BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali)

Salim Sazzed
Old Dominion University
Norfolk, VA, USA
ssazz001@odu.edu

Abstract

Recognizing biomedical entities in the text has significance in biomedical and health science research, as it benefits myriad downstream tasks, including entity linking, relation extraction, or entity resolution. While English and a few other widely used languages enjoy ample resources for automatic biomedical entity recognition, it is not the case for Bangla, a low-resource language. On that account, in this paper, we introduce BanglaBioMed, a Bangla biomedical named entity (NE) annotated dataset in standard IOB format, the first of its kind, consisting of over 12000 tokens annotated with the biomedical entities. The corpus is created by collecting Bangla text from a list of health articles and then annotated with four distinct types of entities: Anatomy (AN), Chemical and Drugs (CD), Disease and Symptom (DS), and Medical Procedure (MP). We provide the details of the entire data collection and annotation procedure and illustrate various statistics of the created corpus. Our developed corpus is a much-needed addition to the Bangla NLP resource that will facilitate biomedical NLP research in Bangla.

1 Introduction

The named-entity recognition (NER) frameworks aim to identify named entities (NE) mentioned in unstructured text documents and then categorize them into predefined domain-specific classes such as person or organization names, medical codes, chemical compounds, and food ingredients. Diverse sets of named entity (NE) annotated datasets have been created by researchers in varied domains such as clinical domain (Doğan et al., 2014), food domain (Stojanov et al., 2021), astronomy (Murphy et al., 2006), biological domain (Hastings et al., 2016). Due to the essence of NE for understanding biomedical concepts such as diseases, chemicals, and proteins, a number of studies focused on building NER corpora for the biomedical domain as it can aid researchers in finding relevant concepts and speed up the process of biomedical scientific discovery.

In English and a few other major languages, various NE corpora representing biomedical entities are publicly available (Kim et al., 2003; Kolárik et al., 2008). However, in Bangla, such a biomedical NE annotated dataset does not exist as NLP research in Bangla is still in infancy except in a few areas such as sentiment analysis (Bodini, 2022; Sazzed and Jayarathna, 2019; Faruque et al., 2021; Sazzed, 2020a; Bhowmick and Jana; Sazzed, 2020b), hate and abusive language detection (Karim et al., 2021; Sazzed, 2021a; Ishmam and Sharmin, 2019; Sazzed, 2021b). With the growing popularity of telemedicine and the availability of health and medical-related data written in Bangla, developing a Biomedical Named Entity Recognition (NER) system in Bangla is a pressing necessity.

For developing a sophisticated NER system, it is essential to have at least a moderate amount of annotated data. In particular, the generalizability and performances of the machine learning approaches (especially the deep learning-based models) heavily rely on the quantity of available annotated data. Hence, in this study, we introduce a biomedical NE dataset, the first of its kind, for the low-resource Bangla. The dataset is created by retrieving biomedical and health-related textual content from a number of health articles. The text data are then tokenized and annotated with four types of entities: Anatomy (AN), Chemical and Drugs (CD), Disease and Symptoms (DS), and Medical Procedure (MP). The final corpus contains around 2000 tokens representing one of the four types of biomedical NE mentioned above and around 10000 non-entity tokens.

1.1 Contributions

The main contributions of this study are:

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 323–329

- To address the lack of annotated data in the Bangla biomedical and health domain, we collect a biomedical corpus, BanglaBio, consisting of around 12000 tokens (i.e., primarily words).

- We manually annotate the corpus in token-level (mainly words) in four different classes of entities, Anatomy (AN), Diseases and Symptoms (DS), Chemical and Drug (CD), and Medical Procedure (MP).

- We provide the statistics of the frequency and structures of various types of entities present in the corpus and make the corpus publicly available for researchers [1].

2 Related Work

Although English and some other languages standardized entity annotated (i.e., IOB format) Biomedical corpora are available for the NER task, to the best of our knowledge, such resources do not exist in Bangla.

2.1 English Biomedical corpus

In English, a number of biomedical corpora exists with various types of entity annotations such as GENIA corpus (Kim et al., 2003), GENETAG corpus (Tanabe et al., 2005), SCAI IUPAC corpus (Kolárik et al., 2008), CellFinder corpus (Neves et al., 2012).

