Abstract

We present in this paper a method to extract geospatial entities and relationships from the unstructured text of the English language Wikipedia. Using a novel approach that applies SVMs trained from purely structural features of text strings, we extract candidate geospatial entities and relationships. Using a combination of further techniques, along with an external gazetteer, the candidate entities and relationships are disambiguated and the Wikipedia article pages are modified to include the semantic information provided by the extraction process. We successfully extracted location entities with an F-measure of 81%, and location relations with an F-measure of 54%.

Introduction

Wikipedia represents an amazing amount of human knowledge and judgement. However, Wikipedia content remains largely unstructured. Content is marked up for display, but not for direct machine understanding. Article titles, links between articles, and infoboxes are structured enough to directly impart basic information for machine understanding, but the majority of the text is not. As the amount of unstructured user-generated content on the Internet increases, the need to refine methods to extract information from it also increases. Because most user content is not marked up for semantic understanding, and it seems naive to expect users to do the extra work of semantic markup themselves, the challenge of automatically extracting machine-understandable data must be addressed. This paper introduces an approach to extract geospatial entities and relationships from Wikipedia articles, providing a base from which to build software that extracts further information from Wikipedia and other free text, with a vision towards enhancing information retrieval and machine reasoning.

While the DBPedia project (Lehmann et al, 2007) currently extracts limited geospatial data from Wikipedia, it is constrained to the content provided in pre-formatted infoboxes. This paper significantly expands the range of the extraction, processing the full text of each article for semantic information.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Getoor and Viechnicki (2006). Additionally, because the cluster of candidate articles are all related, geospatial information from across the cluster is used to disambiguate individual candidate entities within single articles. Between the gazetteer/geocoder lookup and the disambiguation algorithms, the geospatial entities are reduced to a correct set. This geospatial data can then be used to do a number of things:

- Allow the Wikipedia article to be annotated with RDFS that communicates geospatial semantic information.
- Annotate the link structure of the article cluster itself with relational information to capture geospatial relationships between the articles.
- Inform queries to Wikipedia for geospatial information.
- Allow the cluster of articles to be visualized geospatially.

all providing additional semantic information on top of the already significant informational content of Wikipedia, without demanding extra work from the large population of editors who contribute to Wikipedia.

Geospatial Relationship Extraction

While the extraction of named entities, including locations, is a well-researched field, research into the extraction of relations between entities is an up-and-coming field. Adapting the work of Giuliano, Lavelli and Romano (2007) in using NER to aid relation extraction, the geospatial relationships are extracted using a trained SVM based on the structural feature methods above, with the structure definition based on the work of Herskovitz (1998) on the format of English spatial expressions. Herskovitz has done extensive research into locative expressions, and produced one of the seminal works on the semantics and structure of these expressions. Tagging the previously extracted locations as named entities, the structure of the string around the locations and expression terms is used to classify locative expressions. Once extracted, these expressions are used to provide semantic information relating the entities in the articles.

Results

Using the methods discussed above, we have achieved an F-measure of 81% in extracting and disambiguating geospatial entities. We have further achieved an F-measure of 54% in extracting locative expressions relating the location entities. For training, a corpus was extracted from the Reuters article set and the locative relations and geospatial entities tagged. A set of manually-tagged articles from Wikipedia was used for testing.

As an example, consider the following paragraph (from the World War II Article on Wikipedia). Candidate entities are italic, and those disambiguated correctly by our software are bold. Relationships found by the software are also bold.

The starting date of the war is generally held to be September 1939 with the German invasion of Poland and subsequent declarations of war on Germany by the United Kingdom, France and the British Dominions. However, as a result of other events, many belligerents entered the war before or after this date, during a period which spanned from 1937 to 1941. Amongst these main events are the Marco Polo Bridge Incident, the start of Operation Barbarossa and the attack on Pearl Harbor and British and Netherlands colonies in South East Asia.

Conclusion

This work focused on the extraction and disambiguation of geospatial entities and relations. Leveraging previous work done with SVMs for NER using structural features, and geospatial reference disambiguation, we provide advances in accuracy in this area leading to techniques and tools that allow for the extraction, processing and visualization of geospatial attributes embedded in the unstructured text of user-generated media. We can also annotate the existing textual information with semantic markup that communicates semantic information and relationships.

References


Habib, M. 2008. Improving scalability of support vector machines for biomedical named entity recognition. Computer Science Department, University of Colorado at Colorado Springs.


