Basreh or Basra? Geoparsing Historical Locations in the Svoboda Diaries

Jolie Zhou
University of Washington
joliez@uw.edu

Camille Lyans Cole
Illinois State University
clcole5@ilstu.edu

Annie T. Chen
University of Washington
atchen@uw.edu

Abstract

Geoparsing, the task of assigning coordinates to locations extracted from free text, is invaluable in enabling us to place locations in time and space. In the historical domain, many geoparsing corpora are from large news collections. We examine the Svoboda Diaries, a small historical corpus written primarily in English, with many location names in transliterated Arabic. We develop a pipeline employing named entity recognition for geotagging, and a map-based generate-and-rank approach incorporating candidate name augmentation and clustering of location context words for geocoding. Our system outperforms existing map-based geoparsers in terms of accuracy, lowest mean distance error, and number of locations correctly identified. As location names may vary from those in knowledge bases, we find that augmented candidate generation is instrumental in the system's performance. Among our candidate generation methods, the generation of transliterated names contributed the most to increased location matches in the knowledge base. Our main contribution is proposing an integrated pipeline for geoparsing of historical corpora using augmented candidate location name generation and clustering methods – an approach that can be generalized to other texts with foreign or non-standard spellings.

1 Introduction

In the digital humanities, natural language processing tasks such as named entity recognition (NER) and named entity linking (NEL) are valuable for connecting historical information to present-day knowledge. The digitization of historical documents involves additional challenges that can impact the quality of NER and NEL results, including patterns of word usage that can differ from modern-day language, and bias from optical character recognition (OCR) in the digitization process (Linhares Pontes et al., 2020; Ehrmann et al., 2021). Personal documents such as diaries present additional challenges in language processing, such as the personal shorthand of diary authors.

Geoparsing, the task of assigning coordinates to locations extracted from free text, is an important task in the digital humanities for understanding the geospatial information embedded in documents. However, geoparsing usually requires large and heterogeneous data sources (Rupp et al., 2014). Historical NER corpora (Ehrmann et al., 2021) and recommended geoparsing datasets WikToR, Local Global Lexicon (LGL), Tr-NEWS, and GeoWebNews (Gritta et al., 2020) are predominantly news corpora that involve a large volume of data. Frequently used historical geoparsing corpora such as the War of The Rebellion (WoTR) (DeLozier et al., 2016), a collection of official records, and Corpus of Lake District Writing (CLDW) (Rayson et al., 2017), are also large. Geoparsing systems designed for the digital humanities, such as the Edinburgh Geoparser (Filgueira et al., 2020), require large historical gazetteers or news collections for location data (Alex et al., 2015).

Many geoparsers focus on disambiguation, which involves identifying the correct location to match a name from a pool of potential candidates. A challenge is that data may not exist for some locations. This is especially relevant for historical documents because the location name may have changed or is too fine-grained to be identified. Documents can also incorporate foreign words or use non-standard spellings, exacerbating the challenge of retrieving correct entries from gazetteers.

We explore the following primary research questions: how do geoparsing methods perform on a small historical corpus? How can data augmentation via candidate name generation increase the effectiveness of geoparsing methods? We examine geoparsing in a small corpus of personal diaries by Joseph Svoboda. The Svoboda diaries pose a unique challenge in that they are written primarily

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 273–286

in English, but most of the location names originate from Arabic. In combination with the personal nature of the document, the author is likely to spell names differently from modern-day standards for location names in English, making it challenging to successfully retrieve coordinate data from modern-day gazetteers. We tackle this task by generating candidates from knowledge bases and ranking them, assigning the highest-ranking coordinates to the target location.

In this paper, we perform the following tasks: 1) perform geotagging using an NER + NEL pipeline adapted from extant literature; 2) develop a map-based generate-and-rank geoparsing method with enhanced candidate generation and compare it to other methods; and 3) examine the candidate generation portion in our proposed method to elucidate the most important contributors to its performance. We demonstrate that candidate name generation through the generation of orthographic variants of toponym names and clustering of toponym context together, can serve as a powerful approach to identifying suitable coordinates in historical corpora containing location names from other languages.

2 Background

2.1 Corpus

The Svoboda diaries are written by Joseph Mathia Svoboda, a purser on a British steamship in Ottoman Iraq during the late 19th century (Svoboda Diaries Project, 2024). Between the 1860s and 1908, Svoboda kept 61 diaries. The handwritten pages capture aspects of daily life, trade, and culture, and serve as a rich resource for the region and time period. Diaries 47 to 49 cover the period from late 1897 to 1899 and are publicly available as scanned images and text transcriptions on the Svoboda Diaries Project website.[1] We employ diaries 47 and 48 in this study (Table 1). Of the 300 total unique toponyms across both diaries, 92 of the toponyms appear in both diary 47 and diary 48.

[1] https://www.svobodadiariesproject.org/svoboda-diaries-data/

In his capacity as a purser, Svoboda regularly traveled by steamship up and down the Tigris River between Baghdad and Basra. He typically begins each entry with the time, date, and weather, and documents his travels, including when he stops along the river for any period of time. As such, most of the locations in the diaries are clustered near the Tigris River, as in Figure 1.

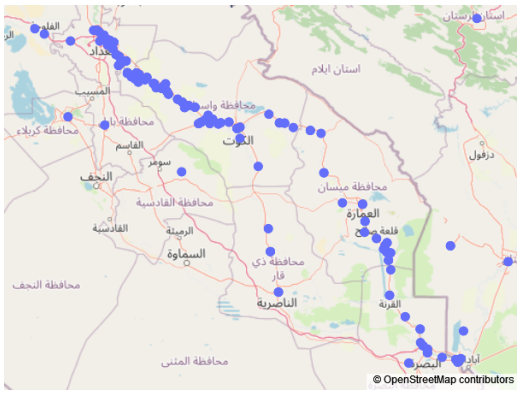

Figure 1: Plot of toponyms near the Tigris River system.

