

A Novel Dataset Towards Extracting Virus-Host Interactions

Rasha Alshawi, Atriya Sen
University of New Orleans

Nathan S. Upham, Beckett Sterner
Arizona State University

Abstract

We describe a novel dataset for the automated recognition of named taxonomic and other entities relevant to the association of viruses with their hosts. We further describe some initial results using pre-trained models on the named-entity recognition (NER) task on this novel dataset. We propose that our dataset of manually annotated abstracts now offers a Gold Standard Corpus for training future NER models in the automated extraction of host-pathogen detection methods from scientific publications, and further explain how our work makes first steps towards predicting the important human health-related concept of viral spillover risk automatically from the scientific literature.

1 Motivation and Related Work

The pace of novel zoonotic diseases is increasing globally (Han et al. 2016), but much of our knowledge about the geography and hosts of zoonotic diseases remains locked in the texts of published scientific articles (Upham et al. 2021). Published studies typically apply one or more methods for pathogen detection in animal hosts, including antibody tests, polymerase chain reaction (PCR) tests, whole genome sequencing, or live pathogen isolation. Similarly, the host species might be identified morphologically or using PCR. These methods of detecting host-pathogen interactions vary in precision and in what they tell us about the ecological relationship being observed; most critically, whether the animal host is a reservoir for pathogen replication and transmission, or else a more transient host.

Distinguishing the confidence in host-pathogen data according to type of detection method has been shown to significantly improve models predicting zoonotic disease risk in rodents (Mull et

al. 2021). However, the information required to incorporate detection method as a variable in zoonotic disease risk models is rarely available from current host-pathogen databases or article metadata, with the important exception of (Olival et al., 2017), which we explore further here.

Named Entity Recognition (NER) methods have the potential to assist in identifying host-pathogen interactions, by the automated extraction of virus-host and other pathogen detection methods from the biological literature, enabling advances in scientific understanding of how and why zoonotic diseases emerge. **There are currently no existing datasets for this purpose, and therefore a comparison cannot be furnished.** In the biomedical domain, NER models such as BioRedditBERT (Basaldella et al., 2020), SapBERT (Liu et al., 2021), and Biobert_ncbi_disease_ner (Doğan et al., 2014), are trained on large datasets (as described in Section 3), whereas we manually curate a novel, much smaller dataset (as described in Section 2), albeit attracting the additional challenges of performing NER effectively on “small” data. Additionally, virus detection methods are manually labelled and recognized via NER after training, which is the first such result in the literature.

The Information Extraction (IE) challenges in the NER task on biological scientific articles are highly similar to those in other domains, such as astrophysics, as exemplified by the *DEAL: Detecting Entities in the Astrophysics Literature* (DEAL, 2022) competition. It is true of many domains that there is a diversity of naming practices, rampant ambiguity, and a highly dynamic vocabulary. We therefore envision our data-collection approach, piloted here on host-pathogen literature, to be highly generalizable to other scientific domains.

2 Dataset description

Our virus dataset was built by manually collecting 1104 articles reporting virus detection results for mammal hosts. The articles were selected from the dataset collected and analyzed as part of a systematic literature review of all known viruses with mammal hosts, as reported in (Olival et al., 2017).

Their review searched the Web of Science, Google Scholar, and PubMed for articles published between 1940 and 2015 that mentioned each of 586 virus species identified as having mammal hosts by the International Committee on Taxonomy of Viruses, 8th Edition (Fauquet et al. 2005). They excluded articles reporting results from experimental infections, zoos, or captive breeding facilities as well as domesticated and peri-domestic mammal species (specifically *Mus musculus* and *Rattus norvegicus*). The final list of articles they analyzed is available online as part of the paper’s supplementary data in the “references.txt” file (<https://zenodo.org/record/807517>). Since the supplementary data did not include article abstracts, we searched by article title in the PubMed search engine (PubMed,2022), and with our additions, the dataset we report here represents a substantial fraction of all published articles reporting viruses detected in wild mammals.

