Enhancing spatial and textual analysis with EUPEG: an extensible and unified platform for evaluating geoparsers

Jimin Wang and Yingjie Hu

GeoAI Lab, Department of Geography, University at Buffalo, NY 14260, USA

Abstract

A rich amount of geographic information exists in unstructured texts, such as Web pages, social media posts, housing advertisements, and historical archives. Geoparsers are useful tools that extract structured geographic information from unstructured texts, thereby enabling spatial analysis on textual data. While a number of geoparsers were developed, they were tested on different datasets using different metrics. Consequently, it is difficult to compare existing geoparsers or to compare a new geoparser with existing ones. In recent years, researchers created open and annotated corpora for testing geoparsers. While these corpora are extremely valuable, much effort is still needed for a researcher to prepare these datasets and deploy geoparsers for comparative experiments. This paper presents EUPEG: an Extensible and Unified Platform for Evaluating Geoparsers. EUPEG is an open source and Web based benchmarking platform which hosts a majority of open corpora, geoparsers, and performance metrics reported in the literature. It enables direct comparison of the hosted geoparsers, and a new geoparser can be connected to EUPEG and compared with other geoparsers. The main objective of EUPEG is to reduce the time and effort that researchers have to spend in preparing datasets and baselines, thereby increasing the efficiency and effectiveness of comparative experiments.

Keywords: geoparsing, benchmarking platform, toponym, spatial and textual analysis, geospatial semantics, geographic information retrieval

1. Introduction

Many studies and applications nowadays need an integration of spatial and textual analysis. In disaster response, it is often necessary to recognize, geo-locate, and analyze the place names mentioned in short text messages or social media posts in order to understand who needs help and where (MacEachren et al., 2011; Gelernter and Balaji, 2013; Lan et al., 2014; Pezanowski et al., 2018). Studying place relations and interactions in virtual or cognitive

Wang, J. and Hu, Y. (2019): Enhancing spatial and textual analysis with EUPEG: an extensible and unified platform for evaluating geoparsers, *Transactions in GIS*, accepted.
DOI: https://www.doi.org/10.1111/tgis.12579
Contact info: Jimin Wang (jiminwan@buffalo.edu); Yingjie Hu (yhu42@buffalo.edu)

spaces usually involves extracting place names from texts, such as Wikipedia pages or news articles, and analyzing their co-occurrences with spatial distances (Hecht and Moxley, 2009; Liu et al., 2014; Geiß et al., 2015; Salvini and Fabrikant, 2016; Hu et al., 2017). To develop place-based GIS, researchers may need to examine human experiences encoded in texts, such as in travel blogs, and how these human experiences are related to spatial locations (Adams and McKenzie, 2013; Ballatore and Adams, 2015; Gao et al., 2017). In addition, there exists various geographic knowledge in Web pages (Jones et al., 2008), housing advertisements (McKenzie et al., 2018), business documents (Faulconbridge et al., 2008), historical archives (Grossner et al., 2016), and other types of texts.

Geoparsing is a critical process for extracting spatial information from textual data. It is recognized as an important research topic in the broader field of geographic information retrieval (GIR) (Jones and Purves, 2008; Purves et al., 2018). A geoparsing system is called a *geoparser* which takes unstructured textual data as the input and outputs a set of recognized place names and their spatial footprints (Freire et al., 2011; Leidner, 2008).

While a number of geoparsers have already been developed, it is difficult to directly compare their performances. Examples of the developed geoparsers include MetaCarta (Frank et al., 2006), GeoTxt (Karimzadeh et al., 2013), the Edinburgh Geoparser (Alex et al., 2015), TopoCluster (DeLozier et al., 2015), and CamCoder (Gritta et al., 2018b). Two factors make the direct comparison difficult. First, many geoparsers were tested on project-specific datasets that are not shared publicly. Besides the additional effort for making data ready for sharing, there also exist policy restrictions (e.g., Twitter forbids sharing the content of tweets) and privacy concerns that prevent researchers from sharing their data publicly. As a result, geoparsers cannot be fairly compared since the same geoparser can have very different performances depending on the testing datasets (Leidner, 2006; Ju et al., 2016). Second, different performance metrics were usually used for evaluating geoparsers. Some researchers used the metrics of *precision*, *recall*, and *F-score* adopted from the field of information retrieval, while some others used metrics based on spatial distances, such as *mean* or *median error distance*. Due to both factors, we cannot compare geoparsers by juxtaposing the performance numbers reported in their papers.

To effectively compare existing geoparsers or to compare a new geoparser with existing baselines, one would ideally find and deploy the geoparsers in the literature and use the same datasets and metrics to test their performances. However, such a process is time-consuming and labor-intensive. First, one needs to find openly shared and annotated datasets. While the community has already made great efforts to share datasets, such as the Local-Global Lexicon (LGL) corpus (Lieberman et al., 2010), WikToR (Gritta et al., 2018c), and GeoCorpora (Wallgrün et al., 2018), researchers still need to spend much time preparing these datasets for experiments. For example, GeoCorpora is a valuable dataset containing human annotated tweets. Due to Twitter's policy restriction, GeoCorpora contains only the IDs of tweets rather than their full content. To use GeoCorpora, one needs to apply for a developer account for using Twitter's Application Programming Interface (API) and write a program to *rehydrate* the dataset. Even for the datasets that are more readily available, they are often in different formats and need to be harmonized into the same format before an experiment. Second, it takes a considerable amount of time to find, deploy, and re-run existing geoparsers. This process can take even longer when no direct download link is provided for a geoparser or when there is a lack of deployment instructions. Third, a set of performance metrics need to

be implemented to compare geoparsers. Although implementing these metrics may not be difficult, one needs to harmonize the heterogeneous output formats of different geoparsers and compare their outputs with ground truth annotations. In sum, conducting an effective comparative experiment of geoparsers costs a lot of time and human resources. While those costs are probably fine for a single research group, the entire community can lose a lot of time if every individual research group has to prepare datasets, geoparsers, and metrics in order to run an experiment.

This paper presents EUPEG: an Extensible and Unified Platform for Evaluating Geoparsers. EUPEG is designed as an open source and Web based platform. It hosts a majority of the open corpora, geoparsers, and performance metrics reported in the literature. One can directly compare these hosted geoparsers on the same datasets using the same metrics, or can connect a new geoparser to EUPEG and compare it with the existing ones. The value of EUPEG can be seen from the perspectives of both geoparser users and researchers. For a user who would like to find a suitable geoparser to process a corpus, EUPEG offers a comprehensive view on the advantages and limitations of different geoparsers (e.g., some may have higher precision while some others may have higher recall) and their performances on different types of corpora (e.g., short messages or long articles). For researchers who would like to develop a new geoparser, they can focus on inventing new methods rather than preparing datasets and baselines. In addition, EUPEG automatically archives the results and configurations of experiments, such as the date and time of an experiment, the selected datasets, used geoparsers, and performance metrics. Researchers can share experiment results with their colleagues or even the general public more easily via an experiment ID. The contributions of this paper are as follows:

- We propose and develop a benchmarking platform, EUPEG, for effective and efficient comparison of geoparsers. EUPEG currently hosts eight annotated geographic corpora, nine geoparsers, and eight performance metrics. New geoparsers and datasets can also be connected to it. Experiment results and configurations are recorded and can be shared via experiment IDs. A demo of EUPEG can be accessed at: `https://geoai.geog.buffalo.edu/EUPEG`.

- We provide a systematic review on the geoparsing resources hosted on EUPEG. The corpora are in four different text genres, ranging from news articles to social media posts; the geoparsers are developed using different methods, such as heuristics and machine learning; and the performance metrics are from information retrieval or based on spatial distances. EUPEG serves as a one-stop platform that unifies the heterogeneous datasets, geoparsers, and performance metrics.

- We share the source code of EUPEG on GitHub, along with the hosted resources under permitted licenses (e.g., GNU General Public License). The code repository can be accessed at: `https://github.com/geoai-lab/EUPEG`. The shared source code enables researchers to run EUPEG on a local computer, or to add more datasets and geoparsers at the source-code level. One can also extend EUPEG with new features, such as new performance metrics suitable for a project.

While EUPEG is designed for geoparsing, the idea of developing benchmarking platforms can be extended to other research topics in geography, such as land use and land

cover (LULC) classification, where different solutions are developed for addressing the same problem. This paper is a major extension of our previous short paper (Hu, 2018). The remainder of this paper is organized as follows. Section 2 reviews related work on geoparsing, corpora building, and benchmarking platforms. Section 3 presents the design details of EUPEG, including the overall architecture and the hosted datasets, geoparsers, and performance metrics. Section 4 demonstrates the implemented EUPEG and provides an analytical evaluation on the approximate time that can be saved by EUPEG for comparative experiments. Finally, Section 5 summarizes this work and discusses future directions.

2. Related Work

In this section, we provide a review on related studies. We start by introducing the background knowledge of geoparsing and major geoparsers developed so far, and continue to discuss the efforts made by the community to create and share open and annotated corpora. We then discuss the recent movement towards developing benchmarking platforms for effective and efficient comparisons of different solutions to the same problems.

2.1. Geoparsing and geoparsers

Geoparsing is a research topic often studied in GIR (Jones and Purves, 2008; Purves et al., 2018). The goal of geoparsing is to recognize place names mentioned in texts and resolve them to the corresponding place instances and location coordinates (Freire et al., 2011; Barbaresi, 2017; Gritta et al., 2018c). Geoparsing is typically performed in two consecutive steps: toponym recognition and toponym resolution. The first step recognizes place names from texts without identifying the particular place instance referred by a name, while the second step aims to resolve any ambiguity of the place name and locates it to the right spatial footprint. Figure 1 illustrates the input and output of geoparsing and its two steps. Many

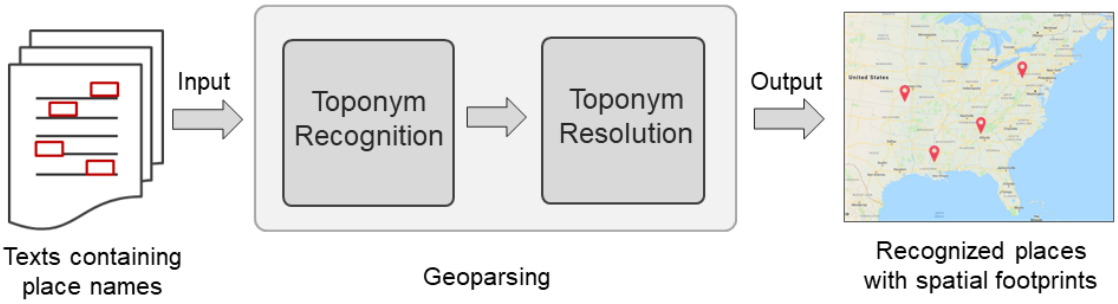

Figure 1: The input and output of geoparsing and its two main steps.

methods have been proposed for these two steps. For toponym recognition, early research used gazetteer-based entry matching and grammatical rules (Woodruff and Plaunt, 1994; Purves et al., 2018), while more recent approaches employed machine learning and natural language processing techniques. Particularly, the Stanford Named Entity Recognition (NER) tool was used in many studies for toponym recognition (Lieberman et al., 2010; Gelernter and Mushegian, 2011; Karimzadeh et al., 2013; DeLozier et al., 2015). For toponym resolution, various heuristics were developed to resolve place name ambiguity (Amitay et al., 2004; Leidner, 2008). A simple method is to resolve a place name to the place instance that has

the highest population or the largest total geographic area (Li et al., 2002; Ladra et al., 2008). Machine learning models were also developed for toponym resolution by exploiting various features, such as toponym co-occurrences (Overell and Rüger, 2008), words in the local context (Speriosu and Baldridge, 2013), distances among toponyms (Santos et al., 2015), topics of the local context (Ju et al., 2016), and a combination of multiple features (Nesi et al., 2016; Gritta et al., 2018b).