Pyysalo et al. (2007) presented BioInfer (Bio Information Extraction Resource), an annotated corpus of biomedical text consisting of 1100 sentences collected from abstracts of biomedical research articles. Kim et al. (2008) introduced single-facet annotation and semantic typing to the existing annotations in the GENIA corpus. The new annotation was performed on half of the GENIA corpus, consisting of 1,000 Medline abstracts. Giorgi and Bader (2020) introduced biomedical named entity recognition (BioNER) system for biomedical information extraction. To improve the generalizing ability of BioNER, the authors proposed an improved regularization technique using variational dropout, transfer learning, and multi-task learning.

Karimi et al. (2015) created CSIRO Adverse Drug Event Corpus (CADEC) consisting of patient-reported Adverse Drug Events (ADEs) collected from various medical forum posts. The authors performed multi-stage annotations for entities such as drugs, adverse effects, symptoms, and diseases. Scepanovic et al. (2020) proposed several approaches to accurately extract a wide variety of medical entities such as symptoms, diseases, and drug names collected from varied social media sources, and validated this approach on a large-scale Reddit dataset.

2.2 Non-English Biomedical corpus

For the French language, the Unified Medical Lexicon for French (UMLF) has been created by Zweigenbaum et al. (2005). For Swedish, an annotated gold standard corpus of medical records was developed by Velupillai (2012). Mowery et al. (2012) proposed a clinical uncertainty and negation taxonomy and mapped an English annotation schema to a Swedish schema.

Mitrofan and Tufiş (2018) presented a biomedical corpus in the Romanian language, which was collected in the contexts of the CoRoLa project, the reference corpus for the contemporary Romanian language. The authors described various statistics about the corpus and data-composition and annotation procedures. Carrino et al. (2021) introduced CoWeSe (the Corpus Web Salud Español), the largest Spanish biomedical corpus to date, consisting of around 750M tokens of clean plain text. The CoWeSe was created by crawling over 3000 Spanish documents.

Sun and Yang (2019) employed two language models, Multilingual BERT and BioBERT, to identify chemical and protein entities from the Spanish biomedical NER corpus PharmaCoNER (Gonzalez-Agirre et al., 2019). The author showed that transferring knowledge learned from large-scale source datasets to the target domain offers an effective solution for the PharmaCoNER task.

3 Creation of BanglaBioMed

3.1 Data Collection and Pre-processing

Unlike English, where a large number of scientific publications are available for extracting biomedical named entities, in Bangla, such resources do not exist, as researchers hardly publish scientific articles in Bangla. Hence, we use alternative sources for extracting biomedical text data. We leverage a set of health articles authored by medical physicians and published in the most popular

[1] https://github.com/sazzadcsedu/BanglaBioMed.git

Structure of Entity	Entity	Sentence
Simple Entity (single and multi-word)	জ্বর, কাশি, গলাব্যথা, শ্বাসকষ্ট	জ্বর, কাশি, গলাব্যথা, শ্বাসকষ্ট হচ্ছে অমিক্রনের মূল উপসর্গ।
	Fever, cough, sore throat, shortness of breath (*English Translation*)	Fever, cough, sore throat, shortness of breath are the main symptoms of Omicron. (*English Translation*)
Complex Entity (Overlapping)	নাক দিয়ে রক্ত আসা, নাক দিয়ে পানি আসা	নাক দিয়ে রক্ত বা পানি আসা
	Blood coming through the nose, Water coming through the nose (*English Translation*)	Blood or water coming through the nose (*English Translation*)

Figure 1: Examples of entities representing varied structures

Bangladeshi daily newspaper, *Prothom Alo* [2]. The health-related articles are chosen from the newspaper's official website. All the text data of the articles are manually excerpted for annotation. The excerpted texts are then segmented into sentences based on the '|' delimiter, which is equivalent to the English 'full stop(.)' delimiter. Afterward, each sentence is tokenized into words and punctuations.

3.2 Entity Types

Similar to Mitrofan (2017), the following four types of entities are considered in the annotation process.

