AGILE: GIScience Series, 5, 12, 2024. https://doi.org/10.5194/agile-giss-5-12-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships

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

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

Correspondence: Joseph Shingleton (joseph.shingleton@glasgow.ac.uk)

Abstract. Geoparsing, the process of linking locations within text to sets of geographic coordinates, plays an important role in the extraction and analysis of information from unstructured textual data. With the rapid growth in availability of user-generated data from online sources, there is increasing demand for reliable geoparsing methods. Central to many of these methods is the accurate identification of toponyms within text. For some applications, however, simple identification of toponyms is insufficient. Problems which require the association of a piece of text containing multiple toponyms to a singular location require a more nuanced approach. In this paper, we show that a transformer based deep learning model, is able to identify the subject toponym within a given text, and classify other toponyms in terms of their spatial relationship with the subject. We curate a dataset of text taken from Wikipedia pages representing 5252 locations, and use OpenStreetMap data to classify toponyms within the text in terms of their spatial relationship with the subject of each article. This dataset is then used to train a transformer based deep-learning model. On a human labelled test set, our model achieves an F1 score of 0.916 when identifying the subject toponym, and 0.884 and 0.793 when identifying toponyms representing parent and child locations of the subject, respectively. We also consider the more complex adjacent and crossing relationships - with the model achieving F1 scores of 0.548 and 0.704 in these categories, respectively.

Keywords. Geoparsing, Natural Language Processing, Toponym Resolution, Transformer Model.

1 Introduction

With the rapidly growing availability of user generated data, there is an increasing need to develop effective tools for the extraction and analysis of information from unstructured textual data. These data represent a significant opportunity for researchers to learn more about an ever-changing world, however, the unstructured nature of text, and the inherent ambiguity of natural language, make this a challenging task. A key part of this process is in the association of textual data with geographic locations - a technique known as geoparsing. Effective geoparsing would allow the geographic information held in social media posts, online news, and open-source data, to be efficiently and reliably extracted. However, significant challenges arise from the complexity of geographic language, and the indeterminate nature of many place names.

The extraction of relevant and accurate geospatial information from text is an ongoing goal in natural language processing. Ambiguities arising from the use of metonymic language, linguistic diversity, common homonyms, and inconsistent grammatical indicators contribute to the complexity of this task (Gritta et al., 2018). Geo-parsing of text typically involves two main steps. First, named entity recognition (NER) is used to identify toponyms within the text, then a geocoding algorithm is used to associate these locations with a set of geographic coordinates (Gritta et al., 2019). Often, these methods use contextual information within the sentence or external geographic information to disambiguate between conflicting locations (Middleton et al., 2018). For example, a sentence which mentions the toponyms London and Canada would associate London with London, Ontario; while a sentence which mentions only the toponym London might resolve to London, UK, due to its higher population.

Such methods have been shown to be highly successful in the task of geotagging (Middleton et al., 2018; Berragan et al., 2023). However, there are many applications for which more detailed tagging is required. For example, the identification of the subject location of a social media post or news article can be crucial in identifying online

misinformation (Kordopatis-Zilos et al., 2017), in aiding disaster relief efforts (Hernandez-Suarez et al., 2019), and in disease surveillance (Ng et al., 2020). In each case, the aim of geoparsing is not to geocode every toponym within a piece of text, but rather to assign a single geographic location to the text. Posts or articles which mention multiple locations can hinder these applications, as it may not be clear which toponyms relate to the subject of the article (the *target* toponym) and which do not (*incidental* toponyms).

Disambiguation of subject and incidental toponyms introduces further complexity to the problem of geoparsing. A text with a single subject location may have several incidental toponyms mentioned. Such texts may appear easy for a human to resolve but can be challenging for geoparsing models. Consider, for example, the text:

"London is a city in Ontario, Canada. Its station, situated on York Street, has rail links to the neighbouring towns of Woodstock, and onward to Toronto."

It is clear to a human reader that this text refers to the city of London in Ontario, Canada. A standard geo-tagging algorithm, however, would identify a number of other place names (Ontario, Canada, York Street, Woodstock, Toronto), without being able to adequately identify London as the target toponym. Further, the identification of London as the target toponym may lead to ambiguities between London in the UK and London in Canada.

In this paper, we present a named-entity recognition model which is able to identify the primary subject location of a piece of text, and specify relationships between other named locations and the target. This approach allows for better resolution of geoparsing problems in which a single location is associated with a text containing multiple toponyms. Further, specifying spatial relationships between the target and other toponyms may allow for better disambiguation between locations matched with the target toponym.

Consider, for example, the sentence: *"I'm on York Street in London, getting ready for my trip to Ontario"*. In this sentence, York Street is the target toponym, and is a location within London. The language used also suggests that the toponym Ontario is incidental to the target. A traditional geoparsing model might recognise the locations London and Ontario and resolve the location of York Street to York Street, London, Ontario, Canada, despite the syntactical information within the text suggesting London and York St are not in Ontario. A model which is able to identify spatial relationships between toponyms would be more likely to resolve York Street to York Street, London, UK, given the grammatical information within the sentence.

Differentiation between types of toponym is an evolving area of study. Gritta et al highlight the importance of differentiating between place names, metonomyic names, geopolitical entities and other commonly misinterpreted toponyms to improve the precision and recall of models (Gritta et al., 2018). The linguistic characteristics of spatial relationships between objects (including but not exclusive to toponyms) are discussed in detail by Pustejovsky, from which the author is able to generate a formal spatial annotation scheme (Pustejovsky, 2017). Syed et al. demonstrate the ability of geoparsing models to identify relative spatial information within text, providing a system for models to resolve statements such as *"80 km south of Paris"* or *"On the border of France"* (Syed et al., 2022). Recent work by Balsbre et al uses a combination of knowledge graphs and transformer based language models to identify geospatial relationships between pairs of locations, based on textual descriptions gathered from online sources (Balsebre et al., 2023). Our work expands on these efforts, combining toponym identification and differentiation with the relational spatial reasoning demonstrated by transformer based models.

