DISSERTATION ABSTRACT

# Toponym Resolution in Text (Annotation, Evaluation and Applications of Spatial Grounding)

**Jochen L. Leidner**
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 8LW, Scotland, UK
*leidner@acm.org*

**Background.** In Information Extraction (IE), processing of named entities in text has traditionally been seen as a two-step process comprising a flat text span recognition sub-task and an atomic classification sub-task; relating the text span to a model of the world has been ignored by evaluations such as DARPA/NIST's MUC or ACE. However, spatial and temporal expressions refer to events in space-time, and the grounding of events is a precondition for accurate reasoning. Thus, automatic grounding can improve many applications such as automatic map drawing (e.g. for choosing a focus) and question answering (e.g., for questions like *How far is London from Edinburgh*, given a story in which both occur and can be resolved). Whereas temporal grounding has received considerable attention in the recent Past [2, 3], robust spatial grounding has long been neglected. Concentrating on geographic names for populated places, I define the task of automatic *Toponym Resolution* (TR) as computing the mapping from occurrences of names for places as found in a text to a representation of the extensional semantics of the location referred to (its referent), such as a geographic latitude/longitude footprint. The task of mapping from names to locations is hard due to insufficient and noisy databases, and a large degree of ambiguity: common words need to be distinguished from proper names (geo/non-geo ambiguity), and the mapping between names and locations is ambiguous *London* can refer to the capital of the UK or to London, Ontario, Canada, or to about forty other Londons on earth). In addition, names of places and the boundaries referred to change over time, and databases are incomplete.

**Objective.** I investigate how referentially ambiguous spatial named entities can be grounded, or resolved, with respect to an extensional coordinate model robustly on open-domain news text. I collect published algorithms and factor out a shared repertoire of linguistic heuristics (e.g. rules, patterns) and extra-linguistic knowledge sources (e.g. population sizes). I then investigate how to combine these sources of evidence to obtain a superior method. I also investigate the noise effect introduced by the named entity tagging step that toponym resolution relies on a sequential system pipeline architecture.

**Scope.** In this thesis, I investigate a present-day snapshot of terrestrial geography as represented in the gazetteer defined and, accordingly, a collection of present-day news text. I limit the investigation to populated places; geo-coding of artifact names (e.g. airports or bridges), compositional geographic descriptions (e.g. *40 miles SW of London, near Berlin*), for instance, is not attempted. Historic change is a major factor affecting gazetteer construction and ultimately toponym resolution. However, this is beyond the scope of this thesis.

**Method.** While a small number of previous attempts have been made to solve the toponym resolution problem, these were either not evaluated, or evaluation was done by manual inspection of system output instead of curating a reusable reference corpus. Since the relevant literature is scattered across several

disciplines (GIS, digital libraries, information retrieval, natural language processing) and descriptions of algorithms are mostly given in informal prose, I attempt to systematically describe them and aim at a *reconstruction in a uniform, semi-formal pseudo-code notation* for easier re-implementation. A systematic comparison leads to an *inventory of heuristics and other sources of evidence*. In order to carry out a comparative evaluation procedure, an evaluation resource is required. Unfortunately, to date no gold standard has been curated in the research community. To this end, a reference gazetteer and an associated novel reference corpus with human-labeled referent annotation are created. These are subsequently used to benchmark a selection of the reconstructed algorithms and a novel re-combination of the heuristics cataloged in the inventory. I then compare the performance of the same TR algorithms under three different conditions, namely applying it to the (i) output of human named entity annotation, (ii) automatic annotation using an existing Maximum Entropy sequence tagging model, and (iii) a näive toponym lookup procedure in a gazetteer.

**Evaluation.** The algorithms implemented in this thesis are evaluated in an intrinsic or *component evaluation*. To this end, we define a task-specific matching criterion to be used with traditional Precision ($P$) and Recall ($R$) evaluation metrics. This matching criterion is lenient with respect to numerical gazetteer imprecision in situations where one toponym instance is marked up with different gazetteer entries in the gold standard and the test set, respectively, but where these refer to the *same* candidate referent, caused by multiple near-duplicate entries in the reference gazetteer.

**Main Contributions.** The major contributions of this thesis are as follows:

- a *new reference corpus* in which instances of location named entities have been manually annotated with spatial grounding information for populated places, and an associated *reference gazetteer*, from which the assigned candidate referents are chosen. This reference gazetteer provides numerical latitude/longitude coordinates (such as 51°32’ North, 0°5’ West) as well as hierarchical path descriptions (such as `London > UK`) with respect to a world wide-coverage, geographic taxonomy constructed by combining several large, but noisy gazetteers. This corpus contains news stories and comprises two sub-corpora, a subset of the REUTERS RCV1 news corpus used for the CoNLL shared task [4], and a subset of the Fourth Message Understanding Contest (MUC-4;[1]), both available pre-annotated with gold-standard;
- a new *method and implemented system to resolve toponyms* that is capable of robustly processing unseen text (open-domain online newswire text) and grounding toponym instances in an extensional model using longitude and latitude coordinates and hierarchical path descriptions, using internal (textual) and external (gazetteer) evidence and a *comparison between a replicated method* as described in the literature, which functions as a baseline, *and a novel algorithm based on minimality heuristics*; and
- an *empirical analysis of the relative utility of various heuristic biases and other sources of evidence* with respect to the toponym resolution task when analyzing free news genre text;

Available for download from URI: http://hdl.handle.net/1842/1849

## References

[1] Nancy A. Chinchor. Overview of MUC-4. In Beth Sundheim, editor, *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, Fairfax, VA, USA, 1995. U.S. Defense Advanced Research Projects Agency (DARPA).

[2] Inderjeet Mani and George Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 69--76, Hong Kong, 2000.

[3] Andrea Setzer. *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. PhD thesis, University of Sheffield, Sheffield, England, UK, 2001.

[4] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the {CoNLL}-2003 shared task: language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Seventh Conference on Natural Language Learning (CoNLL 2003)*, pages 142--147, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics. In association with HLT-NAACL 2003.