

# Approaches to Disambiguating Toponyms

Davide Buscaldi

Laboratoire d'Informatique Fondamentale d'Orléans

Université d'Orléans

Orléans, France

*davide.buscaldi@univ-orleans.fr*

## Approaches

Many approaches have been proposed in recent years in the context of Geographic Information Retrieval (GIR), mostly in order to deal with geographically constrained information in un-structured texts. Most of these approaches share a common scheme: in order to disambiguate a toponym  $t$  with  $n$  possible referents in a document  $d$ , they find a certain number of context toponyms  $c_0, \dots, c_k$  that are contained in  $d$ . A score for each referent is calculated according to the context toponyms, and the referent with the highest score is selected. According to the method used to calculate the score, Toponym Disambiguation (TD) methods may be grouped into three main categories, as proposed by [7]:

- *map-based*: methods that use an explicit representation of toponyms on a map, for instance to calculate the average distance of unambiguous context toponyms from referents;
- *knowledge-based*: methods that exploit external knowledge sources such as gazetteers, Wikipedia or ontologies to find disambiguation clues;
- *data-driven* or *supervised*: methods based on machine learning techniques.

Map-based methods usually do not need any information other than the coordinates of the places appearing in context. These methods are usually very sensitive to changes in context; therefore, it is necessary to remove places that are very far on average from the others [14] from the context, or to include external knowledge, such as the position of the “source” of a text [6]: for instance, if the toponym “Paris” is found in a Texas-based newspaper, it is more likely that it is referring to “Paris, TX” rather than Paris in France. The “source” of information is an important disambiguation feature especially for local text collections: in [6], it has been shown that 76.2% of the places mentioned in an Italian newspaper are located within 400km of the city where the newspaper is published.

The size of a location, for example measured through the number of inhabitants living in that place, is an important clue in knowledge-based methods: more populated places are more likely to be mentioned. Population represents a good rule-of-thumb if other context information is not available. This heuristic was included in the methods of [12, 1, 2]. Structural information derived from the containment, or *part-of*, relationship was also used to develop methods that are based on the idea that the places in the context are usually contained in the same region or geographical area. Some examples of hierarchy-based algorithms are [3, 7]. Wikipedia was also successfully used by [11], who took advantage of article templates, categories and referents (links to other articles in Wikipedia).

Data-driven methods are not commonly used in TD, mostly because of the lack of geographically tagged data and the problem in the classification of unseen toponyms. Nevertheless, supervised classifiers has been implemented with mixed results. The advantage

of these methods is that they can exploit non-geographical content, such as in the work of [13], where events are used to build a probabilistic model, using the spatial relationships between non-geographical entities and places; for instance, if some known person or organisation are based at a place, their presence in the context of the toponym may represent an important clue (for instance, “Google” in the context of “Mountain View” may suggest that it is “Mountain View, CA” rather than “Mountain View” in Arkansas).

### **Toponym Ambiguity**

The ambiguity of a given toponym depends strictly on the specific resource used to represent the world (usually a *gazetteer* or an *ontology*): these resources may be more or less detailed, implying that the number of referents for a toponym vary greatly from a resource to another. For instance, there are two cities named “Cambridge” in the world according to WordNet, 38 according to Yahoo! GeoPlanet and 40 according to Geonames. Gazetteers and geographical ontologies may be characterised according to their *scope*, *coverage* and *detail* or granularity. The scope of a geographic resource indicates whether a resource is limited to a region or a country (for instance, the scope of Ordnance Survey data is limited to Great Britain), or it is a resource covering all the parts of the world. Coverage is determined by the number of toponyms listed in the resource. Detail is related to how fine-grained is the resource with respect to the area covered.

Using one resource rather than another affects not only the degree of ambiguity of toponyms, but also the choice of the disambiguation method that can be used. With an highly detailed resource, the chances of finding an ambiguity within a single state or region increase; therefore, the methods that relies on the containment relationship to resolve ambiguities may be negatively affected by the use of a highly detailed resource. While the resource determines the disambiguation method, the task determines the resource to be used. In [6], the objective was to disambiguate toponyms in a local Italian newspaper, where the granularity of toponyms was intended to be at the level of street names. These requirements made it necessary to create a custom resource by merging Geonames data with data extracted from the Google Maps API, because there was no freely available resource listing street names together with city, region and country names. Street name ambiguities may frequently occur at the province level, with the result that it was impossible to apply disambiguation methods based on hierarchies.

Some classes of toponyms are more ambiguous, on average, than others. In [6] it has been shown that street names are 25% more ambiguous than city names. It can be assumed that the average degree of ambiguity is proportional to the depth of a place in a geographical ontology. This reflects the fact that people usually see the world at a level of detail that decreases with distance (the “Steinberg hypothesis” by [11]).

### **Evaluation**

The lack of a rigid evaluation procedure and a standardised test-set has represented a major issue in the evaluation of toponym disambiguation methods. Most of the initial works published calculated their results under particular conditions and using small test-sets that makes reproduction of their results difficult. In some cases, the same method applied to different test-sets obtained very different results: [15] report precisions between 21% and 87%, depending on the test collection used. Therefore, in recent years there has been an effort to produce test collections specifically aimed to the evaluation of toponym disambiguation methods. In these test collections, toponyms have been manually labelled with the correct referent. These resources include the the TR-CoNLL corpus [8], GeoSemCor [7], the SpatialML [9] and the LGL [10] corpora.

