Toponym Resolution in Text: "Which Sheffield is it?"

Jochen L. Leidner
University of Edinburgh, School of Informatics,
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland, UK.
jochen.leidner@ed.ac.uk

ABSTRACT

Named entity tagging comprises the sub-tasks of identifying a text span and classifying it, but this view ignores the relationship between the entities and the world. Spatial and temporal entities ground events in space-time, and this relationship is vital for applications such as question answering and event tracking. There is much recent work regarding the temporal dimension [13, 10], but no extensive study of the spatial dimension.

I propose to investigate how spatial named entities (which are often referentially ambiguous) can be automatically resolved with respect to an extensional coordinate model (*toponym resolution*), using hybrid heuristic/statistical methods. The major contributions of this research project are a *corpus* of text manually annotated for spatial named entities with their model correlates as a training/evaluation resource [4] and a novel *method to spatially ground toponyms* in text.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.2.8 [**Database Applications**]: Spatial databases and GIS

General Terms

Spatial indexing and retrieval; toponym resolution; disambiguation of place-names

Keywords

Geocoding; geoparsing; georeferencing; place-name disambiguation; spatial retrieval; geographic IR

1. INTRODUCTION

The task of annotating any text with flat (unstructured) named entity annotation of unseen text has recently been successfully automated, achieving near-human performance using machine learning [19]. But for many applications, such as automatic question answering, geographic information retrieval or map generation, the connection between the classified named entity and a *world model* is important. A text string, say "Sheffield" can be distinguished from different text strings (say "London"), but no spatial reasoning can be performed on the basis of strings only. If we want to retrieve a list of events that happened in Britain, or know what the distance is between Sheffield and London, we need to extend NERC (Named Entity Recognition and Classification) to NERC+R (NERC with Resolution) i.e., we must additionally relate the named entities to a correlate in a model of the world:

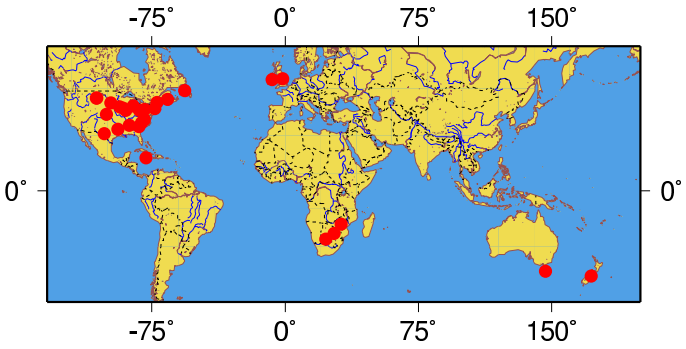

Figure 1: Potential referents for a mention of "Sheffield": but *which* Sheffield is intended?

$$\texttt{Sheffield} \mapsto \langle 53.36; -1.50 \rangle.^{1}$$

This project proposes to develop novel methods to carry out this mapping automatically. Once the mapping is performed, some limited reasoning is possible (such as utilizing entailment relationships: everything that happens in Sheffield also happens in Britain). But attempting to resolve names of places, we are confronted us with the problem of *referential ambiguity*: the mapping is not trivial, as there exist more than one referent with the name "Sheffield" in the world (Figure 1 contains 28 geographic features named "Sheffield" from Canada to New Zealand; most are located in the USA). Humans are very good at determining from context which is the intended referent.[2] But for machines, a mechanical procedure needs to be devised to compute the

[1]We use latitudes/longitudes in decimals rather than degrees here.

[2]Except for the occasional disaster: the author has been confided to various stories of flights taken to the "wrong

referent for each toponym. We call this processing step *toponym resolution*.

