SciNLP (2020)

Normalization of Predominant and Long-tail Bacterial Entities with a Hybrid CNN-LSTM and Knowledge-Driven Model

William Hogan*, **Raghav Mehta***, **Yoshiki Vazquez-Baeza***, WHOGAN@ENG.UCSD.EDU
Yannis Katsis§, **Ho-Cheol Kim**§, **Chun-Nan Hsu*** CHUNNAN@UCSD.EDU
* UC San Diego § IBM Research - Almaden

As part of our ongoing effort to construct a biomedical knowledge base [3], we have recently focused on the normalization of bacterial entities. In contrast to other widely studied biomedical entities, such as diseases, we found that bacteria normalization poses unique challenges, primarily due to the skew of the ground truth data available. In this work, we describe the issues and explain the techniques that we used to address them.

To perform bacteria normalization, we started by employing PubTator [2]—a large dataset of bacterial entities—to train a deep learning normalization model. However, PubTator is mostly comprised of a few predominant bacterial species, as shown in Figures 1 and 2. As a result, our normalization model, while performing well on the common bacteria names appearing in PubTator, failed to correctly map other less common bacteria names. To address this issue we employed two approaches: First, we created *a new annotated dataset*, called *MDAD (Microbes and Diseases Annotation Dataset)*. While significantly smaller than PubTator (containing 1.9K vs the 38K bacteria mentions of the latter), MDAD is more representative of general bacteria names, as it is based on the more uniform Disbiome dataset [1]. Therefore, it serves as a better evaluation dataset. Second, we combined our deep learning model with a knowledge-driven approach into a *hybrid* model that targets both common and rare entities. This is based on our observation that predominant bacteria show notable variability in the naming with 10.97±13.04 surface forms per concept on average, while long-tail bacteria have an average of 1.29±0.65 with most mentions using the preferred name. Our deep learning model is a character-based *CNN-LSTM* to model the variability of predominant bacteria. Our *Knowledge-Driven* method leverages Levenshtein distance and abbreviation resolution to deal with long-tail bacteria. Each of these models was found to perform well for their target bacteria but performed poorly otherwise. Therefore, we created a *hybrid model*, which covers all bacteria by smartly combining the two models. It achieves 96% accuracy for test data containing both predominant and long-tail bacteria, substantially outperforming individual models in isolation. The performance results of the hybrid model and its components are shown in Figure 3.

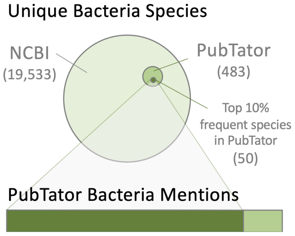

Figure 1: Venn diagram of bacteria species covered by PubTator (drawn to scale)

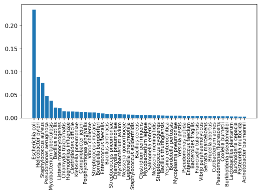

Figure 2: Distribution of the 50 most frequently mentioned bacteria in PubTator

Model	PubTator	MDAD	Mixed
Knowledge-driven	0.829	**0.954**	0.846
CNN + LSTM	**0.972**	0.156	0.905
Hybrid model	0.916	0.887	**0.961**

Figure 3: Performance of bacteria normalization models on different datasets (*Mixed* consists of 3/4 PubTator + 1/4 MDAD)

. This work is supported by IBM Research AI through the AI Horizons Network

References

[1] Yorick Janssens, Joachim Nielandt, Antoon Bronselaer, Nathan Debunne, Frederick Verbeke, Evelien Wynendaele, Filip Van Immerseel, Yves-Paul Vandewynckel, Guy De Tré, and Bart De Spiegeleer. Disbiome database: linking the microbiome to disease. *BMC microbiology*, 18(1):50, 2018.

[2] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41, 07 2013.

[3] Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. Normco: Deep disease normalization for biomedical knowledge base construction. 2019.