Are We There Yet? Evaluating State-of-the-Art Neural Network based Geoparsers Using EUPEG as a Benchmarking Platform

Jimin Wang
jiminwan@buffalo.edu
GeoAI Lab, Department of Geography
University at Buffalo, New York, USA

Yingjie Hu
yhu42@buffalo.edu
GeoAI Lab, Department of Geography
University at Buffalo, New York, USA

ABSTRACT

Geoparsing is an important task in geographic information retrieval. A geoparsing system, known as a *geoparser*, takes some texts as the input and outputs the recognized place mentions and their location coordinates. In June 2019, a geoparsing competition, *Toponym Resolution in Scientific Papers*, was held as one of the SemEval 2019 tasks. The winning teams developed neural network based geoparsers that achieved outstanding performances (over 90% precision, recall, and F1 score for toponym recognition). This exciting result brings the question "are we there yet?", namely have we achieved high enough performances to possibly consider the problem of geoparsing as solved? One limitation of this competition is that the developed geoparsers were tested on only one dataset which has 45 research articles collected from the particular domain of Bio-medicine. It is known that the same geoparser can have very different performances on different datasets. Thus, this work performs a systematic evaluation of these state-of-the-art geoparsers using our recently developed benchmarking platform EUPEG that has eight annotated datasets, nine baseline geoparsers, and eight performance metrics. The evaluation result suggests that these new geoparsers indeed improve the performances of geoparsing on multiple datasets although some challenges remain.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Information systems** → **Retrieval effectiveness**; **Geographic information systems**.

KEYWORDS

Geoparsing, Long Short-Term Memory, Contextual Word Embedding, Deep Learning, Benchmarking Platform, EUPEG, GeoAI

ACM Reference Format:
Jimin Wang and Yingjie Hu. 2019. Are We There Yet? Evaluating State-of-the-Art Neural Network based Geoparsers Using EUPEG as a Benchmarking Platform. In *3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities'19), November 5, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3356991.3365470

GeoHumanities'19, November 5, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6960-2/19/11…$15.00
https://doi.org/10.1145/3356991.3365470

1 INTRODUCTION

Geoparsing is the process of recognizing and geo-locating location mentions from texts. It has been widely applied to various textual data, and is an important task in geographic information retrieval [14]. A geoparsing system, known as a geoparser, usually functions in two steps: toponym recognition and toponym resolution. Toponym recognition detects the place mentions in texts, while toponym resolution resolves any place name ambiguity and assigns the appropriate spatial footprint (e.g., a pair of coordinates). Many geoparsers have been developed, such as CLAVIN[1], the Edinburgh Geoparser [5], GeoTxt [9], and TopoCluster [2].

In June 2019, an important geoparsing competition, *Toponym Resolution in Scientific Papers*, was held as the SemEval 2019 Task 12, in conjunction with the Annual Conference of the North American Chapter of the Association for Computational Linguistics. This competition attracted 29 registered teams and 8 teams eventually submitted a system run [19]. The winning teams all leveraged state-of-the-art neural network based models, such as BiLSTM-CRF and deep contextualized word embeddings, to design their geoparsers. Particularly, the geoparser that won the first place, DM_NLP [18], achieved over 90% precision, recall, and F1 score for toponym recognition. This result is exciting and brings the question "are we there yet?" A 90% performance is not perfect but is probably sufficient for many applications. So have we already made enough progress that we can consider the problem of geoparsing as solved?

A major limitation of the SemEval 2019 Task 12 competition is that the submitted geoparsers were tested on a single dataset which has 45 research articles from one particular domain of Biomedicine. Existing research has shown that the same geoparser can have very different performances when tested on different datasets [4]. Accordingly, answering the question of whether the problem of geoparsing can be considered as solved requires a systematic evaluation of the state-of-the-art geoparsers on multiple datasets which should ideally be in different text genres (e.g., news articles, social media posts, and other types of texts).

In a recent work, we developed an online platform called EUPEG[2] which is an Extensible and Unified Platform for Evaluating Geoparsers [7, 17]. EUPEG hosts a majority of the geopasing resources reported in the literature, including eight annotated datasets, nine geoparsers, and eight evaluation metrics. In addition, the eight annotated datasets are in four different text genres which are news articles, Wikipedia articles, social media posts, and texts on Web pages. The source code of EUPEG and the related geoparsing resources are shared on GitHub[3].