The model used to achieve this is built around a Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). BERT has previously been shown to reliably extract syntactical and semantic information from sentences, including identifying subject/object/verb relationships (Nastase and Merlo, 2023) and capturing structural linguistic information (Jawahar et al., 2019). BERT has also been shown to be effective in identifying spatial relationships within natural language, demonstrating a capacity to parse texts such as *"Tom is on the box"* or *"The cat is in the house"* (Shin et al., 2020). For this investigation, we use an adapted version of Topo-BERT (Zhou et al., 2023), a BERT based model which has been shown to be highly effective in the task of toponym identification.

In this paper, we leverage the capacity of Topo-BERT for accurate toponym extraction, and retrain it on a specially curated dataset describing spatial relationships between toponyms in text gathered from the Wikipedia pages for 5252 locations. To construct the dataset, we use a pretrained Topo-BERT model to identify toponyms within each article, before classifying the identified toponyms into six relational categories. The chosen categories represent distinct topological relationships between polygons Carniel (2023): *target*, indicating that the location is the subject of the article; *parent*, indicating a location which contains the target; *child*, indicating a location that is inside the target; *adjacent*, indicating a location that borders the target; *crossing*, indicating a location which crosses the target; and *incidental*, indicating that there is no relationship to the target within the specified rules. Figure 1 provides a visualisation of these spatial relationships. Relationships such as proximity (e.g. Glasgow is proximal to Edinburgh), or shared parental hierarchy (e.g. Glasgow has shared parental hierarchy with Dundee) are not considered in this paper. We use spatial data from Open Street Map (OpenStreetMap contributors, 2017) to identify these relationships through a rules-based tagging algorithm.

The toponym relationships investigated in this paper represent a broad, but not exhaustive list of the ways loca-

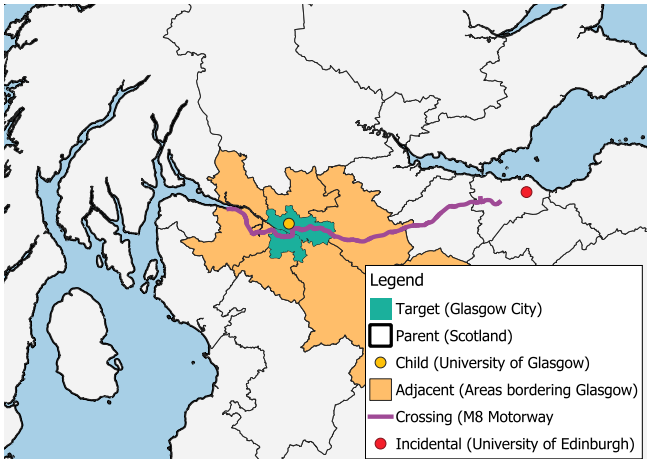

Figure 1. Examples of the different types of spatial relationships investigated in the paper, using Glasgow as the subject location.

tions may relate to one another. We have chosen relationships which are relatively simple to define under the constraints of the available data, and which represent different degrees of grammatical consistency within sentences. Our rules-based tagging approach uses geospatial data from an online database. Identified toponyms which are not represented in the database will yield false negatives in the tagging process. As such, we consider the constructed dataset to be weakly tagged, with a high number of tagging errors. Further, relationships such as adjacency and crossing may not have consistent grammatical representations within the text. A location may be identified as adjacent to the target by the rules-based tagging algorithm, without any indication of adjacency in the text. For example in the sentence *"In 2011, 43 700 people moved from Wales, Northern Ireland or England to live in Scotland"* there is no grammatical indication that England and Scotland share a border, so, beyond any previous examples seen in the training set, it would be unlikely that the trained model would identify the adjacency relationship. Target, parent, and child relationships tend to have much more consistent grammatical representations in the text, and so are likely easier for the model to identify.

Despite the complex nature of geospatial language, and the noise introduced to the dataset through the rules-based tagging process, we are able to show that the trained model can reliably identify toponyms referring to the subject location of a Wikipedia article, and can successfully classify subsequent toponyms in terms of the spatial relationships discussed above. We also show that, for some relationships, the trained transformer model is able to outperform the rules-based algorithm used to build the training set when tested on human tagged data. This may suggest that the model is robust to the noisy tagging regime, and is able to use grammatical indicators in the text to identify the spatial relationships.

2 Methods

We use a three-step method to train our relational-toponym model: toponym identification, rules-based relational tagging, and finally training the model on the rationally tagged data.

2.1 Toponym identification

For toponym identification, we train a Topo-BERT model (Zhou et al., 2023) on Named Entity Recognition (NER) tagged text from both the CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003b) and Wiki-Neural (Tedeschi et al., 2021) datasets. During this step, the model is trained to associate tokens with a collection of NER tags - [LOC] (location), [PER] (person), [ORG] (organisation) and [MISC] (miscellaneous). We use the Beginning-Inside-Outside (B-I-O) format to identify tags spanning multiple tokens (Tjong Kim Sang and De Meulder, 2003a), where [B-...] identifies the tokens comprising the first words of a phrase and [I-...] identifies tokens within additional words inside the same phrase. For example, the toponym "New York" would receive the tags [B-LOC] [I-LOC]. The [O] tag is used to identify tokens which do not fit into any of the stated categories. Words which can not be associated with a token are split into smaller tokens, for example "Los Angeles" might be split into tokens "Los", "Angel" and "##es", where "##" indicates that a token has been split from the previous token; these tokens would then receive the tags [B-LOC] [I-LOC] [I-LOC].

