Fine-Tuning Transformers for Toponym Resolution: A Contextual Embedding Approach to Candidate Ranking

Diego Gomes[1], Ross S. Purves[1] and Michele Volpi[2]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[2]Swiss Data Science Center, ETH Zurich and EPFL, Andreasstrasse 5, 8050 Zurich, Switzerland

Abstract

We introduce a new approach to toponym resolution, leveraging transformer-based Siamese networks to disambiguate geographical references in unstructured text. Our methodology consists of two steps: the generation of location candidates using the GeoNames gazetteer, and the ranking of these candidates based on their semantic similarity to the toponym in its document context. The core of the proposed method lies in the adaption of SentenceTransformer models, originally designed for sentence similarity tasks, to toponym resolution by fine-tuning them on geographically annotated English news article datasets (Local Global Lexicon, GeoWebNews, and TR-News). The models are used to generate contextual embeddings of both toponyms and textual representations of location candidates, which are then used to rank candidates using cosine similarity. The results suggest that the fine-tuned models outperform existing solutions in several key metrics.

Keywords

Geographic Information Retrieval, toponym resolution, transformer, gazetteer

1. Introduction

Toponym resolution, the task of assigning unique identifiers to geographical locations referred to by place names in texts, is an essential yet challenging aspect of geographic information retrieval [1]. The emergence of transformer-based models in natural language processing [2] has opened new avenues to address these challenges, providing sophisticated means to capture the nuanced relationships between textual context and geographical references. In this paper, we present a new approach that leverages the capabilities of transformer models, specifically using the SentenceTransformers framework [3], originally designed for sentence similarity tasks. Our methodology reimagines toponym resolution as a variant of sentence similarity, comparing document-based embeddings to those generated from gazetteers to disentangle the complexities of geographical references within unstructured text.

Our approach consist of two main steps: the generation of location candidates and the ranking of these candidates based on contextual embeddings generated by fine-tuned transformer-

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland

diego.gomes@uzh.ch (D. Gomes); ross.purves@geo.uzh.ch (R. S. Purves); michele.volpi@sdsc.ethz.ch (M. Volpi)
www.geo.uzh.ch/~rsp (R. S. Purves); www.datascience.ch/people/michele-volpi (M. Volpi)
0009-0003-8449-2603 (D. Gomes); 0000-0002-9878-9243 (R. S. Purves); 0000-0003-2771-0750 (M. Volpi)

based Siamese networks. By adapting the SentenceTransformers framework for toponym resolution, we capitalise on the powerful ability of these pre-trained models to compare texts in a semantically meaningful way. The results of our research demonstrate that this approach outperforms recent work in three of four metrics for toponym resolution, offering a scalable, efficient, and accurate solution for the field.

2. Related Work

Recently, transformer-based methods have increasingly influenced toponym resolution methodologies. We define existing approaches as either *localisation-based* or *ranking-based*.

Localisation-based approaches focus on the direct prediction of geographic coordinates or areas from textual input. Radford's [4] method employs DistilRoBERTa for end-to-end probabilistic geocoding, while Cardoso et al. [5] use LSTMs with BERT embeddings to predict probability distributions over spatial regions. Similarly, Solaz and Shalumov [6] use the T5 transformer model in a sequence-to-sequence framework to translate text into hierarchical encodings of geographic cells.

Ranking-based approaches focus on the effective ranking of location candidates. Halterman's [7] Mordecai 3 system integrates spaCy transformer embeddings into a neural model that ranks candidates based on similarity measures. Li et al.'s [8] GeoLM aligns linguistic context with geospatial information through contrastive learning, enhancing language models' understanding of geographic entities. In Zhang and Bethard's [9] GeoNorm framework, a BERT-based transformer model is employed to rerank location candidates, using contextual embeddings to prioritise candidates that best match a toponym's context.

A critical limitation in many transformer-based toponym resolution methods is the lack of task-specific model fine-tuning. Transformer models such as BERT [10], which are pre-trained on tasks such as masked language modelling and next sentence prediction, are trained to produce embeddings optimised for those tasks. Consequently, using these embeddings directly for toponym resolution may limit their effectiveness. Studies such as those by Cardoso et al. [5] and Halterman [7] use embeddings produced by off-the-shelf transformer models without task-specific fine-tuning, potentially limiting the efficacy of these embeddings for toponym resolution. This consideration is consistent with the findings of Reimers and Gurevych [3], who found that while base models such as BERT perform poorly on sentence similarity tasks, their performance improves significantly after task-specific fine-tuning. Thus, fine-tuning these models for toponym resolution could be a crucial aspect in their ability to generate contextually relevant embeddings.

