Geo-WordNet: Automatic Georeferencing of WordNet

Davide Buscaldi, Paolo Rosso

Natural Language Engineering Lab, DSIC
Universidad Politécnica de Valencia, Spain
{dbuscaldi,prosso}@dsic.upv.es

Abstract

WordNet has been used extensively as a resource for the Word Sense Disambiguation (WSD) task, both as a sense inventory and a repository of semantic relationships. Recently, we investigated the possibility to use it as a resource for the Geographical Information Retrieval task, more specifically for the toponym disambiguation task, which could be considered a specialization of WSD. We found that it would be very useful to assign to geographical entities in WordNet their coordinates, especially in order to implement geometric shape-based disambiguation methods. This paper presents Geo-WordNet, an automatic annotation of WordNet with geographical coordinates. The annotation has been carried out by extracting geographical synsets from WordNet, together with their holonyms and hypernyms, and comparing them to the entries in the Wikipedia-World geographical database. A weight was calculated for each of the candidate annotations, on the basis of matches found between the database entries and synset gloss, holonyms and hypernyms. The resulting resource may be used in Geographical Information Retrieval related tasks, especially for toponym disambiguation.

1. Introduction

Almost all the information available on the web contain some kind of geographical reference. In the last few years, researchers have shown a growing interest in the automatic processing of geographical information in text, whereas in the past the handling of geographic information has been based largely on the highly structured map-based representation of space that are used in most Geographical Information Systems (GIS). Geographical Information Retrieval (GIR) is a relatively new research field in Information Retrieval (IR). The technical challenges in GIR are mostly related to the problems of associating a toponym (i.e., a geographical name) to its actual coordinates in a map. Conventional search engines do not make distinctions between toponyms and other classes of words. Therefore, they are able to retrieve only the documents where a query toponym is matched exactly, and not those documents containing alternative versions of the query toponym, nearby places or even within the query toponym itself. Moreover, geographical knowledge often remains implicit in texts: for instance, if we name "Marrakech" in a text, the fact that Marrakech is a city in Morocco is not usually mentioned.

The efforts of the research community in GIR are evidenced by the creation of the GIR series of workshops, that are being held every year since 2004 at important conferences such as ACL SIGIR and CIKM, and the GeoCLEF[1] task at CLEF. These events collect state-of-the-art contributions and constitute a common framework for the comparison of the latest systems and techniques.

Our previous research in GIR has focused until now in the use of the WordNet ontology in order to address the different problems that arise in GIR (Buscaldi et al., 2006b; Buscaldi et al., 2007; Buscaldi et al., 2006a). WordNet was developed at Princeton University as a complex lexical database of general English (Miller, 1995). It contains concepts (*synsets* or sets of synonyms) connected by different conceptual relationships. These relationships can be used to find the geographical information related to a specific toponym. For instance, synonymy allows to identify alternative place names and holonymy (*part-of*) allows to find the entities containing the toponym.

Ambiguity of toponyms is a common problem in GIR (Leidner, 2004). Garbin and Mani (Garbin and Mani, 2005) found that 67.82% of the toponyms found in a corpus that were ambiguous lacked a local discriminator in the text. Overell and Rüger (Overell and Rüger, 2008) showed that high accuracy in the disambiguation of toponyms allows to improve results in GIR. Our experiments with a conceptual density-based disambiguation method (Buscaldi and Rosso, 2008) show that WordNet can also be used in order to address the toponym ambiguity problem.

Unfortunately, WordNet presents some problems as a geographical information resource. First of all, the quantity of geographical information is quite small especially if compared with some of the most known gazetteers. We estimated the number of geographical entities stored in WordNet by means the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. Geographical resources like gazetteers usually contain a much greater quantity of information. For instance, the Geonet Names Server[2] (GNS) contains more than 5 million of place names.

The second problem is that WordNet is not *georeferenced*, that is, the toponyms are not assigned their actual coordinates on earth. Georeferencing WordNet can be useful for many reasons: first of all, it is possible to establish a semantics for synsets that is not vinculated only to a written description (the synset *gloss*, e.g.: "Marrakech, a city in western Morocco; tourist center"). In second place, it can be useful in order to enrich WordNet with other resources; finally, it can improve its effectiveness as a geographical information resource: for instance, it will allow to evaluate some toponym disambiguation methods based on geographical coordinates (Smith and Mann, 2003; Woodruff

[1] http://ir.shef.ac.uk/geoclef/

[2] http://earth-info.nga.mil/gns/html/index.html

and Plaunt, 1994).

