and Schulman, J. (2018). On First-Order Meta-Learning Algorithms. *CoRR*, abs/1803.02999.

Quimbaya, A. P., Múnera, A. S., Rivera, R. A. G., Rodríguez, J. C. D., Velandia, O. M. M., Peña, A. A. G., and Labbé, C. (2016). Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach. *Procedia Computer Science*, 100:55–61.

Ravi, S. and Larochelle, H. (2017). Optimization as a Model for Few-Shot Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical Networks for Few-shot Learning. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Stubbs, A. and Uzuner, O. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics*, 58S, 08.

Sun, W., Rumshisky, A., and Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20, 04.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). Learning to Compare: Relation Network for Few-Shot Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Uzuner, O., South, B., Shen, S., and DuVall, S. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6, 06.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. (2016). Matching Networks for One Shot Learning. In D. Lee, et al., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., Langlotz, C., and Han, J. (2018). Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752, 10.

WHO. (2021). Global tuberculosis report 2021, October.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Yang, Y. and Katiyar, A. (2020). Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

10. Language Resource References

Rezarta Islamaj Dogan and Robert Leaman and Zhiyong Lu. (2014). *NCBI Disease Corpus: A Resource for Disease Name Recognition and Normalization*.

Henry, Sam and Buchan, Kevin and Filannino, Michele and Stubbs, Amber and Uzuner, Ozlem. (2019). *2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records*.

Johnson, Alistair and Pollard, Tom and Shen, Lu and Lehman, Li-wei and Feng, Mengling and Ghassemi, Mohammad and Moody, Benjamin and Szolovits, Peter and Celi, Leo and Mark, Roger. (2016). *MIMIC-III, a freely accessible critical care database*.

Karimi, Sarvnaz and Metke-Jimenez, Alejandro and Kemp, Madonna and Wang, Chen. (2015). *CADEC: A corpus of adverse drug event annotations*.

Li, Jiao and Sun, Yueping and Johnson, Robin J. and Sciaky, Daniela and Wei, Chih-Hsuan and Leaman, Robert and Davis, Allan Peter and Mattingly, Carolyn J. and Wiegers, Thomas C. and Lu, Zhiyong. (2016). *BioCreative V CDR task corpus: a resource for chemical disease relation extraction*.

Stubbs, Amber and Uzuner, Ozlem. (2015). *Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus*.

Sun, Weiyi and Rumshisky, Anna and Uzuner, Ozlem. (2013). *Evaluating temporal relations in clinical text: 2012 i2b2 Challenge*.

Thinh Hung Truong and Mai Hoang Dao and Dat Quoc Nguyen. (2021). *COVID-19 Named Entity Recognition for Vietnamese*.

Uzuner, Ozlem and South, Brett and Shen, Shuying and DuVall, Scott. (2011). *2010 i2B2/VA challenge on concepts, assertions, and relations in clinical text*.