OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Machine Learning for Identifying Relevance to Biosurveillance in Multilingual Text

Qiaochu Chen*[1] and Lauren E. Charles[2]

[1]Tulane University, New Orleans, LA, USA; [2]Pacific Northwest National Laboratory, Richland, WA, USA

Objective

The objective is to develop an ensemble of machine learning algorithms to identify multilingual, online articles that are relevant to biosurveillance. Language morphology varies widely across languages and must be accounted for when designing algorithms. Here, we compare the performance of a word embedding-based approach and a topic modeling approach with machine learning algorithms to determine the best method for Chinese, Arabic, and French languages.

Introduction

Global biosurveillance is an extremely important, yet challenging task. One form of global biosurveillance comes from harvesting open source online data (e.g. news, blogs, reports, RSS feeds). The information derived from this data can be used for timely detection and identification of biological threats all over the world. However, the more inclusive the data harvesting procedure is to ensure that all potentially relevant articles are collected, the more data that is irrelevant also gets harvested. This issue can become even more complex when the online data is in a non-native language. Foreign language articles not only create language-specific issues for Natural Language Processing (NLP), but also add a significant amount of translation costs. Previous work shows success in the use of combinatory monolingual classifiers in specific applications, e.g., legal domain [1]. A critical component for a comprehensive, online harvesting biosurveillance system is the capability to identify relevant foreign language articles from irrelevant ones based on the initial article information collected, without the additional cost of full text retrieval and translation.

Methods

The analysis text dataset contains the title and brief description of 3506 online articles in Chinese, Arabic, and French languages from the date range of August, 17, 2016 to July 5, 2017. The NLP article pre-processing steps are language-specific tokenization and stop words removal. We compare two different approaches: word embeddings and topic modeling (Fig. 1). For word embeddings, we first generate word vectors for the data using a pretrained Word2Vec (W2V) model [2]. Subsequently, the word vectors within a document are averaged to produce a single feature vector for the document. Then, we fit a machine learning algorithm (random forest classifier or Support Vector Machine (SVM)) to the training vectors and get predictions for the test documents. For topic modelling, we used a Latent Dirichlet Allocation (LDA) model to generate five topics for all relevant documents [3]. For each new document, the output is the probability measure for the document belonging to these five topics. Here, we classify the new document by comparing the probability measure with a relevancy threshold.

Results

The Word2Vec model combined with a random forest classifier outperformed the other approaches across the three languages (Fig. 2); the Chinese model has an 89% F1-score, the Arabic model has 86%, and the French model has 94%. To decrease the chance of calling a potentially relevant article irrelevant, high recall was more important than high precision. In the Chinese model, the Word2Vec with a random forest approach had the highest recall at 98% (Table 1).

Conclusions

We present research findings on different approaches of relevance to biosurveillance identification on non-English texts and identify the best performing methods for implementation into a biosurveillance online article harvesting system. Our initial results suggest that the word embeddings model has an advantage over topic modeling, and the random forest classifier outperforms the SVM. Directions for future work will aim to further expand the list of languages and methods to be compared, e.g., n-grams and non-negative matrix factorization. In addition, we will fine-tune the Arabic and French model for better accuracy results.

	Baseline	Word2Vec + random forest	Word2Vec + SVM	LDA
precision	0.68	0.81	0.72	0.56
recall	1.00	0.98	0.98	0.96
F1-score	0.84	0.89	0.83	0.71
10-fold cv accuracy		0.96	0.93	

Table 1. Results of the Chinese model using different methods. Abbreviations in Text.

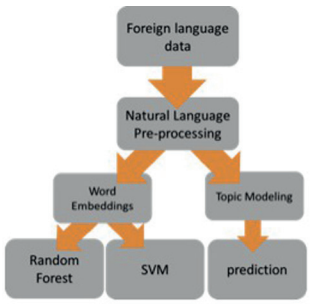

Figure 1: Methodology for comparing different methods to identify the best approach to classifying text data as relevant to biosurveillance.

Figure 1: Methodology for comparing different methods to identify the best approach to classifying text data as relevant to biosurveillance.

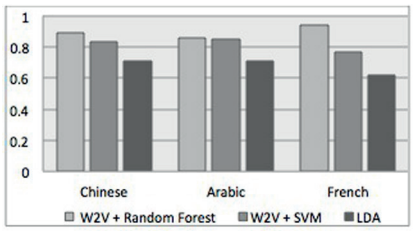

Figure 2: F1-scores of different methods across languages. Abbreviations in Text.

Figure 2. F1-scores of different methods across languages. Abbreviations in Text.

Keywords

Machine learning; biosurveillance; natural language processing; multilingual; articles

Acknowledgments

This work was supported by the Department of Homeland Security Science and Technology Directorate under DOE Contract Number DE-AC05-76RL01830 for the management and operation of Pacific Northwest National Laboratory.

References

[1] Gonalves T, Quaresma P. 2010. Multilingual text classification through combination of monolingual classifiers. Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 29-38

[2] Bojanowski P, Grave E, Joulin A, Mikolov T. 2016. Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.

[3] Blei D, Ng A, Jordan M. 2003. Latent Dirichlet Allocation. The Journal of Machine Learning Research. p.993-1022.

***Qiaochu Chen**
E-mail: qchen7@tulane.edu