Toponym Resolution in Text:

Annotation, Evaluation and Applications of Spatial Grounding of Place Names

Jochen L. Leidner

DISSERTATION.COM

Boca Raton

Toponym Resolution in Text:
Annotation, Evaluation and Applications of Spatial Grounding of Place Names

Dissertation.com
Boca Raton, Florida
USA • 2007

ISBN: 1-58112- 384-1
13-ISBN: 978-1-58112-384-5

Toponym Resolution in Text

Annotation, Evaluation and Applications of Spatial Grounding of Place Names

Jochen Lothar Leidner

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2007

Abstract

Background. In the area of Geographic Information Systems (GIS), a shared discipline between informatics and geography, the term *geo-parsing* is used to describe the process of identifying names in text, which in computational linguistics is known as named entity recognition and classification (NERC). The term *geo-coding* is used for the task of mapping from implicitly geo-referenced datasets (such as structured address records) to explicitly geo-referenced representations (e.g., using latitude and longitude). However, present-day GIS systems provide no automatic geo-coding functionality for *unstructured text*.

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 MUC or ACE (Chinchor (1998); U.S. NIST (2003)).

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 (Mani and Wilson (2000); Setzer (2001)), 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 begin by comparing the few algorithms proposed in the literature, and, comparing semi-formal, reconstructed descriptions of them, I 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 in 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 catalogued 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 naïve 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 $51^\circ 32'$ North,

$0^\circ\, 5'$ 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 (Tjong Kim Sang and De Meulder (2003)), and a subset of the Fourth Message Understanding Contest (MUC-4; Chinchor (1995)), both available pre-annotated with gold-standard. This corpus will be made available as a reference evaluation resource;

- 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;

- an *empirical analysis of the relative utility of various heuristic biases and other sources of evidence* with respect to the toponym resolution task when analysing free news genre text;

- 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

- several exemplary *prototypical applications* to show how the resulting toponym resolution methods can be used to create visual surrogates for news stories, a geographic exploration tool for news browsing, geographically-aware document retrieval and to answer spatial questions (*How far...?*) in an open-domain question answering system. These applications only have demonstrative character, as a thorough quantitative, task-based (extrinsic) evaluation of the utility of automatic toponym resolution is beyond the scope of this thesis and left for future work.

To my family.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jochen Lothar Leidner*)

Table of Contents

Abstract **8**

Acknowledgments **19**

1 Introduction **23**

1.1 Motivation 24

1.2 Problem Statement 28

1.2.1 Research Questions 33

1.3 Methodology 34

1.4 Scope 35

1.5 Contributions 36

1.6 Thesis Plan 38

2 Background **41**

2.1 One Space, Many Geographies 41

2.2 Geographic Information Systems (GIS) and Spatial Databases 42

2.3 Gazetteers 51

2.4 Textual Information Access and Natural Language Processing 52

2.4.1 Digital Libraries 52

2.4.2 Information Retrieval 53

2.4.3 Information Extraction 55

2.4.4 Question Answering 61

2.4.5 Word Sense Disambiguation 65

2.5 The Language of Geographic Space 71

2.6 Chapter Summary 75

3 Previous and Related Work **77**

3.1 Processing of Geographic References 77

3.2 Previous Work in Toponym Resolution 81

3.2.1 Hauptmann and Olligschlaeger (1999): TR for Speech Data 82

3.2.2 Smith and Crane (2001): Centroid-based TR 85
3.2.3 Li et al. (2003): A Hybrid Approach to TR 86
3.2.4 Rauch et al. (2003): Confidence-based TR. 91
3.2.5 Pouliquen et al. (2004, 2006): Multilingual TR and Mapping. 93
3.2.6 Amitay et al. (2004): Web-a-Where. 95
3.2.7 Schilder et al. (2004): Cost Optimisation for German TR. 97
3.2.8 Clough (2005): TR in SPIRIT: Source Preference Order. 97
3.3 Comparative Analysis 101
3.4 Chapter Summary 110

