Lang Resources & Evaluation (2020) 54:683–712
https://doi.org/10.1007/s10579-019-09475-3

ORIGINAL PAPER

A pragmatic guide to geoparsing evaluation

Toponyms, Named Entity Recognition and pragmatics

Milan Gritta[1] · **Mohammad Taher Pilehvar**[1] · **Nigel Collier**[1]

Published online: 19 September 2019
© The Author(s) 2019

Abstract 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.

Keywords Geoparsing · Toponym resolution · Geotagging · Geocoding · Named Entity Recognition · Machine learning · Evaluation framework · Geonames · Toponyms · Natural language understanding · Pragmatics

✉ Milan Gritta
mg711@cam.ac.uk

Mohammad Taher Pilehvar
mp792@cam.ac.uk

Nigel Collier
nhc30@cam.ac.uk

[1] Language Technology Lab (LTL), Department of Theoretical and Applied Linguistics (DTAL), University of Cambridge, 9 West Road, Cambridge CB3 9DP, UK

1 Introduction

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 Sect. 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 (Steinberger et al. 2013), Twitter-based disaster response (de Bruijn et al. 2018) and mapping (Avvenuti et al. 2018), spatio-temporal analysis of tropical research literature (Palmblad and Torvik 2017), business news analysis (Abdelkader et al. 2015), disease detection and monitoring (Allen et al. 2017) as well as analysis of historical events such as the Irish potato famine (Tateosian et al. 2017). 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 (Santos et al. 2015; DeLozier et al. 2015; Karimzadeh et al. 2013; Gritta et al. 2017b; Jurgens et al. 2015), 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 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.[1] A more detailed example analysed in Table 2 and a survey of previous work in Sect. 2.1 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.

[1] 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 Sect. 3.

Table 1 The interplay between context and semantics determines the type

Toponym type	NP semantics indicates	NP context indicates
Literals	Noun literal type	Literal type
Literal modifiers	Noun/adjectival literal	Literal or associative[a]
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

The top five are the literals, the bottom six are the associative types. Examples of each type can be found in Fig. 1

[a] NP **head** must be strongly indicative of a literal type, e.g.: "The *British* **weather** doesn't seem to like us today"

Table 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***"

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.[a]	Literal	Literal	Literal	Literal	Literal
K.of **Jordan** (PostMod)	Assoc.	Person	Literal	Literal	Literal	Organ.	Person
Iraqi militia (AdjMod)	Assoc.	Assoc.[a]	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.

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

[a] Can be derived from the API with a simple rule

Why a Pragmatic Guide Pragmatics (Pustejovsky 1991) 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 gives a background on Geoparsing, NER, GIE and GIR. We present the new taxonomy in Sect. 3, describing and categorising toponym types. In Sect. 4, we conduct a comprehensive review of current evaluation methods and justify the recommended framework. Finally, Sect. 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.

1.1 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 2720 toponyms were annotated with Geonames.[2] Data augmentation was evaluated with an extra 3460 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 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 188 km. 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.[3]

2 Background

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.

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

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

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 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.

2.1 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 (Sekine et al. 2002) was based only on *geographic feature types*[4] 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 (Sang and Tjong 2002) and 2003 (Tjong Kim Sang and De Meulder 2003) 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 (Doddington et al. 2004) 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 (Honnibal and Johnson 2015) use similar schemes (Weischedel et al. 2013) 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 Nothman et al. (2013) is used by NER taggers such as Illinois NER (Redman and Sammons 2016) and Stanford NLP Manning et al. (2014), featured in Table 2. The table shows the limited classification indicating weak and inconsistent usage of context.

The SpatialML (Mani et al. 2010) scheme is focused on spatial reasoning e.g. *X location north of Y*. Metonymy Markert and Nissim (2002), 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

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

which they're being used. The corpus is available at a cost of $500–$1000. The Message Understanding Conferences (MUC) (Hirschman 1998) 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 (Matsuda et al. 2015) has annotated toponyms including locational expressions such as parks, buildings, bus stops and facilities in 10,000 Japanese tweets. Systematic polysemy (Alonso et al. 2013) has been taken into account for *facilities*, but not extended to other toponyms. GeoCLEF Gey et al. (2005) (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 (Jones et al. 2002) (called SPIRIT) project, the focus of which was not the evaluation of Information Extraction or Toponym Semantics but classical GIR.

The WoTR corpus (DeLozier 2016) 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 Sect. 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 (Leveling and Hartrumpf 2008). Section 3 elaborates on the taxonomy of toponyms beyond metonymic cases. Geocorpora Wallgrün et al. (2018) is a Twitter-based geoparsing corpus with around 6000 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 (Lieberman et al. 2010) 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 (Lieberman and Samet 2011) 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 (Kamalloo and Rafiei 2018) 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). 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 (da Graça Martins 2008) 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) has no definition of toponyms and limits the analysis to *nouns only*. The GIR thesis (Andogah 2010) discusses the geographic hierarchy of toponyms as found *in gazetteers*, i.e. feature types instead of linguistic types. A toponym resolution thesis (Buscaldi et al. 2010) 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) provides no toponym taxonomy, nor does the Named Entity Linking thesis (dos Santos 2013). A GIR thesis (Moncla 2015) 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 Domènech 2017) 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), 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) 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 Sect. 2 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.

