Training a Geographic Entity Recognizer on Biomedical Abstracts with the Aid of Embeddings, Metadata, and Linked Data

Xiaoliang Jiang
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL, USA
xjiang36@illinois.edu

Nigel Bosch
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL, USA
pnb@illinois.edu

Vetle I. Torvik
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL, USA
vtorvik@illinois.edu

ABSTRACT

Public access to scientific literature has fueled research in text mining and natural language processing, yet the problem of geographic named entity recognition persists. This paper describes a recognizer that uses candidates from multiple existing Named Entity Recognition (NER) tools to ensure high recall and uses a filtering model trained on sentence embeddings, metadata, and citation data to improve precision. Experimental results on a manually curated set of biomedical abstracts show that this filtering model preserves high recall while achieving much higher precision than all of the individual NER tools. This should enable more effective geography-based analysis of scientific literature, for example, to study the role of place in biomedical discovery.

CCS CONCEPTS

• Information systems ~ Information systems applications ~ Digital libraries and archives • Computing methodologies ~ Artificial intelligence ~ Natural language processing ~ Information extraction

KEYWORDS

Named Entity Recognition, Geographic Entity Recognition, Natural Language Processing, Biomedical Text Mining, Scholarly Document Processing, Sentence Embeddings, Information Extraction, Geoparsing, Linked Data, Metadata

ACM Reference format:

Xiaoliang Jiang, Nigel Bosch, and Vetle Torvik. 2024. Training a Geographic Entity Recognizer on Biomedical Abstracts with the Aid of Embeddings, Metadata, and Linked Data. In *Proceedings of JCDL 2024: ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '24), December 16-20. Hong Kong, China. ACM, New York, NY, USA.*

JCDL'24, December 16-20, 2024, Hong Kong, China

1 Introduction

The intersection of geography and biomedical research has been crucial in understanding disease epidemiology, public health planning, and the global dissemination of medical knowledge [18]. Digital databases have greatly improved access to biomedical literature, enabling data-driven research [22]. PubMed is one such key resource in this field, particularly because of its use of Geographical MeSH (Medical Subject Headings) terms, which assist in retrieving location-specific information [5, 29]. These terms have supported various research areas including geographic information retrieval [27], disaster management [31], and disease surveillance [4], fostering spatial analyses in biomedical research [36]. Some studies utilize geographical names in texts or metadata to focus on regions like Nigeria [2], Morocco [7], Ivory Coast [9], Australia and India [6]. Other research has developed a geographic filter to identify studies involving the Spanish population in PubMed [38]. However, since geographical MeSH terms cover only broad regions, countries, and a few large cities and are assigned at the article level, extracting specific geographic entities from biomedical text remains challenging [32]. Traditional methods often struggle with the specialized biomedical vocabulary, resulting in limited research on geographic terminology extraction in this domain [39]. Additionally, biomedical terminology, author-defined terms, abbreviations, and named entities are obstacles to geographic information extraction [15]. Thus, with the growing volume of biomedical literature, scalable automated methods for geographic entity recognition are increasingly needed [30].

Some scholars have explored the recognition of geographical names through named entity recognition (NER) techniques, using tools like Stanza [28], spaCy [26], FLAIR [3], NLTK [23], and DeBERTa [20] on various corpora including historical texts [40], social media [25], online sources [8], and scientific articles [1]. These studies confirm the importance of geographical names in biomedical research but also reveal limitations in previous work. Much of the research on local areas has depended on metadata, suggesting that including detailed geographical names at the text level could enhance these methods. Additionally, NER research often focuses on biomedical terms or entities like organizations rather than geographical names, and when geographical names are the focus, they are not typically studied within biomedical databases, overlooking challenges related to biomedical terminology.

Building on PubMed's Geographical MeSH terms, our work aims to expand the utility of geographic information within PubMed. This paper focuses on geographic named entities, requiring both "named" and "geographic" attributes. It includes geopolitical entities (GPE) such as cities, states/provinces, and countries, as well as some named locations (LOC) like mountains, rivers, and islands, and certain named organizations (ORG) if a geopolitical entity is implied. However, adjectives containing geographical names, like "Monte Carlo simulation" or "Norway rat," are excluded, similar to the annotation approach by Li et al. [20].

