School of Informatics, University of Edinburgh

Institute for Communicating and Collaborative Systems

Toponym Resolution: A First Large-Scale Comparative Evaluation

by

Jochen L. Leidner

Informatics Research Report

School of Informatics **July 2006**
http://www.informatics.ed.ac.uk/

Toponym Resolution: A First Large-Scale Comparative Evaluation

Jochen L. Leidner

SCHOOL *of* INFORMATICS
Institute for Communicating and Collaborative Systems

July 2006

Copyright © 2006 University of Edinburgh. All rights reserved. Permission is hereby granted for this report to be reproduced for non-commercial purposes as long as this notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed to Copyright Permissions, School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland.

Abstract

Toponym resolution (TR) is the task of mapping the name of a location to a spatial representation of the location referred to, such as the centroid of the location, given as latitude/longitude. While a number of systems for automating the task have been described in the literature, to date no comparative evaluation study has existed, mainly for lack of a standard benchmark (i.e., gazetteer and evaluation corpus).
On the basis of a benchmark methodology and dataset, we present the first systematic account of the utility of different heuristics for the toponym resolution task, based on experimental comparison on two novel, large-scale gold-standard corpora. Each heuristic's utility is evaluated in isolation, and in addition, two previously reported complex methods are replicated in full.

Keywords : Toponym resolution, natural language processing, information extraction (IE), spatial grounding, geo-coding, Named-Entity Recognition (NERC).

Contents

1 Introduction

A recent strategic assessment by the U.S. government has identified bio-technology, geo-spatial technology and nano-computing (the triad "bio, geo, nano" for short) as the three key technology growth areas in the first decades of the 21st century. Geographic Information Systems (GIS) have existed for a long time, but until recently they have lived in a secluded world of their own. Now that the Internet has become pervasive, spatial enabling of Web content is becoming more important.

This paper deals with the automatic mapping from the name of a place in textual form (a *toponym*) to a formal representation of the extension of the location refered to, such as a polygon or centroid given as latitude/longitude (*toponym resolution*). Consider the following example:[1]

(1) The 1666 London fire was one of the country's most tragic accidents in history.
⇝ `London > England > United Kingdom`
(lat./long.: ⟨51.52; -0.10⟩)

(2) The recent fire in London alarmed policy makers from Montreal to Inuvik.
⇝ `London > Ontario > Canada`
(lat./long.: ⟨42.97; -81.24⟩)

In (1), *London* refers to the British capital, as is either known by native speakers or can be infered from the discourse context in which the example occurs. The *London* in (2), on the other hand, refers to a place in Canada. Most people might not realise the degree of toponym ambiguity until they end up in the "wrong one" of the 42 *London*s, dozens of *Berlin*s, *Aberdeen*s, *Sheffield*s or more than 1,500 *Santa Anas* on earth. Toponym resolution as a computational task bears some resemblance to Word Sense Disambiguation (WSD) in that automating it comprises a look-up step to retrieve *candidate referents* (WSD: senses), and a second step that chooses the most likely candidate. As the natural world is perceived in terms of space and time, toponym resolution is the natural counterpart of *time resolution*. But whereas for WSD and time resolution, standard markup languages, corpora, and shared evaluation exercises have been developed (Edmonds and Kilgarriff (2003); Setzer (2001)), for toponym resolution comparable benchmarking resources have only recently become available (Leidner (2004, 2006)).

1.1 Motivation

Automatic processing of text with the aim of utilizing geographic knowledge contained in it was attempted already more than a decade ago (Woodruff and Plaunt (1994)). However, a systematic understanding of the factors contributing to its success (or failure), or comparative empirical evaluations have not been accomplished. Consequently, many have criticized unprincipled system building (Woodruff and Plaunt (1994); Clough and Sanderson (2004); Martins et al. (2005)), and have pointed out the lack of evaluation efforts. A series of workshops on processing geographic references (starting with Kornai and Sundheim (2003)) has brought together the diverse community, which has helped to intensify the discussion, and several groups have called for proposals for an international evaluation along the lines of MUC (Sundheim (1992)) or the CoNLL "shared task" (Tjong Kim Sang and De Meulder (2003)) in the area of named entity tagging (NERC). However, to date no such effort has materialized. As a consequence of this, some production systems presently have to rely on interactive (i.e. *manual* rather than automatic) resolution (Densham and Reid (2003)) to ensure good quality after the (automatic) toponym recognition step. Only when the factors contributing to success in toponym resolution are well understood will we be able to build systems to automate the task with very high accuracy.

