2017

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

VIRGINIA COMMONWEALTH UNIVERSITY

VCU

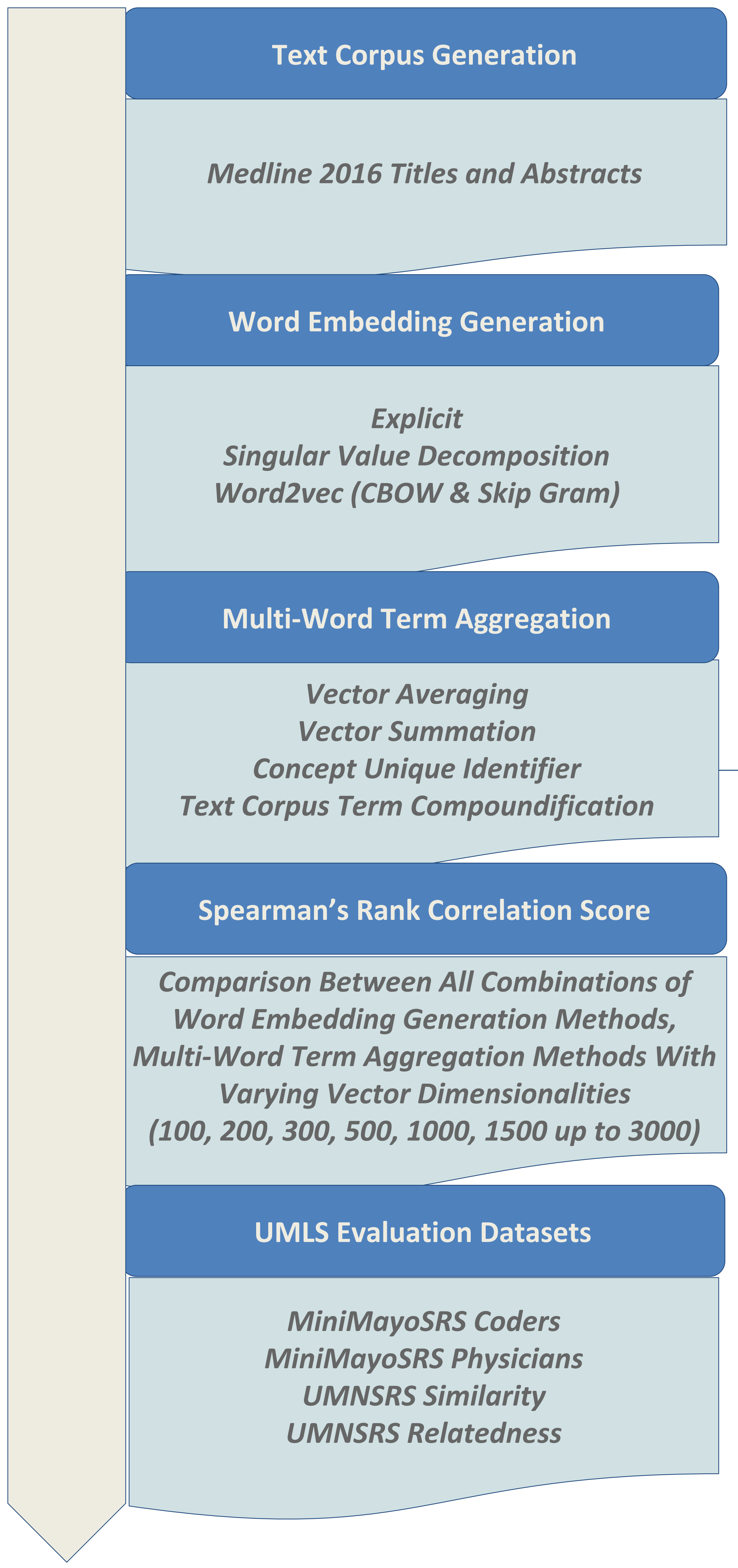
Clint Cuffy
Sam Henry
Dr. Bridget McInnes
Department of Computer Science - Natural Language Processing Lab

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.