2 Method

We use two different labeling standards: the silver standard, which is automatically generated and imperfect but abundant, for model training; and the gold standard, which is manually verified, for model evaluation. The gold standard labels are derived from 1,000 randomly selected abstracts from PubMed from the years 2014 to 2018, a period chosen to ensure comprehensive metadata such as MeSH, citations and affiliations, in which named geographic entities were manually identified. Out of these, 140 abstracts contain a total of 238 sentences with 358 geographic named entities. It was annotated by the first author and verified by the third author. There was a 3.35% disagreement between the two (12/358), and the final version was based on the union of their annotations.

Table 1 shows the performance of several existing NER tools, including Stanza (version 1.6.1), spaCy (version 3.7.2), FLAIR (version 0.13.1), NLTK (version 3.8.1), the union of the first three (Union3) and the union of the four (Union4). The results will determine which models we select for generating the silver standard data.

Table 1: Test Results of Existing NER Tools on 1,000 Manually Checked Abstracts of Gold Standard Data

NER Tools	Detected Candidates #	Prec. (%)	Rec. (%)	Speed (Sent. /s)
Stanza	368	**76.6**	78.8	8.1
SpaCy	796	33.0	73.5	220.5
FLAIR	437	76.0	92.7	2.4
NLTK	2,271	10.1	64.2	**231.3**
Union3*	996	34.7	96.6	1.7
Union4*	2,883	12.4	**99.7**	1.6

*Union3 = Stanza + SpaCy + FLAIR; Union4 = Stanza + SpaCy + FLAIR + NLTK

Among the four tools tested, the accuracy ranked from highest to lowest is FLAIR, Stanza, spaCy, and NLTK, while the speed ranking is exactly the opposite. Although NLTK is very fast, its overall accuracy is the worst, generating many irrelevant results that could severely impact downstream tasks. On the other hand, FLAIR achieves the highest accuracy but is 100 times slower than NLTK. To balance speed and accuracy, spaCy and Stanza are used for the silver standard. Compared to NLTK, spaCy demonstrates 3 times more precision performance with a similar processing speed. Stanza, while providing an intermediate performance level close to FLAIR, processes text four times faster.

We also tried to test some language model based approaches but were unable to reuse the code from Acheson and Purves [1] due to version updates in Python dependencies. However, they use Stanford CoreNLP, and Stanza is also developed by the same Stanford research group. Furthermore, the precision/recall reported by Acheson and Purves [1] is similar to that of Stanza observed in our experiments. We were also keen to reuse the code and data from Li et al. [2] because our precision/recall results differed dramatically, particularly for spaCy. However, the GitHub repository referenced in the paper is currently unavailable. Both Acheson and Purves [1] and Li et al. [2] perform manual annotations, but we were unable to find the corresponding raw data.

Among PubMed abstracts from 2014 to 2018, spaCy identified 10 million sentences containing candidate place names, which were then assessed by Stanza, resulting in a silver-standard dataset of over 4 million candidate sentences. Each candidate place name was wrapped with [locB] and [locE] tags, with label 1 indicating entities recognized by both tools and label 0 indicating those identified only by spaCy. The silver standard process applied to the 1000 abstract gold standard yielding a recall of 70.7% (253/358), and precision of 82.4% (253/307) for label 1 and 98.0% (479/489) for label 0. Overall, the silver standard not only encompasses the majority of true geographic entities with relatively high precision but also captures diverse contexts, including a large number of challenging biomedical terms marked with label 0, addressing unique challenges in geographical NER for biomedical literature.

After generating the silver standard data, we engineered a diverse set of predictive features that are categorized into four types: text-based, metadata-based, linked data-based, and pre-trained sentence embeddings. To better capture potential nonlinear features, we also applied transformations such as the square, square root, and logarithmic functions.

The first major category of features is based on the candidate and its surrounding texts, including the number of words, number of letters, word length, position in the sentence, and the count of uppercase letters, the proportion of uppercase letters in the candidate, whether the candidate contains numbers or symbols, and whether special texts like year or "et al" occurred in candidate sentences.

The second major category involves three sub-types of article metadata information: matching candidate information with affiliation and MeSH metadata, calculating the frequency of the candidate's appearance in different metadata fields (author name, affiliation, title and abstract) across the entire PubMed database, and co-occurrence rate of MeSH with geographical MeSH. MeSH, a controlled vocabulary thesaurus by the National Library of Medicine, standardizes terminology in PubMed, offering a hierarchical structure of terms across 16 trees to enhance search accuracy as shown in https://meshb.nlm.nih.gov/treeView.