[1] https://clavin.bericotechnologies.com
[2] https://geoai.geog.buffalo.edu/EUPEG
[3] https://github.com/geoai-lab/EUPEG

In this paper, we systematically evaluate the top geoparsers from SemEval Task 12 using EUPEG as a benchmarking platform. We focus on the top three end-to-end geoparsers that showed the highest performances in the competition, which are DM_NLP [18], UniMelb [11], and UArizona [22]. We test the performances of these three geoparsers on the datasets hosted on EUPEG, and compare their performances with the other existing geoparsers. The contributions of this paper are as follows:

- We conduct a systematic evaluation experiment on three state-of-the-art geoparsers, and discuss the implications and challenges based on the experiment results.
- We implement the three tested geoparsers based on their papers and share the source code at https://github.com/geoai-lab/GeoAI2019Geoparser to support future research.

2 STATE-OF-THE-ART GEOPARSERS

The top three end-to-end geoparsers from SemEval Task 12 are DM_NLP, UniMelb, and UArizona. They are all designed as pipeline systems comprising of two independent components for toponym recognition and resolution respectively. Accordingly, we describe and compare the three geoparsers based on the two components.

2.1 Toponym Recognition

All three geoparsers adopt the general Bidirectional Long Short Term Memory (BiLSTM) model for toponym recognition. However, their models vary in regard to the selection of word embeddings, integration of character-level embeddings, concatenation with a conditional random field layer, and mechanisms of self attention.

DM_NLP: This model, ranked as the 1st place, is built upon the character and word level BiLSTM model developed by Lample et al. [10]. The authors of DM_NLP also tested the strategies of adding four extra linguistic features into the input layer: Part-of-Speech (POS) tags, NER labels from Stanford NER, Chunking labels, and deep contextualized word representations from the ELMo word embeddings [13], but found that only adding ELMo produces the most performance improvement. In our implementation, we add the ELMo word embeddings as the extra linguistic feature. The final output layer of DM_NLP is a CRF layer.

UniMelb: This model is developed by integrating a word-level BiLSTM [6] and the self-attention mechanism [15]. The authors tested both the GloVe and ELMo word embeddings, and found that the model with ELMo performed better. Thus, our implementation also uses ELMo word embeddings. The final layer of UniMelb is a binary softmax classifier.

UArizona: This model is a re-implementation of a word, character, and affix level LSTM developed by Yadav et al. [21]. In this model, the input of word LSTM is a concatenation of GloVe word embeddings, char embeddings represented by the output of a char BiLSTM, and word affix features. The word LSTM representations are given to the final CRF layer to recognize toponyms.

We train all three toponym recognition models using a general dataset CoNLL 2003. The hyperparameters are set as the same as what reported in their papers. We use 300-dimensional pre-trained GloVe word embeddings and 1024 dimensional pre-trained EMLo embeddings from Tensorflow Hub (https://tfhub.dev/google/elmo/2). We do not update the weights of word embeddings during the training process.

2.2 Toponym Resolution

For toponym resolution, all three geoparsers use the same general workflow of first retrieving place candidates from the GeoNames gazetteer and then identifying the correct place instance among the candidates. However, different techniques were used by each geoparser to identify the right place instance.

DM_NLP: This model constructs four groups of features, which include name string similarity, candidate attributes, contextual features, and mention list features. These features are then used to train a LightGBM model for toponym resolution.

UniMelb: This model also constructs features, including history result in the training dataset, population, GeoNames feature codes, name similarity, and ancestor names, and trains a support vector machine (SVM) for toponym resolution.

UArizona: This model simply uses the population heuristic for toponym resolution. Each place name is resolved to the place instance that has the highest population in GeoNames.