The Wiki-Neural dataset contains 2 193 680 tokens across 92 719 sentences in the training set and a further 267 156 tokens (11 528 sentences) of testing data, all extracted from Wikipedia articles. The ConLL-2003 dataset adds a further 839 238 training tokens (38 367 sentences) and 209 339 tokens (9 591 sentences) in the testing set, predominantly extracted from news articles.

Topo-BERT has previously been shown to be highly effective in location extraction on similar datasets (Zhou et al., 2023). The capability of the convolutional layers in spatial pattern recognition and signal processing tasks allows the model to achieve better results on toponym recognition tasks. As suggested by Zhou et al, we use a large, case-sensitive BERT model, connected to a convolutional layer with 1024 nodes. The output of this is then passed into a max-pooling layer, before being passed through a dense layer with 256 nodes, and a final output layer with a soft-max activation function.

The model is trained over 20 epochs, using a learning rate of 10^{-6} and a batch size of 4. We use an 8GB NVIDIA GTX 1080 graphical processing unit to train the model, which limits the maximum sentence length to 80 tokens. We use a masked categorical cross-entropy loss function, weighted to account for the class imbalance in the dataset. The weightings are applied such that for each category, $c_i \in \{c_1, c_2, ..., c_M\}$, the class weight W_i is given by:

$$W_i = \frac{N}{M \cdot N_i} \quad (1)$$

Where N is the total number of samples, M is the total number of classes in the dataset and N_i is the total number of samples in class c_i (King and Zeng, 2001).

2.2 Constructing a dataset of locations

In this step, we develop a dataset consisting of Wikipedia page summaries of cities and towns. As we will see in later steps, it is important that these locations are well represented in Open Street Map (OSM). To ensure this is the case, we use work by Hertford et al (Hertford et al., 2023) describing the completeness of urban building data in Open Street Map in cities around the world to estimate the completeness of OSM at a national level. We then identify a list of 30 countries which have reasonable OSM completeness, and use the GeoNames dataset of world cities with population over 1000 (GeoNames) to extract a list of cities within these countries.

This process results in a dataset of 72 757 cities. The set is further limited to those cities for which a spatial polygon can be found using OSM's Nominatim API, reducing the total count of cities to 9242. A further 4379 U.S. counties are added to the dataset in order to introduce data at a higher administrative level.

We use the Wikipedia API (Goldsmith, 2014) to extract the summary section from the Wikipedia page associated with each location. Often, however, the name of a location is identical to other locations or other unrelated words. In such cases, simply querying Wikipedia with the toponym is inadequate. We use multiple checks to confirm that the extracted Wikipedia page refers to the expected location. First, we check if the query leads to a disambiguation page, if so, we add the country (in brackets) to the search term. We then check that a word related to a location (e.g., city, town, township, village etc) exists in the first two sentences of the page summary, and if the location's country is mentioned within the first 3 sentences of the summary. If the Wikipedia page has an associated set of coordinates then we check that these coordinates are within the Nominatim polygon associated with the target location. If any of these checks fail, we add the country to the search term and begin the process again. If the page still can not be adequately associated with the target location then it is removed from the dataset. This method allows us to associate the 13 621 locations with a total of 5252 Wikipedia pages.

Once a page has been associated with a target location we extract the summary text from the page. As our model is limited to inputs of up to 80 tokens, we split the text into sentences (or groups of sentences) up to a maximum of 60 words, allowing for a buffer of 20 words to accommodate punctuation and tokens split across multiple words. We also remove any words inside parentheses - this is done to remove the pronunciation guides that are often used at the start of pages, as they typically contain exotic characters which are not present in the NER model training data. This process results in a set of 9243 sentences from 5252 Wikipedia pages, comprising a total of 609 298 tokens.

We then identify any toponyms within the Wikipedia page summaries using the trained Topo-BERT model. Other tags added by the model are removed, as for the purposes of this work we are only interested in toponyms within the text. Of the 609 298 tokens in the constructed dataset, the Topo-BERT model identifies 16 993 as belonging to toponyms.

2.3 Weak Relational Retagging

In this step, we use spatial data provided by OSM to further classify these toponyms according to their relationship with the subject location of the article. The new tags are produced from a set of rules linking true locations in OSM to test locations within the extracted Wikipedia text. These rules are reliant on the toponyms within the text being represented in the Open Street Map corpus, and often rely on a polygon existing for the location. As such, the labels produced by the tagging rules are often noisy and incomplete.

We use six tags to classify spatial relationships between identified locations - target, parent, child, adjacent, crossing and incidental. Thanks to the steps described in the previous section, querying OSM with the target location associated with each Wikipedia article will return a polygon (or multi-polygon) object. Hence, for each true location L_i there is at least one associated polygon, P_i, encoding the geopsatial characteristics of that location. By ensuring that the known coordinates of L_i lie within P_i we can limit this to a single polygon associated with L_i.

However, we can not guarantee that querying OSM with other locations within the text will return a unique result, or that any of the matched locations will be associated with a polygon. Hence for each location, l_j, there will be a set of matched locations $M = \{m_1, m_2, ..., m_N\}$ associated with l_j. Each $m_k \in M$ will have some geospatial data, P_j^k, associated with it, although this may not always include a polygon (i.e. it could be a point or a line-string).

Target locations are identified in two ways. For the true location, L_i, any location, l_j, which shares an identical string representation to L_i will be classed as a target location. For example, if the true location is "Los Angeles", any l_j represented by the tokens "Los Angeles" would be retagged as a target. A further condition compares the polygons associated with L_i and l_j. Querying OSM with location l_j will return a list of matched locations M_j. If any $m_j^k \in M_j$ has a polygon P_j^k for which the area of the intersection of P_i and P_j^k is equal to more than 80% of the area of both P_i and P_j^k, then l_j is also retagged as a target location. Hence, the test location "LA" will be tagged as a target

location if the true location is "Los Angeles", despite the different string representations.