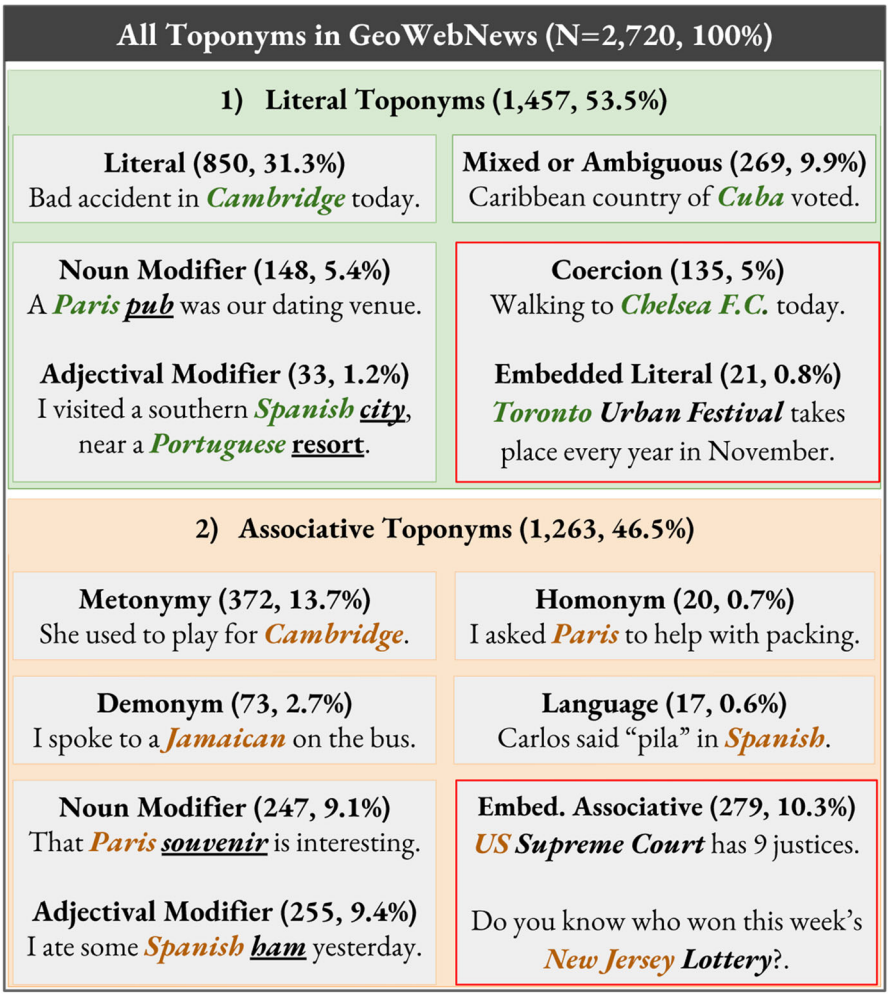

Fig. 1 The Pragmatic Taxonomy of Toponyms. A red border denotes *Non-Toponyms*. Classification algorithm: If the context indicates a literal or is ambiguous/mixed, then the type is literal. If the context is associative, then **a** for non-modifiers the toponym is associative **b** for modifiers, if the head is mobile and/or abstract, then the toponym is associative, otherwise it is literal

3 A pragmatic taxonomy of toponyms

While the evaluation metrics, covered in Sect. 4, are relevant only to geoparsing, Sects. 3 and 5 have implications for Core NLP tasks such as NER. In order to introduce the Toponym Taxonomy, shown in Fig. 1, 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 Sect.

2, 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 (Dong et al. 2015), for example. It is this interplay of semantics and context, seen in Table 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 1). We also assert that for applied NLP, it is sufficient and feasible to distinguish between literal and associative toponyms.

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.

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.

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 40 °C

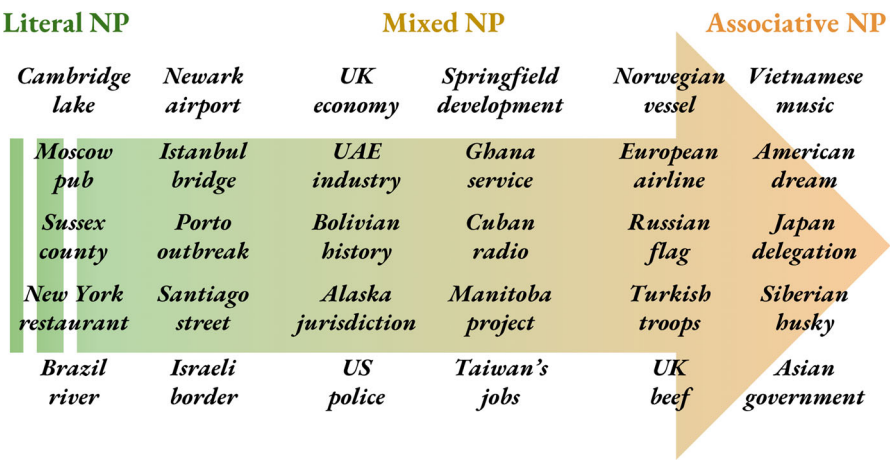

Fig. 2 Example noun phrases ranging from Literal to Mixed to Associative. The further to the right, the more 'detached' the *NP referent* becomes from its physical location. Literal heads tend to be *concrete* (elections, accidents) and *static* (buildings, natural features) while associative heads are 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*