A number of geoparsers were developed which can function as end-to-end systems for completing both steps. *GeoTxt*, developed by Karimzadeh et al. (2013), is a Web-based geoparser that leverages Stanford NER and two other NER tools for toponym recognition and uses GeoNames and a set of heuristic rules for toponym resolution. TopoCluster, by DeLozier et al. (2015), can perform geoparsing without using a gazetteer. It uses Stanford NER to recognize toponyms from texts and then resolves toponyms based on the *geographic profiles* of words in the surrounding context. Cartographic Location And Vicinity INdexer (CLAVIN) is an open source geoparser that employs Apache OpenNLP for toponym recognition and utilizes a gazetteer and fuzzy search for toponym resolution. The Edinburgh Geoparser is a geoparsing system developed by the Language Technology Group at Edinburgh University (Alex et al., 2015). It uses their in-house software tool, called LT-TTT2, for toponym recognition, and the toponym resolution step is based on a gazetteer such as GeoNames. CamCoder is a deep learning based geoparser developed by Gritta et al. (2018b), which integrates convolutional neural networks, word embeddings, and the geographic vector representations of place names. There also exist commercial geoparsers, such as Geoparser.io (`https://geoparser.io`), which often charge a fee. Some commercial geoparsers, such as Yahoo! PlaceSpotter (`https://developer.yahoo.com/boss/geo/docs/PM_KeyConcepts.html`), provide relatively permissive rate limitations for free requests (e.g., 2,000 calls per hour).

2.2. Efforts in sharing open and annotated corpora

While many geoparsers exist, it is difficult to directly compare them due to a lack of open and annotated corpora. In recent years, researchers made great efforts to address this issue. Lieberman et al. (2010) shared a human annotated dataset called Local-Global Lexicon (LGL) containing 588 news articles published by local newspapers from highly ambiguous places. Hu et al. (2014) contributed an automatically annotated corpus containing short sentences retrieved from the home pages of cities with ambiguous names such as *Washington*. Ju et al. (2016) shared a corpus of short sentences from various Web pages, which was automatically collected and annotated using a script based on the Microsoft Bing Search API. Gritta et al. (2018c) contributed WikToR which is a corpus of Wikipedia articles with ambiguous names, such as *Lima, Peru* and *Lima, Oklahoma*, automatically annotated by a Python script. Wallgrün et al. (2018) contributed GeoCopora which is a dataset of tweets manually annotated using a hybrid approach with both users on Amazon's Mechanical Turk and researchers in the domain of geography. Gritta et al. (2018b) and Gritta et al. (2018a) shared two human annotated corpora, GeoVirus and GeoWebNews, which contain 229 and 200 news articles respectively. TR-News is another news article corpus which contains 118 articles manually annotated by Kamalloo and Rafiei (2018). In addition to contemporary corpora, some historical datasets were also made available, such as *War Of The Rebellion* (WOTR) by DeLozier et al. (2016). Leidner (2006) developed the TR-CoNLL corpus which

contains 946 annotated news articles from Reuters; however, it is not publicly available to the best of our knowledge. The ACE 2005 English SpatialML is an annotated news corpus shared on the Linguistic Data Consortium (Mani et al., 2008), but it charges a fee ($1,000) for non-members. While these annotated corpora greatly facilitate the development and testing of geoparsers, finding, downloading, and preparing these corpora require considerable amounts of time and effort.

2.3. Benchmarking platforms

The importance and necessity of evaluating geoparsers in a systematic manner have already been recognized by the research community (Monteiro et al., 2016; Melo and Martins, 2017; Richter et al., 2017; Gritta et al., 2018c; Wallgrün et al., 2018). Melo and Martins (2017) argued that the fact that one geoparser performed worse than another geoparser on one particular dataset did not mean that this geoparser would still perform worse than the other if a different dataset were used. Gritta et al. (2018c) compared the performances of five geoparsers on two corpora using a set of standard performance measures, such as precision, recall, F-score, and median error distance. In their more recent work, the authors further proposed a pragmatic guide to geoparsing evaluation (Gritta et al., 2018a). In addition to publications, Gritta et al. (2018c,a) also released their source code and the annotated corpora which greatly facilitated the reproduction of their experiments. EUPEG is built on the foundational work of (Gritta et al., 2018c,a), but extends their work in three aspects. First, EUPEG provides a benchmarking platform which offers datasets and baseline geoparsers that are ready for use. While Gritta et al. (2018c,a) have shared the source code of comparing five geoparsers, a lot of effort is still needed to understand, deploy, and run these geoparsers. Some geoparsers do not function on certain operating systems (OS) (e.g., the Edinburgh Geoparser is not supported on Windows) or require extra database configurations (e.g., TopoCluster requires PostgreSQL and PostGIS), which can add additional requirements on their deployments. EUPEG directly hosts these geoparsers, along with annotated corpora and performance metrics. Researchers can directly run experiments on EUPEG, and can connect their own geoparsers and datasets to the platform. Second, EUPEG extends the corpora and geoparsers from Gritta et al. (2018c,a). We provide eight annotated corpora in four different text genres, which include news articles, Wikipedia articles, social media posts, and Web pages. We provide nine geoparsing methods which include not only specialized geoparsers (e.g., GeoTxt and CLAVIN) but also a number of geoparsing systems extended from general NER tools, such as DBpedia Spotlight, Stanford NER, and spaCy NER. Third, EUPEG offers the capability of archiving research experiments. Each experiment is assigned a unique ID that allows researchers to share first-hand research outcomes and to search the results of previous experiments.

The demand for benchmarking platforms is also witnessed in other research fields beyond geography. Cornolti et al. (2013) developed a framework for systematically evaluating a number of named entity annotators, such as AIDA, Illinois Wikifier, and DBpedia Spotlight, on the same datasets. Building on the work of Cornolti et al. (2013), Usbeck et al. (2015) developed GERBIL which is a platform for agile, fine-grained, and uniform evaluations of named entity annotation tools. The practice of comparing different methods on the same datasets was also seen in computer science conferences, such as the Message Understanding Conference (MUC) (Sundheim, 1993), the Conference on Computational Natural Language

Learning (CoNLL) (Tjong Kim Sang and De Meulder, 2003), and the Making Sense of Microposts workshop series (MSM) (Cano et al., 2014). Such a practice is especially effective when multiple solutions exist for the same research problem, and can reveal the advantages and limitations of different solutions. Sharing datasets for comparing methods can fuel the advancement in a particular research area as well. For example, the availability of the ImageNet dataset was a critical boost to the remarkable development of deep learning in computer vision (Deng et al., 2009). To the best of our knowledge, EUPEG is the first benchmarking platform for the research problem of geoparsing.

3. EUPEG

3.1. Overall architecture

EUPEG is designed as a Web based and open source benchmarking platform. It provides two main functions:

- It enables effective and efficient comparison of different geoparsers on the same datasets using the same performance metrics.
- It facilitates the sharing of experiments by archiving evaluation results and configurations and supporting the search of previous experiments.

The overall architecture of EUPEG is shown in Figure 2. Two major modules are designed. The *Experiment Module* hosts a majority of openly available resources including annotated corpora, existing geoparsers, and performance metrics. The *Archiving and Search Module* records experiment results and supports the search of previous experiments. For geoparser researchers who would like to develop new geoparsers, they can connect a new geoparser to EUPEG and compare it with others on the hosted datasets using the same performance metrics. Researchers can also upload one or multiple customized datasets and compare the new geoparser with existing baselines on these customized datasets. For users who would like to find a suitable geoparser for processing a corpus, they can compare existing geoparsers directly on EUPEG to see their advantages and limitations. The experiment configurations (e.g., the used datasets, geoparsers, and metrics) and results are recorded in a database in the *Archiving and Search Module*. One can search previous experiments based on their experiment IDs automatically generated by EUPEG. In the following, we present details about the datasets, geoparsers, and metrics.

3.2. Datasets

EUPEG hosts a majority of annotated geographic corpora reported in the literature. Two criteria are used for selecting these datasets. First, they have to be formally described by existing papers; second, the datasets should be openly available without a fee (e.g., a dataset shared on GitHub under the MIT license). In addition, we focus on *geographic* corpora, namely those with toponyms and spatial footprints annotated. There exist general NER corpora whose annotations contain other types of entities (e.g., persons and concepts) and do not provide spatial footprints. Those general NER corpora are not included.

The corpora hosted on EUPEG are in four different text genres: news articles, Wikipedia articles, social media posts, and Web pages. Having multiple genres rather than a single

