# NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction

**Dustin Wright** DBW003@ENG.UCSD.EDU
**Yannis Katsis** YANNIS.KATSIS@IBM.COM
**Raghav Mehta** R3MEHTA@ENG.UCSD.EDU
**Chun-Nan Hsu** CHUNNAN@UCSD.EDU

In this study, we considered the problem of disease entity normalization, an essential task in constructing a biomedical knowledge base. We developed NormCo, a deep coherence model which considers the semantics of an entity mention, as well as the topical coherence of the mentions within a single document. NormCo models entity mentions using a simple semantic model which composes phrase representations from word embeddings, and treats coherence as a disease concept co-mention sequence using an RNN rather than modeling the joint probability of all concepts in a document, which requires NP-hard inference. To overcome the issue of data sparsity, we used distantly supervised data and synthetic data generated from priors derived from the BioASQ dataset. Our experimental results showed that NormCo outperformed state-of-the-art baseline methods on two disease normalization corpora in terms of (1) prediction quality and (2) efficiency, and was at least as performant in terms of accuracy and F1 score on tagged documents. NormCo makes the following contributions:

- A combination of two sub-models which leverage both semantic features and topical coherence to perform disease normalization.
- Addressing the data sparsity problem by augmenting the relatively small existing disease normalization datasets with two corpora of distantly supervised data, extracted through two different methods from readily available biomedical datasets created for non-normalization-related purposes.
- Outperforming state-of-the-art disease normalization methods in efficiency and prediction quality when taking into account the severity of errors, while being at least as performant or better in terms of accuracy and F1 score on tagged documents.

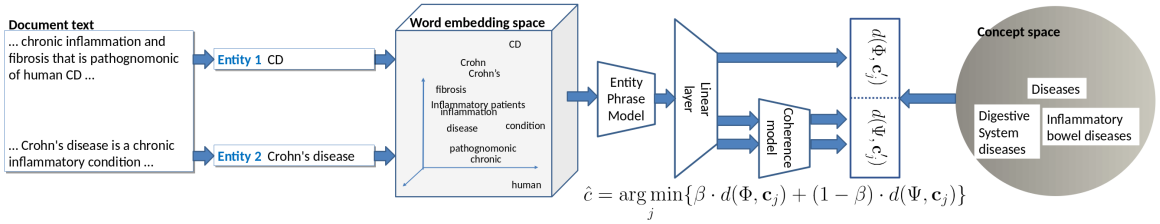


Figure 1: NormCo architecture which utilizes coherence and semantic features for disease normalization.