Where are you talking about?

Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring

Milan Gritta

Supervisor: Dr Nigel Collier

Department of Theoretical and Applied Linguistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Fitzwilliam College — February 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Milan Gritta
February 2019

Where are you talking about? Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring

Abstract - Milan Gritta

The Natural Language Processing task we focus on in this thesis is *Geoparsing*. Geoparsing is the process of extraction and grounding of toponyms (place names). Consider this sentence: "The victims of the *Spanish* earthquake off the coast of *Malaga* were of *American* and *Mexican* origin." Four toponyms will be extracted (called Geotagging) and grounded to their geographic coordinates (called Toponym Resolution). However, our research goes further than any previous work by showing how to distinguish the literal place(s) of the event (*Spain, Malaga*) from other linguistic types/uses such as nationalities (*Mexican, American*), improving downstream task accuracy. We consolidate and extend the Standard Evaluation Framework, discuss key research problems, then present concrete solutions in order to advance each stage of geoparsing. For geotagging, as well as training a SOTA neural Location-NER tagger, we simplify Metonymy Resolution with a novel minimalist feature extraction combined with an LSTM-based classifier, matching SOTA results. For toponym resolution, we deploy the latest deep learning methods to achieve SOTA performance by augmenting neural models with hitherto unused geographic features called Map Vectors. With each research project, we provide high quality datasets and system prototypes, further building resources in this field. We then show how these geoparsing advances coupled with our proposed Intra-Document Analysis can be used to associate news articles with locations in order to monitor the spread of public health threats. To this end, we evaluate our research contributions with production data from a real-time downstream application to improve *geolocation of news events for disease monitoring*. The data was made available to us by the Joint Research Centre (JRC), which operates one such system called MediSys that processes incoming news articles in order to monitor threats to public health and make these available to a variety of governmental, business and non-profit organisations. We also discuss steps towards an end-to-end, automated news monitoring system and make actionable recommendations for future work. In summary, the thesis aims are twofold: (1) Generate original geoparsing research aimed at advancing each stage of the pipeline by addressing pertinent challenges with concrete solutions and actionable proposals. (2) Demonstrate how this research can be applied to news event monitoring to increase the efficacy of existing biosurveillance systems, e.g. European Commission's MediSys.

I would like to dedicate this thesis to my loving parents who worked extraordinarily hard to give me this opportunity. Thank You.

Acknowledgements

I would like to acknowledge the generous funding support of the Natural Environment Research Council (NERC) via the interdisciplinary Centre of Doctoral Training (CDT) called Data, Risk, Environmental and Analytical Methods (DREAM[1]) under the leadership of Dr Stephen Hallett and Co. of Cranfield University and more locally, Dr Mike Bithell of Cambridge University. The CDT is a great model for PhD students, broadening our horizons, offering excellent support and developing non-academic skills through diverse experiences. Naturally, I want to recognise Dr Nigel Collier for his dedicated supervision throughout the 3-4 years, for co-authoring my papers and having the confidence in me as his next PhD student back in 2015. I also want to thank Dr Anna Korhonen for her dedicated (second) supervision and support throughout the PhD process and for expertly examining the thesis (internal). I further want to thank Dr Mark Stevenson[2] from Sheffield University for agreeing to examine my thesis (external) and for providing excellent feedback. I want to acknowledge my fellow authors Dr Nut Limsopatham (2015 - 2017) and Dr Mohammad Taher Pilehvar (2015 - 2019) for their continued advisory support and feedback during manuscript and thesis publication. I extend my gratitude to our industrial partner Dr Jens Linge from the Joint Research Centre of the European Commission for his feedback, collaboration and for hosting me at the JRC in Ispra, Italy in February 2018; the Fitzwilliam College Graduate Officer Sue Free for her support, particularly during my 2018 intermission due to family issues; Savina Velotta and Angela Colclough from DREAM CDT for ensuring the events felt homely and welcoming, enabling us to focus on learning. Last but not least, I thank Cambridge University students Mina Frost and Qianchu (Flora) Liu for providing linguistic expertise; the NVIDIA GPU Grant Program for the free graphics card; my fellow LTL students for their company, support and advice. Finally, *Thank You* to all who helped out in any way big or small, somehow during my thesis submission and new job busyness, I forgot to acknowledge you, you know who you are, you deserve it. Thank You Very Much!

[1] http://www.dream-cdt.ac.uk/

[2] https://www.sheffield.ac.uk/dcs/people/academic/mstevenson

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The aim of this thesis is to advance research into techniques for geoparsing that result in measurable performance improvements on both intrinsic and extrinsic evaluations. Geoparsing is the process of extraction and grounding of toponyms (place names). From the following sentence, "The victims of the *Spanish* earthquake off the coast of *Malaga* were of *American* and *Mexican* origin.", four toponyms will be extracted (this task is called Geotagging) and grounded to their geographic coordinates (this task is called Toponym Resolution). The thesis substantially advances both stages while demonstrating how to utilise the extracted information to improve document geolocation for disease monitoring in breaking news (extrinsic evaluation). In the preceding example, interpretation errors are almost certain, i.e. *if* all four toponyms get extracted, which is by no means a given, no distinction will be made between the *types* of these toponyms. Only *Spanish* and *Malaga* refer to literal locations while *American* and *Mexican* refer to nationalities, which introduces ambiguity into downstream tasks. Other problematic examples include *Chinese clock* versus *Chinese coast*, only one denoting a *physical place*. In addition to high precision and recall errors, *Malaga* can be resolved to several locations in predominantly Spanish-speaking countries, further compounding errors (particularly severe for smaller/infrequent toponyms). In this thesis, we take a deep dive into both stages to significantly reduce such errors while also taking a broad view of geoparsing, critically reviewing the evaluation process given the existing resources/papers. We then communicate our research to the Joint Research Centre to advise on potential implementation of our contributions for their alerting system called *MediSys*. This chapter lays out the essentials of the thesis, major contributions, key partnerships, expected format, published papers and related work in biosurveillance.

An excellent article on *research debt*[1] uses the analogy of *technical debt*, i.e. trading the speed of development and feature addition for the long-term, sound software development principles. In an effort to achieve the next SOTA score while digesting a stream of incoming specialised and general research, we are inevitably accumulating research debt in the form of somewhat disconnected vertical research niches, unfinished ideas and/or incomplete exposition. When we entered the geoparsing field, we noticed this 'research debt' as we struggled to find previous research that could serve as a *convenient, focused yet comprehensive springboard* into geoparsing. To this end, the thesis contributes with original research while trying to distil the latest state of the geoparsing landscape by pointing to the most practically relevant and pressing challenges, providing concrete solutions for many and proposing actionable next steps to address the remaining problems.

1.1 Preamble

1.1.1 Thesis Format

The thesis proceeds in chronological order, describing research *as it happened*, the challenges and discoveries as they were initially perceived and addressed. This approach aims to give the reader a researcher's perspective on the progress made in the last 3.5 years. Chapters are introduced with relevant commentary that gives an up-to-date perspective on earlier research and how that may have changed given today's state of knowledge. Each chapter has a summary with proposals for further work, if appropriate. One may think about Chapters 2, 3, 4 and 5 as *intrinsic evaluation* with each focusing on a separately published paper and Chapter 6 as *extrinsic evaluation* applying research proposals from prior chapters in the context of a specific application (the MediSys news monitoring system). Telling the story of the research progress as it happened means there may be similarities throughout the thesis as topics were revisited and extended in the light of new knowledge. Academic research is not a linear process hence the natural presentation of contributions in the thesis reflects this.

1.1.2 Summary of Contributions

The major contributions of the thesis are outlined below, grouped into six research themes.

The Geoparsing Task and its Evaluation Framework have been clarified, critically reviewed and restated. To that end, we added a comprehensive annotation scheme via a new

[1] https://distill.pub/2017/research-debt/

taxonomy of toponyms describing their linguistic properties in unprecedented depth. The performance metrics have been reviewed and consolidated given the usage of multiple (unequally suitable) evaluation metrics. As part of this theme, we analysed available geoparsing tools and outlined their implications for Named Entity Recognition. Another important aspect of the task that we have tested in multiple publications are the properties of datasets and the hypothetical maximum performance per dataset. These core geoparsing concepts such as the availability of knowledge bases and their alignment for benchmarking purposes instil further rigour, fairness and clarity into the evaluation process. The thesis and its manuscripts shall serve as a convenient, comprehensive and practical entry point into the field that we perceived was missing in somewhat segmented and compartmentalised prior work. This theme of the thesis, published as two papers in the journal *Language Resources and Evaluation in 2017, and 2019*, is split between several chapters due to its thoroughness and depth.

Metonymy Resolution is the substitute of the intended concept with a related one, e.g. "The tariffs were imposed by *Washington*." where *Washington* was substituted in place of the US government. We devised a novel minimalist feature extraction method called *PreWin* based on predicate-argument dependencies. PreWin is designed to extract as few as three words from the context and still outperform the baseline (greedy) methods, which extract 5, 10 or even 50 words. Our approach contrasted with the previously manually engineered features. We then deployed a minimalist neural network, which achieved near SOTA scores on the SemEval 2007 Metonymy Resolution task as well as on our new dataset (with an improved annotation scheme). This research path was our first successful foray into non-literal toponym identification, heavily extended in Chapter 5. The manuscript was published in the 'outstanding paper' track at the ACL Conference[2] in 2017.

The Open-Source Resources we have generated shall help accelerate geoparsing research as new activity may be limited by the lack of quality domain data and SOTA models. Our datasets are suitable for training and evaluation and shall address the dearth of recent geoparsing resources. We published GeoWebNews, WikToR, ReLocaR, GeoVirus and CoNLL 2003 (a subset annotated for metonymy resolution) on GitHub[3], which already inspired subsequent work such as EUPEG [86] and several emails of gratitude for providing the resources and code for possible extensions. We hope that our open data will facilitate easy replication in geoparsing and beyond.

[2] http://acl2017.org/

[3] https://github.com/milangritta

The CamCoder and the Map Vector are novel concepts in Toponym Resolution (TR) presented in Chapter 4. TR is a disambiguation task that requires the correct assignment of geographic coordinates to a toponym (a place name). We present the highly effective CNN-based Geocoder (CamCoder), which was subsequently augmented by generating novel features encoding geographic knowledge hitherto unused in TR and related domains. This representation is called the Map Vector and it is a fully interpretable geographic feature vector enabling SOTA scores on three diverse datasets (and a new MediSys dataset called GeoWebNews) compared with multiple systems and strong baselines. The long paper was published at the ACL Conference[4] in Melbourne, Australia in July, 2018.

The FlexiGraph is a prototype for intra-document analysis designed to enhance document geolocation. This proposal addresses the needs of the MediSys engineers expressed during my visit to the Joint Research Centre in early 2018. There was a need to effectively and efficiently associate the event trigger words e.g. "zika", "earthquake", "terrorist" with the extracted toponyms in order to determine the event location more reliably. This is because news articles and other similar text documents usually contain multiple location references, often unrelated to the event. FlexiGraph is a possible solution with the potential to significantly increase event geolocation capability. It also meets the software engineering constraints of the production version of MediSys.

The Evaluation of MediSys in Chapter 6 focuses on the geoparsing performance of the biosurveillance system[5]. The research provides deep insights into its news feed composition, sources of common errors and their magnitude, their likely impact on event geolocation, followed by detailed actionable recommendations for the configuration of several enhanced systems. The analysis evaluates our 'laboratory research' with production data and makes proposals grounded in a real-time geoparsing application. We show how errors change with different types of locations, datasets and documents (news articles) in order to anticipate and alleviate many predictable challenges. We include comparative SOTA geoparsing systems for a contrasting analysis, test their performance and assess their suitability for a possible inclusion in an automated geolocation model. We think our extensive evaluation of the geoparsing component of MediSys will provide valuable quantitative and qualitative information for its future development.

[4] https://acl2018.org/

[5] http://medisys.newsbrief.eu

1.1.3 Published Papers

2017 (Chapter 2) Gritta, Milan, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. "What's missing in geographical parsing?" Language Resources and Evaluation 52, no. 2 (2017): 603-623.

2017 (Chapter 3) Gritta, Milan, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. "Vancouver Welcomes You! Minimalist Location Metonymy Resolution." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1248-1259.

2018 (Chapter 4) Gritta, Milan, Mohammad Taher Pilehvar, and Nigel Collier. "Which Melbourne? Augmenting Geocoding with Maps." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1285-1296.

2019 (Chapter 5) Gritta, M., Pilehvar, M.T. and Collier, N., 2018. A Pragmatic Guide to Geoparsing Evaluation. arXiv preprint arXiv:1810.12368. The paper was accepted for publication at Springer's Language and Resources journal in June 2019.

1.2 Research Partnerships

1.2.1 DREAM CDT

This research project is generously supported by the interdisciplinary Centre for Doctoral Training[6] (CDT) called *DREAM: Data, Risk, Environmental and Analytical Methods*. The CDT collaborates with many public and private entities and partners working with Big Data and Risk Mitigation. DREAM CDT supports over 30 doctoral students from Cranfield, Birmingham, Cambridge and Newcastle Universities, each with an industrial connection advising ways to apply academic research to existing challenges. DREAM CDT organised regular extracurricular training opportunities such as the DREAM Challenge Week, Conferences, Symposia, Expos, Big Data Training and generously supported field trips, international academic engagements and visits to our industrial partners such as my own visit to JRC.

[6] http://www.dream-cdt.ac.uk/

1.2.2 Joint Research Centre

Our industrial partner is the Joint Research Centre of the European Commission (EC) located in Ispra, Italy (one of several such EC research sites). Our personal advisor is Dr Jens P. Linge[7]. He is a Scientific Officer at the European Commission focusing on research in text information mining and analysis in the area of public health [75]. I had the opportunity to visit the JRC in February 2018, which is a former nuclear research site established in 1960, to present completed and ongoing geoparsing research. The visit also aimed at studying the challenges of maintaining a large-scale news monitoring system, a process that yielded opportunities for further research. Many insights in the later chapters are a direct result of the collaboration, e.g. the data stream analysis and the FlexiGraph proposal in Chapter 6, the New Evaluation Framework in Chapter 5 and the conclusions made in light of the real-time constraints of MediSys. We were delighted with the collaboration as it enabled us to evaluate many aspects of our research on actual system-generated data.

1.2.3 EMM and MediSys

The European Media Monitor [178] is a family of applications for detection, monitoring and analysis of multilingual, real-time news events. Several parts of the system are available for public use. The EMM polls the RSS feeds of thousands of news sources throughout the day, preprocessing, clustering and feeding information forward to downstream applications such as Newsbrief[8], NewsExplorer[9], MediSys[10] and others. MediSys [167, 114], operating since 2005 and used by the European Centre for Disease Prevention and Control (ECDC) and EC's Directorate-General for Health and Food Safety[11] among others, is a public application of the EMM focusing on aberration detection in news events relevant to public health risks such as food fraud, human and animal disease outbreaks, food contamination, radioactive events, bio-terrorism, etc. We have been granted access to the news feed with the aim of researching and proposing new ways of improving the efficacy of the extraction and grounding of toponyms. In 2011, an evaluation report conducted by the ECDC [122] concluded that MediSys was faster at issuing alerts than a human-moderated system. They also made a number of recommendations in terms of the combinations of keywords used by the system to capture epidemiological events. In this thesis, we solely focus on technical recommendations and proposals required for effective information extraction, disambiguation and ranking.

[7] https://scholar.google.co.uk/citations?user=T7eNcLUAAAAJ&hl=en

[8] http://emm.newsbrief.eu/

[9] http://emm.newsexplorer.eu/

[10] http://medisys.newsbrief.eu/

[11] https://ec.europa.eu/info/departments/health-and-food-safety_en

1.3 Related Work in Biosurveillance

A significant amount of prior work in the biosurveillance domain has been published, dating back a few decades due to the ever-present threats to public health, taking advantage of ever-improving and affordable information technology. A Google Scholar search for "automatic disease outbreak monitoring" returns almost 35,000 publications thus we mostly focus on the subset of recent systems that use Natural Language Processing methods (including their data sources). We note that many operational and decomissioned systems do not publish sufficient technical documentation enabling a full (or adequate) replication, which is reflected in this section's narrative. Multiple surveillance systems are in operation, implemented and maintained across several continents, frequently collaborating and exchanging knowledge. Most of the relevant manuscripts get published in medical journals, specialist conferences and governmental research portals rather than at computational linguistics venues. In Europe, there is an entire journal dedicated to surveillance, mitigation, prevention and control, founded in 1995 called Eurosurveillance[12] and funded by the European Centre for Disease Prevention and Control (ECDC). There are a number of additional discontinued services such as Google Flutrends[13], Biocaster [34], EpiSPIDER [79] and others that will not be covered.

Dr Jens Linge, who is our industrial partner and external advisor, described in his 2009 article [115] the three main categories of surveillance systems : 1) *News Aggregators*, which are single portals for easy access but requiring human filtering and analysis. 2) *Automatic Systems* such as MediSys, using NLP methods to categorise, geolocate and filter incoming news, with minimal human supervision. HealthMap[14] [60] is one such system, it is mostly rule-based with curated disease/location lexicons and a clearly defined workflow of processing steps using ProMED-mail and RSS news feeds as data sources. 3) *Moderated Systems*, which are supervised by experts hence offer higher quality output but at a lower volume and with delays due to the verification process. Examples include the Global Public Health Intelligence Network (GPHIN) published in 2008 [18] with an update in 2015 [46]. The project is managed by the Canadian Centre for Emergency Preparedness and Response. Another human-supervised reporting system often referenced as a reliable data source is ProMED-mail[15], founded in 1994 and a well known database and notification platform where reports are carefully curated by expert operators and disseminated via email or web.

[12] https://www.eurosurveillance.org/

[13] https://www.google.org/flutrends/about/

[14] https://www.healthmap.org/en/

[15] https://www.promedmail.org/

Twitter is another popular data source for researchers in Epidemic Intelligence (EI) due to its real-time interactivity, abundant data and a large number of users. This resulted in prolific research reviewed in "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review" featuring 158 papers [129]. A further 249 manuscripts were excluded from the review due to the usage of 'non-Twitter' datasets and a small number of duplicates, showing high activity in this field. Twitter EI research was also the subject of a 2016 systematic review of the use of 'Big Data' for disaster management, which included 76 papers from the SCOPUS database. Papers that use elements of geoparsing for specific applications include mapping events such as flooding in India [6], the spread of Zika in South America [192], studying health and disease in the 19th century [157], parsing tweets in order to map power outages across the US [123] and automatically generating geolocated news reports from Twitter [116] and Reuters News [154]. However, only a few make meaningful *technical advances* towards more accurate location extraction and grounding. Moreover, releasing the evaluation data/code for further work does not seem to be viewed as a matter of course, something we repeatedly encountered in the last 4 years. As a consequence, many one-off experimental laboratory studies on specific events do not provide a working system or a prototype for follow-up work. One way this could be remedied is by following the methods in Chapter 5 of this thesis, i.e. by establishing an open evaluation framework with standard datasets, metrics and other important considerations essential for a fair comparison. Examples of such projects include the online ranking platform EUPEG [86] (under construction) for geoparsing or SQUAD[16] for Natural Language Inference to keep track of the SOTA in this active field of research[17]. Resources are available, however, for a survey that evaluates several distinct geoparsing methods (Statistical language models, Google Geocoder, OpenStreetMap, DBPedia and Geonames entity matching) for *social media text* (Twitter, Flickr) [135] across English and Turkish. The OpenStreetMap entity matching performed best across datasets, which included tweets from earthquakes and blackouts. Note that as Twitter users delete data/accounts over time, the number of tweets retrieved for dataset reconstruction decreases, making robust evaluation difficult. For this reason, the narrow scope and the aforementioned scarcity of resources, in this thesis, we focus more broadly on geoparsing news articles and other generally well-formatted mainstream text documents.

[16] https://rajpurkar.github.io/SQuAD-explorer/

[17] Update: GeoTxt [Karimzadeh et al.] (Chapter 2) released an updated version http://geotxt.org/.

1.3.1 Summary

The thesis presents geoparsing research at multiple levels of abstraction including exploratory work on event geolocation applied to *Disease Monitoring*. We can obtain large quantities of geographic data by automatically converting place names in free-format texts into geographic coordinates. Using these annotated documents, we can monitor the geographic spread of relevant media events over time and at scale. As we shall outline throughout the thesis, current geoparsing technologies (Named Entity Recognition and Toponym Resolution) have relatively high error rates, something we effectively address with technical and linguistic advances. This applies to off-the-shelf and specialised information extraction tools, as evaluated in multiple chapters. The error rate in NER means that a substantial percentage of toponyms become false positives (erroneously tagged as locations) and false negatives (genuine locations mistakenly ignored, especially the less frequent and smaller toponyms), which result in poorly understood geographic ambiguity leading to downstream problems in early detection and alerting. Toponym resolution, the second stage of geoparsing, currently adds further errors as each toponym must be disambiguated to its coordinates. The severity of omitting or misclassifying an entity as a toponym is as adversely impactful as assigning the wrong coordinates to *London*, for example. This thesis presents research that advances both geoparsing stages thus substantially reducing errors leading to more accurate geographic annotation and event geolocation. To this end, we shall introduce the solutions starting with a comprehensive empirical survey of available systems in Chapter 2. In Chapter 3, we investigate whether a toponym refers to a literal place or an entity associated with a place, laying the initial groundwork for more effective toponym extraction by incorporating Metonymy Resolution into NER. This is followed by introducing a SOTA method for Toponym Resolution in Chapter 4. We will revisit metonymy resolution in Chapter 5 but with significant extensions taking a holistic view of the geoparsing task to propose solutions for several issues we discovered throughout the publication process, all the while adding high quality tools and datasets to foster open research. In Chapter 6, we apply these advances to the concrete task of Monitoring Threats to Public Health by building and evaluating models on geoparsed data from an automatic event geolocation and monitoring system called MediSys (maintained by the JRC of the European Commission) and suggesting practical ways of deploying our research insights. Finally, we outline the most pertinent challenges yet to be solved, including specific proposals for Future Work.

Chapter 2

Geoparsing Landscape

2.1 Introduction

When entering a previously unfamiliar research topic, the first port of call is to establish the state of the art, prominent research methods, available resources and relevant articles. A survey paper should be a convenient entry point for new researchers and provide this overview. Naturally, we looked towards the existing literature but were unable to locate any up to date comprehensive surveys. The PhD theses in related research [148, 105, 7, 26, 38] did provide useful insights but were not dedicated to geoparsing. In the following years, a Twitter geolocation survey [207] (2018), a geocoding survey (2017) [133] and some shorter evaluation papers would be published. As we point out in this chapter, prior works have created and evaluated models on useful datasets, however, our ongoing efforts to locate those resources proved unsuccessful. This was yet another reason to contribute an open dataset. Our intention to replicate/trial existing models in the forthcoming survey was expected to uncover avenues of promising research, which is exactly what happened. In order to communicate feasible improvements to MediSys and geoparsing, we employed a pragmatic approach in the following publication: *What's missing in Geoparsing?*[1], which outlines the investigative processes and the resulting discoveries. The paper was published in the Springer journal[2] *Language Resources and Evaluation* in March 2017 and the accompanying resources are available on GitHub[3].

[1] https://link.springer.com/article/10.1007/s10579-017-9385-8

[2] https://link.springer.com/journal/10579

[3] https://github.com/milangritta/WhatsMissingInGeoparsing

2.2 What's Missing in Geoparsing?

Geographical data can be obtained by converting place names from free-format text into geographical coordinates. The ability to geo-locate events in textual reports represents a valuable source of information in many real-world applications such as emergency responses, real-time social media geographical event analysis, understanding location instructions in auto-response systems and more. However, geoparsing is still widely regarded as a challenge because of domain language diversity, place name ambiguity, metonymic language and limited leveraging of context as we show in our analysis. Results to date, whilst promising, are on laboratory data and unlike in wider NLP are often not cross-compared. In this study, we evaluate and analyse the performance of a number of leading geoparsers on a number of corpora and highlight the challenges in detail. We also publish an automatically geotagged Wikipedia corpus to alleviate the dearth of (open source) corpora in this domain.

2.2.1 Introduction to the Survey

With the exponential increase in availability of public, free-format text generated through the accelerating use and adoption of internet-connected devices and the subsequent increase in social media usage, there is a greater opportunity for researchers and developers to utilise geographical information. Geographical data adds an additional dimension to the richness of data enabling us to know not just the what and when, but also the *where* of an event. In geoparsing, the place names containing the geographical information are called *toponyms*, which must first be identified (called geotagging) and resolved to their geographical coordinates (called geocoding), see Fig 2.1. This two stage approach is common in the domain and this is also how we evaluate the geoparsers. In an example sentence, “A publicity stunt in *Boston Common* causes surprise in *Washington D.C.*", the toponyms are “Boston Common" and “Washington D.C." since the action happens in both places.

Geotagging is a special case of Named Entity Recognition (NER), which is very much an open problem in NLP. The difference is, it only retrieves *locations* (no persons, organisations, etc.). Also, an important component of geotagging is Metonymy Resolution, which is discussed in Section 5.3.3. Geocoding, from an NLP point of view, is Named Entity Disambiguation (NED) followed by Named Entity Linking (NEL). Given a list of candidate coordinates for each location and the surrounding context, the goal is to select the correct coordinate i.e. disambiguate. Finally, each toponym is then linked to a record in a geographical knowledge base such as Geonames. This chapter provides the following two main contributions:

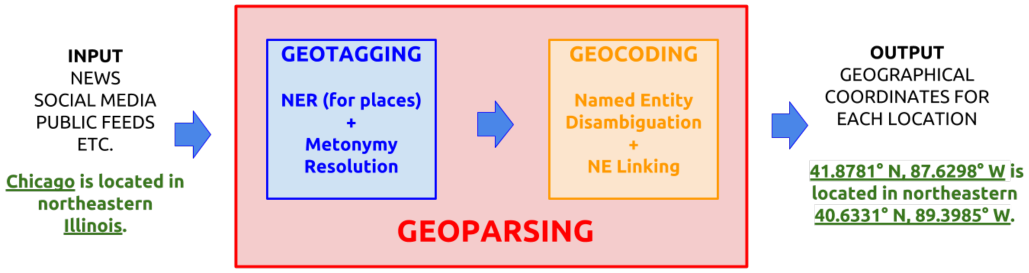

INPUT
NEWS
SOCIAL MEDIA
PUBLIC FEEDS
ETC.
Chicago is located in northeastern Illinois.
GEOTAGGING
NER (for places)
+
Metonymy Resolution
GEOCODING
Named Entity Disambiguation
+
NE Linking
GEOPARSING
OUTPUT
GEOGRAPHICAL COORDINATES FOR EACH LOCATION
41.8781° N, 87.6298° W is located in northeastern 40.6331° N, 89.3985° W.

Fig. 2.1 The geoparsing pipeline comprises two main stages geotagging and geocoding. Geotagging retrieves only literal toponyms (ideally filtering metonymic occurrences) from text and generates candidate coordinates for each. Geocoding then leverages the surrounding context to choose the correct coordinate and link the mention to an entry in a geographical knowledge base.

1. We provide a comprehensive survey and critical evaluation of state-of-the-art geoparsers with heterogeneous datasets. To the best of our knowledge, a detailed geoparsing survey such as this is currently unavailable.

2. We also present WikToR, a novel, large-scale, automatically generated and geographically annotated Wikipedia corpus for the purpose of alleviating the shortage of open source corpora in geoparsing research.

The rest of this chapter is organized as follows. In section 2.2.2, we provide the background to the geoparsing task, definitions and related research. In section 2.2.3, we present the featured systems including related systems that did not qualify for our evaluation. In section 2.2.4, we present our corpus (WikToR), its benefits and limitations, processing steps, quality checking and a sample article. In section 2.2.5, we describe the metrics used, which is followed by result tables and their detailed interpretation in section 2.2.6. We then discuss replicability and our experience in section 2.2.7. Example errors committed by the surveyed systems are analysed in section 2.2.7, which is followed by the conclusion.

2.2.2 Related Work

As mentioned in the Introduction, geographical parsing is the task of identifying and resolving toponyms to their geographical coordinates. The exact definition of a toponym varies among researchers, lexicographers and developers. The most common (and ambiguous) dictionary definition is a *place name*. There are other suitable definitions for a toponym such as Wikipedia's instructions for WikiProject contributors, which (as of June, 2016) contained the following:[4] "In general, coordinates should be added to any article about a location, structure, or geographic feature that is more or less fixed in one place." The United Nations Conference on the Standardization of Geographical Names[5] define a toponym as *the general name for any place or geographical entity*. This is the best definition we found and we further add that a toponym is also *a name for a topographical feature*.

That means lakes, cities, monuments, colleges, and anything static (natural and artificial physical features of an area) are toponyms and require coordinates. In this chapter, we only evaluate inhabited places (countries, cities, villages, regions, neighbourhoods, etc.), which is a subset of toponyms. This is the standard approach [41], [73] in most geoparsing. However, other work [62] evaluated additional toponyms such as landmarks, buildings, street names, points of interest and outdoor areas (parks, etc). The datasets used in our evaluation contain many types of toponyms, although the most common types by far are inhabited places.

Additional Geotagging Challenges

Metonymy Toponyms should exclude metonymic readings i.e. comprise only locative interpretations (geographical territory, the land, the soil, the physical location). In a sentence "*London* voted to stay in the EU.", London is a metonymic reading, substituting for the people of London, hence not a toponym. Another example is "*France* travelled to Spain for their friendly.", which means the French team are heading out. A toponym has no capacity to act like a living entity (humans, organisations, etc.). To that end, any place name substituted in for the actual "actor" should not be considered a location.

In our work, however, we neither have the datasets to evaluate geoparsing performance for metonymic readings nor the geoparsers capable of this distinction. We further discuss the errors caused by the lack of understanding of metonymy in geoparsing in section 5.3.3. There is no doubt that incorporating metonymy resolution into future geoparsing technology will increase performance (higher precision - fewer false positives).

Implementation Choices Another important factor that influences geotagging performance is the choice of software implementation with regards to NER. Most of the geoparsers tested do not implement their own NER, instead opting for open source software such as

[4] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates

[5] http://unstats.un.org/unsd/geoinfo/UNGEGN/

Apache NER[6] or Stanford NER[7]. [139] tested the influence of the changing text time frame on the NER performance and found that over a period of 8 years, the performance of the NER tagger steadily declines, which was also the case in [113]. Section 2.2.7 elaborates on the shortcomings of NER further.

Additional Geocoding Challenges

In addition to the more manageable challenges such as misspellings and case sensitivity (analysed later), processing fictional and/or historical text presents an additional challenge for the geocoder (i.e. translating a location name such as "ancient city of *Troy*" into coordinates). Modern geographical place databases may not contain entries for historical places, many of which changed their names over time. The more prominent changes such as Tsaritsyn - (1925) - Stalingrad - (1961) - Volgograd or Vindobona - Vienna are covered however more obscure changes such as Aegyssus - (1506) - Tulcea (Romania) are not included. This problem can be solved by using Pleiades[8][175], a community-built gazetteer of ancient places. At the time of this publication, it contained 35,000 ancient places. Named Entity Linking algorithms such as REDEN [21] can help with selecting the right candidate from a list of potential entities.

Errors can also occur with places which have historically changed names, such as Belfast, Australia, which became Port Fairy or when Limestone Station, Australia became Ipswich. The harder problem, however, is geocoding fictional places (islands, towns, states, etc.) such as Calisota (California + Minnesota) from Walt Disney Comic Books. The same applies to virtual places like Zion from the film Matrix or other fictional locations. These occurrences, however, should be infrequent in modern popular discourse.

2.2.3 Featured Systems

Selecting Systems

The comparison we conducted here was not restricted to only academic (accompanied by a scientific paper) parsers. In order to produce the broadest survey possible, we included commercial, open-source and academic solutions and used the following selection criteria:

1. The paper (academic geoparsers only) describing the methodology was published in 2010 or later.

[6] https://opennlp.apache.org/
[7] http://nlp.stanford.edu/ner/
[8] https://pleiades.stoa.org/

2. The geoparser is either publicly available (through APIs) or downloadable (as source code/binaries).

3. The geoparser performance (reported in a paper or trialled through a demo version) was near the state-of-the-art.

There are several related NER/NED tools such as OpeNER[9], Thomson Reuters Open Calais[10], however, these do not disambiguate down to the coordinates level, hence do not qualify as geoparsers. AIDA[11] [201] is another excellent NED tool for named entity identification and disambiguation, however similarly to OpeNER, the output is a link to a Wikipedia/DBpedia page, which may or may not contain coordinates. Despite their usefulness, they do not qualify as full geoparsers, hence do not feature in this survey. Finally, we were not able to obtain a working version of the Geolocator by [62], [204].

The Featured Geoparsers

CLAVIN[12](Cartographic Location And Vicinity INdexer) is an open-source geoparser that employs context-based geographic entity resolution. It was downloaded from GitHub on 5/12/15. CLAVIN employs fuzzy search to handle incorrectly-spelled location names. It also recognises alternative names and utilises other open-source technologies such as Apache OpenNLP[13] as well as intelligent heuristics-based combinatorial optimisation to identify precisely which toponym sense was referred to in the text. The data for CLAVIN comes from the GeoNames Knowledge Base[14]. As far as we can tell, there is no academic publication available for this open source project. Additional information can be found on the website[15].

Yahoo!PlaceSpotter[16] is a commercial geoparser that identifies places in unstructured text such as web pages, RSS feeds, news or just plain text files, all with multilingual support. Yahoo!'s proprietary Geo-informatics database information consists of six million (and growing) named places including administrative areas, settlements, postal codes, points of interest, colloquial regions, islands, etc. PlaceSpotter also delivers bounding boxes and centroids of named places. The API was used from January through March 2016 to process the featured corpora. PlaceSpotter identifies locations using WOEID (Where on Earth ID), which always

[9] http://demo2-opener.rhcloud.com/welcome.action
[10] http://www.opencalais.com/opencalais-demo/
[11] https://gate.d5.mpi-inf.mpg.de/webaida/
[12] https://clavin.bericotechnologies.com/about-clavin/
[13] https://opennlp.apache.org/
[14] http://www.geonames.org/
[15] https://clavin.bericotechnologies.com/clavin-core/index.html
[16] https://developer.yahoo.com/boss/geo/docs/key-concepts.html

reference place concepts. For example, “New York", “New York City", “NYC", and “the Big Apple" are all variant names for WOEID 2459115. The geoparser also has an understanding of geographical focus of the document i.e. where is this document talking about, which may be useful for many applications.

The Edinburgh Parser [73], [186] is an end-to-end system, developed in-house and partitioned into two subsystems, a geotagger and a geocoder. The geotagger is a multistep rule-based NER which makes use of lists of place names and person names for personal and location entity recognition. The gazetteers, Unlock[17] (decommissioned in 2016) and GeoNames, which can be chosen at run-time, provide the location candidates together with metadata such as population, coordinates, type, country, etc.

The rule-based geocoder then uses a number of heuristics such as population count, clustering (spatial minimization), type and country and some contextual information (containment, proximity, locality, clustering) to score and rank the candidates and finally choose the correct one. If there is no entry in the gazetteer, the place remains unresolved. As with every other tested geoparser, the evaluation was performed end-to-end (pipeline) to simulate real world usage. The binaries for the geoparser were downloaded[18] on 21/12/2015.

Topocluster [41] models the geographic distribution of words over the earth's surface. The intuition is that many words are strong indicators of a location thus good predictors given contextual cues. The geotagging is done using standard Stanford NER. The geocoding is performed by overlaying the geographic clusters for all words in the context (up to 15 each side) and selecting the strongest overlapping point in the distribution. Topocluster uses a geographically tagged subset of Wikipedia to learn this spatial distribution of word clusters i.e. it is trained to associate words in the vocabulary with particular coordinates. In addition to the single point in the world, a gazetteer entry closest to the predicted coordinates is chosen as the final location candidate.

Although Topocluster can work without a knowledge base, resolving toponyms to a single pair of coordinates, for best performance, it does require a gazetteer (mainly GeoNames plus a small Natural Data[19] dataset of just 500 referents). Domain optimisation to each dataset is required for best results (in our evaluation, we do not adapt to new domains for a fair comparison with other systems, which also use default setups). The experiments outlined

[17] http://edina.ac.uk/unlock/

[18] https://www.ltg.ed.ac.uk/software/geoparser/

[19] http://www.naturalearthdata.com/

in their publication have shown that the domain parameters are volatile and corpus-specific, which may adversely affect performance in the absence of corpus adaptation. The final version of the software was downloaded from GitHub on the 18/12/2015.

GeoTxt[20] [95] is a web-based geoparser specialising in extraction, disambiguation, and geolocation of toponyms in unstructured micro-text, such as Twitter messages (max 3,900 characters, approx 750 words). Identifying and geographically resolving toponyms in micro-text is more challenging than a longer article such as a news report due to fewer contextual features. Much like the previous systems, it works in two steps, geotagging (extracting toponyms) using three NER taggers, Stanford NER[21], ANNIE[22], Illinois NER[23] and geocoding.

Geocoding (disambiguation) is performed using the GeoNames Web Service to generate a list of candidates. To rank and score them, they use: geographic level, e.g. country, province or city of the place name in text (when provided), Levenshtein Distance between the candidate name and the one mentioned in text and population with higher priority given to places with higher population. GeoTxt also uses spatial logic to leverage other place name mentions in context such as “I love *London*, I love Canada!". This will resolve to London, Ontario, Canada. The parser was evaluated from February - March 2016.

2.2.4 Evaluation Corpora

There exists a challenge that all researchers in this field currently face and it is the lack of freely available geotagged datasets. We are aware of a few datasets such as the free and open War Of The Rebellion[24] by [42], which was released at the ACL 2016 conference. A geotagged Twitter corpus was announced by [189], which should help alleviate the problem. Another geotagged dataset called TR-CONLL by [104] exists although the fees, terms and conditions of licensing are unclear. The corpus relies on Reuters data (a subset of the CONLL Shared Task 2003 dataset), which is not publicly available thus not conducive to open research. Finally, the ACE 2005 English SpatialML is a broadcast news and broadcast conversation geotagged corpus, also suitable for geoparsing evaluation. However, it is only available for a fee[25], which once again does not support open research and replication.