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." Ziegeler (2007) 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* (Fig. 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 modifiers are frequently and incorrectly tagged as *nationalities or religious/political groups*[5] and sometimes ignored[6] altogether. Approximately 1 out of 10 adjectival modifiers is literal.

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 Roberts (2011) 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 Lieberman et al. (2010). 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 ..."

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 (Markert and Nissim 2007; Gritta et al. 2017a; Leveling and Hartrumpf 2008). 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%." 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.

[5] https://spacy.io/usage/ and http://corenlp.run/.

[6] http://services.gate.ac.uk/annie/ and IBM NLP Cloud in Table 2.

Adjectival modifiers are sporadically covered by NER taggers (Table 2) or tagging schemes (Hirschman 1998). 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 (Hearst 1991; Gorfein 2001) 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."

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 (DeLozier et al. 2015; Tobin et al. 2010; Karimzadeh et al. 2013; Wing and Baldridge 2014, 2011; Gritta et al. 2018). 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 (DeLozier et al. 2015; Gritta et al. 2017b). Therefore, we recommend the usage of this *strong baseline* as a *necessary component* of evaluation.

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 (Lieberman and Samet 2011).

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.

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 (Santos et al. 2015). 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 5–10 km) 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 (DeLozier 2016; Hulden et al. 2015), 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. Grover et al. (2010) and Tobin et al. (2010) used accuracy within 5 km, (Santos et al. 2015; Dredze et al. 2013) used accuracy at 5, 50, 250 km, related works on tweet geolocation (Speriosu and Baldridge 2013; Zheng et al. 2018; Han 2014; Roller et al. 2012) use accuracy at 161 km. We recommend the more lenient 161 km as it covers errors stemming from database misalignment. *Median Error* is a simple metric to interpret (Wing and Baldridge 2011; Speriosu and Baldridge 2013) but is otherwise uninformative as the error distribution is non-normal hence not recommended. The Area Under the Curve (Gritta et al. 2017b; Jurgens et al. 2015) is another coordinate-based metric, which follows in a separate subsection.

Set-based/categorical metrics and more specifically, the F-Score, has been used alongside coordinates-based metrics (Leidner 2008; Andogah 2010) 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 5 km (Andogah 2010; Lieberman and Samet 2012) to 10 miles (Kamalloo and Rafiei 2018; Lieberman et al. 2010) to all of the previous thresholds (Kolkman 2015) including 100 km and 161 km. In cases where WordNet has been used as the ground truth (Buscaldi et al. 2010) 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,

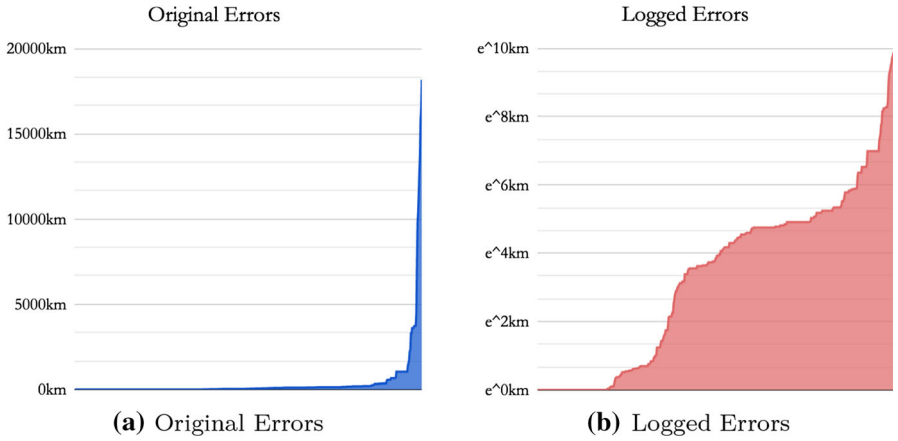

Fig. 3 Computing the area under the curve by integrating the *Logged Errors* in **b**. $AUC = 0.33$ is interpreted as 33% of the maximum geocoding error. 20,039 km is 1/2 of Earth's circumference

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 (Karimzadeh 2016; Craswell 2009). 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 (Gritta et al. 2017b; Jurgens et al. 2015). It is not to be confused with other AUC variants, which include the AUC of ROC, AUC for measuring blood plasma in Pharmacokinetics[7] 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.[8]

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

The original errors, which are highly skewed in Fig. 3a are scaled down using the *natural logarithm* resulting in Fig. 3b. The area under the curve divides into the total area of the graph to compute the final metric value. The logarithm decreases

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

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

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.