Along with affiliation information, MeSH can also be used to assess whether the candidate aligns with the metadata as the first sub-type of metadata features. However, these features are challenging to compute due to place name complexities, including multiple variants, abbreviations, and hierarchical relationships (city, state, country). Consistency between candidates and metadata must account for these relationships; for example, a city like Chicago aligns with metadata listing its state, Illinois, or country, the USA. Since MeSH terms cover limited cities and states, extra information is required for broader matching. To standardize the different variants of place names, we utilized MapAffil, a bibliographic tool that maps PubMed author affiliation strings to geographic locations [34]. The MapAffil dictionary [33] was used to standardize the detected candidates, ensuring consistent naming for entity-to-entity hierarchical matching with MeSH terms and affiliations.

The second sub-type of metadata features calculates the frequency of the candidate's appearance in various metadata fields, such as affiliations, titles, abstracts, and names, to distinguish geographical entities from others. This frequency analysis aims to help the model differentiate place names from personal names and technical terms in biomedical literature.

The third sub-type of metadata-based features includes the co-occurrence rates of non-geographic MeSH terms with geographic MeSH terms across all of PubMed articles. Certain topics, like infectious diseases or health policies, often mention geographic locations, while others, such as DNA or RNA studies, rarely do. Table 2 shows that the co-occurrence with geographic MeSH terms varies dramatically for broad categories of MeSH. For example, Anthropology (I) and Humanities (K) exceed 50%, while Chemicals (D) and Anatomy (A) are below 10%. It is important to note that if an article contains multiple MeSH terms in different categories, the count will be incremented once for each category during the statistics process. Therefore, the sum of article counts across categories will be greater than the overall article count. Table 3 shows the variability for select terms within the same broad Diseases category C; "C24 Occupational diseases" has a high 38.10% rate, while "C04.619 Neoplasms, experimental" has a low 0.80% rate, and the overall rate is 13.83%. The co-occurrence rates for about 30,000 MeSH terms in PubMed 2018 were precalculated and each article was assigned the three most and least frequently co-occurring terms as features for predicting geographic references.

The third major category uses metadata from cited references, similar to the second main category metadata features, but by checking the candidate's presence in cited MeSH terms and affiliations and analyzing co-occurrence rates of the three highest and lowest geo-entity-related terms. This method, supplemented by MapAffil for standardization, generates binary and numeric features to enhance model training, especially for articles with missing metadata.

The fourth major category of features is sentence Embedding Features. Specifically, we explored two pre-trained embedding models, including BioBERT [19] and BioSimCSE [17], which were trained by biomedical literature. Initial testing reveals that the BioSimCSE performed better and there was adopted as a sentence embedder to convert text into vector-type features for inclusion in model training.

Table 2: Co-occurrence of Different Broad Categories with Geographic MeSH Terms

Cat.	Name	Total Article #	Co-occ. Rate (%)
I	Anthropology, education, sociology, and social phenomena	2.95M	51.32
K	Humanities	0.85M	51.13
N	Health care	10.39M	31.90
F	Psychiatry and psychology	4.25M	27.71
H	Disciplines and occupations	3.65M	27.62
J	Technology, industry, and agriculture	2.39M	23.03
M	Named groups	9.18M	22.30
L	Information science	2.85M	20.75
B	Organisms	23.15M	14.55
E	Analytical, diagnostic and therapeutic techniques, and equipment	17.40M	14.38
C	Diseases	13.95M	13.69
G	Phenomena and processes	13.89M	10.29
D	Chemicals and drugs	13.69M	7.75
A	Anatomy	10.35M	4.02
	Overall	**29.15M**	**13.68**

Table 3: Differences in Co-occurrence within the Same Broad Category with Geographic MeSH Terms

Cat.	Name	Total Article #	Co-occ. Rate (%)
C24	Occupational diseases	126K	38.10
*C03**	Parasitic diseases	368K	34.57
*C02**	Virus diseases	867K	31.73
...	...	...	...
C04.697...	Cell transformation...	73K	1.33
C22.232	Disease models, animal	302K	1.08
C04.619	Neoplasms, experimental	144K	0.80
C	**Diseases**	**13.95M**	**13.69**

*Annotated items are valid vocabularies in PubMed 2018 but were moved to other branches from the latest version:

