# Where is the News? Improving Toponym Identification and Differentiation in Online News

Joseph Shingleton[1,*], Ana Basiri[1,2]

[1]School of Geographical and Earth Sciences, The University of Glasgow, United Kingdom
[2]The Alan Turing Institute, London, United Kingdom

**Abstract**

Understanding the geographical context of unstructured textual data is a key challenge in information extraction. In many applications, however, simple identification of toponyms is insufficient and can often lead to ambiguities in the extracted information. One such application is in the geolocation of online news - where a single article may mention multiple locations, with only one location referring to the article's subject. In this paper, we present a transformer based model, trained to identify the subject toponym of news articles. Further, our model identifies likely parents of the subject toponym, potentially helping to improve later geolocation tasks. Our model is able to identify the subject toponym of an article with an F1-score of 0.760 when tested on a human-tagged test dataset.

**Keywords**

Natural language processing, Geoparsing, Toponym identification, Transformer models

## 1. Introduction

Accurate extraction of geographical information from natural language relies on the ability to reliably identify toponyms within text, and to associate those toponyms with unique geographic locations [1, 2, 3]. Modern toponym extraction methods, such as named-entity-recognition, have been shown to be highly effective in this task [4, 5, 6]. In most cases, however, these tools are limited to applications in which a geographic location is expected to be assigned to each toponym in a piece of text. For many applications this may be perfectly acceptable, and even desirable. However, there are some applications which require a more nuanced approach to toponym identification, such as geotagging of news articles [7, 8].

Often, the subject of a news article is associated with a single geographic location referred to in the text. Frequently, however, other toponyms will appear alongside the subject toponym, either as a way to help geographically identify the subject, or due to some interaction between the subject location and other named locations. For example, in the sentence *"Two firefighters have travelled almost 4,000 miles from the USA to confirm their vows at the Calton Community Fire Station in Glasgow"*, it is clear that the subject location of this article is *Calton Community Fire Station*. The article also mentions the toponyms *Glasgow*, and *USA*. In this context, the

*GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland*
*Corresponding author.
joseph.shingleton@glasgow.ac.uk (J. Shingleton); ana.basiri@glasgow.ac.uk (A. Basiri)
0000-0002-1628-3231 (J. Shingleton); 0000-0002-2399-1797 (A. Basiri)

toponym *Glasgow* is used to geographically identify the subject as it is a parent location of the subject. The toponym *USA* is mentioned due to an interaction between it and the subject toponym within the narrative of the article. Such toponyms can be considered *incidental* to the subject.

In this paper, we present a transformer based model which can identify the *subject* location of news articles, and differentiate between *parent* and *incidental* locations to aid in precise geolocation. The model is trained on data scraped from the BBC-Monitoring website [9], an online news platform which collects news from around the world on topics covering terrorism, conflict, misinformation and political extremism.

Various techniques exist for simple toponym recognition in news articles. Samet et al developed a model which combines a rules-based approach to toponym identification with a statistical named entity recognition model, achieving a precision of 0.739 and recall of 0.868 on a corpus of 11,564 news articles, out performing many models existing at the time [10]. Modern transformer based named-entity recognition models, such as Topo-BERT, have been shown to be highly effective in toponym recognition without the need for additional rules-based techniques. The Topo-BERT model achieved an average precision of 0.827 and recall of 0.886 across a range of news and social media data sources [4].

Monteiro et al provide a detailed survey of articles investigating the *geographic scope* of documents [8]. In this context, geographic scope refers to the identification of a single location, or multiple locations, which provide a broad geographic representation of all (or most) toponyms in a document. This differs from subject toponym identification, which aims to assign a single toponym (or multiple toponyms) in the text as the geographic subject of the document. For example, the sentence *"Firefighters from Motherwell and Edinburgh were called in to help fight the fire in Glasgow"* might have the geographic scope of Scotland, as it links the three named toponyms, despite the subject toponym of the article being Glasgow. Monteiro et al allude to this through the identification of *geographic semantic scope* as an area for future research, in which the semantic meaning of the document is considered alongside explicitly mentioned toponyms.