2. PREVIOUS WORK

In the context of a digital video library project, Hauptmann and Olligschlaeger describe a location analysis system [11] used to plot locations mentioned in automatically transcribed news broadcasts on a map. They use a NE tagger on the speech data. Each entity is matched against a global gazetteer, and spatial resolution is attempted using a cascade of decisions. The cues mentioned contribute points to a score for every candidate, and candidates with the highest scores are selected. Their method exploits the fact that repeated mentions of the same place-name, with and without explicit disambiguating cues ("Cambridge" and "Cambridge, MA" in the same text) are likely to refer to the same candidate, so the resolved reading can be propagated to the unresolved one. The partial algorithm does not attempt to resolve all toponyms in a text. In a small evaluation, 269 out of 357 (or 75%) were resolved correctly. Unfortunately, the results reported are of limited value for comparisons since the NE tagger was trained and run on all-uppercase data, and speech recognition errors additionally influence the resolution performance.

Also in a digital library project Smith and Crane proposed the following a toponym resolution method based on storing an explicit map representation [14]: For all possible referents and toponyms, add their coordinates on a map array with one degree resolution with weights given by the number of mentions of each toponym. Compute centroid of the weighted map, and calculate the standard deviation of it from the centroid. Discard all points that are more than two times the standard deviation away from the centroid. They report F-measures between 0.81 and 0.96 in some cases, but found the centroid-based approach to lack robustness.

Smith and Mann use a Naïve Bayes classifier to classify mentions of locations with respect to the underlying U.S. state or (non-U.S.) country [15]. Their definition of the task is a simplified version of the general toponym resolution task, as they do not provide coordinates as output. Considering the toponym types in a gazetteer (rather than tokens in a corpus), they report that 57.1% of US place-names are referentially ambiguous, compared to only 16.6% in Europe [15]. They report 87.38% accuracy in recovering deleted disambiguation cues such as "*Portland, Maine*" in news text. Against a hand-labeled corpus of American biographies and Civil War texts, the same classifier performs at 77.19% and 21.82% accuracy, respectively.

The *Sequoia 2000* project [3] provides storage, indexing, retrieval and browsing of geographic documents based on integrating the *POSTGRES* relational database management system with a full-text IR engine. In this context, Woodruff and Plaunt describe the *GIPSY* subsystem for automatic geo-referencing of text [17]. They incrementally construct a polytope via merging initially flat polygons of the places mentioned in such a way that a third dimension (z-axis increment) is introduced for the intersecting area (polygon stacking). They report runtime problems due to the cost of the polygon manipulations, and issues with noise which might be remedied using a NE tagger before gazetteer lookup, which they do not consider.

The *InfoXtract* IE system [16] has been extended by a component to normalize spatial expressions [7, 8]. Toponym resolution is based on local pattern matching, discourse cooccurrence analysis and default senses. All location names are looked up, then patterns are applied. After applying a "One referent per discourse" heuristic, selected referents are propagated to all other mentions. Then a Minimum Spanning Tree (nodes are toponyms, arcs are relationships) is computed using *Kruskal's Algorithm* to resolve remaining referential ambiguities; it finds a subgraph that (a) contains every vertex of the original graph; (b) has a tree shape and (c) simultaneously maximizes the total weight of the nodes. Senses are acquired from the Web by imposing the *Yahoo!* directory's geographical ontology, thus biasing the system toward a U.S.-American view, which is helpful for the processing of the CNN news stories they evaluate on.

Rauch *et al.* describe the *Metacarta Text Search* (*MTS*) system, which is based on *confidence*. Toponyms are resolved using both supportive and negative contexts [12]. For every candidate referent to a toponym n to a location p, the confidence that n "really" belongs to p is estimated. Features used as evidence or counter-evidence include presence in a location gazetteer, presence of U.S. postal addresses, explicit coordinates local linguistic context, matching of spatial patterns, population heuristics associated with potential referents, and and relative reference cues.

Despite these atttempts and the usefulness of the toponym resolution task, no general, scalable solution has been published so far, and no gold standard for evaluation is available.

3. PROPOSED RESEARCH

Starting from existing proposals of algorithms from the literature, it will be investigated how referentially ambiguous toponyms can be resolved reliably. First a gold standard needs to be devised, as the task in the general form defined here has not been addressed before with respect to systematic large-scale evaluation. Subsequently, several heuristic and supervised machine learning based methods can be implemented and assessed.