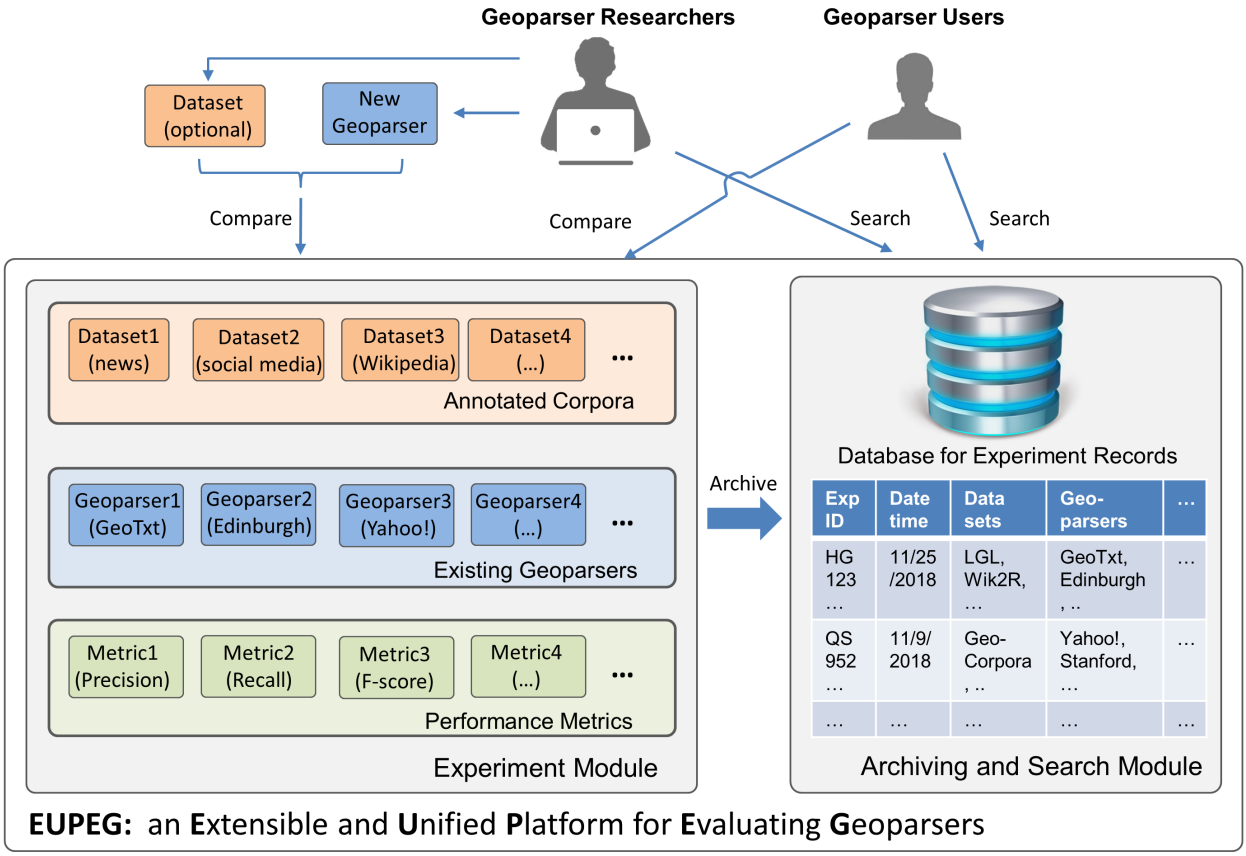

Figure 2: The overall architecture of EUPEG.

text type can provide a more comprehensive evaluation on the performance of a geoparser (Leidner, 2006). These datasets are described as follows.

News articles. Four news article corpora are hosted on EUPEG:

- *LGL*. LGL was developed by Lieberman et al. (2010), which contains 588 human-annotated news articles published by 78 local newspapers from highly ambiguous places, such as *Paris News* (Texas), *Paris Post-Intelligencer* (Tennessee), and *Paris Beacon-News* (Illinois).

- *GeoVirus*. This is a human-annotated dataset shared by Gritta et al. (2018b). It contains 229 news articles from WikiNews during 08/2017 - 09/2017. These news articles cover global disease outbreaks and epidemics, and were collected using keywords, such as *"Ebola"* and *"AIDS"*.

- *TR-News*. This dataset was contributed by Kamalloo and Rafiei (2018), which contains 118 human-annotated news articles from various global and local news sources. The authors deliberately included less dominant place instances, such as *Edmonton, England* and *Edmonton, Australia*, while also keeping articles from general global news sources, such as BBC and Reuters.

- *GeoWebNews*. This dataset was shared by Gritta et al. (2018a) which comprises 200

human-annotated news articles from 200 globally distributed news sites collected from April 1st to 8th in 2018. The authors randomly selected one article from each domain until they reached 200 news articles for creating this dataset.

Wikipedia articles. One Wikipedia article corpus is hosted on EUPEG.

- *WikToR.* This corpus was provided by Gritta et al. (2018c). It was automatically generated by a script using those Wikipedia articles about places. It contains 5,000 articles with ambiguous names, such as *Lima, Peru*, *Lima, Ohio*, and *Lima, Oklahoma*. One limitation of WikToR is that it does not annotate all toponyms in the texts but only those that are the description target of a Wikipedia article. As a result, some performance metrics, such as precision, recall, and F-score, cannot be used for quantifying the performance of geoparsers based on WikToR.

Social media posts. EUPEG hosts one social media dataset, GeoCorpora, contributed by Wallgrün et al. (2018).

- *GeoCorpora.* This is a tweet corpus that contains 1,639 human-annotated tweet posts. These posts were first annotated using a crowdsourcing approach by workers on Amazon's Mechanical Turk and then further annotated (with disagreements resolved) by researchers in geography. It is worth noting that the original paper reported 2,122 tweets with toponyms annotated. Due to the data sharing restriction of Twitter, Wallgrün et al. (2018) could not share the full content of tweets but only tweet IDs. When *rehydrating* tweets, we were able to recover only 1,639 tweets, since some tweets were deleted by their authors. The number of tweets that can be recovered will only decrease as the time goes by. Thus, we believe that another value of EUPEG is its capability of preserving valuable contributions from previous research by sidestepping some policy restrictions (in this case, EUPEG does not provide any direct download of the tweets).

Web pages. Two Web page corpora are hosted on EUPEG.

- *Hu2014.* This is a small corpus contributed by Hu et al. (2014), which was automatically constructed by a script. The authors focused on two highly ambiguous US place names, *Washington* and *Greenville*, and retrieved textual descriptions from the websites of related cities. The texts in these Web pages were then divided into shorter sentences. Overall, this dataset contains 134 entries. Not all toponyms in the sentences are annotated, and therefore precision, recall, and F-score cannot be applied to evaluating the geoparsing results based on this dataset (similar to *WikToR*).

- *Ju2016.* This is another automatically constructed corpus. It was contributed by Ju et al. (2016) who made use of a list of highly ambiguous US place names on Wikipedia, and then used Microsoft Bing Search API to retrieve sentences from various Web pages (Wikipedia articles were removed from these Web pages) that contain the searched place names. This corpus contains 5,441 entries. Similar to *WikToR* and *Hu2014*, this dataset does not annotate all toponyms and cannot use the performance metrics of precision, recall, and F-score.

Table 1: A summary of the open and annotated corpora hosted on EUPEG.

Dataset	Genre	Text Date	Entry Count	Average Words per Entry	Average Toponym per Entry
LGL	News	03/2009	588	315	8.0
GeoVirus	News	08-09/2017	229	276	9.4
TR-News	News	2009-2017	118	324	10.8
GeoWebNews	News	04/2018	200	404	12.6
WikToR	Wikipedia	03/2016	5000	213	6.3
GeoCorpora	Social Media	2014-2015	1639	19	2.1
Hu2014	Web Pages	08/2014	134	27	1.3
Ju2016	Web Pages	11/2016	5441	21	1.2

In total, EUPEG hosts eight geographic corpora in four different text genres. Table 1 summarizes the attributes of these datasets.

There also exist open and annotated historical corpora. For example, WOTR is a US civil war corpus with toponyms focusing on the southern US (DeLozier et al., 2016). However, geoparsing such corpora requires special configurations such as adding historical gazetteers and processing older languages (e.g., the texts of WOTR are from 1860s). The current version of EUPEG aims to compare geoparsers based on their default configurations and does not include historical corpora.

It is worth noting that the definition of toponym can vary across different corpora. It seems that different researchers often have their own opinions on what should be considered as toponyms and what should not. This definition difference affects the ground-truth toponym annotation in a corpus. For example, *LGL* considers demonyms (e.g., Canadian) as toponyms and annotates them to point coordinates (e.g., the center of Canada), whereas *GeoVirus* does not annotate building names, point-of-interest (POI) names, street names, and river names. In a recent work, Gritta et al. (2018a) provided a pragmatic taxonomy of toponyms which further divided toponyms into *literal* and *associative* toponyms with 13 sub categories. In this work, we do not attempt to define toponym from one single perspective, but allow the datasets with different toponym definitions to co-exist. Such diversity allows users and researchers to see the different performances of a geoparser across corpora. One can then choose a geoparser that performs the best on a corpus that has a toponym definition similar to theirs. Table 2 summarizes the different annotations of toponyms contained in each of the corpora hosted on EUPEG. As can be seen, all datasets include administrative units in their annotations but have different coverage on other types of entities. Some of these differences come from the corpus building process (e.g., *WikToR*, *Hu2014*, and *Ju2016* were automatically constructed based on the names of cities and towns only), while some others originate from the different views of researchers on the definition of toponym. It seems that domain knowledge plays a major role in the annotation of toponyms. For example, *GeoCorpora* is a dataset contributed by geographers, and its annotations contain only the names that can be pinned down to a certain location on the surface of the Earth. By contrast, *GeoWebNews*, *TR-News*, and *LGL* are contributed by linguists and computer scientists who tend to annotate any terms that may have a geographic meaning (e.g., "Canadian" and

Table 2: Toponym annotations in different corpora.

Dataset	Admin units (cities, towns, ...)	Natural features (rivers, mountains, ...)	Facilities (buildings, roads, airports, ...)	Demonyms (Canadian, Syrian, American, ...)	Metonymies (London announced a new policy ...)	Modifiers (Spanish sausage, UK beef, ...)
LGL	✓	✓	*	✓	✓	✓
GeoVirus	✓					
TR-News	✓		*			✓
GeoWebNews	✓	✓	✓	✓	✓	✓
WikToR	✓					
GeoCorpora	✓	✓	✓			
Hu2014	✓					
Ju2016	✓					

* indicates that the dataset contains toponym annotations in that category but does not provide geographic coordinates for the annotated toponyms.

"Spanish sausage").

3.3. Geoparsers

We select geoparsers for EUPEG using the following criteria. First, the selected geoparsers should function as end-to-end systems, i.e., they can take textual documents as the input and output spatial coordinates. Second, for academic geoparsers, the accompanying papers should be published after 2010 and they should provide publicly accessible API or downloadable software packages. Due to technological advancements, geoparsers developed before 2010 generally do not have performances close to the state of the art, and their source codes can be hard to obtain and may not run on a modern OS. Third, for industrial geoparsers, they should provide an API that either allows free access or has a permissive rate limitation for free requests. These three criteria follow the foundational work of Gritta et al. (2018c), and the geoparsers below are provided on EUPEG.

- *GeoTxt.* GeoTxt is an academic geoparser developed by the GeoVISTA center of Pennsylvania State University (Karimzadeh et al., 2013, 2019). It was initially designed to geoparse microblogs (e.g., tweets), but can be applied to longer texts as well. GeoTxt provides a publicly accessible and free API at `http://www.geotxt.org`, and is being maintained by its researchers. EUPEG does not host a local instance of GeoTxt but connects to its API. An advantage of such an online connection is that new updates of GeoTxt will be reflected on EUPEG. On the flip side, EUPEG cannot use GeoTxt when its online service is down. EUPEG connects to version 2.0 of GeoTxt which uses its local GeoNames gazetteer deployed in July 2017.

- *The Edinburgh Geoparser.* The Edinburgh Geoparser is an academic geoparser developed by the Language Technology Group (LTG) at The University of Edinburgh (Alex et al., 2015). A publicly available package of this geoparser is provided at: `https://www.ltg.ed.ac.uk/software/geoparser`. EUPEG hosts version

1.1 of the Edinburgh Geoparser which uses the online service of GeoNames as its gazetteer. While supported on Linux and MacOS, it cannot run on Windows.