In this paper we present *Geo-WordNet*, an extension of WordNet that is based on the automatic georeferencing of its synsets. In the following sections we present the resources we used and describe the annotation process in detail.

2. Resources

Gazzetteers are the main sources of geographical coordinates. The GNS, introduced in the previous section, and the Geographic Names Information System (GNIS) gazetteers cover almost any place on earth. The GNIS contains data of U.S. only and is mantained by the U.S. Geological Survey[3], whereas the GNS contains data from any other country in the world and is mantained by the National Geospatial-Intelligence Agency. The coverage of these resources can be observed in Figure 1 (white dots correspond to covered places).

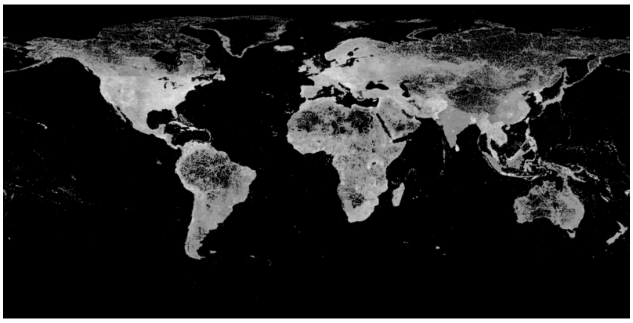


Figure 1: Place coverage provided by the GNS and GNIS gazetteers.

Such a great coverage presented a problem for our task. In fact, by using these resources, it is more probable to find ambiguities between place names and other kind of names (*geo - non geo* ambiguities), such as "church" vs. "Church" (a place in Lancashire, UK), but also between place names. GNS and GNIS together have a mean ambiguity of 4.4 senses per name (Volz et al., 2007).

Therefore, we considered a smaller resource, the Wikipedia-World (WW) project[4] database (in SQL). The coverage of this resource is considerably smaller than the one offered by GNS and GNIS, as it can be observed in Figure 2.

The covered areas are almost the same of the places in WordNet, as it can be observed by comparing Figure 3 to Figure 2.

The WordNet ontology is particularly rich in semantic relationships that connect its *synsets* (i.e., the senses). In this particular case, we are interested in the *hypernymy* (or *is-a* relationship) and the *holonymy* (or *part-of* relationship). For place names, hypernymy allows to find the *class* of a given name (although this has changed with the version 3.0 of WordNet, which introduced the *instance_of* relationship

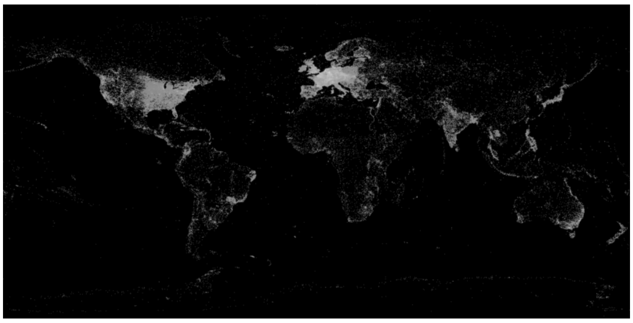


Figure 2: Place coverage provided by the Wikipedia World database.

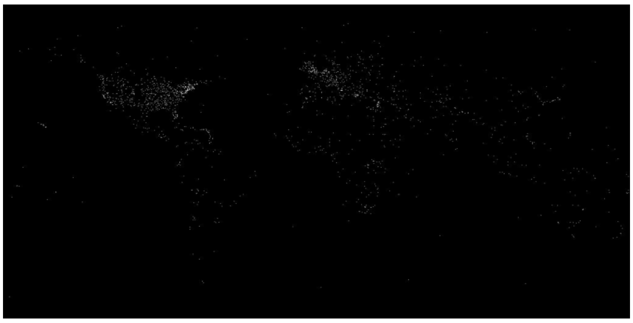


Figure 3: Place coverage provided by WordNet.

- in this work we used WordNet 2.0). For instance, the hypernym of "Armenia" is "country", "Mount St. Helens" is a "volcano". Holonymy can be used to find a geographical entity that contains a given place, such as "Washington (U.S. state)" that is holonym of "Mount St. Helens".