Another issue with machine learning-based toponym resolution methods in general is geographic bias, stemming from the geographic imbalance of training datasets. Trained models tend to favour locations that are overrepresented in the training corpora, as highlighted by Liu et al. [11]. The limited availability and domain diversity of geotagged datasets further exacerbates this bias [12]. Geoparsing methods should aim to generalise the toponym resolution capability acquired from training on geographically biased data to a global context, thus ensuring broad applicability and reliability of the models across different geographical regions.

3. Proposed Method

This study introduces a new method for toponym resolution, centred around the use of transformer-based Siamese networks fine-tuned to discern geographically relevant contextual cues within texts. Our approach unfolds in two key phases: the generation of location candidates and the ranking of these candidates. The first phase involves compiling potential geographical matches for identified toponyms using a gazetteer, while the second phase focuses on ranking these candidates to select the most contextually appropriate location.

Candidate generation involves querying toponyms in a toponym index created from the GeoNames database. This index contains standard and alternate location names, supplemented with externally sourced demonyms, and serves as the primary resource for retrieving location candidates for toponyms. In instances where the index fails to return results using exact string matching, a fallback mechanism initiates API calls to GeoNames using both regular and fuzzy search parameters, ensuring that a list of location candidates is generated for every toponym. This fallback procedure, while simple, effectively broadens the scope of potential matches, albeit with a possible trade-off in precision.

The candidate ranking process involves generating contextual embeddings for toponyms and each of their location candidates using a transformer-based model. Toponym embeddings are created using the toponym's source document as input to the model. To create candidate embeddings, first unique textual representations are created, incorporating textual descriptors of geographical identifiers like country, administrative divisions, and feature types. These representations are then fed into the same model to obtain semantically enriched candidate embeddings. Since both toponym and candidate embeddings reside in the same vector space, cosine similarity scores can be used to rank lists of candidates and ultimately make a prediction about the most likely referent of the toponym.

Our approach aims to reduce geographic bias by emphasising contextual understanding over direct geographic associations. We hypothesise that by training models to detect and interpret geographic cues within texts, rather than learning geographic correlations, they are less likely to inherit biases from geographically skewed training datasets. By focusing on the extraction and comparison of geographically relevant contextual cues, we posit that the models develop a more generalised ability to resolve toponyms in English, less tethered to the geographic distributions present in the training data. We note, however, that this hypothesised reduction in geographic bias is an initial assumption based on the architectural design of the system. Empirical validation across diverse and globally representative datasets will be crucial to substantiate this claim and fully assess the effectiveness of our approach in mitigating geographic bias in toponym resolution. Furthermore, we do not make claims about the direct transferability of our approach to languages other than English.

At the heart of our methodology lies the adaptation of the SentenceTransformers framework [3], originally designed for sentence similarity tasks. We reimagine toponym resolution as a variant of sentence similarity, where the contextual relationship between a toponym and its geographical referent is analogous to the semantic relationship between two sentences. The SentenceTransformer models, known for their efficacy in generating semantically comparable sentence embeddings, are repurposed to generate embeddings for both toponyms and their location candidates. Using these models to encode geographical references in a comparable

Table 1
Dataset properties

	LGL	GWN	TRN
Number of articles	587	199	118
Number of toponyms	4439	2401	1271
Number of unique GeoName IDs	1076	579	349

way seeks to harness their inherent strengths in understanding textual context and nuance.

To use SentenceTransformer models for toponym resolution, we fine-tune them using geographically annotated texts. This entails adapting the pre-trained models using training data that juxtaposes toponyms with both their correct and incorrect geographical matches. During this process, the models are trained to produce pairs of embeddings that are closely aligned for correct toponym-location pairs and distinctly separate for incorrect ones. This ability feeds directly into the ranking of location candidates based on their semantic similarity to the toponym as it appears in the text. By learning to generate embeddings that accurately reflect the relevant geographic information embedded in the context of the toponym, the models gain the ability to effectively discern and prioritise the most likely geographic location.

A key aspect of using the SentenceTransformers framework is the computational efficiency it brings to the methodology. Thanks to the Siamese network architecture, these models act as encoders that can process individual units of text independently. This architectural feature allows for the pre-computation of embeddings for all locations in a gazetteer. This means that during the toponym resolution process, the system only needs to generate embeddings for toponyms, significantly reducing the computational load.