Parent locations are also identified through polygon relationships. For a true location L_i with associated polygon P_i, a test location l_j is a parent of L_i if any polygon P_j^k associated with l_j contains P_i. Since the polygons identified by OSM can occasionally have poor resolution, this condition is weakened to include any polygon P_j^k which has an area larger than the area of P_i, and contains more than 80% of its area. If location l_j meets this condition then it is tagged as a parent location. If P_i contains any P_j^k then l_j is identified as a child location, once again under the same weakened condition. For the child relationship, the geometry of P_j^k can be either a polygon, a point or a line-string.

Polygon relationships are also used to identify locations which cross or are adjacent to the target location. For a true location L_i with associated polygon P_i, a test location l_j is crossing L_i if there is some P_j^k with line-string geometry, and the intersection of P_j^k with the boundary of P_i has point (or multi-point) geometry. Location l_j is adjacent to L_i if the geometry of the intersection of P_i and P_j^k is described by a line-string, and l_j is not crossing P_i. Once again, this condition is weakened to accommodate poor resolution so that l_j is adjacent if the minimum distance between P_i and any P_j^k is less than 1 kilometer, and it is not a parent or child of P_i.

Defining the functions $S(l)$ as the string representation of location l, $A(P)$ as the area of polygon P, $T(P)$ as the type of the geometric object P, $\mathcal{L}$ as denoting line-string (or multi-line-string) geometry, $\mathcal{P}$ as denoting point (or multi-point) geometry, ∂P as the boundary of polygon P, and $D(P_1, P_2)$ as the minimum distance between polygons P_1 and P_2, then these rules can be summarised as follows:

$$l_j \text{ is target if : } \begin{cases} S(l_j) = S(L_i) \text{ or} \\ \left(\frac{A(P_j^k \cap P_i)}{A(P_j^k)} > 0.8 \text{ and } \right. \\ \left. \frac{A(P_j^k \cap P_i)}{A(P_i)} > 0.8 \right) \end{cases}, \tag{2}$$

$$l_j \text{ is parent if : } \begin{cases} P_k \subset P_i \text{ or} \\ \left(A(P_j^k) > A(P_i) \text{ and } \right. \\ \left. \frac{A(P_j^k \cap P_i)}{A(P_i)} > 0.8 \right) \end{cases}, \tag{3}$$

$$l_j \text{ is child if : } \begin{cases} P_i \subset P_j^k \text{ or} \\ \left(A(P_i) > A(P_j^k) \text{ and } \right. \\ \left. \frac{A(P_j^k \cap P_i)}{A(P_j^k)} > 0.8 \right) \end{cases}, \tag{4}$$

$$l_j \text{ is crossing if : } \begin{cases} T(P_i) = \mathcal{L} \text{ and} \\ T(P_j^k \cap \partial P_i) = \mathcal{P} \end{cases}, \tag{5}$$

$$l_j \text{ is adjacent if : } \begin{cases} (T(P_j^k \cap P_i) = \mathcal{L} \text{ and} \\ (P_i \cap P_j^k) \backslash (\partial P_j^k) = \emptyset) \text{ or} \\ D(P_j^k, P_i) < 1\,\text{Km} \end{cases}. \tag{6}$$

In some circumstances a test location l_j may have multiple rules satisfied across its matched locations, M. In this case, we apply a hierarchical system which favours the classification of l_j as a target location first, then a parent location, then an adjacent location, then a crossing location, then a child location. Any location which returns no matches from the Nominatim API, or can not be tagged as either target, parent, child, crossing or adjacent, is tagged as an incidental location.

This set of rules can then be applied to the tagged locations dataset to produce a new dataset in which all tokens initially tagged as toponyms are retagged as either target, parent, child, crossing, adjacent or incidental toponyms. After being retagged, the dataset is split into a training set (8339 sentences comprising of 549 276 tokens) and testing set (904 sentences comprising of 60 022 tokens). We also manually tag 200 sentences from the test set to provide further validation for the methods.

The trained Topo-BERT model is then retrained on the retagged dataset. We train over 20 epochs, using a weighted masked categorical cross entropy loss function. The model has been trained across learning rates $lr \in \{1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$, and batch sizes $b \in \{1, 2, 4, 8, 12\}$. The hyper-parameters which maximized the micro-F1 score of the model on the test data were $lr = 2 \times 10^{-6}$ and $b = 4$.

2.4 Data and Software Availability Section

The code and data associated with this paper can be found in the Open Science Foundation repository: https://osf.io/waf2q/. This includes the software package developed to build and train the BERT model, a series of Jupyter notebooks demonstrating the analytical pipeline, and a Read-Me file detailing the environment requirements necessary to reproduce the results.

3 Results

3.1 Training the initial Topo-BERT model

Following the proposed methods, we trained our initial Topo-BERT model on the CoNLL-2003 and Wiki-Neural datasets. The accuracy of the model after training on each dataset separately, and the results after training on both sets combined, are given in table 1. In each case we have tested the model on the Wiki-Neural test set, as this is most similar to the location data used in the later stages of this

paper. We report the F1 score on the [B-LOC] and [I-LOC] tags(Bramer, 2013), as well as the overall micro and macro averaged F1 score across all tags (Takahashi et al., 2022). Accuracy on other tags has not been reported as we are primarily interested in the model's ability to identify toponyms within the text.

