Geographic Information Retrieval from Disparate Data Sources

Ian Turton, Mark Gahegan, Anuj Jaiswal

GeoVISTA Center, School of Geography, Pennsylvania State University, University Park, 16802, PA

Tel: +1 814-865-5642 Fax: +1 814-863-7943

ijt1,mng1,arj135@psu.edu

1. Introduction

Much information on the Internet is geographic in nature, or at least contains some locational properties. This is also true for other electronic resources which are created and shared among communities of researchers and educators, for example datasets, methods, articles and so forth.

These locational properties may be subtle or obvious, and may refer to a variety of places in different contexts. For example, a journal article might be written about place A, by people who work at place B, and be downloaded at place C; a geocomputational method might be created for use in place X, by people in place Y and used in Place Z. To begin, we can first divide these locational properties into two groups:

- Implicit Geography: has a map coordinate, we know where the information in the resource refers to exactly.
- Explicit Geography: has no coordinates, but contains place names or other semantic clues which we can extract and geolocate to allow us to map the data we can extract from the resource.

Zong et al. (2005) discuss how the ability to perform query by location can be an important and useful addition to a digital library. They conclude that besides simple queries (which could be string based) there are three benefits:

- Providing the location information of events described by the library contents (documents);
- Enabling a map based visualization of documents;
- Mining spatial knowledge from documents or web pages containing both location and semantic concept information. For example, one may want to find the cluster of web pages related to health care in Minnesota.

This paper considers (1) how to take a collection of text documents or other such resources and extract the implicit geographic locations within them and then (2) how to design and build a system which allows an analyst to visualize, explore and discover new knowledge within the data set based on these uncovered geographic relationships.

2.The Example Problem

A medical researcher may wish to keep abreast of developments in avian flu, she might want to be able to visually explore the academic literature on avian flu, explore this temporally, spatially and by concept as well as combine these publications with other sources of data such as WHO outbreak reports and news reports of outbreaks. If on investigating the documents the analyst discovers some interesting new theory or hypothesis the system she is working with should allow her to explore the hypothesis in greater depth possibly by pulling in more data from other sources, to share this discovery with others and to present the theory to outsiders.

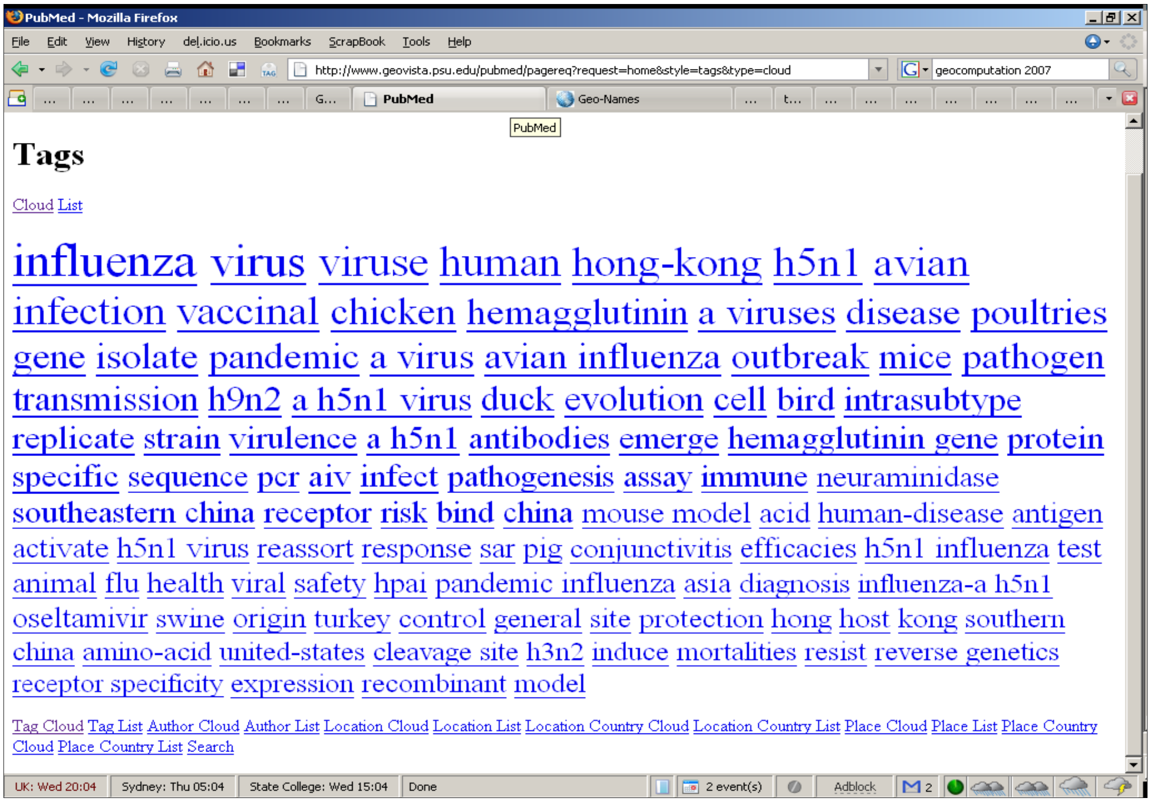
This paper reports on the results of applying the system developed by the authors to a set of 3600 articles related to Avian Flu extracted from the PubMed on line database and the ISI Web of Science database. In each case the bibliographic data, title and abstract of the paper was downloaded and parsed through the system. This provides a large and disparate collection of documents some of which concern places and others which contain no geographic references at all.

