US-SERBIA & WEST BALKAN DATA SCIENCE WORKSHOP
Belgrade, Serbia | August 26-28, 2018

A pilot cognitive computing system to understand immunization programs

Marija Stanojevic[1], Fang Zhou[1], Sarah Ball[2], William Campbell[2], Alison Thaung[2], Jason Brinkley[2], Stacie Greby[3], Alexandra Bhatti[3], Allison Fisher[3], Yoonjae Kang[3], Cynthia Knighton[3], Pamela Srivastava[3], Zoran Obradovic[1]

Temple University[1] Abt Associates[2] Center for Disease Control[3]

Abstract

US national, state, local, and territorial immunization programs use quantitative and qualitative data to ensure vaccinations are provided to prevent diseases. The results of qualitative data analysis are not always available to improve vaccination coverage because of analysis is labor intensive. The Immunization Program Cognitive Computing System (IPCCS) was developed to analyze qualitative data for the Centers for Disease Control and Prevention (CDC).

Data

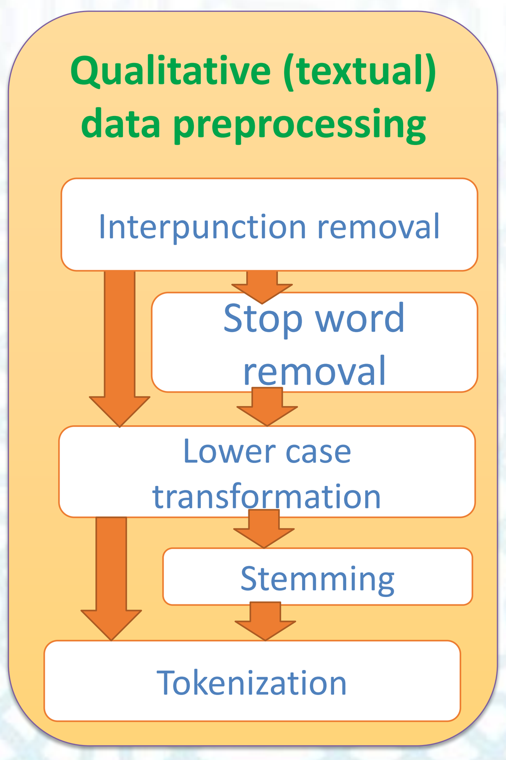
Text from a variety of formal and informal sources was collected to develop the IPCCS lexicon. Formal data included policy documents, vaccine-related websites, scientific journals, textbooks, and state vaccination-related laws. Informal data included Sysomos searches of Twitter, online forums news, and social media feeds from November 2016 to May 2018.

Problem and results description

Main challenges to IPCCS development included: 1) collection and matching spatio-temporal information of formal and informal data, 2) data cleaning and pre-processing to remove references to external documents, non-relevant data, jargon, typos, and misspellings, and 3) fast and well-performing word and paragraph searching. To address these challenges, data were iteratively and thoroughly cleaned and filtered, the best existing algorithms for text understanding were used, and a new algorithm for paragraph searching was developed. Customized features for the lexicon were developed to ensure that the results of the IPCCS are useful to vaccine domain researchers (e.g., ranking of US states to show representation extent of search phrase).