There is a challenge for re-implementing these toponym resolution models, that is, both DM_NLP and UniMelb were trained on the specific training dataset from SemEval Task 12, which consists of 105 research articles in Bio-medicine. While this is fine and even desirable for a competition (since the testing is based on 45 research articles from the same domain), a model trained with one specific type of texts may not generalize well to other types of texts from different domains. Though we have multiple datasets available from the EUPEG platform, training the models with any of these datasets leads to the same bias issue. Ideally, the toponym resolution models of DM_NLP and UniMelb should be trained with a large and general dataset which has labeled place instances (note that CoNLL 2003 cannot be used for training toponym resolution models) so that the general performances of these models can be measured. However, we currently do not have access to such a dataset. Thus, we resort to a simple but general implementation, namely using the population heuristic of UArizona for all three models. Previous research, as well as the experiment result reported by the DM_NLP team [18], has shown that population heuristic is a competent baseline and can sometimes outperform more complex models [2, 20]. Nevertheless, we are aware of the limitations of this simple heuristic and will discuss them with the experiment results.

3 EXPERIMENTS AND RESULTS

3.1 Experiments on EUPEG

The three neural network based geoparsers are tested on EUPEG. As a benchmarking platform, EUPEG provides eight annotated corpora, nine geoparsers, and eight performance metrics. Table 1 summarizes these resources. More detailed descriptions on each of the resources can be found in our full paper about EUPEG [17]. We provide brief descriptions below to make this current paper self-contained.

The eight datasets are in four different text genres: news articles, Wikipedia articles, social media posts, and Web pages. Particularly, *LGL*, *GeoVirus*, *TR-News*, and *GeoWebNews* contain annotated news

Table 1: Datasets, geoparsers, and metrics on EUPEG

Category	Resources
Datasets	LGL, GeoVirus, TR-News, GeoWebNews, WikToR, GeoCorpora, Hu2014, Ju2016
Geoparsers	GeoTxt, The Edinburgh Geoparser, CLAVIN, Yahoo! PlaceSpotter, CamCoder, TopoCluster, StanfordNER+Population, SpaCyNER+Population, DBpedia Spotlight
Metrics	Precision, Recall, F1 score, Accuracy, Mean, Median, AUC, Accuracy@161

articles; *WikToR* is a Wikipedia dataset; *GeoCorpora* is a social media dataset that contains annotated tweets; and *hu2014* and *Ju2016* are two corpora that contain texts retrieved from Web pages. These diverse datasets enable a more comprehensive evaluation on the performance of a geoparser. It is worth noting that these datasets were annotated by researchers from different domains (e.g., geography, linguistics, and computer science). As a result, there exist differences in the words and phrases that are considered as toponyms. All datasets annotate administrative units, such as cities, towns, and countries. However, some datasets, such as *LGL* and *GeoWebNews*, also consider demonyms (e.g., Canadian) as toponyms. The toponyms in the dataset *GeoCorpora*, in addition to administrative units, also include natural features (e.g., lakes and mountains) and facilities (e.g., streets and buildings) which are not included in some other datasets (e.g., *GeoVirus*). This definition difference of toponyms directly affects the performances of the same geoparser on different datasets.

The nine geoparsers hosted on EUPEG use a variety of heuristics and machine learning based methods. Particularly, *GeoTxt*, *The Edinburgh Geoparser*, and *CLAVIN* use a named entity recognition tool for toponym recognition and a number of heuristics (e.g., the level of an administrative unit and population) for toponym resolution. *TopoCluster* uses Stanford NER for toponym recognition and generates geographic profiles of words for toponym resolution. *CamCoder* is a deep learning based geoparser that leverages a Convolutional Neural Network (CNNs) model. *Yahoo! PlaceSpotter* is an industrial geoparser which provides an online REST API (at the time of writing this paper, the online service of *Yahoo! PlaceSpotter* has become unavailable). In addition to the six geoparsers, EUPEG also includes two baseline geoparsers that are developed using Stanford NER and SpaCy NER with a population heuristic, as well as *DBpedia Spotlight*, a general named entity recognition and linking (NERL) tool that can be used as a geoparser.

The eight performance metrics provided on EUPEG include standard metrics from information retrieval as well as geographic distance based metrics designed for measuring the quality of the resolved geographic locations. The metrics of *precision*, *recall*, *F1 score* and *accuracy* evaluate the ability of a geoparser in correctly recognizing toponyms from texts. Particularly, the metric of *accuracy* is used in situations when only some of the mentioned toponyms are annotated. The metrics of *mean* and *median* measures how far the resolved location is away from the ground-truth location (in kilometers). The metric of *accuracy@161* measures the percentage of the resolved locations that are within 161 kilometers (100 miles) of the ground truth. The metric of *AUC* (Area Under the Curve) measures a normalized distance error by calculating the area under a distance error curve.