Locations around the world also appear in the diaries when Svoboda mentions the origin and background of the steamship passengers, and when he corresponds with his son, Alexander Svoboda, who often travels in Europe. Locations also appear when Svoboda writes of the mail and telegrams that he sends and receives, as he often notes the origin, destination, and locations the correspondence was posted through. Thus, a wide geographic spread of infrequent mentions sprinkled throughout in an otherwise regional focused text presents a unique challenge as a geoparsing task.

2.2 Related work

Toponyms are labels to locations and can be realized on a scale from literal, referring to physical location, e.g., proper names or adjectival modifiers, to associative toponyms, which modify non-location concepts, e.g., languages or noun modifiers (Gritta et al., 2020). For example, in "*At 3„30 Am left Amara* (literal toponym) *gave tickets to 24 Amara* (associative toponym) *passengers*", *Amara*

Diary	# Entries	# Tokens	# Vocabulary	# Toponyms	# Unique toponyms
47	273	30979	4430	1671	154
48	210	28321	4195	1485	146

Table 1: Corpus description. Tokens in pre-processed text, unique tokens in vocabulary, toponym instances, and unique toponym instances. There are 208 unique toponyms overall in diaries 47 and 48.

functions as both the literal physical location and is associated with the noun *passengers*.

Geoparsing consists of two primary tasks. The first task is geotagging (toponym extraction), which is a case of named entity recognition (NER). Spans of characters are identified and classified as locations. In digital humanities research, challenges with NER include historical context, language change, and lack of relevant resources (Ehrmann et al., 2021). These challenges can carry over to the second task of geocoding (toponym resolution), which is disambiguating and linking the toponyms to coordinates (Gritta et al., 2020). Geocoding is a named entity linking (NEL) task, which often uses knowledge bases such as Wikipedia and DBPedia for entity linking (Munnelly et al., 2018) and gazetteers such as GeoNames (Wick, 2024).

Toponym resolution approaches can be grouped into three main categories: map-based, knowledge-based, and data-driven or supervised (Buscaldi, 2011). Map-based approaches use external resources such as gazetteers for coordinate data. Previous work includes using co-occurring toponyms in the paragraph and building a weighted map of toponyms in the document to resolve the toponyms (Smith and Crane, 2001).

Knowledge-based approaches incorporate heuristics and hierarchical relationships between toponyms using external resources such as Wikipedia. Aldana-Bobadilla et al. (2020) uses the hierarchy of the toponyms' administrative levels to infer a set of rules to disambiguate each toponym.

Data-driven or supervised approaches rely on machine learning methods, which can be categorized into generate-and-rank systems, vector-space systems, and tile-classification systems (Zhang and Bethard, 2023). Features pertaining to entities include population, geospatial area, geographic entities in common between the candidate and target toponym (Santos et al., 2015), and semantic features, such as historical context (Ardanuy and Sporleder, 2017). Generate-and-rank systems employ a method to rank the candidates, such as a nearest neighbor search leveraging min-hash signatures (Santos et al., 2015), or a neural network with dropout that scores candidates (Halterman, 2023).

Our small historical corpus poses challenges in data availability, both in limited annotated data and insufficient data from knowledge bases. Machine learning and deep learning models require large amounts of annotated data for training, so these types of methods are not always best suited for such corpora. Language change means that locations in historical texts may be spelled differently from those in knowledge bases and be difficult to retrieve. Previous generate-and-rank approaches perform a direct lookup of the toponyms in gazetteers (Alex et al., 2019) or identified textual patterns with parentheses such as "*United States (US)*" (Santos et al., 2015) for alternate names.

Furthermore, gazetteers and knowledge bases may be insufficient and may not contain data for finer-grained toponyms. There is limited research in how to assign coordinates to toponyms that cannot be linked to entries in a gazetteer. Moncla et al. (2014) annotate spatial relations in a corpus of hike descriptions and apply a clustering algorithm, finding collections of spatial points that belong to the same trail and manually resolving the unknown toponyms not in gazetteers using geographic areas of co-occurring toponyms. Moncla et al. (2019) use network analysis to identify neighbors and relations between toponyms. A limitation of this method is reliance on the headwords of the news articles in their data, which does not generalize to other types of historical documents. We address this challenge in our method by performing data augmentation of possible alternate names for toponyms.

3 Methods

We develop a map-based generate-and-rank approach for geoparsing in a small historical corpus of diary entries from Ottoman Iraq written primarily in English but including transliterated Arabic locations. We incorporate a data augmentation step in the pipeline, involving generation of candidate names that could increase potential matches against a knowledge base. Additionally, as most of the toponyms in the corpus are limited to a particular geographical region, we employ a clustering approach, using context words to prefer likely spatial regions for the locations. The small size and restricted geographic scope of the corpus pose severe limitations in terms of suitable training data. As deep learning methods require large amounts of labelled coordinate data or are trained on more global and generalized corpora, we did not find them to be effective for retrieving or inferring viable coordinates for toponyms.[2]

[2]We also conducted an initial exploration of geoparsing with prompting, but found that it could not infer coordinates for smaller locations. This is elaborated upon in Appendix A.

3.1 Annotations

Two annotators created a gold list of annotations for the geotagging task by marking the location entities in diaries 47 and 48 using the brat annotation tool (Stenetorp et al., 2012). We measured inter-annotator agreement by f-score on identical spans for the annotations. We report f-scores of 0.91 for diary 47 and 0.94 for diary 48. The full list of annotation guidelines is in Appendix B. The following are examples of the toponyms annotated:

- Geographical features: "Temreh reach".
- Locations in the context of the postal system: "Posted via Damascus".
- Locations as adjectives: "Amara passengers".