[20] http://www.geotxt.org/api/

[21] http://nlp.stanford.edu/software/CRF-NER.shtml

[22] https://gate.ac.uk/sale/tao/splitch6.html#chap:annie

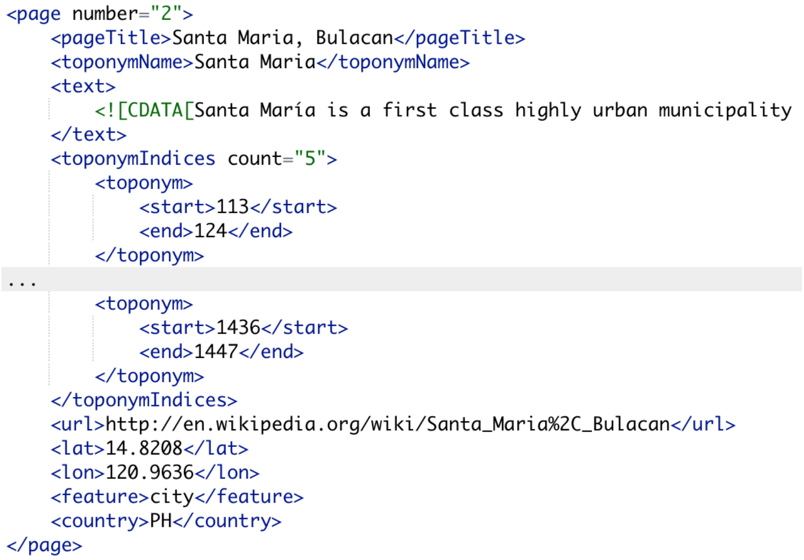
[23] http://cogcomp.cs.illinois.edu/page/demo_view/ner

[24] https://github.com/utcompling/WarOfTheRebellion

[25] https://catalog.ldc.upenn.edu/LDC2011T02

Our Corpus

As part of our contribution, we introduce a large, programmatically created corpus called WikToR (Wikipedia Toponym Retrieval) to allow for a comprehensive and reliable evaluation on a large corpus. The Python programming script[26] uses three sources of data to construct WikToR, the GeoNames[27] database dump, the GeoNames Wikipedia API[28] and the Wikipedia API[29] using a Python wrapper[30](accessed: March 2016).



```
<page number="2">
    <pageTitle>Santa Maria, Bulacan</pageTitle>
    <toponymName>Santa Maria</toponymName>
    <text>
        <![CDATA[Santa María is a first class highly urban municipality
    </text>
    <toponymIndices count="5">
        <toponym>
            <start>113</start>
            <end>124</end>
        </toponym>
...
        <toponym>
            <start>1436</start>
            <end>1447</end>
        </toponym>
    </toponymIndices>
    <url>http://en.wikipedia.org/wiki/Santa_Maria%2C_Bulacan</url>
    <lat>14.8208</lat>
    <lon>120.9636</lon>
    <feature>city</feature>
    <country>PH</country>
</page>
```

Fig. 2.2 A sample article from WikToR. The article has been shortened.

Processing Steps Starting with approximately 10M entries in the GeoNames database, the following steps were used to create the final corpus:

1. Locations with the same name had to be separated by a large (1000km) distance to be distinct and to remove duplicates.

2. The most ambiguous locations (ones with the highest count in step 1) were processed first. For instance, the five most ambiguous GeoNames locations are Santa Maria (26 entries), Santa Cruz (25), Victoria (23), Lima (19) and Santa Barbara (19), although Wikipedia does not have a page for each.

[26] https://github.com/milangritta/WhatsMissingInGeoparsing
[27] http://www.geonames.org/export/
[28] http://www.geonames.org/wikipedia/
[29] https://www.mediawiki.org/wiki/API:Main_page
[30] https://pypi.python.org/pypi/wikipedia

3. For each location, the Wikipedia pages were downloaded using GeoNames API[31].

4. Each page had to be tagged with one of [adm1st, adm2nd, adm3rd, city, country, isle] feature codes[32] to be accepted.

5. The final article is a Wikipedia page with the first paragraph(s) that exceed 200 word tokens of context (see Fig 2.2).

Benefits and Limitations

The benefits: Because WikToR has been created programmatically, it delivers great consistency of annotation (see section 2.2.4 on Quality Checking). The corpus can be updated and extended easily as the code[33] used to generate it was published alongside the corpus. The corpus presents a hard challenge, particularly for the geocoder. The 5,000 unique Wiki pages contain some relatively obscure locations from around the world, at least 1000km apart so a mistake is costly. Locations are deliberately highly ambiguous in that they share their name with other namesake locations for instance Lima, **Peru**, Lima, **Ohio**, Lima, **Oklahoma**, Lima, **New York** and so on. Finally, all articles are resolvable (distinguishable) since they contain the lead section of each Wikipedia article[34] providing enough disambiguation to precisely and confidently resolve each annotated location.

The limitations: The “one sense per discourse" [61] will not hold if processing the whole corpus at once (5,000 articles include many namesake places). We recommend processing one article at a time. Secondly, for geotagging, the corpus can only be used to evaluate the accuracy (rather than F-Score) since not all locations were annotated. For example, a page about “Victoria" will only be annotated with the occurrences of “Victoria", hence no precision or recall. This does not affect the geocoding metrics.

Quality Checking The advantage of a programmatically created corpus is the consistency of annotation. The corpus and the Python code that were used to create it is available for download from GitHub. In terms of data integrity, we rely on GeoNames and Wikipedia for the correctness of the coordinates, titles, URLs, feature and country codes.

[31] http://www.geonames.org/export/wikipedia-webservice.html#wikipediaSearch
[32] www.geonames.org/wikipedia/wikipedia_features.html
[33] https://github.com/milangritta/WhatsMissingInGeoparsing
[34] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

All page text was retrieved using Wikipedia API for Python[35], which is an open source library for accessing Wikipedia. The indexed toponym occurrences in pages include only whole location words i.e. “Brazil" NOT “Brazilian", all of which were unit-tested for their correctness. This includes checking for the presence and correct format of the page title, article numbering, text length, coordinates, URL, toponym indices, country code and toponym type. In the light of the licensing and availability conditions, we chose to benchmark the 5 systems on 2 available corpora (WikToR and LGL). Each parser uses only one best setting to parse all corpora. The datasets we used in our experiments were:

1. Local Global Corpus (LGL) - The corpus contains a set of 588 news articles, collected from smaller, geographically-distributed newspapers. Each article's toponyms were manually tagged and resolved. None of the toponyms in the corpus are larger than a US state (18% are US state or equivalent). The most popular countries in the corpus were USA (3252 locations), Russia (166), UK (140), Canada (129), Georgia (119), Israel (84), Palestine (84), Egypt (72), Sudan (49), Lebanon (33). Around 16% of toponyms in LGL do not have coordinates so we decided not to use those in order to evaluate all systems end-to-end. For a more detailed description, see [112].

2. Wikipedia Toponym Retrieval (WikToR) - is an automatically created and geotagged corpus (full description in the next section). Corpus statistics: 5,000 geographically unique Wiki pages, 1,906 of those lexically unique. Pages are ambiguous such as Lima, *Peru*, Lima, *Ohio*, Lima, *Oklahoma*, Lima, *New York* (2.6 pages per location name on average, median = 1). The average location occurrences per Wiki page is 8.5 (median 7) and the average page word count is 374 (median 302). Finally, all toponyms in WikToR have coordinates.

2.2.5 Evaluation Metrics

Unsuitable Metrics

There is currently no clear agreement on which metrics are best, for instance: **precision@k** (for geotagging) asks: Is the correct result in the first **k** predicted answers? However nearby answers are treated as equally correct as distant answers (and when k=1, it's equivalent to standard precision); **accuracy@k km/miles** (geocoding) is a better metric however it does not distinguish between similar predictions, for instance, 5km away and 50km away are equally correct when k=50. This will hide differences in error distribution.

[35] https://pypi.python.org/pypi/wikipedia

The **mean error** (geocoding) is too sensitive to outliers plus normal distribution of error distances should not be presumed (as this tends to follow the power law, see Fig 2.3) This means that a handful of very large errors has the potential to make the average error go up noticeably thus unfairly punishing a geoparser. In this chapter, we only show mean error and accuracy@k for backward compatibility and comparison with earlier work.

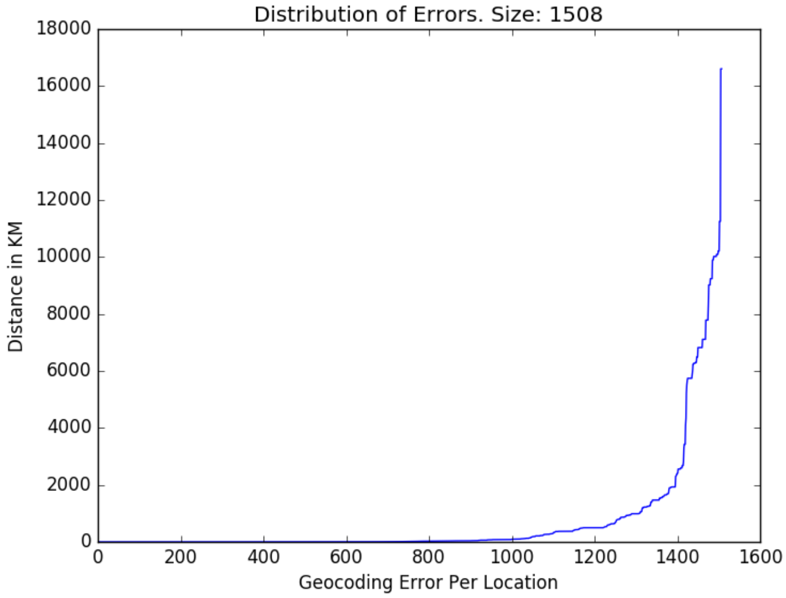


Fig. 2.3 A typical power law distribution of geocoding errors in KMs away from the gold location. Most errors are *relatively* low but increase very fast for the approximately last 20% of locations. This is a typical pattern observed across different systems and datasets.

Selected Metrics

We evaluated geoparsing as a pipeline (Input → Geoparser (as a black box) → Output), then scored each stage separately. The geotagging performance is measured using the **F-Score** for the LGL corpus, **Accuracy** for WikToR (see corpus limitations in section 2.2.4). The geocoding performance is measured using the **AUC** and **Median Error** for both corpora.

AUC (Area Under the Curve) - To the best of our knowledge, [92] first introduced the AUC for geocoding, which is a type of extension of the accuracy@k km/miles. AUC makes a fine-grained distinction between the individual errors rather than just a simple binary evaluation as with accuracy@k (either within k km/miles or not). This is accomplished by quantifying the errors as the area under the curve (see Fig 5.3). The smaller the area under the curve, the more accurate the geocoder.

AUC has a range of 0 to 1 since it assumes that an error cannot be greater than the furthest distance between two points on earth (20,038km). By taking the natural logarithm of the error distance, AUC ensures that the difference between two small errors (say 10km and 50km) is more significant than the same difference between two large errors (say 1000km and 1040km). AUC is superior to the popular accuracy@k because it computes all errors giving a true like-for-like comparison between geoparsers.

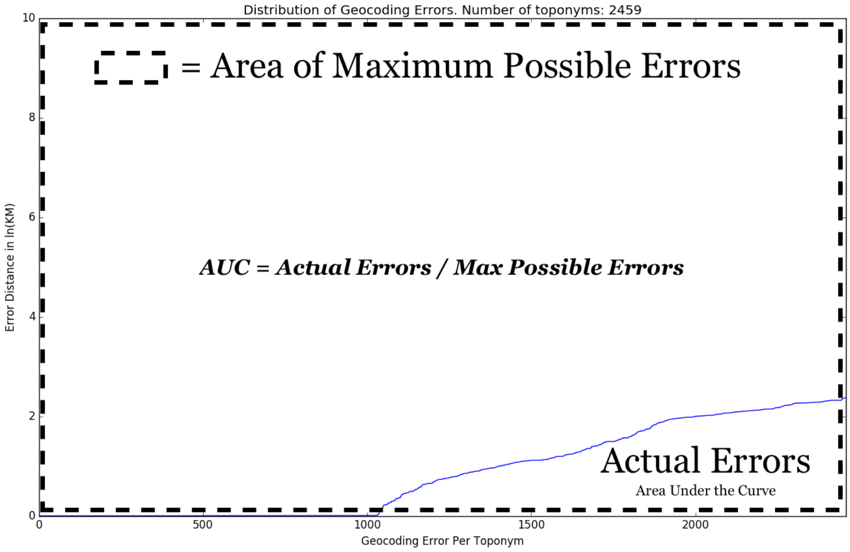


Fig. 2.4 How to calculate the AUC, a visual illustration. Lower scores are better. Fig 2.3 shows the same error data in its original form (before applying the natural logarithm).

Precision is the proportion of entities correctly identified as locations (true positives). In cases where non-location entities were identified as locations, these cases count as false positives. The formula for calculating precision is: true positives / (true positives + false positives) (range: 0 to 1).

Recall is the proportion of all locations identified (true positives). In cases where annotated locations were not identified as such, these count as false negatives. The formula for calculating recall is: true positives / (true positives + false negatives) (range: 0 to 1).

F-score is the harmonic mean of precision and recall and is calculated as (2 * Precision * Recall) / (Precision + Recall).

Accuracy - Accuracy measures the percentage of correct predictions out of a total of expected predictions (ranging from 0 to 1). In the case of WikToR, the accuracy measures the percentage of correctly identified locations by the system out of the total number of gold annotated locations.

Median (natural log) Error Distance - The error distance is a distance from the actual (gold) location (latitude, longitude) to the location (latitude, longitude) predicted by the geocoder. A median error is more informative than a mean error, which only has relevance if error distances are normally distributed (the error distributions follow the power law).

Accuracy@161km (geocoding only) - For backwards compatibility with previous work on geocoding, we also show accuracy@161km for the geoparsers. This metric shows the percentage of geocoding predictions, which were no more than 161km away from their true coordinates. Lower numbers are more desirable. The empirical reason for the choice of 161km as the cut-off distance is not clear from related work [30].

Running the geoparsers

During evaluation, each geoparser was treated as a black box, an end-to-end system receiving a single news article (LGL) or a single Wikipedia article (WikToR), then evaluating separately. NB: Most systems do not allow for geotagging, geocoding to be run individually, hence the “black box". Lastly, the Edinburgh Parser is the only system that allows toponyms to have NIL referents i.e. no coordinates assigned. This only happens in 2% to 4% of the cases. We decided to remove these from evaluation for a fairer comparison.

2.2.6 Results and Analysis

Scoring in Geotagging

Exact and Inexact - Because of subtle differences between test file annotation and real-world geoparsing, we used two types of scoring for the geotagging stage. Exact matching compares the output and annotation exactly. For example, if “Helmand" is annotated, but the output is “Helmand Province", this is both a false positive and a false negative. Same applies for “Birmingham" and “District of Birmingham". Inexact matching for these examples is more lenient and accepts both answers, resulting in a true positive i.e. no errors.

Geotagging Results

Table 2.1 shows the performance on LGL, and Table 2.2 on WikToR (accuracy only, see section 2.2.4). For geotagging, we report the exact score, i.e. only text span matches and inexact (in brackets), i.e. "Avoyelles Parish County" and "Avoyelles Parish" are both correct.

LGL	*Precision*	*Recall*	*F-Score*
GeoTxt	0.80	0.59	0.68 *(0.74)*
Edinburgh	0.71	0.55	0.62 *(0.67)*
Yahoo!	0.64	0.55	0.59 *(0.67)*
CLAVIN	**0.81**	0.44	0.57 *(0.59)*
Topocluster	**0.81**	**0.64**	**0.71** ***(**)***

Table 2.1 Geotagging performance on LGL. (**) - Not available for this system.

Even with inexact matching, no geoparser crosses the F-Score=0.8 threshold. Topocluster and GeoTxt, both of which use Stanford NER, performed almost identically on LGL (0.68 and 0.71 F-Score) although less well on WikToR (0.51 and 0.54 accuracy). The rest of the geoparsers use a combination of techniques for NER, open-source, proprietary and rule-based. The rule-based Edinburgh geoparser performed best on WikToR (3rd best on LGL) while the rest faltered, particularly CLAVIN (only 0.21 accuracy).

In the Introduction, we asserted that NER is very much an open NLP task, particularly for place names. The three geoparsers that integrate external NER software (GeoTxt, Topocluster, CLAVIN), leave a lot of room for improvement as shown in the tables. Wikipedia text proved to be the greater challenge, a finding mirrored by [12]. The ultimate goal for geotagging performance should be F-Scores of 0.9+ or alternatively Fleiss' kappa [59] of 0.81+ in multiple domains, which is approximately human level performance. Until such time, NER, which is a fundamental part of the geographical parsing task, remains an unsolved problem [128] in NLP.

From a parsing speed point of view, we estimated the fastest geoparser (CLAVIN - processing WikToR in $\sim$2 minutes) to be around 10,000 times faster than the slowest (Topocluster - processing WikToR in 2 weeks and 8 hours). However, the speed advantage did not translate into a geotagging advantage. The Edinburgh geoparser did best overall, its processing speed was more than adequate.

WikToR	*Accuracy*	*Accuracy (inexact)*
GeoTxt	0.51	*0.51*
Edinburgh	**0.65**	***0.66***
Yahoo!	0.4	*0.5*
CLAVIN	0.21	*0.22*
Topocluster	0.54	(**)

Table 2.2 Geotagging performance on WikToR. (**) - Not available due for this system.

Geocoding Results

Table 2.3 shows the geocoding results for LGL and Table 2.4 for WikToR. All geoparsers (except Yahoo! - proprietary) use the same knowledge base (GeoNames) for a fair comparison of results. The AUCE metric is the same as AUC, but this time, run on an identical set of toponyms (i.e recognised by all five geoparsers, 787 for the LGL dataset and 2202 for the WikToR dataset). In the equal comparison, the AUC scores improve significantly (most likely due to the selection bias), however, the order of the geoparsers is mostly unchanged.

LGL	*AUC*	*Med*	*Mean*	*AUCE*	A@161
GeoTxt	0.29	0.05	2.9	0.21	0.68
Edinburgh	**0.25**	1.10	**2.5**	0.22	**0.76**
Yahoo!	0.34	3.20	3.3	0.35	0.72
CLAVIN	0.26	**0.01**	**2.5**	**0.20**	0.71
Topocluster	0.38	3.20	3.8	0.36	0.63

Table 2.3 Geocoding results on LGL. Lowest scores are best (except A@161). All figures are exponential (except A@161), so differences between geoparsers grow rapidly.

Our intuition that WikToR is a hard geocoding challenge was confirmed in the poor geocoding scores (N.B. scores are exponential, base **e**). There was a role reversal, CLAVIN (last in geotagging) excelled in this sub-task while Topocluster (first in geotagging) came last. This illustrates an important point, namely the integrity and performance of the entire pipeline. It is not sufficient to excel only at speed (CLAVIN) or geotagging accuracy (Topocluster) or geocoding performance (Yahoo!). A great geoparser must do all three well and the only one to manage that is the Edinburgh Parser. We also emphasise that there is much room for improvement in this technology, see section 2.2.7.

WikToR	*AUC*	*Med*	*Mean*	*AUCE*	A@161
GeoTxt	0.7	7.9	6.9	0.71	0.18
Edinburgh	0.53	6.4	5.3	0.58	0.42
Yahoo!	**0.44**	**3.9**	**4.3**	**0.53**	**0.52**
CLAVIN	0.7	7.8	6.9	0.69	0.16
Topocluster	0.63	7.3	6.2	0.66	0.26

Table 2.4 Geocoding results for WikToR. Lower scores are better (except A@161). Figures are exponential (except A@161), so differences between geoparsers grow fast.

The geocoding task revealed a difference between the two corpora. For LGL, which is a news-based corpus, AUCE scores decreased. For WikToR, which is a Wikipedia-based corpus, AUCE scores increased. We saw easily detectable toponyms in LGL resulting in easier geocoding, however, easily detectable toponyms in WikToR resulted in harder geocoding. This demonstrates the ambiguity of WikToR, which we asserted in section 2.2.4. The corpus was built with location ambiguity as a goal so that prominent locations around the world have many namesakes, and are thus harder to geocode.

With LGL, we hypothesise that prominent locations tend to be the most frequent "senses" of the toponym i.e. "Lima" most likely refers to "Lima, Peru" rather than "Lima, New York". In contrast, WikToR has seven unique entries for "Lima", which is a prominent geographical entity and should be easy to detect. However, resolving "Lima" in WikToR is not a straightforward assignment of the most well-known location entity by that name as seems to be the case in LGL. This makes WikToR the suitable benchmark for the ultimate geocoding challenge on which all other geoparsers should be tested to evaluate how well they can deal with high ambiguity.

Lastly, with respect to fairness of comparison, we briefly address Yahoo! Placemaker's proprietary database, which is the only one not to use GeoNames. It's not possible to empirically measure the differences between private knowledge bases, however, we think this will not unduly affect the results for the following reasons. The number of records in the databases is comparable, around 6M for Yahoo! and 11M for GeoNames, which is easily within an order of magnitude. Secondly, we do not expect the coordinates for any particular location to differ markedly between the knowledge bases. Even a difference of up to two-digit kilometers (which is still adequate) will not significantly alter the final AUC scores. Finally, 82% of LGL and 95% of WikToR locations are no larger than a US state or equivalent, of which most are cities and counties. It is the implementation of the geocoding algorithm that is the greatest determiner of final geocoding performance.

How to improve evaluation further

In future work, evaluation of precision can be further improved by taking the magnitude and type of geocoding errors into consideration. The error should be adjusted based on the size of the area of the location (geographical granularity). For example, a 100km error for a small town (50 sq km) is hugely costly compared to a 100km error for a large country (650,000 sq km). The larger the entity, the more disagreement exists as to its exact geographical centre. This can be mitigated by evaluation becoming more forgiving as the area of the location entity increases. Alternatively, one can stratify the evaluation and present results at each level.

Geocoding errors are all currently treated as equal. Our suggested new kind of error contextualization befits the intuition that tolerance for a geocoding error should grow with the size of the location. An application can utilise the GeoNames (premium subscription[36]) database to retrieve the polygon coordinates for the calculation of the area of the location entity. Then apply a linear scaling factor such as a cube root to calculate an error tolerance. For example, the area of Greater London is $\sim$1500 sq km so $\sqrt[3]{1500} = 11.45$ meaning an error of 11.45km can be considered a “calibration" error for the size of Greater London. Similarly, for France ($\sim$644K sq km), the error tolerance would be $\sqrt[3]{644000} = 86.4$km.

2.2.7 Discussion and Future Work

Replicability

Verifiability and disproof are the cornerstones of the scientific method. In the wider scientific community, it is commonly known that many studies are not replicable [150], [102] and this is also the case with **irreproducible software** [180]. We would like to add value by sharing some key observations from our experience during the making of this publication. It is important that a detailed methodology and/or software be published with the scientific paper (including any data) [63].

This issue repeatedly arose in our experimentation. Some systems did not have a working version of the parser obtainable from the official source. Some of these took many weeks of time and/or a frequent email exchange to rectify in order to enable reproduction of the

[36] http://www.geonames.org/products/premium-data.html

reported results. In other cases, the waiting time for essential software was around a month only after a written request was made. Another software was prohibitively cumbersome to set up, which does not aid replication. On the positive, some parsers were readily runnable with little or no setting up and helpful cooperation from fellow researchers; however notice the substantial disparity.

In a related survey, geo-locating Twitter users, [92] have demonstrated that the real world performance can be much less accurate than the lab-tested system. The performance of the 9 surveyed systems declined over time (the projected performance, i.e. the number of Twitter posts able to be geo-located, would halve every four months). This is another reason for conducting a comprehensive comparative study such as ours, to find out whether reported results hold during independent replication.

Error Analysis

The current geoparsers and NER taggers struggle with three main kinds of NLP challenges giving rise to a need for a deep semantics geoparser to address them. The state-of-the-art in NER is between F=0.7 to F=0.9 [176], [162], depending on the type of corpus (CoNLL gets the highest scores while MUC is more challenging). Twitter NER is the most challenging due to its informal nature [109]. Solving the following problems is going to involve solving common NER challenges as recognising locations in plain text is a special form of NER.

What to do with Metonymy? Perhaps the greatest weakness of current geoparsing technology is the lack of understanding of metonymy on the parsers' side. That is, relying too much on syntactic pattern matching and the dominant sense of the entity in question at the expense of taking important cues from the surrounding words. Understanding metonymy means that the geoparser should be able to predict whether a word at the centre of some context is likely to be a location or not without knowing what the word is. Let's illustrate with a few examples or common context errors.

In a sentence "I think I may be suffering from *Stockholm* syndrome.", Stockholm will likely be tagged as a location because syntactically it matches a place name regardless of the clear semantic context that indicates otherwise. In "I worked with *Milan/Paris/Clinton* for 5 years." the same is true in spite of a human clearly labelling the entity as a personal name. This is a typical example of a lack of understanding of metonymy. "*London* voted to stay in the EU." London denotes the people in this case, not the place. There are more subtle examples such as "She studies *at* Edinburgh." (ORG) versus "She studies *in* Edinburgh."

(LOC) A human annotator would be able to discern Edinburgh as an institution and location given the context. A geoparser has to demonstrate the ability to understand the semantics of the context to avoid making and propagating mistakes to the geocoding stage.

Entity Boundary Errors An example of this is “We had an amazing time at *Six Flags Over Texas* (name of an amusement park)." where only Texas is incorrectly matched as a one-word entity by all five featured geoparsers. The penalty is a huge geocoding imprecision of this toponym. Similarly, failing to recognise “*London* Road" as a whole entity, the parser commits an error if only “London" is tagged. “London Road" is almost certainly not in London so geoparsers need to more accurately recognise the boundary of the entity. However, providing that capitalisation is correct, this is an infrequent occurrence.

Case sensitivity This is particularly impactful when dealing with unstructured domains like social media. When evaluated on LGL, Edinburgh and CLAVIN (using Apache OpenNLP) didn’t resolve any entities when capitalisation was removed, Topocluster (using Stanford NER) suffered a 40% dip in F-Score. For Yahoo! Placespotter, the F-Score decreased by 0.06 (10%), other metrics were unchanged. The only outstanding geoparser is GeoTxt (using three NER taggers), which performed robustly without capitalisation. Too many geoparsers and NER taggers are currently relying on correct capitalisation to capture the full and correct entity. For informal domains such as some social media, this issue will become problematic.

Further Examples of Common Errors

Sentence: “... water main blew on the corner of *Lakeview* and *Harrison* streets, ... " Errors: Too greedy i.e finding and resolving “Lakeview" and “Harrison" as incorrect places (semantics and word boundary error) Correct: Either tagging nothing or for advanced parsers, find and resolve “Lakeview Street" and “Harrison Street" in the wider context of “Pineville, LA".

Sentence: “Twin sisters Lindsay and *Charlotte* Graham, students ..." Errors: Despite clear context, “Charlotte" is tagged as a city by all but one system (semantics error) Correct: No entities found here. Geoparser greed is a common theme costing precision, but may serve to increase recall.

Sentence: “A *Rolling Meadows* resident used his car ..." Errors: All but one geoparser missed this location despite obvious context (semantics error). Correct: Tag and resolve “Rolling Meadows", seems obvious but this is another theme of missing entities despite clear context (such as “*Mansfield* council to vote today ..." - all but one geoparser missed this mention, but not other mentions of “Mansfield" in the same paragraph).

Sentence: “The *Athens* Police Department is asking ... " Errors: Geocoding imprecision, confusing Texas with Greece despite clues in context (Henderson County) Correct: Athens, Texas - the clue is in the next sentence. The same mistake is made with *Berlin, Connecticut* in the context of “*Hartford*, ... *Middletown*" and “Meeting in *Alexandria* will discuss ... " in the context of “Minnesota". This particular error theme shows that the geoparsers don't yet leverage the information generously provided in the context for a more careful and deliberate assignment of coordinates.

This use case is an illustration of incorrect labelling by the NER parsers. Several geoparsers in this chapter use Stanford, Apache NLP, Illinois and/or Gate NER to perform geotagging. Example sentences: “*Bala* is a market town and community in ..." or “*Hamilton* is the seat and most populous city ..." or “*Santa Cruz* is the county seat and largest city ... " Errors: Incorrect label (Person) assigned by most parsers. This is despite the ample and unambiguous context. Correct: Observe the strong cues in context and assign location label.

2.2.8 Closing Remarks

In our detailed analysis, we examined and presented the state-of-the-art systems in the geographical parsing domain. The comparative replication results showed that although useful in providing an additional dimension to the data, the systems still have several limitations. To that end, the current level of output integrity of the featured systems mainly allows for this technology to be used as supplementary input with acknowledgment of its limits rather than a high fidelity, geographical data output stream suitable for further processing “as is". This conclusion does not mean to reflect negatively on the implementation of the software, rather it serves to show the complexity of the task of coordinating the identification of location entities and their accurate disambiguation using the contextual cues in establishing the coordinates.

The new generation of geoparsers needs to employ the understanding of the meaning of context beyond syntactical and word form. It then has to use the cues extracted to intelligently and consistently query the integrated geographical database. An example of that is correctly processing a phrase such as “They are *40 miles west of the Statue of Liberty* ...". To detect and resolve these toponyms requires a more sophisticated AI parsing capability and an additional database of place coordinates. This is the next frontier for the next generation of geoparsers. The challenges of geoparsing make this task a worthwhile problem to solve as the benefits of mass scale text document (geographical) annotation can range from generating high quality social sciences and humanities research to accurately monitoring the world's media for geographical events to facilitating the communication between robots and people. The possibilities for this technology are limitless.

2.3 Summary

The difference between a literature survey and an empirical survey is that the latter has a significant hands-on component allowing for detailed, low-level exploration of the research challenges at hand. Indeed, this publication produced the research plan for the majority of this thesis and is the most cited paper so far. The replication, benchmarking and error analysis of existing systems helped the discovery of problem areas and aided the formulation of hypotheses as well as informed of any implementation issues.

An important lesson in research conduct was learnt after the unsuccessful replication and/or usage of some previous works. The lack of availability of working systems and the quoted resources was an observation that put a high emphasis on open-source standards in all of our subsequent work (later praised in academic email exchanges). The creation of a new resource in a form of a difficult to parse dataset, which would be used for testing the available geoparsers allowed for a direct comparison with our later research. The evaluation on LGL added a persistent thread to our subsequent research and helped ground new evaluation scores in a wider geoparsing perspective. It enabled an evidence-based and quantified claim of methodological progress of this thesis. We are delighted that our survey paper was used in follow-up research as the basis of EUPEG [86], *Extensible and Unified Platform for Evaluating Geoparsers*, which is going to be a web-based framework for fast and convenient system benchmarking where users can upload new datasets and publicly test their geoparsers.

The population heuristic has proven to be a simple but strong baseline, particularly for large toponyms, which would be reaffirmed by later research. The first hurdle a geoparser must clear is to beat this heuristic, which is not always the case. The reasons for this have as much to do with the absolute performance of the geoparsers as with the ambiguity of toponyms. With the benefit of today's knowledge, the geotagging scores observed in this work would be considered low, both in terms of precision and recall. Chapters 3, 4, 5 and 6 outline the research undertaken to improve these figures. Lastly, case sensitivity is only relevant for domains such as social media. We focused mostly on news reports, which are well structured allowing for the use of NLP tools such as parsers.

Chapter 3

Handling Metonymy in Geoparsing

3.1 Introduction

One of the main themes of the error analysis of the previous chapter and a promising route to original research was the problem of metonymy in geoparsing. In hindsight, toponyms were not only metonymic but also occurred as various non-literal types; this is extended in our later work (Chapter 5). As the thesis aims to communicate to relevant researchers proven ways to improve the accuracy of geolocation of news events, one early approach we targeted was be to differentiate metonymic and literal toponyms. For example: "When I travel around *New Zealand*, I feel like *Australia* is watching me." contains both toponym types. Metonymy refers to entities *associated with a place* that were *substituted with a location entity*. We hypothesised that making this distinction would help identify the geographic focus of a news event by eliminating false positives with a minimalist neural network. A new dataset for metonymy resolution was going to form part of our contribution due to a lack of quality evaluation data. The SemEval 2007 dataset was already available, however, a small revision to the annotation scheme was required. In order to action this change, an annotation task was organised in October-November 2016, hiring linguists from the MML faculty as expert consultants for the scheme update. For data creation, we chose to work with expert linguists (including computational) rather than using crowdsourcing. The following sections describe our research in the publication (long paper) titled *Vancouver Welcomes You! Minimalist Metonymy Resolution*[1] published (with oral presentation[2]) at the 55th Annual Meeting of the Association for Computational Linguistics in Vancouver, Canada in August, 2017. The accompanying resources are available on GitHub[3].

[1] https://aclanthology.coli.uni-saarland.de/papers/P17-1115/p17-1115

[2] https://vimeo.com/channels/acl2017/234945689

[3] https://github.com/milangritta/Minimalist-Location-Metonymy-Resolution

3.2 Minimalist Metonymy Resolution

Named entities are frequently used in a metonymic manner. They serve as references to related entities such as people and organisations. Accurate identification and interpretation of metonymy can be directly beneficial to various NLP applications, such as Named Entity Recognition and Geographical Parsing. Until now, metonymy resolution (MR) methods mainly relied on parsers, taggers, dictionaries, external word lists and other handcrafted lexical resources. We show how a minimalist neural approach combined with a novel predicate window method can achieve competitive results on the SemEval 2007 task on Metonymy Resolution. Additionally, we contribute with a new Wikipedia-based MR dataset called *RelocaR*, which is tailored towards locations as well as improving previous deficiencies in annotation guidelines.

In everyday language, we come across many types of figurative speech. These irregular expressions are understood with little difficulty by humans but require special attention in NLP. One of these is metonymy, a type of common figurative language, which stands for the substitution of the concept, phrase or word being meant with a semantically related one. For example, in “*Moscow* traded gas and aluminium with *Beijing*.", both location names were substituted in place of governments. Named Entity Recognition (NER) taggers have no provision for handling metonymy, meaning that this frequent linguistic phenomenon goes largely undetected within current NLP. Classification decisions presently focus on the entity using features such as orthography to infer its word sense, largely ignoring the context, which provides the strongest clue about whether a word is used metonymically.

A common classification approach is choosing the N words to the immediate left and right of the entity or the whole paragraph as input to the model. However, this “greedy" approach also processes input that should in practice be ignored. Metonymy is problematic for applications such as Geographical Parsing [138, 72] and other information extraction tasks in NLP. In order to accurately identify and ground location entities, for example, we must recognise that metonymic entities constitute false positives and should not be treated the same way as regular locations. For example, in “*London* voted for the change.", London refers to the concept of “people" and should not be classified as a location. There are many types of metonymy [174], however, in this paper, we primarily address metonymic location mentions with reference to GP and NER.

3.2.1 Related Work

Some of the earliest work on MR that used an approach similar to our method (machine learning and dependency parsing) was by [145]. The decision list classifier with backoff was evaluated using syntactic head-modifier relations, grammatical roles and a thesaurus to overcome data sparseness and generalisation problems. However, the method was still limited for classifying unseen data. Our method uses the same paradigm but adds more features, a different machine learning architecture and a better usage of the parse tree structure.

Much of the later work on MR comes from the SemEval 2007 Shared Task 8 [125] and later by [126]. The feature set of [145] was updated to include: grammatical role of the potentially metonymic word (PMW) (such as subj, obj), lemmatised head/modifier of PMW, determiner of PMW, grammatical number of PMW (singular, plural), number of words in PMW and number of grammatical roles of PMW in current context. The winning system by [55] used these features and a maximum entropy classifier to achieve 85.2% accuracy. This was also the “leanest" system but still made use of feature engineering and some external tools. [23] achieved 85.1% accuracy using local syntactical and global distributional features generated with an adapted, proprietary Xerox deep parser. This was the only unsupervised approach, based on using syntactic context similarities calculated on large corpora such as the the British National Corpus (BNC) with 100M tokens.

[142] used a Support Vector Machine (SVM) with handcrafted features (in addition to the features provided by [125]) including grammatical collocations extracted from the BNC to learn selectional preferences, WordNet 3.0, Wikipedia’s category network, whether the entity “has-a-product" such as Suzuki and whether the entity “has-an-event" such as Vietnam (both obtained from Wikipedia). The bigger set of around 60 features and leveraging global (paragraph) context enabled them to achieve 86.1% accuracy. Once again, we draw attention to the extra training, external tools and additional feature generation. Similar recent work by [143] which extends that of [141] involved transforming Wikipedia into a large-scale multilingual concept network called WikiNet. By building on Wikipedia’s existing network of categories and articles, their method automatically discovers new relations and their instances. As one of their extrinsic evaluations, metonymy resolution was tested. Global context (whole paragraph) was used to interpret the target word. Using an SVM and a powerful knowledge base built from Wikipedia, the highest performance to date (a 0.1% improvement from [142]) was achieved at 86.2%, which has remained the SOTA until now.

The related work on MR so far has made limited use of dependency trees. Typical features came in the form of a head dependency of the target entity, its dependency label and its role (subj-of-win, dobj-of-visit, etc). However, other classification tasks made good use of dependency trees. [117] used the shortest dependency path and dependency subtrees successfully to improve relation classification (new SOTA on SemEval 2010 Shared Task). [25] show that using dependency trees to generate the input sequence to a model performs well in relation extraction tasks. [49] used dependency parsing for Twitter sentiment classification to find the words syntactically connected to the target of interest. [90] used dependency parsing to explore how features based on syntactic dependency relations can be used to improve performance on opinion mining. In unsupervised lymphoma (type of cancer) classification, [118] constructed a sentence graph from the results of a two-phase dependency parse to mine pathology reports for the relationships between medical concepts. Our methods also exploit dependency parsing to leverage information about the sentence structure.