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 4. Using the Accuracy@161km, we can observe the *percentage* of correctly recognised and resolved toponyms to estimate the performance of the combined system.

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 between global gazetteers have been previously studied (Acheson et al. 2017). The most popular and open-source geoparsers and datasets do use Geonames[9] 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.[10] 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 (Matsuda et al. 2015) versus *Global Coverage*, which is the case with most geoparsers. It is not possible to fairly compare these two types of systems.

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

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

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 five to tenfolds that satisfy the i.i.d. requirement for a parametric test (Dror et al. 2018). 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 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 (Dietterich 1998). 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* (Wilcoxon 1945) for computational efficiency as the number of test samples across multiple datasets can be large (10,000+). Geocoding errors follow a power law distribution (Fig. 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.

4.5 Unsuitable datasets

Previous works in geoparsing (Leidner 2004, 2008; Andogah 2010; Santos et al. 2015) 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 (Hoffart et al. 2011) 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 (Mani et al. 2008, 2010) datasets are primarily focused on spatial expressions in natural language documents and are not freely available ($500–$1000 for a license[11]). Twitter datasets such as GeoCorpora (Wallgrün et al. 2018) experience a gradual decline in completeness as users delete their tweets and deactivate profiles. WoTR DeLozier et al. (2016) and CLDW Rayson et al. (2017) are suitable only for digital humanities due to their historical nature and localised coverage, which is problematic to resolve (Butler et al. 2017). CLUST (Lieberman and Samet 2011) is a corpus of clustered streaming news of global events, similar to LGL. However, it contains only 223 toponym annotations. TUD-Loc2013 (Katz and Schill 2013) provides incomplete coverage, i.e. no adjectival or embedded toponyms, however, it may generate extra training data with some editing effort.

4.6 Recommended datasets

We recommend evaluation with the following *open-source* datasets: (1) WikToR (Gritta et al. 2017b) 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) (Lieberman et al. 2010) is a global collection of local news articles (Geonames coordinates) and likely the most frequently cited geoparsing dataset. (3) GeoVirus (Gritta et al. 2018) is a WikiNews-based geoparsing dataset centred around disease reporting (Wikipedia coordinates) with global coverage though without adjectival toponym coverage. (4) TR-NEWS (Kamalloo and Rafiei 2018) 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 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

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

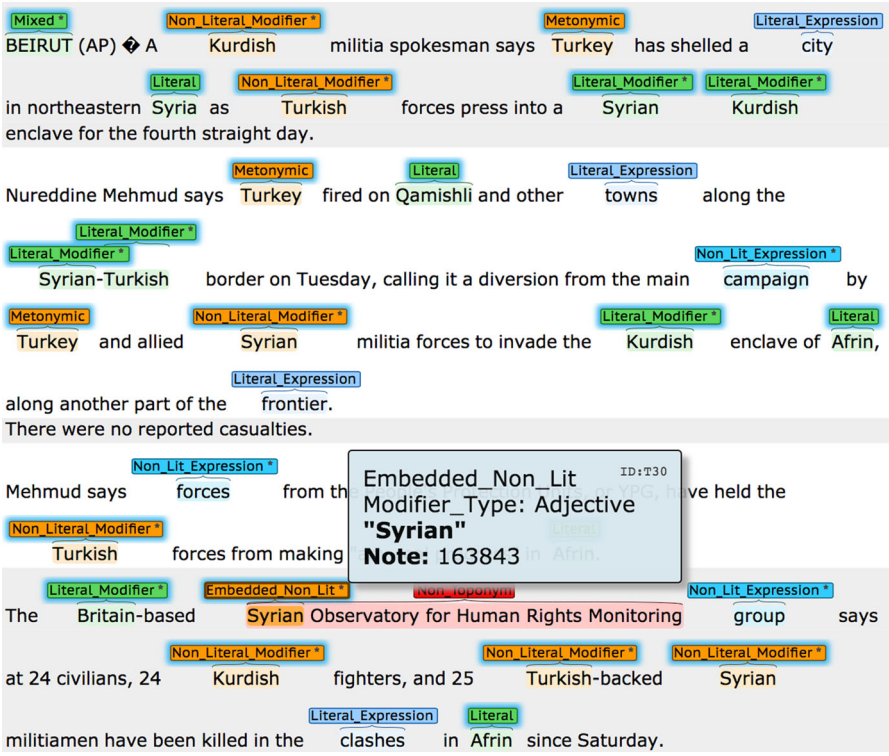

Fig. 4 A GeoWebNews article. An asterisk indicates an attribute, either *a modifier_type* [Adjective, Noun] and/or *a non_locational* [True, False]

collaboration with the European Union's Joint Research Centre,[12] collected during 1st-8th April 2018 from the European Media Monitor (Steinberger et al. 2013) using a wide range of multilingual trigger words/topics.[13] We then randomly selected exactly one article from each domain (English language only) until we reached 200 news stories. We also share the BRAT (Stenetorp et al. 2012) 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 (Gritta et al. 2018), develop and evaluate Machine Learning models for sequence tagging and classification, geographic information retrieval, even used in a Semantic Evaluation (Màrquez et al. 2007) 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.