4 Dataset 115
4.1 Introduction 115
4.2 Corpus Sampling 116
4.2.1 TR-CoNLL: Global News from REUTERS 117
4.2.2 TR-MUC4: FBIS Central American Intelligence Reports 119
4.3 Annotation Desiderata 119
4.3.1 Referent Representation 119
4.3.2 Problems of Gazetteer Selection 121
4.3.3 Gazetteer Ambiguity and Heterogeneity 122
4.4 Gazetteer 126
4.5 Document Annotation 126
4.5.1 A Simple Markup Scheme 126
4.5.2 Tool-Chain and Markup Process 127
4.6 Result: Corpus Profile 131
4.6.1 TR-CoNLL 131
4.6.2 TR-MUC4 133
4.6.3 Inter-Annotator Agreement 133
4.6.4 Toponym Distribution in Documents 137
4.6.5 Referential Ambiguity in the Corpora 137
4.7 Chapter Summary 139

5 Methods 141
5.1 Introduction 141
5.2 The *PERSEUS* Resolver Replicated: Focus & Sliding Window 141
5.3 A New Algorithm Based on Two Minimality Heuristics 142
5.4 Machine Learning Methods 149
5.4.1 Introduction 149
5.4.2 Decision Tree Induction (DTI) 150

5.4.3 Outlook: Learning Voting Weights . . . 152
5.5 The TextGIS® Toolkit: Design and Implementation . . . 153
5.5.1 Introduction: Software Architecture for Language Engineering . . . 153
5.5.2 Design . . . 155
5.5.3 Implementation . . . 155
5.6 Chapter Summary . . . 156

6 Evaluation 159
6.1 Introduction . . . 159
6.2 Evaluation Methodology . . . 160
6.2.1 Adapting Traditional Evaluation Metrics . . . 160
6.2.2 Task-Specific Evaluation Metrics . . . 165
6.3 Component Evaluation Using a Named Entity Oracle ('in vitro') . . . 168
6.4 Component Evaluation Over System Output ('in vivo') . . . 173
6.4.1 Using a Maximum Entropy NERC Model . . . 175
6.5 Discussion . . . 180
6.6 Chapter Summary . . . 181

7 Applications 183
7.1 Visualization: Bridging Text and Space by Hyperlinking to Satellite Images . . . 183
7.2 Summarisation: Generating Map Surrogates for Stories . . . 184
7.3 Exploration: Geo-Spatial News Browsing . . . 188
7.4 Search: Spatial Filtering for Document Retrieval . . . 194
7.4.1 Method: Geo-Filtering Predicates . . . 194
7.4.2 Evaluation in a GEOCLEF Context . . . 200
7.4.3 Discussion . . . 206
7.5 Question Answering: Knowledge-Based Approach . . . 207
7.6 Chapter Summary . . . 210

8 Summary and Conclusion 211
8.1 Summary of Contributions . . . 211
8.2 Future Work . . . 213
8.3 Conclusions . . . 215

A Notational Conventions 217

B Annotation Guidelines 219

C Minimal Bounding Rectangles Extracted from NGA 221

D TR-CoNLL Sample Used in 'Prose Only' Evaluation **223**

E TR-CoNLL Evaluation (All Documents Used) **225**

F Performance Plots for Individual Heuristics **229**

G Stories Used in the Visualization Study **235**

G.1 Story 'Royal Mercy Flight Baby Dies' 235

G.2 Story 'News Crews Wait and Watch as Police Search Home of Missing Woman' 236

H Distance Queries **241**

I ADL Feature Type Thesaurus **249**

Summary (in German) **257**

Bibliography **261**

List of Tables

1.1 Synopsis of the symbols used. 30

2.1 Some gazetteers available in digital form (accessed 2006-08-01). 52
2.2 MUC/MET: achieved performance (modified after Chinchor 1998). 60
2.3 Examples for geo/geo and geo/non-geo ambiguity. 73
2.4 Examples of place-for-event metonymy. 74