SemEval 2007 Dataset

Our main standard for performance evaluation is the SemEval 2007 Shared Task 8 [125] dataset first introduced in [146]. Two types of entities were evaluated, organisations and locations, randomly retrieved from the British National Corpus (BNC). We only use the locations dataset, which comprises a train (925 samples) and a test (908 samples) partition. For *medium* evaluation, the classes are *literal* (geographical territories and political entities), *metonymic* (place-for-people, place-for-product, place-for-event, capital-for-government or place-for-organisation) and *mixed* (metonymic and literal frames invoked simultaneously or unable to distinguish). The metonymic class further breaks down into two levels of subclasses allowing for *fine* evaluation. The class distribution within SemEval is approx 80% literal, 18% metonymic and 2% mixed. This seems to be the approximate natural distribution of the classes for location metonymy, which we have also observed while sampling Wikipedia.

3.2.2 The ReLocaR Dataset and its Improved Annotation

As part of our contribution, we created a new MR dataset called ReLocaR (Real Location Retrieval), partly due to the lack of quality annotated train/test data and partly because of the shortcomings with the SemEval 2007 dataset. Our corpus is designed to evaluate the capability of a classifier to distinguish *literal*, *metonymic* and *mixed* location mentions. In terms of dataset size, ReLocaR contains 1,026 training and 1,000 test instances. The data was sampled using Wikipedia's Random Article API[4]. We kept the sentences that contained at least one of the places from a manually compiled list[5] of countries and capitals of the world. The natural distribution of literal versus metonymic examples is approximately 80/20 so we had to discard the excess literal examples during sampling to balance the classes.

Improvements over SemEval

- We do not break down the metonymic class further as the distinction between the subclasses is subtle and hard to agree on.
- The distribution of the three classes in ReLocaR (literal, metonymic, mixed) is approximately (49%, 49%, 2%) eliminating the high bias (80%, 18%, 2%) of SemEval. We will show how such a high bias transpires in the test results.
- We have reviewed the annotation of the test partition and found that we disagreed with up to 11% of the annotations. [205] disagreed with the annotation 8% of the time. [156] also challenged some annotation decisions. ReLocaR was annotated by 4 trained linguists (undergraduate and graduate) and 2 computational linguists (authors). Linguists were independently instructed to assign one of the two classes to each example with little guidance. We leveraged their linguistic training and expertise to make decisions rather than imposing some specific scheme. Unresolved sentences would receive the mixed class label.
- The most prominent difference is a small change in the annotation scheme (after independent linguistic advice). The SemEval 2007 Task 8 annotation scheme [125] considers the political entity interpretation a literal reading. It suggests that in "*Britain*'s current account deficit...", Britain refers to a literal location, rather than a government (which is an organisation). This is despite acknowledging that "The locative and the political sense is often distinguished in dictionaries as well as in the ACE annotation scheme...". In ReLocaR datasets, we consider a political entity a metonymic reading.

Why government is not a location A government/nation/political entity is semantically much closer to Organisation/Person than a Location. "*Moscow* talks to *Beijing*." does not tell us *where* this is happening. It most likely means a politician is talking to another politician. These are not places but people and/or groups. It is paramount to separate references to "inanimate" places from references to "animate" entities.

[4] https://www.mediawiki.org/wiki/API:Random

[5] https://github.com/milangritta/Minimalist-Location-Metonymy-Resolution/data/

Annotation Guidelines (Summary)

ReLocaR has three classes, *literal*, *metonymic* and *mixed*. Literal reading comprises territorial interpretations (the geographical territory, the land, soil and physical location) i.e. inanimate places that serve to point to a set of coordinates (where something might be located and/or happening) such as “The treaty was signed in *Italy*.", “Peter comes from *Russia*.", “*Britain*’s Andy Murray won the Grand Slam today.", “*US* companies increased exports by 50%.", “*China*’s artists are among the best in the world." or “The reach of the transmission is as far as *Brazil*.". A metonymic reading is any location occurrence that expresses animacy [36] such as “*Jamaica*’s indifference will not improve the negotiations.", “*Sweden*’s budget deficit may rise next year.". The following are other metonymic scenarios: a location name, which stands for any persons or organisations associated with it such as “We will give aid to *Afghanistan*.", a location as a product such as “I really enjoyed that delicious *Bordeaux*.", a location posing as a sports team “*India* beat *Pakistan* in the playoffs.", a governmental or other legal entity posing as a location “*Zambia* passed a new justice law today.", events acting as locations “*Vietnam* was a bad experience for me". The mixed reading is assigned in two cases: either both readings are invoked at the same time such as in “The Central European country of *Slovakia* recently joined the EU." or there is not enough context to ascertain the reading i.e. both are plausible such as in “We marvelled at the art of ancient *Mexico*.". In difficult cases such as these, the mixed class is assigned.

Inter-Annotator Agreement

We now give the IAA for the test set. The full set was annotated by the first author as the main annotator. Two pairs of annotators (all MML[6] linguists) then labelled 25% of the dataset each for a 3-way agreement with the main author. The agreement before adjudication was 91% and 93% for the first and second sample respectively, 97.2% and 99.2% after adjudication. The other 50% of the test set was once again labelled by the main annotator with a 97% agreement with self. The remainder of the sentences without a clear agreement among annotators, even after adjudication, were labelled as a mixed class (1.8% of sentences).

CoNLL 2003 and MR

We have also annotated a small subset of the CoNLL 2003 NER Shared Task data for metonymy resolution (locations only). Respecting the Reuters RCV1 Corpus [108] distribu-

[6] https://www.mml.cam.ac.uk/

tion permissions[7], we make only a heavily processed subset available on GitHub[8]. There are 4,089 positive (literal) and 2,126 negative (metonymic) sentences to assist with algorithm experimentation and model prototyping. Due to the lack of annotated data for MR, this is a valuable resource. The data was annotated by the first author, there are no IAA figures.

3.2.3 Predicate Window (PreWin)

Through extensive experimentation and observation, we arrived at the intuition behind PreWin, our novel feature extraction method. The classification decision of the class of the target entity is mostly informed not by the whole sentence (or paragraph), rather it is a small and focused “predicate window" pointed to by the entity’s head dependency. In other words, most of the sentence is not only superfluous for the task, it actually lowers the accuracy of the model due to irrelevant input.

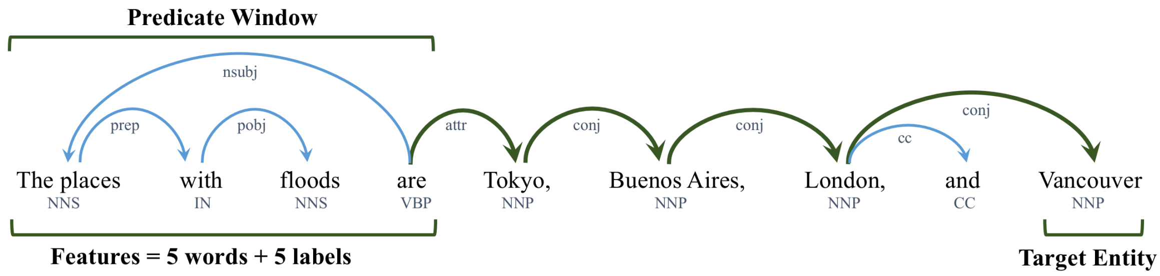


Fig. 3.1 Why it is important for PreWin to always skip the *conjunct* dependency relation.

This is particularly important in metonymy resolution as the entity’s surface form is not taken into consideration, only its context. Figures 3.3, 3.1, 3.2 and 3.4 show the process of extracting the Predicate Window from a number of sample sentences. We start by using the SpaCy dependency parser by [84], which is the fastest in the world, open source and highly customisable. Each dependency tree provides the following features: dependency labels and entity head dependency. Rather than using most of the tree, we only use a single local head dependency relationship to point to the predicate.

Leveraging a dependency parser helps PreWin with selecting the minimum relevant input to the model while discarding irrelevant input, which may cause the neural model to behave unpredictably. Finally, the entity itself is never used as input in any of our methods, we only rely on context. PreWin then extracts up to 5 words and their dependency labels starting at the *head* of the entity (see the next paragraph for exceptions), going in the away (from the entity) direction. The method always skips the *conjunct* (“and", “or") relationships in order

[7] http://trec.nist.gov/data/reuters/reuters.html
[8] https://github.com/milangritta/Minimalist-Location-Metonymy-Resolution

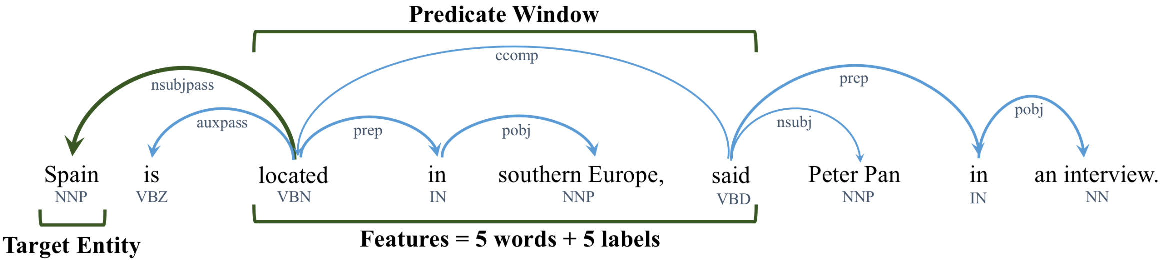


Fig. 3.2 A lot of irrelevant input is skipped such as “is" and “Peter Pan in an interview.".

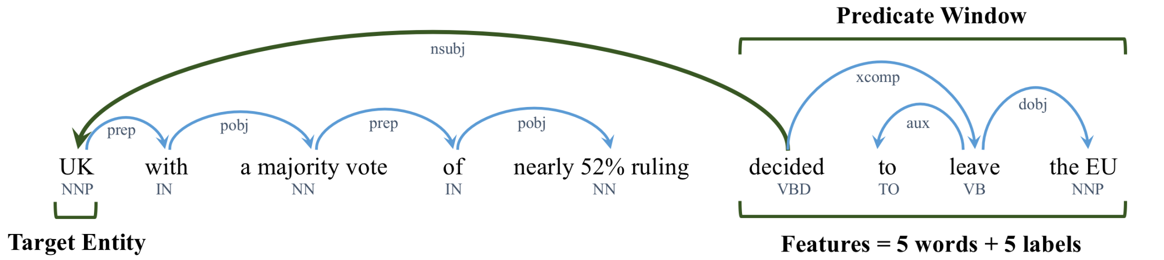


Fig. 3.3 The predicate window starts at the head of the target entity and ends up to 4 words further, going away from the entity. The “conj" relations are always skipped. In the above example, the head of “UK" is “decided" so PreWin takes 5 words plus dependency labels as the input to the model. The left-hand side input to the model is empty and is set to zeroes.

to find the predicate. The reason for the choice of 5 words is the balance between too much input, feeding the model with less relevant context and just enough context to capture the necessary semantics.

We have experimented with lengths of 3-10 words, however 5 words typically achieved the best results. The following are the three types of exceptions when the output will not start with the head of the entity. In these cases, PreWin will include the neighbouring word as well. In a sentence “The pub is located in southern *Zambia*.", the head of the entity is “in”, however in this case PreWin will include “southern" (*adjectival modifier*) as this carries important semantics for the classification. Similarly, PreWin will also include the neighbouring *compound noun* as in: “Lead coffins were very rare in colonial *America*.”, the output will include “colonial” as a feature plus the next four words. In another sentence: “*Vancouver*’s security is the best in the world.’, PreWin will include the “*’s*” (*case*) plus the next four words continuing from the head of the entity (the word “security”).

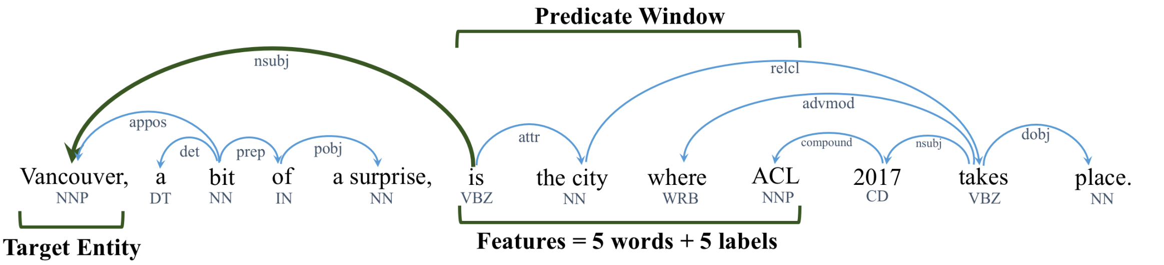


Fig. 3.4 By looking for the predicate window, the model skips many irrelevant words.

3.2.4 Neural Network Architecture

The output of PreWin is used to train the following machine learning model. We decided to use the Long Short Term Memory (LSTM) architecture by Keras[9] [33]. Two LSTMs are used, one for the left and right side (up to 5 words each). Two fully connected (dense) layers are used for the left and right dependency relation labels (up to 5 labels each, encoded as one-hot). You can download the models and data from GitHub[10]. LSTMs are excellent at processing language sequences [81, 168, 68], which is why we use this architecture. It allows the model to encode the word sequences, preserve important word order and provide superior classification performance. Both the Multilayer Perceptron and the Convolutional Neural Network were consistently inferior (typically 5% - 10% lower accuracy) in our earlier performance comparisons. For all experiments, we used a vocabulary of the first (most frequent) 100,000 word vectors in GloVe[11] [151]. Finally, unless explicitly stated otherwise, the standard dimension of word embeddings was 50.

“Immediate" Baseline A common approach in lexical classification tasks is choosing the 5 to 10 words to the immediate right and left of the entity as input to a model [136, 134, 13, 35]. We evaluate this method (its 5 and 10-word variant) alongside PreWin and Paragraph.

Paragraph Baseline The paragraph baseline method extends the “immediate" one by taking 50 words from each side of the entity as the input to the classifier. In practice, this extends the feature window to include extra-sentential evidence in the paragraph. This approach is also popular in machine learning [131, 203].

[9] https://keras.io/
[10] https://github.com/milangritta/Minimalist-Location-Metonymy-Resolution
[11] http://nlp.stanford.edu/projects/glove/

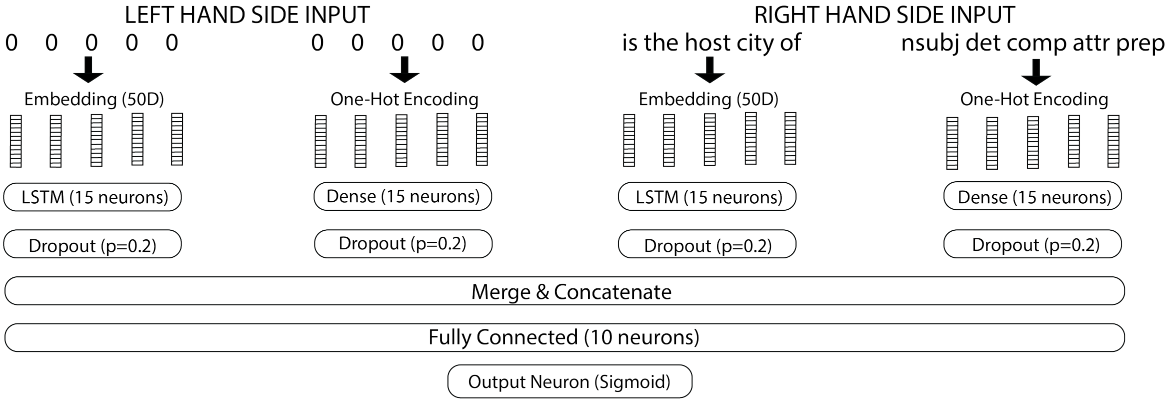


Fig. 3.5 The neural architecture of the final model. The sentence is *Vancouver is the host city of the ACL 2017.* Small, separate sequential models are merged and trained as one. The 50-dimensional embeddings were initiated using GloVe. The *right hand* input is processed from right to left, the *left hand* input is processed from left to right. This is to emphasise the importance of the words *closer* to the entity.

Ensemble of Models In addition to a single best performing model, we have combined several models trained on different data and/or using different model configurations. For the SemEval test, we combined three separate models trained on the newly annotated CoNLL dataset and the training data for SemEval. For the ReLocaR test, we once again let three models vote, trained on CooNLL and ReLocaR data.

3.2.5 Results and Error Analysis

We evaluate all methods using three datasets for *training* (ReLocaR, SemEval, CoNLL) and two for *testing* (ReLocaR, SemEval). Due to inherent randomness in the deep learning libraries, we performed 10 runs for each setup and averaged the figures.

Metrics and Significance

Following the SemEval 2007 convention, we use two metrics to evaluate performance, accuracy and f-scores (for each class). We only evaluate at the *coarse level*, which means literal versus non-literal (metonymic and mixed are merged into one class). In terms of statistical significance, our best score on the SemEval dataset (908 samples) is not significant at the 95% confidence level. However, the accuracy improvements of PreWin over the common baselines are highly statistically significant with 99.9%+ confidence.

Predicate Window

Tables 3.1 and 3.2 show PreWin performing consistently better than other baselines, in many instances, significantly better and with fewer words (smaller input). The standard deviation is also lower for PreWin meaning more stable test runs. Compared with the 5 and 10 window “immediate" baseline, which is the common approach in classification, PreWin is more discriminating with its input. Due to the linguistic variety and the myriad of ways the target word sense can be triggered in a sentence, it is not always the case that the 5 or 10 nearest words inform us of the target entity’s meaning/type. We ought to ask what else is being expressed in the same 5 to 10-word window?

Conventional classification methods (Immediate, Paragraph) can also be seen as prioritising either feature precision or feature recall. *Paragraph* maximises the input sequence size, which maximises recall at the expense of including features that are either irrelevant or mislead the model, lowering precision. *Immediate* baseline maximises precision by using features close to the target entity at the expense of missing important features positioned outside of its small window, lowering recall. PreWin can be understood as an integration of both approaches. It retains high precision by limiting the size of the feature window to 5 while maximising recall by searching anywhere in the sentence, frequently outside of a limited “immediate" window.

Perhaps we can also caution against a simple adherence to [57] *“You shall know a word by the company it keeps"*. This does not appear to be the case in our experiments as PreWin regularly performs better than the “immediate" baseline. Our intuition that most words in the sentence, indeed in the paragraph do not carry the semantic information required to classify the target entity is ultimately based on evidence. The model uses only a small window, linked to the entity via a head dependency relationship for the final classification decision.

Flexibility of Neural Model The top accuracy figures for ReLocaR are almost identical to SemEval. The highest single model accuracy for ReLocaR was 83.6% (84.8% with Ensemble), which was within 0.5% of the equivalent methods for SemEval (83.1%, 84.6% for Ensemble). Both were achieved using the same methods (PreWin or Ensemble), neural architecture and size of corpora. When the models were trained on the CoNLL data, the accuracies were 82.8% and 79.5%. However, when the models trained on ReLocaR and tested on SemEval (and vice versa), accuracy dropped to between 62.4% and 69% showing that what was learnt does not seem to transfer well to another dataset. We think the reason

Method	Train Data	Data Size	Accuracy	St. Dev.
PreWin	SemEval	925	**62.4**	2.30
Immediate 5	SemEval	925	60.6	2.34
Immediate 10	SemEval	925	59.2	2.26
Paragraph	SemEval	925	58.0	2.49
PreWin	CoNLL	6,215	**82.8**	0.46
Immediate 5	CoNLL	6,215	78.2	0.61
Immediate 10	CoNLL	6,215	79.1	0.76
Paragraph	CoNLL	6,215	79.5	1.50
PreWin	ReLocaR	1,026	**83.6**	0.71
Immediate 5	ReLocaR	1,026	81.4	1.34
Immediate 10	ReLocaR	1,026	81.3	1.44
Paragraph	ReLocaR	1,026	80.0	2.25
Ensemble	ReLocaR + CoNLL	7,241	**84.8**	0.34

Table 3.1 Results for ReLocaR. Figures averaged over 10 runs.

for this is the difference in annotation guidelines; the government is a *metonymic* reading, not a literal one. This causes the model to make more mistakes.

Ensemble Method The highest accuracy and f-scores were achieved with the ensemble method for both datasets. We combined three models (previously described in section 3.2.4) for SemEval to achieve 84.6% accuracy and three models for ReLocaR to achieve 84.8% for the new dataset. Training separate models with different parameters and/or on different datasets does increase classification capability as various models learn distinct aspects of the task, enabling the 1.2 - 1.5% improvement.

Dimensionality of Word Embeddings We found that increasing dimension size (up to 300) did not materially improve performance. The neural network tended to overfit, even with fewer epochs, the results were comparable to our default 50-dimensional embeddings. We posit that fewer dimensions of the distributed word representations force the abstraction level higher as the meaning of words must be expressed more succinctly. We think this helps the model generalise better, particularly for smaller datasets. Lastly, learning word embeddings from scratch on datasets this small (around 1,000 samples) is possible but impractical, the performance typically decreases by around 5% if word embeddings are not initialised first.

F-Scores and Class Imbalance Table 3.3 shows the SOTA f-scores, our best results for SemEval 2007 and the best f-scores for ReLocaR. The class imbalance inside SemEval (80%

Method	Train Data	Data Size	Accuracy	St. Dev.
PreWin	SemEval	925	**83.1**	0.64
Immediate 5	SemEval	925	81.3	1.11
Immediate 10	SemEval	925	81.9	0.89
Paragraph	SemEval	925	81.3	0.88
PreWin	CoNLL	6,215	**79.5**	0.34
Immediate 5	CoNLL	6,215	77.8	1.47
Immediate 10	CoNLL	6,215	77.8	1.22
Paragraph	CoNLL	6,215	77.2	2.10
PreWin	ReLocaR	1,026	**69.0**	3.13
Immediate 5	ReLocaR	1,026	63.6	5.42
Immediate 10	ReLocaR	1,026	64.2	4.12
Paragraph	ReLocaR	1,026	64.4	7.76
Nastase et al.	SemEval	925	**86.2**	N/A
Ensemble	SemEval + CoNLL	7,140	84.6	0.43

Table 3.2 Results for SemEval. Figures averaged over 10 runs.

literal, 18% metonymic, 2% mixed) is reflected as a high bias in the final model. This is not the case with ReLocaR and its 49% literal, 49% metonymic and 2% mixed ratio of 3 classes. The model was equally capable of distinguishing between literal and non-literal cases.

Common Errors

Most of the time (typically 85% for the two datasets), PreWin is sufficient for an accurate classification. However, it does not work well in some cases. The typical 15% error rate breaks down as follows (percentages were estimated based on extensive experimentation):

Discarding important context (3%): Sometimes the 5 or 10 word “immediate" baseline method would actually have been preferred such as in the sentence “...REF in 2014 ranked *Essex* in the top 20 universities...". PreWin discards the right-hand side input, which is required in this case for a correct classification. Since "ranked" is the head of "Essex", the rest of the sentence gets ignored and the valuable context gets lost.

More complex semantic patterns (11%): Many common mistakes were due to the lack of the model’s understanding of more complex predicates such as in the following sentences: “ ...of military presence of *Germany*.", “*Houston* also served as a member and treasurer of the..." or "...invitations were extended to *Yugoslavia* ...". We think this is due to a lack of

Dataset / Method	Literal	Non-Literal
SemEval / PreWin	90.6	57.3
SemEval / SOTA	**91.6**	**59.1**
ReLocaR / PreWin	84.4	84.8

Table 3.3 Per class f-scores using the Ensemble method, averaged over 10 runs.

training data (around 1,000 sentences per dataset). Additional examples such as "...days after the tour had exited *Belgium*." expose some of the limitations of the neural model to recognise uncommon ways of expressing a reference to a literal place. Recall that no external resources or tools were used to supplement the training/features, the model had to learn to generalise from what it has seen during training, which was limited in our experiments.

Parsing mistakes (1%): were less common though still present. It is important to choose the right dependency parser for the task since different parsers will often generate slightly different parse trees. We have used SpaCy[12] for all our experiments, which is a Python-based industrial strength NLP library. Sometimes, tokenisation errors for acronyms like "U.S.A." and wrongly hyphenated words may also cause parsing errors, however, this was infrequent.

3.2.6 Discussion

NER, GP and Metonymy We think the next frontier is a NER tagger, which actively handles metonymy. The task of labelling entities should be mainly driven by context rather than the word's surface form. If the target entity looks like "*London*", this should not mean the entity is automatically a location. Metonymy is a frequent linguistic phenomenon (around 20% of location mentions are metonymic and could be handled by NER taggers to enable many innovative downstream NLP applications. Geographical Parsing is a pertinent use case. In order to monitor/mine text documents for geographical information only, the current NER technology does not have a solution. We think it is incorrect for any NER tagger to label "Vancouver" as a location in "*Vancouver welcomes you!*". A better output might be something like the following: *Vancouver = location AND metonymy = True*. This means Vancouver is usually a location but is used metonymically in this case. How this information is used will be up to the developer. Organisations behaving as persons, share prices or products are but a few other examples of metonymy.

[12] https://spacy.io/

Simplicity and Minimalism Previous work in MR such as most of the SemEval 2007 participants [55, 144, 106, 23, 156] and the more recent contributions used a selection of many of the following features/tools for classification: handmade trigger word lists, WordNet, VerbNet, FrameNet, extra features generated/learnt from parsing Wikipedia (approx 3B words) and BNC (approx 100M words), custom databases, handcrafted features, multiple (sometimes proprietary) parsers, Levin's verb classes, 3,000 extra training instances from a corpus called MAScARA[13] by [124] and other extra resources including the SemEval Task 8 features. We managed to achieve comparable performance with a small neural network typically trained in no more than 5 epochs, minimal training data, a basic dependency parser and the new PreWin method by being highly discriminating in choosing signal over noise.

3.2.7 Conclusions

We showed how a minimalist neural approach can replace substantial external resources, handcrafted features and how the PreWin method can even ignore most of the paragraph where the entity is positioned and still achieve competitive performance in metonymy resolution. The pressing new question is: "How much better the performance could have been if our method availed itself of the extra training data and resources used by previous works?" Indeed this may be the next research chapter for PreWin. We discussed how tasks such as Geographical Parsing can benefit from "metonymy-enhanced" NER tagging. We have also presented a case for better annotation guidelines for MR (after consulting with a number of linguists), which now means that a government is not a literal class, rather it is a metonymic one. We fully agreed with the rest of the previous annotation guidelines. We also introduced ReLocaR, a new corpus for (location) metonymy resolution and encourage researchers to make effective use of it (including the CoNLL 2003 subset we annotated for metonymy).

Future work may involve testing PreWin on an NER task to see if and how it can generalise to a different classification task and how the results compare to the SOTA and similar methods such as that of [35] using the CoNLL 2003 NER datasets. Word Sense Disambiguation [200, 155] with neural networks [131] is another related classification task suitable for testing PreWin. If it does perform better, this will be of considerable interest to classification research (and beyond) in NLP.

[13] http://homepages.inf.ed.ac.uk/mnissim/mascara/

3.3 Summary

Metonymy and other types of non-literal toponyms are not covered by Named Entity Recognition. NER models are trained to predict *the most likely class* of an entity, apparently paying little attention to surrounding words. This is because of the polysemous, contextual uses of entities such as demonyms, metonyms, homonyms, languages, modifiers and so on. The only consistent cue the model optimises its objective over is the *entity string itself*. In a way, the training data discourages the use of context features as it proves to be an inconsistent indicator of the entity class. One can explore this behaviour with almost any NER demo interface by changing the context arbitrarily without a change in classification behaviour. It is possible to fool even advanced NER sequence taggers (e.g. Google Cloud NLP) by supplying a random and/or contradictory context and the model prediction will not change.

The addition of more evaluation resources continues with ReLocaR.xml (over 2,000 annotated samples) and a metonymy-annotated version of CoNLL 2003 data for an additional 6,215 training examples. These efforts along with the necessary changes to the SemEval 2007 annotation scheme were made with the consultation of Cambridge MML linguists. Although a correction has been suggested in previous work, it was our research efforts that implemented the changes and generated the open-source data. We showed that latest deep learning tools make the historically engineered features for this task seem costly and slow. Many of the previous solutions featured an elaborate setup of a combination of multiple processes and resources, which can be replaced with an equally effective neural classifier. This is good news for MediSys/EMM and other applications as it means implementing an effective model fast and inexpensively. PreWin was introduced as a simple and effective method of feature extraction coupled with a minimalist neural architecture. The discovery of PreWin emerged chiefly from the analysis of existing datasets and the construction of dataset annotation process. Our classifier had the lowest standard deviation compared to all baselines, i.e., had the most consistent predictions. In an email a from a Google Research Fellow, it was suggested that an attention model/mechanism could be tested to investigate whether we may be able to dispense with the dependency parser step. We think this might constitute a worthy future experiment but in any case, we were delighted to have the paper and ACL slides included in the reading session of Google Research in September 2017.

Chapter 4

Toponym Resolution

4.1 Introduction

The second stage of geoparsing is toponym resolution, also called geocoding, which is the task of grounding toponyms to their geographic coordinates. The work described in this chapter has three parts: 1) Leverage deep learning tools to improve the state of the art in geocoding. The reasoning behind this choice was to let the important features be induced by the model rather than handcrafted and learnt with 'classical' machine learning, e.g. Random Forest and/or Naive Bayes. Machine learning has been employed in toponym resolution before, however, to our best knowledge, the use of GPU-accelerated neural networks has not yet featured in this area. In deep learning, focus shifts from feature design to task objective and architecture design; from supplying features to supplying quality data. 2) Virtually all previous approaches we encountered disambiguated toponyms using only *lexical features*. We began investigating new ways of inducing additional features to encode the deeper semantics of the context. This resulted in Map Vector, a *geographic vector representation* of context, which captures features of geography that word embeddings do not, as shown in multiple experiments. 3) Our earlier empirical survey generated resources and evaluation scores over two datasets, which we used to benchmark the models. As we have done with previous papers, we contributed with a new open-source dataset. High-quality annotated data from heterogeneous domains are not readily available in this field hence we decided to generate resources for a more holistic evaluation. The manuscript[1] was (orally) presented[2] at the 56th Annual Meeting of the Association for Computational Linguistics in Melbourne, Australia in July, 2018. The accompanying resources can be downloaded from GitHub[3].

[1] http://www.aclweb.org/anthology/P18-1119

[2] https://vimeo.com/285803462

[3] https://github.com/milangritta/Geocoding-with-Map-Vector

4.2 Geocoding with Map Vector

The purpose of text geolocation is to associate geographic information contained in a document with a set (or sets) of coordinates, either *implicitly* by using linguistic features and/or *explicitly* by using geographic metadata combined with heuristics. We introduce a geocoder (location mention disambiguator) that achieves state-of-the-art (SOTA) results on three diverse datasets by exploiting the implicit lexical clues. Moreover, we propose a new method for systematic encoding of geographic metadata to generate two *distinct* views of the same text. To that end, we introduce the *Map Vector* (*MapVec*), a sparse representation obtained by plotting prior geographic probabilities, derived from population figures, on a World Map. We then integrate the implicit (language) and explicit (map) features to significantly improve a range of metrics. We also introduce an open-source dataset for geoparsing of news events covering global disease outbreaks and epidemics to help future evaluation in geoparsing.

Geocoding[4] is a specific case of text geolocation, which aims at disambiguating place references in text. For example, *Melbourne* can refer to more than ten possible locations and a geocoder's task is to identify the place coordinates for the intended *Melbourne* in a context such as "*Melbourne* hosts one of the four annual Grand Slam tennis tournaments." This is central to the success of tasks such as indexing and searching documents by geography [15], geospatial analysis of social media [24], mapping of disease risk using integrated data [76], and emergency response systems [8]. Previous geocoding methods have leveraged lexical semantics to associate the implicit geographic information in natural language with coordinates. These models have achieved good results in the past. However, focusing *only* on lexical features, to the exclusion of other feature spaces such as the Cartesian Coordinate System, puts a ceiling on the amount of semantics we are able to extract from text.

Our proposed solution is the *Map Vector* (*MapVec*), a sparse, geographic vector for explicit modelling of geographic distributions of location mentions. As in previous work, we use population data and geographic coordinates, observing that the most populous *Melbourne* is also the most likely to be the intended location. However, MapVec is the first instance, to our best knowledge, of the topological semantics of context locations explicitly isolated into a standardized vector representation, which can then be easily transferred to an independent task and combined with other features. MapVec is able to encode the prior geographic distribution of any number of locations into a single vector. Our extensive evaluation shows how this representation of context locations can be integrated with linguistic features to

[4] Also called Toponym Resolution in related literature.

achieve a significant improvement over a SOTA lexical model. MapVec can be deployed as a standalone neural geocoder, significantly beating the population baseline, while remaining effective with simpler machine learning algorithms.

This paper's contributions are: (1) *Lexical Geocoder* outperforming existing systems by analysing only the textual context; (2) *MapVec*, a geographic representation of locations using a sparse, probabilistic vector to extract and isolate spatial features; (3) *CamCoder*, a novel geocoder that exploits both lexical and geographic knowledge producing SOTA results across multiple datasets; and (4) *GeoVirus*, an open-source dataset for the evaluation of geoparsing (Location Recognition *and* Disambiguation) of news events covering global disease outbreaks and epidemics.

4.2.1 Related Work

Depending on the task objective, geocoding methodologies can be divided into two *distinct* categories: (1) *document geocoding*, which aims at locating a piece of text as a whole, for example geolocating Twitter users [160, 159, 166, 161], Wikipedia articles and/or web pages [30, 10, 196, 52, 195]. This is an active area of NLP research [87, 133, 132, 88]; (2) *geocoding of place mentions*, which focuses on the disambiguation of location (named) entities i.e. this paper and [95, 186, 73, 41, 170, 177, 204]. Due to the differences in evaluation and objective, the categories cannot be directly or fairly compared. Geocoding is typically the second step in *Geoparsing*. The first step, usually referred to as *Geotagging*, is a Named Entity Recognition component which extracts all location references in a given text. This phase may optionally include metonymy resolution, see [205, 71]. The goal of geocoding is to choose the correct coordinates for a location mention from a set of candidates. [72] provided a comprehensive survey of five recent geoparsers. The authors established an evaluation framework, with a new dataset, for their experimental analysis. We use this evaluation framework in our experiments. We briefly describe the methodology of each geocoder featured in our evaluation (names are capitalised and appear in italics) as well as survey the related work in geocoding.

Computational methods in geocoding broadly divide into rule-based, statistical and machine learning-based. *Edinburgh Geoparser* [186, 73] is a fully rule-based geocoder that uses hand-built heuristics combined with large lists from Wikipedia and the Geonames[5] gazetteer. It uses metadata (feature type, population, country code) with heuristics such as contextual

[5] http://www.geonames.org/

information, spatial clustering and user locality to rank candidates. *GeoTxT* [95] is another rule-based geocoder with a free web service[6] for identifying locations in unstructured text and grounding them to coordinates. Disambiguation is driven by multiple heuristics and uses the administrative level (country, province, city), population size, the Levenshtein Distance of the place referenced and the candidate's name and spatial minimisation to resolve ambiguous locations. [52] is a rule-based Twitter geocoder using only metadata (coordinates in tweets, GPS tags, user's reported location) and custom place lists for fast and simple geocoding. *CLAVIN* (Cartographic Location And Vicinity INdexer)[7] is an open-source geocoder, which offers context-based entity recognition and linking. It seems to be mostly rule-based though details of its algorithm are underspecified, short of reading the source code. Unlike the Edinburgh Parser, this geocoder seems to overly rely on population data, seemingly mirroring the behaviour of a naive population baseline. Rule-based systems can perform well though the variance in performance is high (Table 4.1). *Yahoo! Placemaker* is a free web service with a proprietary geo-database and algorithm from Yahoo![8] letting anyone geoparse text in a globally-aware and language-independent manner. It is unclear how geocoding is performed, however, the inclusion of proprietary methods makes evaluation more informative.

The statistical geocoder *Topocluster* [41] divides the world surface into a grid (0.5 x 0.5 degrees, ~60K tiles) and uses lexical features to model the geographic distribution of context words over this grid. Building on the work of [177], it uses a window of 15 words to perform hot spot analysis using Getis-Ord Local Statistic of individual words' association with geographic space. The classification decision was made by finding the grid square with the strongest overlap of individual geo-distributions. [87] used Kernel Density Estimation to learn the word distribution over a world grid with a resolution of 0.5 x 0.5 degrees and classified documents with Kullback-Leibler divergence or a Naive Bayes model, reminiscent of an earlier approach by [196]. [166] used the Good-Turing Frequency Estimation to learn document probability distributions over the vocabulary with Kullback-Leibler divergence as the similarity function to choose the correct bucket in the k-d tree (world representation). [88] combined Gaussian Density Estimation with a CNN-model to geolocate Japanese tweets with Convolutional Mixture Density Networks.

Among the recent machine learning methods, bag-of-words representations combined with a Support Vector Machine [132] or Logistic Regression [195] have also achieved good results. For Twitter-based geolocation [204], bag-of-words classifiers were successfully

[6] http://www.geotxt.org/
[7] https://clavin.bericotechnologies.com
[8] https://developer.yahoo.com/geo/

augmented with social network data [92, 160, 161]. The machine learning-based geocoder by [170] supplemented lexical features, represented as a bag-of-words, with an exhaustive set of manually generated geographic features and spatial heuristics such as geospatial containment and geodesic distances between entities. The ranking of locations was learned with LambdaMART. Unlike our geocoder, the addition of geographic features did not significantly improve scores, reporting: “The geo-specific features seem to have a limited impact over a strong baseline system.” Unable to obtain a codebase, their results feature in Table 4.1. The latest neural network approaches [159] with normalised bag-of-word representations have achieved SOTA scores when augmented with social network data for Twitter document (user’s concatenated tweets) geolocation [11].

4.2.2 Neural Network Geocoder

Figure 4.1 shows our new geocoder *CamCoder* implemented in Keras [33]. The lexical part of the geocoder has three inputs, from the top: Context Words (location mentions excluded), Location Mentions (context words excluded) and the Target Entity (up to 15 words long) to be geocoded. Consider an example disambiguation of *Cairo* in a sentence: “*The Giza pyramid complex is an archaeological site on the Giza Plateau, on the outskirts of Cairo, Egypt.*”. Here, *Cairo* is the Target Entity; *Egypt*, *Giza* and *Giza Plateau* are the Location Mentions; the rest of the sentence forms the Context Words (excluding stopwords). The *context window* is up to 200 words each side of the Target Entity, approximately an order of magnitude larger than most previous approaches.