C02 Virus Diseases were moved to C01.925 Virus Diseases

C03 Parasitic Diseases were moved to C01.610 Parasitic Diseases

Three models were trained on the 500,000-record silver standard dataset: logistic regression (LR), XGBoost (XGB), and a deep neural network (DNN) on an 80/20 split using different feature groups. To enhance the model's ability to capture nonlinear relationships, squaring, and logarithmic transformations were applied to certain non-embedding features. Pairwise multiplicative interactions among selected non-embedding features were created, multicollinearity was addressed using correlation and VIF, and a model was trained on a reduced dataset to select the top 100 features based on importance determined by both tree-based and regression models as well. The DNN in this study is a seven-layer fully connected feedforward network designed for binary classification, processing input features as rank 1 tensors with a length between 768 and 960. Each layer uses the ReLU activation function, with a final single neuron using a sigmoid for outputting the positive class probability. The network is optimized with the Adam optimizer and uses binary cross-entropy as the loss optimizing model performance through various data sizes and validating against both silver and manually verified gold standards. function.

To support reproducibility, the code and data is available from GitHub: https://github.com/XiaoliangJiang/BiomedicalGeoNER

3 Results and Evaluation

Table 4 shows the performance of different configurations of features and classifiers. We added 358 randomly sampled negative examples to the gold standard's 358 positive examples, creating a classifier with 100% recall and 50% precision to simulate a result generated by NER tools. Embeddings alone achieved a 93.2% F1 score while combining them with other features raised the score to 96.1% with logistic regression. The DNN underperformed compared to logistic regression and XGBoost, likely due to overfitting on the noisy instances from the silver standard, whereas logistic regression generalized better.

Table 4: Evaluation Results with Different Configurations on the Gold Standard Data.

Model Settings	Models	Prec. (%)	Rec. (%)	F1 (%)
No Embedding Features	LR	85.7	**95.9**	90.5
	XGB	**88.2**	95.5	**91.7**
Only Embedding Features	LR	90.3	96.3	**93.2**
	XGB	90.3	**96.4**	**93.2**
	DNN	**91.8**	93.8	92.8
With Embedding Features	LR	**95.2**	**97.1**	**96.1**
	XGB	94.1	96.3	95.2
	DNN	92.4	95.8	94.1

To compare changes in geographic named entity recognition before and after applying our model, we tested it on various NER tools, as summarized in Table 5. Since recall depends on the pre-existing NER tools, our model is mainly responsible for improving precision while minimizing recall loss, with spaCy showing the highest precision increase at 58.5%. The best-performing tools with our model were FLAIR and Union3 (Stanza, spaCy, and FLAIR), achieving high precision (94.1%) and recall (93.0%) respectively.

To evaluate feature importance, we used SHAP values [24], revealing that matching with the place name dictionary (MapAffil_dict_matched) was the single most influential positive indicator. However, other features, such as capitalization (cand_capitalized) and sentence position (cand_pos), and some specific embedding features also played crucial roles. The model assigns negative weights to words with a high proportion of capital letters to address challenges in place name recognition, such as distinguishing "US" as the "United States" from "ultrasonography."

Table 5: Performance Comparison of NER Tools Before and After Using Trained Model

NER Tools	NER Recall	+Model Recall	NER Prec.	+Model Prec.	Prec. Improved
Stanza	78.8	77.7	**76.6**	88.0	11.4
spaCy	70.7	70.1	31.8	90.3	**58.5**
FLAIR	92.7	89.3	76.0	**94.1**	18.1
NLTK	64.2	61.1	10.1	53.9	43.8
Union3*	**96.6**	**93.0**	34.7	85.4	50.7

*Union3 = Stanza + SpaCy + FLAIR

4 Discussion

The goal of the presented study was to assess how a broad set of text, metadata, and linked data can contribute to distinguishing from other named entities in biomedical literature, without regard to their relevance to the topic. The recall of the best classifier is limited by the upstream NER tools used to generate candidate, while the classifier dramatically improves precision. Although overall performance metrics exceed 95%, errors persist: false positives often involve terms with common place names that are not necessarily linked to actual locations, such as "Minnesota Job Satisfaction Scale" or "Monte Carlo simulation," while false negatives frequently stem from out-of-dictionary terms like "Mtkvari," a river that is not present in the MapAffil dictionary. Despite this, the model effectively balances dictionary-based evidence with other features, achieving an F1 score of over 93% using only embedding features. Additionally, as the model outputs probabilities along with the results, adjusting the probability thresholds can potentially enhance performance. By fine-tuning the default threshold of 0.5 to other values or introducing a neutral category for handling more challenging classifications near 0.5 while selecting more certain probabilities as the threshold, a balance between precision and recall can be adjusted, resulting in performance changes that cater to the specific requirements of practical applications. Future work may include retraining with different dictionaries or configurations to enhance handling of complex cases.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the US National Institutes of Health (Award Number P01AG039347). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1] E. Acheson and R. S. Purves, “Extracting and modeling geographic information from scientific articles,” PLOS ONE, vol. 16, no. 1, p. e0244918, Jan. 2021, doi: 10.1371/journal.pone.0244918.