3. Building the Dataset

To construct the dataset used for the development of the system a three step process was carried out, which will be described in greater detail in the full paper:

1. **Extract place names from text document using FactXtractor.** FactXtractor is an information extraction web service for Named Entity Recognition (NER) and Entity Relation Extraction (RE) developed at PSU and available at http://julian.mine.nu/snedemo.html. FactXtractor processes a text document using GATE (Cunningham et al., 2002) and identifies entity relations using Stripped Dependency Tree kernels. These results are then returned to the client as an OWL encoded ontology. The ontologies can be viewed using another of our tools—ConceptVista (http://www.geovista.psu.edu/ConceptVISTA).
2. **Disambiguate and geolocate place name**. As each place name is extracted from the document it is returned to the reader which then attempts to disambiguate it and geolocate the resulting place. Disambiguation involves taking a bare place name like 'London' and determining which of the 2683 places called London (http://www.geonames.org/search?q=london) in the GeoNames gazetteer is the correct one. While there is much literature discussing methods and algorithms to carry out this process most assume either that the user has a pre-tagged corpus to train a machine learning method on (Rauch et al., 2003), or that more information than simply London is available to help disambiguate the place name (Amitay et al., 2004). Further details of this technique, and the problems arising are given below. However in the majority of the uses envisaged for the current system the documents under consideration are either very short or contain references to many unrelated places, for example a paper about Avian Flu may refer to Hong Kong, Indonesia and Pennsylvania in the same short abstract if it is discussing the spread of a virus type.
3. **Store article and location in a spatial database to allow further analysis.** Once the system has determined a most likely location for each of the place names in the document, these locations are then stored in a PostGIS database along with any relevant data about the document itself. For instance in the Avian Flu system the authors, the title, abstract, date of publication, journal name and volume are stored in the database to allow users to construct queries about how places and journals or authors are related. While processing the abstracts the system also extracts “keywords” or “tags” from the articles by selecting words (in fact their porter stems) and choosing any word above a specified frequency as important. These keywords were also stored in the database linked to the articles.

