Towards a Reference Corpus for Automatic Toponym Resolution Evaluation (Extended Abstract)

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

ABSTRACT

Spatial named entities ground events in space, and this relationship is essential for advanced text processing applications such as question answering and event tracking. *Toponym resolution* is the task of mapping from an entity to a spatial representation (an extensional coordinate model), given the context. Whereas work on the temporal dimension is ongoing [17], to date no reference corpus exists to evaluate competing algorithms for toponym resolution.
This paper argues that a shareable evaluation resource is necessary, and presents a proposal for the markup and the process of annotating the corpus.
We present TRML, an XML-based markup language, and TAME, the Toponym Annotation Markup Editor, which are both part of a tool-chain developed as part of an ongoing corpus curation effort to address this issue.

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

Geo-coding; geo-parsing; geo-referencing; place-name disambiguation; spatial retrieval; geographic IR

1. INTRODUCTION

Named entity tagging is usually seen as the task of identifying a text span and classifying it. However, this limited 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, event tracking or map generation from text [7]. There is much recent work regarding the temporal dimension [17, 10], but the spacial dimension has only very recently been received more interest [6].
Specifically, in this paper we address the curation of a reference corpus for **toponym resolution**, which can be defined as the task of computing the relation between a place name referring to a location and the location itself (spacial position and extend). For instance, the string "London" is referentially ambiguous between (among many others)

1. London, England, United Kingdom of Great Britain and Northern Ireland in Europe ($51°30'0''N$, $0°7'W$) and
2. London, Ontario, Canada in North America ($42°59'N$, $81°15'W$).

But currently there is no publicly available gold-standard corpus for objective evaluation. Because this prohibits comparison of early works [11, 18] and more recent proposals [8, 13, 7] under controlled conditions (including the use of the same dataset for evaluation), this presents a barrier to progress in the field.
This paper presents an ongoing *effort to provide a new, reusable reference corpus of text manually annotated with spatial named entities with their correlates in a latitude/longitude coordinate model* (grid reference) as a training and evaluation resource.
The remaining parts of this paper are structured as follows. Section 2 describes some design criteria that guide and constrain the curation effort and analyzes some problems. In Section 3, a simple markup scheme for annotation for toponym reference is provided, and Section 4 presents the design and implementation of an annotation tool based on it.

2. TOPONYM ANNOTATION

2.1 Corpus sampling

The first question when curating a reference corpus is the question of *corpus sampling*. For this project, texts from different genres are being collected, ranging from newswire texts provided by global news syndicates and online newspapers to personal narrative. The corpus will mainly be synchronic, but a historic subcollection, part of the *Statistical*

SIGIR 2004 Sheffield, UK. (Geographic IR Workshop)

Annotation with	Type	Structure
latitudes/longitudes	numeric	flat
grid references	symbolic	hierarchical
polygons	numeric	set (of flat)
ISO 631 path identifier	symbolic	hierarchical
Aura Location Identifier	hybrid	hierarchical

Table 1: Different kinds of spatial annotation.

Name	*Distributor*	*Coverage*	*Entries*
Columbia	Columbia UP	Earth	165,000
Digimap	EDINA	UK	258,797
GNIS	U.S. Geogr. Survey	USA	1,836,264
GNS	U.S. NGA	Earth\USA	5,268,934
TGN	J. P. Getty Trust	Earth	1,300,000
UN LOCODE	UNECE	Earth	40,000

Figure 1: Gazetteer profiles.

Accounts of Scotland by the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS),[1], will also be included. However, in this paper, only the synchronic news sub-corpus is considered.
The Reuters **RCV1** corpus[2] [14] was chosen for news, since it contains stories of global scope and interest, and because a subset of it was manually annotated with gold-standard named entities for the CoNLL shared task in named entity tagging [19]. Human gold-standard named entity markup allows to assess toponym resolution accuracy without introducing noise by the named entity recognition sub-task (i.e., a controlled component-based evaluation).

2.2 Referent representation

There have been several proposals for representing locations (Table 1 gives a summary): numerical *latitude/longitude coordinates* are the most widely used system. Some countries define a structured *grid reference* system. *Polygon points* can be used to describe a location associated with a toponym more accurately, since e.g. cities have complex shapes. ISO 631 path identifiers like `de_mag` (for the city of Mageburg, Germany) have also been proposed [16]. Jiang and Steenkiste [5] describe a hybrid notation for the representation of locations in a ubiquitous computing environment.
For this project, numeric point coordinates (latitudes and longitudes) are used with human-readable path descriptions like `London > United Kingdom > Europe`.