We used separate layers, convolutional and/or dense (fully-connected), with ReLu activations [140] to break up the task into smaller, focused modules in order to learn distinct lexical feature patterns, phrases and keywords for different types of inputs, concatenating only at a higher level of abstraction. Unigrams and bigrams were learned for context words and location mentions (1,000 filters of size 1 and 2 for each input), trigrams for the target entity (1,000 filters of size 3). Convolutional Neural Networks (CNNs) with Global Maximum Pooling were chosen for their position invariance (detecting location-indicative words anywhere in context) and efficient input size scaling. The dense layers have 250 units each, with a dropout layer ($p = 0.5$) to prevent overfitting. The fourth input is MapVec, the geographic vector representation of location mentions. It feeds into two dense layers with 5,000 and 1,000 units respectively. The concatenated hidden layers then get fully connected to the softmax layer. The model is optimised with RMSProp [184].

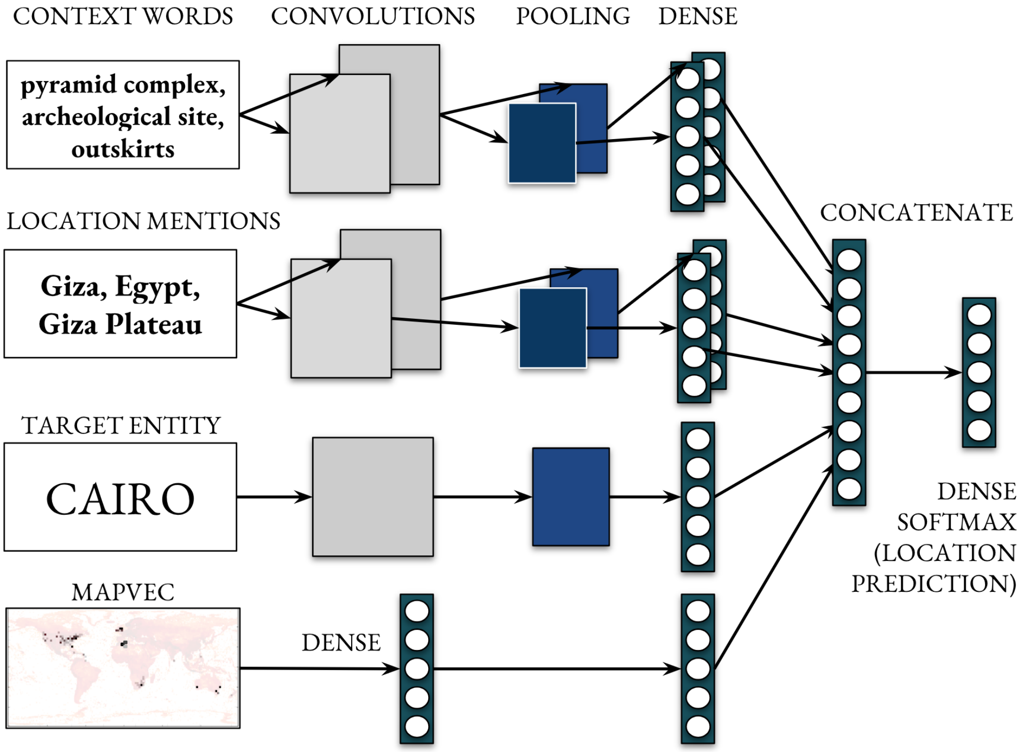

CONTEXT WORDS
CONVOLUTIONS
POOLING
DENSE
pyramid complex, archeological site, outskirts
LOCATION MENTIONS
Giza, Egypt, Giza Plateau
TARGET ENTITY
CAIRO
MAPVEC
DENSE
CONCATENATE
DENSE SOFTMAX (LOCATION PREDICTION)

Fig. 4.1 The *CamCoder* neural architecture. It is possible to split CamCoder into a *Lexical* (top 3 inputs) model and a *MapVec* model (see Table 4.2).

We approach geocoding as a classification task where the model predicts one of 7,823 classes (units in the softmax layer in Figure 4.1), each being a 2x2 degree tile representing part of the world's surface, slightly coarser than MapVec (see Section 4.2.3 next). The coordinates of the location candidate with the smallest *FD* (Equation 4.1) are the model's final output. *FD* for each candidate is computed by reducing the prediction *error* (the distance from predicted coordinates to candidate coordinates) by the value of *error* multiplied by the estimated prior probability (candidate population divided by maximum population) multiplied by the *Bias* parameter. The value of $Bias = 0.9$ was determined to be optimal for highest development data scores and is *identical for all* highly diverse test datasets. Equation 4.1 is designed to bias the model towards more populated locations to reflect real-world data.

$$FD = error - error\,\frac{candidatePop}{maximumPop}\,Bias \tag{4.1}$$

4.2.3 The Map Vector (MapVec)

Word embeddings and/or distributional vectors encode a word's meaning in terms of its *linguistic* context. However, location (named) entities also carry explicit *topological semantic*

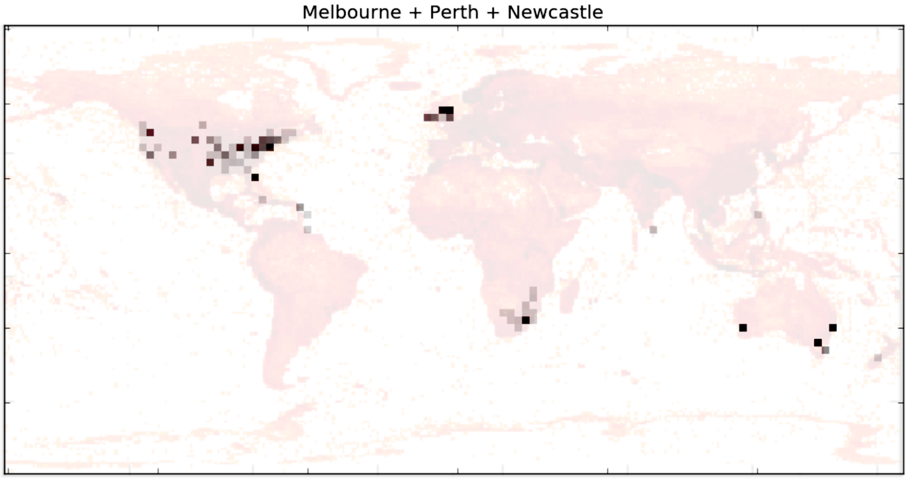


Fig. 4.2 MapVec visualisation (before reshaping into a 1D vector) for *Melbourne*, *Perth* and *Newcastle*, showing their combined prior probabilities. Darker tiles have higher probability.

knowledge such as a coordinate position and a population count for all places with an identical name. Until now, this knowledge was only used as part of simple disparate heuristics and manual disambiguation procedures. However, it is possible to plot this spatial data on a world map, which can then be reshaped into a 1D *feature vector*, or a *Map Vector*, the geographic representation of location mentions. MapVec is a novel standardised method for generating geographic features from text documents *beyond* lexical features. This enables a strong geocoding classification performance gain by extracting additional spatial knowledge that would normally be ignored. Geographic semantics cannot be inferred from language alone (too imprecise and incomplete). Word embeddings and distributional vectors use language/words as an *implicit* container of geographic information. Map Vector uses a low-resolution, probabilistic world map as an *explicit* container of geographic information, giving us two types of semantic features from the same text. In related papers on the generation of location representations, [159] inverted the task of geocoding Twitter users to predict word probability from a set of coordinates. A continuous representation of a *region* was generated by using the hidden layer of the neural network. However, all locations in the same region will be assigned an identical vector, which assumes that their semantics are also identical. Another way to obtain geographic representations is by generating embeddings directly from Geonames data using heuristics-driven DeepWalk [152] with geodesic distances [98]. However, to assign a vector, places must first be disambiguated (catch-22). While these generation methods are original and interesting in theory, deploying them in the real-world is infeasible, hence we invented the Map Vector.

MapVec initially begins as a 180x360 world map of geodesic tiles. There are other ways of representing the surface of the Earth such as using nested hierarchies [132] or k-dimensional trees [166], however, this is beyond the scope of this work. The 1x1 tile size, in degrees of geographic coordinates, was empirically determined to be optimal to keep MapVec's size computationally efficient while maintaining meaningful resolution. This map is then populated with the *prior geographic distribution* of each location mentioned in context (see Figure 4.2 for an example). We use population count to *estimate* a location's prior probability as more populous places are more likely to be mentioned in common discourse. For each location mention and for each of its ambiguous candidates, their prior probability is added to the correct tile indicating its geographic position (see Algorithm 1). Tiles that cover areas of open water (64.1%) were removed to reduce size. Finally, this world map is reshaped into a one-dimensional *Map Vector* of length 23,002.

The following features of MapVec are the most salient: *Interpretability:* Word vectors typically need intrinsic [64] and extrinsic tasks [172] to interpret their semantics. MapVec generation is a fully transparent, human readable and modifiable method. *Efficiency:* MapVec is an efficient way of embedding *any number* of locations using the same standardised vector. The alternative means creating, storing, disambiguating and computing with millions of unique location vectors. *Domain Independence:* Word vectors vary depending on the source, time, type and language of the training data and the parameters of generation. MapVec is language-independent and stable over time, domain, size of dataset since the world geography is objectively measured and changes very slowly.

Data and Preprocessing

Training data was generated from geographically annotated Wikipedia pages (dumped February 2017). Each page provided up to 30 training instances, limited to avoid bias from large pages. This resulted in collecting approximately 1.4M training instances, which were uniformly subsampled down to 400K to shorten training cycles as further increases offer diminishing returns. We used the Python-based NLP toolkit Spacy[9] [84] for text preprocessing. All words were lowercased, lemmatised, any stopwords, dates, numbers and so on were replaced with a special token ("0"). Word vectors were initialised with pretrained word embeddings[10] [151]. We do not employ explicit feature selection as in [19], only a minimum frequency count, which was shown to work almost as well as deliberate selection [187].

[9] https://spacy.io/

[10] https://nlp.stanford.edu/

```
Data: Text ← article, paragraph, tweet, etc.
Result: MapVec location(s) representation

Locs ← extractLocations(Text);
MapVec ← new array(length=23,002);
for each l in Locs do
    Cands ← queryCandidatesFromDB($l$);
    maxPop ← maxPopulationOf(Cands);
    for each c in Cands do
        prior ← populationOf($c$) / maxPop;
        i ← coordinatesToIndex($c$);
        MapVec[i] ← MapVec[i] + prior;
    end
end
m ← max(MapVec);
return MapVec / m;
```

Algorithm 1: MapVec generation. For each extracted location ***l*** in *Locs*, estimate the prior probability of each candidate ***c***. Add ***c***'s prior probability to the appropriate array position at index ***i*** representing its geographic position/tile. Finally, normalise the array (to a $[0-1]$ range) by dividing by the maximum value of the MapVec array.

The vocabulary size was limited to the most frequent 331K words, minimum ten occurrences for words and two for location references in the 1.4M training corpus. A final training instance comprises four types of context information: Context Words (excluding location mentions, up to 2x200 words), Location Mentions (excluding context words, up to 2x200 words), Target Entity (up to 15 words) and the MapVec geographic representation of context locations. We have also checked for any overlaps between our Wikipedia-based training data and the WikToR dataset. Those examples were removed. The aforementioned 1.4M Wikipedia training corpus was once again uniformly subsampled to generate a *disjoint* development set of 400K instances. While developing our models mainly on this data, we also used small subsets of LGL (18%), GeoVirus (26%) and WikToR (9%) described in Section 4.2.4 to verify that development set improvements generalised to target domains.

4.2.4 Evaluation Metrics

Our evaluation compares the geocoding performance of six featured systems, our geocoder (CamCoder) and the population baseline. Among these, our CNN-based model is the only neural approach. We have included all open-source/free geocoders *in working order* we were

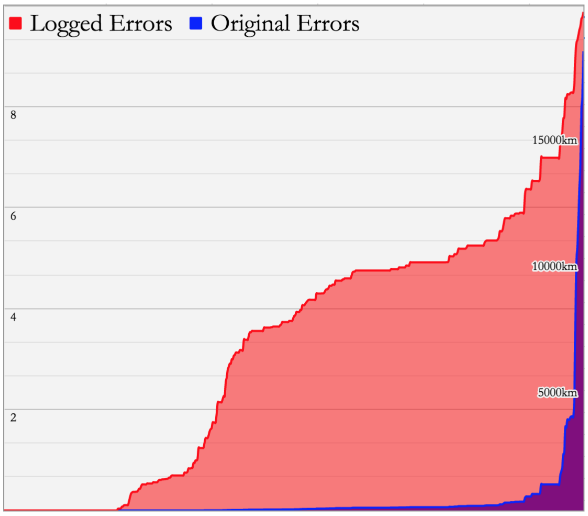


Fig. 4.3 The AUC (range $[0-1]$) is calculated using the Trapezoidal Rule. Smaller errors mean a smaller (blue) area, which means a lower score and therefore better geocoding results.

able to find and they are the most up-to-date versions. Tables 4.1 and 4.2 feature several machine learning algorithms including Long-Short Term Memory (LSTM) [81] to reproduce context2vec [131], Naive Bayes [202] and Random Forest [22] using three diverse datasets.

Geocoding Metrics

We use the three standard and comprehensive metrics, each measuring an important aspect of geocoding, giving an accurate, holistic evaluation of performance. A more detailed cost-benefit analysis of geocoding metrics is available in [94] and [72]. (1) *Average (Mean) Error* is the sum of all geocoding errors per dataset divided by the number of errors. It is an informative metric as it also indicates the total error but treats all errors as equivalent and is sensitive to outliers; (2) *Accuracy@161km* is the percentage of errors that are smaller than 161km (100 miles). While it is easy to interpret, giving fast and intuitive understanding of geocoding performance in percentage terms, it ignores all errors greater than 161km; (3) *Area Under the Curve* (*AUC*) is a comprehensive metric, initially introduced for geocoding in [92]. AUC reduces the importance of large errors (1,000km+) since accuracy on successfully resolved places is more desirable. While it is not an intuitive metric, AUC is robust to outliers and measures *all* errors. A versatile geocoder should be able to maximise all three metrics.

Evaluation Datasets

News Corpus: The Local Global Lexicon (LGL) by [112] contains 588 news articles (4460 test instances), which were collected from geographically distributed newspaper sites. This is the most frequently used geocoding evaluation dataset to date. The toponyms are mostly smaller places no larger than a US state. Approximately 16% of locations in the corpus do not have any coordinates assigned; hence, we do not use those in the evaluation, which is also how the previous figures were obtained. *Wikipedia Corpus*: This corpus was deliberately designed for ambiguity hence the population heuristic is not effective. Wikipedia Toponym Retrieval (WikToR) by [72] is a programmatically created corpus and although not necessarily representative of the real world distribution, it is a test of ambiguity for geocoders. It is also a large corpus (25,000+ examples) containing the first few paragraphs of 5,000 Wikipedia pages. High quality, free and open datasets are not readily available (GeoVirus tries to address this). The following corpora could not be included: WoTR [42] due to limited coverage (southern US) and domain type (historical language, the 1860s), [40] contains fewer than 180 locations, GeoCorpora [190] could not be retrieved in full due to deleted Twitter users/tweets, GeoText [54] only allows for user geocoding, SpatialML [119] involves prohibitive costs, GeoSemCor [27] was annotated with WordNet senses (rather than coordinates).

4.2.5 The GeoVirus Dataset

We now introduce GeoVirus, an open-source test dataset for the evaluation of geoparsing of news events covering global disease outbreaks and epidemics. It was constructed from free WikiNews[11] and collected during 08/2017 - 09/2017. The dataset is suitable for the evaluation of Geotagging/Named Entity Recognition and Geocoding/Toponym Resolution. Articles were identified using the WikiNews search box and keywords such as Ebola, Bird Flu, Swine Flu, AIDS, Mad Cow Disease, West Nile Disease, etc. Off-topic articles were not included. Buildings, POIs, street names and rivers were not annotated.

Annotation (1) The WikiNews contributor(s) who wrote the article annotated most, but not all location references. The first author checked those annotations and identified further references, then proceeded to extract the place name, indices of the start and end characters in text, assigned coordinates and the Wikipedia page URL for each location. (2) A second pass over the entire dataset by the first author to check and/or remedy annotations. (3) A computer program checked that locations were tagged correctly, checking coordinates against the Geonames Database, URL correctness, eliminating any duplicates and validating XML formatting. Places without a Wikipedia page (0.6%) were assigned Geonames coordinates. (4) The second author annotated a random 10% sample to obtain an Inter-Annotator Agreement, which was 100% for geocoding and an F-Score of 92.3 for geotagging. *GeoVirus in Numbers:*

[11] https://en.wikinews.org

Geocoder	**Area Under Curve**†			**Average Error**‡			**Accuracy@161km**		
	LGL	**WIK**	**GEO**	**LGL**	**WIK**	**GEO**	**LGL**	**WIK**	**GEO**
CamCoder	**22 (18)**	**33 (37)**	**31 (32)**	7 **(5)**	**11 (9)**	**3 (3)**	**76 (83)**	**65 (57)**	**82 (80)**
Edinburgh	25 (22)	53 (58)	33 (34)	8 (8)	31 (30)	5 (4)	**76** (80)	42 (36)	78 (78)
Yahoo!	34 (35)	44 (53)	40 (44)	**6 (5)**	23 (25)	**3 (3)**	72 (75)	52 (39)	70 (65)
Population	27 (22)	68 (71)	32 **(32)**	12 (10)	45 (42)	5 **(3)**	70 (79)	22 (14)	80 **(80)**
CLAVIN	26 (20)	70 (69)	32 (33)	13 (9)	43 (39)	6 (5)	71 (80)	16 (16)	79 **(80)**
GeoTxt	29 (21)	70 (71)	33 (34)	14 (9)	47 (45)	6 (5)	68 (80)	18 (14)	79 (79)
Topocluster	38 (36)	63 (66)	NA	12 (8)	38 (35)	NA	63 (71)	26 (20)	NA
Santos et al.	NA	NA	NA	8	NA	NA	71	NA	NA

Table 4.1 Results on LGL, WikToR (WIK) and GeoVirus (GEO). Lower AUC and Average Error are better, higher Acc@161km is better. Figures in *brackets* are scores on identical subsets of each dataset. †Only decimal part shown. ‡Rounded up to nearest 100km.

Annotated locations: 2,167, Unique: 685, Continents: 94, Number of articles: 229, Most frequent places (21% of total): US, Canada, China, California, UK, Mexico, Kenya, Africa, Australia, Indonesia; Mean location occurrence: 3.2, Total word count: 63,205.

4.2.6 New SOTA Results

All tested models (except CamCoder) operate as end-to-end systems; therefore, it is not possible to perform geocoding separately. Each system geoparses its particular majority of the dataset to obtain a representative data sample, shown in Table 4.1 as strongly correlated scores for subsets of different sizes, with which to assess model performance. Table 4.1 also shows scores in *brackets* for the overlapping partition of all systems in order to compare performance on identical instances: GeoVirus 601 (26%), LGL 787 (17%) and WikToR 2,202 (9%). The geocoding difficulty based on the ambiguity of each dataset is: LGL (moderate to hard), WIK (very hard), GEO (easy to moderate).

A population baseline also features in the evaluation. The baseline is conceptually simple: choose the candidate with the highest population, akin to the most frequent word sense in WSD. Table 4.1 shows the effectiveness of this heuristic, which is competitive with many geocoders, even *outperforming* some. However, the baseline is not effective on WikToR as the dataset was deliberately constructed as a tough ambiguity test. Table 4.1 shows how several geocoders mirror the behaviour of the population baseline. This effective heuristic is rarely used in system comparisons, and where evaluated [170, 105], it is inconsis-

Geocoder	**System configuration**			**Dataset**			**Average**
	Language Features	+	**MapVec Features**	**LGL**	**WIK**	**GEO**	
CamCoder	CNN		MLP	**0.22**	**0.33**	**0.31**	**0.29**
Lexical Only	CNN		–	0.23	0.39	0.33	0.32
MapVec Only	–		MLP	0.25	0.41	0.32	0.33
Context2vec†	LSTM		MLP	0.24	0.38	0.33	0.32
Context2vec	LSTM		–	0.27	0.47	0.39	0.38
Random Forest	MapVec features only, no lexical input			0.26	0.36	0.33	0.32
Naive Bayes	MapVec features only, no lexical input			0.28	0.56	0.36	0.40
Population	–		–	0.27	0.68	0.32	0.42

Table 4.2 AUC scores for *CamCoder* and its *Lexical* and *MapVec* components (model ablation). Lower AUC scores are better. †Context2vec model augmented with MapVec.

tent with expected figures (due to unpublished resources, we are unable to investigate further).

We note that no single computational paradigm dominates Table 4.1. The rule-based (Edinburgh, GeoTxt, CLAVIN), statistical (Topocluster), machine learning (CamCoder, Santos) and other (Yahoo!, Population) geocoders occupy different ranks across the three datasets. Due to space constraints, Table 4.1 does not show figures for another type of scenario we tested, a shorter lexical context, using 200 words instead of the standard 400. CamCoder proved to be robust to reduced context, with only a small performance decline. Using the same format as Table 4.1, AUC errors for LGL increased from 22 (18) to 23 (19), WIK from 33 (37) to 37 (40) and GEO remained the same at 31 (32). This means that reducing model input size to save computational resources would still deliver accurate results. Our CNN-based lexical model performs at SOTA levels (Table 4.2) proving the effectiveness of linguistic features while being the outstanding geocoder on the highly ambiguous WikToR.

The Multi-Layer Perceptron (MLP) model using only MapVec with no lexical features is almost as effective but more importantly, it is significantly better than the population baseline (Table 4.2). This is because the Map Vector benefits from wide contextual awareness, encoded in Algorithm 1, while a simple population baseline does not. When we combined the lexical *and* geographic feature spaces in one model (CamCoder[12]), we observed a substantial increase in the SOTA scores. We have also reproduced the *context2vec* model to obtain a continuous context representation using bidirectional LSTMs to encode lexical features, denoted as LSTM[13] in Table 4.2. This enabled us to test the effect of integrating MapVec

[12] Single model settings/parameters for all tests.

[13] https://keras.io/layers/recurrent/

into another deep learning model as opposed to CNNs. Supplemented with MapVec, we observed a significant improvement, demonstrating how enriching various neural models with a geographic vector representation boosts classification results.

Deep learning is the dominant paradigm in our experiments. However, it is important that MapVec is still effective with simpler machine learning algorithms. To that end, we have evaluated it with the Random Forest *without* using any lexical features. This model was well suited to the geocoding task despite training with only half of the 400K training data (due to memory constraints, partial fit is unavailable for batch training in SciKit Learn). Scores were on par with more sophisticated systems. The Naive Bayes was less effective with MapVec though still somewhat viable as a geocoder given the lack of lexical features and a naive algorithm, narrowly beating population. GeoVirus scores remain highly competitive across most geocoders. This is due to the nature of the dataset; locations skewed towards their dominant "senses" simulating ideal geocoding conditions, enabling high accuracy for the population baseline.

GeoVirus alone may not serve as the best scenario to assess a geocoder's performance, however, it is nevertheless important and valuable to determine behaviour in a standard environment. For example, GeoVirus helped us diagnose Yahoo! Placemaker's lower accuracy in what should be an easy test for a geocoder. The figures show that while the average error is low, the accuracy@161km is noticeably lower than most systems. When coupled with other complementary datasets such as LGL and WikToR, it facilitates a comprehensive assessment of geocoding behaviour in many types of scenarios, exposing potential domain dependence. We note that GeoVirus has a dual function, NER (not evaluated but useful for future work) and Geocoding. We made all of our resources freely available[14] for full reproducibility [66].

4.2.7 Discussion and Error Analysis

The Pearson correlation coefficient of the target entity *ambiguity* and the *error size* was only $r \approx 0.2$ suggesting that CamCoder's geocoding errors do not simply rise with location ambiguity. Errors were also not correlated ($r \approx 0.0$) with population size with all types of locations geocoded to various degrees of accuracy. All error curves follow a power law distribution with between 89% and 96% of errors less than 1500km, the rest rapidly increasing into thousands of kilometers. Errors also appear to be uniformly geographically distributed

[14] https://github.com/milangritta/

across the world. The strong lexical component shown in Table 4.2 is reflected by the lack of a relationship between *error size* and the *number of locations* found in the context. The *number of total words* in context is also independent of geocoding accuracy. This suggests that CamCoder learns strong linguistic cues beyond simple association of place names with the target entity and is able to cope with flexible-sized contexts. A CNN Geocoder would expect to perform well for the following reasons: Our context window is 400 words rather than 10-40 words as in previous approaches. The model learns 1,000 feature maps per input and per feature type, tracking 5,000 different word patterns (unigrams, bigrams and trigrams), a significant text processing capability. The lexical model also takes advantage of our own 50-dimensional word embeddings, tuned on geographic Wikipedia pages only, allowing for greater generalisation than bag-of-unigrams models; and the large training/development datasets (400K each), optimising geocoding over a diverse global set of places allowing our model to generalise to unseen instances.

MapVec generation is sensitive to NER performance with higher F-Scores leading to better quality of the geographic vector representation(s). Precision errors can introduce noise while recall errors may withhold important locations. The average F-Score for the featured geoparsers is $F \approx 0.7$ (standard deviation ≈ 0.1). Spacy's NER performance over the three datasets is also $F \approx 0.7$ with a similar variation between datasets. In order to further interpret scores in Tables 4.1 and 4.2, with respect to maximising geocoding performance, we briefly discuss the *Oracle* score. Oracle is the geocoding performance upper bound given by the Geonames data, i.e. the *highest possible score(s)* using Geonames coordinates as the geocoding output. In other words, it quantifies the *minimum error* for each dataset given the *perfect* location disambiguation. This means it quantifies the difference between "gold standard" coordinates and the coordinates in the Geonames database. The following are the Oracle scores for LGL (*AUC=0.04, a@161km=99*) annotated with Geonames, WikToR (*AUC=0.14, a@161km=92*) and GeoVirus (*AUC=0.27, a@161km=88*), which are annotated with Wikipedia data. Subtracting the Oracle score from a geocoder's score quantifies the scope of its *theoretical future improvement*, given a particular database/gazetteer.

4.2.8 Conclusions

Geocoding methods commonly employ lexical features, which have proved to be very effective. Our lexical model was the best language-only geocoder in extensive tests. It is possible, however, to go *beyond* lexical semantics. Locations also have a rich topological meaning, which has not yet been successfully isolated and deployed. We need a means of extracting and encoding this additional knowledge. To that end, we introduced MapVec, an algorithm and a container for encoding context locations in geodesic vector space. We showed how CamCoder, using lexical *and* MapVec features, outperformed both approaches, achieving a new SOTA. MapVec remains effective with various machine learning frameworks (Random Forest, CNN and MLP) and substantially improves accuracy when combined with other neural models (LSTMs). Finally, we introduced GeoVirus, an open-source dataset that helps facilitate geoparsing evaluation across more diverse domains with different lexical-geographic distributions [58, 51]. Tasks that could benefit from our methods include social media placing tasks [32], inferring user location on Twitter [206], geolocation of images based on descriptions [173] and detecting/analyzing incidents from social media [14]. Future work may see our methods applied to *document* geolocation to assess the effectiveness of scaling geodesic vectors from paragraphs to entire documents.

4.3 Summary

Perhaps the most original research idea of this chapter is the Map Vector, a geographic representation of context that can augment a lexical-only model with previously unused features. The representation generation community is a popular research area with a plethora of novel discoveries and tools introduced in the last decade. However, for NLP tasks that require a representation of *several geographic referents* for multiple toponyms in one vector, word embeddings do not capture this geographic distribution as all context information is collapsed into a single distributed representation. The Map Vector is able to capture this explicit distribution thus can be thought of as domain-specific context vector representation. Its addition has been shown to significantly increase performance of toponym resolution.

The combined model managed to set SOTA scores using a CNN-based neural architecture augmented with the Map Vector. CamCoder is quite large in terms of the number of parameters trained and the length of input but can still be used on a good laptop (GPU is recommended for 10,000+ instances). Future work on the architecture may investigate how much performance would be lost by scaling down the size of the model below the levels of our ablation tests. Another promising future experiment may be to use population *rank* instead of population *count* during MapVec generation. This would serve to smooth out the prior probabilities used in the generation process, removing outliers, likely increasing the recall of smaller locations but at a cost to the accuracy of the more populous places.

Despite leading geocoding figures on multiple datasets, we note there is good scope for higher accuracy looking at the *Oracle* scores, particularly for LGL and WikToR. The additional difficulty of the task stems from the source domain shift and the high positive skewness of the toponym referent distribution; 70-90% of the toponym mentions are the most populous places with the rest of the referent probabilities decreasing sharply along the long right tail. In the face of these challenges, the CamCoder's flexibility comes to the fore given it was trained on Wikipedia data and successfully tested on news text. Given enough training data, the CNN model is very capable across similar domains and even performs at 98% accuracy on an unseen Wikipedia *development set* although this statistic was not included in the paper. Finally, we are pleased that our code and data seem to be useful to researchers. According to GitHub Analytics, it's receiving several git clones and other engagements per month. We hope this will encourage new participants to become involved in related research and that using our resources may become a springboard that will accelerate their progress.

Chapter 5

What's Next for Geoparsing?

5.1 Introduction

This chapter was not originally included in the thesis plan, however, it has emerged as a natural and necessary follow-up to the previous chapters. The genesis of this chapter occurred during the peer review of the last ACL paper on Toponym Resolution. One of the reviewers posed a legitimate question (paraphrased): If the paper claims SOTA scores, where is the *standard evaluation framework* that determines the SOTA? This prompted a long reply during the ACL author response period. It also inspired this chapter along with a manuscript that analyses, reviews, clarifies and expedites evaluation in geoparsing. This would significantly extend the work from Chapter 2 and serve as an efficient and informative entry point for new researchers to the field and pull together the best practices in geoparsing evaluation. The need for such contribution was reinforced by the discoveries made during the extensive data annotation in our previous work. We learnt that there exist several different types of toponyms, each with slightly different characteristics. In NER and Geoparsing, toponyms have previously been treated as semantically equivalent, however, we noticed this was not the case and decided to annotate and quantify these linguistic occurrences. This means a new data resource was created via corpus linguistic analysis, our fourth and final dataset (highest quality to date). The raw data came from the European Media Monitor's MediSys feed, which would then be used to analyse the performance of the geoparsing part of the surveillance system in the next chapter. In this chapter, we show that toponyms have a lot more depth to them than previously assumed. The paper was submitted to (and will be published by) Springer's *Language Resources and Evaluation* journal and can be previewed on arxiv.org[1] and the resources downloaded from GitHub[2].

[1] https://arxiv.org/abs/1810.12368

[2] https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation

5.2 A Pragmatic Guide to Geoparsing Evaluation

Empirical methods in geoparsing have thus far lacked a standard evaluation framework describing the task, metrics and data used to compare state-of-the-art systems. Evaluation is further made inconsistent, even unrepresentative of real world usage by the lack of distinction between the *different types of toponyms*, which necessitates new guidelines, a consolidation of metrics and a detailed toponym taxonomy with implications for Named Entity Recognition (NER) and beyond. To address these deficiencies, our manuscript introduces a new framework in three parts. Part 1) Task Definition: clarified via corpus linguistic analysis proposing a fine-grained *Pragmatic Taxonomy of Toponyms*. Part 2) Metrics: discussed and reviewed for a rigorous evaluation including recommendations for NER/Geoparsing practitioners. Part 3) Evaluation Data: shared via a new dataset called *GeoWebNews* to provide test/train examples and enable immediate use of our contributions. In addition to fine-grained Geotagging and Toponym Resolution (Geocoding), this dataset is also suitable for prototyping and evaluating machine learning NLP models.

5.2.1 Current Challenges

Geoparsing aims to translate toponyms in free text into geographic coordinates. Toponyms are weakly defined as "place names", however, we will clarify and extend this underspecified definition in Section 5.3. Illustrating with an example headline, "*Springfield robber escapes from Waldo County Jail. Maine police have launched an investigation.*", the geoparsing pipeline is (1) Toponym extraction [*Springfield, Waldo County Jail, Maine*], this step is called *Geotagging* and is a special case of NER; and (2) Disambiguating and linking toponyms to geographic coordinates [(45.39, -68.13), (44.42, -69.01), (45.50, -69.24)], this step is called *Toponym Resolution* (also *Geocoding*). Geoparsing is an essential constituent of many Geographic Information Retrieval (GIR), Extraction (GIE) and Analysis (GIA) tasks such as determining a document's geographic scope [178], Twitter-based disaster response [39] and mapping [9], spatio-temporal analysis of tropical research literature [149], business news analysis [1], disease detection and monitoring [4] as well as analysis of historical events such as the Irish potato famine [183]. Geoparsing can be evaluated in a highly rigorous manner, enabling a robust comparison of state-of-the-art (SOTA) methods. This manuscript provides the end-to-end *Pragmatic Guide to Geoparsing Evaluation* for that purpose. End-to-end means to (1) critically review and extend the definition of toponyms, i.e. *what* is to be evaluated and *why* it is important; (2) review, recommend and create high-quality open resources to expedite research; and (3) outline, review and consolidate metrics for each stage of the geoparsing pipeline, i.e. *how* to evaluate.

Due to the essential NER component in geoparsing systems [170, 41, 95, 72, 92], our investigation and proposals have a strong focus on NER's *location extraction* capability. We demonstrate that off-the-shelf NER taggers are inadequate for location extraction due to the lack of ability to extract and classify the pragmatic types of toponyms (Table 5.1). In an attempt to assign coordinates to an example sentence, "*A French bulldog bit an Australian tourist in a Spanish resort.*", current NER tools fail to differentiate between the literal and associative uses of these adjectival toponyms[3]. A more detailed example analysed in Table 5.2 and a survey of previous work in Section 5.2.4 show that the definition and handling of toponyms is inconsistent and unfit for advanced geographic NLP research. In fact, beyond a limited "place name" definition, a deep pragmatic/contextual toponym semantics has not yet been defined in Information Extraction, to our best knowledge. This underspecification results in erroneous and unrepresentative real-world extraction/classification of toponyms incurring both *precision errors* and *recall errors*. To that end, we propose a *Pragmatic Taxonomy of Toponyms* required for a rigorous geoparsing evaluation, which includes the recommended datasets and metrics.

Why a Pragmatic Guide Pragmatics [158] is the linguistic theory of generative approach to word meaning, i.e. how context contributes to and changes the semantics of words and phrases. This is the first time, to our best knowledge, that the definition of fine-grained toponym types has been quantified in such detail using a representative sample of general topic, globally distributed news articles. We also release a new *GeoWebNews* dataset to challenge researchers to develop Machine Learning (ML) algorithms to evaluate classification/tagging performance based on deep pragmatics rather than shallow syntactic features. Section 2.2.2 gives a background on Geoparsing, NER, GIE and GIR. We present the new taxonomy in Section 5.3, describing and categorising toponym types. In Section 5.4, we conduct a comprehensive review of current evaluation methods and justify the recommended framework. Finally, Section 5.5 introduces the GeoWebNews dataset, annotation and resources. We also evaluate geotagging and toponym resolution on the new dataset, illustrating the performance of several sequence tagging models such as SpacyNLP and Google NLP.

[3] Throughout the paper, we use the term *Literal* to denote a toponym and/or its context that refers directly to the *physical* location and the term *Associative* for a toponym and/or its context that is only *associated* with a place. Full details in Section 5.3.

5.2.2 Summary of the most salient findings

Toponym semantics have been underspecified in NLP literature. Toponyms can refer to physical places as well as entities associated with a place as we outline in our proposed taxonomy. Their distribution in a sample of 200 news articles is 53% literal and 47% associative. Until now, this type of fine-grained toponym analysis was not conducted. We provide a dataset annotated by linguists (including computational) enabling immediate evaluation of our proposals. GeoWebNews.xml can be used to evaluate Geotagging, NER, Toponym Resolution and to develop ML models from limited training data. A total of 2,720 toponyms were annotated with Geonames[4]. Data augmentation was evaluated with an extra 3,460 annotations although effective implementation remains challenging. We also found that popular NER taggers appear not to use contextual information, relying instead on the entity's primary word sense (see Table 5.2). We show that this issue can be addressed by training an effective geotagger from limited training data (F-Score=88.6), outperforming Google Cloud NLP (F-Score=83.2) and Spacy NLP (F-Score=74.9). In addition, effective 2-class (Literal versus Associative toponyms) geotagging is also feasible (F-Score=77.6). The best toponym resolution scores for GeoWebNews were 95% accuracy@161km, AUC of 0.06 and a Mean Error of 188km. Finally, we provide a critical review of available metrics and important nuances of evaluation such as database choice, system scope, data domain/distribution, statistical testing, etc. All recommended resources are available on GitHub[5].

5.2.3 Related Work

Before we critically review *how* to rigorously evaluate geoparsing and introduce a new dataset, we first need to clarify *what* is to be evaluated and *why*. We focus on the pragmatics of toponyms for fine-grained geoparsing of events described in text. This requires differentiating literal from associative types as well as increasing toponym recall by including entities ignored by current models. When a word spells like a place, i.e. shares its orthographic form, this does not mean it *is* a place or has equivalent meaning, for example: "*Paris* (a person) said that *Parisian* (associative toponym) artists don't have to live in *Paris* (literal toponym)." and "*Iceland* (a UK supermarket) doesn't sell *Icelandic* (associative toponym) food, it's not even the country of *Iceland* (literal toponym)." In order to advance research in toponym extraction and other associated NLP tasks, we need to move away from the current practice of seemingly ignoring the context of a toponym, relying on the entity's dominant word sense

[4] https://www.geonames.org/

[5] https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation

and morphological features, treating toponyms as semantically equivalent. The consequences of this simplification are disagreements and incompatibilities in toponym evaluation leading to unrepresentative real-world performance. It is difficult to speculate about the reason for this underspecification, whether it is the lack of available quality training data leading to lower traction in the NLP community or the satisfaction with a simplified approach. However, we aim to encourage active research and discussions through our contributions.