[2] I. Adamson, “Access and retrieval of information as coordinates of scientific development and achievement in Nigeria,” Scientometrics, vol. 23, no. 1, pp. 191–199, Aug. 2005, doi: 10.1007/bf02020922.

[3] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), W. Ammar, A. Louis, and N. Mostafazadeh, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 54–59. doi: 10.18653/v1/N19-4010.

[4] T. Allen et al., “Global hotspots and correlates of emerging zoonotic diseases,” Nat Commun, vol. 8, no. 1, Art. no. 1, Oct. 2017, doi: 10.1038/s41467-017-00923-8.

[5] M. E. Anders and D. P. Evans, “Comparison of PubMed and Google Scholar Literature Searches,” Respiratory Care, vol. 55, no. 5, pp. 578–583, May 2010.

[6] R. Arulanandam, B. T. R. Savarimuthu, and M. Purvis, Extracting crime information from online newspaper articles. 2014.

[7] H. Badrane and M. Alaoui-el-Azher, “Biomedical research in developing countries: the case of Morocco in the 1990s: La Tunisie medicale,” Tunis Med, vol. 81, no. 6, pp. 377–382, Jun. 2003.

[8] C. Berragan, A. Singleton, A. Calafiore, and J. Morley, “Transformer based named entity recognition for place name extraction from unstructured text,” International Journal of Geographical Information Science, vol. 37, no. 4, pp. 747–766, Apr. 2023, doi: 10.1080/13658816.2022.2133125.

[9] Y. Chatelin and R. Arvanitis, “Representing scientific activity by structural indicators: The case of Cote d'Ivoire 1884–1968,” Scientometrics, vol. 23, no. 1, pp. 235–247, Aug. 2005, doi: 10.1007/bf02020925.

[10] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, “A study of active learning methods for named entity recognition in clinical text,” Journal of Biomedical Informatics, vol. 58, pp. 11–18, Dec. 2015, doi: 10.1016/j.jbi.2015.09.010.

[11] J. P. C. Chiu and E. Nichols, “Named Entity Recognition with Bidirectional LSTM-CNNs,” TACL, vol. 4, pp. 357–370, Dec. 2016, doi: 10.1162/tacl_a_00104.

[12] H. Cho and H. Lee, “Biomedical named entity recognition using deep neural networks with contextual information,” BMC Bioinformatics, vol. 20, no. 1, p. 735, Dec. 2019, doi: 10.1186/s12859-019-3321-4.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, arXiv: arXiv:1810.04805. Accessed: Aug. 02, 2022. http://arxiv.org/abs/1810.04805

[14] R. Islamaj et al., “NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature,” Sci Data, vol. 8, no. 1, p. 91, Mar. 2021, doi: 10.1038/s41597-021-00875-1.

[15] X. Jiang and V. I. Torvik, “On the Ambiguity and Relevance of Place Names in Scientific Text,” in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event China: ACM, Aug. 2020, pp. 401–404. doi: 10.1145/3383583.3398618.

[16] S. Jonnalagadda and P. Topham, “NEMO: Extraction and normalization of organization names from PubMed affiliation strings,” J Biomed Discov Collab, vol. 5, pp. 50–75, Oct. 2010.

[17] K. raj Kanakarajan, B. Kundumani, A. Abraham, and M. Sankarasubbu, “BioSimCSE: BioMedical Sentence Embeddings using Contrastive learning,” in Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 81–86. Accessed: Jun. 20, 2023. https://aclanthology.org/2022.louhi-1.10

[18] R. S. Kirby, E. Delmelle, and J. M. Eberth, “Advances in spatial epidemiology and geographic information systems,” Annals of Epidemiology, vol. 27, no. 1, pp. 1–9, Jan. 2017, doi: 10.1016/j.annepidem.2016.12.001.

[19] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, p. btz682, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[20] W. Li, K. Sun, S. Wang, Y. Zhu, X. Dai, and L. Hu, “DePNR: A DeBERTa-based deep learning model with complete position embedding for place name recognition from geographical literature,” Transactions in GIS, doi: 10.1111/tgis.13170.