Previous approaches to subject-location identification tend to rely on heuristic models [7]. Such approaches use syntactical and contextual clues, such as a toponym's occurrence in a headline, its position and/or frequency within in the text, or the relative prominence of an associated location. These approaches, however, can not account for all of the grammatical nuance within natural language which help to identify subject locations, and are unable to identify spatial relationships between locations. Further, rules based approaches may suffer from reduced generalizability, as domain-specific language can lead to rules failing to translate across different types of text [8, 11].

Modern transformer based language models may help to address the poor generalizability of heuristic approaches. A recent paper by Tahmasebzadeh *et al.* [12] implemented a BERT based transformer model to identify the subject location of news articles. The implementation, however, limits the model to predicting locations from a pre-defined pool, reducing its utility in real world geo-parsing applications. Our model improves on the utility of this approach by classifying toponyms within the text, allowing the model to make predictions on previously unseen locations.

In this paper, we propose a transformer based model which is able to identify the subject

toponym of a news article and differentiate secondary toponyms in terms of their spatial relationship with the subject. To do this, we use a simple heuristic model to automatically tag a dataset of news articles. This noisy data is used to train a transformer based model, before fine-tuning the model on a smaller manually annotated dataset. The process of fine-tuning, along with the use of a relatively noise-robust transformer model, helps to alleviate some of the noise introduced by the heuristic tagging method.

## 2. Methods

The aim of this paper is to develop a Topo-BERT model trained on the task of subject toponym identification. To achieve this, we first need to construct a dataset of news articles with appropriate toponym tags. This process consists of three steps: accessing and downloading relevant news articles; generic identification of toponyms within articles via named entity recognition, and subsequent re-categorisation of the identified toponyms in terms of their relationship with an article's subject. The constructed dataset can then be used to train a Topo-BERT model on the task of subject identification and toponym differentiation.

### 2.1. Collecting news data

We use the BBC-Monitoring API [9] to access news articles. In order to link each article to a specific location, we search the API for articles by headline. We use GeoNames [13] to construct a list of 14,696 global cities with population greater than 100,000, and search for any BBC-Monitoring articles which mention each city in the headline. The coordinates of each city are also recorded. Capital cities are removed from the list to avoid metynomic use of capital cities as a reference to a country's government. In cases where a single name refers to two or more locations, the cities with the smaller population are removed from the list. Headlines which mention more than one place name are also removed from our dataset.

Geographic information, including coordinates, and (if available) bounding boxes or spatial polygons, associated with each city are obtained by querying OpenStreetMap's Nominatim API [14] with each city name. If multiple matches are found then we select only the match which contains the known coordinates of the city and which has the highest OpenStreetMap *importance* score, since a higher importance is associated with higher populations.

This process does introduce some noise into the training data. In particular, we can not be certain that the place referred to in the headline is always the article's subject location. For example, articles referencing the Nagoya Protocol may be identified as having Nagoya, Japan as a subject. Further, articles referring to places with ambiguous names may result in erroneous geographic information collected from OpenStreetMap. Using the city with the highest population goes some way to addressing this as more news articles are written about places with high populations [15].

### 2.2. Toponym identification in news articles

For our initial toponym identification, we train a Topo-BERT model to perform NER tagging on the Wiki-Neural [16] and CoNNL-2003 datasets [17]. Our Topo-BERT model is constructed

from a large, cased BERT model, which outputs into a one-dimensional convolutional layer with 16 nodes, connected to max-pooling layer. The output of this layer is passed into a fully connected layer with 512 nodes, before finally passing through a soft-max activated output layer [4]. We train the model over 20 epochs, using a weighted masked categorical loss function and an Adam optimizer. The loss function is weighted to help account for the class imbalance in the dataset [18].

### 2.3. Relational tagging of news articles

All toponyms identified in the previous step are again searched for using the Nominatim API, and the spatial information (points, polygons, or bounding boxes) of all matches are stored. If the toponym cannot be found using the Nominatim API then a further search is completed using the Geonames API, yielding only point information.

The extracted geographical information is used to identify spatial relationships between the subject toponym and other toponyms in the text. If the subject location is contained within any polygon (or bounding box) associated with a secondary toponym (across multiple possible matches), then the secondary toponym is tagged as a parent of the subject location. If the subject location contains any polygon (or point) associated with a secondary toponym, then we assume that the more geographically specific location is the actual subject of the article. Hence, the secondary toponym is tagged as the subject toponym, and the previously identified subject toponym is reassigned as a parent. By doing this, we ensure that the subject toponym relates to the most geographically specific location in the text.