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

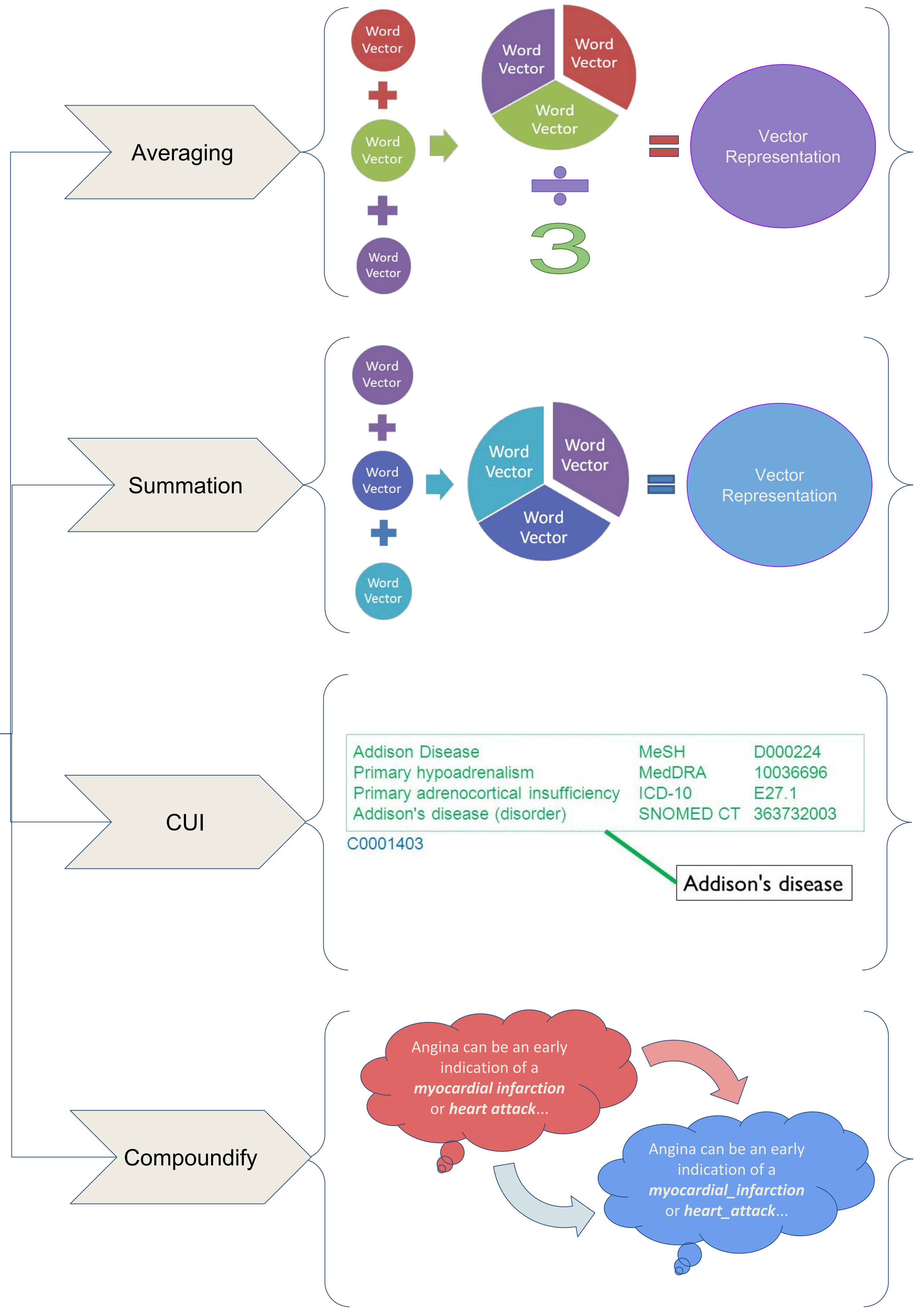
Spearman's Rank Correlation Score

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