- **Anatomy (AN)**: This entity label portrays the structure of the human body, especially as revealed by dissection and the separation of parts. This type of entity is common in health and medical text. Some examples include-

  মাথা (Head), হাত (Hand), পা (Leg), কোমর (Waist)

- **Chemicals and Drugs (CD)**: This entity label indicates the presence of chemical and drug-related terms in the tokens. Some examples are- ইনসুলিন (insulin), ফলিক অ্যাসিড (Folic Acid), ভিটামিন সি (Vitamin C), হাইড্রোকুইনোন (Hydroquinone)

- **Disease and Symptom (DS)**: This entity category includes names and descriptions of various diseases and symptoms (i.e., features appearing to the patients as conditions of the diseases). The following entities are some of the examples of this category- ক্যানসার (Cancer), হাঁপানি, প্রেশার, (Pressure) শ্বাসকষ্ট, (Shortness of breath) স্থূলতা , (Obesity)

- **Medical Procedures (MP)**: The entity of this group indicates laboratory procedures, the therapeutic or preventive procedures used for medical treatment. The followings are some examples- অস্থিমজ্জা প্রতিস্থাপন, (Bone marrow transplantation) কলোনস্কোপি (Colonoscopy), ব্লাড ট্রান্সফিউশন (Blood transfusion).

3.3 Entity Annotation Guidelines

We perform the entity annotation at the sentence level. Duplicate entities within a sentence or the corpus are annotated independently (all the occurrences of the same entity are labeled). We observe that most entities constitute single or multiple words without intervening with other entities (i.e., simple entities). Nevertheless, there exist entities that partially overlap with another; these types of entities can be referred to as complex entities (Examples shown in Figure 1).

Besides, we find that some entities are entirely embedded (nested) within another entity. Especially, the entities from the Anatomy (AN) class often are embedded into the Disease and Symptom (DS) category. To give an example, the DS entity *back pain* contains *back* entity from AN class. For this type of overlapping scenario, the longer entity is considered as the "top-level" entity, while its sub-part(s) is deemed as the "nested" entity. Most of the well-known NE annotated corpora employed the non-nested approach, where the words are annotated based on the top-level entity (Sang and De Meulder, 2003).

We do not consider co-referential or anaphoric references to the entity during annotation. The intensifier (e.g., slightly/severe) or possessive adjectives are not included in the entity to keep the annotation consistent across the corpus. The annotation is performed by an annotator who possesses a university-level education. The annotated label is

[2] https://www.prothomalo.com/life/health

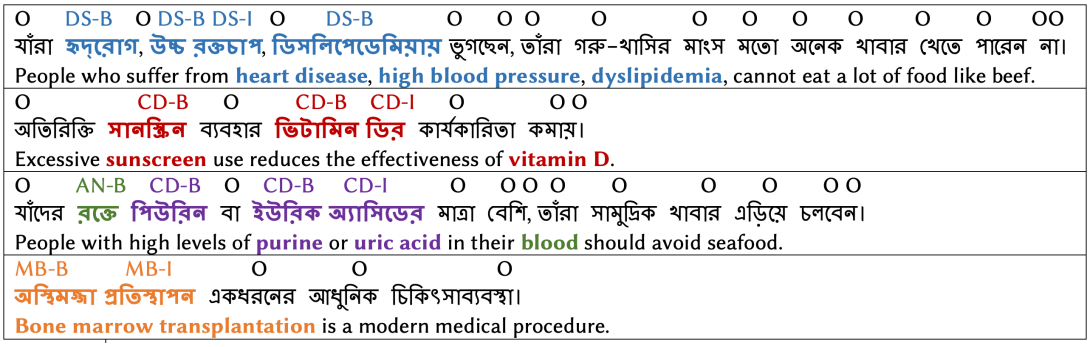

Figure 2: Examples of entity annotation within the sentences

further verified by a medical professional.

3.4 Entity Tagging

To make the annotated corpus suitable for the automatic NER task, we follow the standard IOB2 format (Tjong Kim Sang and Veenstra, 1999). The IOB2 format is described below,

B: The term 'B' indicates the beginning of a particular type of entity (i.e., the first token of an entity)

I: 'I' represents a token is a part of an already initiated entity,

O: 'O' indicates a token is not part of any entity of interest. All tokens outside the entity of interest are labeled as O.

3.5 Corpus Statistics

Table 1: The length distributions of unique entities of various types (in words)

Entity Type	Entity Length (# words)	Total
	1 / 2 / 3 / >= 4	
AN	105 /16 /0 /0	121
DS	190 /167 /64 /52	473
CD	79 /28 /1 /0	108
MP	41 /29 /3 /0	73

Table 1 shows the word length distributions of various types of entities. We find most of the entities contain a single word, while some comprising of two words. For example, the AN group contains close to 90% entities having a single word. The lengthy entities of over two words primarily belong to the DS category.