The three neural network based geoparsers from SemEval Task 12 are tested using the datasets from EUPEG. We quantify their performances using the discussed metrics, and compare their performances with those of the other geoparsers hosted on EUPEG.

3.2 Results

The experiment results contain the performances of the three state-of-the-art geoparsers on the eight datasets in comparison with the other existing geoparsers. In the following, we present and discuss the experiment results on three datasets, namely *GeoVirus*, *GeoCorpora*, and *Ju2016*. We provide the results on the other five datasets in Appendix A.

3.2.1 Results on GeoVirus. GeoVirus is a corpus that contains 229 news articles. This dataset was originally developed by Gritta et al. [3], and the news articles were collected during 08/2017 - 09/2017, covering the topics about global disease outbreaks and epidemics. GeoVirus is a relatively easy dataset since most location mentions refer to prominent place instances (e.g., major cities or countries) and the texts from news articles are well formatted. The evaluation results on GeoVirus are summarized in Table 2. Since the online service of Yahoo! PlaceSpotter has become unavailable, its performance is not included in the experiment results.

Table 2: Evaluation results on GeoVirus

Geoparser	precision	recall	f_score	mean	median	acc@161	AUC
DM_NLP+Pop	0.917	0.916	**0.917**	770.337	48.676	0.655	0.378
StanfordNER	0.927	0.903	0.915	791.296	48.676	0.655	0.378
UniMelb+Pop	0.882	**0.936**	0.908	777.234	48.466	0.657	0.379
UArizona	0.887	0.859	0.873	769.810	55.635	0.640	0.386
CamCoder	**0.940**	0.802	0.866	619.397	33.945	0.770	0.336
TopoCluster	0.877	0.813	0.844	599.632	63.858	0.673	0.407
GeoTxt	0.857	0.726	0.786	487.874	36.255	0.787	0.338
CLAVIN	0.913	0.637	0.750	522.176	35.503	0.786	0.320
DBpedia	0.792	0.616	0.693	1272.937	122.314	0.533	0.406
Edinburgh	0.860	0.559	0.678	**435.799**	**33.187**	**0.807**	**0.319**
SpaCyNER	0.721	0.382	0.499	788.231	40.653	0.698	0.367

The geoparsers in the table above are ordered by their F1 scores. The metrics of *precision*, *recall*, and *f_score* evaluate the performances of a geoparser for toponym recognition. The other four metrics evaluate the performance of a geoparser in resolving a toponym to its correct geographic location. It can be seen that the three top geoparsers from SemEval Task 12 indeed rank very high based on their F1 scores for the task of toponym recognition. However, the off-the-shelf StanfordNER also shows very competitive performance on this simple dataset. In terms of toponym resolution, *The Edinburgh Geoparser* performs the best, although the median error distance for most geoparsers are within 100 km. Since most place mentions refer to their prominent instances, the population heuristic works well. It is worth noting that toponym resolution is performed based on only the toponyms recognized in the previous step. Thus, the metrics of *mean*, *median*, *acc@161*, and *AUC* are measured based on different numbers of toponyms that need to be resolved.

3.2.2 Results on GeoCorpora. GeoCorpora is a social media corpus that contains annotated tweets. GeoCorpora was developed by Wallgrün et al. [16], and their original paper reported 2,287 annotated tweets. Due to deletions, only 1,639 tweets are recovered on EUPEG. Compared to GeoVirus, GeoCorpora has two unique characteristics. First, the texts in GeoCorpora are short sentences (tweets within 140 characters) which provide only limited contextual information. Second, the content of tweets does not strictly follow grammatical rules and often contains abbreviations. Accordingly, GeoCorpora presents a more difficult dataset than GeoVirus. The evaluation results are summarized in Table 3.