2.3 Problems of gazetteer selection

Existing gazetteers (see Figure 1 for some examples) vary along a large number of dimensions. The following seven key criteria for gazetteer selection were taken into account for selecting a gazetteer compatible with the goals of this project:

1. Gazetteer availability: the need to be able to share experimental data posits that free resources be given preference in research.

2. Gazetteer scope: gazetteers vary in range from small communal (cadastral) databases over regional/national list to worldwide scope. In this project, grounding shall be attempted on global scope, which requires earth-wide scope.

3. Gazetteer completeness: no gazetteer to date covers all places in existence; but whereas some are very comprehensive, others only have a very limited coverage.

4. Gazetteer correctness Gazetteer precision (measurements are inherently imprecise). Gazetteers typically contain many wrong or outdated entries: for example, in 1996, South Africa changed its administration from four provinces to nine.[3] However, at the time of writing (2004-05), the current GNS edition still features a London, Transvaal, South Africa,although Transvaal does not exist any more. Indeed, there are circa 20,000 changes per month in the GNS gazetteer alone.

5. Gazetteer granularity: not all gazetteers aim to achieve completeness; some merely list the more popular or relevant places. A less fine-grained gazetteer might actually facilitate the toponym resolution task by providing a useful bias (in the same way that average humans living in New York are not familiar with Siberian villages), and too fine-grained databases yield "noise", but sometimes unpopular places are in the media spotlight for a short term due to an important event, and it is then desirable for a system to have very fine-grained geographic knowledge.

6. Gazetteer balance: a gazetteer that is balanced provides uniform degree of detail and correctness across all continents and regions.

7. Gazetteer richness of annotation: the amount and detail of information associated with the name of a place varies from mere longitude/latitude numbers to detailed type and population information.

For this study, the GNIS gazetteer of the U.S. Geographic Survey and the GNS corpus of the National Geospatial Intelligence Agency (NGA)[4] were used. If used together, they have world-wide scope, very good coverage, and the data can be freely shared. However, the quality of the data is only modest, and it is much less suited for studies of grounding historic text.
The "Schrödinger's Tag" Paradoxon. Unlike in traditional text span classification tasks, a grounding task must reply on an external knowledge source and thus suffers from a interdependence between gazetteer/ontology on the one hand and the document with instances to be marked up and grounded on the other hand: *the gazetteer is not simply an interchangeable system component, it gains reference status together with the corpus in which it is employed to look up the set of potential referents.*
This means that the gazetteer chosen to curate a reference corpus influences the outcome of any experiment: there can be a potential bias towards systems using the same gazetteer

[1] `http://www.rcahms.gov.uk/`
[2] `http://about.reuters.com/researchandstandards/corpus/`
[3] personal communication, Douglas E. Ross, National Geospatial Intelligence Agency, 2004-04-23.
[4] formerly known as NIMA

for resolving the toponyms. However, if systems are designed in a modular fashion, they could be provided with the gazetteer used for gold-standard curation for the purpose of evaluation of the resolution method only.[5]

3. A SIMPLE MARKUP SCHEME

This section describes *Toponym Resolution Markup Language* (TRML), the XML-based markup scheme, which is implemented by the tool-chain described in the next section.

Appendix B gives an example fragment of valid TRML. XML was chosen because of its standard status and the widespread tool support.[6] An important desideratum for the design of a successful toponym markup was that *document structure* should be preserved. Otherwise, discourse conventions, such as introducing a news story by specifying the main location and the source of the information below the headline cannot be utilized by the resolution method.

Thus, TRML offers markup for documents (`<doc>`), optional paragraphs[7] (`<p>`) and sentences (`<s>`). Sentences comprise either word tokens (sometimes referred to as *w-unit*, `<w>`) or toponyms (`<toponym>`), which in turn contain one or more `<w>` elements followed by a `<candidates>` element that contains a set of the alternative candidate referents (`<cand>`). Each of these locations has an identifier and carries information about origin of the data (e.g. whether an entry is from the NIMA gazetteer), decimal longitude/latitude coordinates as well as a hierarchical geographic path description for the human annotator (`humanPath`). A '`select`' attribute stores the referent chosen by the annotator.