4. Experiments

4.1. Datasets

In this pilot study, we used three existing annotated datasets of English news articles (Table 1). These are the Local Global Lexicon (LGL) [13], the GeoWebNews (GWN) [14], and the TR-News (TRN) [15]. The LGL dataset is heavily concentrated in the United States, with moderate coverage in Europe and the Middle East, and sparse coverage in other regions (Figure 1). The GWN dataset shows a similar pattern, but with a broader European coverage and notable coverage in Africa, the Middle East, and some Asian regions. The TRN dataset, while also focused heavily on the United States, presents a more balanced distribution across Europe, the Middle East, East Asia, and Australia.

The choice of these specific datasets aligns our work with that of Zhang and Bethard [9]. By using the same datasets and mirroring the data splits into training, evaluation, and testing segments (70%, 10%, and 20%, respectively), we aim to provide a direct comparison. For training and interim evaluations, data from the three datasets were pooled, while for final testing, they were kept separate, allowing separate performance assessments on each dataset.

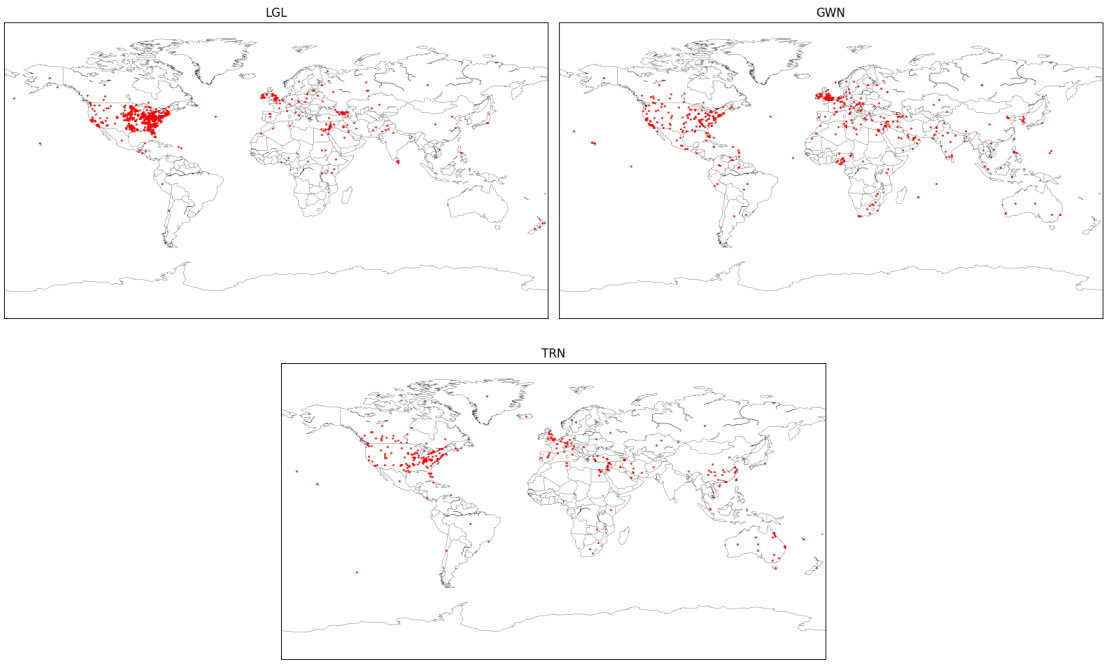

Figure 1: Geographical distribution of toponyms in the 3 datasets

4.2. Data Preparation

Text documents containing toponyms were truncated to comply with the input sequence limits of the SentenceTransformer models, taking care to preserve the integrity of sentences and to keep the toponyms centred within the truncated text. A small number of toponyms with outdated or invalid GeoName IDs (n = 27) were removed from the datasets.

Candidate generation involved compiling a list of location candidates for each toponym. The recall rates for this process were 97.5% for LGL, 90.2% for GWN, and 98.5% for TRN. In some cases, the correct location was not included in the compiled lists, thus setting a performance ceiling for subsequent toponym resolution.

Textual representations of location candidates were constructed using attributes retrieved from candidate's GeoNames entries, incorporating the name, feature type, and relevant administrative and geographical identifiers in a pseudo-sentence format, formulated as: "[name] ([feature type]) in [admin2], [admin1], [country]". This approach was designed to provide geographically distinct descriptors of location candidates that can be processed in the same way as text documents.

To create training examples, texts containing toponyms were paired with the textual representations of location candidates. For positive training examples, the correct locations were retrieved from the toponym labels provided in the datasets. For negative examples, the toponym's list of location candidates was used to generate incorrect pairings from all items in the lists, excluding the correct location. This was done without taking into account the proportion of positive to negative examples.