3.1 Weight function in InfoXtract (after Li et al. (2003)). 89
3.2 Summary of the state of the art in toponym resolution. 112
3.3 Synopsis of heuristics used by previous work. 114

4.1 Different kinds of spatial annotation. 120
4.2 Comparison of gazetteer density. 124
4.3 Most referentially ambiguous toponyms with respect to four different gazetteers. 125
4.4 Evaluation corpus profiles. 131

5.1 Implemented heuristics. 156

6.1 Toponym resolution evaluation: calculation example. 164
6.2 Micro-averaged evaluation results for TR-CoNLL on human oracle NERC tags (subset). MINIMALITY and LSW03 have the highest absolute scores (for R and F1) and significantly outperform RAND, but not MAXPOP. LSW03 outperforms PERSEUS in absolute terms (for all three metrics). 168
6.3 Micro-averaged evaluation results for TR-MUC4 on human oracle NERC tags. Despite its simplicity MAXPOP significantly outperforms all heuristics and complex systems on this dataset. 169
6.4 *C&C* NERC performance on CoNLL 2003 `eng.train` trained on default MUC-7 model. 176

6.5 Micro-averaged evaluation results for TR-CoNLL (subset) on MaxEnt-tagged data. MAXPOP has highest absolute precision and F1-score overall, though not significantly different from LSW03 (at the 5% level), while PERSEUS' 5% lower absolute F1 performance means it is outperformed by MAXPOP. . . 179
6.6 Micro-averaged evaluation results for TR-MUC4 on MaxEnt-tagged data. MAXPOP has highest absolute performance also on automatic named entity tags (F1-score statistically on par with PERSEUS, but significantly outperforming every other method). MAXPOP and PERSEUS are significantly superior to LSW03 in this setting. 180

7.1 List of the most frequent toponyms in the GEOCLEF corpus. Toponyms in bold type are artifacts of the Glasgow/California bias of the corpus. 195
7.2 Minimal bounding rectangles (MBRs) from the Alexandria and ESRI gazetteers. MBRs are given as pairs of points, each with lat/long in degrees. A dash means that no result was found or that a centroid point was available only. 199
7.3 Result summary: Average Precision and R-Precision. 201

E.1 Micro-averaged evaluation results for TR-CoNLL collection (all documents used) for automatic toponym resolution on human oracle NERC results ('in vitro'). LSW03 is the strongest method on the whole, unfiltered dataset, outperforming both MAXPOP and PERSEUS. 228
E.2 Micro-averaged evaluation results for TR-CoNLL collection (all documents used) for automatic toponym resolution on automatic (MaxEnt) NERC results ('in vivo'). LSW03 outperforms all other methods in terms of F1-score, but it is not significantly different from MAXPOP. PERSEUS performs significantly worse than LSW03 ($p < 0.001$) in this setting, which is not shown but can be inferred. 228

List of Figures

1.1 Disciplines concerned with geographic space. 24
1.2 Discourse model. 31
1.3 Types of grounding in language processing. 32
1.4 The dual role of toponym resolution and toponym generation in connecting text and geographic space. 33

2.1 Data representation: RDBMS versus GIS: (a) relational data, (b) raster data, and (c) vector data. 43
2.2 Quad-tree example. 44
2.3 Internet GIS application showing parts of Hunterdon (New Jersey, USA). . . . 45
2.4 Mercator projection of the earth (created with GMT). 47
2.5 The geo-coding process, modified after (Crosier, 2004, p. 38). 48
2.6 Example geo-coder: Eagle. 49
2.7 Sample locality descriptions from herbarium specimen records from (Beaman and Conn, 2003, page 48). 51
2.8 The International Children's Digital Library: Meta-data (top) and a sample page (bottom). 54
2.9 Spatial pipeline in a geography-aware retrieval system. 56
2.10 Message Understanding Contest history. 57
2.11 MUC-7 named entity sub-types. 58
2.12 Toponym resolution as an additional NE processing layer. 61
2.13 The *QED* Q&A system architecture (Leidner et al. (2004)). 64
2.14 SENSEVAL3: the majority of systems performed between 40% and 70% precision and recall. 67
2.15 Yarowsky's algorithm at work: (a) initial state, (b) intermediate state, (c) final state (Yarowsky, 1995, p. 191-2). 69