524 of these abstracts were preprocessed and manually annotated in the form of Gold Standard Corpus (GSC) for the Name Entity Recognition (NER). The annotations were conducted following well-defined guidelines to ensure consistency and accuracy. NER annotation necessitates explicit rules to mark the boundaries and class of each entity. We adopted the widely recognized Inside, Outside, Beginning (IOB) tagging format (Perera et al., 2020). According to this format, entities are denoted by 'B-' at the beginning, 'I-' for tokens inside the entity, and 'O' to indicate tokens without any entity association. The 'B-' prefix signifies the start of an entity, while 'I-' indicates tokens within the entity. Tokens with 'O' tags do not belong to any entity. This clear and standardized approach to annotation was diligently applied a clear and standardized annotation methodology to identify virus names, host names, and other pertinent entities within the abstracts. Following the Inside, Outside, Beginning (IOB) format, we maintained consistent definitions of entity boundaries and

classes. This meticulous approach resulted in a well-annotated Gold Standard Corpus (GSC), comprising 1104 annotated abstracts, serving as a valuable resource for training and evaluating Name Entity Recognition (NER) models.

The manual annotation process was inherently time-consuming, amplified by challenges in disambiguating entities. Ambiguities, especially when a term represented both a virus name and a host name, required careful contextual analysis. To address this, we implemented a straightforward rule: incorporating the term 'virus' as an entity marker ('B-virus name' for the beginning and 'I-virus' for tokens inside). Despite these challenges, our systematic approach and integration of specific entity markers ensured precise annotations within our GSC. These efforts establish a solid foundation for developing accurate NER models tailored to extract host-pathogen detection methods from scientific publications.

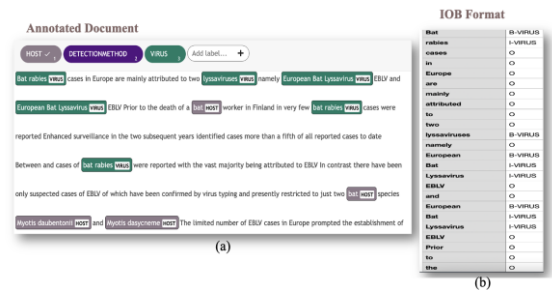


Figure 1: A sample from the dataset demonstrating the annotation process according to Inside, Outside, Beginning (IOB) gold standard corpus. (a) shows the annotation process that includes highlighting and classifying the entities, (b) shows the output annotated in form of IOB standard.

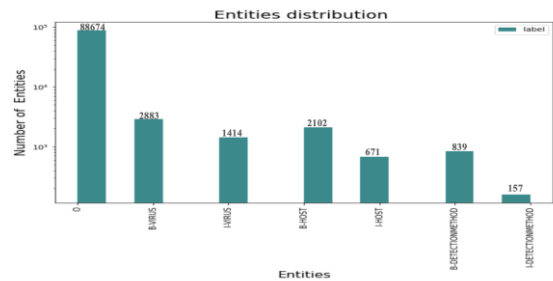


Figure 2: Entity Distribution Graph of Our Dataset.

3 Transformer-based Model

This section describes the model architectures, training, and evaluation procedures for the named-entity recognition (NER) task we performed. All our code, written in Python & using Google TensorFlow, will be made freely available online upon publication.

Transformer models were the first deep neural network-based sequence transduction model based entirely on the concept of *attention*. The model architecture is composed of the transformer’s *encoder*, based on the original implementation described in (Vaswani et al., 2017), and followed by a classification model. Similar to other sequence processing models, the architecture first uses an embedding layer to convert the input tokens into a feature vector representation and a positional encoding layer to provide information about the order of the sequence. The encoder block consists of self-attention layers, normalization layers, and feed-forward layers (i.e., a multilayer perceptron (MLP)), and outputs a vector for each time step of an input sequence. The classification model uses a feed-forward network to classify these sequences into predefined named entities, therefore performing a *sequence classification task*.

We deployed a *Bidirectional Encoder Representations from Transformers* (BERT) model, a transformer-based model that leverages a fine-tuning-based approach for applying a pretrained language model, i.e., a model trained on a generic task in a semi-supervised manner, and then fine-tuned on a specific task in a supervised manner (Devlin et al., 2018). Leveraging pretrained language models significantly improves performance on many tasks, especially when labeled data is scarce, as in our use-case.

Three distinct pretrained BERT models were used, each followed by a classifier model to project the output onto predefined named entities. Since there is no available BERT model that is pretrained on virus and host-related biological literature, available models pretrained on general biological and biomedical literature were used: (1) **BioRedditBERT**, pretrained on large biomedical documents and health-related Reddit posts (Basaldella et al., 2020), (2) **SapBERT**, pretrained on abstracts from PubMed and full-text articles from PubMed Central (Liu et al., 2021), and (3)

BioBERT_ncbi_disease_ner, fine-tuned for NER task on NCBI disease dataset. The NCBI dataset

Table 1: Performance evaluations of the transformer and the three pretrained BERT models finetuned on our novel dataset.