4.3. Fine-Tuning

The fine-tuning of the SentenceTransformer models involved training on the prepared examples using a contrastive loss function. This taught the models to generate pairs of embeddings that were similar for correct toponym-location pairs and dissimilar for incorrect pairs.

Models were evaluated at regular intervals during the training process using the separate evaluation set. They were assessed every 10% of the training steps, with the most accurate model on the evaluation set being selected for final testing. The entire training phase was completed in a single epoch, using a batch size of 8. No hyperparameter optimisation was performed.

4.4. Evaluation Metrics

In line with Zhang and Bethard [9] and guided by Gritta et al. [14], we adopt four metrics to evaluate our methodology. Accuracy (A) measures the exact match rate of predicted and labelled GeoName IDs, providing a binary assessment of correctness. Accuracy@161km (A161) provides a broader perspective, assessing the proportion of toponyms that are correctly resolved within a 161 km (100 mile) radius, thus allowing for minor geographical deviations. Mean error distance (MED) quantifies the average geographical deviation of predictions from true locations. Finally, the area under the curve (AUC) assesses the error distribution, particularly accounting for outliers, by integrating the area under a curve of scaled logarithmic error distances.

5. Results

In the evaluation of two SentenceTransformer models for toponym resolution, the base models, originally designed for sentence similarity tasks, showed better than random but generally poor performance, and were outperformed by a simple population-based method (Table 2). However, a substantial increase in performance was observed after the models were fine-tuned with every model outperforming the population baseline across all datasets and metrics.

Compared to the recently introduced GeoNorm model by Zhang and Bethard [9], the fine-tuned SentenceTransformer models showed comparable or superior performance across all evaluated corpora and all metrics except for mean error distance.

6. Discussion

In this study, we have demonstrated the efficacy of adapted SentenceTransformer models for toponym resolution. Our results underline the viability of this approach, with the models achieving state-of-the-art performance in the context of the datasets used.

Nevertheless, it is important to note the constraints and limitations of this study. The training data, exclusively sourced from news articles, was limited in both volume and diversity. This limitation was partly intentional, to align our methodology with that of Zhang and Bethard [9] for direct comparability, however, it also reflects a broader challenge in the field: the scarcity of geographically annotated text corpora spanning diverse domains [12]. Going forward, enriching

Table 2
Comparison of model performances on test sets using accuracy (A), accuracy at 161 km (A161), mean error distance (MED), and area under the curve (AUC)

Dataset	Model	A	A161	MED	AUC
LGL	Random	0.229	0.278	3579	0.588
	Population	0.650	0.732	1149	0.229
	GeoNorm	0.799	0.828	**52**	0.136
	all-distilroberta-v1	0.417	0.518	1922	0.398
	all-mpnet-base-v2	0.398	0.472	1660	0.411
	all-distilroberta-v1 (fine-tuned)	**0.843**	**0.887**	280	**0.096**
	all-mpnet-base-v2 (fine-tuned)	0.825	0.880	320	0.107
GWN	Random	0.288	0.348	3585	0.551
	Population	0.727	0.850	723	0.153
	GeoNorm	0.832	0.876	**54**	0.104
	all-distilroberta-v1	0.406	0.481	2782	0.441
	all-mpnet-base-v2	0.429	0.496	2335	0.421
	all-distilroberta-v1 (fine-tuned)	0.845	0.915	438	0.089
	all-mpnet-base-v2 (fine-tuned)	**0.862**	**0.925**	325	**0.075**
TRN	Random	0.253	0.308	4209	0.594
	Population	0.778	0.859	609	0.126
	GeoNorm	0.897	0.911	**36**	0.073
	all-distilroberta-v1	0.414	0.490	3352	0.446
	all-mpnet-base-v2	0.480	0.530	2126	0.383
	all-distilroberta-v1 (fine-tuned)	**0.939**	**0.975**	61	**0.021**
	all-mpnet-base-v2 (fine-tuned)	0.934	**0.975**	61	0.022

the training datasets with more numerous and varied text sources could potentially improve the models' robustness and applicability across different contexts.

Another limitation of our experimental setup was the relatively simplistic representation of location candidates with attributes sourced from GeoNames. Although the employed strategy was effective, there is substantial room for enrichment. Given the capacity of the models to process text sequences of 256-512 tokens, there is untapped potential for augmenting location descriptions. Incorporating additional information from knowledge bases or integrating spatial data, such as nearby landmarks or geographical features, could improve the models' ability to more accurately match toponyms to their geographical referents. Such an enhancement could lead to more nuanced associations between contextual cues in texts and specific location attributes, potentially increasing the resolution accuracy.