Once the service has built up a collection of papers with locations resolved for both the authors' addresses and the places mentioned in the text of the abstract the average user will wish to be able to query the dataset to answer questions about the data. Initially a simple web based interface was constructed to facilitate simple queries. For many of the attributes of the papers stored in the data base a “tag cloud” was constructed that a user could click on to find papers related to that concept, place or author (see fig. 1). Each word is displayed at a size that is proportional to the number of papers in the database that are related to the concept or place. A user can then drill down for more information by following the links on the web page. As figure 2. shows after selecting a keyword (in this case “pig”) the user is presented with a list of all papers that have been tagged with the word pig or pigs, and an overview map that shows their locations.


PubMed - Mozilla Firefox
File Edit View History del.icio.us Bookmarks ScrapBook Tools Help
http://www.geovista.psu.edu/pubmed/pagereq?request=home&style=tags&type=cloud
geocomputation 2007
G... PubMed Geo-Names t...
PubMed
UK: Wed 20:04 Sydney: Thu 05:04 State College: Wed 15:04 Done
2 event(s) Adblock 2

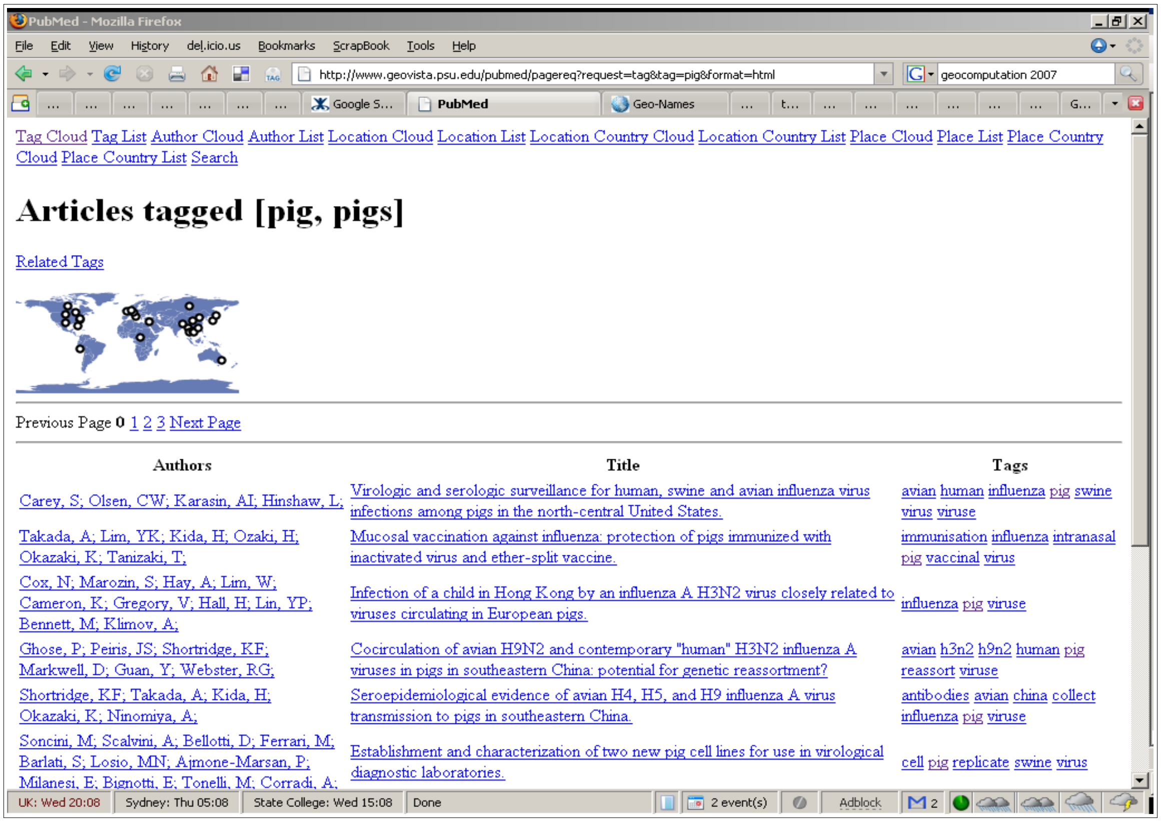
Tags

Cloud List

influenza virus viruse human hong-kong h5n1 avian infection vaccinal chicken hemagglutinin a viruses disease poultries gene isolate pandemic a virus avian influenza outbreak mice pathogen transmission h9n2 a h5n1 virus duck evolution cell bird intrasubtype replicate strain virulence a h5n1 antibodies emerge hemagglutinin gene protein specific sequence pcr aiv infect pathogenesis assay immune neuraminidase southeastern china receptor risk bind china mouse model acid human-disease antigen activate h5n1 virus reassort response sar pig conjunctivitis efficacies h5n1 influenza test animal flu health viral safety hpai pandemic influenza asia diagnosis influenza-a h5n1 oseltamivir swine origin turkey control general site protection hong host kong southern china amino-acid united-states cleavage site h3n2 induce mortalities resist reverse genetics receptor specificity expression recombinant model

Tag Cloud Tag List Author Cloud Author List Location Cloud Location List Location Country Cloud Location Country List Place Cloud Place List Place Country Cloud Place Country List Search

Figure 1: The Tag Cloud for keywords, the larger a word the more papers are associated with it.


PubMed - Mozilla Firefox
File Edit View History del.icio.us Bookmarks ScrapBook Tools Help
http://www.geovista.psu.edu/pubmed/pagereq?request=tag&tag=pig&format=html
geocomputation 2007
Google S... PubMed Geo-Names t... G...
UK: Wed 20:08 Sydney: Thu 05:08 State College: Wed 15:08 Done
2 event(s) Adblock 2

Tag Cloud Tag List Author Cloud Author List Location Cloud Location List Location Country Cloud Location Country List Place Cloud Place List Place Country Cloud Place Country List Search

Articles tagged [pig, pigs]