4. TOOL-CHAIN AND MARKUP PROCESS

This section describes the implementation of the *Toponym Annotation Markup Editor* (TAME), the tool (Figure 2) that constitutes the annotation system, and the marked up process it supports.

The CoNLL data comes in tabular plain-text format (1), where the first column contains a token (word or part of a multi-token word), the second column contains a part-of-speech tag, the third column contains a chunk-tag and the fourth and final column contains a named entity tag in BIO-format [19]. Sentence boundaries are represented by empty lines, and the start of a new document is indicated by an idiosyncratic `-DOCSTART-` token. This format does not lend itself to elegant extension or processing with modern tools based on structured data modeling standards [15].

A markup language for toponym resolution called TRML was specified based on XML [21], and a converter was implemented in Perl which transforms the CoNLL format into TRML (2). During this conversion, an SQL-based gazetteer server is consulted on the fly (3) to look up the set of candidate referents for each toponym (i.e. named entity instance of type `LOC`). This server (which takes about 1.5 GB of persistent storage) delivers entries from the gazetteers provided by NGA and USGS very efficiently (4). The result of the process is a set of independent XML document instances that can be served to annotators anywhere on the Internet over HTTP [3] by a Web server (5-8). However, raw XML data containing numerical coordinates would be of little use to human annotators. This is traditionally solved by converting a set of XML document instances to HTML. Here, another route was taken: an XSLT style sheet [1] was implemented (9) that translates TRML into XHTML [12] dynamically on the client.

XHTML forms are used to offer the actual annotation interaction to the human annotator, who simply selects a referent from a list of candidates presented in a drop-down menu (Figure 3). Selecting one out of a set of textual path descriptions such as `London > United Kingdom > Europe` hides the numerical longitude/latitude coordinates from the user, which are associated with the paths in the TRML internally, but not rendered visibly.

Very rarely it might happen that a toponym has more than one referent even within the smallest administration region, such as 'London', which has three several potential referents in South Africa alone, one in the Northern Province, and two in Mpumalanga.[8] For such cases, the TAME editor automatically flags such cases for expert moderation, since human annotators lack cues and typically also expertise to cope with these instances. However, it is not expected that such cases are encountered often (if at all) in the RCV1 data.

When displayed in a standard Web browser the document instances are also validated automatically on the fly to ensure correctness (10), making the system immune to syntactic conversion errors. The advantage of this procedure is that no spurious files have to be maintained and kept up to date after system modification.[9]

Cases where human annotators are uncertain about their annotation decision can be flagged for moderation using a check-box. After the annotation of a document is complete, submission transfers the results back via CGI [2], which completes the cycle (11-12).

5. CONCLUSIONS AND FUTURE WORK

This paper describes an ongoing effort to create a reference corpus for the toponym resolution task. Design issues regarding corpus sampling, gazetteer influence, and markup schemes were discussed. TRML, a new proposal for a markup language, and TAME, an editor which implements document annotation supporting it, were presented.

The annotation is currently in progress, and it is hoped that the resulting corpus will prove to be a useful resource for the evaluation of toponym resolution algorithms. Specifically, it will be used to evaluate a set of heuristics proposed in the literature, including [7]. It will also be used to train and evaluate classifiers in standard supervised statistical machine learning regimes [4].

Toponym resolution is a prerequisite for high-quality geographic information retrieval, event-oriented (especially spatial) question answering [20], and grounding events in topic detection and tracking systems [9]. But successful toponym resolution research requires controlled evaluation of the state of the art, and this in turn requires a substantial evaluation

[5]This would be a method evaluation rather than a component evaluation, since the system would still be deployed with another gazetteer.

[6]For these same reasons, stand-off XML, which we consider a superior modeling approach, was discarded: for instance, Web browsers do not at the time of writing support stand-off XML.

[7]The CoNLL subset of RCV1 does not contain paragraph information.

[8]personal communication, Douglas E. Ross, 2004-04-23.

[9]On the client, TRML is shown when 'view source' is selected in a typical Web browser instead of the XHTML seen by the user.

Figure 3: TAME, the Toponym Annotation Markup Editor (screen shot).