[21] X. Li, T. Wang, Y. Pang, J. Han, and J. Shi, “Review of Research on Named Entity Recognition,” in Advances in Artificial Intelligence and Security, X. Sun, X. Zhang, Z. Xia, and E. Bertino, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2022, pp. 256–267. doi: 10.1007/978-3-031-06761-7_21.

[22] D. a. B. Lindberg and B. L. Humphreys, “Rising Expectations: Access to Biomedical Information,” Yearb Med Inform, vol. 17, no. 1, pp. 165–172, 2008, doi: 10.1055/s-0038-1638596.

[23] E. Loper and S. Bird, “NLTK: The Natural Language Toolkit,” May 17, 2002, arXiv: arXiv:cs/0205028. Accessed: Aug. 02, 2022. [Online]. Available: http://arxiv.org/abs/cs/0205028

[24] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: May 07, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[25] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, “Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging,” ACM Trans. Inf. Syst., vol. 36, no. 4, p. 40:1-40:27, 2018, doi: 10.1145/3202662.

[26] C. Pearson, N. Seliya, and R. Dave, “Named Entity Recognition in Unstructured Medical Text Documents,” in 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2021, pp. 1–6. doi: 10.1109/ICECET52533.2021.9698694.

[27] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock, “Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text,” INR, vol. 12, no. 2–3, pp. 164–318, Feb. 2018, doi: 10.1561/1500000034.

[28] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” Apr. 23, 2020, arXiv: arXiv:2003.07082. doi: 10.48550/arXiv.2003.07082.

[29] D. Rosselli, “Geography of biomedical publications,” The Lancet, vol. 354, no. 9177, p. 517, Aug. 1999, doi: 10.1016/S0140-6736(05)75555-0.

[30] N. A. Sanati and M. Sanati, “Growing interest in use of geographic information systems in health and healthcare research: a review of PubMed from 2003 to 2011,” JRSM Short Reports, vol. 4, no. 6, p. 2042533313478810, Jun. 2013, doi: 10.1177/2042533313478810.

[31] G. Scalia, C. Francalanci, and B. Pernici, “CIME: Context-aware geolocation of emergency-related posts,” Geoinformatica, vol. 26, no. 1, pp. 125–157, Jan. 2022, doi: 10.1007/s10707-021-00446-x.

[32] J. Tamames and V. de Lorenzo, “EnvMine: A text-mining system for the automatic extraction of contextual information,” BMC Bioinformatics, vol. 11, no. 1, p. 294, Jun. 2010, doi: 10.1186/1471-2105-11-294.

[33] V. Torvik, “MapAffil 2018 dataset -- PubMed author affiliations mapped to cities and their geocodes worldwide with extracted disciplines, inferred GRIDs, and assigned ORCIDs,” 2021, doi: 10.13012/B2IDB-2556310_V1.

[34] V. I. Torvik, “MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide,” Dlib Mag, vol. 21, no. 11–12, p. 10.1045/november2015-torvik, 2015.

[35] V. I. Torvik, “MapAffil 2016 dataset -- PubMed author affiliations mapped to cities and their geocodes worldwide,” 2018, doi: 10.13012/B2IDB-4354331_V1.

[36] O. A. Uthman and M. B. Uthman, “Geography of Africa biomedical publications: An analysis of 1996–2005 PubMed papers,” African Journal of Food, Agriculture, Nutrition and Development, vol. 8, no. 2, Art. no. 2, 2008, doi: 10.4314/ajfand.v8i2.19192.

[37] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,” Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 552–556, Sep. 2011, doi: 10.1136/amiajnl-2011-000203.

[38] J. M. Valderas, J. Mendivil, A. Parada, M. Losada-Yáñez, and J. Alonso, “Development of a Geographic Filter for PubMed to Identify Studies Performed in Spain,” Revista Española de Cardiología (English Edition), vol. 59, no. 12, pp. 1244–1251, Jan. 2006, doi: 10.1016/S1885-5857(07)60080-2.

[39] D. Weissenbacher et al., “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration,” Bioinformatics, vol. 31, no. 12, pp. i348–i356, Jun. 2015, doi: 10.1093/bioinformatics/btv259.

[40] M. Won, P. Murrieta-Flores, and B. Martins, “Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora,” Front. Digit. Humanit., vol. 5, Mar. 2018, doi: 10.3389/fdigh.2018.00002.