5.2.4 Geographic datasets and the pragmatics of toponyms

Previous work in annotation of geographic NLP datasets constitutes our primary source of enquiry into recent research practices, especially the lack of linguistic definition of toponym types. An early specification of an Extended Named Entity Hierarchy [171] was based only on *geographic feature types*[6] i.e. address, country, region, water feature, etc. Geoparsing and NER require a deeper contextual perspective based on how toponyms are used in practice by journalists, writers or social media users, something a static database lookup cannot determine. CoNLL 2002 [169] and 2003 [185] similarly offer no semantic definition of a toponym beyond what is naively thought of as a location, i.e. an entity spelled like a place and a location as its primary word sense. Schemes such as ACE [47] bypass toponym type distinction, classifying entities such as governments via a simplification to a single tag *GPE: A Geo-Political Entity*. Modern NER parsers such as Spacy [84] use similar schemes [193] to collapse different taxonomic types into a single tag avoiding the need for a deeper understanding of context. A simplified tag set (LOC, ORG, PER, MISC) based on Wikipedia [147] is used by NER taggers such as Illinois NER [164] and Stanford NLP [121], featured in Table 5.2. The table shows the limited classification indicating weak and inconsistent usage of context.

The SpatialML [119] scheme is focused on spatial reasoning e.g. *X location north of Y*. Metonymy [124], which is a substitution of a related entity for a concept originally meant, was acknowledged but not annotated due to the lack of training of Amazon Mechanical Turk annotators. Facilities were always tagged in the SpatialML corpus *regardless of the context* in which they're being used. The corpus is available at a cost of $500-$1,000. The Message Understanding Conferences (MUC) [80] have historically not tagged adjectival forms of locations such as *"American* exporters". We assert that there is no difference between that and *"U.S.* exporters", which would almost certainly be annotated. The Location Referring Expression corpus [130] has annotated toponyms including locational expressions such as

[6] https://nlp.cs.nyu.edu/ene/version7_1_0Beng.html

parks, buildings, bus stops and facilities in 10,000 Japanese tweets. Systematic polysemy [5] has been taken into account for *facilities*, but not extended to other toponyms. GeoCLEF [65] (Geographic Cross Language Evaluation Forum) focused on Multilingual GIR evaluation. Geoparsing specifically, i.e. Information Extraction was not investigated. Toponym types were not linguistically differentiated despite the multi-year project's scale. This conclusion also applies to Spatial Information Retrieval and Geographical Ontologies [89] (called SPIRIT) project, the focus of which was not the evaluation of Information Extraction or Toponym Semantics but classical GIR.

The WoTR corpus [43] of historical US documents also did not define toponyms. However, browsing the dataset, expressions such as "Widow Harrow's house" and "British territory" were annotated. In Section 5.3, we shall claim this is beyond the scope of toponyms, i.e. "house" and "territory" should not be tagged. The authors do acknowledge, but *do not annotate* metonymy, demonyms and nested entities. Systematic polysemy such as metonymy should be differentiated during toponym extraction and classification, something acknowledged as a problem more than ten years ago [107]. Section 5.3 elaborates on the taxonomy of toponyms beyond metonymic cases. Geocorpora [191] is a Twitter-based geoparsing corpus with around 6,000 toponyms with buildings and facilities annotated. The authors acknowledge that toponyms are frequently used in a metonymic manner, however, these cases have not been annotated after browsing the open dataset. Adjectival toponyms have also been *left out*. We show that these constitute around 13% of all toponyms thus should be included to boost recall.

The LGL corpus [112] loosely defines toponyms as "spatial data specified using text". The evaluation of an accompanying model focused on toponym resolution. Authors agree that standard Named Entity Recognition is inadequate for geographic NLP tasks. It is often the case that papers emphasise the geographic ambiguity of toponyms but not their semantic ambiguity. The CLUST dataset [110] by the same author, describes toponyms simply as "textual references to geographic locations". Homonyms are discussed as is the low recall and related issues of NER taggers, which makes them unsuitable for achieving high geotagging fidelity. Metonymy was not annotated, some adjectival toponyms have been tagged though sparsely and inconsistently. There is no distinction between literal and associative toponyms. Demonyms were tagged but with no special annotation hence treated as ordinary locations with no descriptive statistics offered. TR-News [93] is a quality geoparsing corpus despite the paucity of annotation details or IAA figures in the paper. A brief analysis of the open dataset showed that embedded toponyms, facilities and adjectival toponyms were annotated, which substantially increases recall, although no special tags were used hence unable to gather descriptive statistics. Homonyms, coercion, metonymy, demonyms and languages were not annotated and nor was the distinction between literal, mixed and associative toponyms. With that, we still recommended it as a suitable resource for geoparsing in the latter sections.

PhD Theses are themselves comprehensive collections of a large body of relevant research and therefore important sources of prior work. Despite this not being the convention in NLP publishing, we outline the prominent PhD theses from the past 10+ years to show that toponym types have not been organised into a pragmatic taxonomy and that evaluation metrics in geocoding are in need of review and consolidation. We also cite their methods and contributions as additional background for discussions throughout the paper. The earliest comprehensive research on toponym resolution originated in (Leidner, 2008) [105]. Toponyms were specified as “names of places as found in a text". The work recognised the ambiguity of toponyms in different contexts and was often cited by later research papers though until now, these linguistic regularities have not been formally and methodically studied, counted, organised and released as high fidelity open resources. A geographic mining thesis (Martins, 2008) [38] defined toponyms as “geographic names" or “place names". It mentions homonyms, which are handled with personal name exclusion lists rather than learned by contextual understanding. A Wikipedia GIR thesis (Overell, 2009) [148] has no definition of toponyms and limits the analysis to *nouns only*. The GIR thesis (Andogah, 2010) [7] discusses the geographic hierarchy of toponyms as found *in gazetteers*, i.e. feature types instead of linguistic types. A toponym resolution thesis (Buscaldi, 2010) [26] describes toponyms as “place names", once again mentions metonymy without handling these cases citing lack of resources, which our work provides.

The Twitter geolocation thesis (Han, 2014) [74] provides no toponym taxonomy, nor does the Named Entity Linking thesis (Santos, 2013) [50]. A GIR thesis (Moncla, 2015) [137] defines a toponym as a spatial named entity, i.e. a location somewhere in the world bearing a proper name, discusses syntactical rules and typography of toponyms but not their semantics. The authors recognise this as an issue in geoparsing but no solution is proposed. The GIA thesis (Ferrés, 2017) [56] acknowledges but doesn’t handle cases of metonymy, homonymy and non-literalness while describing a toponym as “a geographical place name". Recent Masters theses also follow the same pattern such as a toponym resolution thesis (Kolkman, 2015) [99], which says a toponym is a “word of phrase that refers to a location". While none of these definitions are incorrect, they are very much underspecified. Another Toponym Resolution thesis (DeLozier, 2016) [43] acknowledges relevant linguistic phenomena such as metonymy and demonyms, however, no resources, annotation or taxonomy is given. Toponyms were established as “named geographic entities". This background section presented a multitude of research contributions using, manipulating and referencing toponyms, however, without a deep dive into their pragmatics, i.e. what *is* a toponym from a *linguistic point of view* and the practical NLP implications of that. Without an agreement on the *what*, *why* and *how* of geoparsing, the evaluation of SOTA systems cannot be consistent and robust.

5.3 A Pragmatic Taxonomy of Toponyms

While the evaluation metrics, covered in Section 5.4, are relevant only to geoparsing, Sections 5.3 and 5.5 have implications for Core NLP tasks such as NER. In order to introduce the Toponym Taxonomy, we start with a location. *A location* is any of the potentially infinite physical points on Earth identifiable by *coordinates*. With that in mind, *a toponym is any named entity that labels a particular location.* Toponyms are thus a subset of locations as most locations do not have proper names. Further to the definition and extending the work from the Background section, toponyms exhibit various degrees of *literalness* as their *referents* may not be physical locations but other entities as is the case with metonyms, languages, homonyms, demonyms, some embedded toponyms and associative modifiers.

Structurally, toponyms occur within clauses, which are the smallest grammatical units expressing a full proposition. Within clauses, which serve as the context, toponyms are embedded in noun phrases (NP). A toponym can occur as the *head* of the NP, for example “Accident in *Melbourne*.” Toponyms also frequently *modify* NP heads. Modifiers can occur before or after the NP head such as in “President of *Mongolia*" versus “*Mongolian* President" and can have an *adjectival* form “*European* cities" or a *noun* form “*Europe*'s cities". In theory, though not always in practice, the classification of toponym types is driven by (1) *the semantics of the NP*, which is conditional on (2) *the NP context* of the surrounding clause. These types may be classified using a hybrid approach [48], for example. It is this interplay of semantics and context, seen in Table 5.1, that determines the type of the following toponyms (literals=**bold**, associative=*italics*): “The **Singapore** project is sponsored by *Australia*." and “He has shown that in **Europe** and last year in **Kentucky**." and “The soldier was operating in **Manbij** with *Turkish* troops when the bomb exploded." As a result of our corpus linguistic analysis, we propose *two top-level* taxonomic types (a) *literal*: where something is happening or is physically located; and (b) *associative*: a concept that is associated with a toponym (Table 5.1). We also assert that for applied NLP, it is sufficient and feasible to distinguish between literal and associative toponyms.

Toponym Type	NP Semantics Indicates	NP Context Indicates
Literals	Noun Literal Type	Literal Type
Literal Modifiers	Noun/Adjectival Literal	Literal or Associative†
Mixed	Noun/Adjectival Literal	Ambiguous or Mixed
Coercion	Non-Toponym	Literal Type
Embedded Literal	Non-Toponym	Literal Type
Embedded NonLit	Non-Toponym	Associative Type
Metonymy	Noun Literal Type	Associative Type
Languages	Adjectival Literal Type	Associative Type
Demonyms	Adjectival Literal Type	Associative Type
Non-Lit Modifiers	Noun/Adjectival Literal	Associative Type
Homonyms	Noun Literal Type	Associative Type

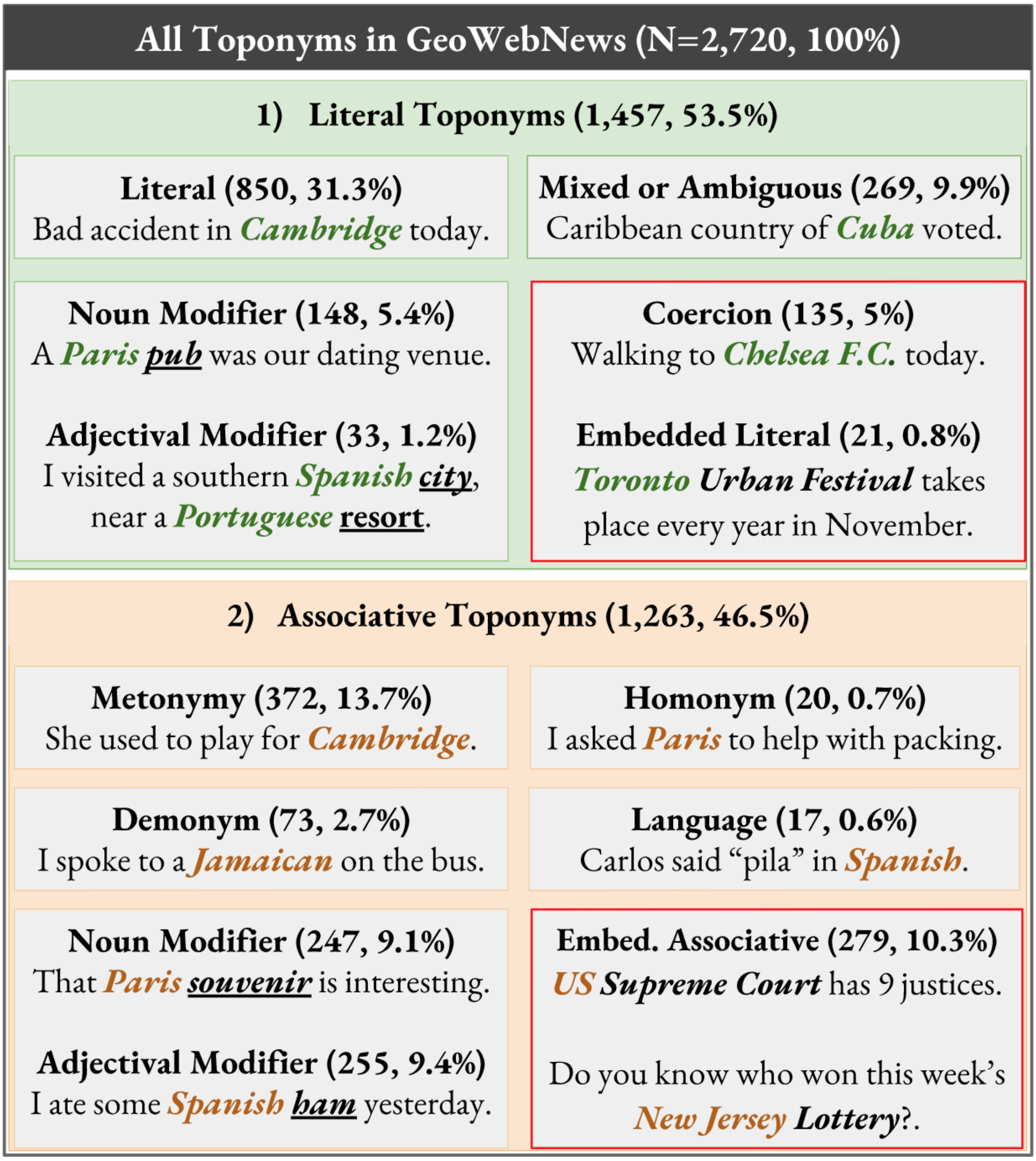
Table 5.1 The interplay between context and semantics determines the type. The top five are the literals, the bottom six are the associative types. Examples of each type can be found in Figure 5.1. †NP **head** must be strongly indicative of a literal type, e.g.: “The *British* **weather** doesn’t seem to like us today."

5.3.1 Non-Toponyms

There is a group of entities that are currently not classified as toponyms, denoted as *Non-Toponyms* in this paper. We shall assert, however, that these are in fact equivalent to “regular" toponyms. We distinguish between three types: a) *Embedded Literals* such as “The *British* Grand Prix" and *“Louisiana* Purchase" b) *Embedded Associative* toponyms, for example *“Toronto* Police" and *“Brighton* City Council” and c) *Coercion*, which is when a polysemous entity has its less dominant word sense *coerced* to the *location* class by the context. Failing to extract Non-Toponyms lowers real-world recall, missing out on valuable geographical data. In our diverse and broadly-sourced dataset, Non-Toponyms constituted a non-trivial 16% of all toponyms.

5.3.2 Literal Toponyms

These types refer to places *where something is happening or is physically located.* This subtle but important distinction from associative toponyms allows for higher quality geographic analysis. For instance, the phrase “*Swedish* people" (who could be anywhere) is *not* the same as “people *in Sweden*" so we differentiate this group from the associative group. Only the latter mention refers to Swedish “soil” and can/should be processed separately.



All Toponyms in GeoWebNews (N=2,720, 100%)

1) Literal Toponyms (1,457, 53.5%)

Literal (850, 31.3%)
Bad accident in *Cambridge* today.

Mixed or Ambiguous (269, 9.9%)
Caribbean country of *Cuba* voted.

Noun Modifier (148, 5.4%)
A *Paris* ***pub*** was our dating venue.

Adjectival Modifier (33, 1.2%)
I visited a southern *Spanish* ***city***, near a *Portuguese* **resort**.

Coercion (135, 5%)
Walking to *Chelsea F.C.* today.

Embedded Literal (21, 0.8%)
Toronto ***Urban Festival*** takes place every year in November.

2) Associative Toponyms (1,263, 46.5%)

Metonymy (372, 13.7%)
She used to play for *Cambridge*.

Homonym (20, 0.7%)
I asked *Paris* to help with packing.

Demonym (73, 2.7%)
I spoke to a *Jamaican* on the bus.

Language (17, 0.6%)
Carlos said "pila" in *Spanish*.

Noun Modifier (247, 9.1%)
That *Paris* ***souvenir*** is interesting.

Adjectival Modifier (255, 9.4%)
I ate some *Spanish* ***ham*** yesterday.

Embed. Associative (279, 10.3%)
US ***Supreme Court*** has 9 justices.

Do you know who won this week's *New Jersey* ***Lottery***?.

Fig. 5.1 The Pragmatic Taxonomy of Toponyms. A red border denotes *Non-Toponyms*. Classification algorithm: If the context indicates a literal or is ambiguous/mixed, then the type is literal. If the context is associative, then (a) for non-modifiers the toponym is associative (b) for modifiers, if the head is mobile and/or abstract, then the toponym is associative, otherwise it is literal.

A Literal is what is most commonly and too narrowly thought of as a location, e.g. "Harvests in *Australia* were very high." and "*South Africa* is baking in 40C degree heat." For these toponyms, the semantics and context both indicate it is a literal toponym, which refers directly to a physical location.

Coercion refers to polysemous entities typically classified as Non-Toponyms, which in a *literal context* have their word sense coerced to (physical) *location*. More formally, coercion is "an observation of grammatical and semantic incongruity, in which a syntactic structure places requirements on the types of lexical items that may appear within it."[208] Examples include "The *University of Sussex, Sir Isaac Newton (pub), High Court* is our meeting place." and "I'm walking to *Chelsea F.C., Bell Labs, Burning Man.*" Extracting these toponyms increases recall and allows for a *very precise location* as these toponyms tend to have a small geographic footprint.

Mixed Toponyms typically occur in an ambiguous context, e.g. "*United States* is generating a lot of pollution." or "*Sudan* is expecting a lot of rain." They can also simultaneously activate a literal *and* an associative meaning, e.g. "The north African country of *Libya* announced the election date." These cases sit somewhere between literal and associative toponyms, however, we propose to include them in the literal group.

Embedded Literals are Non-Toponyms nested within larger entities such as *"Toronto* Urban Festival", *"London* Olympics", *"Monaco* Grand Prix" and are often extracted using a 'greedy algorithm'. They are semantically, though not syntactically, equivalent to Literal Modifiers. If we ignored the case, the meaning of the phrase would not change, e.g. "*Toronto* urban festival".

Noun Modifiers are toponyms that modify *literal heads* (Figure 5.2), e.g. "You will find the UK [*lake, statue, valley, base, airport*] there." and "She was taken to the South Africa [*hospital, border, police station*]". The context, however, needn't always be literal, for instance "An *Adelaide* court sentenced a murderer to 25 years." or "The *Vietnam* office hired 5 extra staff." providing the head is literal. Noun modifiers can also be placed after the head, for instance "We have heard much about the stunning caves of *Croatia*."

Adjectival Modifiers exhibit much the same pattern as noun modifiers except for the *adjectival form* of the toponym, for example, "It's freezing in the *Russian* tundra.", "*British* ports have doubled exports." or "*American* schools are asking for more funding." Adjectival

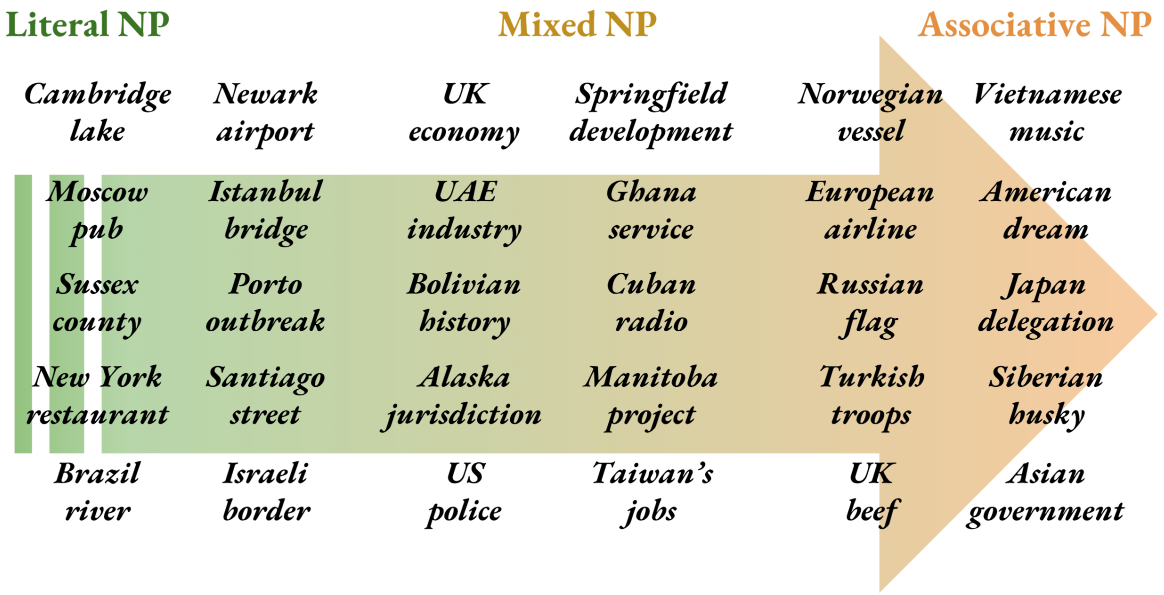


Fig. 5.2 Example noun phrases ranging from Literal to Mixed to Associative. The further to the right, the more 'detached' the *NP referent* becomes from its physical location. Literal heads tend to be *concrete* (elections, accidents) and *static* (buildings, natural features) while associative heads are more *abstract* (promises, partnerships) and *mobile* (animals, products). In any case, *context* is the main indicator of type and needs to be combined with *NP semantics*.

modifiers are frequently and incorrectly tagged as *nationalities or religious/political groups*[7] and sometimes ignored[8] altogether. Approximately 1 out of 10 adjectival modifiers is literal.

5.3.3 Associative Toponyms

Toponyms frequently refer to or are used to modify *non-locational concepts* (NP heads), which are *associated* with locations rather than directly referring to their physical presence. This can occur by substituting a non-locational concept with a toponym (metonymy) or via a demonym, homonym or a language reference. Some of these instances look superficially like modifiers leading to frequent NER errors.

Demonyms [165] are derived from toponyms and denote the inhabitants of a country, region or city. These persons are *associated* with a location and have been on occasion, sparsely rather than exhaustively, annotated [112]. Examples include "I think he's *Indian*.", which is equivalent to "I think he's an *Indian* citizen/person." or "An *American* and a *Briton* walk into a bar ..."

[7] https://spacy.io/usage/ and http://corenlp.run/
[8] http://services.gate.ac.uk/annie/ and IBM NLP Cloud in Table 5.2.

Languages can sometimes be confused for adjectival toponyms, e.g. “How do you say pragmatics in *French, Spanish, English, Japanese, Chinese, Polish?*” Occurrences of languages should not be interpreted as modifiers, another NER error stemming from a lack of contextual understanding. This is another case of a concept associated with a location that should not require coordinates.

Metonymy is a figure of speech whereby a concept that was originally intended gets *substituted* with a *related* concept, for example “*Madrid* plays *Kiev* today.", substituting sports teams with toponyms. Similarly, in “*Mexico* changed the law.”, the likely latent entity is the *Mexican government*. Metonymy was previously found to be a frequent phenomenon, around 15-20% of place mentions are metonymic [125, 71, 107]. In our dataset, it was 13.7%.

Noun Modifiers are toponyms that modify associative noun phrase heads in an associative context, for instance “*China* exports slowed by 7 percent.” or “*Kenya’s* athletes win double gold.” Noun modifiers also occur after the head as in “The President *of Armenia* visited the Embassy of *the Republic of Armenia* to the *Vatican*.". Note that the event did *not* take place in Armenia but the Vatican, potentially identifying the wrong event location.

Adjectival Modifiers are sporadically covered by NER taggers (Table 5.2) or tagging schemes [80]. They are semantically identical to associative noun modifiers except for their adjectival form, e.g. “*Spanish* sausages sales top €2M.”, “We’re supporting the *Catalan* club." and “*British* voters undecided ahead of the Brexit referendum."

Embedded Associative toponyms are Non-Toponyms nested within larger entities such as *“US* Supreme Court", *“Sydney* Lottery" and *“Los Angeles* Times". They are semantically, though not syntactically, equivalent to Associative Modifiers. Ignoring case would not change the meaning of the phrase *“Nigerian* Army" versus *“Nigerian* army". However, it *will* wrongly change the shallow classification from ORG to LOC for most NER taggers.

Homonyms and more specifically *homographs*, are words with identical spelling but different meaning such as *Iceland* (a UK grocery chain). Their meaning is determined mainly by contextual evidence [77, 67] as is the case with other types. Examples include: “*Brooklyn* sat next to *Paris*." and “*Madison, Chelsea, Clinton, Victoria, Jamison and Norbury* submitted a Springer paper."

TOPONYM (TYPE)	**LABEL**	**GOOG.**	**SPACY**	**STANF.**	**ANNIE**	**ILLIN.**	**IBM**
Milan (Homomymy)	**Assoc.**	**Literal**	**Literal**	**Literal**	**Literal**	**Organ.**	**Literal**
Lebanese (Language)	**Assoc.**	**Literal**	**Demon.**	**Demon.**	−−	**Misc.**	−−
Syrian (Demonym)	**Assoc.**	**Literal**	**Demon.**	**Demon.**	−−	**Misc.**	−−
UK origin (NounMod)	**Assoc.**	**Assoc.**†	**Literal**	**Literal**	**Literal**	**Literal**	**Literal**
K.of **Jordan** (PostMod)	**Assoc.**	**Person**	**Literal**	**Literal**	**Literal**	**Organ.**	**Person**
Iraqi militia (AdjMod)	**Assoc.**	**Assoc.**†	**Demon.**	**Demon.**	−−	**Misc.**	−−
US Congress (Embed)	**Assoc.**	**Organ.**	**Organ.**	**Organ.**	**Organ.**	**Organ.**	**Organ.**
Turkey (Metonymy)	**Assoc.**	**Literal**	**Literal**	**Literal**	**Literal**	**Literal**	−−
city in **Syria** (Literal)	**Literal**	**Literal**	**Literal**	**Literal**	**Literal**	**Literal**	**Literal**
Iraqi border (AdjMod)	**Literal**	**Literal**	**Demon.**	**Demon.**	−−	**Misc.**	−−
Min.of Defense (Fac)	**Literal**	**Organ.**	**Organ.**	**Organ.**	**Organ.**	**Organ.**	**Organ.**

Table 5.2 Popular NER taggers tested in June 2018 using official demo interfaces (incorrect labels underlined) on the sentence: *"**Milan**, who was speaking **Lebanese** with a **Syrian** of **UK** origin as well as the King of **Jordan**, reports that the **Iraqi** militia and the **US** Congress confirmed that **Turkey** has shelled a city in **Syria**, right on the **Iraqi** border near the **Ministry of Defense**."* A distinction is made only between *a location* and *not-a-location* since an *associative label* is unavailable. The table shows only a weak agreement between tagging schemes. †Can be derived from the API with a simple rule.

5.4 Standard Evaluation Metrics

The previous section established *what* is to be evaluated and *why* it is important. In this part, we focus on critically reviewing existing geoparsing metrics, i.e. *how* to assess geoparsing models. In order to reliably determine the SOTA and estimate the practical usefulness of these models in downstream applications, we propose a holistic, consistent and rigorous evaluation framework. Considering the task objective and available metrics, the recommended approach is to evaluate geoparsing as separate components. Researchers and practitioners do not typically tackle both stages at once [41, 186, 95, 195, 196, 70]. More importantly, it is difficult to diagnose errors and target improvements without this separation. The best practice is to evaluate geotagging first, then obtain geocoding metrics for the true positives, i.e. the subset of correctly identified toponyms. We recommend evaluating with a minimum of 50% of geotagged toponyms for a representative geocoding sample. Finally, population has not consistently featured in geocoding evaluation but it is capable of beating many existing systems [41, 72]. Therefore, we recommend the usage of this *strong baseline* as a *necessary component* of evaluation.

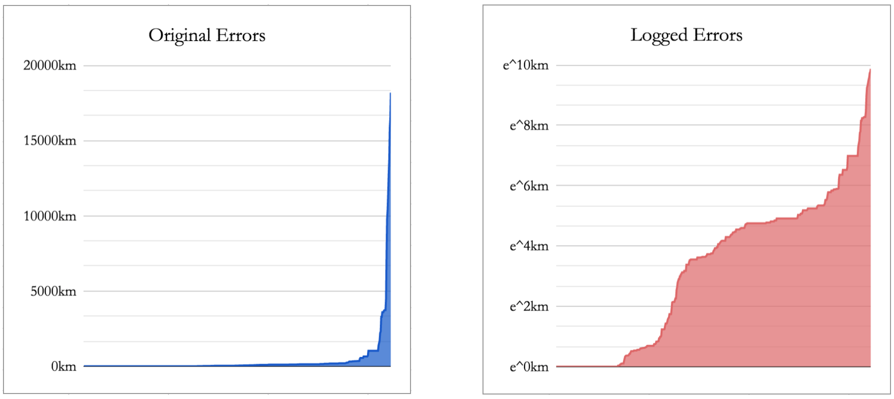
5.4.1 Geotagging Metrics

There is a strong agreement on the appropriate geotagging evaluation metric so most attention will focus on toponym resolution. As a subtask of NER, geotagging is evaluated using the *F-Score*, which is also our recommended metric and an established standard for this stage of geoparsing [110]. Figures for precision and recall may also be reported as some applications may trade precision for recall or may deem precision/recall errors more costly.

5.4.2 Toponym Resolution Metrics

Several geocoding metrics have been used in previous work and can be divided into *three groups* depending on their output format. We assert that the most 'fit for purpose' output of a geoparser is a *pair of coordinates*, not a categorical value or a ranked list of toponyms, which can give unduly flattering results [170]. Ranked lists may be acceptable if subjected to further human judgement and/or correction but not as the final output. With set-based metrics such as the *F-Score*, when used for geocoding, there are several issues: (a) Database incompatibility for geoparsers built with different knowledge bases that cannot be aligned to make fair benchmarking feasible. (b) The all-or-nothing approach implies that every incorrect answer (e.g. error greater than 5-10km) is equally wrong. This is not the case, geocoding errors are *continuous* variables, not categorical variables hence the F-Score is unsuitable for toponym resolution. (c) Underspecification of recall versus precision, i.e. is a correctly geotagged toponym with an error greater than *Xkm* a false positive or a false negative? This is important for accurate precision and recall figures. Set-based metrics and ranked lists are prototypical cases of trying to fit the wrong evaluation metric to a task. We now briefly discuss each metric group.

Coordinates-based (continuous) metrics are the recommended group when the output of a geoparser is a *pair of coordinates*. An error is defined as the distance from predicted coordinates to gold coordinates. *Mean Error* is a regularly used metric [43, 87], analogous to a sum function thus informs of the total error as well. *Accuracy@Xkm* is the percentage of errors resolved within *Xkm* of gold coordinates. [73] and [186] used accuracy within 5km, [170, 52] used accuracy at 5, 50, 250km, related works on tweet geolocation [177, 207, 74, 166] use accuracy at 161km. We recommend the more lenient 161km as it covers errors stemming from database misalignment. *Median Error* is a simple metric to interpret [196, 177] but is otherwise uninformative as the error distribution is non-normal hence not recommended. The Area Under the Curve [72, 92] is another coordinate-based metric, which follows in a separate subsection.



(a) Original Errors

(b) Logged Errors

Fig. 5.3 Computing the area under the curve by integrating the *Logged Errors* in Figure (b). $AUC = 0.33$ is interpreted as 33% of the maximum geocoding error. 20,039 km is 1/2 of Earth’s circumference.

Set-based/categorical metrics and more specifically, the F-Score, has been used alongside coordinates-based metrics [105, 7] to evaluate the performance of the full pipeline. A true positive was judged as a correctly geotagged toponym *and* one resolved to within a certain distance. This ranges from 5km [7, 111] to 10 miles [93, 112] to all of the previous thresholds [99] including 100km and 161km. In cases where WordNet has been used as the ground truth [26] an F-Score might be appropriate given WordNet’s structure but it is not possible to make a comparison with a coordinates-based geoparser. Another problem with it is the all-or-nothing scoring. For example, *Vancouver, Portland, Oregon* is an acceptable output if *Vancouver, BC, Canada* was the expected answer. Similarly, the implicit suggestion that *Vancouver, Portland* is equally wrong as *Vancouver, Australia* is erroneous. Furthermore, using F-Score exclusively for the full pipeline does not allow for evaluation of individual geoparsing components making identifying problems more difficult. As a result, it is not a recommended metric for toponym resolution.

Rankings-based metrics such as Eccentricity, Cross-Entropy, Mean Reciprocal Rank, Mean Average Precision and other variants (Accuracy@k, Precision@k) have sometimes been used or suggested [94, 37]. However, due to the aforementioned output format, ranked results are not recommended for geocoding. These metrics have erroneously been imported from Geographic Information Retrieval and should not be used in toponym resolution.

Area Under the Curve (AUC) is a recent metric used for toponym resolution evaluation [72, 92]. It is not to be confused with other AUC variants, which include the AUC of ROC, AUC for measuring blood plasma in Pharmacokinetics[9] or the AUC of the Precision/Recall curve. The calculation uses the standard calculus method to integrate the area under the curve of geocoding errors denoted as x, using the Trapezoid Rule[10].

$$Area\ \ Under\ \ the\ \ Curve = \frac{\int_0^{dim(x)} \ln(x)dx}{dim(x) * ln(20039)}$$

The original errors, which are highly skewed in Figure 5.3(a) are scaled down using the *natural logarithm* resulting in Figure 5.3(b). The area under the curve divides into the total area of the graph to compute the final metric value. The logarithm decreases the effect of *outliers* that tend to distort the Mean Error. This allows for evaluation of the majority of errors that would otherwise be suppressed by outliers.

5.4.3 Recommended Metrics for Toponym Resolution

There is no single metric that covers every important aspect of geocoding, therefore based on the previous paragraphs, we make the following recommendations. (1) *The AUC* is a comprehensive metric as it accounts for *every error*, it is suitable for a rigorous comparison but needs some care to be taken to understand. (2) *Accuracy@161km* is a fast and intuitive way to inform of “correct" resolutions (error within 100 miles of gold coordinates) but ignores the rest of the error distribution. (3) *Mean Error* is a measure of average and total error but it hides the full distribution, treats all errors as equal and is prone to distortion by outliers. Therefore, using *all three metrics* gives a holistic view of geocoding performance as they compensate for each others’ weaknesses while testing different aspects of toponym resolution. The SOTA model should perform well across all three metrics. As a final recommendation, an informative and intuitive way to assess the full pipeline would be to indicate how many toponyms were successfully extracted and resolved as in Table 5.4. Using the Accuracy@161km, we can observe the *percentage* of correctly recognised and resolved toponyms to estimate the performance of the combined system.

5.4.4 Important Considerations for Evaluation

The Choice of the Database of geographic knowledge used by the geoparser and/or for labelling datasets must be clearly noted. In order to make a fair comparison between models *and* datasets, the toponym coordinates must be a close match. Incompatibilities

[9] The branch of pharmacology concerned with the movement of drugs within the body.

[10] https://docs.scipy.org/doc/numpy/reference/generated/numpy.trapz.html

between global gazetteers have been previously studied [2]. The most popular and open-source geoparsers and datasets do use Geonames[11] allowing for an "apples to apples" comparison (unless indicated otherwise). In case it is required, we also propose a database alignment method for an empirically robust comparison of geoparsing models and datasets with incompatible coordinate data[12]. The adaptation process involves a post-edit to the output coordinates. For each toponym, retrieve its nearest candidate by measuring the distance from the predicted coordinates (using a different knowledge base) to the Geonames toponym coordinates. Finally, output the Geonames coordinates to allow for a reliable comparison.

Resolution scope also needs to be noted when comparing geoparsers, although it is less likely to be an issue in practice. Different systems can cover different areas, for example, geoparsers with *Local Coverage* such as country-specific models [130] versus *Global Coverage*, which is the case with most geoparsers. It is not possible to fairly compare these two types of systems.

The train/dev/test data source domains, i.e. the homogeneity or heterogeneity of the evaluation datasets is a vital consideration. The distribution of the evaluation datasets must be noted as performance will be higher on *in-domain data*, which is when all partitions come from the same corpus. When training data comes from a different distribution from the test data, for example News Articles versus Wikipedia, the model that can *generalise to out-of-domain test data* should be recognised as superior even if the scores are similar.

Statistical significance tests need to be conducted when making a comparison between two geoparsers unless a large performance gap makes this unnecessary. There are two options (1) *k-fold cross-validation followed by a t-test* for *both* stages or (2) the *McNemar's test* for Geotagging and the *Wilcoxon Signed-Rank Test* for Geocoding. The k-fold cross-validation is only suitable when a model is to be *trained from scratch* on *k-1* folds, *k* times. For evaluation of trained geoparsers, we recommend using the latter options with similar statistical power, e.g. when it is infeasible to train several deep learning models.

K-Fold Cross-Validation works by generating 5-10 folds that satisfy the i.i.d. requirement for a parametric test [53]. This means folds should (a) come from disjoint files/articles and *not* be randomised to satisfy the independent requirement and (b) come from *the same domain* such as news text to satisfy the identically distributed requirement. GeoWebNews

[11] https://www.geonames.org/export/

[12] The code can be found in the project's GitHub repository.

satisfies those requirements by design. The number of folds will depend on the size of the dataset, i.e. fewer folds for a smaller dataset and vice versa. Following that, we obtain scores for each fold, perform a t-test and report the p-value. There is a debate as to whether a p-value of 0.05 is rigorous enough. We think 0.01 would be preferred but in any case, the lower the more robust. Off-the-shelf geoparsers should be tested as follows.

For *Geotagging*, use *McNemar's test*, a non-parametric statistical hypothesis test suitable for matched pairs produced by binary classification or sequence tagging algorithms [45]. McNemar's test compares the disagreement rate between two models using a contingency table of the outputs of two models. It computes the probability of two models 'making mistakes' at the same rate, using chi-squared distribution with one degree of freedom. If the probability of obtaining the computed statistic is less than 0.05, we reject the null hypothesis. For a more robust result, a lower threshold is preferred. This test is not well-approximated for contingency table values less than 25, however, if using multiple of our recommended datasets, this is highly unlikely.