Any locations which can not be found using either OpenStreetMap or GeoNames, or which have no relationship with the subject under the specified rules, are tagged as incidental locations. This is a common source of noise in the dataset, as there may be locations which have a parent/child relationship with the subject, but which cannot be found using the API tools. In such cases, the tagging method will introduce false negatives to the training data. To address this and other sources of labelling noise, we manually label a subset of 1343 sentences from the training data and use these to fine-tune the model in a final training step. This allows the model to retain some of the relationships learned during the initial noisy training step, while correcting for some of the inaccuracies introduced [19].

For both the noisy training step and the fine-tuning step we use the same model as described in the previous subsection. The model is again trained for 20 epochs in each case. The model which achieves the lowest weighted loss on a validation set (a random sample of 10% of our training set) is saved. As this work serves to act as a proof-of-concept we have not performed any hyper-parameter tuning and do not consider our results to be optimal.

## 3. Results

We use a test set of 200 human-tagged articles to assess the accuracy of the heuristic tagging method and the Topo-BERT model. For the Topo-BERT model, we present the accuracy after training on just the noisy data, just the fine-tuning data, and the noisy data plus the fine-tuning data. The recall, precision, and F1 score for each toponym type are given in table 1.

| Model | Label | Precision | Recall | F1 |
|---|---|---|---|---|
| Heuristic | Subject | 0.834 | 0.636 | 0.728 |
| | Parent | 0.798 | 0.634 | 0.707 |
| | Incidental | 0.468 | 0.961 | 0.629 |
| Noisy data only | Subject | 0.807 | 0.687 | 0.742 |
| | Parent | 0.687 | 0.602 | 0.642 |
| | Incidental | 0.410 | 0.669 | 0.509 |
| Fine-tuning data only | Subject | 0.712 | 0.768 | 0.739 |
| | Parent | 0.637 | 0.623 | 0.630 |
| | Incidental | 0.489 | 0.433 | 0.459 |
| Noisy data + fine-tuning | Subject | 0.814 | 0.713 | 0.760 |
| | Parent | 0.669 | 0.768 | 0.715 |
| | Incidental | 0.462 | 0.646 | 0.539 |

Table 1: Accuracy of the heuristic tagging method and the Topo-BERT model with and without fine-tuning.

The heuristic model achieves an F1 score of 0.728 when identifying subject toponyms and 0.708 when identifying parent toponyms. The precision of the model is generally very good in this category (0.834), with much of the error coming from poor recall (0.646), indicating a low false positive rate and a high false negative rate. A similar pattern is observed in the parent category (F1: 0.702, precision: 0.798, recall: 0.626), indicating that many of the false negatives might be misidentified as belonging to the incidental class. This is expected due to the limitations of the Open Street Map database used to perform the heuristic tagging.

Allowing the model to first train on the noisy data, before fine-tuning on the high-quality data improves the accuracy of the model. The final model outperforms the heuristic model on subject toponym identification (F1: 0.760) and parent toponym identification (F1: 0.715). Identification of incidental toponyms is reduced, however. Figure 1 shows that 17% of the toponyms tagged as incidental by the human reviewer are tagged as subject toponyms by the model, and 26% of model tagged incidental locations are tagged as subject locations by the human reviewer. The inability to differentiate between subject and incidental toponyms is likely due to the limited size of the human-tagged fine-tuning data. By increasing the number of high quality samples available in this step we will likely see improvements in the model.

## 4. Discussion

The methods developed in this paper provide a promising indication of the capacity for transformer based models in the geoparsing of news media. Existing approaches to subject toponym identification tend to rely on purely heuristic models [7], similar to the initial noisy tagging model presented in our work. Such models use specific structural features within text to make predictions. For many articles, however, such structural information (such as a name within a headline, or an expected order or frequency of toponyms) may be missing, leading to misleading results and reduced generalizability. A trained machine learning model, however, can