Our exploration was confined to the SentenceTransformers framework, which presented both advantages and limitations. The intuitiveness of the framework and the availability of pre-trained models for sentence similarity tasks provided a solid foundation for our experiments. Nonetheless, this choice came with certain architectural constraints. In particular, the generation of embeddings via mean pooling of whole text sequences raises questions about the optimal representation of toponyms, especially in sentences containing multiple toponyms. Further experiments will be necessary to explore whether single token embeddings might be more

effective when applying transformer models to the task of toponym resolution.

While our experiments were designed for comparability with the work of Zhang and Bethard [9], it is important to acknowledge the broader landscape of toponym resolution research. Hu et al. [16] provide a comprehensive overview of toponym resolution approaches, including their novel spatial clustering-based voting approach that combines several individual methods. Our method showed superior performance compared to all of these approaches on the tested datasets. However, this comparison may not be entirely fair, given that our models were trained on data sourced from the same domain used for testing. This scenario potentially provided our models with an inherent advantage over others evaluated by Hu et al. In future work we will attempt to replicate these frameworks using an out-of-domain dataset for training.

Finally, we are unsure why our models outperformed GeoNorm for all metrics except mean error distance. One possible explanation would relate to the possible candidate matches - by including alternative names and fuzzy matching we may penalise our approach, but this discrepancy again points to the difficulties in effectively comparing toponym resolution methodologies [12].

7. Conclusion

This paper has presented a proposed new approach to the application of transformer-based models, specifically the SentenceTransformers framework, to the task of toponym resolution. While the proposed methodology has shown promising results, achieving state-of-the-art performance, we currently view it as a proof of concept. Several elements of the proposed methodology, such as configurations and training paradigms, are preliminary and require further research and more rigorous evaluation. As such, the true extent and applicability of this novel approach remain to be fully realised and validated.

Project repository: https://github.com/dguzh/SemTopRes

References

[1] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, V. Murdock, Geographic information retrieval: Progress and challenges in spatial search of text, Foundations and Trends® in Information Retrieval 12 (2018) 164–318. doi:10.1561/1500000034.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv (2017). doi:10.48550/arXiv.1706.03762.

[3] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[4] B. J. Radford, Regressing location on text for probabilistic geocoding, in: Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political

Events from Text (CASE 2021), Association for Computational Linguistics, Online, 2021, pp. 53–57. doi:10.18653/v1/2021.case-1.8.

[5] A. B. Cardoso, B. Martins, J. Estima, A novel deep learning approach using contextual embeddings for toponym resolution, ISPRS International Journal of Geo-Information 11 (2022) Article No. 28. doi:10.3390/ijgi11010028.

[6] Y. Solaz, V. Shalumov, Transformer based geocoding, arXiv (2023). doi:10.48550/arXiv.2301.01170.

[7] A. Halterman, Mordecai 3: A neural geoparser and event geocoder, arXiv (2023). doi:10.48550/arXiv.2303.13675.

[8] Z. Li, W. Zhou, Y.-Y. Chiang, M. Chen, GeoLM: Empowering language models for geospatially grounded language understanding, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5227–5240. doi:10.18653/v1/2023.emnlp-main.317.

[9] Z. Zhang, S. Bethard, Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution, in: Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 48–60. doi:10.18653/v1/2023.starsem-1.6.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, United States, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, M. Shi, Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing, AGILE: GIScience Series 3 (2022) 1–13. doi:10.5194/agile-giss-3-9-2022.

[12] M. Gritta, M. T. Pilehvar, N. Limsopatham, N. Collier, What's missing in geographical parsing?, Language Resources and Evaluation 52 (2018) 603–623. doi:10.1007/s10579-017-9385-8.

[13] M. D. Lieberman, J. Sankaranarayanan, H. Samet, Geotagging with local lexicons to build indexes for textually-specified spatial data, in: 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), IEEE Computer Society, Los Alamitos, California, United States, 2010, pp. 201–212. doi:10.1109/ICDE.2010.5447903.

[14] M. Gritta, M. T. Pilehvar, N. Collier, A pragmatic guide to geoparsing evaluation, Language Resources and Evaluation 54 (2020) 683–712. doi:10.1007/s10579-019-09475-3.

[15] E. Kamalloo, D. Rafiei, A coherent unsupervised model for toponym resolution, in: Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 2018, pp. 1287–1296. doi:10.1145/3178876.3186027.

[16] X. Hu, Y. Sun, J. Kersten, Z. Zhou, F. Klan, H. Fan, How can voting mechanisms improve the robustness and generalizability of toponym disambiguation?, International Journal of Applied Earth Observation and Geoinformation 117 (2023) Article No. 103191. doi:10.1016/j.jag.2023.103191.