For *Toponym Resolution,* use a two-tailed *Wilcoxon Signed-Rank Test* [194] for computational efficiency as the number of test samples across multiple datasets can be large (10,000+). Geocoding errors follow a power law distribution (Figure 5.3a) with many outliers among the largest errors hence the non-parametric test. This sampling-free test compares the matched samples of geocoding errors. The null hypothesis assumes that the ranked differences between models' errors are centred around zero, i.e. model one is right approximately as much as model two. Finally, report the p-value and z-statistic.

5.4.5 Unsuitable Datasets

Previous works in geoparsing [103, 7, 170, 105] have evaluated with their own labelled data but we have been unable to locate those resources. For those that are freely available, we briefly discuss the reasons for their unsuitability. AIDA [82] is a geo-annotated CoNLL 2003 NER dataset, however, the proprietary CoNLL 2003 data is required to build it. Moreover, the CoNLL file format does not allow for original text reconstruction due to the missing whitespace. SpatialML [119, 120] datasets are primarily focused on spatial expressions in natural language documents and are not freely available ($500-$1,000 for a license[13]). Twitter datasets such as GeoCorpora [191] experience a gradual decline in completeness as

[13] https://catalog.ldc.upenn.edu/LDC2011T02

users delete their tweets and deactivate profiles. WoTR [42] and CLDW [163] are suitable only for digital humanities due to their historical nature and localised coverage, which is problematic to resolve [28]. CLUST [110] is a corpus of clustered streaming news of global events, similar to LGL. However, it contains only 223 toponym annotations. TUD-Loc2013 [97] provides incomplete coverage, i.e. no adjectival or embedded toponyms, however, it may generate extra training data with some editing effort.

5.4.6 Recommended Datasets

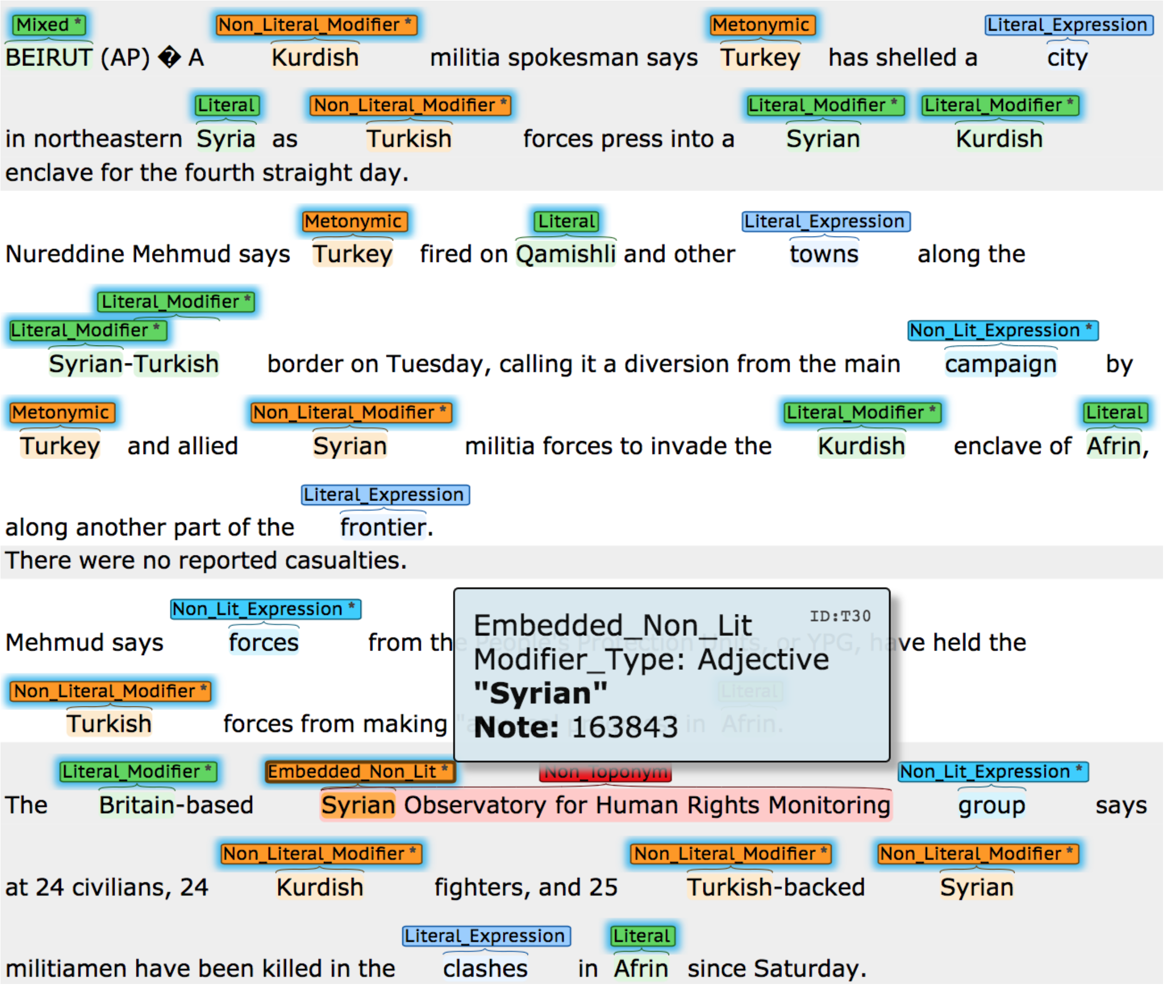
We recommend evaluation with the following *open-source* datasets: (1) WikToR [72] is a large collection of programmatically annotated Wikipedia articles and although quite artificial, to our best knowledge, it's the most difficult test for handling *toponym ambiguity* (Wikipedia coordinates). (2) Local Global Lexicon (LGL) [112] is a global collection of local news articles (Geonames coordinates) and likely the most frequently cited geoparsing dataset. (3) GeoVirus [70] is a WikiNews-based geoparsing dataset centred around disease reporting (Wikipedia coordinates) with global coverage though without adjectival toponym coverage. (4) TR-NEWS [93] is a new geoparsing news corpus of local and global articles (Geonames coordinates) with excellent toponym coverage and metadata. (5) Naturally, we also recommend GeoWebNews for a complete, fine-grained, expertly annotated and broadly sourced evaluation dataset.

5.5 GeoWebNews

As our final contribution, we introduce a new dataset to enable evaluation of fine-grained tagging and classification of toponyms. This will facilitate an immediate implementation of the proposals from previous sections. The dataset comprises 200 articles from 200 globally distributed news sites. Articles were sourced via a collaboration with the European Union's Joint Research Centre[14], collected during 1st-8th April 2018 from the European Media Monitor [178] using a wide range of multilingual trigger words/topics[15]. We then randomly selected exactly one article from each domain (English language only) until we reached 200 news stories. We also share the BRAT [179] configuration files to expedite future data annotation using the new scheme. GeoWebNews can be used to evaluate the performance of NER (locations only) known as Geotagging and Geocoding/Toponym Resolution [70], develop and evaluate Machine Learning models for sequence tagging and classification, geographic

[14] https://ec.europa.eu/jrc/en

[15] http://emm.newsbrief.eu/



BEIRUT (AP) � A Kurdish militia spokesman says Turkey has shelled a city in northeastern Syria as Turkish forces press into a Syrian Kurdish enclave for the fourth straight day.

Nureddine Mehmud says Turkey fired on Qamishli and other towns along the Syrian-Turkish border on Tuesday, calling it a diversion from the main campaign by Turkey and allied Syrian militia forces to invade the Kurdish enclave of Afrin, along another part of the frontier.

There were no reported casualties.

Mehmud says forces from th ve held the Turkish forces from making in Afrin.

Embedded_Non_Lit ID:T30
Modifier_Type: Adjective
"Syrian"
Note: 163843

The Britain-based Syrian Observatory for Human Rights Monitoring group says at 24 civilians, 24 Kurdish fighters, and 25 Turkish-backed Syrian militiamen have been killed in the clashes in Afrin since Saturday.

Fig. 5.4 A GeoWebNews article. An asterisk indicates an attribute, either *a modifier_type* [Adjective, Noun] and/or *a non_locational* [True, False].

information retrieval, even used in a Semantic Evaluation [127] task. GeoWebNews is a web-scraped corpus hence a few articles may contain duplicate paragraphs or some missing words from improperly parsed web links, which is typical of what might be encountered in practical applications.

5.5.1 Annotation Procedure and Inter-Annotator Agreement (IAA)

The annotation of 200 news articles at this level of granularity is a laborious and time-consuming effort. However, annotation quality is paramount when proposing changes/extensions to existing schemes. Therefore, instead of using crowd-sourcing, annotation was performed by the first author and two linguists from Cambridge University's Modern and Medieval

Languages Faculty[16]. In order to expedite the verification process, we decided to make the annotations of the first author available to our linguists as 'pre-annotation'. Their task was then twofold: (1) *Precision Check*: verification of the first author's annotations with appropriate edits; (2) *Recall Check*: identification of additional annotations that may have been missed. The F-Scores for the Geotagging IAA were computed using BratUtils[17], which implements the MUC-7 scoring scheme [31]. The Geotagging IAA after adjudication were 97.2 and 96.4 (F-Score), for first and second annotators respectively, computed on a 12.5% sample of 336 toponyms from 10 randomly chosen articles (out of a total of 2,720 toponyms across 200 articles). The IAA for a simpler binary distinction (literal versus associative types) were 97.2 and 97.3.

5.5.2 Annotation of Coordinates

The Geocoding IAA with the first annotator on the same 12.5% sample of toponyms expressed as accuracy [correct/incorrect coordinates] was 99.7%. An additional challenge with this dataset is that some toponyms ($\sim$8%) require either an extra source of knowledge such as Google Maps API, a self-compiled list of businesses and organisations names such as [130] or even human-like inference to resolve correctly. These toponyms are facilities, buildings, street names, park names, festivals, universities and other venues. We have estimated the coordinates for these toponyms, which do not have an entry in Geonames using Google Maps API. These toponyms can be excluded from evaluation, which is what we did, due to the geoparsing difficulty. We have excluded 209 of these toponyms plus a further 110 demonyms, homonyms and language types without coordinates, evaluating with the remaining 2,401. We did not annotate the articles' geographic focus as was done for Twitter [54, 166] and Wikipedia [100].

5.5.3 Evaluation

Sections 5.3 and 5.4 have established GeoWebNews as a new standard dataset for fine-grained geoparsing grounded in real-world pragmatic usage. In the remainder of this section, we shall evaluate the SOTA Geoparsing and NER models to assess their performance on the linguistically nuanced location dataset, which should aid future comparisons with new NLP models. For a broad comparison, we have also included the Yahoo! Placemaker[18], the Edinburgh Geoparser [73] and our own CamCoder [70] resolver as the main geoparsing

[16] https://www.mml.cam.ac.uk/

[17] https://github.com/savkov/BratUtils

[18] The service was officially decommissioned but some APIs remain accessible.

benchmarks. We have also considered GeoTxt [95], however, due to low performance, it was not included in the tables. Further related geoparsing evaluation with diverse datasets/systems can be found in our previous papers [71][72].

Geotagging GeoWebNews

For toponym extraction, we selected the two best models from Table 5.2, Google Cloud Natural Language[19] and SpacyNLP[20]. We then trained an NCRF++ model [199], which is an open-source Neural Sequence Labeling Toolkit[21]. We evaluated models using 5-Fold Cross-Validation (40 articles per fold, 4 train and 1 test fold). Embeddings were initialised with 300D vectors[22] from GloVe [151] in a simple form of transfer learning as training data was limited. The NCRF++ tagger was trained with default hyper-parameters but with two additional features, the dependency head and the word shape, both extracted with SpacyNLP. For this custom model, we prioritised fast prototyping and deployment over meticulous feature/hyper-parameter tuning hence there is likely more performance to be found using this approach.

NER Model / Geoparser	**Precision**	**Recall**	**F-score**
NCRF++ (Literal & Associative labels)	79.9	75.4	77.6
Yahoo! Placemaker	73.4	55.5	63.2
Edinburgh Geoparser	81	52.4	63.6
SpacyNLP	82.4	68.6	74.9
Google Cloud Natural Language	**91.0**	76.6	83.2
NCRF++ (“Location" label only)	90.0	**87.2**	**88.6**

Table 5.3 Geotagging F-Scores for GeoWebNews featuring the best performing models. The NCRF++ models’ scores were averaged over 5 folds (σ=1.2-1.3).

There were significant differences in precision and recall between off-the-shelf and custom models. SpacyNLP and Google NLP achieved a precision of 82.4 and 91 respectively while achieving a lower recall of 68.6 and 76.6 respectively. The NCRF++ tagger exhibited a balanced classification behaviour (90 precision, 87.2 recall). It achieved the highest F-Score of 88.6 despite only a modest amount of training examples.

[19] https://cloud.google.com/natural-language/
[20] https://spacy.io/usage/linguistic-features
[21] https://github.com/jiesutd/NCRFpp
[22] Common Crawl 42B - https://nlp.stanford.edu/projects/glove/

Geotagging with two labels (physical location versus associative relationship) was evaluated with a custom NCRF++ model. The mean F-Score over 5 folds was 77.6 (σ=1.7), which is higher than SpacyNLP (74.9) with a single label. This demonstrates the feasibility of geotagging on two levels, treating toponyms separately in downstream tasks. For example, literal toponyms may be given a higher weighting for the purposes of geolocating an event. In order to incorporate this functionality into NER, training a custom sequence tagger is currently the best option for a two-label toponym extraction.

Geocoding GeoWebNews

For the evaluation of toponym resolution, we have excluded the following examples from the dataset. (a) the most difficult to resolve toponyms such as street names, building names, festival venues and so on, which account for ∼8% of the total, without an entry in Geonames and often requiring a reference to additional resources. (b) demonyms, languages and homonyms, accounting for ∼4% of toponyms as these are not locations hence do not have coordinates. The final count was 2,401 (∼88%) toponyms in the test set. Several setups were evaluated for a broad indication of expected performance. For geotagging, we used SpacyNLP to extract a *realistic* subset of toponyms for geocoding, then scored the true positives with a matching entry in Geonames. The second geotagging method was Oracle NER, which *assumes* perfect NER capability. Although artificial, it allows for geocoding of all 2,401 toponyms. We have combined these NER methods with the CamCoder [70] default model[23]. The population heuristic was also evaluated as it was shown to be a strong baseline in our previous work. In practice, one should expect to lose up to 30-50% toponyms during geotagging, depending on the dataset and NER. This may be seen as a disadvantage, however, in our previous work as well as in Table 5.4, we found that a ∼50% sample is representative of the full dataset.

The overall errors are low indicating low toponym ambiguity, i.e. low geocoding difficulty of the dataset. Other datasets [72] can be more challenging with errors 2-5 times greater. When provided with a database name for each extracted toponym (Oracle NER), it is possible to evaluate the whole dataset and get a sense of the pure disambiguation performance. However, in reality, geotagging is performed first, which reduces that number significantly. Using the geoparsing pipeline of SpacyNLP + CamCoder, we can see that 94-95% of the 1,547 correctly recognised toponyms were resolved to within 161km. The number of recognised toponyms could be increased with a “normalisation lexicon" that maps non-standard surface forms such as adjectives (“Asian", “Russian", “Congolese") to their canonical/database names. SpacyNLP provides a separate class for these toponyms called *NORP*, which stands for

[23] https://github.com/milangritta/Geocoding-with-Map-Vector

Setup/Description	Mean Err	Acc@161km	AUC	# of Toponyms
SpacyNLP + CamCoder	**188**	**95**	**0.06**	1,547
SpacyNLP + Population	210	**95**	0.07	1,547
Oracle NER + CamCoder	232	94	**0.06**	2,401
Oracle NER + Population	250	94	0.07	2,401
Yahoo! Placemaker*	203	91	0.09	1444
Edinburgh Geoparser*	338	91	0.08	1363

Table 5.4 Toponym Resolution scores for the GeoWebNews data. **This geoparser provides both Geotagging and Geocoding steps.*

nationalities, religious or political groups. Such lexicon could be assembled with a gazetteer-based statistical n-gram model such as [3] that uses multiple knowledge bases or a rule-based system [188]. For unknown toponyms, approximating the geographic representation from places that co-occur with it in other documents [78] may be an option. Finally, not all errors can be evaluated in a conventional setup. Suppose an NER tagger has 80% precision. This means 20% of false positives will be used in downstream processing. In practice, this subset carries some unknown penalty that NLP practitioners hope is not too large. For downstream tasks, however, this is something that should be considered during error analysis.

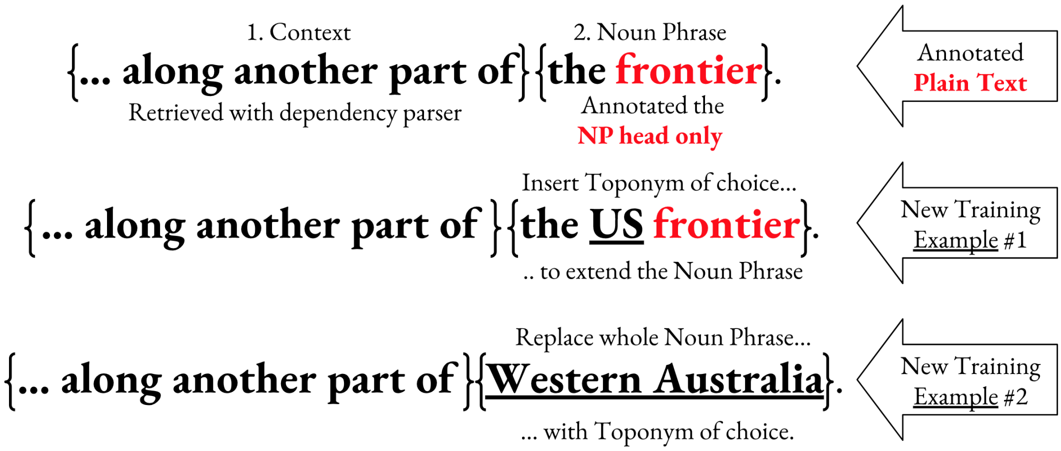

1. Context
2. Noun Phrase
{... along another part of}{the frontier}.
Retrieved with dependency parser
Annotated the
NP head only
Annotated
Plain Text
Insert Toponym of choice...
{... along another part of }{the US frontier}.
.. to extend the Noun Phrase
New Training
Example #1
Replace whole Noun Phrase...
{... along another part of }{Western Australia}.
... with Toponym of choice.
New Training
Example #2

Fig. 5.5 An augmentation of a literal training example. An associative augmentation equivalent might be something like **{The deal was agreed by} {the chief engineer.}** replacing "the chief engineer" by a toponym.

(1) No Aug.	(2) Partial Aug.	(3) Full Aug.	(4) Ensemble of (1, 2, 3)
88.6	88.2	88.4	88.5

Table 5.5 F-Scores for NCRF++ models with 5-Fold Cross-Validation. No improvement was observed for the augmented or ensemble setups over baseline.

Training Data Augmentation

We have built the option of data augmentation right into GeoWebnews and shall now demonstrate its possible usage in a short experiment. In order to augment the 2,720 toponyms to double or triple the training data size, two additional lexical features (*NP heads*) were annotated, denoted *Literal Expressions* and *Associative Expressions*[24]. These annotations generate two separate components (a) the NP context and (b) the NP head itself. In terms of distribution, we have literal (N=1,423) versus associative (N=2,037) *context* and literal (N=1,697) versus associative (N=1,763) *heads*, indicated by a binary *non-locational* attribute. These two interchangeable components give us multiple permutations from which to generate a larger training dataset[25] (see Figure 5.5 for an example). The associative expressions are deliberately dominated by ORG-like types because this is the most frequent metonymic pair [5].

Table 5.5 shows three augmentation experiments (numbered 2, 3, 4) that we have compared to the best NCRF++ model (1). We hypothesised that data augmentation, i.e. adding additional modified training instances would lead to a boost in performance, however, this did not materialise. An ensemble of models (4) also did not beat the baseline NCRF++ model (1). Due to time constraints, we have not extensively experimented with elaborate data augmentation and *encourage further research* into other implementations.

5.5.4 Conclusion and Future Work

In this manuscript, we introduced a detailed pragmatic taxonomy of toponyms as a way to increase Geoparsing recall and to differentiate literal uses (53%) of place names from associative uses (47%) in a corpus of multi-source global news data. This helps clarify the task objective, quantifies type occurrences, informs of common NER mistakes and enables innovative handling of toponyms in downstream tasks. In order to expedite future research, address the lack of resources and contribute towards replicability and extendability [66, 29], we shared the annotation framework, recommended datasets and any tools/code

[24] Google Cloud NLP already tags common nouns in a similar manner.

[25] https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation

required for fast and easy extension. The NCRF++ model trained with just over 2,000 examples showed that it can outperform SOTA taggers such as SpacyNLP and Google NLP for location extraction. The NCRF++ model can also achieve an F-Score of 77.6 in a two-label setting (literal, associative) showing that fine-grained toponym extraction is feasible. Finally, we critically reviewed current practices in geoparsing evaluation and presented our best recommendations for a fair, holistic and intuitive performance assessment. Future work may focus on generalising our pragmatic evaluation to other NER classes as well as extend the coverage to languages other than English. As we conclude this section, here is a summary of the recommended evaluation.

1. Review (and report) important geoparsing considerations in Section 5.4.4.
2. Use a proprietary or custom NER tagger to extract toponyms using the recommended dataset(s) as demonstrated in Section 5.5.3.
3. Evaluate geotagging using F-Score as the recommended metric and report statistical significance with McNemar's Test (Section 5.4.4).
4. Evaluate toponym resolution using Accuracy@161, AUC and Mean Error as the recommended metrics, see Section 5.5.3 for an example.
5. *Optional*: Evaluate geocoding in "laboratory setting" as per Section 5.5.3.
6. Report the number of toponyms resolved and the statistical significance using the Wilcoxon Signed-Rank Test (Section 5.4.4).

5.5.5 Closing Thoughts

The Principle of Wittgenstein's Ruler from Nassim N. Taleb's book, Fooled by Randomness [181] deserves a mention as we close this paper with some philosophical remarks for the future of IE. It says: "*Unless you have confidence in the ruler's reliability, if you use a ruler to measure a table you may also be using the table to measure the ruler.*" In the case of Geoparsing and NER, this translates into: "*Unless you have confidence in the dataset, if you use the dataset to evaluate the real-world, you may also be using the real-world to evaluate the dataset.*" If research is optimised towards and benchmarked on a possibly unreliable or unrepresentative dataset, does it matter how it performs in the real world? The biases in machine learning reflect human biases so unless we improve training data annotation and generation, machine learning algorithms will not reflect linguistic reality. We must pay close attention to the correctness, diversity, timeliness and particularly the *representativeness* of the dataset, which also influences the soundness of the task objective. Many models still tune their performance on CoNLL NER '03 data [197], for example at COLING 2018 [198] and ACL 2018 [69], which comes from a single source (Reuters News) [185] and is shallowly annotated (not open-source either). It is important to ask whether models trained and evaluated on data that does not closely mirror the real-world is the goal to aim for.

Chapter 6

Geoparsing for Disease Monitoring

6.1 Introduction

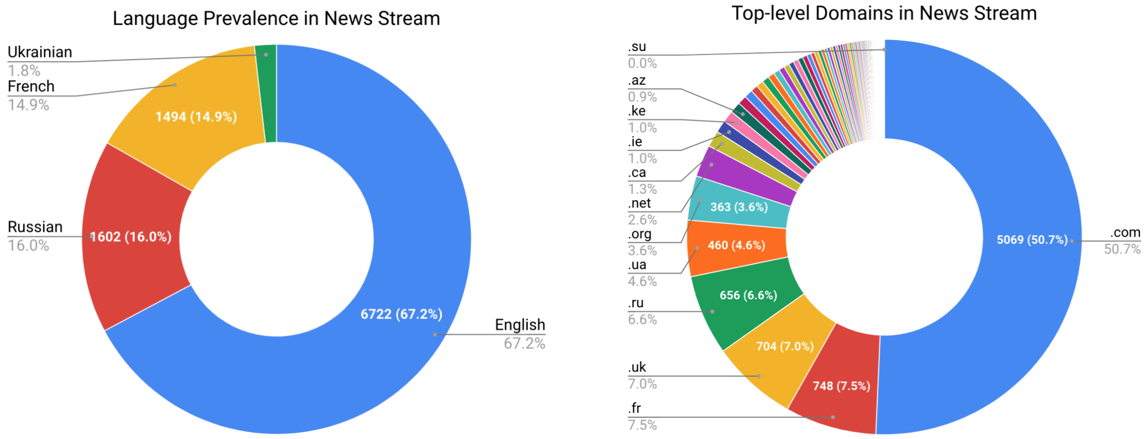
The previous chapter served a dual purpose; we have introduced a Standard Evaluation Framework for geoparsing while producing the annotated files that would be used for this chapter's evaluation of MediSys' geoparsing capability. As we mentioned in the GeoWebNews section of the previous chapter, the text data in the form of 200 articles was made available to us via the MediSys news feed from the Joint Research Centre. The maintenance and development of a complex production system such as MediSys means may not always feasible to periodically assess the geoparsing quality of the system in great detail. Therefore, we think our analysis of EMM's geoparsing behaviour will be informative not only to the reader, but above all, to JRC researchers and other groups in the health surveillance space. This chapter leverages the research presented thus far in the thesis to propose and demonstrate new ways to reduce errors at each stage of geoparsing using proven and/or theoretical methods, actionable under the constraints of a production system. Some of these software engineering constraints are multilingual support, high scalability, inexpensive implementation and model interpretability. The visit to JRC was an enlightening demonstration of the differences between laboratory studies in research settings versus custom requirements in a real-time production environment. We have taken on board the feedback and suggestions from Dr Jens Linge when making recommendations for geoparsing extensions in this chapter, which comprises three major themes: 1) Evaluation of the geoparsing stage of MediSys with actionable suggestions for error reduction given the often challenging nature of the task and data; 2) Proposal of the FlexiGraph, a geoparsing extension for improved document geolocation and keyword ranking; and 3) Discussion of specific configurations of geoparsing systems based on evaluation of MediSys data.

6.2 Geoparsing Evaluation

This section analyses the geoparsing performance of the MediSys[1] part of the European Media Monitor as sampled in April 2018 via a dedicated RSS feed provided by the Joint Research Centre. We compare its scores with easily accessible geoparsers and NER taggers, discuss errors, provide feasible solutions, analyse toponym types and determine the most effective paths to improvements and/or system extensions. MediSys was evaluated previously such as 10 years ago [167], however, not for geoparsing but for its ability to serve as an early warning system for different types of hazards such as food-borne illnesses.

6.2.1 MediSys News Stream

We now give an overview of the news feed (a subset of the EMM data stream) we have been granted access to for the geoparsing evaluation of MediSys. For this overview, we have sampled 10,000 news articles (200 per batch) during October/November 2018 to obtain essential stream statistics shown in Figures 6.1 and 6.2. The available languages and source domain distributions are shown in Figure 6.1. The articles come from almost 100 diverse top-level domains with two thirds of the news text written in English. We will come back to this when discussing the availability of NLP tools for a limited number of languages. Despite resources provided only for the most frequent languages, we could feasibly cover a large percentage of MediSys articles with just a few separate unilingual models.



(a) Languages of the news stream.

(b) Domains of the news stream.

Fig. 6.1 Languages and Domain sources of our MediSys data stream (subset).

[1] https://ec.europa.eu/jrc/en/scientific-tool/medical-information-system

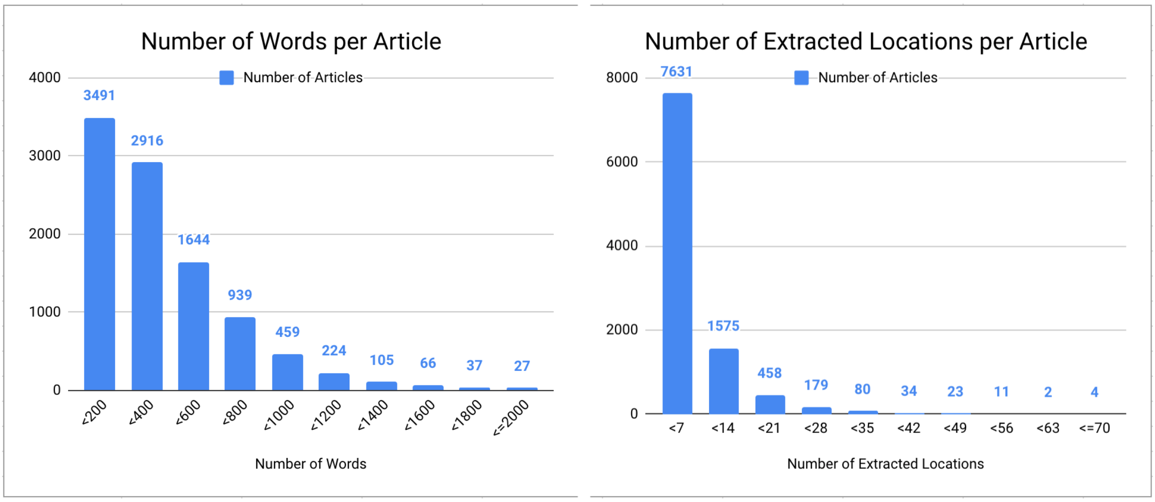


Fig. 6.2 The distributions of word counts and extracted toponyms per article. The counts in both histograms were clipped at 70 locations and 2,000 words respectively.

There are only four languages represented in our data feed, unlike the full EMM service News Brief[2]. In terms of word count, 80% of articles have fewer than 600 words (Figure 6.2, left histogram). Over 90% of news articles have fewer than 14 *extracted* toponyms. With EMM's geotagging precision of 58% and recall of 56%, we may cautiously estimate the extracted location counts in Figure 6.2 (right histogram) to be close to the actual figures based on the sampling process for the GeoWebNews dataset. *Please, note: MediSys/EMM is being upgraded at the start of 2019 with a new geoparsing component (NERone). Unless stated otherwise, all scores reported in this thesis are for the 2018 system.*

6.2.2 MediSys Geotagging

It is important to state that MediSys is required to be a fast, language-independent news analysis tool. For the software implementation to be scalable, transparent and interpretable, this necessarily means lower performance and simpler models due to these constraints. With this in mind, we shall be making specific implementation suggestions, which can be incorporated into an existing infrastructure, primarily in a unilingual setting. Any comparisons shown intend to provide additional context/options rather than a direct benchmarking test.

The key conclusion from Table 6.1 is that it is feasible to train a high-quality custom NER tagger from limited training data showing a high performance gain in terms of Recall and F-Score. Precision is approximately the same as the best NER tagger (Google NLP),

[2] http://emm.newsbrief.eu/

NER Model	**Precision**	**Recall**	**F-Score**
Custom NCRF++ (Literal & Assoc. labels)	79.9	75.4	77.6
European Media Monitor (MediSys)	57.8	55.9	56.8
Yahoo! Placemaker	73.4	55.5	63.2
Edinburgh Geoparser	81	52.4	63.6
Spacy NER	82.4	68.6	74.9
Google NLP	**91**	76.6	83.2
Custom NCRF++ (single label)	90	**87.2**	**88.6**

Table 6.1 Geotagging F-Scores on GeoWebNews data featuring the best performing NER models and Geoparsers. The NCRF++ scores were averaged over 5 folds (σ=1.2-1.3).

however, the resulting F-Score is ∼5 points higher. MediSys has a lower precision/recall than our custom trained model although it is mostly in line with other geoparsers. Even in a *two-label* NER setup, the NCRF++ still commands the third highest F-Score but *with* the added literalness distinction. From our experience, the NCRF++ framework is fast, effective and convenient, given carefully annotated data. Geotagging performance is better than the SOTA NER model by Google Cloud NLP, which is a proprietary service. It is also possible to customise other models such as Spacy NLP[3], which achieved good scores with the English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. We also included the two best geoparsers from our previous work for an additional perspective. Edinburgh and Yahoo! have higher precision scores than MediSys and do not deviate from their performance on LGL in Table 6.2.

Geotagging Error Analysis

False Positives would be expected to be high in a production system with demands for a scalable, multilingual, context-independent and fully interpretable geoparsing. Indeed, EMM's precision runs at 57.8% meaning that on average, 4 out of 10 tagged entities are not toponyms. Most of the false positives could be categorised as one of the following four types: 1) *First names*: Tara, Paula, Beryl, Olga, Mitchell, etc. 2) *Surnames*: Santiago, Metcalfe,

LGL	*Pre.*	*Rec.*	*F-Score*
GeoTxt	0.80	0.59	0.68
Edinburgh	0.71	0.55	0.62
Yahoo!	0.64	0.55	0.59
CLAVIN	0.81	0.44	0.57
Topocluster	0.81	0.64	0.71

Table 6.2 Geotagging scores on LGL.xml data as an additional comparison/context.

[3] https://spacy.io/usage/training

Garcia, Hansen, Pitman, etc. 3) *Acronyms*: TIGER, BA, ISIS, NAFTA, HART, etc. 4) *Common words*: More, Band, Reliance, Video, Summer, Justice, Legal, etc. All of them get resolved to coordinates introducing a significant amount of geographic noise and uncertainty. Many *homonyms* (types 1 and 2) could be handled with exclusion lists as these entities rarely refer to locations as their primary word sense. N.B. Practitioners must deal with the full consequences of the output while researchers may primarily aim to achieve a high score without overly concerning themselves with the impact of the false positives.

False Negatives were also expected to be higher for two reasons; a) the aforementioned system constraints and b) the presence of embedded and coercion toponym types, which comprise around 16% of the annotations. EMM's recall was 55.9% on GeoWebNews with the missing toponyms following a few patterns: "U.S." was one of the most frequent missed toponyms for reasons not immediately apparent but which may be due to the tokenisation protocol. Some state abbreviations were also overlooked, e.g. "SC" (South Carolina), "BC" (British Columbia) as well as "Calif." (California), "Ill." (Illinois), "N.J." (New Jersey) and "Va." (Virginia). "EU", which is an example of Coercion from Chapter 5 and often referring to the European *continent*, also did not get captured. Adjectival toponyms such as "Maltese", "Arab", "Serb", "Havanese", "Venezuelan" and "Ukrainian" were also not extracted. Another relatively frequent toponym causing recall to diminish was "Korea" and its variants: "Korean", "South Korea", "Korean Peninsula", "DPRK", etc. Of these variants, only "Korea" was being tagged in the absence of the other parts of the full toponym.

6.2.3 MediSys Toponym Resolution

This subsection evaluates several geoparsers alongside MediSys using multiple metrics to assess geocoding performance in a holistic manner. The initial EMM scores, marked '(raw)' in Table 6.3, indicate low performance. However, the partial reason for this quickly emerges, i.e., the coordinate selection policy, discussed in greater detail in section 6.4.1. MediSys resolves countries to their *capitals' coordinates* resulting in large errors. This is not the default policy in geoparsing and a careful error analysis should inform of this choice. Such policies can make it harder to establish SOTA. Table 6.3 also shows the manually and/or automatically '(aligned)' scores for *parsers using different knowledge bases*. However, even if we remove these penalties and align the coordinates, the errors are still high relative to the population baseline and the geocoders from Chapter 5.

In terms of the number of resolved toponyms, all geocoders except *Oracle NER* (the laboratory setup from Chapter 5) resolved a similar number of toponyms. The best overall scores

Model Description	Mean Error	A@161km	AUC	Toponyms
EMM NER + EMM TR (raw)	1,143	47	0.47	1,451
Yahoo NER + Yahoo TR (raw)	307	69	0.37	1,444
EMM NER + Population	1,110	82	0.18	1,359
EMM NER + EMM TR (aligned)	692	78	0.22	1,451
Yahoo NER + Yahoo TR (aligned)	203	91	0.09	1,444
Edinburgh NER + Edinburgh TR	338	91	0.08	1,363
Oracle NER + Population	250	94	0.07	2,401
Spacy NER + Population	210	**95**	0.07	1,547
Oracle NER + CamCoder	232	94	0.06	2,401
Spacy NER + CamCoder	**188**	**95**	**0.06**	1,547

Table 6.3 Toponym resolution scores for the GeoWebNews dataset.

were achieved by CamCoder, narrowly beating the strong population baseline. This heuristic performs exceedingly well on datasets covering global events, i.e. containing references to large geographic entities, which are less ambiguous. Most of MediSys' errors stem from the common ambiguous toponyms whose candidates have a wide geographic spread. *Typical Errors:* adjectival toponyms such as *Spanish* and *Canadian* resolved to cities bearing that name; Countries resolved to cities, e.g. *Egypt*, *Syria* and *Malta*; capitals resolved to their smaller counterparts, e.g. *Cairo*, *Manila*, *Hanoi*, this also happened with administrative areas such as *Texas*, *Florida* and *Berkshire*, quickly accumulating large penalties. Out of 1,451 toponyms, 78% were resolved to within 161km (aligned). Other errors included resolving *UK* to a place in Russia, *Irish* to an area somewhere in Ontario, Canada, *German* to a village on the Isle of Man and *White House* to a town in Tennessee. For an even greater perspective, we evaluated the resolution of EMM's retrieved toponyms using the *population heuristic*. The performance was equivalent with somewhat improved scores but a lower resolved toponym count (Table 6.3).