Model	[B-LOC] F1	[I-LOC] F1	Macro F1	Micro F1
CoNLL-2003	0.881	0.751	0.625	0.951
Wiki-NEuRal	0.913	0.897	0.917	0.978
Combined	0.920	0.899	0.922	0.978

Table 1. Accuracy of the Topo-BERT model trained on the CoNLL and Wiki-Neural datsets, and tested on the Wiki-Neural test set.

3.2 Retagging Wikipedia location data with relational tags

The trained Topo-BERT model is able to identify toponyms with a high degree of accuracy. As such, we can be confident that using the model to identify toponyms in our new dataset of Wikipedia articles will not introduce significant noise to the relationally tagged dataset. After tagging the collected Wikipedia articles with NER tags, we re-tag all locations using the previously described rules. To assess the accuracy of the tagging method, we manually tag a sample of 200 sentences (13 395 tokens) from the re-tagged data. In order to mimic the eventual machine-learned tagging, the human tagging is completed using only the grammatical indicators within the sentence, and is independent of any external geographical knowledge. Table 2 shows the accuracy of the rules-based tagging process compared to the human tagging. The abbreviations *TAR*, *PAR*, *CHI*, *ADJ*, *CRO* and *INC* refer to target, parent, child, adjacent, crossing and incidental toponyms, respectively. The human tagging process does not use the B-I-O format and so the B and I prefixes in algorithm applied tags have been ignored when comparing the two.

Tag	N_{human}	N_{alg}	Precision	Recall	F1
TAR	796	816	0.866	0.948	0.905
PAR	1472	1178	0.976	0.817	0.889
CHI	350	183	0.905	0.543	0.679
ADJ	380	360	0.697	0.713	0.705
CRO	50	74	0.500	0.640	0.561
INC	498	941	0.448	0.850	0.586

Table 2. The accuracy of the rule-based relational re-tagging algorithm compared to human tagging of a sample of 200 sentences from the Wikipedia locations dataset. The columns N_{human} and N_{alg} show the total number of tags assigned in each category by the human reviewer and rules-based algorithm respectively.

The rules based algorithm can effectively identify target and parent toponyms (F1 0.905 and 0.889 respectively), and is reasonably effective at identifying child and adjacent toponyms (F1 0.679 and 0.705, respectively). Identification of crossing and incidental toponyms is less consistent (F1 0.561 and 0.586). Figure 2 helps to explain some of the sources of error. For each human tagged token, figure 2 shows the proportion of times the algorithm guessed each of the six tags. The central diagonal line is the proportion of human tagged tokens which were given the same tag by the algorithm - hence this is the recall of the algorithm (assuming the human tagging to be a ground truth).

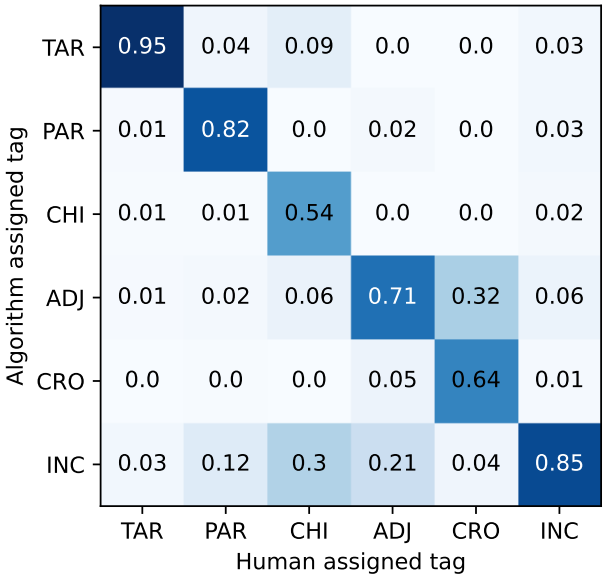

Figure 2. The distribution of tags applied by the retagging algorithm given the initial tag assigned by a human reviewer.

The algorithm has poor precision on crossing and incidental toponyms (0.500 and 0.448 respectively), indicating a large number of false positives. This is to be expected for incidental locations, as this group includes locations which could not be identified within the OSM data. As shown in figure 2, 32% of crossing locations are mislabeled as adjacent, explaining the poor precision in this category.

The main source of error for child toponyms comes from poor recall (recall 0.543). The reason for this error is suggested in figure 2, which shows that 30% of human-tagged child locations are tagged by the algorithm as incidental. This is expected, since the category naturally represents smaller, less prominent places which are less likely to be represented in the OSM database. This is not an easy problem to overcome, and is likely a significant source of noise in the relationally tagged dataset.

3.3 Retrained relational Topo-BERT model

Table 3 shows the accuracy of the Topo-BERT model after retraining on the relationally tagged location dataset. Figure 3 illustrates the source of error in the model, with figure 3a showing the distribution of predicted tags for each

assigned tag, and the inverse shown in figure 3b. Hence, the central horizontal of figure 3a indicates recall of the model, while figure 3b shows its precision.

The model is able to accurately identify both target and parent locations (weighted average F1 score 0.937 and 0.912 respectively). This is likely thanks to the high accuracy of the rules-based tagging method when classifying these types of locations, leading to less noise in the training data. Additionally, the parent and target tags often have consistent grammatical indicators which the model is able to interpret, such as target locations occurring at the start of a sentence, or parent locations proceeding the word 'in'.

The model has a higher recall rate than precision rate for most tags, indicating a significant number of false positive predictions. Figure 3b helps to explain the source of the poor model precision. Tokens tagged by the model as adjacent or child toponyms were frequently tagged as incidental by the rules-based algorithm. This may indicate that the model is using the grammatical context within the text to assign the correct token to toponyms which were not represented in OSM. Tokens which received the adjacent tag from the rules-based algorithm were frequently mislabeled as crossing or incidental by the model. For crossing locations, this is likely a reflection of the poor precision of the algorithm in applying the adjacent tag, whereas toponyms predicted as belonging to the incidental category likely did not provide enough syntactical context for the model to confidently assign the adjacent tags.