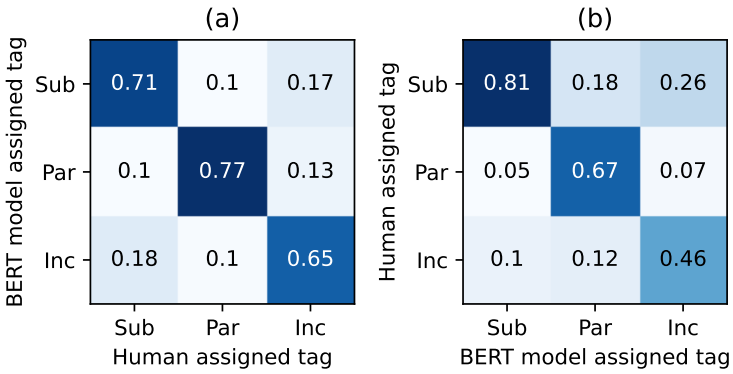


Figure 1: (a) the proportion of Topo-BERT model guesses in the subject (Sub), parent (Par) and incidental (Inc) classes, given the human assigned tag and (b) the proportion of human assigned tags in each category given the model prediction.

use grammatical indicators within the text which may be missed by rules-based approaches [20, 21, 22]. Because of this, the Topo-BERT model benefits from wider generalizability to a more diverse set of problems compared to the heuristic model. As such, the utility of the model exceeds the marginal improvements on identification of subject and parent toponyms. Further improvements to model accuracy may be achieved through hyper-parameter tuning, use of noise-robust loss functions, or through increasing the size of the manually annotated fine-tuning dataset.

Our model performs well on a human tagged test set, correctly identifying the subject toponym in 71% of cases. Differences in testing data and performance metrics means that it is difficult to draw comparison to existing models. The CLIFF-CLAVIN model [7] achieved an accuracy of 74.1% when identifying the subject country of an article, but has not been tested on city-level extraction. A more recent transformer-based model [12] achieved an accuracy of 48.1% and 53.4% when predicting the city and region of focus respectively. Our model appears to outperform this, however, this should be validated by testing both models on the same test data.

A further benefit of our model is its capacity to identify spatial relationships between the subject toponym and other toponyms in the text. This provides more spatial information for later geocoding steps and may improve geocoding accuracy, however this remains to be demonstrated fully. Other spatial relationships which aid in disambiguation have not been considered at this stage. Further work may try to further differentiate between toponyms through identification of locations which are near the subject, or have shared parental lineage. Including these more complex geographical relationships may improve the accuracy of later disambiguation methods [8].

A more nuanced approach to noise handling will likely further improve model accuracy. Approaches which attempt to identify mislabeled data [23], or establish robust classification boundaries [24] can help to reduce the effect of labeling noise on transformer based models.

Further work will aim to apply these noise handling techniques to our model to improve classification accuracy.

## Acknowledgments

The authors acknowledge the support from The UK Research and Innovation (UKRI) Future Leaders Fellowship on "Indicative Data", MR/S01795X/2, and the Alan Turing Institute-DSO partnership project on "Multi-Lingual and Multi-Modal Location Information Extraction".

## References

[1] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, Y. Kompatsiaris, Location extraction from social media: Geoparsing, location disambiguation, and geotagging, ACM Transactions Information Systems 36 (2018). URL: https://doi.org/10.1145/3202662. doi:10.1145/3202662.

[2] M. Gritta, M. Pilehvar, N. Collier, A pragmetic guide to geoparsing evaluation, Language Resources and Evaluation 54 (2019) 683–712. doi:10.1007/s10579-019-09475-3.

[3] M. Karimzadeh, S. Pezanowski, A. M. MacEachren, J. O. Wallgrün, Geotxt: A scalable geoparsing system for unstructured text geolocation, Transactions in GIS 23 (2019) 118–136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12510. doi:https://doi.org/10.1111/tgis.12510. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12510.

[4] B. Zhou, L. Zou, Y. Hu, Y. Qiang, D. Goldberg, Topobert: a plug and play toponym recognition module harnessing fine-tuned bert, International Journal of Digital Earth 16 (2023) 3045–3064. URL: https://doi.org/10.1080/17538947.2023.2239794. doi:10.1080/17538947.2023.2239794. arXiv:https://doi.org/10.1080/17538947.2023.2239794.