Figure 6.3 shows the *average error per feature type* for the population heuristic. This baseline can thus be used to measure the average ambiguity of toponyms of different feature types. Lower average error means easier toponym resolution due to the *high prior probability of the most populous candidate*. The bar chart shape in Figure 6.3 confirms the hypothesised shape, an expectation borne out of multiple annotation exercises. The lowest errors came from *PCLI* (independent political entity, i.e. country), *CONT* (continent), *PPLA/PPLA2* (seat of a first/second-order administrative division, e.g. county seat or state capital), *PPLC* (capital of a political entity) and *ADM1/ADM2* (first/second-order administrative division, e.g. US or Australian state). The moderate errors are generated from *PPL/PPLX*, which are

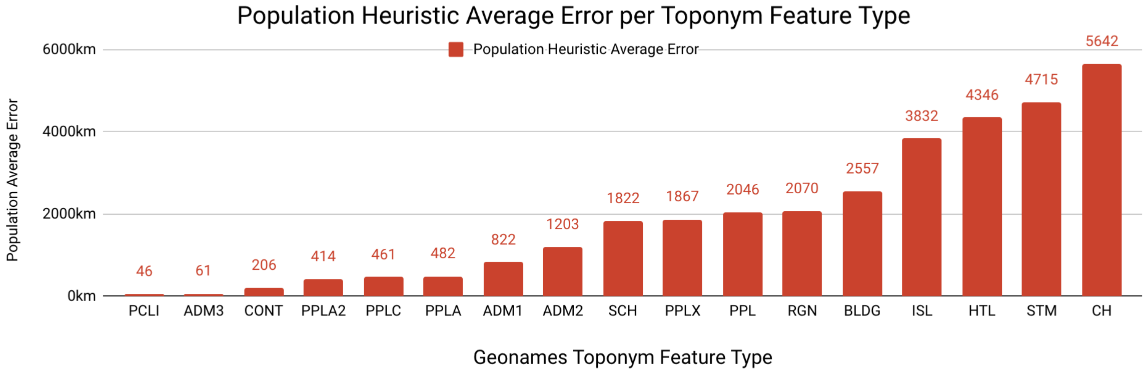


Fig. 6.3 Mean error of the population heuristic for each Geonames feature type.

populated places and their sections, i.e. small towns and villages. The feature types causing *on average* the highest errors are *BLDG* (buildings), *ISL* (islands), *CH* (churches) and *HTL* (hotels). This relationship is so strong that if we only knew the *feature type composition* of a given dataset, we can quite confidently predict the toponym resolution scores for that dataset. This is reflected by the highest geocoding errors for WikToR, which is dominated by the ambiguous *PPL* and *ADM2/3* types. GeoVirus and GeoWebNews are on the opposite end, low geocoding difficulty due to dominant *PCLI*, *ADM1* and *PPLC* types. LGL was designed to be a composition of both, which is reflected in its moderate geocoding errors.

6.3 FlexiGraph

Despite the progress in Geotagging and Toponym Resolution presented thus far, in production environments such as MediSys/EMM, we need to solve another important problem. During my visit to the European Commission's Joint Research Centre, I had the opportunity to discuss one of several active challenges of the monitoring system with resident engineers and researchers. It was the need for a type of *Intra-Document Analysis*. At the time of the visit, toponyms found in news articles were processed at the *document level*, i.e. not distinguished or differentiated *within* the document. The keyword or phrase of interest that triggered the analysis of an article would be associated with all extracted toponyms to the same degree. The researchers were actively looking for a method to allow the system to assess the pertinence of the topic of interest to each toponym found in the document. We know that documents frequently contain multiple toponyms (Figure 6.2). If the target keyword being tracked is "Zika", for example, a *naive method* would treat all toponyms contained in the retrieved news article as *equally* relevant. This approach is vulnerable to *Type 1 Errors*, i.e. false positives present in the article with little of no relevance to the target keyword(s). In order to advance beyond document-level geolocation of events, we require an adequate *intra-document analysis*. A system could benefit from 'relevance ranking' based on the distance from node *A*, such as "Zika" to node *B*, e.g. "Turkey". We will use the following short article from GeoWebNews as an example throughout the section.

***BEIRUT** (AP) A **Kurdish** militia spokesman says **Turkey** has shelled a city in north-eastern **Syria** as **Turkish** forces press into a **Syrian Kurdish** enclave for the fourth straight day. Nureddine Mehmud says **Turkey** fired on **Qamishli** and other towns along the **Syrian-Turkish** border on Tuesday, calling it a diversion from the main campaign by **Turkey** and allied **Syrian** militia forces to invade the **Kurdish** enclave of **Afrin**, along another part of the frontier. There were no reported casualties. Mehmud says forces from the People's Protection Units, or YPG, have held the **Turkish** forces from making "any real progress" in **Afrin**. The **Britain**-based **Syrian** Observatory for Human Rights Monitoring group says at 24 civilians, 24 **Kurdish** fighters, and 25 **Turkish**-backed **Syrian** militiamen have been killed in the clashes in **Afrin** since Saturday.*

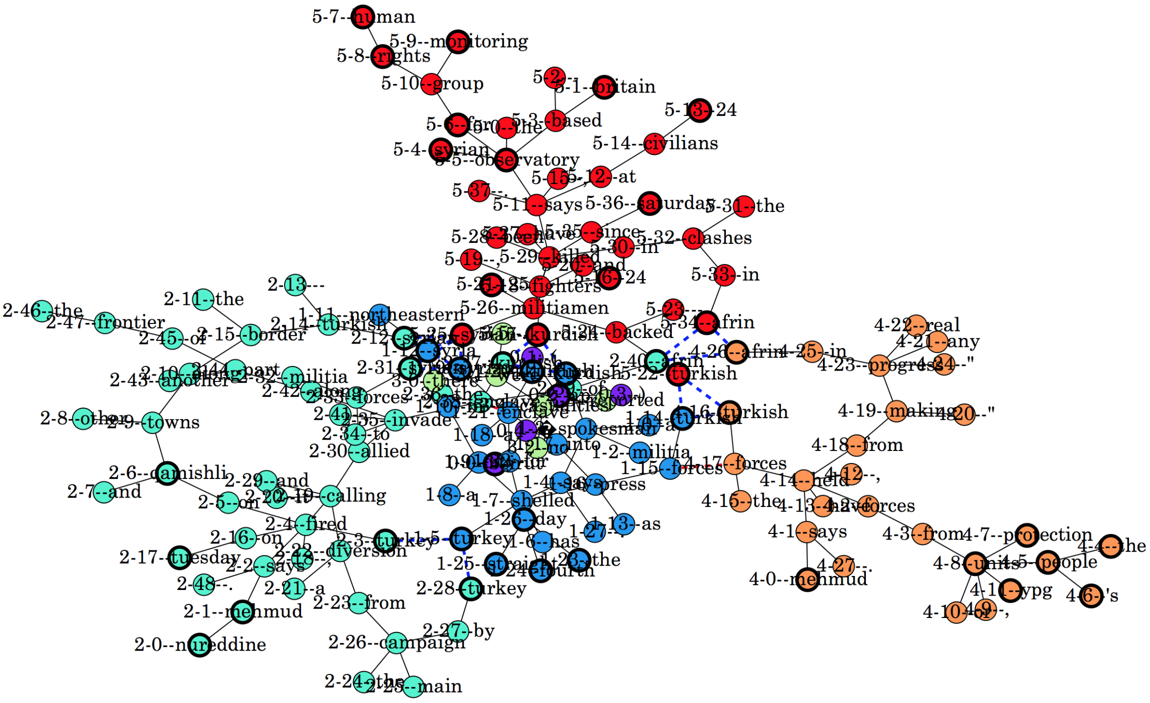


Fig. 6.4 The FlexiGraph prototype of the example article above.

FlexiGraph Construction We use the *dependency trees* of the document's sentences to construct a Sparse Undirected Affinity Graph seen in Figure 6.4. The affinity between any

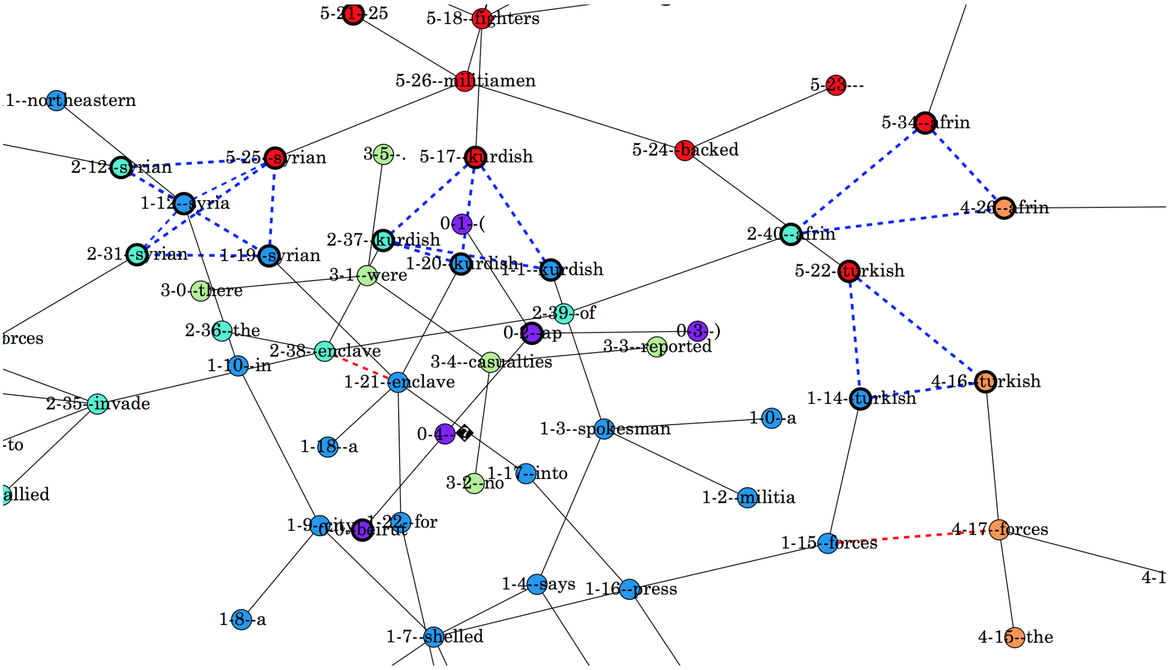


Fig. 6.5 A zoomed-in view of the FlexiGraph of the news article.

two nodes in *the same sentence* is their dependency distance with a default weight of 1.0 for each edge. The affinity of nodes *between sentences* can be defined as *a)* an edge between *entity mentions* with an edge weight of 0.0, which joins the sentences on an exact entity match (assuming the One Sense Per Discourse [61]). In Figures 6.4 and 6.5, this is shown as *blue dashed lines*, e.g. *Turkish*, *Kurdish*, *Turkey* and *Afrin*; *b)* an edge between *similar noun phrases* defined as the *cosine similarity* of the sum of the noun phrase word vectors, shown as red dashed lines in Figures 6.4 and 6.5, e.g. *Turkish forces* and *Kurdish enclave*. The edge weight is given as *1.0 - the cosine similarity*; *c)* an edge with a weight of 1.0 to connect *neighbouring sentences*, i.e., denoting a sentence border. We can then calculate the shortest paths between nodes as a heuristic to find the 'nearest' toponym to the trigger word. If toponym *A* is 10 units away from a word, e.g. "Zika" in node *Z* and toponym *B* is 20 units away from node *Z*, we hypothesise that *Z* is more likely to be relevant to toponym *A* than toponym *B*. This implementation has a simple software engineering setup suitable for requirements for a multilingual, fast, scalable and interpretable system. Even if a fast dependency parser is not available in a multilingual setting, the method could be simplified to a '*longest ngram match*' creating an edge between neighbouring ngrams in the sentence. While the code[4] for the FlexiGraph prototype is available, we have not been able to evaluate its effectiveness hence recommend this as possible future work.

[4] https://github.com/milangritta/FlexiGraph

6.4 Geolocation System Proposals

In this section, we bring together research themes covered thus far to propose several permutations of the geoparsing component of event monitoring systems. We then assess their potential performance in the wider context with recommendations focusing on proven and/or hypothetical improvements. Figure 6.6 shows the last sentence of the annotated news article from the earlier (FlexiGraph) example. Assuming we have an early warning system that is tracking the trigger word *’clashes’* as an example, we now wish to determine the most likely location of the relevant news event. Note that it is possible to illustrate greater or lesser benefits of our research using different examples. There are six toponyms in this paragraph and the event location is Afrin, Syria[5].

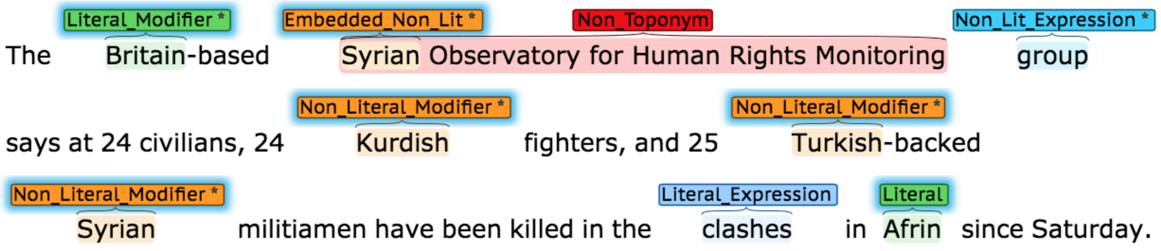

The Britain-based Syrian Observatory for Human Rights Monitoring group says at 24 civilians, 24 Kurdish fighters, and 25 Turkish-backed Syrian militiamen have been killed in the clashes in Afrin since Saturday.

Fig. 6.6 A snippet of an annotated news article from the previous section.

Scenario 1: Current System Beginning with geotagging, the F-Score for MediSys is 56.8 with a precision of 57.8% and a recall of 55.9%. In the above snippet, this means that ***on average***, three toponyms would go undetected and three false positives would be introduced. All toponyms will be treated as literal instead of only *Britain* and *Afrin*. Without any means to perform intra-document analysis, the trigger word ’*clashes*’ will be equally associated with the three remaining toponyms and three false positives, which may or may not refer to the event location of *Afrin*. The (aligned) geocoding accuracy to 161km is 78%, which may be seen as the probability of resolving the toponym to the correct location. This accuracy is adversely affected by *a)* three false negatives and three false positives; *b)* no literalness distinction as an additional filter; and *c)* no relevance ranking of the trigger word ’*clashes*’.

Scenario 2a: One-Label NER We have previously shown that for monolingual geotagging, we can efficiently train a custom NER tagger on a limited number of annotations resulting in a improved F-Score of 88.6 with a precision of 90% and a recall of 87.2%. In this configuration, we would be expect to retrieve (on average) at least 5 out of 6 toponyms with

[5] GeoWebNews was not annotated with articles’ geographic focus but can be retrofitted.

a chance of a 1 false positive. Deploying CamCoder for toponym resolution increases the accuracy at 161km to 95% leading to a significantly higher chance of correct geolocation. We could then employ FlexiGraph or some other intra-document heuristic to rank the relevance of the trigger word '*clashes*' to the event location of *Afrin*.

Scenario 2b: Two-Label NER The option of two-label geotagging (literal and associative) allows the system to determine which toponyms refer to physical locations and use this knowledge as an additional filter in event geolocation. The F-Score of 77.6 (88.6 for a single label) is lower for this configuration with a precision of 79.9% (90% for a single label) and a recall of 75.4% (87.2% for a single label). In the example paragraph, we would (*on average*) expect 2 false positives and some 2 false negatives. Whether the lower geotagging scores are likely to be offset by the literalness distinction will be discussed in the next section. CamCoder's toponym resolution capability does not change from the previous paragraph, remaining at 95%. Once again, we could deploy FlexiGraph as yet another filtering mechanism to identify *Afrin* as the event location.

Scenario 3: Other Geoparsers For an even broader comparison, we have also evaluated the two best geoparsers from Chapter 4; Edinburgh Geoparser and Yahoo! Placemaker on GeoWebNews. The Edinburgh Geoparser achieved an F-Score of 63.6 with a precision of 81% and a recall of 52.4%, which could have been higher if, like Yahoo! Placemaker and others, it duly extracted *adjectival toponyms*. Edinburgh Geoparser's geocoding accuracy to 161km is 91% with a low mean error. Yahoo! Placemaker achieved a similar F-Score of 63.2 with a precision of 73.4% and a recall of 55.5%. Its toponym resolution scores were high once the coordinates were aligned to use Geonames' data. Accuracy@161km increased from 69% to 91% and AUC decreased from 0.37 to just 0.09. Both NER components performed markedly worse than Spacy NLP, Google NLP and NCRF++, however, better than MediSys for both NER and Toponym Resolution. Not all geoparsers are open-source, however, it is possible to incorporate many open technologies in a production system.

6.4.1 Literal and Associative Types

This section discusses the feasibility of the inclusion of what was hitherto an open problem in geographic text analysis, i.e., the distinction between literal and associative toponyms, which can help answer the question: "*Is the toponym literalness distinction worth having?*" Figure 6.7 shows a map of the earlier news article from Section 6.3 marking the MediSys and the human annotations. The purple point markers denote the MediSys output, the green point markers are the literal gold toponyms, the yellow star circles are the associative gold

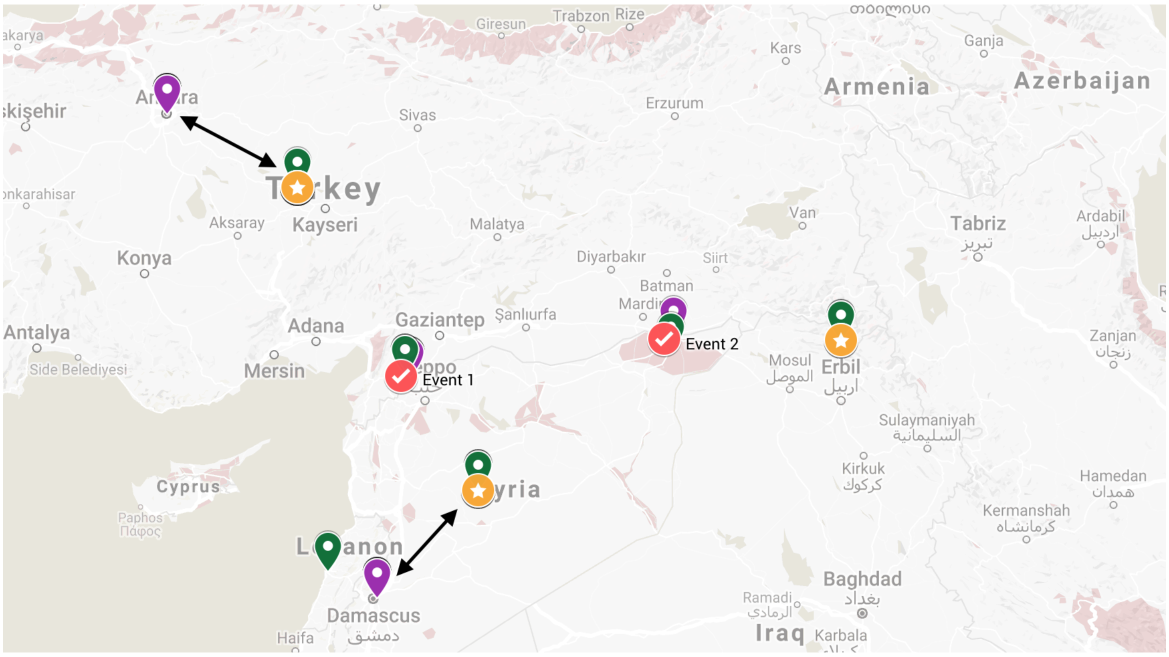


Fig. 6.7 A map view of the news article from Section 6.3.

toponyms and the red tick markers are the geographic foci of this news article, labelled *Event 1* and *Event 2*. The bidirectional black arrows show the differences in coordinate assignment policy, i.e. choosing to resolve *Turkey* and *Syria* to their capitals *Ankara* and *Damascus* instead of the countries' geographic centres. This shows the need for database alignment[6] prior to evaluating with geoparsers/datasets using different knowledge bases. In terms of toponym types, there is significant overlap between literal and associative toponyms with literal mentions coinciding with the actual event location. Associative toponyms do not add any false positives to the analysis of this document, however, it is easy to show these cases by adding the following sentence to the article: "The *American* and *Russian* presidents are negotiating on the best course of action together with their *French*, *British* and *German* counterparts". These associative toponyms would introduce many false positives making geolocation of this event more difficult.

We have seen that the F-Score decreases by 11 points for the two-label NER configuration thus it is important to establish the likely effect of this feature. Figure 6.8 shows how likely associative toponyms are to mislead a geoparser when they occur in the same article as other literal toponyms. Consider a short example: "The *Japanese* president is in *Botswana* this week." There are two toponyms in the 'news article'. If we naively try to geolocate this

[6] Provided as a convenience function on GitHub in the *dataset.py* file. The script does not cover the coordinate policy alignment as this is rare in geoparsing and the policies may not be standardised.

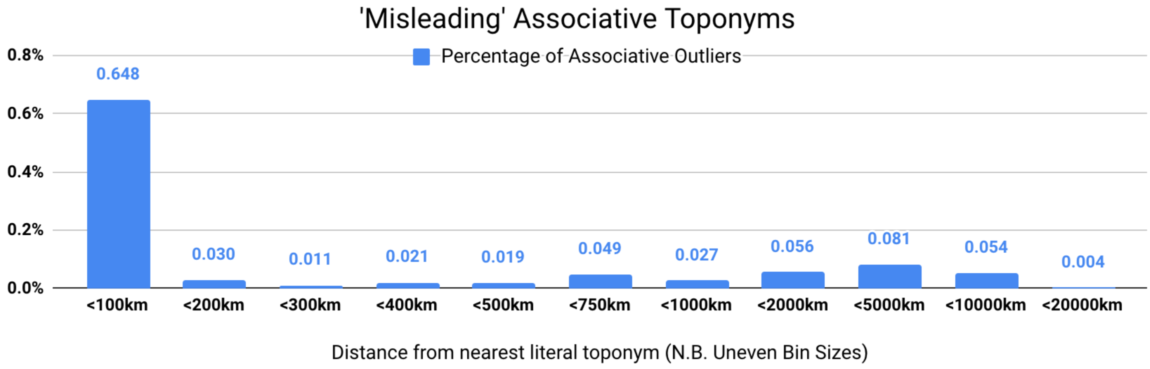


Fig. 6.8 Histogram showing the distribution of associative outliers, which are non-literal toponyms in the article that refer to locations *Xkm* from the nearest literal toponym.

document, we determine *Japan* and *Botswana* to be the event location, which is misleading (for *Japan*) with an error distance of around 13,500 kilometres.

The histogram above shows the percentage of associative toponyms that can result in such false positives and their respective error distances. It tells us that 65% of associative toponyms point to the same locations as the literals. Another ∼15% will be no more than 1000km away from the nearest literal. However, that lets ∼20% of associative toponyms mislead a surveillance system by not resolving the event location to physical places but concepts associated with them. In order to definitively test which configuration is more effective, we would need an extensive end-to-end task evaluation, which has not been conducted. However, we can conclude that despite the decreased geotagging scores of the two-label setup, there are benefits to event geolocation that may offset other losses in end-to-end tasks.

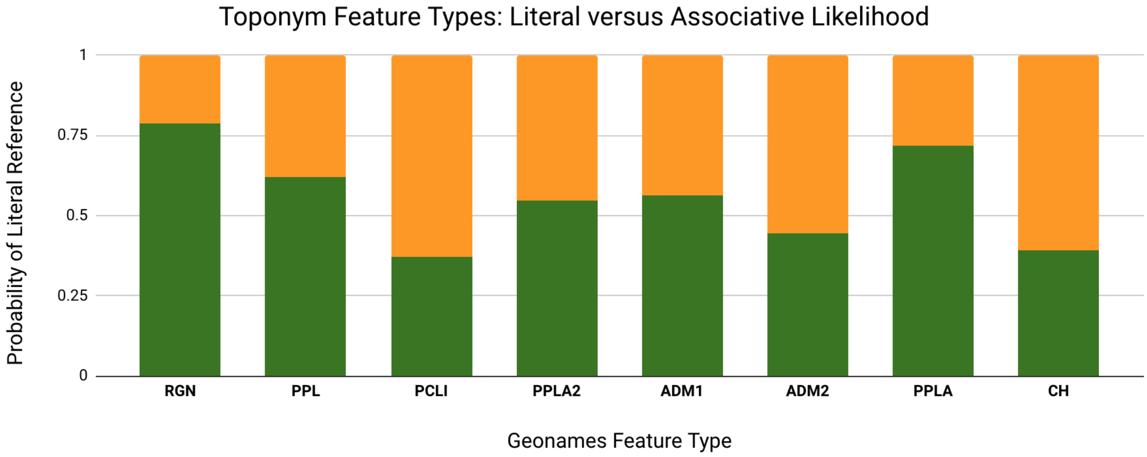
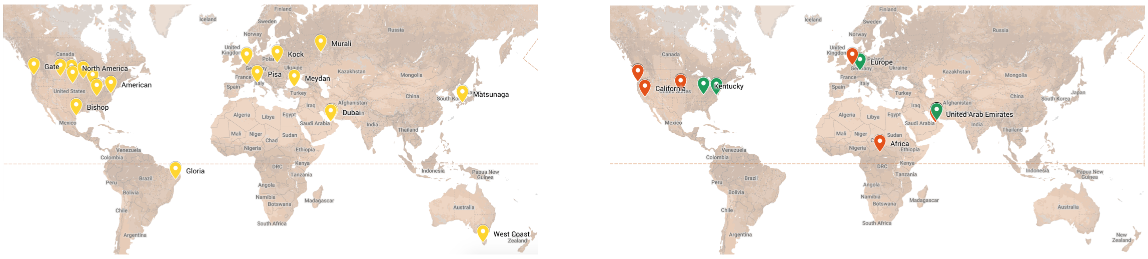


Fig. 6.9 Probability of literal vs. associative toponyms for frequent types.

Figure 6.9 shows toponyms of frequent feature types and their probability of being literal. We can see that countries (*PCLI*) are associative more often than not. Back in figure 6.3, we saw that PCLIs were the easiest to resolve, however, they were also more likely to mislead. On the other hand, capitals (*PPLA*) are more often used in a literal sense while also being resolved with a lower average error. Therefore, a simple and effective heuristic would be to *prioritise a capital* over a country to geolocate an event. As for the rest, we can generally conclude that the larger the toponym, the easier it is to geocode, but it is also more likely to mislead. Let us discuss another short example. Figure 6.10 shows the MediSys annotation on the left *(a)* and the human annotation on the right *(b)*. Event geolocation starts with correct *toponym identification*. This is significant since errors made at this stage propagate to geocoding and beyond. From the information in Figure 6.10a, assigning an event location to the article (full text in Appendix A.1) will be error-prone. Knowing the toponym mention frequency would be of limited impact as there are many repetitions on different continents.



(a) MediSys annotation of the news article.

(b) Gold annotation of the news article.

Fig. 6.10 Map views of the example news article. The event location is Dubai, UAE.

In Figure 6.10b, knowing the frequency *and* toponym type would make grounding this news event in geography relatively straightforward. With the repetitive false positives removed, "*Dubai*" is the obvious frequent literal type and also the human labelled geographic focus of the article. For the purposes of *event geolocation*, we may conclude that in the presence of the aforementioned challenges, the most promising *overall strategy* is to train a custom NER model on carefully annotated data. We also recommend incorporating a variant of FlexiGraph into the framework as we hypothesise that linking toponyms to trigger words *within a document* would have a high positive impact. We would expect the toponyms closest to "*Zika*" or "*terror attack*" to be the most indicative of the event's true location. One may then use the knowledge of toponym types to further prioritise certain locations, e.g. by modulating the population bias during toponym resolution. Completing our recommendations is the *literalness distinction*, which could provide yet another filtering mechanism for grounding a news event in geography.

Fast and Effective NER The NCRF++ model is an excellent demonstration of cheap and fast (over 1,000 sentences per second) Named Entity Recognition from limited data. Open-source annotated datasets such as TR-News, GeoVirus, LGL and others are already available and having tested NCRF++ on GeoWebNews, we found this framework to be highly effective. Therefore, it is now possible to selectively train a model for the desired annotation scheme/task and to suit custom system requirements. One may also extend the existing BRAT annotation to better resemble the target domain. In the pursuit of the best results, we tested (mostly Web-based demo versions of) more NLP frameworks than reported in the thesis and can conclude that *Google Cloud NLP* is by a good margin the *SOTA off-the-shelf NER tagger* available today (for a price). However, using open-source solutions and data, we have been able to achieve a *5-point improvement* in F-Score, which makes this type of framework one of our key recommendations.

6.5 Summary

In this chapter, we put the theory from chapters 2, 3, 4 and 5, evaluated intrinsically, into practice by using data from the MediSys news stream as a form of extrinsic evaluation. Our contributions focused mainly on the geoparsing pipeline, however, the post-geoparsing stage leaves significant room for research innovation in future work. These insights should primarily benefit the researchers at the Joint Research Centre in Ispra whose valuable feedback (Dr Jens Linge) was included. After describing the nature of the news feed articles, we discussed the usefulness of *toponym literalness* given the occurrence of various toponym types and the penalties incurred for different systems under different conditions. We have undertaken a detailed assessment of MediSys' geoparsing performance including error analysis, suggestions for improvements and comparisons with similar diversely sourced tools. Having ascertained the needs of system developers and researchers at the agency, we prototyped *FlexiGraph*, which is a simple method of transforming a short document into a graph that can adequately capture its semantic/syntactic structure. It can then serve as an additional filter for event geolocation by measuring the distance from trigger words to toponyms inside the article. Finally, we made proposals for the implementation of several future geoparsers, each with clearly stated advantages and likely downsides, taking into account the knowledge and insights from the full thesis.

There is further progress to be made to attain *human geoparsing performance*, i.e. an expert familiar with the task. We think that the lower amateur-level performance has already been surpassed. From our experience of having annotated several datasets and computed

IAAs, we estimate the geotagging human performance to be around the *F-Score=95* mark for frequent toponyms (countries, known regions and larger cities) and around the *F-Score=85* mark for less frequent toponyms, allowing for mostly recall errors. Geocoding accuracy for an expert is estimated at *near 100%* for frequent toponyms and *∼90%* for less known or more ambiguous toponyms, which can sometimes be *unresolvable without appropriate reference*. The scores will be worse for datasets of higher ambiguity. In addition to the suggestions for future work in the next chapter, we also encourage fellow researchers to extend annotation of some of our recommended datasets, e.g. to assign a *geographic focus to each article* for the evaluation of full pipeline systems. With regards to document geolocation (full text classification), we remind the reader of the cited work on Twitter from chapters 3 and 4. This is because news reports sometimes *do not contain any toponyms* but we still wish to try to locate their likely origin. For this task, we recommended starting with these papers [160, 159, 166, 161, 204, 11] for *full text geolocation*, regardless of the presence of any toponyms, which may indeed be of interest to EMM/MediSys developers and researchers.

Chapter 7

Conclusion

7.1 Synopsis

The thesis presents the latest geoparsing research with application to Automatic Disease Monitoring via breaking news analysis. Text documents frequently contain multiple references to geographic entities, which can be converted into sets of coordinates and used to track the geographic spread of news events. To this end, we have also undertaken exploratory work on event geolocation in order to identify more effective methods of using the geoparsed text. The contributions are almost entirely technical rather than epidemiological, which have been and continue to be actively studied as separate research topics. The research is presented at multiple levels of abstraction; from deep vertical work on Metonymy Resolution and Toponym Resolution to a broad empirical survey of prominent research. This is followed by a comprehensive critical discussion of the standard geoparsing evaluation framework, finally turning specific again by evaluating a real-time EU surveillance system to propose compelling avenues of future work. The thesis aims to serve as a convenient, comprehensive and practical entry point into the field, which we perceived was missing in the somewhat segmented and compartmentalised prior work. Hence it may solve any ’chicken-egg problem’ and encourage further participants into this domain.

The major contributions of the thesis cover six cohesive themes: *1. The Geoparsing Task and its Evaluation Framework* have been clarified, technically and linguistically reviewed and restated in multiple forms such as annotation scheme corrections, evaluation framework introduction, a new taxonomy of toponyms, performance metrics consolidation, analysis of available tools, replication suggestions, dataset characteristics review, knowledge base choice and alignment, hypothetical maximum performance per dataset/system and other process/task refinements. This will instil further rigour, fairness and clarity into geoparsing.

2. Metonymy Resolution with PreWin, which achieved near SOTA scores on the SemEval 2007 dataset (and our new MR dataset) using a minimalist neural approach in preference to the previously manually engineered features. *3. The Open-Source Datasets*, suitable for both training and evaluation, address the historical lack of high-quality resources that we hope will help accelerate geoparsing research. *4. The CamCoder + Map Vector* combination presents a novel way of generating new features encoding geographic knowledge hitherto unused. The CNN model achieved SOTA scores on three diverse datasets including our new MediSys dataset when compared with multiple systems and strong baselines. *5. The FlexiGraph* was prototyped as an intra-document analysis tool, addressing the needs of the MediSys engineers discussed during the visit to the Joint Research Centre in Italy. FlexiGraph has the potential to significantly increase overall event geolocation capability by computing the strength of the association between the target/trigger word(s) and the extracted locations. *6. A Detailed Evaluation* of MediSys' geoparsing capability was conducted with deep process insights and actionable system recommendations drawing interest from international biosurveillance organisations. The performance analysis allowed us to test the 'laboratory research' on real-time data and make proposals grounded in a specific application. We anticipate our empirical and highly pragmatic insights to inspire the development of the next generation geographic text analysis tools.

7.2 Discussion: Towards Robust Biosurveillance

Let us discuss the likely technical challenges and expected benefits of making current semi-supervised systems more robust, perhaps even transitioning to faster, weakly-supervised monitoring systems. By *more robust*, it is meant that known sources of errors will be substantially reduced. By *semi-supervised*, it is meant that the machine output is interpreted by humans before any conclusions, recommendations or decisions can be made, i.e. output needs further inspection to correct known errors. By *monitoring systems*, it is meant that topics of interest are retrieved using keyword searches of relevant streaming text documents, e.g. news articles. We then proceed to geoparse each news article to geolocate the event, in this instance, for the purposes of disease monitoring. We shall outline technical areas where semi-supervised systems tend to fall short and how NLP methods can alleviate these problems. N.B. Moderated systems are already robust but require painstaking and expensive human annotation. In this section, we aim to describe methods that will accelerate biosurveillance processing while preserving robustness.

Source	**Precision Errors**		**Recall Errors**	
Matrix	Semi-supervised	NLP Benefits	Semi-supervised	NLP benefits
Locations (Geoparsing)	Corrections are easy but slow.	Speed gains. Low error rate.	Corrections are costly *and* slow.	Speed gain and higher recall.
Keywords (Retrieval)	Corrections are easy and fast.	Negligible gains.	No discoveries of new diseases.	Can help discover emergent threats.

Table 7.1 The four quadrants of anticipated technical challenges and opportunities.

We know from previous error analyses that false positives and false negatives can be misleading for the successful geolocation of news events. *False positives* are not actual toponyms or existing diseases (for keyword searches). These introduce noise into the system possibly leading to false alarms. *False negatives*, however, are the more serious instances as users don't necessarily know how much information is missing from the output hence these errors go unnoticed (without excessive investment in detection). This is the problem of silent evidence, a type of survivorship bias in the data where we don't know what isn't included in the sample and how consequential that may be. Table 7.1 briefly summarises the sources of challenges and opportunities discussed in this section. Let us first describe the four quadrants of the table before determining the greatest risks of semi-supervised systems and discuss how to mitigate them using various NLP methods including geoparsing.

Location Precision errors are caused by false positives in the geoparsed output, entities incorrectly assumed to be toponyms. Toponym precision is a known problem for many systems, evidenced in our earlier analyses. In semi-supervised monitoring systems, these are easily excluded by humans, however, at a non-trivial cost to the user (Table 7.1). SOTA geoparsers have a single-digit error rate and a high throughput requiring only weak supervision.

Keyword Precision errors are cases where the monitored symptoms and disease names overlap with regular words. Ambiguity, however, is less of an issue with keyword precision as diseases do not frivolously feature in common discourse and do not typically overlap with generic proper nouns (unlike toponyms). Corrections are therefore easy and fast.

Location Recall errors, caused by false negatives, can be problematic as the true event location may be missing and unavailable for later processing or for presentation to the user. For semi-supervised systems, the option of each user reading 100's of relevant news articles is available but it requires a long and careful correction of missed toponyms. SOTA geoparsers are approaching 90% recall hence can provide fast processing with high accuracy.

Keyword Recall errors cause the news articles, which should have been retrieved, to be missing from the output. A relevant document that was not matched against a database of known diseases/symptoms does not mean we are not dealing with an emergent threat. New diseases, which have not yet been studied will not be detected, geoparsed, analysed and reported. More on the risk impact in the next subsection.

Toponym Resolution error rates for global streaming news are also in the single digits (Chapters 5 and 6). Deploying a SOTA geoparser for disease monitoring would accelerate processing speed with high accuracy. Users can skilfully perform geocoding error correction, however, the aim is to use technology to enable experts to focus on epidemic intelligence, decision making, communication and scientific research, not debugging the output.

7.2.1 Disease Monitoring, NLP and System Risks

Now that we understand the sources of the main geoparsing and keyword search errors, let us consider their costs and risks in a dynamic setting and offer possible solutions. We wish to anticipate the likely technical risks and suggest applied NLP research that may protect against the *tail risk* of black swan effects[1]. The magnitude of any risk is just as important as its probability. Even though the probability of emergent health threats is low, e.g. new disease strain or unseen symptoms for current diseases, their magnitude is of great significance. While keyword and location *precision errors* incur a financial penalty, the humanitarian risks of the system, in addition to inefficiency, will come from two sources: 1) keyword recall: events and diseases that should be monitored but get ignored; and 2) location recall: toponyms missed during geoparsing that will not be associated with a public health threat.

Assuming that streaming news articles have already been retrieved (typically by keyword matching), processed and stored, we first discuss how users of semi-supervised systems may search the database by location and/or by keyword. Location search is problematic because of recall errors, i.e. we cannot retrieve and analyse articles for public health threats if the event locations were not geoparsed. Keyword search can mitigate location recall problems *if* users regularly verify that toponyms are not missing from the system output, which is a type of crowdsourcing, i.e. distributing the workload among thousands of system users. A

[1] Read more about Black Swans in Nassim N. Taleb's book[182].

greater risk, however, comes from keyword recall, which may cause a failure to discover emerging health threats. Trigger words that are not being monitored will not recall relevant articles for examination. A compound search by keyword *and* location incurs both penalties. We might therefore expect semi-supervised monitoring systems to miss important cues or misidentify event locations as a result of these anticipated issues. However, we think this may not materialise as humans do much of the information/error filtering themselves thus a moderately performing system can be a useful and informative event monitoring resource for human operators, assuming their willingness to invest in regular error checking. The tasks humans are good at can compensate for the mistakes made by machines.