For the geocoding task, one researcher identified the gold standard coordinates for the locations and was assisted by other research team members in identifying the locations' coordinates from various sources, including Lorimer's Gazetteer of the Persian Gulf (Lorimer, 1915), a map of Lower Mesopotamia (East India Company, 1919), a map of the Middle East (East India Company, 1924), and Google Maps. The locations were verified by a historian who is an expert on the region and time period and is an author on this paper.

3.2 Geotagging

We train a model using spaCy for the geotagging task (Montani et al., 2023). We use Fields et al.'s (2023) human-in-the-loop named entity recognition and named entity linking pipeline developed on the Svoboda Diaries corpus. We adapt the pipeline by training the model with location annotations and updating the coreference resolution rules to resolve different spellings of the same toponym to the same entity. Since the spellings of words may be inconsistent for transliterated Arabic names, and the diary author may not consistently spell words the same way, it is valuable to link entities having the same referent. We use the output with the resolved coreferences of the geotagging task as input for the geocoding task.

3.3 Geocoding

Our map-based generate-and-rank approach involves: 1) a novel incorporation of data augmentation in generating candidates for each toponym to query against the knowledge bases, and 2) ranking the candidates to find the most likely candidate for the toponym (Figure 2). We use GeoNames (Wick, 2024) and Wikidata as the knowledge bases.

3.3.1 Candidate Generation

A main challenge of a small historical corpus is that location names will be different from those in modern-day gazetteers. Data augmentation, including augmenting words with similar morphology, is an approach used in low-resource situations for natural language processing (Hedderich et al., 2021). As one toponym may be referred to by different names, we generate alternate names for each entity and use the names to retrieve candidate toponyms from the knowledge base. We consider the constructions in Table 2 to generate alternate names.

Long names break up the toponym name, since not every part of the name may be used in present-day gazetteers. **Close names** account for spelling variation or transcription errors. **Consecutive names** combine adjacent entities. As Svoboda includes excerpts from postal mail and telegrams in his writing, it is possible that two adjacent entities may be grouped together to form an alternate name. For example, *Berggasse* is a neighborhood in *Vi-*

Construction	Description and Example
Long	All ngrams of entities with a large (3) number of tokens, **e.g.**, `Um El Aroog` → `Um, El, Aroog, Um El, El Aroog`
Close	Similar (more than 70%) names using Ratcliff and Obershelp algorithm (Ratclif and Metzener, 1988), **e.g.**, `Shetra` → `Shatra`
Consecutive	Combining adjacent entities, **e.g.**, `Berggasse, Vienna` → `Berggasse Vienna`
Translated	Generated by back-translation through Arabic, **e.g.**, `Basreh` → `Basra`
EngNORM	Generated using rules adapted from EngNORM algorithm for Arabic name variants (Nwesri and Shinbir, 2009), **e.g.**, `Gorna` → `Gornah, Jorna, Gurna`

Table 2: Description of alternate name constructions with examples.

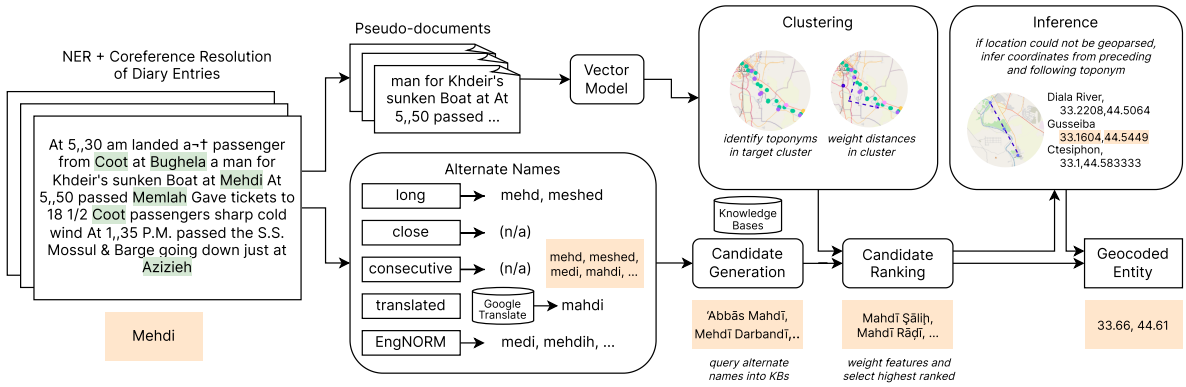

Figure 2: Geocoding toponyms through a generate-and-rank approach. In orange is an example of the outputs at each step for the toponym "Mehdi" in this passage from diary 48.

enna, which may be too fine-grained to geocode on its own, so grouping it with *Vienna* can help retrieve a location that is close to its true position. **Translated** and **EngNORM names** account for linguistic differences. Since Svoboda uses a non-standard Arabic Romanization convention, we generate possible alternate Romanizations of the name so that it may match with names in a resource.

We use these alternate names to query GeoNames and Wikidata, selecting the top 10 results of each query as candidates for the target toponym. If multiple locations exist for the entity, we extract the most recent coordinate entry.

3.3.2 Candidate Ranking

We present **Cluster+Rank** (see Algorithm 1), a method to rank candidates by first clustering word vectors, and then ranking them with the features in Section 3.3.3 to select the best candidate. We adapt Moncla et al.'s (2014)'s clustering and network analysis approach for location referent ambiguity, modifying it to select the best toponym by distance.