Table 3: Evaluation results on GeoCorpora

Geoparser	precision	recall	f_score	mean	median	acc@161	AUC
DM_NLP+Pop	0.888	**0.669**	**0.763**	1249.865	0.000	0.661	0.288
UniMelb+Pop	0.852	0.661	0.745	1245.992	0.000	0.659	0.289
UArizona	0.892	0.598	0.716	1079.012	0.000	0.668	0.278
GeoTxt	**0.926**	0.521	0.667	714.94	0.000	0.876	0.116
StanfordNER	0.899	0.526	0.664	1063.473	0.000	0.676	0.270
CamCoder	0.904	0.503	0.647	1024.723	0.000	0.820	0.163
TopoCluster	0.882	0.506	0.643	575.225	32.948	0.698	0.361
DBpedia	0.865	0.500	0.633	669.105	33.816	0.654	0.352
Edinburgh	0.832	0.505	0.628	958.401	0.000	0.848	0.139
SpacyNER	0.705	0.467	0.562	982.137	0.000	0.752	0.224
CLAVIN	0.907	0.341	0.496	**373.563**	0.000	**0.913**	**0.084**

As can be seen, the F1 scores of all three geoparsers drop considerably on this more difficult dataset. However, it is worth emphasizing that DM_NLP increases the best possible F1 score from about 0.66 (by GeoTxt) to about 0.76, which is a large improvement. For toponym resolution, population heuristic is still a relatively effective approach on this dataset based on the zero median error distances achieved by the three new geoparsers. However, population heuristic is not as effective as other models, such as CLAVIN, GeoTxt, Edinburgh, and CamCoder (based on their higher acc@161 and lower AUC). Again, when we interpret the values of *mean*, *median*, *acc@161* and *AUC*, it is necessary to take into account the factor that toponym resolution is evaluated based on the different numbers of recognized toponyms from the previous step.

3.2.3 Results on Ju2016. Ju2016 is a corpus containing short sentences retrieved from various Web pages. This dataset was created by Ju et al. [8] who developed a script using Microsoft Bing Search API to automatically retrieve sentences containing highly ambiguous US place names (e.g., "Washington"). This corpus contains 5,441 entries in total and the average length of each entry is 21 words. This is a very difficult dataset, because the sentences are short (limited contextual information), place names are ambiguous, and upper and lower cases are not differentiated (all words are converted to lower case). Since this is an automatically created dataset, not all place mentions are annotated and as a result, precision, recall, and F1 score cannot be used as performance metrics. Following previous research [4], we use *accuracy* which measures the percentage of place names that are correctly recognized among all annotated place names. The results on Ju2016 are provided in Table 4.

As can be seen, many geoparsers show dramatically decreasing performances on this very difficult dataset. Two geoparsers, CLAVIN and Edinburgh, completely fail on this dataset which does not have word capitalization. Many other geoparsers, including DM_NLP and UArizona, also largely fail on this dataset due to their use of case-sensitive features, such as separate character-level embeddings for upper and lower case characters. UniMelb is an exception among the three geoparsers that performs still relatively well. Its performance can be attributed to its model design that does not include case sensitive character-level embeddings as DM_NLP and UArizona do. The highest accuracy is achieved by GeoTxt and DBpedia Spotlight, but all geoparsers show very low performances for toponym resolution based on the low acc@161 and high AUC scores. *Ju2016* is an artificially created dataset whose difficulty was deliberately increased for the purpose of testing geoparsers. It is less likely for a real world corpus to contain so many different place instances all sharing the same name (e.g., the many "Washington"s in this dataset). However, many real world corpora are likely to have irregular case alternations, and a robust geoparser should be able to accommodate such variations.

Table 4: Evaluation results on Ju2016

Geoparser	accuracy	mean	median	acc@161	AUC
GeoTxt	**0.463**	2609.734	1616.741	0.032	0.731
DBpedia	0.447	3101.087	1417.795	**0.111**	**0.698**
UniMelb+Pop	0.379	3301.993	2081.599	0.020	0.758
TopoCluster	0.158	4026.270	1547.266	0.036	0.752
DM_NLP+Pop	0.097	3357.802	2266.718	0.020	0.760
UArizona	0.036	2433.890	1966.937	0.029	0.739
StanfordNER	0.01	2027.016	2459.841	0	0.745
CamCoder	0.004	**1559.437**	**1389.716**	0.042	0.709
SpacyNER	0.004	3330.696	3478.187	0	0.818
CLAVIN	0	0	0	0	0
Edinburgh	0	0	0	0	0