Multi-Word Term Aggregation Methods



Averaging

Summation

CUI

Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

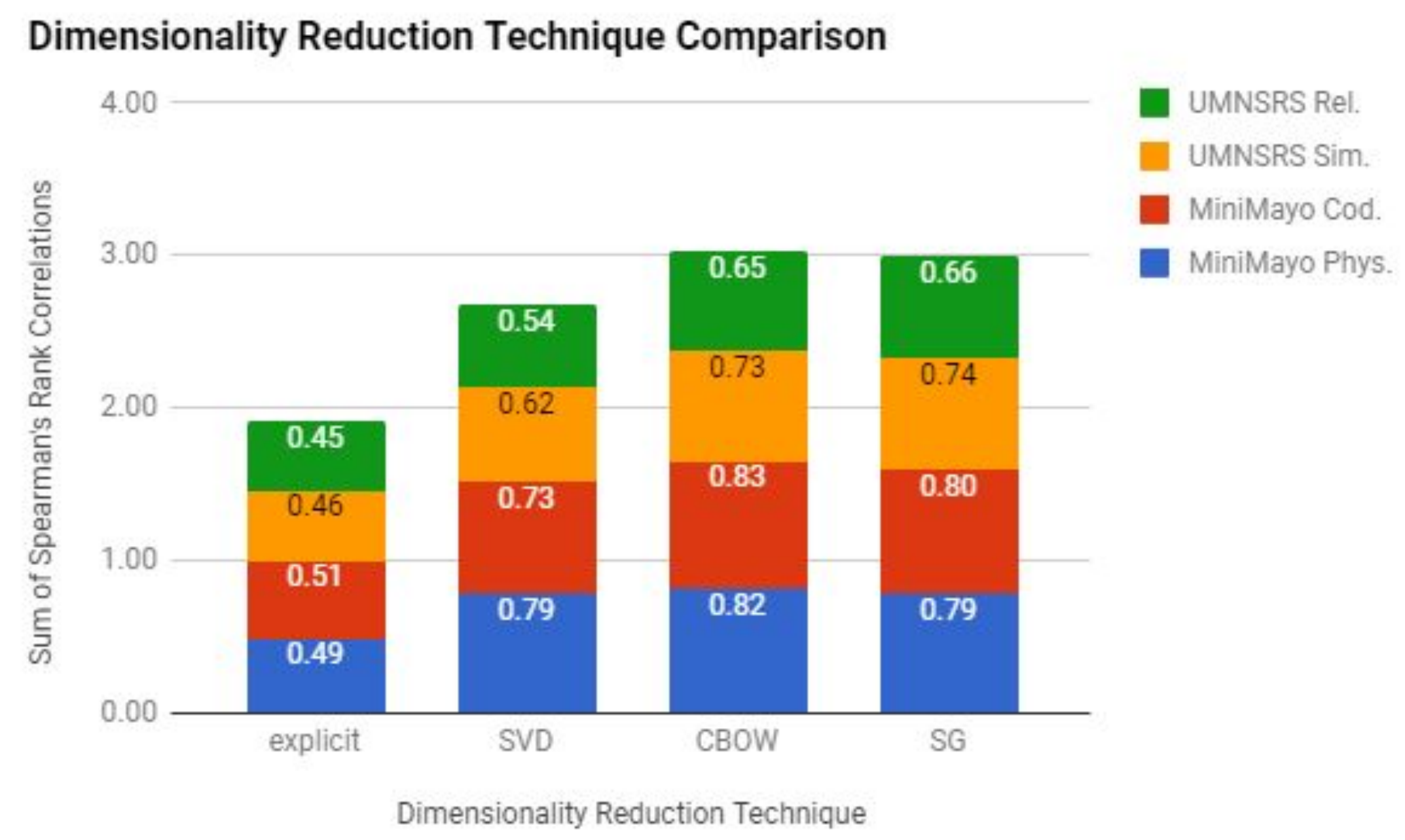
C0001403

Addison's disease

Compoundify