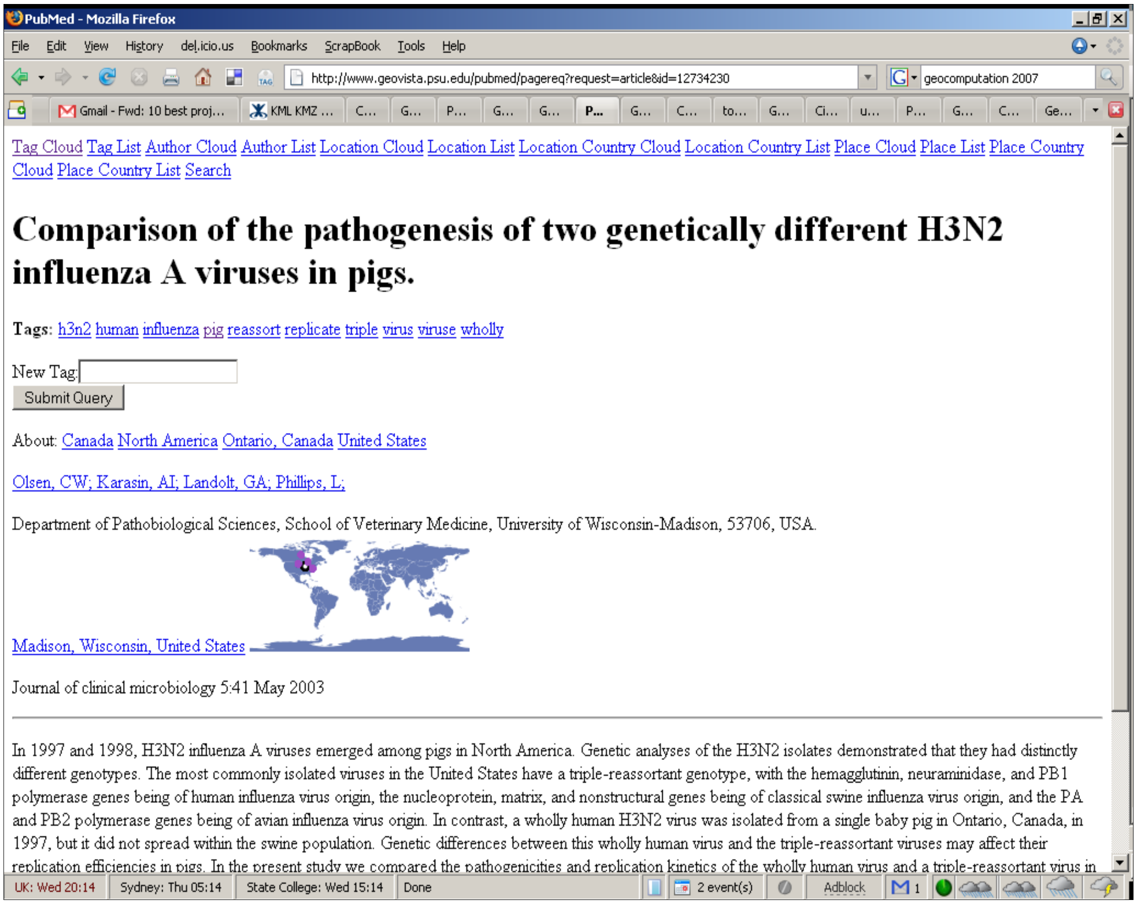
Related Tags

Previous Page **0** 1 2 3 Next Page

Authors	Title	Tags
Carey, S; Olsen, CW; Karasin, AI; Hinshaw, L;	Virologic and serologic surveillance for human, swine and avian influenza virus infections among pigs in the north-central United States.	avian human influenza pig swine virus viruse
Takada, A; Lim, YK; Kida, H; Ozaki, H; Okazaki, K; Tanizaki, T;	Mucosal vaccination against influenza: protection of pigs immunized with inactivated virus and ether-split vaccine.	immunisation influenza intranasal pig vaccinal virus
Cox, N; Marozin, S; Hay, A; Lim, W; Cameron, K; Gregory, V; Hall, H; Lin, YP; Bennett, M; Klimov, A;	Infection of a child in Hong Kong by an influenza A H3N2 virus closely related to viruses circulating in European pigs.	influenza pig viruse
Ghose, P; Peiris, JS; Shortridge, KF; Markwell, D; Guan, Y; Webster, RG;	Cocirculation of avian H9N2 and contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: potential for genetic reassortment?	avian h3n2 h9n2 human pig reassort viruse
Shortridge, KF; Takada, A; Kida, H; Okazaki, K; Ninomiya, A;	Seroepidemiological evidence of avian H4, H5, and H9 influenza A virus transmission to pigs in southeastern China.	antibodies avian china collect influenza pig viruse