[1] read "⇝" as "resolves to"

We attempt to solve this problem. To this end, we introduce a systematic methodology for toponym resolution evaluation and present empirical results based on applying our methodology.

1.2 Contributions

Our main contributions are the following:

- a first systematic analysis of heuristics and knowledge sources used for TR in the past, derived from an extensive analysis of the literature, which is scattered across the fields of natural language processing, information retrieval and GIS;
- the first quantitative account of the relative utility of different heuristics for the toponym resolution task, based on empirical evaluation using two large-scale gold-standard news corpora, one global in focus (TR-CoNLL), one regional (TR-MUC4);
- the first large-scale comparative evaluation (on more than 1,000 documents) benchmarking two previously reported complex methods, which we replicated in full on newspaper prose and evaluated on the same datasets (test corpora and gazetteer);
- TextGIS®, a robust software platform for toponym resolution experimentation, which allows rapid (re-)implementation and evaluation of methods;

1.3 Report Outline

The remaining sections of this report are organized as follows: Section 2 introduces our methodology and presents an analysis previously reported methods with respect to heuristics and knowledge sources used. Section 4 describes the gazetteer and the two evaluation corpora used in our experiments. Section 3 describes two methods that we replicated in full and the TextGIS® software platform that forms the basis for the implementation and evaluation experiments. Section 5 presents our evaluation results and Section 7 summarizes our findings and outlines future work.

$\mathcal{H}_0$	(**Resolve unambiguous**)
$\mathcal{H}_1$	**"Contained-in" qualifier following**
$\mathcal{H}_2$	**Superordinate mention**
$\mathcal{H}_3$	**Largest population**
$\mathcal{H}_4$	**One referent per discourse**
$\mathcal{H}_5$	**Geometric minimality**
$\mathcal{H}_6$	Singleton capitals
$\mathcal{H}_7$	Ignore small places
$\mathcal{H}_8$	Focus on geographic area
$\mathcal{H}_9$	**Dist. to unambiguous text neighbors**
$\mathcal{H}_{10}$	**Discard off-threshold**
$\mathcal{H}_{11}$	**Frequency weighting**
$\mathcal{H}_{12}$	**Prefer higher-level referents**
$\mathcal{H}_{13}$	Feature type disambiguator
$\mathcal{H}_{14}$	Textual-Spatial Correlation
$\mathcal{H}_{15}$	Default Referent

Table 1: Collected Inventory of Heuristics (Heuristics in Bold are Evaluated in this Paper).

2 Methodology

In order to achieve a better understanding of the factors contributing to performance in toponym resolution, we pursue the following methodology:

1. Analysis of the existing research literature:
   - (re-)**construct pseudo-code** in unified notation (Section 3);
   - **extract inventory of heuristics** and other evidence sources (this section);
2. **Implementation of a software platform** for experimentation (Section 3.3);
3. Procurement and/or curation of a **re-usable evaluation dataset** (comprising a reference gazetteer and benchmark corpora) (Section 4);
4. Empirical **evaluation** (Section 5) of:
   - the relative utility of heuristics;
   - complete replicated systems.

Table 1 shows the inventory of types of evidence proposed or used for TR over the last decade. For example, $\mathcal{H}_1$ stands for the use of local context patterns: `London, UK` contains two toponyms that match a regular pattern like "`X, Y`" or "`X ( Y )`"), and there is a candidate referent for `X` that is contained in (meronymy) a candidate referent for `Y`. When analyzing the distribution of heuristics (Table 2), our first finding is that there is no concensus about what knowledge contributes most to the task (with the exception of $\mathcal{H}_1$, which is applied almost universally). The reason for this somewhat unprincipled system development is of course the absence of a standard benchmark discussed earlier. However, implementing heuristics whose utility we do not know leads to a potential waste of resources.