- *TopoCluster.* TopoCluster is an academic geoparser developed by DeLozier et al. (2015) at the University of Texas at Austin. It performs geoparsing based on the geographic profiles of words characterized by the local Getis-Ord G_i^* statistic. While their methodology focuses on toponym resolution, their source code (`https://github.com/grantdelozier/TopoCluster`) provides an end-to-end system for completing both steps of geoparsing. TopoCluster does not provide an official version number. We host its latest version shared on GitHub which was updated in November 2016.

- *CLAVIN.* Cartographic Location And Vicinity INdexer is an open source geoparser that employs Apache OpenNLP Name Finder for toponym recognition, and a number of heuristics and fuzzy search for toponym resolution. CLAVIN does not come with an academic paper, but its descriptions and source code can be obtained from GitHub (`https://github.com/Berico-Technologies/CLAVIN`) and Maven Central. We host CLAVIN 2.1.0 on EUPEG and it employs a local GeoNames gazetteer deployed in April 2019.

- *Yahoo! PlaceSpotter.* Yahoo! PlaceSpotter is an industrial geoparser which offers an online REST API. As a proprietary geoparser, PlaceSpotter does not describe the exact methods behind but provides some descriptions on its functions and outputs at: `https://developer.yahoo.com/boss/geo/docs/PM_KeyConcepts.html`. PlaceSpotter is requested via YQL (Yahoo! Query Language), and its rate limit for free requests is relatively permissive (2,000 calls per hour; a corpus with 5,000 entries can be parsed within 3 hours). EUPEG connects to Yahoo! PlaceSpotter via its online REST API which employs its Where-on-Earth ID (WOEID) for referencing places.

- *CamCoder.* CamCoder is an academic geoparser developed by Gritta et al. (2018b) at the Language Technology Lab of the University of Cambridge. CamCoder is a deep learning based geoparser that leverages Convolutional Neural Networks (CNNs) with global maximum pooling and map-based word vector representations. The source code of CamCoder is available at: `https://github.com/milangritta/Geocoding-with-Map-Vector`. Running CamCoder requires configurations on the local computing environment to include deep learning libraries, such as Tensorflow and Keras. EUPEG hosts the latest version of CamCoder shared on GitHub (updated in September 2018) which uses its local GeoNames gazetteer prepared in July 2018.

In addition to the six specialized geoparsers above, EUPEG also provides three geoparsing systems extended from general NER tools (e.g., Stanford NER). These systems are included because previous research argued that geoparsing can be considered as a sub task of NER, and a geoparser can be developed by limiting the entities recognized by an NER tool to toponyms and adding spatial footprints via a gazetteer (Inkpen et al., 2017). Thus, including these NER-based geoparsing systems can help provide more comprehensive comparisons. The following three systems are hosted on EUPEG:

- *Stanford NER + Population.* Stanford NER is a powerful and open source NER tool developed by the Stanford Natural Language Processing Group. It has been used in numerous previous studies, including some specialized geoparsers, such as GeoTxt and TopoCluster. We extend Stanford NER to a geoparsing system by assigning the recognized toponyms to the place instances with highest populations. This simple heuristic is used because previous research has shown that assigning place names to the instances with the highest populations is a strong baseline for geoparsing and can sometimes surpass more complex models (Speriosu and Baldridge, 2013; DeLozier et al., 2015; Gritta et al., 2018a). We use Stanford CoreNLP toolbox (version 3.9.2) integrated with the online service of GeoNames.

- *spaCy NER + Population.* spaCy is a free and open source library for natural language processing tasks in Python. Released in 2014, it has already been used in many studies and applications due to its good performance (Choi et al., 2015; Jiang et al., 2016). We integrate spaCy NER (version 2.0.18) with the online service of GeoNames, and assign the recognized toponyms to the place instances with the highest populations.

- *DBpedia Spotlight.* DBpedia Spotlight is a general named entity recognition and linking (NERL) tool (Mendes et al., 2011; Daiber et al., 2013). This type of tool not only recognizes entities from texts but also links them to the corresponding URLs in a knowledge base (such as DBpedia) which provides geographic coordinates for the recognized places. We convert DBpedia Spotlight into a geoparser by limiting the output to toponyms and extracting their coordinates from DBpedia pages via the *geo:lat* and *geo:long* properties. While there exist other NERL tools such as AIDA (Hoffart et al., 2011) and TagMe (Ferragina and Scaiella, 2010), DBpedia Spotlight is a widely-used NERL tool whose performance is among the state of the art (Van Erp et al., 2013; Cornolti et al., 2013; Usbeck et al., 2014). EUPEG hosts DBpedia Spotlight 1.0.0 with coordinates retrieved from online DBpedia pages.

In total, EUPEG provides nine geoparsing systems for comparative experiments. Table 3 summarizes these systems and their main components. These systems include six specialized geoparsers that can be directly deployed, and three baseline systems that are extended from general NER or NERL tools via further developments and gazetteer configurations.

3.4. Performance metrics

EUPEG provides a number of performance metrics based on which different geoparsers can be evaluated and compared. There is no general agreement on which metrics should be used for evaluating geoparsers. As a result, we select eight metrics that were used in a variety of previous studies. In the following, we discuss these metrics individually.

- *Precision.* Precision measures the percentage of correctly identified toponyms (true positives) among all the toponyms recognized by a geoparser. Precision was used in previous studies, such as (Leidner, 2008; Lieberman et al., 2010; Inkpen et al., 2017). Precision is calculated using the following equation:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

Table 3: Geoparsing systems hosted on EUPEG and their main components.

Geoparser	Toponym Recognition	Toponym Resolution	Gazetteer
GeoTxt (Version 2.0)	Stanford NER	Heuristic rules	GeoNames (July 2017)
Edinburgh (Version 1.1)	LT-TTT2	Heuristic rules	GeoNames (Online)
TopoCluster (Nov. 2016)	Stanford NER	Geo-profiles of words	GeoNames+ Natural Earth (Nov. 2016)
CLAVIN (Version 2.1.0)	Apache OpenNLP	Heuristic rules	GeoNames (Apr. 2019)
Yahoo! PlaceSpotter (Online)	Proprietary	Proprietary	WOEID (Where on Earth ID) (Online)
CamCoder (Sept. 2018)	spaCy NER	CNNs+Map-based word vectors	GeoNames (July 2018)
Stanford NER + Population (Version 3.9.2)	Stanford NER	Highest population	GeoNames (Online)
spaCy NER + Population (Version 2.0.18)	spaCy NER	Highest population	GeoNames (Online)
DBpedia Spotlight (Version 1.0.0)	LingPipe Exact Dictionary Chunker	Context similarity	DBpedia (Online)

where tp represents *true positive* and fp represents *false positive*.

- *Recall*. Recall measures the percentage of correctly identified toponyms among all the toponyms that should be identified (i.e., the toponyms that are annotated as ground truth). Recall was used in previous studies, such as (Leidner, 2008; Lieberman et al., 2010; Inkpen et al., 2017). Recall is calculated using the equation as below:

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

where fn represents *false negative*.

- *F-score*. F-score is the harmonic mean of precision and recall. F-score is high when both precision and recall are fairly high and is low if either of the two is low. F-score was used in previous studies, such as (Leidner, 2008; Lieberman et al., 2010; Inkpen

et al., 2017). F-score is calculated using the equation below:

$$F\text{-}score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

F-score is also called F-measure or F_1-score.

- *Accuracy.* This metric is suitable for measuring performances on those corpora that do not have all toponyms annotated. For example, both *WikToR* and *Ju2016* only annotate a subset of all the toponyms mentioned in the text. In these situations, precision, recall, and F-score are no longer suitable, since we do not have all toponyms annotated. Accuracy can be used to quantify the percentage of the annotated toponyms that are also recognized by a geoparser. Accuracy was used in previous studies, such as (Gelernter and Mushegian, 2011; Karimzadeh, 2016; Gritta et al., 2018c). It is calculated using the equation below:

$$Accuracy = \frac{|Annotated \cap Recognized|}{|Annotated|} \tag{4}$$

where *Annotated* represents the set of toponyms provided in the annotation, and *Recognized* represents the set of toponyms recognized by the geoparser.

Precision, recall, F-score, and accuracy quantify the ability of a geoparser in correctly recognizing place names from texts rather than geo-locating these names. Accordingly, they measure the performance of a geoparser in the toponym recognition step. The establishment of matching between ground-truth annotations and geoparsing outputs is a topic that is worth discussing since it can directly affect the obtained measures. Previous work has discussed both *exact matching* and *inexact matching* (Gritta et al., 2018c). For a sentence such as "The Town of Amherst has been a leader in providing online geographic information", a geoparser may recognize "Amherst" as a toponym, while the ground-truth annotation may be "Town of Amherst". For *exact matching*, this will be considered as both a fp and a fn, since the output of the geoparser does not match the ground truth. For *inexact matching*, it will be considered as a tp. We adopt *inexact matching* to accommodate such syntactically inconsistent but semantically meaningful outputs, and use the same implementation as in (Gritta et al., 2018c) for determining matches.

To measure the performance of a geoparser in geo-locating toponyms, the following four metrics are provided on EUPEG.

- *Mean Error Distance (MED).* MED computes the mean of the Euclidean distances between the annotated location and the location output by a geoparser. MED was used in previous studies, such as (Cheng et al., 2010; Speriosu and Baldridge, 2013; Santos et al., 2015). It is calculated using the equation below:

$$MED = \frac{\sum_{i=1}^{N} \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}}{N} \tag{5}$$

where N is the number of annotated toponyms that are recognized and geo-located by a geoparser. (x_i, y_i) is the annotated coordinates, and (x_i', y_i') is the geoparsed

coordinates. The toponyms, which are only in the geoparsing output or only in the annotations, are not included in computing MED; those mismatches are evaluated by the previous four metrics.

- *Median Error Distance (MdnED).* MED is sensitive to outliers which means a small number of geoparsed toponyms that are located far away from their ground-truth locations can largely distort the evaluation result. MdnED computes the median value of the error distances and is robust to outliers. MdnED was used in previous studies, such as (Speriosu and Baldridge, 2013; DeLozier et al., 2015; Santos et al., 2015). MdnED is calculated as below:

$$MdnED = Median(\{ed_i | ed_i = \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}, i \in [1, N]\}) \tag{6}$$

  where ed_i represents the ith error distance.