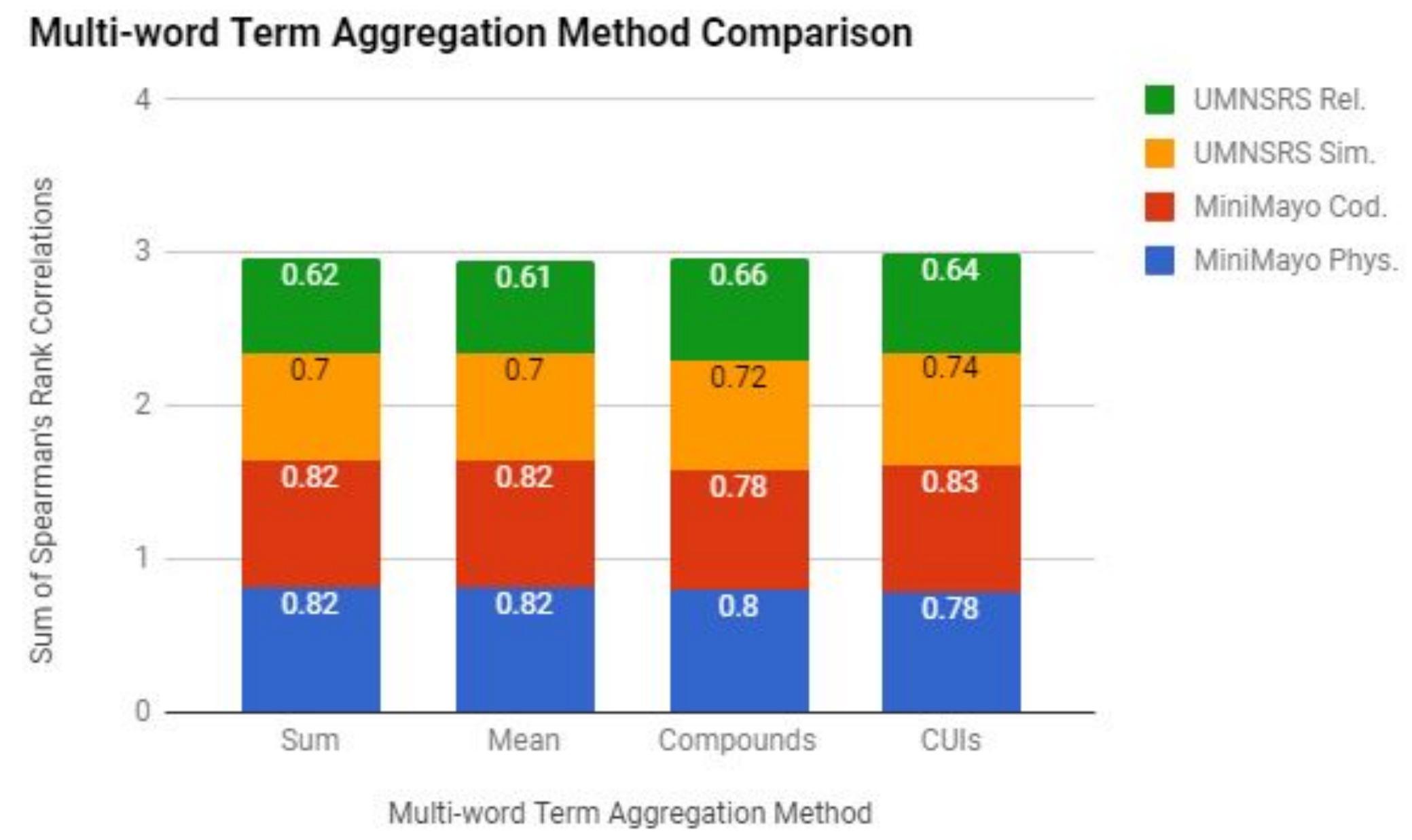
Angina can be an early indication of a *myocardial infarction* or *heart attack*...
Angina can be an early indication of a *myocardial_infarction* or *heart_attack*...