Paper/Heuristic	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$	$\mathcal{H}_7$	$\mathcal{H}_8$	$\mathcal{H}_9$	$\mathcal{H}_{10}$	$\mathcal{H}_{11}$	$\mathcal{H}_{12}$	$\mathcal{H}_{13}$	$\mathcal{H}_{14}$	$\mathcal{H}_{15}$
Hauptmann and Olligschlaeger (1999)	■	■		■											
Smith and Crane (2001)	■								■	■	■	■			
Leidner et al. (2003)	■			■	■							■			
Li et al. (2003)	■			■	■						■		■	■	■
Rauch et al. (2003)	■		■										■	■	
Schilder et al. (2004)	■			■		■							■		
Pouliquen et al. (2004)		■	■	■			■	■							
Amitay et al. (2004)	■		■	■	■										
Garbin and Mani (2005)								■				■	■		

Table 2: Distribution of TR Heuristics in the Published Literature.

3 Toponym Resolution Methods

We describe two methods that we have replicated in full, and the TextGIS® software platform which our implementation is based on.

3.1 *PERSEUS* – Focus & sliding window

Smith and Crane's method in the *PERSEUS* digital library system (Smith and Crane (2001)) works as follows (cf. Algorithm 1 for a pseudo-code re-construction). First a bitmap representing the globe is populated with all referents for all mentioned toponyms in a document, weighted by frequency of mention. Then the geometric centroid of all potential referents is computed, and all candidates with a distance greater than two standard deviations from it are discarded. After this pruning, the centroid is updated. Then for each toponym instance in the document, a sliding window containing four toponyms—unambiguous or previously uniquely resolved—to the left and to the right, is constructed. For each referent, a score based on the spatial distance to other resolved toponyms in the context window, the distance to the document centroid, and its relative importance is computed. Relative importance is determined using an order of feature types (country interpretations carry more weight than city interpretations). Finally, the candidate with the highest score is selected.

3.2 *LSW03* – Two minimality heuristics

Leidner et al. propose a method—here called *LSW03* for short—based on *minimality heuristics* (Gardent and Webber (2001)), which combines two interpretational biases (Leidner et al. (2003)):

- $\mathcal{H}_4$: "Assume One Referent per Discourse", the pragmatic version of Yarowsky's principle (Gale et al. (1992)), which postulates that a resolved toponym propagates its interpretation to other instances of the same toponym in the same discourse or discourse segment:

  $\ldots$ London$_{\boxed{1}}$ $\ldots$ London$_{\boxed{2}}$, UK $\ldots$ London$_{\boxed{3}}$ $\ldots$
  $\Rightarrow \boxed{1} \equiv \boxed{2} \equiv \boxed{3} \rightsquigarrow$ `London > England > UK`; and

- $\mathcal{H}_5$: "Assume Spatial Minimality" Leidner et al. (2003), which postulates that the interpretation that (in the absence of explicit evidence to the contrary) minimizes the bounding polygon that contains all candidate referents be selected:

Algorithm 1 Smith and Crane (2001): Centroid-Based Toponym Resolution (PERSEUS).

[Initialize + $\mathcal{H}_0$.]
resolve trivial (unambiguous) toponyms
["Contained-in" qualifier following ($\mathcal{H}_1$).]
match patterns that resolve some toponyms based on local context (e.g. *Oxford, England, UK*)
let M be a 2-dimensional, 1°-resolution map $[\pm 180; \pm 90]$
for all possible toponyms t in a document **do**
  for all possible referents t_r of t **do**
    store $freq(t)$ in M at coordinates for t_r
  end for
end for
[Centroid and pruning ($\mathcal{H}_{10}$).]
compute the centroid c of weighted map M
calculate standard deviation σ from c
for each point associated with any t_r in M **do**
  Discard all points that are more than 2σ away from c
end for
[Centroid re-computation.]
re-compute centroid c
[Sliding window.]
for each toponym instance t in document **do**
  construct a context window w with ± 4 unambiguous or uniquely resolved toponym to the left and to the right of t.
  for each candidate referent t_r of t **do**
    [Scoring ($\mathcal{H}_{9,11,12}$).]
    compute candidate score $s(t_r)$ based on:
    – proximity to other toponyms in w,
    – proximity to c, and
    – relative salience
    (i.e. $s(Spain) > s(Madrid)$)
  end for
  pick as referent un-discarded candidate
  $t_r^* = \arg\max_{t_r} s(t_r)$ unless $s < \theta$
end for

{ *Paris*; Gennevillier; Versailles }
⇒ Paris ⇝ `Paris > France`
{ Bonham; *Paris*; Windom }
⇒ Paris ⇝ `Paris > TX > USA`

Algorithm 2 gives the pseudo-code for the method.[2]

Algorithm 2 Leidner et al. (2003): Minimality-Based Toponym Resolution (LSW03).