Model	Acc.	Prec.	Recall	F1
Transformer	0.9826	0.9793	0.9826	0.9795
BioReddit BERT	0.9814	0.9772	0.9814	0.9770
SapBERT	0.9857	0.9870	0.9857	0.9853
BioBERT_ncbi	0.9846	0.9840	0.9846	0.9832

consists of 793 PubMed abstracts and contains 6,892 disease mentions (Doğan et al., 2014). All three models are hosted on the HuggingFace model repository.

The pretrained model may be used as a feature extractor by freezing the model’s weights and training only the classification model on the target dataset, or, the weights of some neural layers may be unfrozen and updated on the target task, which is known as *fine-tuning*. Since these models were pretrained on a different corpus, we obtained slightly better results using fine-tuning.

4 Results

An NER experiment was performed to evaluate the quality of the novel dataset. To evaluate and compare NER models using Gold Standard Corpora, it is required to use standardized evaluation scores. A frequently used error measure is the F-Score, a combination of Recall and Precision. NER models were also evaluated using the accuracy metric. Table 1 shows the evaluation performance of the basic transformer model and performance after fine-tuning of the three pretrained BERT models. Table 2 shows the evaluation performance of the models using the feature extraction learning method described in the previous section. SapBERT obtains the best performance in both fine-tuning and feature extraction learning, probably due to its relatively general nature. Table 3 shows the loss metrics after training the models for 20 epochs.

The visualized annotations in **Error! Reference source not found.** show that the SapBERT model

was able to detect and classify almost all the entities of interest: both taxonomic names and detection method names (the latter is a novel result), that appeared in the abstracts.



Figure 3: The visual results of the transformer and two pretrained BERT models of a sample drawn from the dataset: (a) transformer (b) BioRedditBERT. All the BERT models were pretrained on biological and health-related literature, and then fine-tuned on our novel dataset. Red lines underscore unrecognized entities.

Table 2: Performance evaluations of the three pretrained BERT models trained on our novel dataset using the feature extraction approach.

Model	Acc.	Prec.	Recall	F1
BioReddit BERT	0.965 5	0.9376	0.9655	0.9508
SapBERT	0.984 9	0.9859	0.9849	0.9844
Biobert_ncbi	0.984 5	0.9840	0.9846	0.9832

Table 3: Performance evaluations of the pre-trained models on the virus dataset

Model	Loss
Transformer	0.0624
BioRedditBERT	0.0911
SapBERT	0.0736
Biobert_ncbi_disease_ner	0.0904

5 Conclusions, Impact & Potential

We have presented a novel dataset of significance to the important concept of virus-host association, and therefore to the emergence of pandemics such as the COVID-19 pandemic, and promising initial results on the NER task of identifying both taxonomic names and experimental detection methods. We claim that our dataset of manually annotated abstracts now offers a Gold Standard Corpus for training future NER models in the automated extraction of virus-host and other pathogen detection methods from the biological literature. Several other entities, particularly geographical entities and entities describing species migration, are also relevant to the virus-host association. As a result, immediate next steps will consist of recognizing these entities, and also automatically annotating the full text of the article using semi-supervised methods, in lieu of manually annotating the abstracts.

Recognized taxonomic entities in particular can be linked with knowledge graphs representing taxonomic synonymy as well as more complex taxonomic relationships. These graphs have been used (ATCR, 2022) to reason using automated reasoning and inference techniques such as SMT solving and answer-set programming about relationships expressed in a qualitative spatial logical calculus (such as a form of the region connection calculi), with the goals of resolving taxonomic ambiguity or inferring unspecified relationships. This has been used to align and disambiguate published taxonomies of primates and other species (Franz, N.M. et al., 2016). Further, the approach has the potential to be used in biodiversity conservation applications (Sen, A., Sterner, B., et al., 2021). Such inference may be seen as a generalized form of querying or question-answering over taxonomic graphs, and moreover provides a highly intuitive and visual representation of taxonomic flux over time.

Augmenting these graphs of logical taxonomic relationships with automatically extracted context from the biological literature will have the important benefits of serving to identify novel application domains and providing extra-biological context (e.g., geospatial context) to known & inferred taxonomic relationships.