3.1 Geographic focus computation using polygon intersection (after Woodruff and Plaunt (1994)): the polygon model shows California (North is left). 79

3.2 Representation for 'Tous le départment du nord de la France' (after Bilhaut et al. (2003)). 81
3.3 Screen capture of the Web interface to the Perseus digital library. 85
3.4 Maximum-weight spanning tree applied to toponym resolution. 90
3.5 Entries for 'Cambridge' in the TIPSTER gazetteer. 92
3.6 Taxonomy of sources of evidence for toponym resolution decision making. . . 109

4.1 Example document D307. Toponyms (named entities tagged `LOCATION`) in the original corpus are underlined. 118
4.2 CoNLL format (excerpt). 120
4.3 Gazetteer profiles. 121
4.4 Gazetteer ambiguity (number of gazetteer entries as a function of the number of candidate referents). 123
4.5 TRML format (excerpt). 128
4.6 The TAME system architecture. 132
4.7 TAME, the Toponym Annotation Markup Editor (screen capture). 133
4.8 Geographic distribution of the locations in TR-CoNLL (top) and TR-MUC4 (bottom). 135
4.9 TR-CoNLL: from unresolved (top) to resolved (bottom). 136
4.10 Toponym distribution in discourse of TR-CoNLL document D19. 138
4.11 Number of documents in the sample as a function of toponym occurrences. . . 139
4.12 Distribution of referent frequency. 140

5.1 Illustration of the spatial minimality principle. 145
5.2 Spatial minimality algorithm at work (trace). 148
5.3 Toponym resolution with spatial minimality: examples. 149
5.4 Supervised machine learning for toponym resolution. 150
5.5 TR-CoNLL feature vectors: structure (top) and examples (bottom). 151
5.6 Heuristic ensembles with weights. 152
5.7 Developer productivity versus re-usability in software systems (Leidner (2003a)). 154
5.8 System architecture of the TextGIS® toolkit. 155
5.9 All resolution strategies implement the `ToponymResolver` C++ interface. . . . 156
5.10 TextGIS, the main class of the TextGIS API. 157

6.1 Three cases in TR evaluation. 161

6.2 Performance of the two systems and two baselines on TR-CoNLL as a function of the F-score's β parameter on gold standard data. The more emphasis is placed on recall, the closer (and lower) MAXPOP and PERSEUS get in F-score. Obviously, LSW03 outperforms both baselines as well as PERSEUS for $\beta > 0.9$, and the more so the higher recall is weighted. LSW03 is thus superior in applications where missing a relevant item comes at a high price, such as patent search or intelligence analysis. 171
6.3 Performance of the two baselines and two systems on TR-MUC4 as a function of the F-score's β parameter on gold standard data. MAXPOP by far outperforms other methods for all weightings between P and R considered. LSW03 beats PERSEUS for $\beta > 1.65$. 172
6.4 Errors can be introduced at three levels. 174
6.5 Performance of the two systems and two baselines on TR-CoNLL ('clean prose'-only subset) as a function of the F-score's β parameter using MaxEnt toponym tagging. PERSEUS' performance stays well under the MAXPOP baseline, but they show convergent behaviour for large values of β. LSW03 outperforms the MAXPOP baseline in scenarios where high recall is vital. 177
6.6 Performance of the two systems on TR-MUC4 as a function of the F-score's β parameter using MaxEnt toponym tagging. MAXPOP outperforms its competitors for all β settings. 178