[Initialize + $\mathcal{H}_0$**.]**
resolve trivial (unambiguous) toponyms
let S be the cross-product of all candidate referents for each of the N toponyms in a document
["Country" ($\mathcal{H}_{12}$**).]**
for each toponym t **do**
  if t_i has a country interpretation **then**
    pick the country interpretation
  end if
end for
["Contained-in" qualifier following ($\mathcal{H}_1$**).]**
match patterns that resolve some toponyms based on local context (e.g. *Oxford, England, UK*)
["One-referent-per-discourse" ($\mathcal{H}_4$**).]**
for each toponym t **do**
  if t appears resolved elsewhere **then**
    Propagate the resolvent to all unresolved instances
  end if
end for
[Search.]
for each N-tuple $C \in S$ **do**
  [Scoring.]
  create MBR H_C that contains all centroids in tuple C
  compute area $A(H_C)$
end for
[Spatial minimality ($\mathcal{H}_5$**).]**
pick candidate tuple C^* with minimal MBR area:
$C^* = \arg\min_C A(H_C)$ as referents

3.3 Implementation: TextGIS® platform

To re-implement the aforementioned algorithms, we designed a robust and flexible software platform for experimentation with toponym resolution methods and for building applications. Figure 1 shows the system architecture of the resulting *TextGIS*® platform for geo-spatial text mining. An *Infrastructure Layer* provides access to functionality for database access, mapping, named entity tagging and some generic tools (generic API). An *Interface Layer* provides a useful abstraction over details of the representation of data and linguistic markup. It also offers access to non-linguistic knowledge such as population information. The *Resolution Strategy Layer* provides a repertoire of pre-defined resolution strategies, including those compared in this paper. Finally, an *Application Layer* offers tools to perform conversion to RDF, XHTML with links to satellite images, and performance evaluation.

[2] For this evaluation, we use a Minimum Bounding Rectangle (MBR) approximation (Leidner et al., 2003, Footnote 8, p. 33).

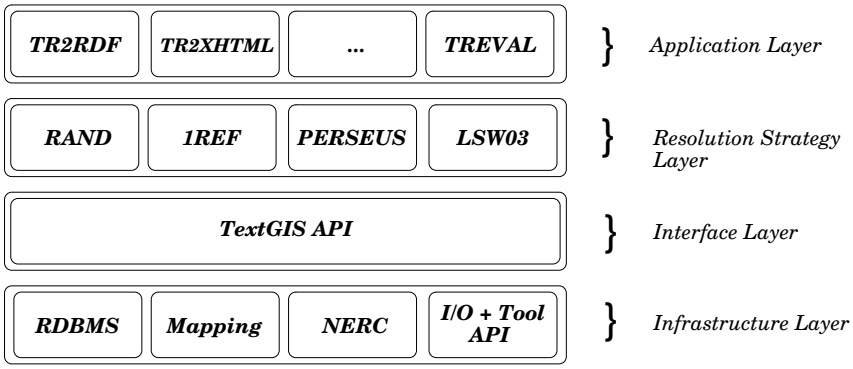

Figure 1: System Architecture of the TextGIS® platform.

4 Dataset

We now characterize the toponym resolution evaluation dataset presented in Leidner (2006), which was supplemented with another corpus for the evaluation described in this report. The aforementioned dataset comprises:

- a large-coverage, short-form **reference gazetteer** with global (earth-wide) focus to look up all candidate referents for each toponym (e.g. `London`). These are represented as hierarchical path (`London > England > United Kingdom`) and decimal latitude/longitude of the location centroids (e.g. lat./long.: ⟨51.52; -0.10⟩). This gazetteer, compiled from US GS, US NGA and US CIA sources, has ≈7.1 million entries.
- **a gold-standard corpus (TR-CoNLL)**, nearly 1,000 news articles from CoNLL (Tjong Kim Sang and De Meulder (2003)) with the correct referents annotated by humans. This corpus was sampled from a well-known source, REUTERS RCV1 (Lewis et al. (2004)), and covers news prose all over the globe.