Further, taxonomic automated reasoning systems have previously been combined (Sen, A., Sterner, B., et al., 2021) with statistical features extracted from biological image repositories (such as citizen-sourced or herbarium-sourced images) to further facilitate the taxonomic relationship discovery task. While we have only considered textual abstracts in our work so far, further useful context may thus be added by augmenting taxonomic knowledge graphs with images or tables extracted from the full text of the publications.

The recognition of a variety of intermediary entities (e.g., locations, methods, migration patterns) is likely to facilitate the discovery of the relevant ecological contexts of the host-virus associations, which, in turn, are subjectively known to be dependent (in some currently undiscovered manner) upon these entities. The extraction of such scientifically informative relationships is a further tangible step ahead.

Finally, these extracted relationships may be considered as background structure for *learning an explainable theory of viral spillover* (from other mammals to humans), when taken together with known examples of such spillover, and known negative examples. Symbolic machine learning techniques such as Inductive Logic Programming (ILP) may be able to exploit such structured data and background knowledge to learn logical relationships that generalize from these data, expressed in a subset of first-order logic and interpretable directly by humans: it is in this sense that we use the term *explainable*.

Acknowledgment

This research was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health. The award number has been withheld for anonymity.

References

- Upham, Nathan S, Jorrit H Poelen, Deborah Paul, Quentin J Groom, Nancy B Simmons, Maarten P M Vanhove, Sandro Bertolino, et al. "Liberating Host-Virus Knowledge from Biological Dark Data." *The Lancet Planetary Health* 5, no. 10 (October 1, 2021): e746–50. [https://doi.org/10.1016/S2542-5196\(21\)00196-0](https://doi.org/10.1016/S2542-5196(21)00196-0).
- Han, Barbara A., Andrew M. Kramer, and John M. Drake. "Global Patterns of Zoonotic Disease in Mammals." *Trends in Parasitology* 32, no. 7 (July 1, 2016): 565–77. <https://doi.org/10.1016/j.pt.2016.04.007>.
- Mull, Nathaniel, Colin J. Carlson, Kristian M. Forbes, and Daniel J. Becker. "Viral Competence Data Improves Rodent Reservoir Predictions for American Orthohantaviruses." *BioRxiv*, January 4, 2021, 2021.01.01.425052. <https://doi.org/10.1101/2021.01.01.425052>.
- Olival, J, et al. "Host and viral traits predict zoonotic spillover from mammals" *Nature* 546, pp. 646–650 (2017).
- Fauquet, C., Mayo, M.A., Maniloff, J., Desselberger, U. & Ball, L.A. *Virus taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses*. (Elsevier Academic Press, 2005)
- PubMed, 2022 URL: <https://pubmed.ncbi.nlm.nih.gov/>.
- Perera, N., Dehmer, M., Emmert-Streib, F., 2020. Named entity recognition and relation detection for biomedical information extraction. *Front. Cell Dev. Biol.* 673.
- UBIAI, 2022 URL: <https://ubiai.tools/>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr. ArXiv181004805*.
- Basaldella, M., Liu, F., Shareghi, E., Collier, N., 2020. COMETA: A corpus for medical entity linking in the social media. *ArXiv Prepr. ArXiv201003295*.
- Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N., 2021. Self-Alignment Pretraining for Biomedical Entity Representations. *Association for Computational Linguistics*, pp. 4228-4238.
- Doğan, R.I., Leaman, R. and Lu, Z., 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47, pp.1-10.
- DEAL, 2022 URL: [DEAL Shared Task | WIESP \(harvard.edu\)](https://deal.shared-task.wiesp.harvard.edu)
- A. Sen, N. Franz, B. Sterner, and N. Upham, "Automated Taxonomic Concept Reasoner and Learner." <http://atcrl.herokuapp.com> (accessed Sep. 02, 2022).
- Franz NM, Pier NM, Reeder DM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B. Two Influential Primate Classifications Logically Aligned. *Syst Biol.* 2016 Jul;65(4):561-82. doi: 10.1093/sysbio/syw023. Epub 2016 Mar 22. PMID: 27009895; PMCID: PMC4911943.
- A. Sen, B. Sterner, N. Franz, C. Powel, and N. S. Upham, "Combining Machine Learning & Reasoning for Biodiversity Data Intelligence," presented at the Thirty-Fifth AAAI Conference on Artificial Intelligence, Held virtually, 2021.