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

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

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.[14] An annotated article sample can be viewed in Fig. 4. 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,[15] which implements the MUC-7 scoring scheme (Chinchor 1998). 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.

5.2 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 (Matsuda et al. 2015) 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 2401. We did not annotate the articles' geographic focus as was done for Twitter (Eisenstein et al. 2010; Roller et al. 2012) and Wikipedia (Laere et al. 2014).

5.3 Evaluation

Sections 3 and 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,[16] the Edinburgh Geoparser (Grover et al. 2010) and our

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

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

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

own CamCoder (Gritta et al. 2018) resolver as the main geoparsing benchmarks. We have also considered GeoTxt (Karimzadeh et al. 2013), 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 (Gritta et al. 2017a, b).

5.3.1 Geotagging GeoWebNews

For toponym extraction, we selected the two best models from Table 2, Google Cloud Natural Language[17] and SpacyNLP.[18] We then trained an NCRF++ model (Yang and Zhang 2018), which is an open-source Neural Sequence Labeling Toolkit.[19] We evaluated models using fivefold Cross-Validation (40 articles per fold, 4 train and 1 test fold). Embeddings were initialised with 300D vectors[20] from GloVe (Pennington et al. 2014) 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. The results are shown in Table 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.

Geotagging with two labels (physical location versus associative relationship) was evaluated with a custom NCRF++ model. The mean F-Score over fivefolds was 77.6 ($\sigma = 1.7$), 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.

5.3.2 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 $\sim 8\%$ of the total, without an entry in Geonames and often requiring a reference to additional

[17] https://cloud.google.com/natural-language/.

[18] https://spacy.io/usage/linguistic-features.

[19] https://github.com/jiesutd/NCRFpp.

[20] Common Crawl 42B—https://nlp.stanford.edu/projects/glove/.

Table 3 Geotagging F-Scores for GeoWebNews featuring the best performing models

NER model/geoparser	Precision	Recall	F-Score
NCRF++ (literal and 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**

The NCRF++ models' scores were averaged over fivefolds ($\sigma = 1.2$–1.3)

Table 4 Toponym resolution scores for the GeoWebNews data

Setup/description	Mean Err	Acc@161km	AUC	# of toponyms
SpacyNLP + CamCoder	**188**	**95**	**0.06**	1547
SpacyNLP + Population	210	**95**	0.07	1547
Oracle NER + CamCoder	232	94	**0.06**	2401
Oracle NER + Population	250	94	0.07	2401
Yahoo! Placemaker[a]	203	91	0.09	1444
Edinburgh Geoparser[a]	338	91	0.08	1363

[a] This geoparser provides both Geotagging and Geocoding steps.

resources. (b) demonyms, languages and homonyms, accounting for $\sim 4\%$ of toponyms as these are not locations hence do not have coordinates. The final count was 2401 ($\sim 88\%$) 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 2401 toponyms. We have combined these NER methods with the CamCoder (Gritta et al. 2018) default model.[21] 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 4, we found that a $\sim 50\%$ 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 (Gritta et al. 2017b) can be more challenging with errors 2–5 times greater. When provided with a database name for each

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

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 1547 correctly recognised toponyms were resolved to within 161 km. 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 nationalities, religious or political groups. Such lexicon could be assembled with a gazetteer-based statistical n-gram model such as Al-Olimat et al. (2017) that uses multiple knowledge bases or a rule-based system (Volz et al. 2007). For unknown toponyms, approximating the geographic representation from places that co-occur with it in other documents (Henrich and Lüdecke 2008) 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.

5.3.3 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 2720 toponyms to double or triple the training data size, two additional lexical features (*NP heads*) were annotated, denoted *Literal Expressions* and *Associative Expressions*.[22] 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 = 2037) *context* and literal (N = 1697) versus associative (N = 1763) *heads*, indicated by a binary *non-locational* attribute. These two interchangeable components give us multiple permutations from which to generate a larger training dataset[23] (see Fig. 5 for an example). The associative expressions are deliberately dominated by ORG-like types because this is the most frequent metonymic pair (Alonso et al. 2013).

Table 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.