Soncini, M; Scalvini, A; Bellotti, D; Ferrari, M; Barlati, S; Losio, MN; Ajmone-Marsan, P; Milanesi, E; Bignotti, E; Tonelli, M; Corradi, A;	Establishment and characterization of two new pig cell lines for use in virological diagnostic laboratories.	cell pig replicate swine virus

Figure 2: A list of papers relating to pigs.

Once a user has found a paper of interest they can again drill down to the paper and see (fig. 3) the data from the database in more detail. Note that the map shows where the paper was written and which places are mentioned in the abstract.


PubMed - Mozilla Firefox
File Edit View History del.icio.us Bookmarks ScrapBook Tools Help
http://www.geovista.psu.edu/pubmed/pagereq?request=article&id=12734230
geocomputation 2007
Gmail - Fwd: 10 best proj... KML KMZ ... C... G... P... G... G... P... G... C... to... G... Ci... u... P... G... C... Ge...
UK: Wed 20:14 Sydney: Thu 05:14 State College: Wed 15:14 Done 2 event(s) Adblock 1

Tag Cloud Tag List Author Cloud Author List Location Cloud Location List Location Country Cloud Location Country List Place Cloud Place List Place Country Cloud Place Country List Search

Comparison of the pathogenesis of two genetically different H3N2 influenza A viruses in pigs.