Algorithm 1 Cluster+Rank

```
t ← target toponym
C ← clusters of toponym context vectors
C_t ← cluster C_t ∈ C such that t ∈ C_t
T ← generated candidates for t from KB
if T ≠ ∅ then
    for each candidate t′ in T do
        compare distance for t′ with every c ∈ C
        record minimum distance as a feature
    T ← linear ranking of features
    return highest-ranked in T
else
    return inferred from context toponyms
```

First, we develop pseudo-documents based on the context words for each location, similar to that as in Molina-Villegas et al. (2021). The corpus is tokenized using the Penn Treebank tokenizer and pre-processed to remove English stop words and punctuation using the NLTK library (Bird et al., 2009). The pseudo-documents are created by collecting all context words in a word window of size 20 around each target entity. The word window size was set by a search over 10, 15, 20, 25, and 30. We then encode the toponyms as vectors by creating a vector using the set union of the one-hot encoding vectors of the context words in the pseudo-documents. Initially, we also experimented with training a Doc2Vec model (Řehůřek and Sojka, 2010) to generate word embeddings for each pseudo-document that represent the location, but the vector method performed better empirically.

We use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) to perform clustering, identifying the cluster that the target toponym belongs in. The textual and geographical features are computed for each entity-candidate pair and weighted. The weights are selected to maximize accuracy, and a linear ranking method is used, ranking candidates of minimal distance to the toponyms in that cluster higher.

If there are no viable candidates, we infer the coordinates of the unknown location. We use the immediately preceding and following toponyms to infer a line on the Earth's surface, and then interpolate the position of the unknown location.

For each toponym t, the runtime of the system scales linearly with the amount of generated candidates in the set T for the toponym t. Although we implemented the system sequentially, the method

is also parallelizable since we only record the minimum distance as a feature. The system does not require GPUs or other substantial computing power.

3.3.3 Features Used in Ranking Candidates

We use text similarity-based and geographic features for each entity to rank potential candidates, similar to the categories used by Ardanuy and Sporleder (2017) and Santos et al. (2015).

Textual features include text similarity and similarity in phonetic encoding. The text similarity is computed using the Gestalt pattern matching algorithm developed by Ratcliff and Obershelp (Ratclif and Metzener, 1988). The phonetic encoding uses the double metaphone phonetic encoding developed by Philips (2000) and is computed for each entity and candidate independently. The double metaphone encoding attempts to account for more phonetic variation in foreign languages, which is preferable compared to phonetic encoding algorithms designed for American English names or other Western European languages. Svoboda travels along the Tigris River, so many of the locations mentioned in the diaries are in Romanized Arabic that are not typical in most English lexicons.

Geographic features include latitude, longitude, and population. The distance from the Tigris River is computed as the cross-track distance (Chris Veness, 2022), the distance of the geographical coordinate point from the line segment with the source and mouth coordinates of the Tigris River as endpoints. Additionally, considering the distribution of locations in Svoboda's diaries, we consider the location type of the candidates generated. This is coded as "feature code" in GeoNames, and distinguishes between regions, countries, and continents, among others (Wick, 2024). Svoboda mostly travels near the Tigris River, but may occasionally mention larger and faraway locations, such as America. As such, we prioritize continents and administrative regions by limiting our candidates to these types of locations if such a type exists among the candidates generated for a toponym.

3.4 Experimental Setup

We compare our approach with end-to-end geoparsers **CLAVIN** (Cartographic Location And Vicinity INdexer)[3] (Greenbacker, 2021), a heuristics-based geoparser employing fuzzy search and document context, and the **Edinburgh Geoparser**,[4] a heuristics-based geoparser designed for adaptation to historical collections (Grover et al., 2010). These systems were evaluated in other geoparsing works (Gritta et al., 2018; Halterman, 2023). In initial experiments, we evaluated CLAVIN and the Edinburgh Geoparser as end-to-end geoparsers, but they both did not perform well in the geotagging step. Our approach had substantially higher fine-grained accuracy compared to the two systems as a result of better geotagging performance, which made geocoding evaluation difficult.[5] As both systems had used the output of geotagging as the input to their respective geocoding component, we adapted both systems to use our geotagging output and subsequently compare the geocoding approaches in isolation. The adaptations of both systems are elaborated upon in Appendix E.

We compare against additional geocoders: **Nominatim**,[6] an out-of-the-box geocoder to search OpenStreetMap by name (Nominatim, 2023); **Random**, which randomly selects an entity from the candidates generated; **Population**, which chooses the one with the highest population from the candidates generated; and **Cluster+Rank**. For Cluster+Rank, the models are trained on diary 47 and evaluated on diary 48.

3.5 Evaluation

3.5.1 Geotagging

For the geotagging task, the results are evaluated using standard evaluation metrics, precision, recall, and F-score (Gritta et al., 2020). Precision (P) measures the accuracy of the model's predictions, which is the ratio of true positives to all predictions. Recall (R) measures the ratio of true positive predictions to all true positives in the dataset. F-score (F) is the harmonic mean of precision and recall.

Since the evaluation metrics are calculated based on all instances of toponyms, we also count the total instances (**#T**) and unique toponyms (**#U**) geotagged.

3.5.2 Geocoding

For the geocoding task, we evaluate the output coordinates of the pipeline using several metrics. Some considerations for the selection and interpretation of coordinate-based metrics include the distribution of locations and how outliers impact

[3] https://github.com/bigconnect/clavin

[4] https://www.ltg.ed.ac.uk/software/geoparser/

[5] These initial end-to-end experiments are documented in Appendix D.