The limitations of the feature set are also highlighted in figure 3a. Many tokens originally assigned child and adjacent tags by the rules-based algorithm are identified as incidental locations by the Topo-BERT model. This again suggests that there is insufficient information provided by the feature set to successfully disambiguate the relationship between the given location and the assumed target.

Tag	N_{alg}	N_{mod}	Precision	Recall	F1
B-TAR	3116	2852	0.894	0.977	0.934
I-TAR	681	710	0.928	0.968	0.948
B-PAR	4683	4364	0.891	0.956	0.922
I-PAR	963	1102	0.807	0.923	0.861
B-CHI	1109	954	0.716	0.832	0.770
I-CHI	283	313	0.546	0.604	0.574
B-ADJ	1175	1253	0.546	0.512	0.528
I-ADJ	588	660	0.486	0.546	0.514
B-CRO	215	230	0.696	0.804	0.746
I-CRO	70	98	0.643	0.900	0.750
B-INC	3074	2391	0.698	0.543	0.610
I-INC	1431	1176	0.675	0.555	0.609
O	43226	43180	0.994	0.993	0.994
Macro Avg	-	-	0.732	0.778	0.751
Micro Avg	-	-	-	-	0.935

Table 3. Accuracy of the Topo-BERT model after retraining on the relationally tagged dataset. The columns N_{alg} and N_{mod} give the total number of tags assigned in each category by the rules-based algorithm and by the BERT model respectively.

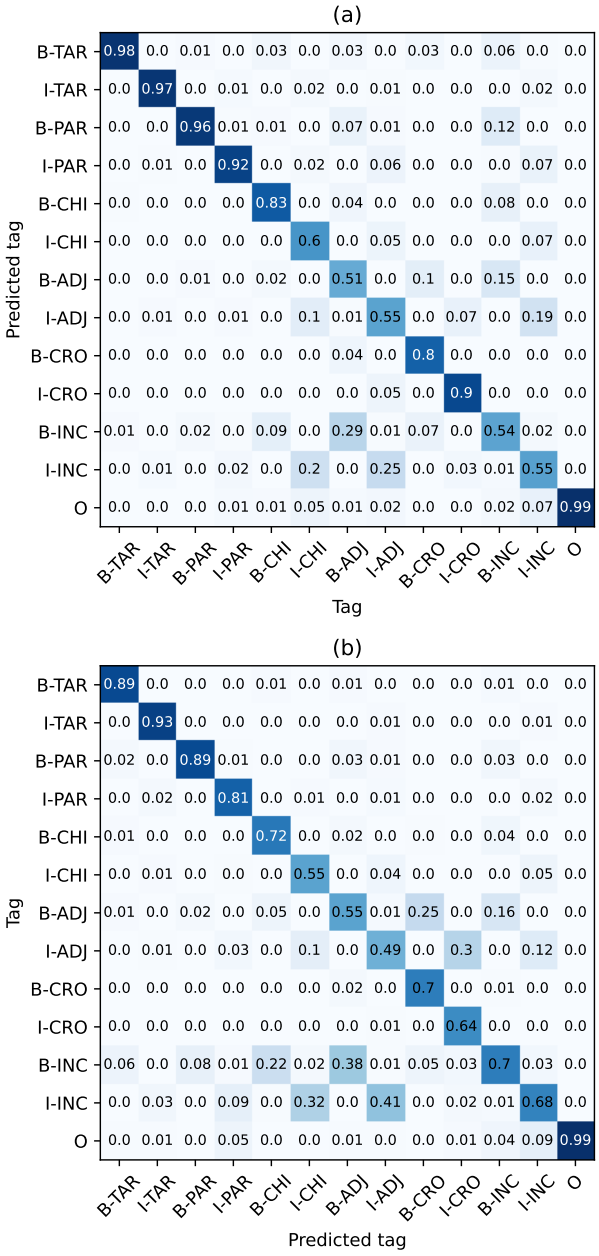

Figure 3. (a) The distribution of tags predicted by the Topo-BERT model in the test dataset given the tag applied by the rules-based tagging algorithm. (b) the distribution of tags applied by the rules-based tagging algorithm on the test dataset given the predicted tag from the Topo-BERT model.

3.4 Comparison of BERT model with human tagging

Table 4 gives the accuracy metrics of the BERT model when tested against the 200 human reviewed sentences. None of these sentences were included in the training set for the model. The classification errors are highlighted in figure 4, which once again shows the distribution of BERT model applied tag, given the initial tag assigned by the human reviewer.

Tag	N_{human}	N_{mod}	Precision	Recall	F1
TAR	796	849	0.987	0.946	0.916
PAR	1472	1249	0.944	0.830	0.884
CHI	350	260	0.931	0.691	0.793
ADJ	380	364	0.560	0.537	0.548
CRO	50	58	0.655	0.760	0.704
INC	498	756	0.443	0.673	0.534

Table 4. Accuracy of the trained BERT model compared to human tagging of a sample of 200 sentences from the Wikipedia locations dataset. The columns N_{human} and N_{model} give the total number of tags assigned in each category by the human reviewer and BERT model respectively.