Tags: h3n2 human influenza pig reassort replicate triple virus viruse wholly

New Tag:

Submit Query

About: Canada North America Ontario, Canada United States

Olsen, CW; Karasin, AI; Landolt, GA; Phillips, L;

Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, 53706, USA.

Madison, Wisconsin, United States

Journal of clinical microbiology 5:41 May 2003

In 1997 and 1998, H3N2 influenza A viruses emerged among pigs in North America. Genetic analyses of the H3N2 isolates demonstrated that they had distinctly different genotypes. The most commonly isolated viruses in the United States have a triple-reassortant genotype, with the hemagglutinin, neuraminidase, and PB1 polymerase genes being of human influenza virus origin, the nucleoprotein, matrix, and nonstructural genes being of classical swine influenza virus origin, and the PA and PB2 polymerase genes being of avian influenza virus origin. In contrast, a wholly human H3N2 virus was isolated from a single baby pig in Ontario, Canada, in 1997, but it did not spread within the swine population. Genetic differences between this wholly human virus and the triple-reassortant viruses may affect their replication efficiencies in pigs. In the present study we compared the pathogenicities and replication kinetics of the wholly human virus and a triple-reassortant virus in

Figure 3: A view of a paper with location map.

Selecting one of the place links will bring back a list of papers either written at that location or about that location allowing the user to explore research teams (this can also be done through cross-author analysis by following the author links) and to find a group of papers about a particular area of interest.

4. Using Ontologies to Build Better Queries.

It is possible to look at the data stored with in the system as an ontology with which users can construct semantic queries. Each location in the GeoNames database can be queried as an RDF object with semantic relationships to other locations such as it's parent, children and neighbours. This allows a user that is interested in papers relating to Pennsylvania to widen their search to either neighbouring states or to the whole of the United States by following a link in the ontology. Alternatively a query about Asia can be narrowed to just some of the countries within Asia to reduce the number of papers returned by a broad search query.

Implicit linguistic knowledge can further be utilized and included within information retrieval by utilizing various semantic lexicons for languages for example WordNet (Miller, G.A., 1995). Queries can be made semantically richer by use of synonym and hyponym relationships which would allow discovery of similar concepts within retrieved information. Further, when the retrieved document space is constrained, the linguistic domain can be expanded by utilizing hypernym and holonym relationships potentially allowing greater information to be extracted. In addition, when the number of documents returned is huge the semantic range can be reduced by utilizing meronymy, thus restricting the information extracted. Figure 4 illustrates a sub set of the different query expansions that could be performed on the keyword “chicken” by using WordNet dictionaries.