The TR-CoNLL corpus consists in a collection of documents from the Reuters news agency labelled with toponym referents. A set of 946 documents was manually annotated with co-

ordinates from a custom gazetteer derived from Geonames. The resulting resource contains 6,980 toponym instances. GeoSemCor was obtained from SemCor, the most used corpus for the evaluation of WSD methods. SemCor is a collection of texts extracted from the Brown Corpus of American English, where each word has been labelled with a WordNet sense (synset). It contains 1,210 toponym instances in its final version. The SpatialML corpus is a manually tagged collection of documents drawn mainly from broadcast news, newsgroups, and weblogs, which contains 4,783 toponyms instances. LGL (Local/Global Lexicon) is a corpus of 588 articles collected from 78 different data sources, containing 4,793 toponyms. This corpus is focused on smaller newspapers with a localised audience, which have a large presence on the Internet, in contrast with the other resources which have been derived from news wires, which is usually written for a broadly distributed geographical audience.

The primary measures used for the evaluation of toponym disambiguation are *precision* and *recall*, where precision is calculated as the number of correctly disambiguated toponyms divided by the number of disambiguated toponyms, and recall is calculated as the number of correctly disambiguated toponyms divided by the number of toponyms in the test collection. Correctness is determined by comparing the location ID assigned by the disambiguation process to the location ID label contained in the resource. A fuzzy criterion for correctness could use the inverse of the distance between the assigned location and the correct one. This criterion reduces the importance of errors such as mismatching the province with the city of the same name [10] though problems still exist since point-based geometry assigns a single location (for example a centroid) irrespective of the spatial footprint related to a toponym.

### **Areas with Success and Failure**

In the last years, the relationship between Toponym Disambiguation and Information Retrieval has been object of many studies. This is also a result of the introduction in 2005 of the GeoCLEF evaluation campaign (see also the notes from Mandl and Cardoso in this issue). [17] carried out experiments on a reduced document set from GeoCLEF and showed that halving the disambiguation accuracy appears to have a greater negative effect on MAP score than halving NERC Recall. In [5] an experiment using the whole collection showed that disambiguation improved results only in the case of short queries and with a relatively detailed geographical resource (GeoNames) was used, a configuration more similar to Web retrieval than to the IR ad-hoc task. Recent work by [16] shows an improvement in the retrieval performance when applying TD to IR over geological texts.

Some tasks where toponym disambiguation has as of yet failed to prove useful are the scope resolution task and Question Answering (QA). In the first case, [2] observed that content-based features, such as person names, are more useful for the resolution of geographical scopes than the disambiguation of ambiguous toponyms in text. In [5], the experiments in QA showed that the errors derived by the wrong labelling of ambiguous toponyms do not alter significantly the process of answer extraction.

### **Conclusions**

In the last decade, toponym disambiguation has progressively gained importance within the field of Natural Language Processing. The number of works dealing with this task has been increasing: some have proposed new methods and compared them with existing ones, others have dealt with the production of evaluation frameworks, and finally some areas with success have appeared. However, there is still room for further work: toponym disambiguation may be used to tag ambiguous toponyms in Web documents in order to compensate the “lack of explicit spatial knowledge within the Web” [4]. Moreover, given that the average performance of existing methods usually do not exceed 85% in accuracy, and it may vary depending on the collection used and the final task, researchers may be

interested in developing new methods or adapting the existing ones to new tasks.

## References

- [1] E. Amitay, N. Harel, R. Sivan, and A. Soffer, 2004. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, Sheffield, UK, 2004.
- [2] G. Andogah, 2010. *Geographically Constrained Information Retrieval*. PhD thesis, University of Groningen.
- [3] I. Bensalem and M.-K. Kholadi, 2010. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6(6):653–659.
- [4] S. Boll, C. Jones, E. Kansa, P. Kishor, M. Naaman, R. Purves, A. Scharl, and E. Wilde, 2008. Location and the web (LocWeb 2008). In *Proceeding of the 17th international conference on World Wide Web, WWW '08*, pages 1261–1262, New York, NY, USA, 2008. ACM.
- [5] D. Buscaldi, 2010. *Toponym Disambiguation in Information Retrieval*. PhD thesis, Universidad Politécnic de Valencia.
- [6] D. Buscaldi and B. Magnini, 2010. Grounding toponyms in an italian local news corpus. In *Proceedings of GIR'10 Workshop on Geographical Information Retrieval*.
- [7] D. Buscaldi and P. Rosso, 2008. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 22(3):301–313.
- [8] J. L. Leidner, July 2006. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417.
- [9] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner, 2008. SpatialML: Annotation Scheme, Corpora, and Tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.
- [10] J. S. Michael D. Lieberman, Hanan Samet, 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE'10)*, pages 201–212.
- [11] S. Overell, 2009. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. PhD thesis, Imperial College London.
- [12] E. Rauch, M. Bukatin, and K. Baker, 2003. A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–54, Edmonton, Alberta, Canada, 2003.
- [13] K. Roberts, C. A. Bejan, and S. Harabagiu, 2010. Toponym disambiguation using events. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*.
- [14] D. A. Smith and G. Crane, 2001. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, volume 2163 of *Lecture Notes in Computer Science*, pages 127–137. Springer, Berlin.
- [15] D. A. Smith and G. S. Mann, 2003. Bootstrapping toponym classifiers. In *HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 45–49, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [16] N. V. Sobhana, A. Barua, M. Das, P. Mitra, and S. K. Ghosh, 2010. Co-occurrence based place name disambiguation and its application to retrieval of geological text. In *Recent Trends in Networks and Communications*, volume 90 of *Communications in Computer and Information Science*, pages 543–552. Springer Berlin Heidelberg.
- [17] N. Stokes, Y. Li, A. Moffat, and J. Rong, 2008. An empirical study of the effects of nlp components on geographic ir performance. *International Journal of Geographical Information Science*, 22(3):247–264.