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

Uyen T.P. Phan[1,2], Phuong N.V. Nguyen[3] and 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, nvanphuong@pnt.edu.vn, nhung.nguyen@manchester.ac.uk

Introduction

- Tuberculosis (TB) is one of the leading causes of death from a single infectious agent.
- TB burden in Vietnam remains high → A system to support TB treatment in the Vietnamese language can be valuable in helping to improve the situation.
- There is no publicly available biomedical named entity recognition (NER) tool and corpus in Vietnamese.
- This paper introduces a novel Vietnamese NER corpus for biomedical texts, called VietBioNER.

Corpus Annotation

Document Sources

- Manually selecting 220 documents consist of both scientific articles and theses related to TB.
- Most of collected documents are hard copies → we scanned and used VietOCR[1] to digitise them.

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

Manual Annotation

- 5 entity categories: Organisation, Location, DateTime, Symptom & Disease, Diagnostic Procedure.
- Randomly selecting 63 segments → 2 annotators—Pham Ngoc Thach University of Medicine senior students.
- Using brat—a web-based annotation tool.

Corpus Statistics

Inter-Annotator Agreement (IAA)

- Randomly selecting 7 segments for double annotation.
- The agreement between annotators is reported in Table 1.

Table 1. Agreement between annotators for each entity category (F-score).

Entity Category	IAA (%)
DiagnosticProcedure	70.59
DateTime	76.19
Symptom&Disease	81.96
Organisation	95.00
Location	95.89
All	80.69

The corpus annotation is sufficiently reliable

Statistics

- The final annotated corpus contains 1706 sentences with an average length of 31 tokens.
- 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.

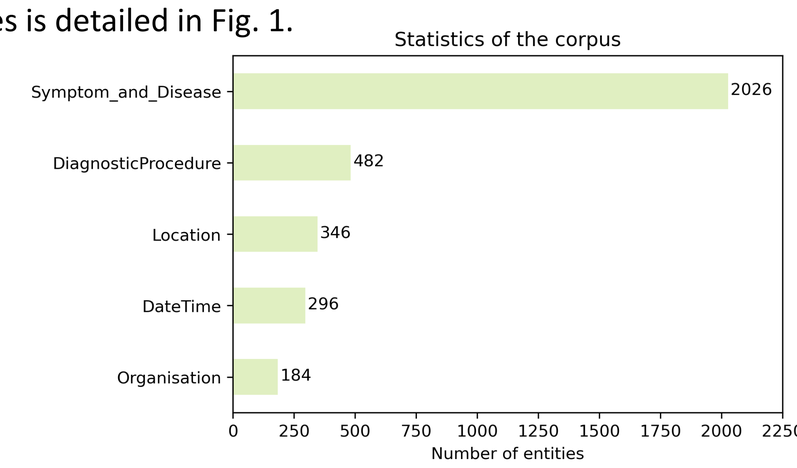

Figure 1. Distribution of each entity type in VietBioNER.

Experiments

Benchmark Settings

- Standard supervised learning: the training set: 706 sentences, the validation set: 300 sentences, and the test set: 700 sentences.
- Few-shot learning: we use the Greedy Sampling algorithm [1] → 1-shot, 5-shot, and 10-shot support sets.

Experimental Methods

- Dictionary-based method: we used simple left-right maximum matching (LRMM) to match entities against a dictionary.
- Supervised learning
  - Bi-LSTM: we employed the Bi-LSTM NER by Lample et al. [2].
  - BERT: we used BERT-based NER model with pre-trained Multilingual BERT [3] and PhoBERT [4].
- Few-shot learning: we experimented with NNShot and StructShot [1]. For the meta training phase, we used PhoNER_COVID19 [5] as a source domain.

Results

Table 2. 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		Precision (%)	Recall (%)	F1 (%)
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

Conclusion & Future Work

Conclusion

- We constructed VietBioNER, a novel Vietnamese NER corpus for the biomedical domain.
- We also reported performance of baseline systems using three different approaches to NER.

Future Work

- We plan to label additional entity categories such as Drug, Laboratory Procedure, Therapeutic or Preventive Procedure.

References

[1] Yang, Y. and Katiyar, A. 2020. “Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning”. In EMNLP.

[2] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. 2016. “Neural Architectures for Named Entity Recognition”. In NAACL.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In NAACL-HLT.

[4] Nguyen, D. Q. and Nguyen, A. T. 2020. “PhoBERT: Pre-trained language models for Vietnamese”. In Findings of the Association for Computational Linguistics: EMNLP 2020.

[5] Truong, T. H., Dao, M. H., and Nguyen, D. Q. 2021. “COVID-19 Named Entity Recognition for Vietnamese”. In NAACL.