- *Accuracy@161.* This metric calculates the percentage of the toponyms that are geolocated within 161 kilometers (100 miles) of the ground truth locations. Accuracy@161 was used in previous studies, such as (Cheng et al., 2010; DeLozier et al., 2015; Gritta et al., 2018c). A main motivation of having this metric is that the geographic coordinates of a place in a gazetteer used for geoparsing may be different from the annotated coordinates. Thus, an error distance can exist even when a geoparser correctly resolves a toponym to the right place instance. Accuracy@161 considers the result as correct as long as the resolved location is within 100 miles of the annotated location. This metric is calculated as below.

$$Accuracy@161 = \frac{|\{ed_i | ed_i = \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}, i \in [1, N], ed_i \leqslant 161\ km\}|}{N} \tag{7}$$

- *Area Under the Curve (AUC).* AUC is a metric that quantifies the overall deviation between geoparsed locations and ground-truth annotations. AUC is computed by first plotting a curve of the normalized log error distance and then calculating the total area under the curve. AUC was used in previous studies, such as (Jurgens et al., 2015; Gritta et al., 2018c,b). Figure 3 shows an example of the error distance curve. The horizontal axis represents the index of the toponyms ranked from small to large error distances. A majority of toponyms are typically located at the correct locations, and therefore have errors as zero. However, once the error distance starts to appear, it can increase rapidly. The vertical axis represents the normalized log error distance of the geoparsed toponyms. AUC is the total area under the curve calculated using Equation 8, where *Max_Error* is the maximum possible error distance (half of the Earth's circumference) between the ground truth and the geoparsed location. A better geoparser should have a lower AUC.

$$AUC = \int_{i=1}^{N} \frac{\ln(ed_i + 1)}{\ln(Max\_Error)} di \tag{8}$$

Calculating the error distance is a fundamental step for the four metrics above. Currently,

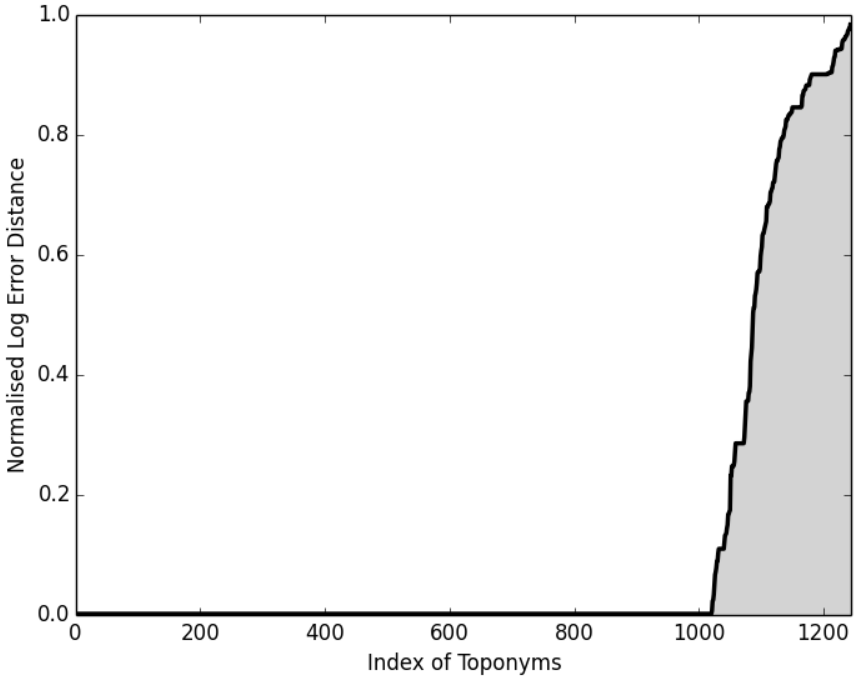

Figure 3: An illustration of AUC for quantifying the overall error distance of a geoparser.

the error distance is calculated based on point locations only. This is because all the geographic corpora we have reviewed contain only point-based annotations, and all the discussed geoparsing systems only output point-based locations. To some degree, this point-based annotation and geoparsing facilitate the comparison of geoparsing outputs and ground-truth annotations: it may be more difficult to reach an agreement on how to compare a geoparsing output that contains points, lines, and polygons with a ground-truth dataset that contains, e.g., points only. However, geo-locating a toponym to a single point is not ideal when the toponym refers to a large geographic area (e.g., a country or a river). Geographic scale further complicates this issue: it may be fine to locate a city name to a point if a study focuses on the country scale, but we may want a city name or even a neighborhood name to be represented as a polygon if the study is at the city scale. While existing geoparsers only output point-based locations, future geoparsers could provide other geometries to represent the spatial footprints of the recognized toponyms. When geoparsers and annotated datasets with various footprints have become available, EUPEG could be extended with error distances calculated using other methods such as Fréchet distance and Hausdorff distance.

In sum, EUPEG provides eight metrics for quantifying the performances of geoparsers. Four of these metrics examine the percentage of the toponyms correctly recognized from texts, while the other four metrics are based on the error distances between the geoparsed locations and their ground-truth locations. There also exist ranking-based metrics, such as Mean Reciprocal Rank (MRR) and normalized Discounted Cumulative Gain (nDCG) (Purves et al., 2018). However, these metrics require a geoparser to output a ranked list of candidate places. Some end-to-end systems only output one single result rather than a list

of places. Thus, these ranking-based measures are not included.

3.5. Resource unification

The corpora and geoparsers hosted on EUPEG are from different sources and are highly heterogeneous. They can be annotated using different data formats. For example, LGL, GeoVirus, TR-News, GeoWebNews, and WikToR use Extensible Markup Language (XML), and organize data into hierarchical structures; Hu2014 and Ju2016 use TXT or Comma-Separated Values (CSV), and organize data using simple line-by-line text annotations with each line representing one data record; GeoCorpora uses the format of Tab-Separated Values (TSV) and one data record can be put into multiple lines if it contains more than one toponym. The outputs of geoparsers are also in different formats. GeoTxt and Yahoo! PlaceSpotter use JavaScript Object Notation (JSON) to format their outputs; the Edinburgh Geoparser employs XML; TopoCluster uses its own text-based geoparsing output; and CLAVIN provides an API that allows a user to format the output in a customized manner. Even if the resources are in the same format, they can still use different vocabularies to organize similar content (which will be discussed in Section 3.6).

EUPEG serves as a platform for unifying these heterogeneous geoparsing resources. Building on the foundational work of Gritta et al. (2018c), we unify these resources in the following steps. First, we write a customized computer script for each geographic corpus to convert it into two parts: a collection of individual text files (with each file containing one text entry) and a single ground-truth text file (with each line containing the ground-truth annotation for one file in the collection). Such a design was used in Gritta et al. (2018c). While it seems to be rather an engineering design, we re-use it since EUPEG is built on the work of Gritta et al. (2018c) and doing so can avoid reinventing the wheel. Second, we write a customized wrapper for each geoparser hosted on EUPEG. These wrappers convert the heterogeneous geoparsing outputs into the same format in which each line contains the recognized toponyms from one text file. Third, a comparison function is developed to compare the standardized geoparsing outputs with the ground-truth files, and measure the performances of the geoparsers by computing the eight metrics. In sum, EUPEG unifies the heterogeneous resources by first converting them into the same formats and then comparing the performances of geoparsers based on the same metrics.

3.6. Resource extension

The resources on EUPEG can be extended with new corpora and geoparsers. A newly created geographic corpus can be uploaded to EUPEG for testing the performances of geoparsers. To enable the upload of any new corpus to EUPEG, we need an agreed format for organizing the text entries and ground-truth annotations in a new corpus. Although some toponym annotation languages, such as TRML (Leidner, 2006) and SpatialML (Mani et al., 2010), have been proposed, many publicly shared corpora, such as LGL and WikToR, use their own formats, probably due to a lack of access to example datasets of TRML and SpatialML. Here, we specify the format of a new corpus to be connected to EUPEG based on LGL, GeoVirus, TR-News, GeoWebNews, and WikToR. Although these five corpora are all in XML format, they employ different XML tags for organizing their content. For example, GeoVirus uses the XML tag $\langle article \rangle$ to represent each text entry, while WikToR uses the tag $\langle page \rangle$ to represent each entry (since the text entries are Wikipedia pages). Similarly,

TR-New uses the tag ⟨*gaztag*⟩ to provide location information obtained from a gazetteer, while GeoWebNews does not use the tag ⟨*gaztag*⟩ at all. Learning from these existing corpora, we build an XML format that has a small number of required core tags and offers the flexibility of including optional tags. Listing 1 shows this format.

```
<?xml version="1.0" encoding="utf-8"?>
<entries>
  <entry>
     <text>Paris is a city in Texas...</text>
     <toponyms>
       <toponym>
         <start>0</start>
         <end>4</end>
         <phrase>Paris</phrase>
         <place>
            <footprint>-95.5477 33.6625</footprint>
            <placename>City of Paris</placename>  #optional
            <placetype>ADM3</placetype>  #optional
        </place>
        ...  # other optional attributes
      </toponym>
      <toponym>
         ...  # another annotated toponym
      </toponym>
      ...
    </toponyms>
  </entry>
  <entry>
    ...  # another entry in the dataset
  </entry>
                ...
</entries>
```

Listing 1: The format for a new corpus to be uploaded to EUPEG.

A corpus to be uploaded to EUPEG will be organized into one XML file using the format above. This file can contain multiple text ⟨*entries*⟩, and the ⟨*entry*⟩ tag is used to organize each individual data entry. The ⟨*text*⟩ tag contains the text to be geoparsed, which can be a news article, a tweet, a Web page, or others. The ⟨*toponyms*⟩ tag contains the toponyms in the ground-truth annotation. For each ground-truth ⟨*toponym*⟩, it should contain the ⟨*start*⟩ position (in character index) and the ⟨*end*⟩ position of the toponym in the text. The ⟨*phrase*⟩ tag contains the name of the place mentioned in the text, which can be not only an official name but also a name abbreviation, a colloquial name, or other aliases. The ⟨*place*⟩ tag contains the annotated information of the place. The required core information for a ⟨*place*⟩ is ⟨$footprint$⟩ which is in the form of longitude and latitude. Other optional information, such as ⟨*placename*⟩ and ⟨*placetype*⟩, can also be included.

A newly developed geoparser can be connected to EUPEG and compared with other hosted geoparsers. To do so, one needs to make the new geoparser accessible via a REST API and organize the geoparsing output using an agreed format. As far as we know, there is no standard way for organizing geoparsing output. Accordingly, we specify the output

format based on that of an existing geoparser, GeoTxt. As an academic geoparser, GeoTxt is available freely and publicly with a REST API and accompanied by scholarly publications (Karimzadeh et al., 2013, 2019). GeoTxt uses JSON to format its output. The original output of GeoTxt contains elements specific to the used gazetteer, GeoNames, such as *geoNameId* and *featureCode*. These elements are not required in this format since a new geoparser may not necessarily employ GeoNames as its gazetteer. Similar to the format of a new corpus, we also classify the information elements in the geoparsing output as required core elements and optional ones. A geoparser can output only the four core elements for simple implementation, or can include additional and optional information for a comprehensive output. The proposed output format for a new geoparser is shown in Listing 2.

This format organizes the geoparsing output into a JSON object. It starts with a root attribute *toponyms* whose value is an array of JSON objects. Each JSON object contains the information for a toponym recognized by the geoparser. The attribute *start* contains the start position (in character index) of the toponym, while the attribute *end* contains the end position of the recognized toponym. The attribute *phrase* represents the toponym mentioned in the text which could be an official name or other alternative names. The attribute *place* contains more detailed information about the recognized place. The required element is *footprint* which takes the value of a JSON array following the format of GeoJSON. For a typical point-based footprint, the JSON array contains the longitude and latitude of the place. Other optional information, such as *placename* and *placetype*, can also be included.

```
{
  toponyms:
  [
        {
                start:0,
                end:4,
                phrase:"Paris",
                place:
                {
                        footprint:[[-95.5477,33.6625]],
                        placename:"City of Paris", # optional
                        placetype:"ADM3" # optional
                },
                ...  # other optional attributes
        },
        {
                ...  # another recognized toponym

        },
        ...
  ]
}
```

Listing 2: The format for the output of a new geoparser to be connected to EUPEG.