However, we were also interested in studying the robustness of TR methods by comparing news of varying difficulty. We conjectured that the TR-CoNLL corpus, as global news, would be simpler to deal with than more regional news items. Consequently, we created created a second corpus (**TR-MUC4**) by taking 100 MUC-4 documents (Sundheim (1992)), whose focus is on Central America, and annotating them in a way compatible with the aforementioned corpus (Table 3 compares the two corpora).[3] Note that the human inter-annotator agreement is remarkably lower for TR-MUC4 than for TR-CoNLL. This is caused by the mention of small Central American villages that the annotators had difficulty disambiguating, despite the fact that they were aided with an Internet search engine to retrieve additional information where necessary.

[3] This annotation effort was financially supported by MetaCarta Inc., whose contribution is gratefully acknowledged.

	TR-CoNLL	TR-MUC4
Corpus size (in tokens)	204,566	30,051
Number of documents	946	100
Toponym instances	6,980	278
Unique toponyms	1,299	135
Annotator agreement κ	0.9350	0.7840
Human annotators employed	4	2

Table 3: Evaluation Corpus Profiles.

5 Evaluation

5.1 Metrics

We now present how some traditional performance metrics can be re-cast and used for measuring the quality of a toponym resolution method.
An instance of *London*, after having been identified by the NERC stage (or by an oracle as in this study) as being a toponym[4], is either found in the gazetteer or not, resulting in 0...n readings. If the number of candidate referents is 0 (due to incomplete coverage of the gazetteer), the toponym remains *unresolved*. Otherwise, a mapping to coordinates is attempted, which can either be correct (coordinates represent the intended referent of the toponym) or incorrect (coordinates do *not* represent the intended referent of the toponym). We can thus use standard *Precision P*:

$$P = \frac{\text{\#toponyms resolved} \wedge \text{correct}}{\text{\#toponyms resolved}} \tag{1}$$

Unlike in part-of-speech tagging or text categorization, where the number of categories is small (typically 20-50 tags or document topics), in real-world toponym resolution it is not uncommon for the number of candidate referents (labels, tags) to exceed the the length of the whole document measured in tokens. In addition, the categories (referents) are not shared across toponyms. Metaphorically speaking, each toponym type comes with its very own tag-set, which may be bigger than the corpus itself. In this preliminary evaluation, we work with *Coverage* and define a combined *Toponym Score T*, (similar to F-Score) by relating Precision P to Coverage C using the geometric mean:

$$T_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{C}} \tag{2}$$

$$C = \frac{\text{\#toponyms resolved}}{\text{\#total number of toponyms}} \tag{3}$$

We report $T_{\alpha=0.5}$, which gives equal weight to Precision and Coverage.

5.2 Results

Table 4 shows the TR component evaluation results for the two corpora. The lines give the performance results for a random baseline (RAND), a naïve strategy that only (trivially) resolves non-ambiguous toponyms (1REF), six heuristics from Table 1, and the two complete systems.
Utility of Heuristics. The random baseline has 100% Coverage as it always makes a choice, but its Precision is of course low. 1REF, since is resolves only trivial toponyms, has a Precision of 1; more interestingly, the YAROWSKY assumption also holds for 100% of cases in both corpora. The Precision of the "maximum population" heuristics is high, but its Coverage limits its usefulness (lack of available

[4] Since we rely on two corpora that have gold-standard named entity recognition we control for NERC errors.

TR-CoNLL (gold NERC)	P	C	$T_{\alpha=0.5}$
RAND	0.2982	1.0000	0.3929
$\mathcal{H}_0$ (1REF)	1.0000	0.1179	0.2110
$\mathcal{H}_{0+1}$ (LOCAL)	**0.9382**	0.1305	0.2292
$\mathcal{H}_{0+2}$ (SUPER)	0.3120	0.0744	0.1202
$\mathcal{H}_{0+3}$ (MAXPOP)	**0.6463**	0.2032	0.3093
$\mathcal{H}_{0+4}$ (MINIMALITY)	**0.6290**	0.2324	0.3394
$\mathcal{H}_{0+5}$ (YAROWSKY)	**1.0000**	0.1179	0.2110
$\mathcal{H}_{0+12}$ (COUNTRY)	0.3264	0.3076	0.3167
$\mathcal{H}_{0+1,9-12}$ (PERSEUS)	0.3228	0.3663	0.3431
$\mathcal{H}_{0+1,4+5,12}$ (LSW03)	**0.3679**	**0.6644**	**0.4736**

