Vector Representations of Multi-Word Terms for Semantic Relatedness

Clint A. Cuffy  
Virginia Commonwealth University

Sam Henry  
Virginia Commonwealth University

Bridget T. McInnes  
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/uresposters

© The Author(s)
Vector Representations of Multi-Word Terms for Semantic Relatedness

Introduction

Semantic similarity and relatedness measures quantify the degree to which two concepts are similar (e.g. liver-organ) or related (e.g. headache-aspirin). These metrics are critical to improving many natural language processing tasks involving retrieval and clustering of biomedical and clinical documents and developing biomedical terminologies and ontologies. Numerous ways exist to quantify these measures between distributional context vectors but to date there has not been a direct comparison between these metrics nor an exploration of representing multi-word context vectors. We explore several multi-word aggregation methods of distributional context vectors for the task of semantic similarity and relatedness in the biomedical domain.

Multi-Word Term Aggregation Methods

Overall

Text Corpus Generation
- Medline 2016 Titles and Abstracts

Word Embedding Generation
- Explicit
- Singular Value Decomposition
- Word2vec (CBOW & Skip Gram)

Multi-Word Term Aggregation
- Vector Averaging
- Vector Summation
- Concept Unique Identifier
- Text Corpus Term Compoundification

Spearman’s Rank Correlation Score

Comparison Between All Combinations of Word Embedding Generation Methods, Multi-Word Term Aggregation Methods With Varying Vector Dimensionalities (100, 200, 300, 500, 1000, 1500 up to 3000)

UMLS Evaluation Datasets
- MiniMayoSRS Coders
- MiniMayoSRS Physicians
- UMNSRS Similarity
- UMNSRS Relatedness

Results

We found that vector dimensionality of 200 is best for skip-gram and continuous bag of words, and a dimensionality of 1000 is best for SVD. SG and CBOW created better vector representations than explicit and SVD, but there is no significant increase in correlation using SG versus CBOW. In regards to multi-word term aggregation methods including the summation and averaging of component word vectors, creating multi-word term vectors using the compoundify tool, and creating concept vectors using the MetaMap tool. Concept vectors achieved the highest sum of correlations across all four datasets, but only marginally. No statistical significance was found between any multi-word term aggregation method across all dimensionality reduction techniques, and dimensions tested.

Conclusion

We found that vector dimensionality of 200 is best for skip-gram and continuous bag of words, and a dimensionality of 1000 is best for SVD. SG and CBOW created better vector representations than explicit and SVD, but their is no significant increase in correlation using SG versus CBOW. In regards to multi-word term aggregation methods including the summation and averaging of component word vectors, creating multi-word term vectors using the compoundify tool, and creating concept vectors using the MetaMap tool. Concept vectors achieved the highest sum of correlations across all four datasets, but only marginally. No statistical significance was found between any multi-word term aggregation method across all dimensionality reduction techniques, and dimensions tested.