Then the referential ambiguity can be measured (to get an idea of the task difficulty). A simple baseline can be defined by always assigning the globally most common referent, ignoring any contextual cues.

A set of new techniques will be devised and evaluated on the gold standard. In [6], I have proposed a new method based on two minimality heuristics: "one referent per discourse" [2] and "minimal bounding box"; their evaluation is postponed until the construction of the evaluation corpus.

3.1 Methodology

A *component evaluation* of the new resolution methods requires

(a) an *evaluation metric*: a slight modification of the traditional F-measure (as well as precision and recall) will be necessary, since Mercator coordinates in various gazetteers used can differ due to imprecision, which could be countered by replacing equality check by a proximity threshold test; and

(b) a *gold-standard*, which is currently being designed [4] using texts from the Reuters RCV1 news corpus , WWW sources and texts from the historic *Statistial Accounts of Scotland*. This resource needs to be annotated with coordi-

referent" or hotels book in the "'right' place in the wrong country".

nates (in addition to adding partially already existing named entity information). The first dataset is to be contructed by the author with the criterion of being sharable across researchers in mind, the second dataset is available inhouse.
For the place-name recognition step, I propose to use off-the-shelf classifiers trained for the NERC task. An important part is the use of corpus data already marked up with gold-standard named entities for the component evaluation of the resolution step, without potential noise introduced by NERC.

3.2 Proposed Experiments

What kind of information could be used to resolve place-names beyond the simple heuristics suggested, and how can it be used?
Heuristics. Beyond systematic replication of heuristics used in the aforementioned literature, we have proposed two **minimality heuristics** for toponym resolution [6], but they will likely need to be supplemented by other sources of evidence, if robust resolution across genres is sought.
Linguistic cues. Although linguistic cues create a cross-language portability barrier if utilized, they are too important to neglected. Thus, they should ideally be incorporated as modular *features* in a supervised machine learning regime wherever available. I'm planning to use shallow finite-state patterns over surface strings, POS and chunk tags as binary features (match/non-match). For example, metaphoric use as in "Washington said..." suggests that a match of the pattern feature "⟨*toponym*⟩ `+ said`" might be is a strong predictor for the reading *Washington, DC, USA* rather than for any other *Washington*.[3] Furthermore, exploiting recent successes in the robust extraction of deep semantic relationships [1], a finer-grained model could take into account the logical structure of the sentence in which a toponym occurs. Defining a good set of features based on linguistic intuition, and informed by empirical analysis of a training portion of the corpus, is one of the main challenges of this project.
Co-occurrence statistics. Statistical association measures from collocation extraction can also be used to compute ties between toponym-toponym pairs (co-locations quite in the literal, but discontinuous, sense) as well as toponym-term pairs which are good predictors for sought toponym referents. For example, calculating Pointwise Mutual Information (PMI) or Log-Likelihood Ratio (LLR), we can determine that the tie between ⟨Pennsylvania;USA⟩ is closer than between ⟨Pennsylvania;Australia⟩, and use this knowledge as soft evidence supporting a decision.
Discourse and position information. Published text is usually written by professionals who follow genre-specific structural conventions. *News items*, for example, start with a grounding indicators ("New York (CNN).") and have an underlying "Christmas tree" structure, where the first paragraph sums up the news. Then details are laid out in more detail, and a final paragraph often places the event described into a wider context and relates it to similar events in the past. *Biographies* often have a single location remains prominently associated with a longer text fragment, e.g. a whole chapter ("my childhood years in North Yorkshire"), with occasional statements referring forwards ("I would never feel so happy again after moving to the City to pursue my career") or backwards ("This reminded me of my wonderful Edinburgh years."). *Geographic descriptions* (surveys such as the *Statistical Accounts of Scotland* or travel guides), on the other hand, tend to iterate over all regions and focus in on supposedly interesting spots ("To the South of it, the pretty peninsula of Dingle offers a typical impression of Irish rural life.").
The utilization of these heterogeneous information sources presents us with an evidence integration problem, which can be solved by inducing a data-driven decision procedure in a supervised regime that learns how evidence from heuristics and the other features mentioned here have to be weighted to select intended referents.