TR-MUC4 (gold NERC)	P	C	$T_{\alpha=0.5}$
RAND	0.2440	1.0000	0.3923
$\mathcal{H}_0$ (1REF)	1.0000	0.1295	0.2293
$\mathcal{H}_{0+1}$ (LOCAL)	**0.8409**	0.1365	0.2349
$\mathcal{H}_{0+2}$ (SUPER)	0.3922	0.0810	0.1342
$\mathcal{H}_{0+3}$ (MAXPOP)	0.6601	**0.4469**	**0.5330**
$\mathcal{H}_{0+4}$ (MINIMALITY)	0.4194	0.2524	0.3152
$\mathcal{H}_{0+5}$ (YAROWSKY)	**1.0000**	0.1295	0.2292
$\mathcal{H}_{0+12}$ (COUNTRY)	0.6182	0.2881	0.3931
$\mathcal{H}_{0+1,9-12}$ (PERSEUS)	**0.6195**	0.2979	0.4023
$\mathcal{H}_{0+1,4+5}$ (LSW03)	0.3971	**0.5460**	**0.4598**

Table 4: Micro-Averaged Evaluation Results for TR-CoNLL (Top) and TR-MUC4 (Bottom).

population data). However, even the fact that population data may only be available for one of our 25 referents can be a salience indicator for the one referent. "superordinate mention" has the poorest Coverage. The Spatial Minimality Heuristic on its own has a surprisingly high Precision.

Performance of Systems. Both systems perform at low Precision; LSW03 scores higher overall in T_α, not so much because of its slightly higher Precision, but because its Coverage exceeds PERSEUS by a factor of almost two.

Robustness. We are surprised that the country heuristic is stronger on TR-MUC4 than on TR-CoNLL, as intuitively speaking, *Spain* is more likely to always refer to the country in global news than in a regional news. On the (harder) TR-MUC, system performance changes dramatically: PERSEUS doubles its Precision compared to its own performance on TR-CoNLL, and scores one third higher than LSW03, which still has much higher Coverage.

Discussion. The performance of both methods still leaves open much room for improvement, as our evaluation on world-scale scope shows. It would be interesting to combine the two methods implemented so as to inherit from LSW03 high robustness and high Coverage properties and from PERSEUS its superior Precision on regional data, respectively. Also, methods could be applied selectively taking into account the nature of the data (global versus regional).

6 Related Work

After reporting on a pioneering GIR (Geographic IR) system for California, Woodruff and Plaunt conclude that "*although benchmarking is a daunting task, evaluation [of toponym resolution] is extremely significant. Consequently, future work should include the development of a benchmark.*" (Woodruff and Plaunt (1994)). However, in the subsequent decade, no such benchmark materialized. Consequently, Smith and Mann used pseudo-disambiguation instead of a realistic evaluation corpus (Smith and Mann (2003)); they report an accuracy of 87% on the artificial task of using a Naïve Bayes Classifier to recover deleted local disambiguation cues such as "MA" in "Springfield, MA". While this yields some insights into the task difficulty, evaluating a toponym resolver in a realistic scenario is arguably more important. Rauch et al. present MetaCarta Inc.'s *Geographic Text Search* (*GTS*), an industrial-strength GIR system capable of toponym resolution, search and mapping (Rauch et al. (2003)). Unfortunately, no evaluation has been published to date. Li et al. describe Cymfony Inc.'s *InfoXtract*, another commercial system, which is based on minimum spanning tree graph search (Li et al. (2003)). They evaluate their system, on 49 Web pages, containing just 180 toponyms in total. Amitay et al. evaluate their system *Web-a-Where* for the spatial processing of Web pages (Amitay et al. (2004)) on three Web collections of 2k toponyms each and report 2.9% to 3.7% toponym resolution error rate. However, their evaluation is done *a posteriori*, by judging system output, rather than *a priori*, by curating an evaluation corpus with well-known inter-annotator agreement. It is not clear to what extent this methodology adversely affects the results, beyond limiting re-usability. Furthermore, while certainly an interesting application area, Web pages are different in nature from newspaper prose. Garbin and Mani present some interesting weakly supervised learning experiments, inducing decision lists for toponym type disambiguation (Garbin and Mani (2005)), but unlike our experiments using a U.S.-only gazetteer. For the German language, Schilder et al. present a toponym resolver (Schilder et al. (2004)), which is evaluated, but on just 12 newspaper articles (they report a resolution Accuracy of 64%). Unfortunately, besides being for a different language, the articles are not available, and the gazetteer used is smaller by a factor of 1,000 than the one used in this paper. Pouliquen et al. describe a multilingual system for processing geographic names in text documents and look at its named entity recognition capabilities in 8 languages. However, its toponym resolution performance remains un-assessed (Pouliquen et al. (2004)).