Results



Dimensionality Reduction Technique Comparison

		MiniMayo Phys.					MiniMayo Cod.				
		Dimensionality					Dimensionality				
Aggreg.	**Red.**	100/e	200	500	1000	1500	100/e	200	500	1000	1500
sum	SG	0.78/29	0.79/29	0.74/29	0.76/29	0.74/29	0.79/29	0.80/29	0.78/29	0.79/29	0.78/29
	CBOW	0.81/29	**0.82**/29	0.79/29	0.75/29	-	**0.82**/29	**0.82**/29	0.79/29	0.78/29	-
	SVD	0.38/28	0.57/28	0.56/28	0.79/28	0.66/28	0.36/28	0.53/28	0.52/28	0.54/28	0.71/28
	explicit	0.37/28	-	-	-	-	0.34/28	-	-	-	-
mean	SG	0.78/29	0.79/29	0.74/29	0.76/29	0.74/29	0.79/29	0.80/29	0.78/29	0.79/29	0.78/28
	CBOW	0.81/29	**0.82**/29	0.79/29	0.75/29	-	**0.82**/29	0.81/29	0.79/29	0.78/29	-
	SVD	0.37/29	0.52/29	0.54/29	0.77/29	0.65/29	0.36/29	0.53/29	0.53/29	0.54/29	0.71/29
	explicit	0.34/28	-	-	-	-	0.36/29	-	-	-	-
compound	SG	0.78/28	0.78/28	0.77/28	0.76/28	0.75/28	0.75/28	0.76/28	0.76/28	0.75/28	0.76/28
	CBOW	0.79/28	**0.80**/28	0.79/28	0.77/28	-	0.76/28	**0.78**/28	**0.78**/28	**0.78**/28	-
	SVD	0.65/28	0.74/28	0.75/28	0.72/28	0.70/28	0.65/28	0.73/28	0.70/28	0.72/28	0.72/28
	explicit	0.49/28	-	-	-	-	0.51/28	-	-	-	-
cui	SG	0.76/29	0.76/29	0.77/29	0.76/29	0.76/29	0.77/29	0.77/29	0.78/29	0.77/29	0.79/29
	CBOW	0.77/29	0.75/29	**0.78**/29	0.76/29	-	**0.83**/29	**0.83**/29	**0.83**/29	0.82/29	-
	SVD	0.41/28	0.42/28	0.50/28	0.40/28	0.38/28	0.35/28	0.39/28	0.58/28	0.48/28	0.35/28
	explicit	0.37/28	-	-	-	-	0.26/28	-	-	-	-



Multi-word Term Aggregation Method Comparison

		UMNSRS Rel.					UMNSRS Sim.				
		100/e	200	500	1000	1500	100/e	200	500	1000	1500
sum	SG	**0.70**/374	**0.70**/374	0.68/374	0.69/374	0.68/374	0.59/396	0.61/396	**0.62**/396	**0.62**/396	**0.62**/396
	CBOW	0.68/374	0.69/374	0.66/374	0.61/374	-	0.55/396	0.61/396	0.61/396	0.58/396	-
	SVD	0.53/331	0.52/331	0.55/331	0.56/331	0.52/331	0.41/343	0.36/343	0.45/343	0.47/343	0.45/343
	explicit	0.46/331	-	-	-	-	0.42/343	-	-	-	-
mean	SG	**0.70**/374	**0.70**/374	0.68/374	0.69/374	0.68/374	0.58/397	0.60/397	**0.61**/397	**0.61**/397	**0.61**/397
	CBOW	0.68/374	0.69/374	0.66/374	0.61/374	-	0.55/397	0.59/397	0.59/397	0.57/397	-
	SVD	0.53/332	0.52/332	0.55/332	0.55/32	0.52/332	0.39/346	0.34/346	0.46/346	0.47/346	0.43/346
	explicit	0.33/400	-	-	-	-	0.36/430	-	-	-	-
compound	SG	**0.72**/373	0.71/373	0.70/373	0.69/373	0.70/373	0.63/393	0.64/393	0.64/393	0.65/393	**0.66**/393
	CBOW	0.70/373	0.70/373	0.68/373	0.65/373	-	0.62/393	0.64/393	0.65/393	0.65/393	-
	SVD	0.49/328	0.51/328	0.58/328	0.60/328	0.58/328	0.39/335	0.38/335	0.48/335	0.54/335	0.52/335
	explicit	0.45/328	-	-	-	-	0.45/335	-	-	-	-
cui	SG	**0.74**/388	**0.74**/388	**0.74**/388	**0.74**/388	**0.74**/388	0.62/413	0.62/413	0.63/413	**0.64**/413	**0.64**/413
	CBOW	0.72/388	0.73/388	0.73/388	0.72/388	-	0.56/413	0.56/413	0.59/413	0.60/413	-
	SVD	0.41/362	0.45/362	0.50/362	0.53/362	0.57/362	0.26/380	0.31/380	0.30/380	0.34/380	0.38/380
	explicit	0.35/362	-	-	-	-	0.20/380	-	-	-	-

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.

[1] Research in press at the International Journal of Biomedical Informatics.