Data preprocessing

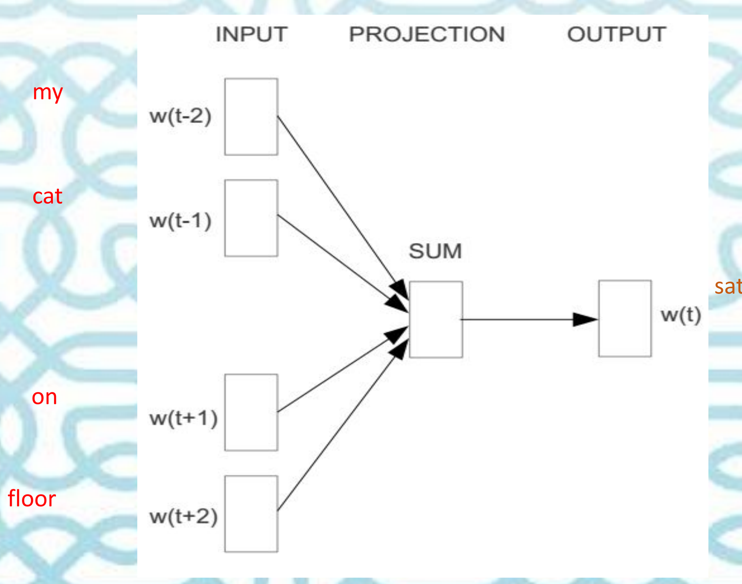
Only textual information were used for creating lexicons, even though quantitative data such as location, time, and other features were available. Text was cleaned from links, user-names, special characters (except dash and apostrophe symbols), and numbers and then split into words/tokens. Stop-words were removed from tokens. Jargon terms, connected words, and misspellings were kept in the data to allow routinely used language describing immunization to be used in future analysis.



Lexicon creation

Several algorithms and related versions were used to develop the lexicon: word2vec [1] and its variation doc2vec [2], glove [3] and WTM [4]. Word2vec algorithm is a small neural network with very little parameters that allows fast learning of vector representation of each word. It has four different versions distinguished by sampling of output (hierarchical softmax or negative sampling) and by learning architecture (continuous bag of words which is learning word from context and skip-gram which is learning context from word).

Doc2vec is also able to learn vectors of documents. It has two versions based on architecture that can learn word from paragraph and words (PV-DM) or can learn words from paragraph (PV-DBOW). All of those six versions were evaluated on both formal and informal datasets using different number of vector sizes.



Lexicon creation

Glove algorithm builds matrix of words co-occurrences X_{ij} and then optimizes cost function:

$$\hat{J} = \sum_{i,j} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j - logX_{ij}\right)^2 \text{where } w \in R^d \text{ are word vectors}$$

and $\tilde{w} \in R^d$ are context word vectors and $f(x) = \begin{cases} (x/x_{max})^{3/4}, if\, x < x_{max} \\ 1, otherwise \end{cases}$.

This function prevents influence from large co-occurrence of common words (e.g., "on the", "at that", "he is"). The glove algorithm generally shows better results than word2vec [3], but we found that algorithm was not always stable.

WTM model is a purely statistical model which combines ideas from latent Dirichlet allocation (LDA) and word2vec algorithms to learn both representation of words and topics simultaneously. The WTM model generally shows better results than word2vec [4], but it is more complex model and slower at producing output.

Conclusions

Evaluation of word2vec/doc2vec algorithms showed the best outcomes were given by continuous bag of word with negative sampling followed by PV-DM on both datasets. References used different vectors lengths – number of features to describe words. For versions of word2vec algorithm, vectors of length 50 gave best results, even though recommended setting in the references are different. Change in vector size for glove algorithm wasn't that crucial as in word2vec algorithms.

Analysis showed that different aspects of similarity are brought to the top by different algorithms. The word2vec result was usually better in understanding relationships between query and paragraph, while glove algorithm results were more influenced by frequency of occurrence of the most significant words in query and paragraph.

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[2] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