[6] https://nominatim.org/

the distribution. This corpus has limited geographical scope compared to many geocoding tasks using news article corpora (Ehrmann et al., 2021), so the use of an accuracy metric with stricter tolerance is necessary. Svoboda frequents many cities near each other along the Tigris River; thus, we report accuracy at two different distances. As accuracy metrics treat errors as equally problematic (Gritta et al., 2020), mean distance error is necessary in addition to accuracy.

- **Accuracy@10km (A10)**: ratio of the number of correctly geocoded locations to the total number of locations predicted, within 10 kilometers of the true location.
- **Accuracy@161km (A161)**: ratio of the number of correctly geocoded locations to the total number of predicted locations, within 161 kilometers of the true location (Gritta et al., 2018; DeLozier et al., 2015).
- **Mean distance error (MDE)**: mean of the distances between the true and predicted coordinate locations (DeLozier et al., 2015).

If the system fails to identify coordinates for a location, it is counted as an incorrect prediction and excluded from the MDE calculation.

Lastly, as our primary objective is the identification of coordinates for locations, we consider the number of correct locations within 10km (**C10**), and the number correct within 161km (**C161**), and **# Geocoded**, the number of entities to which coordinates are assigned.

3.5.3 Candidate Generation

We calculate metrics from the candidate generation step to better understand the successes and limitations of the candidate generation step. This approach bears similarities to Heino et al. (2017). We count the number of toponyms in the gold list with a known location (**# Known**) and how many of these toponyms have coordinates in the knowledge base of GeoNames and Wikidata within 10 kilometers of the known position (**# in KB**), which indicates the potential for the location to be identified in the knowledge base. Among the candidates generated for each toponym, we count the number of toponyms for which the system retrieves candidates that have coordinates within 10 kilometers of the known position (**# in Generation**).

Geoparser	Geotagging				
	P	R	F	#T	#U
CLAVIN	0.70	0.17	0.27	366	28
Edinburgh	0.54	0.25	0.34	688	58
Fields et al. (2023)	**0.93**	**0.93**	**0.93**	**1489**	**101**

Table 3: Geotagging results for diary 48, as described in Section 3.5.1.

4 Results

4.1 Geotagging

There are 154 and 146 unique locations in diaries 47 and 48, respectively (Table 1). Unknown locations in each diary are excluded from evaluation [unknown(47)=7, unknown(48)=14].

We revised the pipeline in Fields et al. (2023) to perform geotagging and coreference resolution of locations. We used LEA from the CoVal package (Moosavi and Strube, 2016) to evaluate coreference relations, and report high precision (0.97), recall (0.87), and f-score (0.92). Additional coreference resolution metrics are reported in Appendix C.

We evaluated CLAVIN and the Edinburgh Geoparser as end-to-end systems and found that they suffered in geotagging. In Table 3, CLAVIN and the Edinburgh Geoparser exhibit low recall of the locations at 0.17 and 0.25, respectively. Our system identified more than twice as many of the toponym instances (#T) as the Edinburgh Geoparser.

4.2 Geocoding

4.2.1 System Comparison

We use our human-in-the-loop NER pipeline output and report the geocoding accuracies in Table 4. Of all the systems compared, Cluster+Rank exhibited the strongest performance. It has the highest accuracy, geocodes the most locations (# Geocoded), and gets the most correct locations (C10, C161) overall. The MDE is also substantially lower than the other systems. As all three of our methods outperform CLAVIN, Edinburgh, and Nominatim which are heuristics- and gazetteer-based methods, the performance of our method illustrates the value of candidate generation in geoparsing.

4.2.2 Candidate Generation

In Table 5, we explore the contributions of the candidate generation step. The difference in the number known (# Known) and the number in the

Geoparser	Geocoding					
	A10	A161	MDE	C10	C161	# Geocoded
CLAVIN	0	0.04	1578	0	5	6
Edinburgh	0.17	0.22	1852	22	30	54
Nominatim	0.31	0.47	2351	18	27	58
Population	0.31	0.51	1816	25	41	81
Random	0.19	0.33	2880	15	27	81
Cluster+Rank	**0.41**	**0.56**	**945**	**39**	**53**	**95**

Table 4: Geoparsing results for diary 48, in terms of the metrics and the ratio of unique toponyms successfully geocoded described in Section 3.5.2. The methods use our geotagger combined with different geocoding approaches.

Diary	# Known	# in KB	# in Generation
47	141	110	76
48	132	88	50

Table 5: Candidate generation, from Section 3.5.3.

knowledge base (# in KB) depicts the limitation of the knowledge base. Our approach attempts to close the difference between # in KB and # in Generation. Ideally, # in Generation should equal # in KB, showing that the correct candidate is always in the set of possible candidates. The difference indicates that there are cases in which a location close to the target entity exists in the knowledge base but is not retrieved. For example, *Gherrara*'s known location is at latitude and longitude 33.30, 44.47, but the closest entry in the knowledge base *Ar Rustamīyah* at 33.28, 44.52 is not retrieved.

Our alternate name generation step directly contributes to successfully geocoding toponyms, as shown by the number of candidates generated by the alternate names that were correct predictions (C10 and C161 in Table 6). Considering the number correct (C10 and C161 in Table 4), these names contribute substantially to the performance of the system. The translation and EngNORM constructions, which aim to generate various transliterated Arabic names, generate the largest number of candidates and directly increase the number of correct predictions, demonstrating the potential of these alternate name generation methods for geoparsing pipelines.

The translated construction may be more helpful than the EngNORM construction because EngNORM does not necessarily construct words that exist in lexicons, whereas translation libraries such as Google Translate are trained using large amounts of examples (Isaac Caswell and Bowen Liang, 2020), and so output more commonly used variants of names that match those in knowledge bases.