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

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

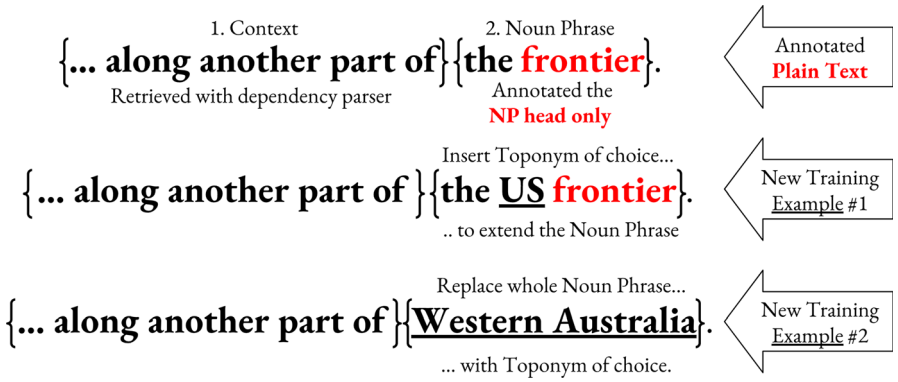

Fig. 5 An augmentation of a literal training example. An associative augmentation equivalent might be something like **{The deal was agreed by} {the chief engineer.}** replacing "the chief engineer" by a toponym

Table 5 F-Scores for NCRF++ models with fivefold cross-validation

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

No improvement was observed for the augmented or ensemble setups over baseline

6 Conclusions

6.1 Future work

Geoparsing is a special case of NER and often the initial step of an information extraction pipeline used by downstream applications. A detailed use case of the benefits of geoparsing and our pragmatic taxonomy of toponyms can be seen in Chapter 6 (page 95) of this PhD thesis (Gritta 2019). Geoparsing is a key step for the monitoring of the spread of public health threats such as epidemics and food-borne diseases using public news text. The chapter shows how fine-grained toponym extraction enables a deeper understanding and classification of geographic events using deep learning models with SOTA performance, significantly improving upon previous approaches. This methodology lets researchers automatically learn about entities associated with particular geographic areas. The ideas proposed in our paper can therefore enable a more accurate analysis of geographic events described in free text. Whether it is public health risks or other domains of interest, in the age of Big Data, there is a need for automated information processing of relevant events at scale.

We also expect to see more (1) reproduction/replication studies to test and/or revise evaluation setups, (2) dataset/model probing to test the validity of SOTA results, and (3) the annotation of multilingual and multi-domain resources for a wider range of tasks. Examples include a recently published paper on Twitter *user*

geolocation (Mourad et al. 2019) where the authors provide a critical review of current metrics, systems, datasets and SOTA claims. Similar to our work, the authors also recommend the use of several metrics for a holistic evaluation of user geolocation. Another consideration that applies beyond geoparsing is the construction of standard dataset splits for evaluation, investigated in 'We need to talk about standard splits' (Gorman and Bedrick 2019). The authors reproduced several SOTA part-of-speech models evaluated on standard 80-10-10 train-dev-test splits. However, when the splits were *randomly* generated, the SOTA rankings were not reliably reproduced. With that in mind, our practical guide to geoparsing evaluation complies with this recommendation as the cross-validation was performed with fivefolds generated from randomly sampled news articles.

It is critical that automatic evaluation is closely aligned with human evaluation and that this is periodically examined as we have done in this paper. Incorrectly structured datasets can also produce misleading comparisons with human performance. In 'Probing Neural Network Comprehension of Natural Language Arguments' (Niven and Kao 2019), the authors carefully examined BERT's (Devlin et al. 2018) peak performance on the argument reasoning task. The 77% accuracy was *only 3 points* below an untrained human baseline. However, it transpired that this performance came from exploiting the dataset's patterns, rather than the model's language understanding ability. The authors then created an adversarial version by removing those regularities resulting in just 53% accuracy, slightly above random choice. It is therefore prudent to ensure the robustness of evaluation and caution against any premature claims of near-human or superhuman performance.[24]

In Sect. 5, we introduced a new dataset for fine-grained geoparsing. However, we also encourage future efforts to be focused on corrections to existing datasets (with the consultation of expert linguists, if possible) such as CoNLL 2003 (Tjong Kim Sang and De Meulder 2003). Many models still benchmark their performance on the original (non-random) splits (Yadav and Bethard 2018), for example at COLING (Yang et al. 2018) and ACL (Gregoric et al. 2018). A survey/review could keep the original 3-class annotation, utilise our taxonomy to make the dataset suitable for geoparsing evaluation or even extend the taxonomy to other NER classes. An example of a dataset correction is MultiWOZ 2.1 (Eric et al. 2019), which is frequently used for training and evaluation of dialogue systems. The authors made changes to over 32% of state annotations across 40% of dialogue turns, which is a significant correction to the original dataset (Budzianowski et al. 2018). The final future work proposal is a Semantic Evaluation task in the Information Extraction track to close the gap to human (expert) baselines, almost 100% for geocoding (95% for SOTA) and around 97 F-Score for geotagging (87 for SOTA). GeoWebNews is most suitable for sequence labelling evaluation of the latest machine learning models. It comes with an added constraint of limited training samples, which could be overcome with transfer learning via pretrained language models such as BERT or ELMo (Peters et al. 2018).

[24] https://gluebenchmark.com/leaderboard/.

6.2 Closing thoughts

The Principle of Wittgenstein's Ruler from Nassim N. Taleb's book, Fooled by Randomness (Taleb 2005) deserves a mention as we reflect on the previous paragraphs. 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 field of NLP and beyond, this translates into: "*Unless you have confidence in the reliability of the evaluation, if you use the tools (data, metrics, splits, etc.) to evaluate models, you may also be using the models to evaluate the tools.*" We must pay close attention to the representativeness of the evaluation methods. It is important to ask whether models 'successfully' evaluated with tools that do not *closely mirror* real-world conditions and human judgement is the goal to aim for in NLP.