3.7. Experiment archiving and search

Another important function of EUPEG is archiving experiments. A database is created to store information about an experiment, such as *Experiment ID*, *Date and Time*, *Datasets*,

Geoparsers, *Metrics*, and *Experiment Results*. An experiment ID is a 16-digit serial number that uniquely identifies an experiment. All other information is based on the configurations specified by an user at the time of running an experiment. One can search experiments based on their IDs and see their results.

The value of this function can be seen in two aspects. First, it facilitates the sharing of experiment results. A researcher or a geoparser user can quickly share the result of an experiment with colleagues by embedding the experiment ID in, e.g., an email. The colleagues who receive this experiment ID can check it on EUPEG and see the experiment results and configurations themselves. Second, the independently-recorded experiment results provide further evidence for researchers to demonstrate their work, and allow others to verify the outcome of a study more easily. Accordingly, EUPEG can help enhance the reproducibility and replicability of scientific research.

3.8. Summary

We have presented the overall architecture, resources, and functions of EUPEG. In summary, EUPEG has the following features:

- **Comprehensiveness.** EUPEG provides eight annotated corpora, nine geoparsing systems, and eight performance metrics for evaluating geoparsers. The annotated corpora are in four different text genres; the geoparsing systems include both specialized geoparsers and those extended from general named entity recognizers; and the performance metrics include both information retrieval based metrics and those based on error distances.

- **Unification.** EUPEG can be considered as a one-stop platform where corpora, geoparsers, and performance metrics are unified. EUPEG also unifies geoparser users and geoparser researchers: users can use EUPEG to select the most suitable geoparser for their own corpora, while researchers can leverage the hosted resources to perform effective and efficient evaluation experiments.

- **Extensibility.** EUPEG offers extensibility for the hosted geoparsing resources. A newly created corpus can be uploaded to EUPEG for testing the hosted geoparsers. A newly developed geoparser can be connected to EUPEG and compared with other geoparsers. We also provide the source code of EUPEG, and one can further extend EUPEG by adding new performance metrics or other features for evaluating geoparsers.

- **Documentation.** EUPEG documents experiment results and configurations, and provides a search function for retrieving previous experiments. Such an archiving feature provides researchers with further evidence to demonstrate their research outcome. It also enables researchers and users to share experiment results more easily, e.g., by embedding the experiment ID in an email.

4. Implementation and Analytical Evaluation

4.1. Implementation and demonstration

Based on the proposed architecture, we have implemented EUPEG as a Web-based platform that can be accessed online at: `https://geoai.geog.buffalo.edu/EUPEG`.

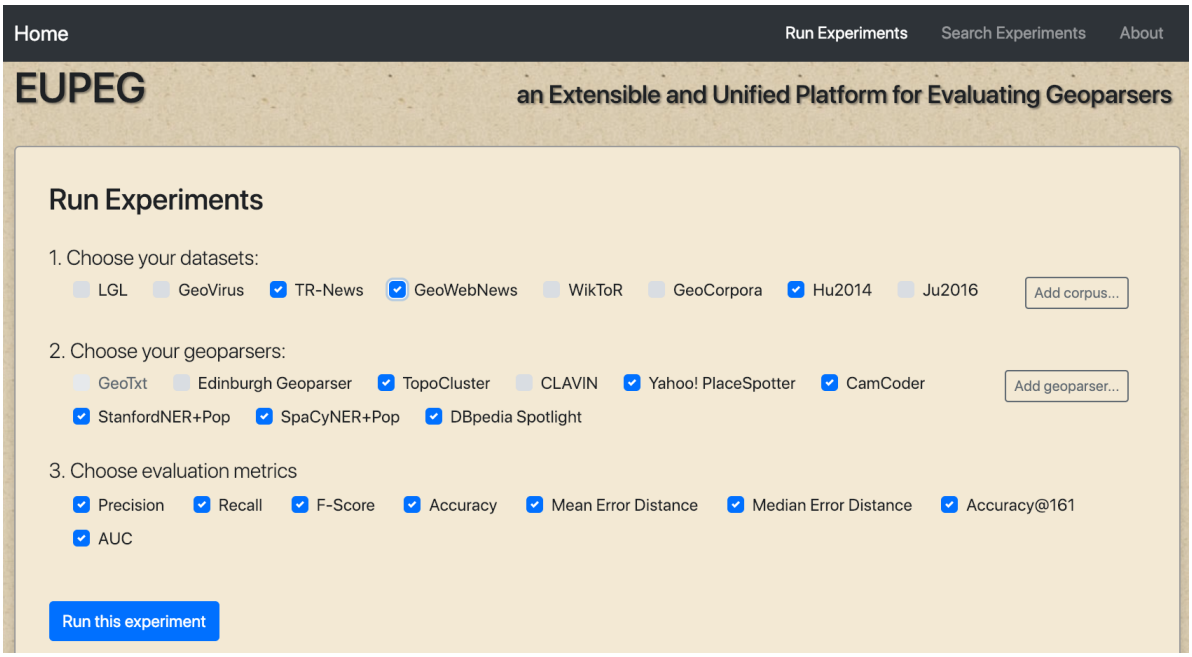

Figure 4: A screenshot of EUPEG and the (1)-(2)-(3) workflow for running an experiment.

Figure 4 shows a screenshot of its main interface. EUPEG offers a (1)-(2)-(3) workflow for conducting an experiment: a user selects (1) datasets, (2) geoparsers, and (3) metrics, and then clicks the "Run this experiment" button to start the experiment (Figure 4). One can also click the "Add corpus..." or "Add geoparser..." buttons to add their own resources. Once an experiment is finished, the user will be provided with an experiment ID which can then be used by the user or others to search for this experiment. Figure 5 shows an example of searching a previous experiment and seeing its results. The returned results contain not only the performance information of the compared geoparsers based on the selected corpora and metrics, but also the date and time of this experiment and the versions of the geoparsers and their used gazetteers. Such information allows one to see the detailed configuration of an experiment. A reader can also try this example by searching the experiment ID "8380NII17XEKM0GD" on EUPEG.

EUPEG is implemented using a technology stack of multiple programming languages, software libraries, and development tools. Java JDK 11 is used on the server side for implementing servlets, database connections, and external API requests. Javascript , HTML5, CSS3, and other libraries, such as Bootstrap and JQuery, are employed on the client side for implementing the user interface and AJAX-based HTTP requests and responses. SQLite 3 is used for storing the experiment records, which is a light-weight, high-reliability, and public-domain database. To reduce the time of experiments and avoid running the same experiment many times, we store and re-use the results if the datasets and geoparsers selected by a user were tested in a previous experiment. Such an implementation increases experiment efficiency and decreases computational cost, since running an experiment can take from hours to days depending on the selected datasets and geoparsers. In addition, we use the following approaches to keep the hosted geoparsers up-to-date. For the geoparsers that are connected to EUPEG via online APIs, a computational thread is developed

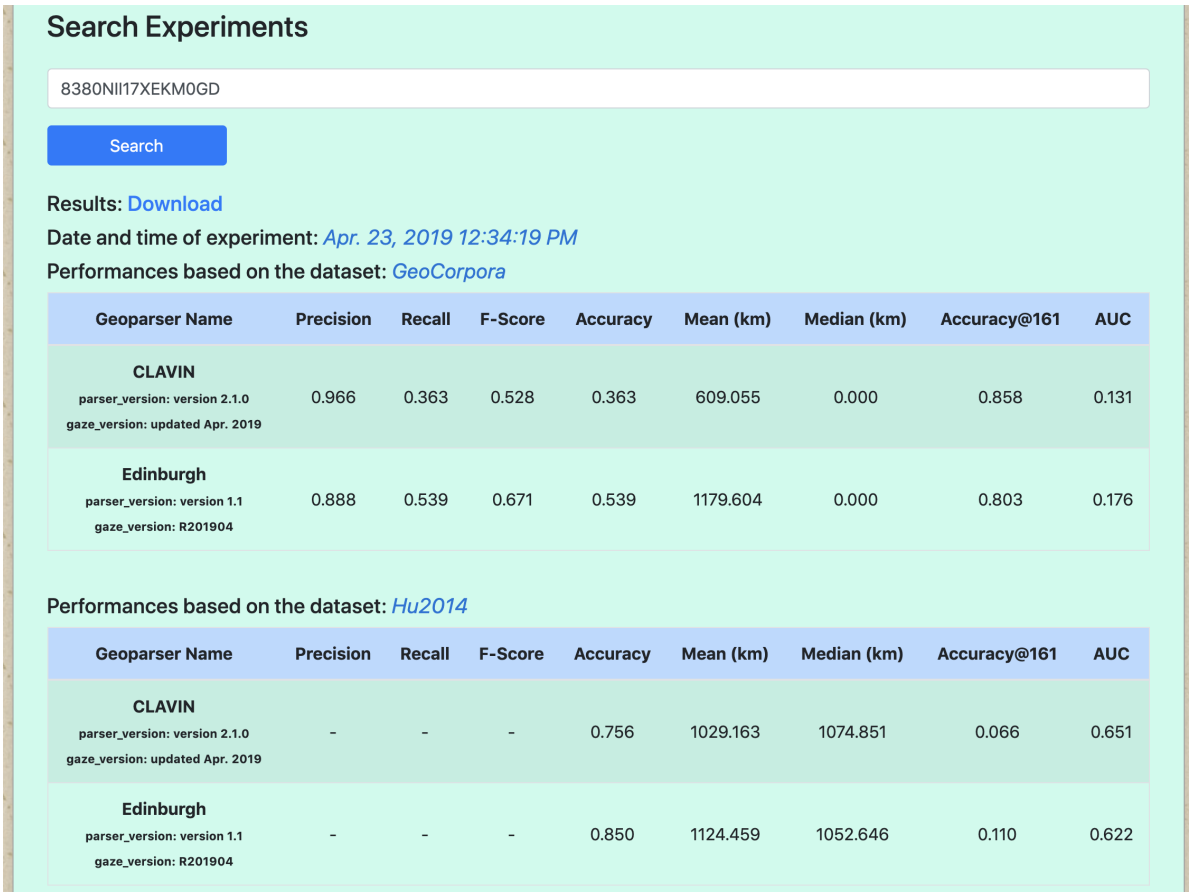

Figure 5: Searching a previous experiment and seeing its result.

which runs in the background and automatically updates the geoparsing results of these geoparsers once per month to reflect any possible changes. For the geoparsers that are deployed locally on our server, we plan to check their websites once every three to six months and will update our local instances when new stable versions have become available. While we plan to maintain EUPEG for the next few years, resource limitation may not allow us to maintain it for a long time. Thus, we also share the source code of EUPEG, along with the datasets under permitted licenses (e.g., GNU General Public License), on GitHub at: `https://github.com/geoai-lab/EUPEG`, and invite the community to further enhance and extend it.