Location recall (third quadrant in Table 7.1) has been duly addressed in this thesis, reducing errors into the single digits. One way to mitigate the adverse impacts of keyword recall (fourth quadrant) would be to use Topic Modelling, of which the most popular type is Latent Dirichlet Allocation [17] and its later extension, Dynamic Topic Models [16]. These are generative probabilistic topic models that represent documents as mixtures of latent topics with each being a set of typically co-occurring words. Topics need not be semantically coherent or complete. Documents are represented as vectors of probabilities over some fixed number of latent topics. Dynamic topic models can be used to study the evolution of topics and their probabilities over time. Another variant of topic models is the Online Learning for LDA [83] allowing for efficient processing of a large number of *streaming documents*. This methodology would be especially suitable for a disease monitoring system.

In order to mitigate the aforementioned risks of keyword recall errors, we first need to compute the *class vectors* for each cluster of news articles using LDA. We may obtain the classes by clustering by disease, symptom or some other attribute/criterion, as an example, or use a single class. The generated feature vectors record the topic distribution for each document in each group and can be combined by averaging to produce the *class vector(s)*. If the incoming stream contains articles with *similar class vectors* that haven't triggered a keyword match, these would be subject to close scrutiny by EI experts to verify that we are not dealing with an emergent phenomenon that has not yet been named or registered as a search term. The annotated data could then be used to modulate the parameter value of the *similarity threshold*, i.e. which documents without known keywords are similar to the ones that do contain them. Increasing this parameter would present more documents for inspection and vice versa, depending on available resources. We think the adverse impact of keyword recall errors would be substantially reduced using this or a similar method. The paragraph serves as a practical illustration, an in-depth solution is reserved for future work.

7.3 Future Work

This section leverages my collective knowledge and experience to illuminate the most promising avenues for researchers wishing to extend the work in this thesis. The proposals presented in subsequent paragraphs and the outstanding "local" research problems outlined in previous chapters can be understood as the "*challenges*" part of the thesis title "Advances and *Challenges* of Geographic Analysis of Text with Application to Disease Monitoring".

Multilingual Applications Our research has been limited to monolingual contributions (including models and data available for English only). Our collaboration with the European Commission's Joint Research Centre showed the need for multilingual geoparsing advances. To that end, we propose an inexpensive and promising solution in a form of a Zero-Shot Transfer, which should be possible not only for Toponym Resolution but possibly Geotagging/NER. To enable this knowledge transfer, we were inspired by a brand new paper on *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond* [101], with a particular focus on the '*Beyond*' part. The methodology can be broadly distilled into two steps: *1.* Learn a single Encoder of stacked BiLSTMs on top of Byte Pair Embeddings using concatenated multilingual corpora (machine translation task). The concatenated outputs of the BiLSTMs generate the *sentence embedding*, encoding it in any of the 93 languages. *2.* Learn a single Decoder on *monolingual training data* (English) only, then transfer without modification to dozens of different languages. Figure 7.1 shows the architecture of the system. While a zero-shot cross-lingual transfer naturally causes accuracy to decrease compared to a separately trained model, the loss is on average only ∼5%, which is a strong result. This architecture could be straightforwardly adapted for *Multilingual Toponym Resolution* where we have plentiful training data, but only for a few languages. We

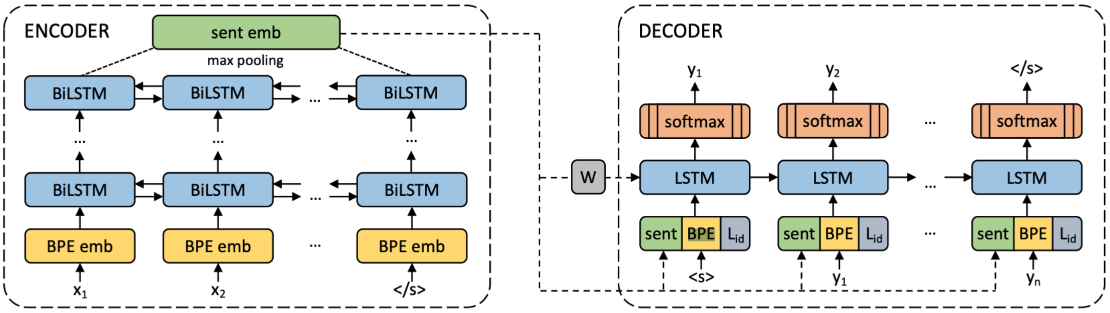


Fig. 7.1 Screenshot of the system architecture from the arxiv.org paper titled: *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond.*

only need to train the Decoder (a CNN classifier or similar) on the available data, then deploy for dozens of languages maintaining a high accuracy.

Representation and Transfer Learning We have seen great progress recently (2016-2018) in the field of representation learning. This ranges from subword-level representations such as FastText[2] classifiers [91, 20] to contextualised word vectors like ELMo[3] [153], pretrained language models with fine-tuning such as ULMFiT [85] to deep bidirectional transformer-based models for language understanding such as BERT[4] [44]. As our recommended and/or proposed models for geotagging and toponym resolution are also deep neural networks, we think that these improved representations could positively impact both stages. Our models have been prototyped with 'older' GloVe embeddings (2014-2015) for the initialisation of the first layer for both Geotagging and Geocoding. The latest representation and transfer learning breakthroughs have shown SOTA results on multiple NLP tasks, e.g. BERT by Google as well as improved scores on a range of tasks with ELMo contextualised embeddings. In the geoparsing domain, training data is still not easily available despite our additions thus *transfer learning* is the next logical option to explore. CamCoder uses CNNs to encode the salient features of the 400 words in the context of the toponym being resolved. This model should benefit from the encoding capability of the deep language models, which are particularly well suited for classification tasks, of which CamCoder is an example. For NER, the NCRF++ tagger can incorporate subword information into word embeddings to manage vocabulary size and obtain representations for infrequent/unseen words.

Word Probing Task Continuing with distributed representations for geoparsing, can we learn better word embeddings with an external task? Word vectors have proven to be very capable of capturing the linguistic context of words resulting in strong performance in downstream tasks. However, can these distributed word representations accurately (or even adequately) encode geographic information? We propose a simple word probing task that investigates whether we can build a shallow, interpretable geographic location classifier (linear model or a shallow MLP) using nothing but word vectors as input. More precisely, can we train a classifier to predict the coordinates for unseen toponyms given a word embedding as a feature vector? Evaluation would include reporting the loss, goodness of fit and/or accuracy scores for train, validation and test splits (for trainable and fixed word embeddings). We may use the Map Vectors from Chapter 4 as a competitive baseline. Map Vectors explicitly encode geographic information thus their word probing task performance would be expected

[2] https://fasttext.cc/
[3] https://allennlp.org/elmo
[4] https://github.com/google-research/bert

to be high. For extrinsic evaluation, we propose using the trained embeddings from the word probing task for Toponym Resolution with our results from Chapter 4 as a reference point. We hypothesise that embeddings trained on a geographic prediction task would outperform the standard GloVe vectors we used with the CNN-based geocoder. As a further experiment (sanity check), we would need to verify that these geographically enhanced embeddings still perform as expected on a semantic downstream task, e.g. NER or SRL. This is to ensure that the original contextual information has not adversely degraded during the word probing task. Understanding which information word embeddings already encode (or are able to encode with training/augmentation) would contribute to increased efficacy on specific downstream tasks such as geoparsing.

The Geographic Interpretation of directional/spatial phrases such as "*on the Turkish-Syrian border*" is another avenue of geoparsing research. The phrase currently gets resolved to the geographic centres of Turkey *and* Syria. However, the event location lies approximately at the centroid of these locations. Other variants of locational references include “*65 miles north of X*”, “*northeast Y*”, “*central X*”, "*on the border of X and Y*", "*between X and Y*", "*just outside X*" and many more. The correct interpretation of these phrases in a fully automatic event geolocation pipeline constitutes a difficult composite NLP task. From our experience and observation, these cases are relatively sparse in news text. Therefore, we assert that the priority is to develop ever more accurate geoparsing models as these provide the data for subsequent tasks, then research better geographic interpretation methods.

Deep Localised Geoparsing is a similar method to the previously cited work of Local-Global Lexicons [112] and warrants further development with the benefit of 8 years of subsequent research. The idea was to infer a local lexicon of toponyms for each *news source*, then use it to help resolve the typically dominant referents of toponyms such as "*Paris*", "*London*" or "*Moscow*" to their local counterparts. The other related work is a 2010 PhD thesis [7], which uses *rules* to estimate document location to help disambiguate each toponym. A machine learning extension could make a way for a type of *local toponym candidate weighting* with individual parameters for each toponym resolution decision. If we know the source of the document, we can restrict the toponym candidates to a smaller area and improve results. Geocoding is greatly simplified for country-level, region-level or other restricted area since ’shrinking the world’ automatically reduces the maximum permissible error. Therefore, building a single global system for all decision making may not be the optimal solution. Alternatively, we may learn parameters for each news source, which would allow for a flexible framework in cases where news sources change their publication content.

An evolving system that actively learns over time could control for these local publishing patterns/biases.

End-to-end Geoparser The preferred machine learning approach nowadays is *end-to-end* Deep Learning (DL), e.g., in speech recognition, object identification + counting, autonomous agent learning, machine translation and reinforcement learning in computer games. This means that the task is no longer separated into discrete steps with possibly hand-engineered features and explicit intermediate input representations throughout. Instead, modern DL systems endeavour to structure multiple processing steps into three parts: input data, a deep neural network and output (loss function). Future work may evaluate the feasibility of an end-to-end DL geoparser, merging geotagging with toponym resolution, even assigning the event location for the entire document. While this model would have to manage a more complex task, perhaps too complex, it would benefit from sharing input data, internal states and a shared training objective. This approach would remove any intermediate connections between components but would likely require a lot of annotated data, which is currently an issue. Error and performance analysis would be also more difficult due to the added model complexity. However, for experts in DL, this might be compelling experiment.

The final suggestion for future work is *Meta-Geoparsing*. We coined this term as it does not seem to exist in current literature. This proposal may also be seen as addressing what we see as the main limitation of our work. As strong as the evidence of our results has been so far, we think that meta-geoparsing has the potential to further improve the effectiveness of our technical advancements. The thesis has focused on *sequential, single document processing*, treating each toponym independently, however, once a pool of articles has been geoparsed, this creates clusters of data we can use for error mitigation, taking a global (meta) view of the task. We encourage experimentation with *clusters of 85-95% correct answers*, which is the SOTA for geotagging/geocoding on GeoWebNews from Chapter 5. The hypothesis is that pooling high probability decisions would help identify outliers and correct any local (single document/toponym) errors. Meta-geoparsing strategies could be developed using methods such as Monte Carlo simulation as we no longer require the textual part of documents, only the meta data. Annotating tens of thousands of documents is infeasible, therefore, a *simulated portfolio of documents by topic* could to be generated instead. We can use the relatively well-known and stable distributions of toponyms in news articles and in the source feed to generate a larger sample of *pseudo-geoparsed news articles*. This could be achieved by estimating the probability mass function for geotagging and the kernel density for toponym resolution. The parameters for the generator may include geotagging precision and recall, geocoding accuracy, the number, type and geographic distribution of toponyms inside documents for each topic, then using the annotated articles as the development and test sets.

Researchers would be able to test original meta-geoparsing strategies on hundreds of simulated clusters of pseudo-documents (*to study the forest, not just the trees*) and evaluate performance on annotated news articles. Strategies could range from simple heuristics such as choosing the most frequently occurring *top N toponyms* to more sophisticated statistical modelling or machine learning approaches. The workload for this experiment would likely be significant, however, we anticipate a high return on research effort. We hypothesise that studying the dynamics of large clusters of annotated documents will provide new scope for future research, which is missing in current literature, to our best knowledge.

7.4 Summary

In this chapter, we presented a synopsis of the thesis and made proposals for specific future work, which in our most informed judgement, is expected to produce the most return on invested research time. As we looked towards more robust biosurveillance, we have outlined the most likely risks stemming from the remaining technical challenges, some of which lie outside the scope of this research piece. We have also shown how NLP methods can significantly improve the speed and accuracy of execution, reducing time and costs but more importantly, help discover emergent public health threats and identify hitherto undetected event locations. For a future monitoring system to make decisions *autonomously*, we need more research into even more reliable system output. Even if a fully-automated end-to-end system is not feasible, we share the sentiments of the founders of Palantir Technologies[5], i.e. that technology is '*augmenting human intelligence, not replacing it*'.

In a compelling article[6] on the nature of algorithmic breakthroughs in AI (original article[7] by Alexander Wissner-Gross), it was hypothesised that *the availability of quality training data* rather than the algorithm's publication led to landmark breakthroughs such as Gary Kasparov's defeat by IBM's Deep Blue machine, human-level classification performance on ImageNet, rapid improvements in machine translation or Deep Mind's reinforcement learning agent playing Atari games at superhuman capability. All of these were achieved 3 years (on average) after quality data became available and 15 years after the algorithms

[5] https://www.palantir.com/about/

[6] http://www.spacemachine.net/views/2016/3/datasets-over-algorithms

[7] https://www.edge.org/response-detail/26587

were *introduced*. Recent breakthroughs have also been aided by a steep decline in the cost of computing power. We hope that our datasets (annotated by Cambridge linguists incl. computational) and accompanying resources will help accelerate deep learning research applied to geoparsing and other IE tasks, e.g. NER. At the beginning of this journey, we found that one of the obstacles to faster research was the lack of freely available resources, which we have duly provided for what we hope will be prolific future work. The technical improvements may free up human capital to perform further work to benefit humankind.

This has been a Department of Theoretical and Applied Linguistics, University of Cambridge PhD thesis on Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring. In the future, which almost certainly lies in the industry, please reach me at milangritta@gmail.com for questions, comments and suggestions. The thesis has already attracted the attention of international Public Health agencies and researchers, whom we hope our proposals and resources shall duly serve, if utilised. Thank you ever so much for investing your time in reading this deep research piece.

References

[1] Abdelkader, A., Hand, E., and Samet, H. (2015). Brands in newsstand: spatio-temporal browsing of business news. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 97. ACM.

[2] Acheson, E., De Sabbata, S., and Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320.

[3] Al-Olimat, H. S., Thirunarayan, K., Shalin, V., and Sheth, A. (2017). Location name extraction from targeted text streams using gazetteer-based statistical language models. *arXiv preprint arXiv:1708.03105*.

[4] Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J., and Daszak, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature communications*, 8(1):1124.

[5] Alonso, H. M., Pedersen, B. S., and Bel, N. (2013). Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 725–730.

[6] Anbalagan, B. and Valliyammai, C. (2016). # chennaifloods: Leveraging human and machine learning for crisis mapping during disasters using social media. In *High Performance Computing Workshops (HiPCW), 2016 IEEE 23rd International Conference on*, pages 50–59. IEEE.

[7] Andogah, G. (2010). *Geographically constrained information retrieval*. University Library Groningen][Host].

[8] Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. In *ISCRAM*.

[9] Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., and Tesconi, M. (2018). Crismap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers*, pages 1–19.

[10] Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM.

[11] Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., and Bahran, R. (2018). Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(3):34.

[12] Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., and Curran, J. R. (2009). Named Entity Recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 10–18. Association for Computational Linguistics.

[13] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

[14] Berlingerio, M., Calabrese, F., Di Lorenzo, G., Dong, X., Gkoufas, Y., and Mavroeidis, D. (2013). Safercity: a system for detecting and analyzing incidents from social media. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 1077–1080. IEEE.

[15] Bhargava, P., Spasojevic, N., and Hu, G. (2017). Lithium nlp: A system for rich information extraction from noisy user generated text on social media. *arXiv preprint arXiv:1707.04244*.

[16] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

[17] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[18] Blench, M. (2008). Global public health intelligence network (gphin). In *Proceedings of the conference of the American machine translation association (AMTA). Waikiki, Hawai'i: AMTA*.

[19] Bo, H., Cook, P., and Baldwin, T. (2012). Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING*, pages 1045–1062.

[20] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

[21] Brando, C., Frontini, F., and Ganascia, J.-G. (2016). REDEN: Named Entity Linking in Digital Literary Editions using linked datasets. *Complex Systems Informatics and Modeling Quarterly*, (7):60–80.

[22] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[23] Brun, C., Ehrmann, M., and Jacquet, G. (2007). XRCE-M: A hybrid system for named entity metonymy resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 488–491.

[24] Buchel, O. and Pennington, D. (2017). Geospatial analysis. *The SAGE Handbook of Social Media Research Methods*, pages 285–303.

[25] Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731.

[26] Buscaldi, D. et al. (2010). *Toponym disambiguation in information retrieval*. PhD thesis.

[27] Buscaldi, D. and Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313.

[28] Butler, J. O., Donaldson, C. E., Taylor, J. E., and Gregory, I. N. (2017). Alts, abbreviations, and akas: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, 13(1):58–81.

[29] Cacho, J. R. F. and Taghva, K. (2018). Reproducible research in document analysis and recognition. In *Information Technology-New Generations*, pages 389–395. Springer.

[30] Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM.

[31] Chinchor, N. (1998). Appendix b: Muc-7 test scores introduction. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

[32] Choi, J., Thomee, B., Friedland, G., Cao, L., Ni, K., Borth, D., Elizalde, B., Gottlieb, L., Carrano, C., Pearce, R., et al. (2014). The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, pages 27–31. ACM.

[33] Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

[34] Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., et al. (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

[35] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

[36] Coulson, S. and Oakley, T. (2003). Metonymy and conceptual blending. *Pragmatics and beyond - new series*, pages 51–80.

[37] Craswell, N. (2009). Mean reciprocal rank. In *Encyclopedia of Database Systems*, pages 1703–1703. Springer.

[38] da Graça Martins, B. E. (2008). *Geographically aware web text mining*. PhD thesis, Universidade de Lisboa (Portugal).

[39] de Bruijn, J. A., de Moel, H., Jongman, B., Wagemaker, J., and Aerts, J. C. (2018). Taggs: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis*, 2(1):2.

[40] De Oliveira, M. G., de Souza Baptista, C., Campelo, C. E., and Bertolotto, M. (2017). A gold-standard social media corpus for urban issues. In *Proceedings of the Symposium on Applied Computing*, pages 1011–1016. ACM.

[41] DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *AAAI*, pages 2382–2388.

[42] DeLozier, G., Wing, B., Baldridge, J., and Nesbit, S. (2016). Creating a novel geolocation corpus from historical texts. *LAW X*, page 188.

[43] DeLozier, G. H. (2016). *Data and Methods for Gazetteer Independent Toponym Resolution*. PhD thesis.

[44] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[45] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

[46] Dion, M., AbdelMalik, P., and Mawudeku, A. (2015). Big data: Big data and the global public health intelligence network (gphin). *Canada Communicable Disease Report*, 41(9):209.

[47] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.

[48] Dong, L., Wei, F., Sun, H., Zhou, M., and Xu, K. (2015). A hybrid neural model for type classification of entity mentions. In *IJCAI*, pages 1243–1249.

[49] Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL (2)*, pages 49–54.

[50] dos Santos, J. T. L. (2013). *Linking entities to Wikipedia documents*. PhD thesis, PhD thesis, Instituto Superior Técnico, Lisboa.

[51] Dredze, M., Osborne, M., and Kambadur, P. (2016). Geolocation for twitter: Timing matters. In *HLT-NAACL*, pages 1064–1069.

[52] Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45.

[53] Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1383–1392.

[54] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

[55] Farkas, R., Simon, E., Szarvas, G., and Varga, D. (2007). Gyder: maxent metonymy resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 161–164.

[56] Ferrés Domènech, D. (2017). Knowledge-based and data-driven approaches for geographical information access.

[57] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

[58] Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., and Kanza, Y. (2015). On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 127–136. ACM.

[59] Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*.

[60] Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157.

[61] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

[62] Gelernter, J. and Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667.

[63] Gentleman, R. and Lang, D. T. (2012). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*.

[64] Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

[65] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., and Petras, V. (2005). Geoclef: the clef 2005 cross-language geographic information retrieval track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 908–919. Springer.

[66] Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12.

[67] Gorfein, D. S. (2001). An activation-selection view of homograph disambiguation: A matter of emphasis. *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*, pages 157–173.

[68] Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

[69] Gregoric, A. Z., Bachrach, Y., and Coope, S. (2018). Named entity recognition with parallel recurrent neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 69–74.

[70] Gritta, M., Pilehvar, M. T., and Collier, N. (2018a). Which melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1285–1296.

[71] Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2017). Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1248–1259.

[72] Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018b). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

[73] Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

[74] Han, B. (2014). *Improving the utility of social media with Natural Language Processing*. PhD thesis.

[75] Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., et al. (2010). The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(1):7096.

[76] Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., Brownstein, J. S., Collier, N., Myers, M. F., George, D. B., et al. (2013). Global mapping of infectious disease. *Phil. Trans. R. Soc. B*, 368(1614):20120250.

[77] Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. *Using Corpora*, pages 185–188.

[78] Henrich, A. and Lüdecke, V. (2008). Determining geographic representations for arbitrary concepts at query time. In *Proceedings of the first international workshop on Location and the web*, pages 17–24. ACM.

[79] Herman Tolentino, M., Raoul Kamadjeu, M., Michael Matters PhD, M., Marjorie Pollack, M., and Larry Madoff, M. (2007). Scanning the emerging infectious diseases horizon-visualizing promed emails using epispider. *Adv Dis Surveil*, 2:169.

[80] Hirschman, L. (1998). The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech & Language*, 12(4):281–305.

[81] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[82] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

[83] Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

[84] Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

[85] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339.

[86] Hu, Y. (2018). Eupeg: Towards an extensible and unified platform for evaluating geoparsers. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, page 3. ACM.

[87] Hulden, M., Silfverberg, M., and Francom, J. (2015). Kernel density estimation for text-based geolocation. In *AAAI*, pages 145–150.

[88] Iso, H., Wakamiya, S., and Aramaki, E. (2017). Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750*.

[89] Jones, C., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., and Weibel, R. (2002). Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of 25th ACM Conference of the Special Interest Group in Information Retrieval*, pages 389–390. ACM.

[90] Joshi, M. and Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 313–316.

[91] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[92] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM*, 15:188–197.

[93] Kamalloo, E. and Rafiei, D. (2018). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1287–1296. International World Wide Web Conferences Steering Committee.

[94] Karimzadeh, M. (2016). Performance evaluation measures for toponym resolution. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, page 8. ACM.

[95] Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., Mitra, P., and MacEachren, A. M. (2013). Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73. ACM.

[Karimzadeh et al.] Karimzadeh, M., Pezanowski, S., MacEachren, A. M., and Wallgrün, J. O. Geotxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*.

[97] Katz, P. and Schill, A. (2013). To learn or to rule: two approaches for extracting geographical information from unstructured text. *Data Mining and Analytics 2013 (AusDM'13)*, 117.

[98] Kejriwal, M. and Szekely, P. (2017). Neural embeddings for populated geonames locations. In *International Semantic Web Conference*, pages 139–146. Springer.

[99] Kolkman, M. C. (2015). Cross-domain textual geocoding: the influence of domain-specific training data. Master's thesis, University of Twente.

[100] Laere, O. V., Schockaert, S., Tanasescu, V., Dhoedt, B., and Jones, C. B. (2014). Georeferencing wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS)*, 32(3):12.

[101] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

[102] Leek, J. T. and Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646.

[103] Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, Sheffield, UK*.

[104] Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417.

[105] Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers.

[106] Leveling, J. (2007). Fuh (fernuniversität in hagen): Metonymy recognition using different kinds of context for a memory-based learner. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 153–156.

[107] Leveling, J. and Hartrumpf, S. (2008). On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289–299.

[108] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

[109] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., and Lee, B.-S. (2012). Twiner: Named Entity Recognition in targeted Twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM.

[110] Lieberman, M. D. and Samet, H. (2011). Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 843–852. ACM.

[111] Lieberman, M. D. and Samet, H. (2012). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740. ACM.

[112] Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 201–212. IEEE.

[113] Lingad, J., Karimi, S., and Yin, J. (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web*, pages 1017–1020. ACM.

[114] Linge, J. P., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M., Al-Khudhairy, D., Yangarber, R., and van der Goot, E. (2010). Medisys.

[115] Linge, J. P., Steinberger, R., Weber, T., Yangarber, R., van der Goot, E., Al Khudhairy, D., and Stilianakis, N. (2009). Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13):19162.

[116] Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R., and Duprey, J. (2017). Reuters tracer: Toward automated news production using large scale social media data. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 1483–1493. IEEE.

[117] Liu, Y., Wei, F., Li, S., Ji, H., Zhou, M., and Wang, H. (2015). A dependency-based neural network for relation classification. *arXiv preprint arXiv:1507.04646*.

[118] Luo, Y., Sohani, A. R., Hochberg, E. P., and Szolovits, P. (2014). Automatic lymphoma classification with sentence subgraph mining from pathology reports. *Journal of the American Medical Informatics Association*, 21(5):824–832.

[119] Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S., and Clancy, S. (2010). Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

[120] Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools. In *LREC*.

[121] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

[122] Mantero, J., Belyaeva, J., and Linge, J. P. (2011). How to maximise event-based surveillance web-systems the example of ecdc/jrc collaboration to improve the performance of medisys. *Luxembourg: Publications Office of the European Union*.

[123] Mao, H., Thakur, G., Sparks, K., Sanyal, J., and Bhaduri, B. (2018). Mapping near-real-time power outages from social media. *International Journal of Digital Earth*, pages 1–15.

[124] Markert, K. and Nissim, M. (2002). Metonymy resolution as a classification task. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 204–213.

[125] Markert, K. and Nissim, M. (2007). Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 36–41. Association for Computational Linguistics.

[126] Markert, K. and Nissim, M. (2009). Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

[127] Màrquez, L., Villarejo, L., Martí, M. A., and Taulé, M. (2007). Semeval-2007 task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 42–47. Association for Computational Linguistics.

[128] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named Entity Recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

[129] Martínez-Rojas, M., del Carmen Pardo-Ferreira, M., and Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, 43:196–208.

[130] Matsuda, K., Sasaki, A., Okazaki, N., and Inui, K. (2015). Annotating geographical entities on microblog text. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 85–94.

[131] Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of CONLL*.

[132] Melo, F. and Martins, B. (2015). Geocoding textual documents through the usage of hierarchical classifiers. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, page 7. ACM.

[133] Melo, F. and Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1):3–38.

[134] Mesnil, G., He, X., Deng, L., and Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

[135] Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., and Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)*, 36(4):40.

[136] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

[137] Moncla, L. (2015). *Automatic reconstruction of itineraries from descriptive texts*. PhD thesis, Université de Pau et des Pays de l'Adour; Universidad de Zaragoza.

[138] Monteiro, B. R., Davis, C. A., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.

[139] Mota, C. and Grishman, R. (2008). Is this NE tagger getting old? In *LREC*.

[140] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[141] Nastase, V., Judea, A., Markert, K., and Strube, M. (2012). Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 183–193.

[142] Nastase, V. and Strube, M. (2009). Combining collocations, lexical and encyclopedic knowledge for metonymy resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 910–918.

[143] Nastase, V. and Strube, M. (2013). Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

[144] Nicolae, C., Nicolae, G., and Harabagiu, S. (2007). Utd-hlt-cg: Semantic architecture for metonymy resolution and classification of nominal relations. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 454–459.

[145] Nissim, M. and Markert, K. (2003a). Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 56–63.

[146] Nissim, M. and Markert, K. (2003b). Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 56–63.

[147] Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

[148] Overell, S. E. (2009). *Geographic information retrieval: Classification, disambiguation and modelling*. PhD thesis, Citeseer.

[149] Palmblad, M. and Torvik, V. I. (2017). Spatiotemporal analysis of tropical disease research combining europe pmc and affiliation mapping web services. *Tropical medicine and health*, 45(1):33.

[150] Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

[151] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[152] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

[153] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[154] Petroni, F., Raman, N., Nugent, T., Nourbakhsh, A., Panić, Ž., Shah, S., and Leidner, J. L. (2018). An extensible event extraction system with cross-media event resolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 626–635. ACM.

[155] Pilehvar, M. T. and Navigli, R. (2014). A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*.

[156] Poibeau, T. (2007). Up13: Knowledge-poor methods (sometimes) perform poorly. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 418–421.

[157] Porter, C., Atkinson, P., and Gregory, I. N. (2018). Health and disease in nineteenth-century newspaper: A textual and geographical analysis of a large newspaper corpus. *International Journal of Humanities and Arts Computing*, 12(2).

[158] Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

[159] Rahimi, A., Baldwin, T., and Cohn, T. (2017). Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *arXiv preprint arXiv:1708.04358*.

[160] Rahimi, A., Cohn, T., and Baldwin, T. (2016). pigeo: A python geotagging tool. *Proceedings of ACL-2016 System Demonstrations*, pages 127–132.

[161] Rahimi, A., Vu, D., Cohn, T., and Baldwin, T. (2015). Exploiting text and network context for geolocation of social media users. *arXiv preprint arXiv:1506.04803*.

[162] Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

[163] Rayson, P., Reinhold, A., Butler, J., Donaldson, C., Gregory, I., and Taylor, J. (2017). A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 9–15. ACM.

[164] Redman, T. and Sammons, M. (2016). Illinois named entity recognizer: Addendum to ratinov and roth'09 reporting improved results. Technical report, Technical report. http://cogcomp. cs. illinois. edu/papers/neraddendum-2016. pdf.

[165] Roberts, M. (2011). Germans, queenslanders and londoners: The semantics of demonyms. In *ALS2011: Australian Linguistics Society Annual Conference: Conference proceedings*.

[166] Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

[167] Rortais, A., Belyaeva, J., Gemo, M., Van der Goot, E., and Linge, J. P. (2010). Medisys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards. *Food Research International*, 43(5):1553–1556.

[168] Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

[169] Sang, K. and Tjong, E. (2002). Introduction to the conll-2002 shared task: Language-independent named entity recognition. Technical report, cs/0209010.

[170] Santos, J., Anastácio, I., and Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3):375–392.

[171] Sekine, S., Sudo, K., and Nobata, C. (2002). Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

[172] Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T. (2017). Semantic structure and interpretability of word embeddings. *arXiv preprint arXiv:1711.00331*.

[173] Serdyukov, P., Murdock, V., and Van Zwol, R. (2009). Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM.

[174] Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):11.

[175] Simon, R., Isaksen, L., Barker, E., and De Soto Cañamares, P. (2015). The Pleiades Gazetteer and the Pelagios Project.

[176] Speck, R. and Ngomo, A.-C. N. (2014). Ensemble learning for Named Entity Recognition. In *International Semantic Web Conference*, pages 519–534. Springer.

[177] Speriosu, M. and Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1466–1476.

[178] Steinberger, R., Pouliquen, B., and Van der Goot, E. (2013). An introduction to the europe media monitor family of applications. *arXiv preprint arXiv:1309.5290*.

[179] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

[180] Sufi, S., Hong, N. C., Hettrick, S., Antonioletti, M., Crouch, S., Hay, A., Inupakutika, D., Jackson, M., Pawlik, A., Peru, G., et al. (2014). Software in reproducible research: Advice and best practice collected from experiences at the collaborations workshop. In *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, page 2. ACM.

[181] Taleb, N. (2005). *Fooled by randomness: The hidden role of chance in life and in the markets*, volume 1. Random House Incorporated.

[182] Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*, volume 2. Random house.

[183] Tateosian, L., Guenter, R., Yang, Y.-P., and Ristaino, J. (2017). Tracking 19th century late blight from archival documents using text analytics and geoparsing. In *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*, volume 17, page 17.

[184] Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

[185] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

[186] Tobin, R., Grover, C., Byrne, K., Reid, J., and Walsh, J. (2010). Evaluation of georeferencing. In *proceedings of the 6th workshop on geographic information retrieval*, page 7. ACM.

[187] Van Laere, O., Quinn, J., Schockaert, S., and Dhoedt, B. (2014). Spatially aware term selection for geotagging. *IEEE transactions on Knowledge and Data Engineering*, 26(1):221–234.

[188] Volz, R., Kleb, J., and Mueller, W. (2007). Towards ontology-based disambiguation of geographical identifiers. In *I3*.

[189] Wallgrün, J. O., Hardisty, F., MacEachren, A. M., Karimzadeh, M., Ju, Y., and Pezanowski, S. (2014). Construction and first analysis of a corpus for the evaluation and training of Microblog/Twitter geoparsers. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, page 4. ACM.

[190] Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., and Pezanowski, S. (2017). Geocorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, pages 1–29.

[191] Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., and Pezanowski, S. (2018). Geocorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1):1–29.

[192] Wang, M. (2017). *Following the Spread of Zika with Social Media: The Potential of Using Twitter to Track Epidemic Disease*. PhD thesis, Concordia University.

[193] Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

[194] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

[195] Wing, B. and Baldridge, J. (2014). Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348.

[196] Wing, B. P. and Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964. Association for Computational Linguistics.

[197] Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

[198] Yang, J., Liang, S., and Zhang, Y. (2018). Design challenges and misconceptions in neural sequence labeling. *arXiv preprint arXiv:1806.04470*.

[199] Yang, J. and Zhang, Y. (2018). Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626*.

[200] Yarowsky, D. (2010). Word sense disambiguation. In *Handbook of Natural Language Processing, Second Edition*, pages 315–338. Chapman and Hall/CRC.

[201] Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.

[202] Zhang, H. (2004). The optimality of naive bayes. *AA*, 1(2):3.

[203] Zhang, J., Meng, F., Wang, M., Zheng, D., Jiang, W., and Liu, Q. (2016). Is local window essential for neural network based chinese word segmentation? In *China National Conference on Chinese Computational Linguistics*, pages 450–457. Springer.

[204] Zhang, W. and Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70.

[205] Zhang, W. and Gelernter, J. (2015). Exploring metaphorical senses and word representations for identifying metonyms. *arXiv preprint arXiv:1508.04515*.

[206] Zheng, X., Han, J., and Sun, A. (2017). A survey of location prediction on twitter. *arXiv preprint arXiv:1705.03172*.

[207] Zheng, X., Han, J., and Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*.

[208] Ziegeler, D. (2007). A word of caution on coercion. *Journal of Pragmatics*, 39(5):990–1028.

Appendix A

Supplementary Materials

A.1 Quoted News Text

Full News Article from Chapter 6

Dubai may have been hot and humid over the past week but thunder and snow rained on the city on Saturday night. And the royal blue silks, the colours of Godolphin, hung over the spectacular Meydan Racecourse as Thunder Snow won the 23 rd renewal of the $10 million Dubai World Cup. His Highness Sheikh Mohammed bin Rashid Al Maktoum, Vice-President and Prime Minister of the UAE and Ruler of Dubai, and Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum, Crown Prince of Dubai, along with other members of the Royal family after Thunder Snow won the Dubai World Cup. - Photo by Neeraj Murali. The four-year-old bay colt from Helmet, the mount of Christophe Soumillon, ended Emirati handler Saeed bin Suroor's two-year drought, in sensational style, winning the 10-furlong affair by a comprehensive five-and-three-quarter of a length over strong favourite, the USA's West Coast. Thunder Snow aced the five-horse American challenge to win in record time, putting Arrogate's 2:02:53 seconds to shade with a time of 2:01:38 seconds. It was Thunder Snow's seventh win in 18 starts but the biggest Group 1 victory of his fledging career. The win also increased Saeed bin Suroor's tally to an astonishing record eight at the Dubai World Cup, the most by any trainer. Thunder Snow's victory also brought up Suroor's 38 th win after Benbatl had triumphed in the Dubai Turf earlier on the night. And while Thunder Snow, Godolphin and Saeed bin Suroor all racked up the numbers, it was a life-long dream becoming a reality for Christophe Soumillon. The Belgian jockey notched his first Dubai World Cup in nine attempts. The 36-year-old's best result was the runner-up finish to California Chrome, on board Mubtaahij. Thunder Snow was drawn on an unfavourable Gate 10, right on the outside, but that didn't deter him as it was to be his night. Thunder Snow was out of the gates in a blink of an eye and with North America, his UAE rival, with whom he tussled in the Al Maktoum Challenge, missing the start, it made it all the more easier. But it was just one contender down and many more to go as the Americans lurked. Thunder Snow still had a job to do and he did it in some style. He kept West Coast, with whom he had exchanged the lead briefly at the start, at bay over the course of the 2000-metre contest. And Thunder Snow then went on to deny American legendary trainer Bob Baffert a second win on the trot and a fourth at the World Cup. "We won two years in a row and now we have come back and won it again. It is a great and a brilliant result," an elated Saeed bin Suroor said, moments after the race. "The jockey did a great job despite being drawn from Gate 10. What he has done, nobody has done. To take Thunder Snow from the Gate 10 and to take him to a position from where he can win is superb," added the Emirati, whose last win came with Prince Bishop, ridden by William Buick. Meanwhile, Soumillon revealed that a pre-race pep talk helped him win. "I don't know if it was Sheikh Mohammed's daughter, a little girl, she told me: 'It is small track and if you go in front then, you are going to win it.' I never thought I can do that running with that draw. He jumped quite well and I saw nobody trying to challenge me and then West Coast let me go. And when I arrived at the first corner, my horse was in front and, on the back straight, I was just cantering. He is a very funny horse and very talented but when he doesn't want to do, he doesn't and when he wants, it is just amazing. He was in great shape and pretty fit. He has shown that in Europe and last year in Kentucky. "It is difficult to say how I'm feeling because it has not sunk in. I had finished second one time but winning this was like a dream come true"; said Soumillon. SOUMILLON'S FIRST. 2010: 11 on Red Desire (Mikio Matsunaga), won by Gloria De Campeo. 2011: 7 on Musir (Mike de Kock), won by Victoire Pisa. 2012: 8 on Master Of Hounds (Mike de Kock), won by Monterosso. 2013: 8 on Treasure Beach (Mike de Kock), won by Animal Kingdom. 2014: 7 on Sanshaawes (Mike de Kock), won by African Story. 2015: 9 on Epiphaneia (Katsuhiko Sumii), won by Prince Bishop. 2016: 2 on Mubtaahij (Mike de Kock), won by California Chrome. 2017: 4 on Mubtaahij (Mike de Kock), won by Arrogate. 2018: Winner on Thunder Snow (Saeed bin Suroor)