7.1 Hyperlinking toponyms in a text document with maps and satellite images. . . . 185
7.2 Textual geo-spatial document surrogates for the stories in Appendices G.1 and G.2. 186
7.3 Map surrogate generation process. 187
7.4 Automatic visualization of story G.1: a baby flown from London to Glasgow for medical treatment dies there. 189
7.5 Automatic visualization of story G.2: a pregnant woman is missing in Modeno, CA (local view; final paragraph excluded). 190
7.6 Story G.2: the final paragraph places the event in context (global view; complete story). 191
7.7 Generated KML for TR-MUC4 document D35. 193
7.8 TextGIS® integration with Google Earth. 194
7.9 Toponym resolution using the maximum-population heuristic. 196
7.10 Performance of the run `LTITLE` (average precision). 202
7.11 Performance of the run `LCONCPHRSPATANY` (average precision). 203
7.12 Individual topic performance (1-25) relative to the median across participants: run `LTITLE`. 204

7.13 Individual topic performance (1-25) relative to the median across participants: run LCONCPHRSPATANY. 205
7.14 Google fails to answer a distance question (Q-503678). 209

C.1 Bounding Rectangles for countries in Europe (left) and North America (right). . 221
C.2 Bounding Rectangles for countries in Central America (left) and South America (right). 222
C.3 Bounding Rectangles for countries in Africa (left) and Australia (right). 222
C.4 Bounding Rectangles for countries in Asia. 222

E.1 Performance of the two systems against two baselines on TR-CoNLL (complete corpus) as a function of the F-score's β parameter on gold standard data. LSW03 is able to outperform MAXPOP for $\beta > 0.75$ and PERSEUS for all weightings between P and R considered. F1(MAXPOP) drops below F1(MAXPOP) in high-recall settings ($\beta > 1.75$). 226
E.2 Performance of the two systems against two baselines on TR-CoNLL (complete corpus) as a function of the F-score's β parameter using MaxEnt toponym tagging. Obviously, reporting F1 only conceals the fact that LSW03 as very good recall, which lets it outperform MAXPOP and all other methods reported here beyond $\beta = 1$. 227

F.1 Plot of the performance of the heuristics and two baselines on TR-CoNLL (subset) as a function of the F-score's β parameter on gold standard NERC. The heuristics used in isolation perform almost identically at $\beta = 0.5$. The MINIMALITY heuristic is very competitive for $\beta > 0.75$. 230
F.2 Plot of the performance of the heuristics and two baselines on TR-CoNLL (all documents) as a function of the F-score's β parameter on gold standard NERC. 231
F.3 Plot of the performance of the heuristics and two baselines on TR-CoNLL (subset) as a function of the F-score's β parameter on MaxEnt NERC. 232
F.4 Plot of the performance of the heuristics and two baselines on TR-CoNLL (all documents) as a function of the F-score's β parameter on MaxEnt NERC. . . . 233

Acknowledgements

First and foremost, I would like to express my gratitude to Claire Grover and Bonnie Webber, who have been great advisers. While my wide-ranging interests and activities have probably stretched their patience at times, they were always supportive, extremely helpful and ready to provide feedback and guidance. I am well aware that getting written comments on an email-ed paper draft within just a few hours is not something that many other PhD students can benefit from. Their experience, knowledge and kindness will serve as a role model for me beyond this thesis project. Thanks!

Steve Clark has been a supportive third adviser until his departure to Oxford and (no longer formally, but no less supportive) beyond. Bruce Gittings in the Department of Geography was always happy to chat about all matters geographical. Also many thanks to Ewan Klein and Dave Robertson for being such helpful internal examiners. Thanks to Richard Tobin and John Tait for valuable discussions on evaluation methodology, and to Mark Sanderson and Jon Oberlander, who form my thesis committee, for reading this thesis.

Edinburgh is a wonderfully vibrant and magically productive place, and the Potteresque maze of Buccleuch Place is home to some of the most brilliant researchers in natural language processing. Many people have wondered why this is so, and one possible explanation could be the phenomenon locally known as the Blind Poet (not a Potter novel—yet). Unlike in many other places, research in Edinburgh is always seen as something fun, and fun things are more fun if they are shared, so the boundaries of work and play are blurred.

