Unsupervised Text Similarity

Clint Cuffy, Sam Henry, Ph.D., Bridget T. McInnes, Ph.D.
Virginia Commonwealth University, Richmond, VA, USA
George Mason University, Fairfax, VA, USA

Introduction

The induction of Electronic Health Records (EHR) provides a challenging avenue for data exploration. These documents accommodate patient information fields including medical history, diagnosis, treatment and prognosis information. Their systematic use has led to an increase in the quality of patient care, however also introduces challenges due to the lack of document standardization. Challenges include poor organization, inclusion of unnecessary information and population with erroneous data. These drawbacks contribute to increased difficulty for health-care providers in providing viable treatment options to their patients. As EHRs are synchronized across a variety of differing sources without regard to documentation standards, this compounds the challenges presented to health-care providers in administering rapid quality care. Automated methods of extracting relatedness between EHRs will simultaneously reduce data redundancy and the burden of health-care providers' clinical-decision making abilities. In this work, we explore several unsupervised textual analysis approaches in computing semantic relatedness among EHR records by computing semantic similarity between EHR text snippets.

Data

The data in this study is composed of 1642 de-identified pairs of clinical text snippets, or sentences, extracted from a variety of EHRs. These sentence pairs are ordinally rated from 0 to 5 with respect to their semantic equivalence where 0 indicates no semantic overlap and 5 indicates complete semantic overlap. Likewise, the testing data is composed of 412 de-identified pairs of sentences; however contains no semantic equivalence scores.

Methods

We compare pairs of sentences in a variety of ways. First, we compare sentence pairs using the $lesk$[1,2] approach over the terms in the sentences and the terms mapped to Concept Unique Identifiers (CUIs). Next, we represent the individual sentences as embeddings. With these embeddings we quantify the degree of semantic correlation between each sentence pair using cosine similarity. Lastly, we combine these two methods. Lesk is used to locate phrase overlap between both sentences, a score is generated and overlapping phrases are removed from each sentence. Relatedness between remaining phrase data in each sentence is computed using embeddings. The initial lesk and embedding correlation scores are weighted then summed providing an over all score for each sentence pair. We evaluate $word$[3], $document$[4] and $concept$[5] $embeddings$ generated over three distinct training corpora: 1) EHRs from Mimic III data; 2) titles and abstracts from the 2015 MEDLINE baseline ranging in year from 1975 to 2015; 3) Metamapped Medline Baseline from Concept Unique Identifiers[5].

Table 1: Pearson's Correlation Coefficient Scores

Training Data-set			**Testing Data-set**		
Method	**Dataset**	**Correlation Score**	**Method**	**Dataset**	**Correlation Score**
Lesk (Terms)		0.71	Lesk (Terms)		0.49
Lesk (CUI)		0.68	Lesk (CUI)		0.58
CUI Embeddings	Medline	0.62	CUI Embeddings	Medline	0.46
Document Embeddings	Mimic III	0.67	Document Embeddings	Mimic III	0.40
Word Embeddings (WE)	Medline	0.65	Word Embeddings (WE)	Medline	0.40
Word Embeddings (WE)	Mimic III	0.66	Word Embeddings (WE)	Mimic III	0.42
Combined (Lesk + WE)	Medline	0.68	Combined (Lesk + WE)	Medline	0.47
Combined (Lesk + WE)	Mimic III	0.69	Combined (Lesk + WE)	Mimic III	0.45

WE: Word Embeddings

Results

We evaluate our methods using Pearson's correlation. Table 1 contains each method and Pearson's correlation score for the training and test data. High positive correlation was noted using Lesk (Terms) for the training data and Lesk (CUI) for the test data.

Conclusions

Results show that our methods obtained overall higher accuracy with the training versus testing data for this task. Lesk (Terms) captures high overlap between text snippets while comparing training data. With Lesk (CUI) the mapping of terms to CUIs reduces the vocabulary size by reducing specific terms to general concepts within text snippets to achieve higher testing results.

References

[1] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone; 1986. p. 24–26.

[2] Banerjee S, Pedersen T. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. 2002;p. 136–145.

[3] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[4] Le Q, Mikolov T. Distributed Representations of Sentences and Documents. ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning. 2014;32:1188–1196.

[5] Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium. Washington, DC; 2001. p. 17–21.