[3]One might argue that *Washington* is here used as an ORGANIZATION rather than LOCATION according to some NERC guidelines and should thus be ignored altogether, but the spatial resolution of it might still contribute to the resolution of other toponyms found in the same text.

3.3 Main Research Questions

How can we resolve toponyms to coordinates reliably and robustly in open-domain text such as news? Which method works best for historic text like the RCHAMS data? Which features and what supervised machine learning setting are useful to induce a component that can perform the task?

3.4 Issues for Discussion

One problem for the curation of a gold standard is the dependence of gazetteers to look up the coordinates of candidate referents: different gazetteers have different densities, and there is a certain imprecision in the coordiates due to measurement, data conversion/representation and different definitions of "centroid of a geographic feature" (such as a city). Imprecision also impacts the evaluation metric; a more lenient version of Precision/Recall is required.

4. CONCLUSIONS

Successful toponym resolution is expected to help increase precision in applications such as geographic information retrieval, topic detection and tracking [9] and question answering [18, 5]. Geographic information retrieval is especially interesting as current Web search engines do not to date support a notion of space, and keyword-based attempts to constrain a seach spatially cannot discriminate between the various toponym referents. Toponym resolution will finally provide us with the right Sheffield.

Acknowledgments. The author is grateful to Claire Grover, Bonnie Webber and Steve Clark for discussions and guidance and to the U.S. National Geo-spatial Intelligence Agency (NGA) for providing the data and support. This research is funded by a doctoral scholarship from the German Academic Exchange Service (DAAD), and a research grant from Linguit GmbH.

5. REFERENCES

[1] J. Bos, S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier. Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004.

[2] W. Gale, K. Church, and D. Yarowsky. One sense per

discourse. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.

[3] R. R. Larson, C. Plaunt, A. G. Woodruff, and M. Hearst. The Sequoia 2000 electronic repository. *Digital Technical Journal of Digital Equipment Corporation*, 7(3):50–65, 1995.

[4] J. L. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval held at SIGIR-2004*, Sheffield, UK, submitted. ACM SIGIR.

[5] J. L. Leidner, J. Bos, T. Dalmas, J. R. Curran, S. Clark, C. J. Bannard, M. Steedman, and B. Webber. The QED open-domain answer retrieval system for TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, pages 595–599, Gaithersburg, MD, 2003.

[6] J. L. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In A. Kornai and B. Sundheim, editors, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*, pages 31–38, Edmonton, Alberta, Canada, May 2003. Association for Computational Linguistics.

[7] H. Li, K. R. Srihari, C. Niu, and W. Li. Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In A. Kornai and B. Sundheim, editors, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*, pages 39–44, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[8] H. Li, R. K. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *COLING 2002*, Taipei, Taiwan, 2002.

[9] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Topic detection and tracking with spatio-temporal evidence. In *Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003)*, pages 251–265, Pisa, Italy, 2003.

[10] I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000.

[11] A. M. Olligschlaeger and A. G. Hauptmann. Multimodal information systems and GIS: The Informedia digital video library. In *1999 ESRI User Conference*, San Diego, CA, 1999.

[12] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In A. Kornai and B. Sundheim, editors, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*, pages 50–54, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[13] A. Setzer and R. Gaizauskas. On the importance of annotating temporal event-event relations in text. In *LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language*, Las Palmas, Gran Canaria, Spain, 2002.

[14] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001*, pages 127–136, 2001.

[15] D. A. Smith and G. S. Mann. Bootstrapping toponym classifiers. In A. Kornai and B. Sundheim, editors, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*, pages 45–49, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[16] R. K. Srihari, C. Niu, and W. Li. Hybrid approach for named entity and sub-type tagging. In *ANLP 2000*, Seattle, WA, 2000.

[17] A. Woodruff and C. Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9):645–655, 1994.

[18] H. Yang, T.-S. Chua, S. Wang, and C.-K. Koh. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 33–40. ACM Press, 2003.

[19] G. Zheng and J. Su. Named entity tagging using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 209–219, Philadelphia, 2002.