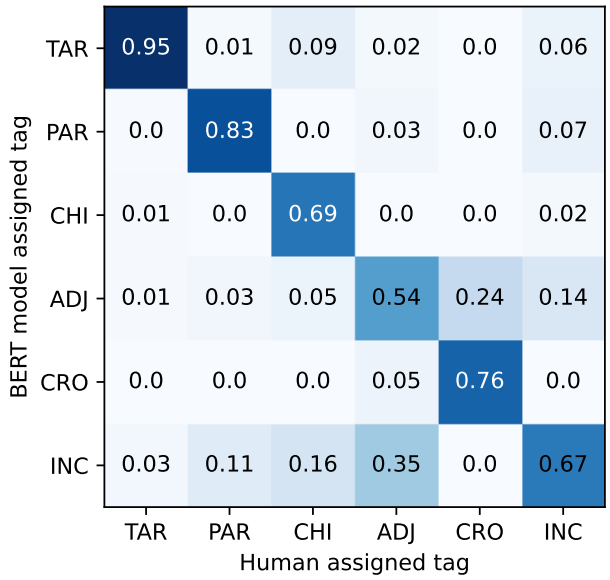

Figure 4. The distribution of tags applied by the Topo-BERT model, given the initial tag assigned by a human reviewer.

The trained model identifies target and parent toponyms with accuracy comparable to that of the rules-based algorithm (F1 for the BERT model of 0.916 and 0.884 respectively, compared to 0.905 and 0.889 for the rules-based algorithm). Child toponyms are more reliably identified by the BERT model (BERT model F1: 0.793, algorithm F1: 0.679), indicating that the model is correcting for some of the algorithm's misclassifications in this category. Figure 4 shows that some of the error in this category is attributable to misclassification to the incidental category, suggesting that grammatical indications of the relationship are occasionally not present in the text. Similarly, crossing toponyms are identified more reliably by the model (F1 0.704) compared to the algorithm (0.561), although as shown in fig 4, a significant proportion (24%) of crossing locations are misclassified as adjacent locations.

Adjacent toponyms are identified with less accuracy compared to the rules-based algorithm (F1 0.548 and 0.713). Figure 4 indicates that much of this error is due to the misclassification of adjacent locations as incidental. The poor performance in this category is likely attributable to both the introduction of noise during the rules-based tagging step and, in some cases, a lack of grammatical indications of the relationship.

4 Discussion

Identification of spatial relationships between toponyms in text is a complex, and sometimes unsolvable problem. Syntactical indicators can consistently be associated with some relationships, such as the target toponym, or parent and child toponyms, while other relationships are less reliably sign posted. The results described in this paper reflect these solvability issues. In this paper, we have demonstrated the suitability of a BERT based transformer model in differentiating between toponyms in text, and in identifying spatial relationships between identified toponyms. Our model is able to reliably identify the subject toponym of text extracted from Wikipedia, and can consistently identify mentions of parent and child locations of the subject.

The model is less reliable at identifying adjacent and incidental locations. Much of this inaccuracy can be attributed to two key sources: limitations in the data sources used to build the training set, and limitations in the modelling and methodology.

The relationally tagged dataset used to train the model is constructed using open-source data from both Wikipedia and Open Street Map. The size and diversity of these data sources have been crucial in developing a sufficiently large training set. Open source data, however, are known to have issues with completeness (Hertford et al., 2023) and accuracy (Haklay, 2010; Zhou et al., 2022). By considering only locations within countries which have high OSM completeness, we have attempted to mitigate some of these effects, however this has not always been possible. The most significant effect of OSM incompleteness is in the misclassification of child locations as incidental, due to them not being found in the OSM database. Similarly, under the rules used to produce our tagged dataset, adjacent or crossing locations which do not have associated polygons or line-strings are also unable to be accurately classified.

These inaccuracies introduce significant noise to the dataset and ultimately limit the power of the model. The BERT model used in this paper, however, has previously been shown to be reasonably robust to label noise, even without the introduction of noise correcting techniques (Zhu et al., 2022). Our results suggest that the model is able to account for the introduction of noise to a certain degree, especially when classifying the commonly mislabeled child toponyms.

The ability of the model to identify relationships between toponyms is dependent on its ability to associate syntacti-

cal indicators, such as *"Glasgow is in Scotland"* or *"Scotland neighbours England"* with spatial relationships. As previously discussed, however, these indicators are not always present. Our data labelling method uses geographic relationships as a ground truth, rather than grammatical indicators in the text. This can lead to potentially unresolvable labelling. For example, the toponym Lanarkshire in the sentence *"Glasgow Central Station serves the southern suburbs of the city, as well as Lanarkshire and the Clyde coast."*, would be labeled as adjacent to the target (since Lanarkshire shares a border with Glasgow), despite no grammatical indication of the relationship existing in the text. As such, some of the labels in the dataset, while geographically true, may not provide sufficient grammatical evidence to be resolved by the model. This is particularly true for the adjacent class, as reflected in the model's tendency to consistently assign such toponyms to the incidental class.

The definitions used to classify spatial relationships between toponyms are a further limitation of the approach. For the target, child and parent relationships, we include a tolerance parameter which helps to account for inconsistencies in the geospatial data acquired from OSM. We have set this parameter such that two polygons are assumed to represent the same location if they share 80% of each others area; and that a polygon is a parent of a child polygon if it is larger than the child and contains 80% of its area. These choices have not been optimized and may introduce further labelling errors into the training data. Further work may look at fine-tuning these parameters over a subset of human tagged data to better quantify and minimize this interference.

The training data for our model consists entirely of text taken from Wikipedia articles. As previously discussed, this allows for a large and diverse set of locations to be represented. The highly structured and formal linguistic style of Wikipedia articles, however, may result in poor performance when applying the model to out-of-domain problems. While this has not been investigated in this paper, further work might aim to incorporate news articles and social media posts into the modelling process. This will likely require adjustment to the labelling algorithm, however, as their may be ambiguities around the subject location of the text, particularly in the case of social media posts.

