Named Entity Boosted Topic Models for like Clinical Trial Strategies in Multiple Myeloma

Philip Robinson

OHSU: Center for Speach and Learning Understanding

January 31, 2019

1 Proposal

When a patient is forced down the track of clinical trials, their ability to make informed medical and life decisions can be greatly limited by the vast amount of information in both structured and unstructured text. Inability to efficiently skim through this information also limits their communication bandwidth with medical professionals.

In the case of Myeloma, several trial immunotherapy strategies exist. Traditional treatments of multiple Myeloma focus on persistent reduction in plasma cells with maintenance therapy. More recently, strategies are broken down to specific modifications or implementations CAR-T, BiTEs, and Immuno Checkpoint Inhibitors, for example. Specific cases of CAR-T (chimeric antigen receptor) treatments include strategies targeting BCMA, CD138, and CD19 (and others) antigens. As an example, Chimeric antigen receptor T-cell (CAR-T) strategies genetically modify T-Cells to include a linker designed to bind with a specific antigen expressed on the target diseased cell. The modified T-Cell recognizes or has specificity to the antigen on the diseased cell, and, once bound, is activated thereby killing the cancer.

Given that these clinical trials are described by natural text, there is advantage in applying Natural Language Processing techniques to aid patients and doctors. In particular automatic organization by grouping like documents can significantly reduce human research time, commonly called topic modeling [1].

Unfortunately, most of these strategies aren't able to prioritize and discover important or influential words. To identify important words, there is a second task called named entity recognition where domain knowledge is used to identify important words. CliNER [2], is a named entity recognizer specifically trained to recognize medical terms. Biasing our learned topics, by prioritizing named entities, results in far more informative clusters [3, 4].

I hypothesis that mixing these two strategies, and applying them to the clinical trials database, would allow for informed biasing without manually setting theoretical priors on our topics. In particular, CliNER also annotates entities with a category for the found entities. This should allow for different upweighting on a per-category basis.

References

[1] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.

[2] W. Boag, E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, and T. Naumann. CliNER 2.0: Accessible and Accurate Clinical Concept Extraction. *arXiv e-prints*, March 2018.

[3] Hyungsul Kim, Yizhou Sun, Julia Hockenmaier, and Jiawei Han. Etm: Entity topic models for mining documents associated with entities. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 349–358, Washington, DC, USA, 2012. IEEE Computer Society.

[4] Katsiaryna Krasnashchok and Salim Jouili. Improving topic quality by promoting named entities in topic modeling. In *ACL*, 2018.