3. Automatic Referencing

The heuristic we developed is pretty simple and is based on contributions from the following components:

- Match between a synset wordform and a database entry;
- Match between the holonym of a geographical synset and the containing entity of the database entry;
- Match between a second level holonym and a second level containing entity in the database;
- Match between holonyms and containing entities at different levels (0.5 weight); this corresponds to a case in which WordNet or the WW lacks the information about the first level containing entity.
- Match between the hypernym and the class of the entry in the database (0.5 weight);
- A class of the database entry is found in the gloss (i.e. the description) of the synset (0.1 weight).

The reduced weights were introduced in the cases we observed that a match could lead either to a correct or wrong assignment. This is true especially for gloss comparison, since WordNet glosses usually include example sentences

[3] http://geonames.usgs.gov

[4] http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en

that are not related with the definition of the synset, but instead provide a “use case” example.

The algorithm is as follows:

1. Pick a synset s in WordNet and extract all of its wordforms $w_1, \ldots, w_n$ (i.e., the name and its synonyms)
2. Check whether a wordform w_i is in the WW database
3. If w_i appears in WW: find the holonym h_s of the synset s. Else: goto 1.
4. If $h_s = \{\}$: goto 1. Else: find the holonym h_{hs} of h_s
5. Find the hypernym H_s of the synset s.
6. $L = \{l_1, \ldots, l_m\}$ is the set of locations in WW that correspond to the synset s
7. A weight is assigned to each l_i depending on the weighting function f
8. The coordinates related to $\max_{l_i \in L} f(l_i)$ are assigned to the synset s
9. Repeat until the last synset in WordNet

A final step was carried out manually and consisted in reviewing the labeled synsets, removing those which were mistakenly identified as locations.

The weighting function is defined as:

$$\begin{aligned} f(l) = m(w_i, l) + m(h_s, c(l)) + m(h(h_s), c(c(l))) + \\ +0.5 \cdot m(h_s, c(c(l))) + 0.5 \cdot m(h(h_s), c(l)) + \\ +0.1 \cdot g(D(l)) + 0.5 \cdot m(H_s, D(l)) \end{aligned}$$

where $m : \Sigma^* \times \Sigma^* \rightarrow \{1, 0\}$ is a function returning 1 if the string x matches l from the beginning to the end or from the beginning to a comma, and 0 in the other cases. $c(x)$ returns the containing entity of x, for instance it can be $c(\text{“}Abilene\text{”}) = \text{“}Texas\text{”}$ and $c(\text{“}Texas\text{”}) = \text{“}US\text{”}$. In a similar way, $h(x)$ retrieves the holonym of (x) in WordNet. $D(x)$ returns the class of location x in the database (e.g. a mountain, a city, an island, etc.). $g : \Sigma^* \rightarrow \{1, 0\}$ returns 1 if the string is contained in the gloss of synset s. Country names obtain an extra $+1$ if they match with the database entry name and the country code in the database is the same as the country name.

Example

For instance, consider the following synset from WordNet: *(n) Abilene (a city in central Texas)*; in Figure 4 we can see its first level and second level holonyms (“Texas” and “USA”, respecrively) and its direct hypernym (“city”).

The search in the WW database returns the results in Figure 5. The fields have the following meanings: *Titel_en* is the English name of the place, *lat* is the latitude, *lon* the longitude, *country* is the country the place belongs to, *subregion* is an administrative division of a lower level than country.

Subregion and *country* fields are processed as first level and second level containing entities, respectively. In the case the *subregion* field is empty, we use the specialization in the *Titel_en* field as first level containing entity. Note that styles fields (in this example *city k* and *city e*) were normalized to fit with WordNet classes. In this case, we transformed *city k* and *city e* into *city*. The calculated weights can be observed in Table 1.

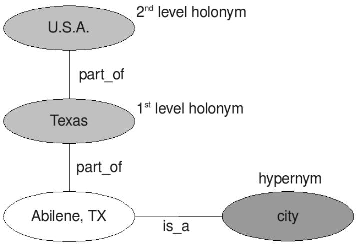
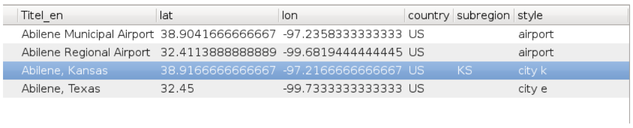


Figure 4: Portion of WordNet related to the *Abilene* example.