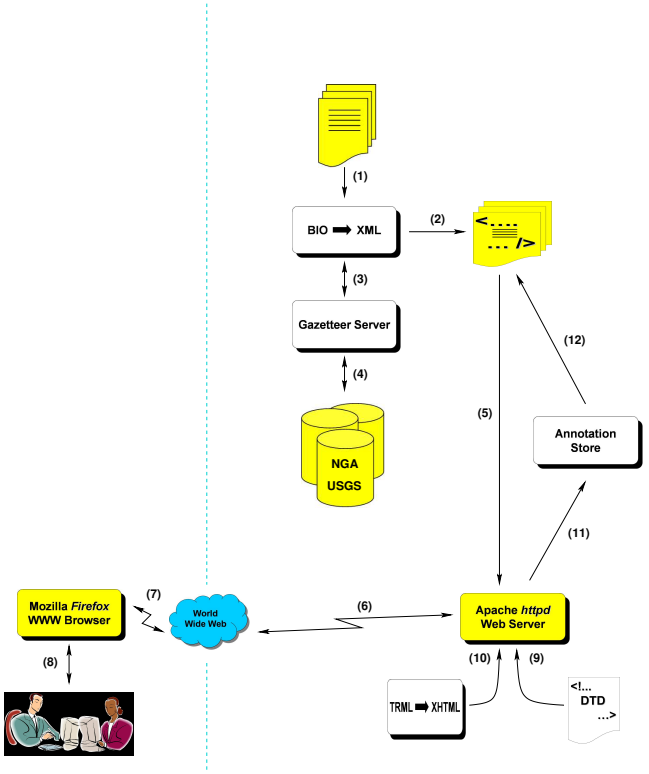

Figure 2: The TAME system architecture.

corpus such as the one being constructed in the project described here.

Acknowledgments. The author is grateful to the U.S. National Geospatial Intelligence Agency (NGA) and the U.S. Geographic Survey (USGS) for providing the gazetteer data, without which this research project would not be possible in its present scope. Steve Clark, Jean Carletta, Claire Grover, Bruce Gittings, András Kornai, Yuval Krymolowski, Colin Matheson, Malvina Nissim, Douglas E. Ross, Yannick Versley and Bonnie Webber provided valuable input in numerous discussions.
This research is funded by the German Academic Exchange Service (DAAD) under scholarship D/02/01831 and by Linguit GmbH under grant UK-2002/2. The support by the School of Informatics, University of Edinburgh, is also gratefully acknowledged.

6. REFERENCES

[1] J. Clark. XSL Transformations (XSLT) Version 1.0. W3C Recommendation 16 November 1999 [online], 1999. http://www.w3.org/TR/xslt.

[2] K. A. L. Coar and D. R. T. Robinson. The WWW Common Gateway Interface Version 1.1. W3C Recommendation [online], 1999. http://cgi-spec.golux.com/draft-coar-cgi-v11-03-clean.html.

[3] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. Network Working Group 2616 [online], 1999. http://www.ietf.org/rfc/rfc2616.txt.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[5] C. Jiang and P. Steenkiste. A hybrid location model with a computable location identifier for ubiquitous

computing. In *Proceedings of UbiComp 2002: Ubiquitous Computing: 4th International Conference*, pages 246–263, Goteborg, Sweden, 2002. Springer.

[6] A. Kornai and B. Sundheim, editors. *Analysis of Geographic References. Workshop held at the HLT/NAACL Conference 2003*, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[7] J. L. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In Kornai and Sundheim [6], pages 31–38.

[8] H. Li, K. R. Srihari, C. Niu, and W. Li. *InfoXtract location normalization: a hybrid approach to geographic references in information extraction*, pages 39–44. In Kornai and Sundheim [6], 2003.

[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*, pages 69–76, Hong Kong, 2000.

[11] A. M. Olligschlaeger. *Spatial Analysis of Crime Using GIS-Based Data: Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1997.

[12] S. Pemberton. XHTML 1.0 The Extensible HyperText Markup Language (Second Edition). W3C Recommendation, 26 Januar 2000, Revised 1 August 2002 [online], 2002. http://www.w3.org/XML/.

[13] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In Kornai and Sundheim [6], pages 50–54.

[14] T. G. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus Volume 1 – from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, volume 3, pages 827–833, Las Palmas de Gran Canaria, Spain, 2002.

[15] K. B. Sall. *XML Family of Specifications: A Practical Guide*. Addison-Wesley, Boston, MA, 2002.

[16] F. Schilder, Y. Versley, and C. Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. in preparation.

[17] A. Setzer and R. Gaizauskas. On the importance of annotating temporal event-event relations in text. In *LREC 2000 Workshop on Annotation Standards for Temporal Information in Natural Language*, volume 3, pages 1281–1286, Athens, Greece, 2000.

[18] 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.

