Biomedical Synonym Normalization for Knowledge Graph Insight Generation

Eric Tanalski (tellic LLC)

Extracting insights from unstructured scientific text has lagged other advancements in natural language processing, as recent calls for NLP solutions surrounding COVID-19 published data have highlighted. While dictionary- and machine learning-powered NER approaches can identify document concepts, this alone does not lead to interconnected, actionable information. To serve such information for a scientist to search, the typical path is to push entity-entity relationships into the form of a knowledge graph. However, many search terms have numerous synonyms or acronyms with degenerate definitions, which all need to be resolved into the concepts within a standard ontology to surface correct and complete results to scientists. With annual biomedical publications counted in the millions, a lack of proper standardization will result in a significant number of missed insights from a knowledge graph.

Prior approaches to pre-graph entity normalization include large rule-based dictionaries and string-powered filters/sieves for normalization. These are effective at a first pass, but lead to failures with degenerate terms or acronyms. While the accuracy of relationship generation has grown dramatically with the development of transformer models, actionable relationship results are still dependent on accurate NER endpoints and term normalization.

We present an entity normalization approach that utilizes classifiers fed by cascading text vectorization in parallel with string feature extraction. This allows the context of the sentence, document, title and author's prior art to balance string matching features and merge synonyms into a single ontological concept. Additionally, it provides a mechanism to identify NER false positives, reducing spurious relationships in the graph. The resulting normalized concepts are fed into a series of transformer-powered classifiers acting at the sentence level. The classifiers determine if pairs of concepts in the text are directionally causal, positively correlated, neutral or negatively correlated. Overall normalization performance provided an average F1 score >0.75 across 10 different unstructured data sources, with 750k mentions on average per class.

This normalization led to a notable increase in relationships for insight generation at the graph level. Table 1 shows the impact of normalization on the generation of positive relationships within the knowledge graph. Overall, normalizing the entity mentions provides roughly a 2x-10x increase in the number positive relationships mined from the sources, dramatically increasing the number of potential actionable insights.

Term	Literal Mentions	Normalized Mentions	Positive Relationships (literal mentions)	Positive Relationships (normalized mentions)
Neuroendocrine tumors	15,477	130,015	4,214	37,481
IL6	67,950	968,680	87,095	1,220,880
COVID-19	194,167	316,858	19,132	34,854
Entire HGNC gene ontology	43 million	154 million	10 million	35 million

TABLE 1. **Normalizing each search term into a concept within a single ontology surfaces more relationships for insight generation.** Literal mentions are string matches, while normalized mentions are all resolved to a single ontology: MeSH for diseases and HGNC for genes. Sources: bioRxiv, medRxiv, arXiv, Pubmed, Pubmed Central, ClinicalTrials.gov and USPTO.