

COMPUTER SCIENCE
DEFENSE

*Machine Learning Methods for Biomedical Keyphrase
Extraction*

Zelalem Gero
Emory University

Abstract: Due to the increased generation and digitization of text documents on the Internet and digital libraries, automated methods that can improve search, discovery and mining of the vast body of literature are more essential than ever. Efficient automated methods that extract keywords to retrieve the salient concepts of a document are shown to be of a paramount importance in text analysis, document summarization, topic detection, and recommendation systems among others. One of the largest scientific databases, PubMed, contains more than 33 million citations and abstracts of biomedical literature to facilitate searching across several National Library of Medicine literature resources. The search results mainly depend on the effective indexing of the PubMed citations with MeSH (Medical Subject Headings) and author keywords. While indexing is enormously important in facilitating searching and clustering documents, automated software systems are still far behind human level performance. In this dissertation, we focused on the two tasks of indexing PubMed citations with keywords and MeSH terms. To that end, we proposed 1) an unsupervised extraction method based on phrase-embedding and modified PageRank algorithm which converges faster and performs better than related baseline methods; 2) A Sequence tagging deep learning method based on attending to words that are central to the documents semantics; 3) A semi-supervised deep learning approach to harness vastly available unannotated biomedical data that improves keyword extraction based on uncertainty estimation. 4) A reinforcement Learning-based encoder-decoder method for MeSH indexing.

Tuesday, October 5, 2021, 1:00 pm
<https://emory.zoom.us/j/8187241545>

COMPUTER SCIENCE
EMORY UNIVERSITY