[19] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In W. Daelemans and M. Osborne, editors, *Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 142–147, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics. In association with HLT-NAACL 2003.

[20] 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.

[21] F. Yergeau, J. Cowan, T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler. XML 1.1: Extensible Markup Language (XML) 1.1. W3C Recommendation, 4th February 2004 [online], 2004. http://www.w3.org/TR/xml11.

APPENDIX

A. SAMPLE CONLL FORMAT

```
-DOCSTART- -X- -X- O

EU NNP I-NP I-ORG
rejects VBZ I-VP O
German JJ I-NP I-MISC
call NN I-NP O
to TO I-VP O
boycott VB I-VP O
British JJ I-NP I-MISC
lamb NN I-NP O
. . O O

Peter NNP I-NP I-PER
Blackburn NNP I-NP I-PER

BRUSSELS NNP I-NP I-LOC
1996-08-22 CD I-NP O

The DT I-NP O
European NNP I-NP I-ORG
Commission NNP I-NP I-ORG
said VBD I-VP O
on IN I-PP O
Thursday NNP I-NP O

[...]
```

Figure 4: CoNLL format (excerpt).

B. SAMPLE TRML DOCUMENT INSTANCE

```
<doc id="d1">
 <s id="s1">
  <w tok="EU" pos="NNP" chk="I-NP" ne="I-ORG" />
  <w tok="rejects" pos="VBZ" chk="I-VP" ne="O" />
  <w tok="German" pos="JJ" chk="I-NP" ne="I-MISC" />
  <w tok="call" pos="NN" chk="I-NP" ne="O" />
  <w tok="to" pos="TO" chk="I-VP" ne="O" />
  <w tok="boycott" pos="VB" chk="I-VP" ne="O" />
  <w tok="British" pos="JJ" chk="I-NP" ne="I-MISC" />
  <w tok="lamb" pos="NN" chk="I-NP" ne="O" />
  <w tok="." pos="." chk="O" ne="O" />
 </s>
 <s id="s2">
  <w tok="Peter" pos="NNP" chk="I-NP" ne="I-PER" />
  <w tok="Blackburn" pos="NNP" chk="I-NP" ne="I-PER" />
 </s>
 <s id="s3">
  <toponym did="1" sid="3" tid="1" term="BRUSSELS">
    <w tok="BRUSSELS" pos="NNP" chk="I-NP" ne="I-LOC" />
    <candidates>
      <cand id="c1" src="NIMA" lat="-23.3833333" long="29.15"
            humanPath="Brussels &gt; (SF04) &gt; South Africa" />
      <cand id="c2" src="NIMA" lat="-24.25" long="30.95"
            humanPath="Brussels &gt; (SF04) &gt; South Africa" />
      <cand id="c3" src="NIMA" lat="-24.6833333" long="26.6833333"
            humanPath="Brussels &gt; (SF04) &gt; South Africa" />
      <cand id="c4" src="NIMA" lat="-27.1" long="24.6666667"
            humanPath="Brussels &gt; (SF01) &gt; South Africa" />
      <cand id="c5" src="NIMA" lat="-27.15" long="24.75"
            humanPath="Brussels &gt; (SF01) &gt; South Africa" />
      <cand id="c6" src="NIMA" lat="50.8333333" long="4.3333333"
            selected="yes"
            humanPath="Brussels &gt; (BE02) &gt; Belgium" />
      <cand id="c7" src="USGS_PP" lat="38.94944" long="-90.58861"
            humanPath="Brussels &gt; Calhoun &gt; IL &gt; US &gt; North America" />
      <cand id="c8" src="USGS_PP" lat="44.73611" long="-87.62083"
            humanPath="Brussels &gt; Door &gt; WI &gt; US &gt; North America" />
    </candidates>
  </toponym>
  <w tok="1996-08-22" pos="CD" chk="I-NP" ne="O" />
 </s>
 <s id="s4">
  <w tok="The" pos="DT" chk="I-NP" ne="O" />
  <w tok="European" pos="NNP" chk="I-NP" ne="I-ORG" />
  <w tok="Commission" pos="NNP" chk="I-NP" ne="I-ORG" />
  <w tok="said" pos="VBD" chk="I-VP" ne="O" />
  <w tok="on" pos="IN" chk="I-PP" ne="O" />
  <w tok="Thursday" pos="NNP" chk="I-NP" ne="O" />

[...]
```

Figure 5: TRML format (excerpt).