Despite the inherent complexity of geospatial language, and the difficulty in producing clean, noise-free training data, the Topo-BERT model is able to identify spatial relationships between toponyms with a good degree of accuracy. Comparing the results against human tagged data suggests that the model is able to use grammatical indicators within the text to achieve this. While future research may consider reducing labelling errors and applying the model to out-of-domain examples, the results presented here provide a promising indication that toponym differentiation with respect to spatial relationships is achievable using modern transformer based models.

5 Acknowledgment

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"

6 Author Contribution

JS contributed to conceptualisation and scoping, data collection and wrangling, software development, methodology and analysis, writing, visualisation, editing and reviewing. AB contributed to ideation, conceptualisation, scoping, editing and reviewing.

References

Balsebre, P., Yao, D., Cong, G., Huang, W., and Hai, Z.: Mining Geospatial Relationships from Text, Proc. ACM Manag. Data, 1, https://doi.org/10.1145/3588947, 2023.

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

Bramer, M.: Measuring the Performance of a Classifier, in: Principles of Data Mining, pp. 175–187, Springer, London, https://doi.org/10.1007/978-1-4471-4884-5_12, 2013.

Carniel, A.: Defining and designing spatial queries: the role of spatial relationships, Geo-spatial Information Science, 0, 1–25, https://doi.org/10.1080/10095020.2022.2163924, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: 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), pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423, 2019.

GeoNames: www.geonames.org, accessed on 2023-12-19.

Goldsmith, J.: Wikipedia API for Python, https://pypi.org/project/wikipedia/, 2014.

Gritta, M., Pilehvar, M., and Limsopatham, N.: What's missing in goegraphical parsing?, Language Resources and Evaluation, 52, https://doi.org/10.1007/s10579-017-9385-8, 2018.

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

Haklay, M.: How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets, Environment and Planning B: Planning and Design, 37, 682–703, https://doi.org/10.1068/b35097, 2010.

Hernandez-Suarez, A., Snachez-Perez, G., Toscano-Medina, K., and H., P.-M.: Using Twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with

word embeddings and kernel density estimation, Sensors, 19, e1746, https://doi.org/10.3390/s19071746, 2019.

Hertford, B., Lautenbach, S., and Porto de Albuquerque, J. e. a.: A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap, Nature Communications, 14, 3985, https://doi.org/10.1038/s41467-023-39698-6, 2023.

Jawahar, G., Sagot, B., and Seddah, D.: What Does BERT Learn about the Structure of Language?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657, Association for Computational Linguistics, https://doi.org/10.18653/v1/P19-1356, 2019.

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

Kordopatis-Zilos, G., Papadopoulos, S., and Kompatsiaris, I.: Geotagging Text Content With Language Models and Feature Mining, Proceedings of the IEEE, 105, 1971–1986, https://doi.org/10.1109/JPROC.2017.2688799, 2017.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., and Kompatsiaris, Y.: Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging, ACM Transactions Information Systems, 36, https://doi.org/10.1145/3202662, 2018.

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

Ng, V., Rooe, E., Niu, J., Zaghlool, A., Ghiasbeglou, H., and Verster, A.: Application of natural language processing algorithms for extracting information from news articles in event-based surveillance, Canada Communicable Disease Report, 46, 186–91, https://doi.org/10.14745/ccdr.v46i06a06, 2020.

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

Pustejovsky, J.: ISO-Space: Annotating static and dynamic spatial information, in: Handbook of Linguistic Annotation, edited by Ide, N. and Pustejovsky, J., chap. 37, pp. 989–1024, Springer, Dordecht, 2017.

Shin, H. J., Park, J. Y., Yuk, D. B., and Lee, J. S.: BERT-based Spatial Information Extraction, in: Proceedings of the Third International Workshop on Spatial Language Understanding, pp. 10–17, https://doi.org/10.18653/v1/2020.splu-1.2, 2020.

Syed, M. A., Arsevska, E., Roche, M., and Teisseire, M.: GeoXTag: Relative Spatial Information Extraction and Tagging of Unstructured Text, AGILE: GIScience Series, 3, 16, https://doi.org/10.5194/agile-giss-3-16-2022, 2022.

Takahashi, K., Yamamoto, K., Kuchiba, A., and Koyama, T.: Confidence interval fro micro-averaged F1 and macro-averaged F1 scores, Applied Intelligence, 52, 4961–4972, https://doi.org/https://doi.org/10.1007/s10489-021-02635-5, 2022.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R.: WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER, in: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2521–2533, Association for Computational Linguistics, Punta Cana, Dominican Republic, https://aclanthology.org/2021.findings-emnlp.215, 2021.

Tjong Kim Sang, E. F. and De Meulder, F.: 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, pp. 142–147, https://aclanthology.org/W03-0419, 2003a.

Tjong Kim Sang, E. F. and De Meulder, F.: 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, pp. 142–147, https://.aclweb.org/anthology/W03-0419, 2003b.

Zhou, B., Zou, L., Hu, Y., Qiang, Y., and Goldberg, D.: TopoBERT: a plug and play toponym recognition module harnessing fine-tuned BERT, International Journal of Digital Earth, 16, 3045–3064, https://doi.org/10.1080/17538947.2023.2239794, 2023.

Zhou, Q., Wang, S., and Liu, Y.: Exploring the accuracy and completeness patterns of global land-cover/land-use data in OpenStreetMap, Applied Geography, 145, 102 742, https://doi.org/10.1016/j.apgeog.2022.102742, 2022.

Zhu, D., Hedderich, M. A., Zhai, F., Adelani, D., and Klakow, D.: Is BERT Robust to Label Noise? A Study on Learning with Noisy Labels in Text Classification, in: Proceedings of the Third Workshop on Insights from Negative Results in NLP, pp. 62–67, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.insights-1.8, 2022.