Kisuh Ahn, Beatrix Alex, Amittai Axelrod, Markus Becker, Johan Bos, Jean Carletta, Heriberto Cuayahuitl, Johannes Flieger, Ben Hachey, Harry Halpin, Pei-yun Hsueh, Amy Isard, Frank Keller, Yuval Krymolowski, Mirella Lapata, Colin Matheson, Johanna Moore, Malvina Nissim, Jon Oberlander, Miles Osborne, David Reitter, Gail Sinclair, Andrew Smith, Mark Steedman, David Talbot, Tim Willis and many, many others provided feedback in numerous discussions, were fun to hang out with in the pub and on parties, but often both, even at the same time.

Ewan Klein, Harry Halpin and Sebastian Riedel were fascinating discussion partners in the GridNLP group; hopefully, some of the things we brainstormed will become standard.

Many thanks also to the system group for maintaining the DICE computing evnironment and to the administration staff for being so well-organised and proactive.

I am grateful for having such great friends during my PhD, in Edinburgh and elsewhere. Thanks to (in alphabetical order) Andrew, Annette, Anita, Andreas, Daniel, Carsten, Chia-Leong, Claudine, Christine (two of them), David, Dörthe, Francesca, Hannele, Jeanette, Maciej, Matthias, Michael, Vera, Priscilla, Rana, Rob, Sibylle, Tiphaine, and Yves, for your friendship, for keeping me sane, and for reminding me there is more to Edinburgh, the world and to life than academia.

Colin Bannard, Chris Callison-Burch, Nicole Kim, Manolis Mavrikis, Mark McConwell, Rafael

Morales, Victor Tron and Verena Rieser were not only pleasant office mates to have, they also ensured wide topic diversity of our daily office chat. Thanks to my flatmates Alice, Gabrielle, Georgios, and Manolis, for accepting a German bloke into their predominantly French and Greek flats, respectively, and for being great pals. Special thanks to Vera for her dear companionship, for many cups of peppermint tea together, for advice on *R*, and for discussions on statistics.

With Kisuh Ahn, Tiphaine Dalmas, Johan Bos, James Curran and Steve Clark (also known as the two 'C's in *C&C*) I have shared the unique experience of building the *QED* open-domain question answering system and 'doing TREC' together, which involved all things true geeks need to function, such as overnight hacking sessions or pizza-and-DVD self-rewards, plus some less common features such as Settler boardgame sessions, Scottish dancing (online and offline), and accordeon-accompanied evaluation runs.

Dietrich Klakow very kindly hosted me in his speech signal processing lab (LSV) at Saarland University, first as part of an International Graduate College 8-month exchange between Edinburgh and Saarbücken, and then gave me a research job that allowed me to continue to work on this thesis after my DAAD scholarship. Not only that, Dietrich was very supportive and the source for interesting conversations, always helpful and happy to provide feedback on any issue. In Saarbrücken, Andreas Beschorner was my amicable (and sometimes composing) office mate, and many items containing chocolate in one form or another were jointly disposed off while writing papers, debugging code, or marking student exercises. Many thanks also to Barbara Rauch, Irene Cramer and Andreas Merkel for many pleasant dinners, discussions, and ubiquitous joint walks to the campus supermarket to obtain ice cream to fight the summer heat, and to the rest of the LSV group for having me around.

Another round of thanks to my academic co-authors during my PhD period: Kisuh Ahn, Beatrice Alex, Colin Bannard, Johan Bos, Chris Callison-Burch, Steve Clark, James Curran, Irene Cramer, Tiphaine Dalmas, Claire Grover, Dietrich Klakow, Ewan Klein, Yuval Krymolowski, Harry Halpin, Stephen Potter, Sebastian Riedel, Sally Scrutchin, Matthew Smillie, Mark Steedman, Richard Tobin and Bonnie Webber; it was (and is) fun to work with you.
Travelling to (or hanging out at) conferences with András, Chris, James, Johan, Markus, Miles, Olga, Steve and Tiphaine was always lots of fun (and often full of adventures).

