

# A System for the Automatic Comparison of Machine and Human Geocoded Documents

Ian Turton  
GeoVISTA Center  
Pennsylvania State University,  
University Park, PA 16802, USA  
+1 814 865 5642  
ijt1@psu.edu

## ABSTRACT

This paper describes an initial experiment in testing a geocoding system by comparing geocoded documents to locations assigned by human indexers as part of the MeSH indexing process of the PUBMED abstracting system. Preliminary results indicate that this is a useful check on the geocoding system and provides useful feedback to developers of geocoding systems.

## Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing] *Linguistic processing*

## General Terms

Reliability, Experimentation, Verification.

## Keywords

Geocoding, MeSH

## 1. INTRODUCTION

The task of geocoding unstructured text documents is a hard problem with a long history of research [3,5]. Much of this research has concentrated on the task of detecting and extracting place names from the text (Named Entity Extraction) rather than the disambiguation of the place names found. Disambiguation is a problem in all types of named entity extraction work but is particularly difficult in the field of geography as toponyms are reused across the world (so called Geo-Geo ambiguity) and often places share names with people or every day objects (Geo-NonGeo ambiguity) [1]. However there has been little research on how to evaluate the named entity extraction and geocoding systems that have been developed. Leidner [2] describes one of the few systematic attempts to develop an evaluation data set for the toponym resolution problem. In his paper describes his use of the Reuters Corpus to develop a standard text base to test geocoding systems against. As the Reuters corpus was already human indexed to provide a “gold standard” named entity markup assessing the

toponym resolution task was easy, however the harder task of toponym annotation (or geocoding) was carried out separately using the NGA/USGS gazetteer and a group of human volunteers selecting from the possible matches found in the database of locations which proved laborious and potentially error prone.

This paper describes a related methodology which makes use of another commonly available public domain text corpus with human indexed terms, abstracts from papers in the PUBMED system. Each abstract in the PUBMED system has been indexed using a controlled vocabulary known as MeSH which was developed by the National Library of Medicine. It consists of sets of terms naming descriptors in a hierarchical structure. One of the higher level terms included in the MeSH hierarchy is “Geographic Location”, which contains child terms for regions and countries and some cities below it. The complete geographic hierarchy contains 372 distinct regions, countries or populated places. There is no ambiguity in the location of these regions as they are defined in terms of a “is a part of” hierarchy.

## 2. METHOD

As part of a wider project on health geography a subset of papers relating to avian influenza were extracted from PUBMED and named entity extraction and geocoding was carried out on each abstract [4]. This gave a working set of 4738 papers, of those 1982 papers were found to contain at least one geographic named entity, while 1871 papers were indexed with a geographic MeSH term. For each geographic location in the list of MeSH terms all papers that were indexed with that term were extracted from the dataset. Since when the papers were geocoded the system preferred results that were more specific over more general results it was necessary to extract all papers that had been geocoded as including any part of the region being compared. So for example a paper may be indexed as Pennsylvania in the MeSH terms if it includes a reference to an avian flu outbreak in State College, PA, where as the geocoding system will locate this accurately as State College, PA (40.802,-77.8564). Thus for comparison purposes the analysis was carried out using all papers that were geocoded with Pennsylvania or any child location of Pennsylvania.

## 3. RESULTS

As can be seen in Figure 1 the automatic geocoding system performs reasonably well (70% average correct) for the range of countries shown with only a few obvious problems. Figure 1 also shows that the geocoder (FactX) geocoded more papers than the MeSH indexers. This is because the geocoder locates any place mention, whereas the indexers only tagged papers which were predominately about a location.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIR'08, October 29–30, 2008, Napa Valley, California, USA.

Copyright 2008 ACM 978-1-60558-253-5/08/10...\$5.00.