To sum up, despite the widely recognized importance of geographic text processing, we do not know of any other large-scale comparative evaluation of toponym resolution methods published to date. In addition, where (mostly toy) evaluations are reported, they are not comparable due to the lack of control for important factors such as influence of a particular gazetteer chosen, a geo-focus or due to use of a corpus that cannot be freely shared. We have attempted to improve this situation by proposing a standard benchmark for the task, and have provided comparative evaluation results.

7 Summary, Conclusions and Future Work

Summary & Conclusions. We have presented a systematic analysis yielding an inventory of heuristics and evidence sources previously used in various approaches to toponym resolution. Using two novel, reusable gold-standard datasets (a gazetteers and two corpora), which we propose as a standard benchmark for the task, we have carried out an empirical evaluation to determine the relative utility of heuristics both individually and in combinations as used in two systems described in the literature, which we have replicated using a new experimental software platform, TextGIS®.

We found that LSW03 lacks Precision, but performs robustly across datasets at high Coverage levels, whereas PERSEUS has low Precision on the less hard dataset TR-CoNLL than on the more difficult TR-MUC4. The reason for this is that it requires a high per-document density of toponyms to be present in order to utilize its strengths (which was only the case in TR-MUC).

To the best of our knowledge, this is both the first comparative TR evaluation and the largest experimental study of toponym resolution on news prose (as opposed to Web pages), using two corpora with over 1,000 documents in total to compare six heuristics and two complex methods across two test sets.

Future Work. In future work, we plan to evaluate more heuristics and replicate further systems to complete our understanding of their relative performance. The scope of this study was restricted to English news prose and a notion of toponym as populated place such as a city, state, country or continent. Future work should include studies that use other languages, text genres other than news, and different location types (including cultural artefacts such as airports, historic monuments etc.). By using machine learning to induce weights for the evidence in a more principled way than by combining heuristics solely based on human intuition, we expect to outperform the state of the art.[5]

Acknowledgements

Thanks to Claire Grover, Bonnie Webber and Dietrich Klakow for their support and guidance, to the German Academic Exchange Service (DAAD) and to the School of Informatics, University of Edinburgh, and MetaCarta Inc. for financial support; to András Kornai for discussions; to David A. Smith for sharing implementation details; to my annotators Annette, Claudine, Darren, Ian and Vasilis; and to the U.S. National Geo-spatial Agency, the U.S. Geographic Survey, and the U.S. Central Intelligence Agency for providing the gazetteer data.

[5] We are making the reference gazetteer and corpora used in this study available (email the first author for details).

References

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-Where: Geotagging Web content. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 273–280. ACM.

Paul Clough and Mark Sanderson. 2004. A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *Workshop on Geographic Information Retrieval held at the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page unnumbered. Association for Computing Machinery, Sheffield, England, UK.

Ian Densham and James Reid. 2003. System demo: A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In Kornai and Sundheim (2003), pages 79–80.

Phil Edmonds and Adam Kilgarriff. 2003. Introduction. *Natural Language Engineering*, 9(1). Special Issue on SENSEVAL-2.

W. Gale, K. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 233–237. Defense Advanced Research Projects Agency, Morgan Kaufmann, San Mateo, CA.

Eric Garbin and Inderjeet Mani. 2005. Disambiguating toponyms in news. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 363–370. Association for Computational Linguistics, Vancouver, British Columbia, Canada.

Claire Gardent and Bonnie Webber. 2001. Towards the use of automated reasoning in discourse disambiguation. *Journal of Logic, Language and Information*, 10(4):487–509.

Alexander G. Hauptmann and Andreas M. Olligschlaeger. 1999. Using location information from speech recognition of television news broadcasts. In Tony Robinson and Steve Renals, editors, *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pages 102–106. University of Cambridge, Cambridge, England.

András Kornai and Beth Sundheim, editors. 2003. *Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL-2003)*. Association for Computational Linguistics, Edmonton, Alberta, Canada.