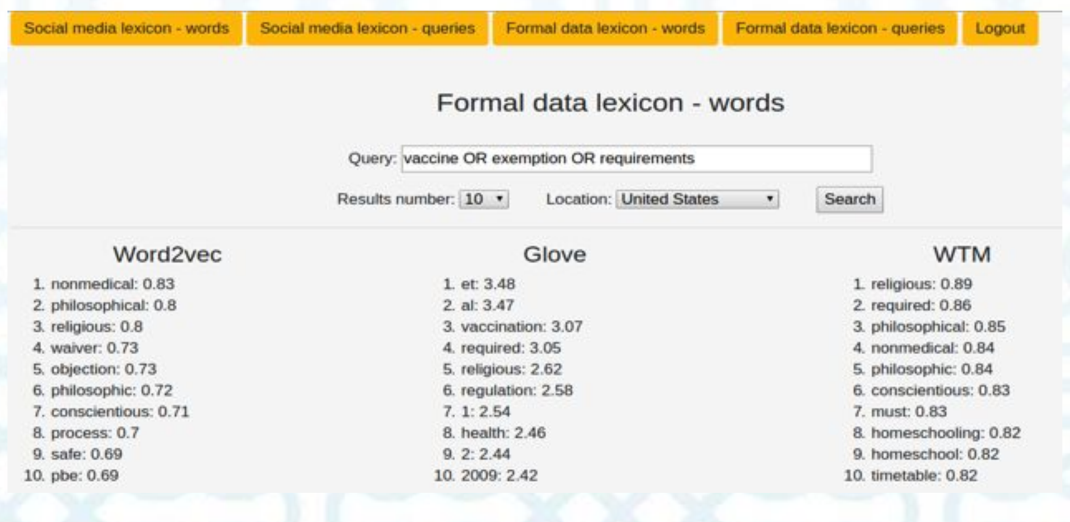
[3] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[4] Fu, X., Wang, T., Li, J., Yu, C., & Liu, W. (2016, November). Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model. In Asian Conference on Machine Learning (pp. 190-205).

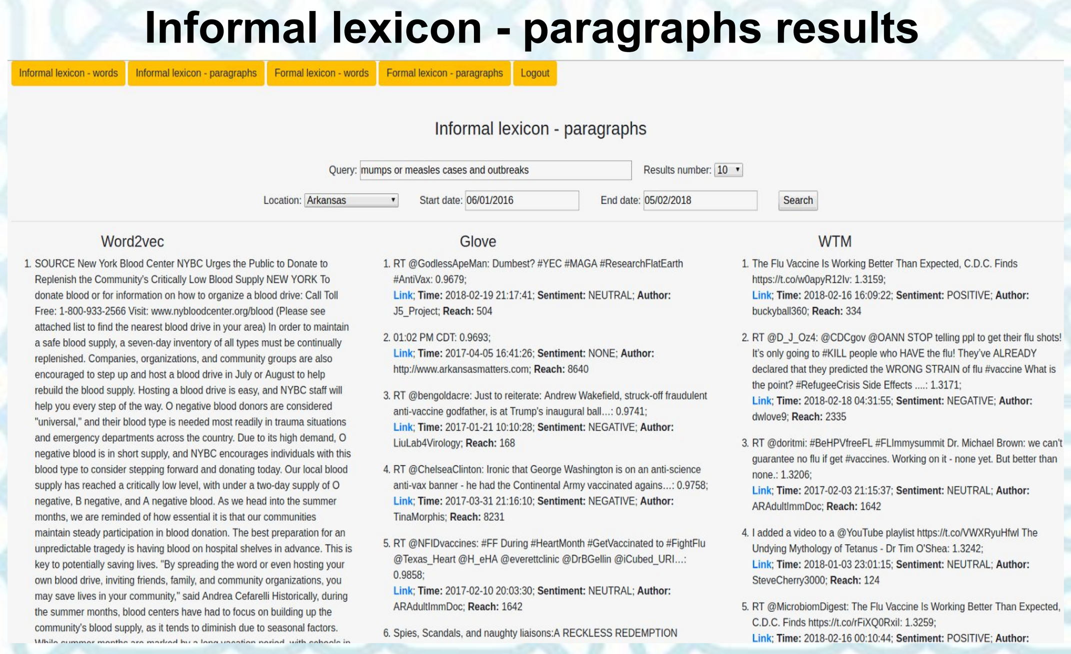
Acknoledgment

This research was supported by the Centers for Disease Control through project Pilot to Assess Cognitive Computing to Analyze Immunization Program Data.

Formal lexicon - words results



Informal lexicon - paragraphs results



Formal lexicon - results aggregated by state