Expanding a Gazetteer-based Approach for Geo-Parsing Text from Media Reports on Global Disease Outbreaks

Mikaela Keller, Clark C. Freifeld, John S. Brownstein,

Children's Hospital Informatics Program, Children's Hospital Boston, Department of Pediatrics, Harvard Medical School, Boston, United States

OBJECTIVE

Discovering geographic references in text is a task that human readers perform using both their lexical and contextual knowledge. Automating this task for real-time surveillance of informal sources on epidemic intelligence therefore requires efforts beyond dictionary-based pattern matching. Here, we describe an automated approach to learning the particular context in which outbreak locations appear and by this means extending prior knowledge encoded in a gazetteer.

BACKGROUND

HealthMap (www.healthmap.org) is a freely accessible, automated real-time system that monitors, organizes, integrates, filters, and maps online news about emerging diseases [1]. The system performs geographic parsing ("geo-parsing") [2] of disease outbreaks by assigning incoming alerts to low resolution geographic descriptions, such as country, with the help of a purposely crafted gazetteer. However, the system is limited by the size of the gazetteer, precluding high resolution assignment of place. In this study, we use the prior knowledge encoded in the gazetteer to expand the capabilities of the geo-parsing system.

METHODS

A dataset of 2,500 articles on disease outbreaks, retrieved by HealthMap in 2007, is used as the training dataset. The words in the articles are tagged with the targeted tags 'loc' if they match geographical references found in the gazetteer and 'none' otherwise. In addition to those tags, general linguistic knowledge (part-of-speech tags, words' capitalization status, etc.) is also applied to the articles. With the geographic lexicon partially hidden, we use the additional linguistic knowledge to teach a parsing neural network, similar to the one proposed in [3], which relies on a window of words surrounding the word to parse. Because it takes context into account, this approach provides a generalization of the gazetteer's rule-based geo-parsing.

RESULTS

The observed increase in performance for hidden words proportionally to the size of the lexicon cut-off suggests the potential for discovering phrases not in the initial gazetteer without substantial loss in performance for unhidden words. A preliminary evaluation using a commercial geo-parser (MetaCarta), reveals that up to 50% of the words labeled as 'none' which the system identified as location are in fact location references that are outside of the gazetteer.

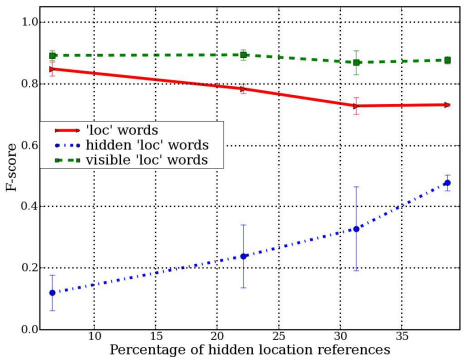

FIGURE 1: F-score performance measure on a validation dataset of neural networks trained over datasets with varying lexicon masking.

CONCLUSIONS

We have demonstrated that the described model has the ability to discover geographic references based solely on their context. The experiments also reveal that providing additional training material improves the performance of the model, suggesting that despite the fabricated nature of the data it is still able to generate interesting information. This technique could also be integrated with a more conventional method for geo-parsing based on a geographically annotated dataset. This approach may provide further improvement in the precision of indexing disease outbreak reports for geographic information retrieval.

REFERENCES

[1] Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc, 15(2)*, 150–157.

[2] Woodruff AG, Plaunt C. 1994. Gipsy: Automated geographic indexing of text documents. *Journal of the American Society for Information Science, 45:9*, 645–655.

[3] Collobert, R, Weston, J. 2007. Fast semantic extraction using a novel neural network architecture. *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics.*

Further Information: mikaela.keller@childrens.harvard.edu