Jochen L. Leidner. 2004. Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR)*, page (pages unnumbered). Sheffield, UK.

Jochen L. Leidner. 2006. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417. URL http://dx.doi.org/10.1016/j.compenvurbsys.2005.07.003, special Issue on Geographic Information Retrieval. Elsevier Science.

Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. 2003. Grounding spatial named entities for information extraction and question answering. In Kornai and Sundheim (2003), pages 31–38.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. 2003. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In Kornai and Sundheim (2003), pages 39–44.

Bruno Martins, Mário J. Silva, and Marcirio Silveira Chaves. 2005. Challenges and resources for evaluating geographical IR. In Ross Purves and Chris Jones, editors, *Workshop on Geographic Information Retrieval held at CIKM 2005*, pages 65–69. Association for Computing Machinery, ACM Press, Bremen, Germany.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. 2004. Geographical information recognition and visualization in texts written in various languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 1051–1058. ACM Press.

Erik Rauch, Michael Bukatin, and Kenneth Baker. 2003. A confidence-based framework for disambiguating geographic terms. In Kornai and Sundheim (2003), pages 50–54.

Frank Schilder, Yannick Versley, and Christopher Habel. 2004. Extracting spatial information: grounding, classifying and linking spatial expressions. In *Workshop on Geographic Information Retrieval held at the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page unnumbered. Association for Computing Machinery, Sheffield, England, UK.

Andrea Setzer. 2001. *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield, Sheffield, UK.

David A. Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries: Fifth European Conference (ECDL 2001), Darmstadt, Germany, September 4-9, 2001*, pages 127–136.

David A. Smith and Gideon S. Mann. 2003. Bootstrapping toponym classifiers. In Kornai and Sundheim (2003), pages 45–49.

Beth Sundheim, editor. 1992. *MUC-4 — Proceedings of the Fourth Message Understanding Conference*. U.S. Defense Advanced Research Projects Agency (DARPA), Fairfax, VA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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. Association for Computational Linguistics, Edmonton, Alberta, Canada. In association with HLT-NAACL 2003.

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.

A Example Documents

A.1 TR-CoNLL Example (D25)

PRESS DIGEST - France - Le Monde Aug 22 . PARIS 1996-08-22
These are leading stories in Thursday 's afternoon daily Le Monde , dated Aug 23 . FRONT PAGE – Africans seeking to renew or obtain work and residence rights say Prime Minister Alain Juppe's proposals are insufficient as hunger strike enters 49th day in Paris church and Wednesday rally attracts 8,000 sympathisers . – FLNC Corsican nationalist movement announces end of truce after last night 's attacks . BUSINESS PAGES – Shutdown of Bally 's French factories points up shoe industry crisis , with French manufacturers und ercut by low-wage country competition and failure to keep abreast of trends . – Secretary general of the Sud-PTT trade union at France Telecom all the elements are in place for social unrest in the next few weeks . – Paris Newsroom +33 1 42 21 53 81

A.2 TR-MUC4 Example (D27)

4-0017 San Salvador, 17 Sep 88 (DIARIO LATINO)
In recent military operations, Armed Forces units killed one rebel, seized 45,000 cartridges and other war material, and destroyed an underground hideout. The terrorist died in a clash with 6th Infantry Brigade units. The soldiers had detected an FMLN column deploying near La Arana Hill in Estanzuelas, Usulutan Department. The 1st Infantry Brigade also reported that Libertad Battalion troops found an underground warehouse near the El Castano farm in Nejapa, north of San Salvador. In the warehouse, the brigade units found 45,000 cartridges of various calibers and a large amount of material for manufacturing explosives. In addition, Atlacatl Battalion counterinsurgency units recently seized 15 fragmentation grenades, 5 Claymore mines, 5 booby traps, 10 units of TNT, loaded cartridge clips, and 7 knapsacks containing civilian clothing, olive-green uniforms, and communist propaganda.
[...] These items were apparently left behind by wounded rebels who managed to flee after a clash with the battalion members at the foot of El Chino Hill in San Francisco Morazan , Chalatenango Department . Furthermore , 2d Infantry Brigade units patrolling the zone of El Rodeo in El Congo , Santa Ana Department , discovered a 200 - meter underground hideout big enough to conceal at least 250 insurgents . The brigade units proceeded to destroy the hideout .