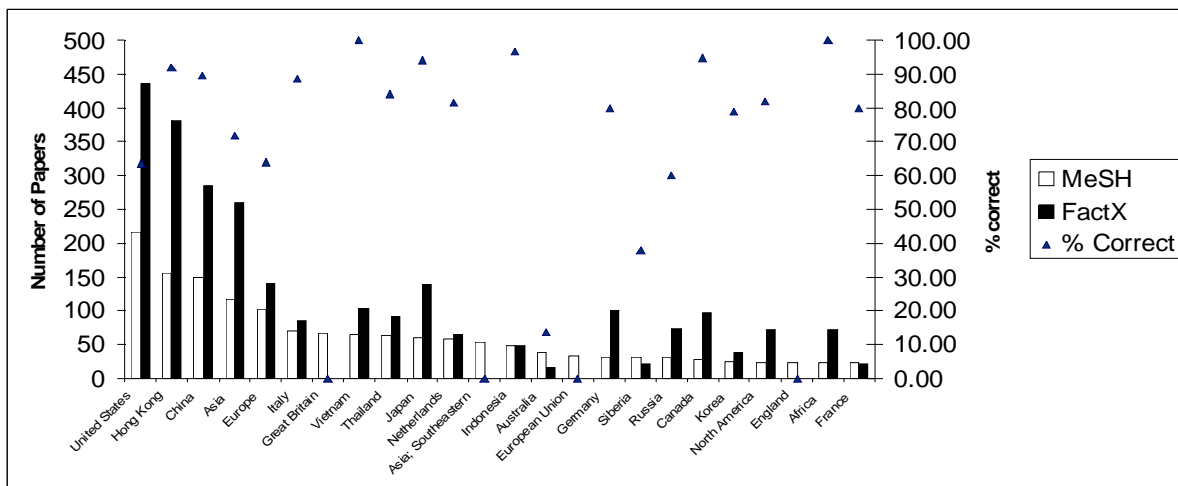


Figure 1: Number of papers about a country and % correctly geocoded

#### 4. PROBLEMS

The first group of errors occurs for MeSH terms which are not real countries such as Asia; Southeastern and the European Union which do not occur in the gazetteer used and so can not be matched. The other group of problem countries that can be seen are Great Britain and England, this is again an artifact of the gazetteer<sup>1</sup> used and the normalization process applied to locations. All locations returned from the geocoder were an ISO country name which means that Great Britain is properly known as the United Kingdom. While England is a region of the UK but is not a generally used administrative unit and is again removed by the geocoder in favor of more commonly used counties. Once this problem is corrected the success rate for the geocoder rises to 58.8% for the United Kingdom. Closer inspection of results for Australia, showed that the named entity parser was unable to correctly extract “New South Wales” as the word new was considered to be a stop word when found next to a direction, this reduced the entities extracted to “Wales” a whole hemisphere away from the correct location. This also had the effect of lowering the success rate for the UK as well. Another problem which was detected was a tendency for the geocoder to consider “de” to be Delaware when it occurred in the middle of place names such as Rio de Janeiro, causing Delaware to be over represented in the initial output.

Initial results of the test system described in this paper indicate that geocoder does a good job of determining the region of interest of a paper. However, in some cases, because the human indexer has access to the whole paper (especially with foreign language papers) where as the geocoding system is limited to only the abstract (or sometimes the title) of the paper if that is all that is publicly available, it is sometimes possible for the human indexer to determine a location when the computer system is unable to make any attempt. This reflects the design goals of the overarching system which is used for the automated location of abstracts from PUBMED. The results above have been adjusted to allow for this by removing papers with no abstract from the counts used to calculate the percentages of correctly geocoded papers.

<sup>1</sup> <http://www.geonames.org>

#### 5. CONCLUSIONS

The system that has been described in this paper shows a useful automated method to provide a first pass check on the accuracy of an unsupervised geocoding system. By exploring the errors flagged up by this system developers were able to fix systematic errors in both the named entity extraction routines and the geolocation routines.

#### ACKNOWLEDGMENTS

This work is supported, in part, by the National Visualization and Analytics Center, a U.S Department of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL). PNNL is a U.S. Department of Energy Office of Science laboratory.

#### REFERENCES

- [1] E. Amitay, N. Har’el, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In SIGIR ’04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 273–280, New York, NY, USA, 2004. ACM Press.
- [2] J. L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417, July 2006.
- [3] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In ECDL ’01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, pages 127–136, London, UK, 2001. Springer-Verlag.
- [4] I. Turton, M. Gahegan, and A. Jaiswal. Geographic information retrieval from disparate data sources. In *GeoComputation’07*, Maynoth, Ireland, September 2007.
- [5] W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H. Goh. On assigning place names to geography related web pages. In JCDL ’05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, pages 354–362, New York, NY, USA, 2005. ACM Press.