[5] C. Berragan, A. Singleton, A. Calafiore, J. Morley, Transformer based named entity recognition for place name extraction from unstructured text, International Journal of Geographical Information Science 37 (2023) 747–766. URL: https://doi.org/10.1080/13658816.2022.2133125. doi:10.1080/13658816.2022.2133125. arXiv:https://doi.org/10.1080/13658816.2022.2133125.

[6] L. Tao, Z. Xie, D. Xu, K. Ma, Q. Qiu, S. Pan, B. Huang, Geographic named entity recognition by employing natural language processing and an improved bert model, ISPRS International Journal of Geo-Information 11 (2022). URL: https://www.mdpi.com/2220-9964/11/12/598. doi:10.3390/ijgi11120598.

[7] C. D'Ignazio, R. Bhargava, E. Zuckerman, L. Beck, Cliff-clavin: Determining geographic focus for news articles, in: Proceddings of the NewskDD: Data Science for News Publishing, 2014.

[8] B. R. Monteiro, C. A. Davis, F. Fonseca, A survey on the geographic scope of textual documents, Computers Geosciences 96 (2016) 23–34. URL: https://www.sciencedirect.com/science/article/pii/S0098300416301972. doi:https://doi.org/10.1016/j.cageo.2016.07.017.

[9] British Broadcasting Corporation, BBC Monitoring, https://monitoring.bbc.co.uk/, 2024.

[10] M. D. Lieberman, H. Samet, Multifaceted toponym recognition for streaming news, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 843–852. URL: https://doi.org/10.1145/2009916.2010029. doi:10.1145/2009916.2010029.

[11] X. Li, W. Zhang, Y. Wang, Y. Tan, J. Xia, Spatio-temporal information extraction and geoparsing for public chinese resumes, ISPRS International Journal of Geo-Information 12 (2023). URL: https://www.mdpi.com/2220-9964/12/9/377. doi:10.3390/ijgi12090377.

[12] G. Tahmasebzadeh, E. Müller-Budack, S. Hakimov, R. Ewerth, Mm-locate-news: Multimodal focus location estimation in news, in: MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I, volume 13833 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 204–216. URL: https://doi.org/10.1007/978-3-031-27077-2_16. doi:10.1007/978-3-031-27077-2_16.

[13] GeoNames, GeoNames, http://geonames.org/, 2024.

[14] OpenStreetMap contributors, Planet dump retrieved from https://planet.osm.org , https://www.openstreetmap.org, 2017.

[15] E. Avraham, Cities and their news media images, Cities 17 (2000) 363–370. URL: https://www.sciencedirect.com/science/article/pii/S0264275100000329. doi:https://doi.org/10.1016/S0264-2751(00)00032-9.

[16] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, R. Navigli, WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2521–2533. URL: https://aclanthology.org/2021.findings-emnlp.215.

[17] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[18] G. King, L. Zeng, Logistic regression in rare event data, Political Analysis 9 (2001) 137–163. doi:https://doi.org/10.1093/oxfordjournals.pan.a004868.

[19] S. Ahn, S. Kim, J. Ko, S.-Y. Yun, Fine tuning pre trained models for robustness under noisy labels, 2023. arXiv:2310.17668.

[20] V. Nastase, P. Merlo, Grammatical information in BERT sentence embeddings as two-dimensional arrays, in: Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), 2023, pp. 22–39. URL: https://aclanthology.org/2023.repl4nlp-1.3. doi:10.18653/v1/2023.repl4nlp-1.3.

[21] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 3651–3657. URL: https://aclanthology.org/P19-1356. doi:10.18653/v1/P19-1356.

[22] H. J. Shin, J. Y. Park, D. B. Yuk, J. S. Lee, BERT-based spatial information extraction, in: Proceedings of the Third International Workshop on Spatial Language Understanding, 2020, pp. 10–17. URL: https://aclanthology.org/2020.splu-1.2. doi:10.18653/v1/2020.splu-1.2.

[23] S. Wang, Z. Tan, R. Guo, J. Li, Noise-robust fine-tuning of pretrained language models

via external guidance, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 12528–12540. URL: https://aclanthology.org/2023.findings-emnlp.834. doi:10.18653/v1/2023.findings-emnlp.834.

[24] R. Liu, S. Mo, J. Niu, S. Fan, CETA: A consensus enhanced training approach for denoising in distantly supervised relation extraction, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 2247–2258. URL: https://aclanthology.org/2022.coling-1.197.