3.3 Discussion

So are we there yet? Have we achieved sufficient progress on geoparsing to possibly consider the problem as solved? In our view, the answer is "it depends". It depends on the characteristics of the textual corpus on which geoparsing is performed. If the dataset contains well-formatted articles and is mostly about prominent places throughout the world (e.g., international news articles), then the answer is probably "yes" since the state-of-the-art geoparser, DM_NLP can achieve over 0.91 in precision, recall, and F1 score, and a relatively low toponym resolution error using a simple population heuristic. In fact, for such a dataset, one can even use the off-the-shelf StanfordNER combined with a population heuristic, saving the time for training a complex deep neural network model. On the other hand, if the dataset contains mostly short and informally-written sentences with ambiguous place names, then the answer is "no" since many of our current geoparsers will largely fail on such a dataset. In addition to handling toponym ambiguity, typos, name variations, case alterations, and limited contexts in short texts, future geoparsing research could also explore a number of directions, which are discussed as follows.

Geoparsing without population information. As shown in our experiment results, an off-the-shelf NER tool combined with a simple population heuristic can already provide competent performance for geoparsing. However, there are situations in which population information is not available in the gazetteer, or the toponyms to be parsed do not have population (e.g., toponyms about streets or

mountains). Methods that do not rely on population information need to be employed in these situations. For example, Moncla et al. [12] leveraged clustering techniques to disambiguate toponyms contained in a hiking description corpus.

Geoparsing fine-grained locations. A majority of geoparsing research so far has focused on recognizing and resolving toponyms at a geographic level higher than cities, towns, and villages. Sometimes, we may want to geoparse fine-grained locations within a city, such as street names, or the names of parks and monuments. A geoparser based on a large and general gazetteer will not be able to geo-locate such fine-grained locations. In a recent work, Alex et al. adapted the Edinburgh Geoparser to process literary text containing fine-grained place names located in and around the City of Edinburgh, and also released a non-copyrighted gold standard datasets to support research in this direction [1].

Geoparsing with gazetteers beyond GeoNames. Gazetteer plays a critical role in linking recognized toponyms and their geographic locations. However, most existing geoparsers only use GeoNames as their gazetteer. This, to some extent, can be attributed to the fact that many corpora are annotated based on GeoNames, and as a result, geoparsers are also developed based on GeoNames for evaluation convenience. As discussed in the previous point, a geoparser based on GeoNames will not be able to parse fine-grained place names. Besides, such a geoparser cannot process the historical texts in the context of digital humanities applications. An ideal geoparser, therefore, should allow users to switch the underlying gazetteer to one beyond GeoNames.

4 CONCLUSION AND FUTURE WORK

Geoparsing is an important research problem. This paper presents our work on evaluating the three state-of-the-art geoparsers coming out from the SemEval-2019 Task 12 competition in June 2019. This work is motivated by the outstanding performances of these geoparsers in the competition. As a result, we set out to examine whether we have made enough progress to possibly consider the problem of geoparsing as solved. We systematically tested the top three geoparsers on our benchmarking platform EUPEG. The results suggest that these new geoparsers indeed improve the highest possible scores on multiple datasets, and the problem of geoparsing well-formatted texts referring to prominent place instances could be considered as solved. Meanwhile, some challenges remain, such as geoparsing toponyms from informally-written texts with ambiguous place names.

This work can be extended in several directions. As discussed previously, we used a simple population heuristic for the toponym resolution component of the three geoparsers. Therefore, a next step is to develop a general toponym resolution dataset and use it to train the machine learning models described in the papers of DM_NLP and UniMelb. Second, EUPEG currently does not contain historical corpora. As a result, it cannot be used for testing the performances of geoparsers on historical texts for humanities applications. An extension of EUPEG with historical corpora (e.g., 19th century newspapers and fictional works) can make this platform even more useful for researchers in digital humanities. A similar idea can be applied to extending EUPEG with non-English corpora. Third, EUPEG currently evaluates only end-to-end geoparsers, and it could be useful to extend EUPEG with the capability of evaluating software tools designed for toponym recognition or resolution only. We have shared the source code of EUPEG, along with the datasets under open licenses, on GitHub at: https://github.com/geoai-lab/EUPEG. The source code of the three implemented neural network geoparsers tested in this work is also shared on GitHub at: https://github.com/geoai-lab/GeoAI2019Geoparser. We hope that these resources can help support the future work of the community to further advance geoparsing.