4.2. Analytical evaluation

A main goal of EUPEG is to reduce the time that researchers have to spend in preparing datasets and baselines for experiments. This section attempts to estimate the amount of the time that could be saved by EUPEG. One possible approach to providing such an estimate is to invite a number of researchers, ask them to prepare all the corpora and geoparsers hosted on EUPEG by themselves, and measure the average time they spend. Such a process, however, can be very tedious for the invited researchers, and depending on their particular fields and technical skills, their used time may not represent the time that others may need for preparing these experiment resources. Here, we provide an analytical evaluation on the

amount of time that could be saved based on our own experience of developing EUPEG and focus on the *lower bound* of the time. In the following, we first analyze the steps that a research group typically has to complete if they were to prepare datasets and baselines for an experiment themselves. These identified steps are shown in Figure 6. We then estimate the minimum amount of time that is necessary to complete each step.

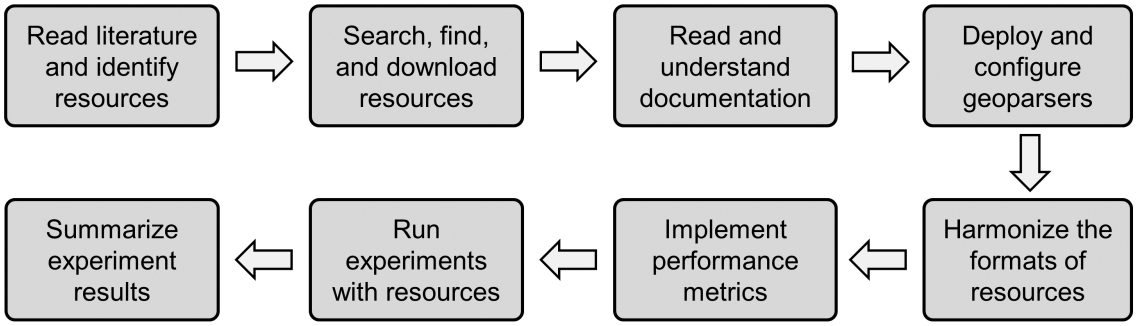

Figure 6: Typical steps for preparing datasets and geoparsers for experiments.

Read literature and identify resources. This is generally the first step, in which one studies previous research and identifies resources that can be re-used. For this step, EUPEG does not save much time. Although it makes various resources ready for use, researchers may still need to read the related publications to understand the methods under the hood. EUPEG and this paper, however, can serve as an entry point for new researchers. The time that can be saved in this step is estimated as zero.

Search, find, and download resources. After identifying resources from the literature, one needs to obtain them. For most datasets hosted on EUPEG, we were able to obtain each of them within half an hour, thanks to the authors who shared relevant URLs in their papers. For GeoCorpora, while the authors have kindly provided its URL, much time is still needed to rehydrate this dataset due to the data sharing restriction of Twitter. It took us more than 5 person-hours to recover this dataset, and additional time has to be spent in applying for a Twitter developer account before one can start to recover the dataset. We estimate a minimum of 8.5 person-hours for obtaining the datasets hosted on EUPEG. For the geoparsers, we were able to download the source codes or compiled versions of the Edinburgh Geoparser, TopoCluster, CLAVIN, and CamCoder within half an hour each. The other five geoparsers are either connected to EUPEG via their APIs or are further developed based on general NER tools. About half an hour is needed for finding each of these resources. In total, about 13 person-hours are needed for this step.

Read and understand documentation. After the source codes of previous geoparsers are obtained, one needs to read documents and understand how to deploy and run them. Background knowledge on different programming languages (e.g., Python and Java) and system architectures (e.g., REST Web services) is necessary for understanding the installation instructions. Based on our own experience, we estimate an average of two person-hours for an experienced developer to read and understand the documents of one geoparsing system. Thus, this step takes about 18 person-hours.

Deploy and configure geoparsers. This step is particularly time-consuming and requires a lot of expertise. First, different geoparsers can be implemented in different programming

languages. Accordingly, a researcher needs to have some basic knowledge on the multiple languages in order to deploy them. Second, there exist specific configuration requirements for some geoparsers. For example, geoparsers available via Web APIs require one to have the expertise of handling HTTP requests and responses; a geoparser (e.g., TopoCluster) may require the installation of a database and its spatial extension, or may require a researcher to be familiar with certain deep learning libraries (e.g., CamCoder). Third, including general NER tools as baselines requires further developments and gazetteer configurations to convert these general tools into geoparsers. We estimate an average of 24 person-hours to successfully deploy and configure one geoparser and thus a total of 216 person-hours.

Harmonize the formats of resources. The annotated datasets and the outputs of geoparsers are often in different formats and structures. To conduct an experiment on these heterogeneous resources, one needs to harmonize these datasets and geoparser outputs by writing programs to convert them into the same format. We estimate an average of three person-hours for processing one resource (a dataset or a geoparser), and in total, this step takes about 51 person-hours.

Implement performance metrics. Performance metrics, such as precision, recall, and AUC, need to be implemented for evaluating geoparsers. In addition, some programming work is necessary for comparing geoparsing outputs to ground-truth annotations. In total, we estimate eleven person-hours for completing this step (eight hours for eight metrics plus three hours for developing the code for comparing outputs with ground-truth annotations).

Run experiments with resources. Once everything is prepared , we can run experiments to obtain evaluation results. The running time of different geoparsers can vary largely. Figure 7 reports the empirical time of the nine geoparsers on the same machine for processing the GeoCorpora dataset. TopoCluster took the longest time (660.51 minutes), while CLAVIN is the fastest geoparser that took only 0.28 minute to process the same dataset. Longer pro-

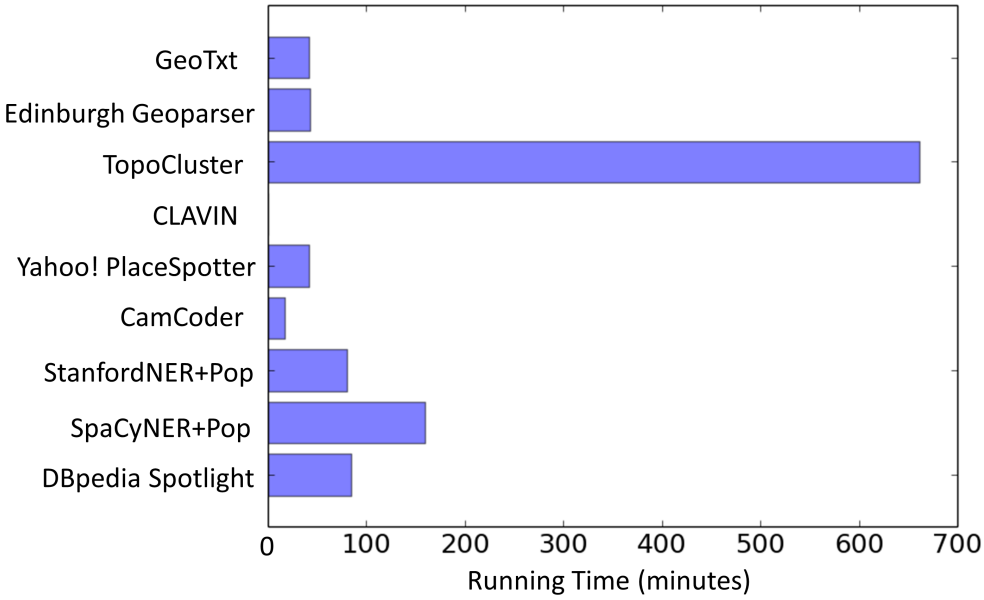

Figure 7: Running time of different geoparsers on GeoCorpora.

cessing time, however, does not mean better performance. Figure 8 shows the performances

of different geoparsers on GeoCorpora. The time that can be saved by EUPEG in this step is estimated as zero, since one can work on other tasks when an experiment is running.

Geoparser Name	Precision	Recall	F-Score	Accuracy	Mean (km)	Median (km)	Accuracy@161	AUC
GeoTxt	0.978	0.550	0.704	0.550	786.140	0.000	0.848	0.137
Edinburgh	0.888	0.539	0.671	0.539	1179.604	0.000	0.803	0.176
TopoCluster	0.950	0.545	0.693	0.545	746.734	38.764	0.670	0.381
CLAVIN	0.966	0.363	0.528	0.363	609.055	0.000	0.858	0.131
Yahoo	0.871	0.640	0.738	0.640	564.471	49.980	0.681	0.390
CamCoder	0.931	0.518	0.666	0.518	1095.508	0.000	0.790	0.186
StanfordNER	0.970	0.568	0.716	0.568	1269.950	0.456	0.649	0.296
SpaCyNER	0.799	0.530	0.637	0.530	1372.139	0.000	0.690	0.279
DBpedia	0.912	0.526	0.667	0.526	760.484	33.816	0.646	0.356

Figure 8: The performances of different geoparsers on GeoCorpora.

Summarize experiment results. When experiments are finished, one often needs to collect the obtained results and organize them into a report. For this step, we estimate the saved time as zero, since it has to be done with or without EUPEG.

Table 4 provides a summary of the approximate number of person-hours that can be saved by EUPEG.

Table 4: The estimated amount of time that can be saved by EUPEG.

Task for preparing experiments	Estimated time (person-hours)
Read literature and identify resources	0
Search, find, and download resources	13
Read and understand documentation	18
Deploy and configure geoparsers	216
Harmonize the formats of resources	51
Implement performance metrics	11
Run experiments with resources	0
Summarize experiment results	0
Total	309

In total, we estimate 309 person-hours if another research group were to prepare the same resources hosted on EUPEG. This estimate is close to a *lower bound*, as it is based on the assumption that researchers have all the necessary knowledge and technical skills and does not take into account the time spent on trials and errors.

5. Conclusions and Future Work

In this work, we present EUPEG, an Extensible and Unified Platform for Evaluating Geoparsers. With large amounts of textual data available from various sources, geoparsers have become increasingly important given their capabilities of extracting geographic information from textual documents. Many studies in spatial data science and digital humanities have leveraged geoparsers under various contexts (e.g., disaster responses, platial studies, and event detection) to integrates spatial and textual analysis. While a number of geoparsers were developed, they were tested on different datasets using different performance metrics. Consequently, the reported evaluation results cannot be directly compared. In addition, a new geoparser often needs to be compared with existing baselines to demonstrate its merits. However, preparing baselines and testing datasets can take much time and effort from different research groups. In this context, we propose and develop EUPEG as a benchmarking platform for evaluating geoparsers and eventually enhancing spatial and textual analysis. It is implemented as a Web based and open source platform with four major features. (1) Comprehensiveness: EUPEG provides eight open corpora, nine geoparsing systems, and eight performance metrics for evaluating geoparsers; (2) Unification: EUPEG can be considered as a one-stop platform where heterogeneous corpora, geoparsers, and metrics are unified; (3) Extensibility: EUPEG allows the hosted resources to be extended with new corpora and geoparsers; (4) Documentation: EUPEG documents experiment results and configurations, and allows the search of previous experiments.