Construction	N	Count	C10	C161
Long	27	10	0	0
Close	43	13	0	1
Consecutive	4	0	0	0
Translated	177	600	7	10
EngNORM	870	910	4	5

Table 6: Relative contributions of alternate name constructions (from Table 2) on diary 48. N is the number of alternate names created, 'count' is the number of candidates retrieved using the alternate names (exclusive from all original spans), C10 and C161 is the number of predictions correct within 10km and 161km from querying an alternate name.

Another challenge is when locations in the knowledge base are morphologically similar but unrelated to the target entity. For example, Hai, near the Haî river in Iraq, elicited *Shanghai*, *Haiphong*, and *Haikou* as candidates. All include "hai" in the spelling but transcribe different phonemes from different languages. These misleading candidates make it difficult to geocode the location.

5 Conclusion

We present an approach to address the challenge of geoparsing on a small historical corpus. Our approach combines named entity recognition and coreference resolution to identify location entities, then augments candidate name generation using multiple methods, and lastly, employs a map-based cluster-and-rank approach to identify appropriate geographic coordinates. Compared to existing systems, our approach substantively increases viable

candidate locations generated, and in doing so, facilitates the identification of appropriate coordinates for finer-grained entities.

Aside from the substantive performance increase over existing methods, our approach holds some promise in terms of translation to other contexts. As society becomes increasingly globalized, with more communication between people of different cultures, more foreign words are included in text and exchanged. Language also appears to become more standardized, but minority variants of language remain important to include in natural language processing tasks.

The geocoding approach that we used can be leveraged in the context of other research involving identification of non-standard spellings of locations, particularly in situations where data is limited. In addition, the methods we employed in candidate name generation can make working with texts including words of foreign origin easier, which is particularly important in an increasingly multilingual society.

Limitations & Future Work

A possible challenge of this method is that it requires tailoring to the corpus and requires additional overhead, such as manual parameter tuning, to scale to other corpora. However, the method still has great potential to generalize to other contexts, particularly in low-resource, multilingual text data settings, or texts that incorporate foreign words.

While our approach employed manual review, our procedure first involved multiple team members identifying coordinates for locations, which were then subsequently verified by a historian. We found this approach to be manageable given the limited size of our corpus (two diaries). Applying a similar approach to other multilingual datasets could also be scalable. In the future, we can also explore semi-supervised learning approaches that can further reduce manual involvement.

Future work can continue enhancing candidate generation, such as including the use of neural language models, using phonological data, and modeling the orthographic features of a text to generate more spellings. In texts such as diaries that involve highly individual writing patterns, machine learning methods may help to learn these patterns, though the trade-off in training such a model must also be considered.

In addition, our findings demonstrate that systems can leverage context to increase geoparsing performance but are still limited by gazetteer knowledge. One possible approach might be to leverage resources which do not include coordinates (e.g., Lorimer's Gazetteer) to derive additional candidate names.

Ethics

The corpus we employ in this project is publicly available on the Svoboda Diaries Project website.[7] We share our gold NER annotations and coordinate labels as well.[8]

Two volunteer annotators created the gold list of annotations for the geotagging task: one of them is an author on this paper, and the other is working on a different project with the diaries.

We use GeoNames, which is licensed CC-BY 4.0 and Wikidata, which is licensed CC0-1.0. We follow the Wikidata API[9] etiquette by querying sequentially and limiting the number of requests necessary for generating candidates. All queries on GeoNames and Wikidata were made on the free option.

Acknowledgements

Thank you to the anonymous reviewers for providing helpful feedback! We would like to thank the members of the Svoboda Diaries Project for their support and assistance in proofing the location coordinates and providing general feedback, and Rachel Hu in particular for her work with the NER of locations in the corpus. This research was supported by the Mary Gates Research scholarship.

References

Edwin Aldana-Bobadilla, Alejandro Molina-Villegas, Ivan Lopez-Arevalo, Shanel Reyes-Palacios, Victor Muñiz-Sanchez, and Jean Arreola-Trapala. 2020. Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. *Remote Sensing*, 12(18):3041.

Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. 2015. Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, 9(1):15–35.

[7] https://www.svobodadiariesproject.org/svoboda-diaries-data/

[8] https://github.com/svobodadiaries/SvobodaGeoparsing

[9] https://www.wikidata.org/wiki/Wikidata:Data_access

Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander. 2019. Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation*, 53(4):651–675.

Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, pages 175–180, New York, NY, USA. Association for Computing Machinery.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Chris Veness. 2022. Calculate distance and bearing between two Latitude/Longitude points using haversine formula in JavaScript.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. ISSN: 2374-3468, 2159-5399 Issue: 1 Journal Abbreviation: AAAI.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a Novel Geolocation Corpus from Historical Texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198, Berlin, Germany. Association for Computational Linguistics.

East India Company. 1919. 'Lower Mesopotamia between Baghdad and the Persian Gulf' [55r].

East India Company. 1924. 'THE MIDDLE EAST.' [3v].

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. ArXiv:2109.11406 [cs].

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231, Portland, Oregon. AAAI Press.

Sam Fields, Camille Lyans Cole, Catherine Oei, and Annie T Chen. 2023. Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman–Iraqi personal diaries. *Digital Scholarship in the Humanities*, 38(1):66–86.

Rosa Filgueira, Claire Grover, Melissa Terras, and Beatrice Alex. 2020. Geoparsing the historical Gazetteers of Scotland: accurately computing location in mass digitised texts. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 24–30, Marseille, France. European Language Ressources Association.

Charlie Greenbacker. 2021. BigConnect CLAVIN. Original-date: 2019-12-17T08:53:55Z.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which Melbourne? Augmenting Geocoding with Maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54(3):683–712.

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

Andrew Halterman. 2023. Mordecai 3: A Neural Geoparser and Event Geocoder. ArXiv:2303.13675 [cs].