The NLP, IR and GIS communities were a pleasant environment for study and research. Jean Carletta, Paul Clough, Fred Gey, Alex Hauptmann, Chris Jones, András Kornai, Marcus Kracht, Ray Larson, Douglas Oard, Andreas 'Olli' Olligschläger, John Prange, Ross Purves, Douglas E. Ross, Mark Sanderson, David A. Smith, Ralf Steinberger, John Tait, Erik Tjong Kim Sang, Yannick Versley, Richard Waldinger and countless others provided valuable input in numerous discussions in emails and on conferences spanning the very globe that is dealt with in this thesis.

The author is also grateful to the U.S. National Geospatial Intelligence Agency (NGA), the U.S. Geographic Survey (USGS) and the U.S. Central Intelligence Agency (CIA) for providing the gazetteer datasets, without which this research project would not be possible in its present form (especially regarding its scale). The Freedom of Information Act that made the data release possible is one of the greatest pieces of legislation since the Geneva Convention on Human Rights (if only the latter was as consistently applied as the former). I also thank my annotators Annette, Claudine, Darren, Ian and Vasilis.

A very practical 'toponym resolver', Fred Bull, from Aberdeen, Scotland, travelled 95,438 miles to visit most *other* Aberdeens on our globe from Jamaica to Hong Kong. Thanks to Fred for inviting me to his book launch party where he shared his experiences when meeting members of the global family of 'Aberdonians worldwide' (Bull (2004)). Unlike this thesis, which is bound to concentrate on a narrow technical topic, his book emphasises that places are founded by people, many of whom share a common history.

This research was funded by the German Academic Exchange Service (DAAD) under the three-year scholarship D/02/01831 and by Linguit GmbH under grant UK-2002/2. Financial support by the School of Informatics, University of Edinburgh, and a Socrates scholarship by the European Union are also gratefully acknowledged. The contribution of MetaCarta Inc. to the funding for annotating the TR-MUC4 corpus is likewise gratefully acknowledged. The author is further grateful to ACM SIGIR for a generous travel stipend as well as a useful stack of books accompanying an ACM SIGIR Doctoral Consortium Award.

Last but not least, I'm grateful for the endless love and support of my mother and grandparents; I dedicate this thesis to them.

Chapter 1

Introduction

Then I found out that there was a place called Black in every state in the country, and actually in almost every country in the world.
– Jonathan Safran Foer (2005), Extremely Loud & Incredibly Close, p. 42

[Parts of this chapter have been published as Leidner et al. (2003) and Leidner (2004a).]

Space and time are two fundamental dimensions of human perception that we use to organize our experiences. Consequently, documents, as textual artefacts of human experience (real or fictitious), make frequent use of expressions of space and time as points of reference.

With the availability of large amounts of textual data on computer networks and the parallel availability of increasingly powerful computing devices, information systems for spatial and textual processing have been developed. To date, the automatic processing of text is investigated by the discipline of Natural Language Processing (NLP, comprising sub-fields such as automatic information extraction and automatic question answering), whereas the processing of spatial information is investigated by the discipline of geographic information systems. The existing split into research disciplines is perhaps understandable, given the different nature of textual and spatial data at the surface level, and the heritage of the disciplines, rooted in linguistics and geography, respectively (Figure 1.1). However, as a negative consequence of this organizational divide, the full power of the data remains under-utilised: conventional information systems are unable to relate a text document reporting on a riot in Somalia with sensor data (such as satellite imagery) covering exactly the spot where the riot took place.

It is the aim of this thesis to contribute to a wider effort to *'bridge text and space'*, which I call for, and the objective of which should be to overcome the technical and organisational divides that prohibit the co-computation of textual and spatial relationships, mutually supporting