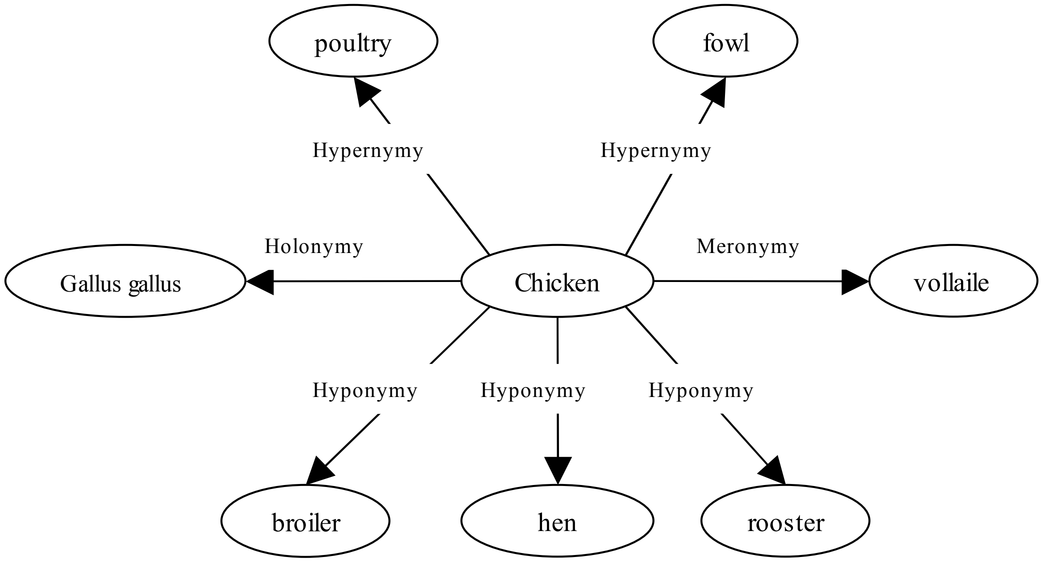


Figure 4: Semantic range expansion/ restriction of a keyword (chicken) using various linguistic relationships by use of the WordNet lexicon.

5. The Disambiguation Algorithm

Amitay et al. (2004) discuss how place names extracted from text documents can be disambiguated by applying a series of heuristic rules which they use to determine the geographic focus of a text document (in their case a web page). They define the two types of ambiguity that can occur in this sort of process as geo/non-geo and geo/geo. A geo/non-geo ambiguity is one where a place name is also a person (Washington) or is a common word (turkey), while a geo/geo ambiguity is where a place name occurs for many distinct places in the world, e.g. London, UK and London, Ontario; Springfield – as in “The Simpsons'” – is also a very popular choice, having been used as a place name in at least 25 US states. In their Web-a-where system Amitay et al. (2004) initially built a dictionary of likely geographic and non-geographic words from the geo/non-geo group of words from a corpus of documents and the number times a word was capitalized, implying it was a proper noun. Then they make use of the fact that when several places are mentioned in a document they are most likely to be near each other to resolve geo/geo ambiguities.

In the system described in this paper we found that the abstracts were too short and often about too many places to be able to apply these and other more complex disambiguation algorithms. Therefore a simpler heuristic was applied which was found to work well in most cases (see fig. 5). It works with the GeoNames feature codes (http://www.geonames.org/export/codes.html) which are unique to each location.

Given two locations A and B:

```
Choose A if A is a Political Entity and B is not,
Choose B if B is a Political Entity and A is not,
Choose A if A is a Region and B is not,
Choose B if B is a Region and A is not,
Choose A if A is an Ocean and B is not,
Choose B if B is an Ocean and A is not,
Choose A if A is a Populated Place and B is not,
Choose B if B is a Populated Place and A is not,
Choose A if A's population is greater than B's,
Choose B if B's population is greater than A's,
Choose A if A is an Administrative Area and B is not,
Choose B if B is an Administrative Area and A is not,
Choose A if A is a Water Feature and B is not,
Choose B if B is a Water Feature and A is not,
Choose A.
```

Figure 5: The Disambiguation Algorithm

The reason for choosing A in the last line is that this makes the comparison stable if used in sort functions; that is, if two places are “equal” they will remain in the same order in the list after sorting as before. For a more specialized use of the system it might be useful to add some extra domain knowledge to this system to bias it towards feature types that are more likely to be mentioned in the particular documents being processed. For example in a geological legend, or field report, one is likely to find names of places where prototypical rock formations appear used as the names of rock categories, in Avian Flu documents, one is likely to find references to different kinds of birds (e.g. turkeys).

6.References

Amitay, E., Har'el, N., Sivan, R. & Soffer, A., 2004. Web-a-where: geotagging web content. SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 273--280.

Cunningham, H., Maynard, D. & Bontcheva, K. & Tablan, V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.. 40th Anniversary Meeting of the Association for Computational Linguistics, .

Miller, G.A. ,1995, WordNet: a lexical database for English. *Communications of the ACM* 38(11),39-41.

Rauch, E., Bukatin, M. & Baker, K., 2003. A confidence-based framework for disambiguating geographic terms. HLT-NAACL 2003 Workshop on Analysis of Geographic References, .

Zong, W., Wu, D., Sun, A., Lim, E. & Goh, D.H., 2005. On assigning place names to geography related web pages. JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 354--362.