ACKNOWLEDGMENTS

The authors would like to thank the four anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander. 2019. Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation* 0, 0 (2019), 1–25.

[2] Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, USA, 2382–2388.

[3] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which Melbourne? Augmenting Geocoding with Maps. In *Proceedings of the 56th Annual Meeting of the ACL*, Vol. 1. ACL, Stroudsburg, PA, USA, 1285–1296.

[4] Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation* 52, 2 (2018), 603–623.

[5] Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, 1925 (2010), 3875–3889.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] Yingjie Hu. 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, Article 3, 2 pages.

[8] Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. 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, Cham, 353–367.

[9] Morteza Karimzadeh, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. 2019. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS* 23, 1 (2019), 118–136.

[10] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*. ACL, Stroudsburg, PA, USA, 260–270.

[11] Haonan Li, Minghan Wang, Timothy Baldwin, Martin Tomko, and Maria Vasardani. 2019. UniMelb at SemEval-2019 Task 12: Multi-model combination for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. ACL, Stroudsburg, PA, USA, 1313–1318.

[12] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*. ACM, Dallas, Texas, 183–192.

[13] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the NAACL-HLT*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237.

[14] Ross S Purves, Paul Clough, Christopher B Jones, Mark H Hall, Vanessa Murdock, et al. 2018. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval* 12, 2-3 (2018), 164–318.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon et al. (Ed.). NIPS Foundation, Inc., San Diego, USA, 5998–6008.

[16] Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. 2018. GeoCorpora: building a corpus to test and train microblog

geoparsers. *International Journal of Geographical Information Science* 32, 1 (2018), 1–29.

[17] Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: an extensibleand unified platform for evaluating geoparsers. *Transactions in GIS* 0, 0 (2019), accepted.

[18] Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM_NLP at SemEval-2018 Task 12: A Pipeline System for Toponym Resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. ACL, Stroudsburg, PA, USA, 917–923.

[19] Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez. 2019. Semeval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. ACL, Stroudsburg, PA, USA, 907–916.

[20] Davy Weissenbacher, Tasnia Tahsin, Rachel Beard, Mari Figaro, Robert Rivera, Matthew Scotch, and Graciela Gonzalez. 2015. Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics* 31, 12 (2015), i348–i356.

[21] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2145–2158.

[22] Vikas Yadav, Egoitz Laparra, Ti-Tai Wang, Mihai Surdeanu, and Steven Bethard. 2019. University of arizona at semeval-2019 task 12: Deep-affix named entity recognition of geolocation entities. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. ACL, Stroudsburg, PA, USA, 1319–1323.

APPENDIX A: EXPERIMENT RESULTS ON THE OTHER FIVE DATASETS

Table 5: Evaluation results on LGL

Geoparser	precision	recall	f_score	mean (km)	median (km)	acc@161	AUC
DBpedia Spotlight	**0.813**	**0.635**	**0.713**	1465.669	7.953	0.643	0.361
DM_NLP+Pop	0.730	0.630	0.677	1517.406	12.852	0.582	0.373
StanfordNER	0.744	0.622	0.677	1485.954	11.56	0.59	0.367
UniMelb+Pop	0.694	0.653	0.673	1527.597	13.340	0.581	0.375
TopoCluster	0.763	0.577	0.657	1209.39	18.959	0.625	0.379
CamCoder	0.811	0.548	0.654	837.126	0.02	0.717	0.248
UArizona	0.717	0.533	0.611	1570.982	13.477	0.582	0.372
GeoTxt	0.747	0.503	0.601	1544.283	0.044	0.633	0.312
CLAVIN	0.808	0.444	0.573	1261.408	0.012	0.701	0.262
Edinburgh Geoparser	0.723	0.383	0.501	**611.06**	**0.005**	**0.819**	**0.172**
SpacyNER	0.493	0.371	0.423	1702.214	7.685	0.561	0.381