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 (Goodman et al. 2016; Cacho and Taghva 2018), we shared the annotation framework, recommended datasets and any tools/code 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 holistic and intuitive performance assessment. As we conclude this section, here are the recommended evaluation steps.

1. Review (and report) important geoparsing considerations in Sect. 4.4.
2. Use a proprietary or custom NER tagger to extract toponyms using the recommended dataset(s) as demonstrated in Sect. 5.3.1.
3. Evaluate geotagging using F-Score as the recommended metric and report statistical significance with McNemar's Test (Sect. 4.4).
4. Evaluate toponym resolution using Accuracy@161, AUC and Mean Error as the recommended metrics, see Sect. 5.3.2 for an example.
5. *Optional*: Evaluate geocoding in "laboratory setting" as per Sect. 5.3.2.
6. Report the number of toponyms resolved and the statistical significance using the Wilcoxon Signed-Rank Test (Sect. 4.4).

Acknowledgements We gratefully acknowledge the funding support of the Natural Environment Research Council (NERC) PhD Studentship (Milan Gritta NE/M009009/1), EPSRC (Nigel Collier EP/M005089/1) and MRC (Mohammad Taher Pilehvar MR/M025160/1 for PheneBank). We also

acknowledge Cambridge University linguists Mina Frost and Qianchu (Flora) Liu for providing expertise and verification (IAA) during dataset construction/annotation.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Abdelkader, A., Hand, E., & 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* (p. 97). New York: ACM.

Acheson, E., De Sabbata, S., & 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.

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

Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., et al. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature communications*, *8*(1), 1124.

Alonso, H. M., Pedersen, B. S., & 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)* (vol. 2, pp. 725–730).

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

Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. (2018). Crismap: A big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers*, *20*, 1–19.

Budzianowski, P., Wen, T. H., Tseng, B. H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). *Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling*. arXiv preprint arXiv:1810.00278.

Buscaldi, D., et al. (2010). *Toponym disambiguation in information retrieval*. Ph.D. thesis.

Butler, J. O., Donaldson, C. E., Taylor, J. E., & 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.

Cacho, J. R. F., & Taghva, K. (2018). Reproducible research in document analysis and recognition. *Information technology-new generations* (pp. 389–395). Berlin: Springer.

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.

Craswell, N. (2009). Mean reciprocal rank. *Encyclopedia of database systems* (pp. 1703–1703). Berlin: Springer.

da Graça Martins, B. E. (2008). *Geographically aware web text mining*. Ph.D. thesis, Universidade de Lisboa (Portugal).

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

DeLozier, G. H. (2016). *Data and methods for gazetteer independent toponym resolution*. Ph.D. thesis.

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Association for the advancement of artificial intelligence* (pp. 2382–2388).

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a novel geolocation corpus from historical texts. *LAW X* (p. 188).

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

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

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *International conference on language resources and evaluation* (vol. 2, p. 1).

Dong, L., Wei, F., Sun, H., Zhou, M., & Xu, K. (2015). A hybrid neural model for type classification of entity mentions. In *International joint conferences on artificial intelligence* (pp. 1243–1249).

dos Santos, J. T. L. (2013). *Linking entities to wikipedia documents*. Ph.D. thesis, Instituto Superior Técnico, Lisboa.

Dredze, M., Paul, M. J., Bergsma, S., & 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)* (vol. 23, p. 45).

Dror, R., Baumer, G., Shlomov, S., & 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)* (vol. 1, pp. 1383–1392).

Eisenstein, J., O'Connor, B., Smith, N. A., & 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* (pp. 1277–1287). Stroudsburg: Association for Computational Linguistics.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., & Hakkani-Tur, D. (2019). *Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines*. arXiv preprint arXiv:1907.01669.

Ferrés Domènech, D. (2017). Knowledge-based and data-driven approaches for geographical information access. Universitat Politècnica de Catalunya. http://hdl.handle.net/2117/114615.

Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., & 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* (pp. 908–919). Berlin: Springer.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12.

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* (pp. 157–173). Washington: American Psychological Association.

Gorman, K., & Bedrick, S. (2019). We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, (pp. 2786–2791). Association for Computational Linguistics, Florence, Italy. https://www.aclweb.org/anthology/P19-1267.

Gregoric, A. Z., Bachrach, Y., & 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)* (vol. 2, pp. 69–74).

Gritta, M. (2019). *Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring*. Ph.D. thesis, University of Cambridge.

Gritta, M., Pilehvar, M. T., & Collier, N. (2018). 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).

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017a). 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).

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017b). What's missing in geographical parsing? *Language Resource Evaluation*, *52*, 603–623.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., et al. (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.

Han, B. (2014). *Improving the utility of social media with natural language processing*. Ph.D. thesis.

Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre f07' the New OED and Text Research* (pp. 1–22). Oxford

Henrich, A., & Lüdecke, V. (2008). Determining geographic representations for arbitrary concepts at query time. In *Proceedings of the first international workshop on Location and the web* (pp. 17–24). New York: ACM.

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

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., et al. (2011). In *Robust disambiguation of named entities in text* (pp. 782–792). Stroudsburg: Association for Computational Linguistics.

Honnibal, M., & 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* (pp. 1373–1378). Lisbon: Association for Computational Linguistics. https://aclweb.org/anthology/D/D15/D15-1162.

Hulden, M., Silfverberg, M., & Francom, J. (2015). Kernel density estimation for text-based geolocation. In *Association for the advancement of artificial intelligence* (pp. 145–150).

Jones, C., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., & 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* (pp. 389–390). New York: ACM.

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

Kamalloo, E., & Rafiei, D. (2018). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 world wide web conference on world wide web, international world wide web conferences steering committee* (pp. 1287–1296).

Karimzadeh, M. (2016). Performance evaluation measures for toponym resolution. In *Proceedings of the 10th workshop on geographic information retrieval* (p. 8). New York: ACM.

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

Katz, P., & 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*,

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

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

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on geographic information retrieval*, Sheffield, UK.

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

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

Lieberman, M. D., & 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* (pp. 843–852). New York: ACM.

Lieberman, M. D., & 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* (pp. 731–740). New York: ACM.

Lieberman, M. D., Samet, H., & 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)* (pp. 201–212). New York: IEEE.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., et al. (2010). Spatialml: Annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, *44*(3), 263–280.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., & Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools. In: *International conference on language resources and evaluation*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations* (pp. 55–60). http://www.aclweb.org/anthology/P/P14/P14-5010.

Markert, K., & 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* (pp. 204–213). Stroudsburg: Association for Computational Linguistics.

Markert, K., & Nissim, M. (2007). Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the 4th international workshop on semantic evaluations* (pp. 36–41). Stroudsburg: Association for Computational Linguistics.

Màrquez, L., Villarejo, L., Martí, M.A., & Taulé, M. (2007). Semeval-2007 task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of the 4th international workshop on semantic evaluations* (pp. 42–47). Stroudsburg: Association for Computational Linguistics.

Matsuda, K., Sasaki, A., Okazaki, N., & Inui, K. (2015). Annotating geographical entities on microblog text. In *Proceedings of the 9th linguistic annotation workshop* (pp. 85–94).

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

Mourad, A., Scholer, F., Magdy, W., & Sanderson, M. (2019). *A practical guide for the effective evaluation of twitter user geolocation*. arXiv preprint arXiv:1907.12700.

Niven, T., & Kao, H. Y. (2019). *Probing neural network comprehension of natural language arguments*. arXiv preprint arXiv:1907.07355.

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual Named Entity Recognition from wikipedia. *Artificial Intelligence*, *194*, 151–175.

Overell, S. E. (2009). *Geographic information retrieval: Classification, disambiguation and modelling*. Ph.D. thesis, Citeseer.

Palmblad, M., & 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.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

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

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, *17*(4), 409–441.

Rayson, P., Reinhold, A., Butler, J., Donaldson, C., Gregory, I., & 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* (pp. 9–15). New York: ACM.

Redman, T., & 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.

Roberts, M. (2011). Germans, queenslanders and londoners: The semantics of demonyms. In *ALS2011: Australian linguistics society annual conference: conference proceedings*.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & 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* (pp. 1500–1510). Stroudsburg: Association for Computational Linguistics.

Sang, K., & Tjong, E. (2002). Introduction to the conll-2002 shared task: Language-independent Named Entity Recognition. Technical report. cs/0209010.

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

Sekine, S., Sudo, K., & Nobata, C. (2002). Extended named entity hierarchy. *LREC*.

Speriosu, M., & Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. *ACL*, *1*, 1466–1476.

Steinberger, R., Pouliquen, B., & Van der Goot, E. (2013). *An introduction to the Europe media monitor family of applications*. arXiv preprint arXiv:1309.5290.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & 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* (pp. 102–107). Stroudsburg: Association for Computational Linguistics.

Taleb, N. (2005). *Fooled by randomness: The hidden role of chance in life and in the markets* (Vol. 1). New York: Random House Incorporated.

Tateosian, L., Guenter, R., Yang, Y. P., & 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* (vol. 17, p. 17).

Tjong Kim Sang, E. F., & 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* (pp. 142–147). Stroudsburg: Association for Computational Linguistics.

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh, J. (2010). Evaluation of georeferencing. In *proceedings of the 6th workshop on geographic information retrieval* (p. 7). New York: ACM.

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

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

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., et al. (2013). *Ontonotes release 5.0 ldc2013t19*. Philadelphia: Linguistic Data Consortium.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wing, B. P., & 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* (pp. 955–964). Stroudsburg: Association for Computational Linguistics.

Wing, B., & Baldridge, J. (2014). Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 336–348).

Yadav, V., & 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* (pp. 2145–2158).

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

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

Zheng, X., Han, J., & Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, *30*, 1652–1671.

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.