The main goal of EUPEG is to enable effective and efficient comparisons of geoparsers while reducing the time that researchers and end users have to spend in preparing datasets and baselines. Based on our analytical evaluation, EUPEG can save one single research group approximately 309 hours for preparing the same datasets, geoparsers, and metrics. The number of saved hours will be multiplied when multiple research groups attempt to develop and compare geoparsers. While EUPEG serves as a benchmarking platform, it is not to replace project-specific evaluations necessary for highlighting certain unique features of a geoparser but to supplement existing evaluations.

The development of EUPEG also reveals several issues that may need future work. First, there is a lack of commonly-agreed standards on corpus annotation. While languages, such as TRML (Leidner, 2006) and SpatialML (Mani et al., 2010), were proposed, they were not adopted by the recently available and publicly shared corpora, such as LGL, WikToR, and GeoCorpora. Having a commonly-agreed standard can facilitate the development of datasets that are more readily usable by others in future experiments. Second, a similar situation happens to the outputs of geoparsers, where a commonly-agreed output format is not available. Most geoparsers organize their outputs in a format that they consider suitable. While it is feasible to harmonize these heterogeneous outputs by writing wrapper programs (as done in EUPEG), a standard output format can make it easier for others to use a geoparser or to combine multiple geoparsers. In this work, we have developed a simple format based on GeoTxt to allow new geoparsers to be connected to EUPEG. However, further efforts are needed from the community to develop an agreed and standard output format for geoparsers. Third, the current version of EUPEG focuses on English-based geoparsers and corpora only. Resources for other languages could be added in the future to support multilingual geoparsing evaluations. With the source code shared, new extensions

could be added to EUPEG to further enhance it and help it better serve our community.

Acknowledgments

The authors would like to thank Dr. Morteza Karimzadeh and Dr. Alan M. MacEachren for providing further technical information about GeoTxt. We thank the anonymous reviewers for their constructive comments and suggestions.

References

Adams, B., McKenzie, G., 2013. Inferring thematic places from spatially referenced natural language descriptions. In: Crowdsourcing geographic knowledge. Springer, pp. 201–221.

Alex, B., Byrne, K., Grover, C., Tobin, R., 2015. Adapting the Edinburgh geoparser for historical georeferencing. International Journal of Humanities and Arts Computing 9 (1), 15–35.

Amitay, E., Har'El, N., Sivan, R., Soffer, A., 2004. Web-a-where: geotagging web content. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA, pp. 273–280.

Ballatore, A., Adams, B., 2015. Extracting place emotions from travel blogs. In: Proceedings of AGILE. Vol. 2015. pp. 1–5.

Barbaresi, A., 2017. Towards a toolbox to map historical text collections. In: Proceedings of the 11th Workshop on Geographic Information Retrieval. ACM, p. 5.

Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.-S., 2014. Making sense of microposts: (#microposts2014) named entity extraction & linking challenge. In: CEUR Workshop Proceedings. Vol. 1141. pp. 54–60.

Cheng, Z., Caverlee, J., 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. ACM, pp. 759–768.

Choi, J. D., Tetreault, J., Stent, A., 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Vol. 1. pp. 387–396.

Cornolti, M., Ferragina, P., Ciaramita, M., 2013. A framework for benchmarking entity-annotation systems. In: Proceedings of the 22nd international conference on World Wide Web. ACM, pp. 249–260.

Daiber, J., Jakob, M., Hokamp, C., Mendes, P. N., 2013. Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems. ACM, pp. 121–124.

DeLozier, G., Baldridge, J., London, L., 2015. Gazetteer-independent toponym resolution using geographic word profiles. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, Palo Alto, CA, USA, pp. 2382–2388.

DeLozier, G., Wing, B., Baldridge, J., Nesbit, S., 2016. Creating a novel geolocation corpus from historical texts. In: Proceedings of The 10th Linguistic Annotation Workshop. Association for Computational Linguistics, pp. 188–198.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 248–255.

Faulconbridge, J. R., Hall, S. J., Beaverstock, J. V., 2008. New insights into the internationalization of producer services: organizational strategies and spatial economies for global headhunting firms. Environment and Planning A 40 (1), 210–234.

Ferragina, P., Scaiella, U., 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, pp. 1625–1628.

Frank, J. R., Rauch, E. M., Donoghue, K., October 2006. Spatially coding and displaying information. US Patent 7,117,199.

Freire, N., Borbinha, J., Calado, P., Martins, B., 2011. A metadata geoparsing system for place name recognition and resolution in metadata records. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. ACM, New York, NY, USA, pp. 339–348.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., Yan, B., 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science 31 (6), 1245–1271.

Geiß, J., Spitz, A., Strötgen, J., Gertz, M., 2015. The wikipedia location network: overcoming borders and oceans. In: Proceedings of the 9th workshop on geographic information retrieval. ACM, p. 2.

Gelernter, J., Balaji, S., 2013. An algorithm for local geoparsing of microtext. GeoInformatica 17 (4), 635–667.

Gelernter, J., Mushegian, N., 2011. Geo-parsing messages from microtext. Transactions in GIS 15 (6), 753–773.

Gritta, M., Pilehvar, M. T., Collier, N., 2018a. A pragmatic guide to geoparsing evaluation. arXiv preprint arXiv:1810.12368.

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

Grossner, K., Janowicz, K., Keßler, C., 2016. Place, period, and setting for linked data gazetteers. Placing Names: Enriching and Integrating Gazetteers, 80–96.

Hecht, B., Moxley, E., 2009. Terabytes of tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. In: International conference on spatial information theory. Springer, pp. 88–105.

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

Hu, Y., 2018. EUPEG: Towards an Extensible and Unified Platform for Evaluating Geoparsers. In: Proceedings of the 12th Workshop on Geographic Information Retrieval.

GIR'18. ACM, New York, NY, USA, pp. 3:1–3:2.

Hu, Y., Janowicz, K., Prasad, S., 2014. Improving Wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In: Proceedings of the 8th workshop on geographic information retrieval. ACM, pp. 1–8.

Hu, Y., Ye, X., Shaw, S.-L., 2017. Extracting and analyzing semantic relatedness between cities using news articles. International Journal of Geographical Information Science 31 (12), 2427–2451.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., Ghazi, D., 2017. Location detection and disambiguation from twitter messages. Journal of Intelligent Information Systems 49 (2), 237–253.

Jiang, R., Banchs, R. E., Li, H., 2016. Evaluating and combining name entity recognition systems. In: Proceedings of the Sixth Named Entity Workshop. pp. 21–27.

Jones, C. B., Purves, R. S., 2008. Geographical information retrieval. International Journal of Geographical Information Science 22 (3), 219–228.

Jones, C. B., Purves, R. S., Clough, P. D., Joho, H., 2008. Modelling vague places with knowledge from the web. International Journal of Geographical Information Science 22 (10), 1045–1065.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., McKenzie, G., 2016. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In: European Knowledge Acquisition Workshop. Springer, pp. 353–367.

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. ACM, New York, NY, USA, p. 8.

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. ACM, New York, NY, USA, pp. 72–73.

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., Wallgrün, J. O., 2019. Geotxt: A scalable geoparsing system for unstructured text geolocation. Transactions in GIS 23 (1), 118–136.

Ladra, S., Luaces, M. R., Pedreira, O., Seco, D., 2008. A toponym resolution service following the ogc wps standard. In: International Symposium on Web and Wireless Geographical Information Systems. Springer, pp. 75–85.

Lan, R., Adelfio, M. D., Samet, H., 2014. Spatio-temporal disease tracking using news articles. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health. ACM, pp. 31–38.

Leidner, J. L., 2006. An evaluation dataset for the toponym resolution task. Computers,

Environment and Urban Systems 30 (4), 400–417.

Leidner, J. L., 2008. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Universal-Publishers, Irvine, CA, USA.

Li, H., Srihari, R. K., Niu, C., Li, W., 2002. Location normalization for information extraction. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, pp. 1–7.

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). IEEE, Long Beach, CA, USA, pp. 201–212.

Liu, Y., Wang, F., Kang, C., Gao, Y., Lu, Y., 2014. Analyzing relatedness by toponym co-occurrences on web pages. Transactions in GIS 18 (1), 89–107.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., Blanford, J., 2011. Senseplace2: Geotwitter analytics support for situational awareness. In: Visual analytics science and technology (VAST), 2011 IEEE conference on. IEEE, pp. 181–190.

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

Mani, I., Hitzeman, J., Richer, J., Harris, D., 2008. ACE 2005 english spatialML annotations. Linguistic Data Consortium, Philadelphia.

McKenzie, G., Liu, Z., Hu, Y., Lee, M., 2018. Identifying urban neighborhood names through user-contributed online property listings. ISPRS International Journal of Geo-Information 7 (10), 388.

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

Mendes, P. N., Jakob, M., García-Silva, A., Bizer, C., 2011. Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. ACM, pp. 1–8.

Monteiro, B. R., Davis Jr, C. A., Fonseca, F., 2016. A survey on the geographic scope of textual documents. Computers & Geosciences 96, 23–34.

Nesi, P., Pantaleo, G., Tenti, M., 2016. Geographical localization of web domains and organization addresses recognition by employing natural language processing, pattern matching and clustering. Engineering Applications of Artificial Intelligence 51, 202–211.

Overell, S., Rüger, S., 2008. Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science 22 (3), 265–287.

Pezanowski, S., MacEachren, A. M., Savelyev, A., Robinson, A. C., 2018. Senseplace3: a geovisual framework to analyze place–time–attribute information in social media. Cartography and Geographic Information Science 45 (5), 420–437.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., Murdock, V., et al., 2018. Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends® in Information Retrieval 12 (2-3), 164–318.

Richter, L., Geiß, J., Spitz, A., Gertz, M., 2017. Heidelplace: An extensible framework for geoparsing. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 85–90.

Salvini, M. M., Fabrikant, S. I., 2016. Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design 43 (1), 228–248.

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

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

Sundheim, B. M., 1993. Tipster/MUC-5: information extraction system evaluation. In: Proceedings of the 5th conference on Message understanding. Association for Computational Linguistics, pp. 27–44.

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. Association for Computational Linguistics, pp. 142–147.

Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S., Both, A., 2014. Agdistis-graph-based disambiguation of named entities using linked data. In: International Semantic Web Conference. Springer, pp. 457–471.

Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al., 2015. Gerbil: general entity annotator benchmarking framework. In: Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1133–1143.

Van Erp, M., Rizzo, G., Troncy, R., 2013. Learning with the web: Spotting named entities on the intersection of nerd and machine learning. In: Proceedings of the 3rd workshop on Making Sense of Microposts (#MSM'13). pp. 27–30.

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.

Woodruff, A. G., Plaunt, C., 1994. GIPSY: Automated geographic indexing of text documents. Journal of the American Society for Information Science 45 (9), 645–655.