As shown in Table 2, the corpus has an unbalanced distribution regarding various types of entities. The most dominant entity type is DS,

Table 2: Statistics of various metrics in the annotated corpus

Metric	Count
#Tokens	11196
#Sentences	818
#Words with entity tag	1968
#Non-entity Words	9228
Average sentence length	13.68
Average number of entity per sentence	1.62
Entity Tag	**Count**
AN-B	259
AN-I	16
DS-B	699
DS-I	510
CD-B	102
CD-I	45
MP-B	269
MP-I	68
Total	1968

which is expected since these source articles contain more information related to various diseases and related symptoms. Among the 2000 biomedical entity annotated tokens present in the corpus, around 60% represent the DS category. The lowest presence is observed for the entities belonging to the CD category.

4 Summary and conclusion

In this study, we introduce a Bangla biomedical named entity annotated corpus created from a number of Bangla health articles. To the best of our knowledge, this is the first biomedical NE annotated corpus in Bangla (Bengali) in standard IOB format created for biomedical text mining. We report detailed annotation guidelines and procedures of the annotation. Moreover, we pro-

vide the various statistics of four different types of biomedical entities: AN, DS, CD, and MP, in the annotated corpus. We have made the corpus publicly available for the researchers. The future work will focus on enhancing the size of the annotated corpus and creating strong baselines for automatic NER tasks by incorporating transformer-based language models. Besides, we will investigate how to leverage cross-lingual resources from other languages, such as English, to improve the performance of the NER task.

References

Anirban Bhowmick and Abhik Jana. Sentiment analysis for bengali using transformer based models.

Matteo Bodini. 2022. Opinion mining from machine translated bangla reviews with stacked contractive auto-encoders. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *arXiv preprint arXiv:2109.07765*.

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.

M Abdullah Faruque, Saifur Rahman, Partha Chakraborty, Tanupriya Choudhury, Jung-Sup Um, and Thipendra Pal Singh. 2021. Ascertaining polarity of public opinions on bangladesh cricket using machine learning techniques. *Spatial Information Research*, pages 1–8.

John M Giorgi and Gary D Bader. 2020. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–D1219.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 555–560. IEEE.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9(1):1–25.

Corinna Kolárik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. 2008. Chemical names: terminological resources and corpora annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*.

Maria Mitrofan. 2017. Bootstrapping a romanian corpus for medical named entity recognition. In *RANLP*, pages 501–509.

Maria Mitrofan and Dan Tufiş. 2018. Bioro: The biomedical corpus for the romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Danielle L Mowery, Sumithra Velupillai, and Wendy Chapman. 2012. Medical diagnosis lost in translation–analysis of uncertainty and negation expressions in english and swedish clinical texts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 56–64.

Tara Murphy, Tara McIntosh, and James R Curran. 2006. Named entity recognition for astronomy literature. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 59–66.

Mariana Neves, Alexander Damaschun, Andreas Kurtz, and Ulf Leser. 2012. Annotating and evaluating text for stem cell research. In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC). Istanbul, Turkey*, pages 16–23. Citeseer.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Salim Sazzed. 2020a. Cross-lingual sentiment classification in low-resource bengali language. In *Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020)*, pages 50–60.

Salim Sazzed. 2020b. Development of sentiment lexicon in bengali utilizing corpus and cross-lingual resources. In *2020 IEEE 21st International conference on information reuse and integration for data science (IRI)*, pages 237–244. IEEE.

Salim Sazzed. 2021a. Abusive content detection in transliterated bengali-english social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130.

Salim Sazzed. 2021b. Identifying vulgarity in bengali social media textual content. *PeerJ Computer Science*, 7:e665.

Salim Sazzed and Sampath Jayarathna. 2019. A sentiment classification in bengali and machine translated english corpus. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)*, pages 107–114. IEEE.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 170–181.

Riste Stojanov, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, Tome Eftimov, et al. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research*, 23(8):e28229.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: an evaluation of bert in the pharmaconer task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):1–7.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179, Bergen, Norway. Association for Computational Linguistics.

Sumithra Velupillai. 2012. *Shades of certainty: annotation and classification of swedish medical records*. Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University.

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyere, et al. 2005. Umlf: a unified medical lexicon for french. *International Journal of Medical Informatics*, 74(2-4):119–124.