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. 2017. Named Entity Linking in a Complex Domain: Case Second World War History. In *Language, Data, and Knowledge*, Lecture Notes in Computer Science, pages 120–133, Cham. Springer International Publishing.

Isaac Caswell and Bowen Liang. 2020. Recent Advances in Google Translate.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity Linking for Historical Documents: Challenges and Solutions. In *Digital Libraries at Times of Massive Societal Transition*, Lecture Notes in Computer Science, pages 215–231, Cham. Springer International Publishing.

John Gordon Lorimer. 1915. *'Gazetteer of the Persian Gulf. Vol I. Historical. Part IA & IB. J G Lorimer.*

1915'. British Library: India Office Records and Private Papers.

Alejandro Molina-Villegas, Victor Muñiz-Sanchez, Jean Arreola-Trapala, and Filomeno Alcántara. 2021. Geographic Named Entity Recognition and Disambiguation in Mexican News using word embeddings. *Expert Systems with Applications*, 176:114855.

Ludovic Moncla, Katherine McDonough, Denis Vigier, Thierry Joliveau, and Alice Brenon. 2019. Toponym disambiguation in historical documents using network analysis of qualitative relationships. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities '19, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 183–192, New York, NY, USA. Association for Computing Machinery.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Nafise Sadat Moosavi and Michael Strube. 2016. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Gary Munnelly, Harshvardhan J. Pandit, and Séamus Lawless. 2018. Exploring Linked Data for the Automatic Enrichment of Historical Archives. In *The Semantic Web: ESWC 2018 Satellite Events*, Lecture Notes in Computer Science, pages 423–433, Cham. Springer International Publishing.

Nominatim. 2023. Nominatim.

Abdusalam F. Ahmad Nwesri and Nabila Al-Mabrouk S. Shinbir. 2009. Capturing Variants of Transliterated Arabic Names in English Text.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18(6):38–43.

John W. Ratclif and David E. Metzener. 1988. Pattern Matching: the Gestalt Approach.

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. A deeply annotated testbed for geographical text analysis: The Corpus of Lake District Writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities '17, pages 9–15, New York, NY, USA. Association for Computing Machinery.

C.J. Rupp, Paul Rayson, Ian Gregory, Andrew Hardie, Amelia Joulain, and Daniel Hartmann. 2014. Dealing with heterogeneous big data when geoparsing historical corpora. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 80–83.

João Santos, Ivo Anastácio, and Bruno Martins. 2015. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3):375–392.

David A. Smith and Gregory R. Crane. 2001. Disambiguating geographic names in a historical digital library. In *European conference on research and advanced technology for digital libraries*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Svoboda Diaries Project. 2024. About the Diaries.

Marc Wick. 2024. GeoNames.

Zeyu Zhang and Steven Bethard. 2023. 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)*, pages 48–60, Toronto, Canada. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, pages 45–50, Valletta, Malta. ELRA.

A Prompting Exploration

We experimented with prompting approaches on OpenAI's ChatGPT (GPT-3.5) for the geoparsing task. We selected diary 47's day 212 because of the nuance in the passage with how locations are entailed and positioned relative to each other.

We provide the model with the following context about the corpus:

> `The passage is from the Svoboda Diaries, where the author, Joseph Svoboda, is a British steamship purser who frequently travels along the Tigris River. Joseph travels between Basra and Baghdad.`

We provided the following prompt, and then pasted the diary entry text below.

```
Tell me the coordinates of the
locations in this passage:
```

A.1 Observations

We document the following qualitative observations on the results.

Geotagging. The model successfully identifies the relevant locations in the passage most of the time, even if it is not able to identify coordinates for the location. In the case for diary 47's day 212, it identified all the locations except for "Basreh".

Alternate names. For the locations geocoded by the model, it links the locations to existing entities, providing appropriate modern-day spellings such as "Kut" for "Coot" and "Amarah" for "Amara".

Geocoding. The model provides coordinates for locations when possible, but for many locations such as "Azair", "Ali Gherbi", and "Aboo Sedra" (Abu Sidra), the model responds that more information is necessary to determine the specific location.

Inference. We examine locations for which we were unable to identify coordinates for, and refer to them as unknown locations. The model does not attempt to infer coordinates for the unknown locations, though it sometimes describes the location relative to other places in the text. Even with additional prompting, the model will not provide possible coordinates of any kind. A possible explanation for the lack of attempt at location inference is that the model has guardrails in its prompt to avoid responding with inaccurate or hallucinated information.

With regard to the model's description of an unknown location relative to other locations, it tends to describe the location relative to known geocoded locations, and does not keep track of previous information in the passage. For example, in diary 47's day 212 entry:

```
3rd Frid 1898 June At 4 AM.
We left Coot, took 10 1/2
Passengers↩
The Khalifah had 215,000 Okes
& over 220 passengers (127 1/2
Return Jews from Azair all that
remained there~The SS. Ressafah
had just left Coot last night
bound up when we got there↩
N. Erly wind blowing fresh At
10„30 Am passed Ali Gherbi~Henry
writes to me that Mr. Gladstone
the Ex Premier & Minister for
foreign Affairs has died on
the 19th of May, At 5„30 P.M.
arrived at Amara landed 21 1/2
passengers & 48 packages We began
to Ship Pressed Bales of Wool
from Lynch's wool Press, Finished
shipping of the wool of 274 Bales
all for Basreh to Asfar & Kassim
Khdery~At 9 P.M. we left Amara,
Light N.W. & fine Cool weather.
I sleep still in my cabin  At
10„ We dropped Anchor above Aboo
Sedra~
```

We had previously input the entry from day 19, where Azair is between Elbow and Gorna. The system infers that Azair is between Coot and Amara instead, although the prepositional phrase modifies "Jews" and is not described relative to Coot or Amara. This behavior may also demonstrate some of the challenges prompting approaches have with compositional tasks, such as deducing the position of particular locations given multiple facts of the unknown location relative to others.

