Gazetteer Linkage to WordNet

Beth M. Sundheim
SPAWAR Systems Center
beth.sundheim@navy.mil

Scott Mardis, John Burger
The MITRE Corporation
mardis@mitre.org, john@mitre.org

Abstract

One of the new WordNet features to be found in version 2.1 is the *instance* relation, which replaces the hypernym relation for noun synsets that denote instances rather than types (Miller and Hristea, forthcoming). The creation of this distinction serendipitously coincided with project work at SPAWAR Systems Center (SSC) and the MITRE Corporation to produce a tailored gazetteer database of place names for use in research on question answering by participants in a U.S. government-sponsored research program (Irie and Sundheim 2004). The millions of place names contained in this database, called the Integrated Gazetteer Database (IGDB) (Mardis and Burger), are drawn from publicly available sources provided by the National Geospatial-Intelligence Agency, the U.S. Geological Survey, the CIA World Factbook, and the Tipster Text research program. The IGDB project includes a task that is being carried out in collaboration with Princeton University to incorporate the instance synsets that define places into the database as an additional source of gazetteer information.

At the completion of this project, IGDB users should have access to the WordNet location information and WordNet users should have access to related IGDB data. Also, a process should be in place to facilitate updating the IGDB as WordNet continues to evolve. While there are only approximately 3,000 place instance synsets in WordNet, there are a number of benefits to be gained by this combination of resources. There is partial overlap between the resources both in the set of places defined as well as the type of information provided. Examining the places and attributes they share provides a general strategy for defining a general map between the two resources.

Many of the places defined in WordNet are of a sort not well represented in traditional gazetteers. Examples include unofficially named regions (such as *Middle East*, *New England* and *The South*) and other hard-to-define areas (such as large natural features: seas, mountain ranges, rivers). On the other hand, gazetteers such as the IGDB contain entries for millions of officially named places, and it would be a advantageous if the need for creating synsets for those places were eliminated by enabling WordNet users to access the gazetteer entries as a supplement to WordNet.

We originally viewed WordNet as completely separate from the IGDB, with cross-references between the two resources establishing the equivalence between synsets and gazetteer-defined places. More recently, the view has shifted to capturing the gazetteer-relevant synsets as gazetteer entries themselves. This enables us to develop the linkages between WordNet places and ones from other sources as part of the general problem of representing inter-gazetteer entry coreference. Having a general coreference representation in the IGDB, we would be able to respond to a user's query for information on a particular place name with one composite set of information per place, rather than one set of information for each source of information.

For example, *Frisian Islands* is found in both WordNet and the current IGDB. The WordNet entry makes it clear that the place defined by that synset is specifically an archipelago, while the entry in the current IGDB classifies it as ISLS (*islands*), defined more generally as *tracts of land, smaller than a continent, surrounded by water at high tide*. The WordNet entry also explicitly identifies the part-whole relation between the islands and the countries of Netherlands, Germany, and Denmark, a fact which is also supported by the

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 103–104.

current IGDB. The current IGDB provides additional name variants for the place (*Friesische Inseln*, *Friesian Islands*, *Friese Eilanden*), as well as latitude/longitude and a variety of other information.

In other cases, places identified in WordNet are not represented in the IGDB. This often happens with places that are composites of politically-defined entities, such as *the Carolinas* (North and South Carolina together) or *the Gulf States* (the U.S. states bordering the Gulf of Mexico); or physically-defined regions such as the *the Upper Peninsula* (the region of Michigan between Lake Superior and Lake Michigan). If a WordNet entry has meronyms, corresponding IGDB entries can be successfully inferred by comparing meronyms with the containment (part-of) relations in the IGDB. If an entry has only a holonym relation, its meaning is less clear, though it may still be possible to place it in the IGDB containment hierarchy. For some of the latter cases, it may be beneficial for the WordNet developers to consider adding appropriate meronym entries. A summary of some possible heuristics useful when there is no direct correspondence in the IGDB appears below:

- *Upper Peninsula***:** Holonym in WordNet (*Michigan*) can be coreferenced with an IGDB entry.
- *Carolina***:** Holonym ($South_1$) cannot be coreferenced. There are no explicit meronyms, but some are indicated in the gloss.
- *Gulf States***:** Holonym ($South_1$) cannot be coreferenced. There is a full set of meronyms, however, that can be coreferenced (*Alabama*, etc.).

A database of WordNet synsets and annotations documenting their expected linkage to specific IGDB entries was created as part of a study of linkage issues for all WordNet places. A summary report was completed in February 2005. Particular classes of cases and instances that warrant specific attention are being reviewed. The data can also be used to support development of automated methods to perform intergazetteer coreference.

Some issues identified thus far in the review process concern the representation of particular entries in WordNet:

- Some synsets ambiguously refer to both modern and ancient places.
- Some "instance" labels may be in error, and some may be missing.

More thorough/consistent treatment of aggregate/area term definitions and meronym (part-whole/member-of) relations may be needed. There are also more general issues in the design and use of the IGDB that we hope to address as development continues.

References

Scott Mardis and John Burger. *Design for an Integrated Gazetteer Database*. MITRE Technical Report 05B0000085.

Robert Irie and Beth Sundheim (2004). Resources for place name analysis. *Proceedings of the 2004 Language Resources and Evaluation Conference*, pp. 317–320.

George A. Miller and Florentina Hristea (forthcoming). WordNet nouns: classes and instances. *Computational Linguistics*.