Table 6: Evaluation results on TR-News

Geoparser	precision	recall	f_score	mean (km)	median (km)	acc@161	AUC
TopoCluster	0.883	**0.714**	**0.79**	1140.551	29.336	0.623	0.387
CamCoder	0.897	0.638	0.746	863.856	0	0.824	0.161
StanfordNER	0.89	0.731	0.803	1170.53	0	0.711	0.261
DBpedia Spotlight	0.861	0.631	0.728	1702.697	115.055	0.53	0.434
UniMelb+Pop	0.842	0.621	0.715	1151.496	0.000	0.729	0.244
UArizona	0.871	0.580	0.696	1217.766	0.000	0.697	0.265
GeoTxt	0.824	0.596	0.692	1017.3	0	0.801	0.168
DM_NLP+Pop	0.749	0.618	0.677	1313.811	0.000	0.688	0.280
CLAVIN	**0.908**	0.505	0.649	955.49	0	0.829	0.149
Edinburgh Geoparser	0.709	0.538	0.612	**770.227**	0	**0.85**	**0.127**
SpacyNER	0.659	0.402	0.5	1249.594	0	0.739	0.239

Table 7: Evaluation results on GeoWebNews

Geoparser	precision	recall	f_score	mean (km)	median (km)	acc@161	AUC
StanfordNER	0.885	**0.635**	**0.739**	818.282	0	0.698	0.257
UniMelb+Pop	0.851	0.628	0.722	877.210	0.003	0.691	0.268
DM_NLP+Pop	0.865	0.612	0.717	866.754	0.003	0.697	0.265
CamCoder	0.895	0.562	0.691	723.122	0	0.839	0.15
TopoCluster	0.838	0.559	0.67	597.082	42.46	0.68	0.357
Edinburgh Geoparser	0.819	0.538	0.65	346.873	0	0.921	0.071
DBpedia Spotlight	0.847	0.51	0.637	736.677	94.298	0.564	0.396
GeoTxt	0.771	0.479	0.591	421.073	0	0.903	0.086
CLAVIN	**0.909**	0.394	0.549	**210.905**	0	**0.937**	**0.06**
SpacyNER	0.784	0.415	0.543	1053.063	55.555	0.661	0.396
UArizona	0.860	0.357	0.504	928.186	1.046	0.648	0.290

Table 8: Evaluation results on Hu2014

Geoparser	accuracy	mean (km)	median (km)	acc@161	AUC
GeoTxt	**0.85**	928.839	1074.851	0.044	0.653
UArizona	0.813	2353.558	2575.671	0.000	0.763
TopoCluster	0.794	926.444	1074.851	0.008	0.674
StanfordNER	0.787	2277.44	2575.671	0	0.759
DM_NLP+Pop	0.700	2285.899	2575.671	0.000	0.759
DBpedia Spotlight	0.688	8846.334	10154.526	0.018	0.883
UniMelb+Pop	0.681	11007.875	11040.163	0.000	0.939
SpacyNER	0.681	2322.062	2575.671	0	0.762
Edinburgh Geoparser	0.656	**854.222**	1030.441	0.114	0.607
CLAVIN	0.65	951.498	1074.851	0.048	0.653
CamCoder	0.637	1250.964	**655.799**	**0.294**	**0.536**

Table 9: Evaluation results on WikToR

Geoparser	accuracy	mean (km)	median (km)	acc@161	AUC
UniMelb+Pop	**0.681**	4775.239	2804.149	0.171	0.712
DM_NLP+Pop	0.673	4842.807	2882.810	0.167	0.715
DBpedia Spotlight	0.604	2272.737	5.226	0.545	0.391
StanfordNER	0.54	4602.864	2513.181	0.184	0.702
SpacyNER	0.518	4785.691	2917.174	0.157	0.72
GeoTxt	0.506	4706.664	2644.041	0.179	0.701
TopoCluster	0.47	3800.378	1531.454	0.26	0.628
CamCoder	0.424	**1150.051**	33.967	0.588	**0.37**
UArizona	0.383	4940.599	3172.620	0.148	0.730
Edinburgh Geoparser	0.298	2165.389	**3.49**	**0.591**	0.378
CLAVIN	0.215	4220.027	2331.119	0.154	0.702