	Titel_en	lat	lon	country	subregion	style
	Abilene Municipal Airport	38.9041666666667	-97.2358333333333	US		airport
	Abilene Regional Airport	32.4113888888889	-99.6819444444445	US		airport
	Abilene, Kansas	38.9166666666667	-97.2166666666667	US	KS	city k
	Abilene, Texas	32.45	-99.7333333333333	US		city e

Figure 5: Results of the query SELECT Titel_en, lat, lon, country, subregion, style FROM pub_CSV_test3 WHERE Titel_en like ``Abilene%" on the WW database.

Entity	Weight
Abilene Municipal Airport	1.0
Abilene Regional Airport	1.0
Abilene, Kansas	2.0
Abilene, Texas	3.6

Table 1: Resulting weights for the “Abilene” example.

The weight of the two airports derive from the match for “US” as the second level containing entity ($m(h(h_s), c(c(l))) = 1$). “Abilene, Kansas” benefits also from an exact name match ($m(w_i, l) = 1$). The highest weight is obtained for “Abilene, Texas” since there are the same matches as before, but also they share the same containing entity ($m(h_s, c(l)) = 1$) and there are matches in the class part both in gloss (a *city* in central Texas) and in the direct hypernym.

4. Geo-WordNet

The final resource is constituted by two plain text files: the most important is a single text file that contains 2,012 labeled synsets, where each row is constituted by an offset (WordNet version 2.0) together with its latitude and longitude, separated by tabs. This file is named `WNCoord.dat`. A small sample of the content of this file can be found in Figure 6.

The other file contains a human-readable version of the database, where each line contains the synset description and the entry in the database: *Acapulco a port and fashionable resort city on the Pacific coast of southern Mexico; known for beaches and water sports (including cliff diving)*

```
08294059 7.06666666667 171.266666667
08294488 9.19388888889 167.459722222
08294965 -7.475 178.005555556
```

Figure 6: Portion of the resulting resource, corresponding to the following synsets: *Marshall Islands*, *Kwajalein* and *Tuvalu*.

('Acapulco', 16.851666666666699, -99.9097222222222, 'MX', 'GRO', 'city c').
The resource is available from the 'downloads' section of the Natural Language Engineering Lab website: *http://www.dsic.upv.es/grupos/nle*.

5. Conclusions

We created an expansion for the WordNet ontology, especially aimed to researchers working on toponym disambiguation and in the Geographical Information Retrieval field. The heuristic used was particularly simple. Some errors were manually removed at the end of the automatic process. We hope that this resouce will reveal itself a valuable one. We are currently planning to extend the experiments presented in (Buscaldi and Rosso, 2008) with algorithms that take into account coordinates and geometrical methods in order to carry out an exhaustive review of currently available toponym disambiguation methods.

Acknowledgements

We would like to thank the TIN2006-15265-C06-04 and PCI-AECI A/7067/06 research projects for partially supporting this work. Figures 1 and 2 courtesy from *http://commons.wikimedia.org*.

6. References

Davide Buscaldi and Paolo Rosso. 2008. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 22(3):301–313.

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. 2006a. Using the wordnet ontology in the geoclef geographical information retrieval task. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, Maarten de Rijke, and Danilo Giampiccolo, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 939–946. Springer, Berlin.

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. 2006b. Wordnet as a geographical information resource. In *3rd Global WordNet Conference (GWC06)*, pages 299–304, Cheju, South Korea.

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. 2007. A wordnet-based indexing technique for geographical information retrieval. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 954–957. Springer.

Eric Garbin and Inderjeet Mani. 2005. Disambiguating toponyms in news. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Jochen L. Leidner. 2004. Toponym resolution in text:" which sheffield is it?". In *Proceedings of the the 27th Annual International ACM SIGIR Conference (SIGIR 2004)*, pages 602–606, Sheffield, UK. ACM Press.

George A. Miller. 1995. WordNet: A Lexical Database for English. In *Communications of the ACM*, volume 38, pages 39–41.

Simon Overell and Stefan Rüger. 2008. Using cooccurrence models for placename disambiguation. *International Journal of Geographical Information Systems*, 22(3):265–287.

David A. Smith and Gideon S. Mann. 2003. Bootstrapping toponym classifiers. In *HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 45–49, Morristown, NJ, USA. Association for Computational Linguistics.

Rafael Volz, Joachim Kleb, and Wolfgang Mueller. 2007. Towards ontology-based disambiguation of geographical identifiers. In *16th International World Wide Web Conference (WWW2007)*, Banff, Alberta, Canada.

Allison Woodruff and Christian Plaunt. 1994. Gipsy: Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9):645–655.