B Annotation Guidelines

Two annotators created the gold standard for geotagging in diaries 47 and 48 following a set of agreed-upon guidelines. In general, we capture the longest possible annotation that refers to an individual entity.

1. Annotate locations as geographical features, including both the natural and the built environment: "Diala river", "Ctesiphon", "Khalenberg hill".

2. Annotate locations including those in the context of the postal system: "posted Via Bombay", "Persian Gulf Post Offices", "Damascus Mail".

3. Annotate locations in the content of the letters in French in Svoboda's entries: "preparez depart Vienne", "Telegraphiez Consul Baghdad", "from Alexandre Paris".

4. Annotate locations as adjectives when they are complements of a noun in a noun phrase: "Amara passengers", "the Wali of Basreh", "the Emperor of Germany".

5. Annotate abbreviations of locations: "78 Constple Oke".

6. Some cities share names with the ships. Do not annotate the ship names: "S.S. Baghdad, S.S. Koordistan, S.S. Mossul".

Disagreements were resolved after the initial annotations were completed.

C Coreference Resolution Results

We use the CoVal package (Moosavi and Strube, 2016) to measure four coreference resolution metrics and report the results in Table 7.

	P	R	F
MUC	0.99	0.95	0.97
BUC^3	0.97	0.87	0.92
CEAF	0.94	0.77	0.85
LEA	0.97	0.87	0.92

Table 7: Coreference resolution results for diary 48

MUC is based on the minimum number of missing or extra links in the response to the key entities. **BUC^3** considers the fraction of correct mentions included in the response entity. **CEAF** measures the similarity of two entities. **LEA** evaluates coreference relations rather than mentions.

D End to End System Performance

We examine the performance of CLAVIN and the Edinburgh Geoparser as full end-to-end systems. There are significant performance disparities in the geotagging step that make the geocoding evaluation challenging. In Table 8, we report precision, recall, f-score, and the counts for geotagging, and the accuracy metrics for geocoding, with the denominator being the predictions made by the system from the geotagging system.

Regarding the geocoding performance, CLAVIN appears to exhibit the best performance, with the highest accuracies and the lowest MDE. However, since CLAVIN and Edinburgh identified fewer locations overall, the denominators used in the accuracy calculations are lesser than that of Cluster+Rank, which explains why CLAVIN and Edinburgh have higher accuracies compared to our method. CLAVIN and Edinburgh identify fewer locations from the text, and of those locations, the systems are able to geocode them successfully. CLAVIN and Edinburgh generally succeed at geotagging well known place names and other large and modern locations, e.g., Baghdad, Vienna. Our system identifies more locations correctly (C10 and C161) overall, and subsequently assigns coordinates to more entities than the other methods.

CLAVIN exhibits a low recall and high accuracy because there are fewer entities matched. It geocodes the geotagged toponyms that are larger locations including countries, such as Germany, and populous cities, such as Marseille. Edinburgh performs at a more similar level to Cluster+Rank. This may be because of how Edinburgh ranks candidates and selects candidates that tend to be closer to other locations in the text. Our method outperforms both methods in attempting to geocode more finer-grain toponyms, with more correct locations than the Edinburgh Geoparser and with comparable distance accuracy.

Geoparser	Geotagging					Geocoding					
	P	R	F	#T	#U	A10	A161	MDE	C10	C161	# Geocoded
Full											
CLAVIN	0.70	0.17	0.27	366	28	**0.50**	**0.71**	**902**	14	20	28
Edinburgh	0.54	0.25	0.34	688	58	0.43	0.59	949	23	32	54
Combined											
Nominatim	**0.93**	**0.93**	**0.93**	**1489**	**101**	0.31	0.47	2351	18	27	58
Random						0.19	0.33	2880	15	27	81
Population						0.31	0.51	1816	25	41	81
Cluster+Rank						0.41	0.56	945	**39**	**53**	**95**

Table 8: Geoparsing results for diary 48, divided into geotagging and geocoding in terms of the accuracy metrics and the ratio of unique toponyms successfully geocoded. The first section is the full end-to-end geoparsing systems, while the second is our geotagger combined with different geocoding approaches.

E CLAVIN and Edinburgh Adaptation

As listed in Table 1, diary 47 consists of 273 text files and diary 48 consists of 210 text files. Each diary has one corresponding comma-separated file of the brat NER annotations. To simplify the system comparison, the individual text files are merged into one single file per diary, and the indices of the words in the text are updated accordingly in the comma-separated file.

Both systems have separate modules for the geotagging and geocoding tasks, so we feed in our geotagging output into the geocoding module.

CLAVIN is a heuristics-based geocoder that uses GeoNames as its gazetteer. The system is implemented in Java. Its pipeline uses the geotagging output as the geocoding input. The intermediate data structure is a list of spans, which keeps track of the string span and the position of the span, e.g., "France" at position 10231. The entire diary text is input into the system, and the output from the console is converted into the comma-separated file format.

The Edinburgh Geoparser is a heuristics-based geoparser that uses LT-XML 2 markup on the document for geoparsing. The gazetteer in the system is able to be customized, but currently Geonames is the only gazetteer that can be used as the Unlock gazetteer can no longer be accessed. We convert our list of placenames to the required XML markup format and input it to the geocoder. Note that all the surrounding context non-location words are lost in this process converting the placenames to a list of XML elements. The output in the XML file is then parsed back into the comma-separated file format.