Semantic Similarities between Locations based on Ontology

Moiz Khan Sherwani*[1], Petr Sojka[1], and Francesco Calimeri[2]

[1] Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czechia
mksherwani@mail.muni.cz, ORCID: 0000-0001-6061-6753
sojka@fi.muni.cz, ORCID: 0000-0002-5768-4007
[2] Dept. of Mathematics and Computer Science, University of Calabria, Calabria, Italy
calimeri@mat.unical.it, ORCID: 0000-0002-0866-0834

Abstract. Toponym disambiguation or location names resolution is a critical task in unstructured text, articles or documents. Our research explores how to link ambiguous locations mentioned in documents, news and articles with latitude/longitude coordinates. We designed an evaluation system for toponym disambiguation based on annotated GEO-CLEF data. We implemented a node-based approach taking population into account and a geographic distance-based approach. We have proposed new approach based on edges between the pairs of toponyms in ontology, taking also population attribute into account. Our edge-based approach gave better results than population and distance-based only approaches. The results could be used in any information system dealing with texts containing geographic locations, such as news texts.

Key words: toponym disambiguation, geonames, geographic text retrieval, ontology based geoname relations, toponym similarity

Everything has to do with geography. (Judy Martz)

1 Introduction

Toponym disambiguation or *place name resolution* is a process of assigning location names (toponyms) that appear in article by normalizing them with the help of their respective coordinates and their context of appearance. This process turns out to be quite difficult for locations that are highly ambiguous and in short texts. [14] Gazetteers are the main source for the identification and disambiguation of the location names. Gazetteer serves as the dictionary for the geographical entities, usually with the set of properties (`City`, `Country`, `Continent`, `Coordinate`, `Population`, `Alternative Names`, `Administration Division` etc.) about every location. GeoNames[3] is very well known Gazetteer with all ambiguous location names structured according to the respective classes.

* Reported work has been done during Erasmus+ stay at the NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czechia

[3] `http://www.geonames.org`

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 85–94, 2017.

Toponym disambiguation is regularly performed in two stages. The initial step, *toponym recognition*, discovers all occurrences of names and put them in an archive. There are, for example, over one hundred places in the world with the name Alexandria. The second step, *toponym disambiguation*, assigns latitude and longitude and scope to all names found in the initial step.

Many people prefer to read articles, news and blogs online, so it is important to provide structured, location-based reading sources for the geographic entities. The process of understanding geography from any type of unstructured content, articles or text is called *geoparsing*, *geocoding* or *geotagging* [8,10,9]. Once the geonames are disambiguated, their results could be used for indexing and search, for document similarity computations, for document filtering, infosystems alerting an the like.

The paper is structured as follows: In Section 2 we review related work. In Section 3, our experimental setup to perform the evaluation of the research is discussed. We describe our methodology, evaluation metrics and datasets in Sections 4 and 5. Section 6 serves for reporting the results achieved. We conclude in Section 7 by summarizing our outcomes, and suggest future work.

2 Related Work

The idea of toponym disambiguation is to identify all location names stated in an article and to specify these location names with the coordinates latitude and longitude. In this research, we are not considering the references to some location names, e.g. "1 km south of Brno" or "around the University of Calabria", rather this research focuses on the use of specific location names. Identifying toponyms has been widely studied in *named entity recognition* (NER) research: location names were one of the main classes of named entities to be distinguished in article [12]. Most of the approaches are based on the toponym disambiguation are driven by the physical properties of the toponyms. Some methods rely on external sources [4]. Properties depicting Geo-spatial areas, as well as their relations on the Earth are utilized for disambiguation. This approach is supported by several heuristics as explained by Leidner [7].

One approach utilizes the attributes of the Geo-spatial areas to resolve any uncertainty between toponyms. The significance of an area is frequently computed by having the location with the biggest population. Another approach assumes that an article is likely to refer to places within a constrained geographical zone, so it picks places that are near to each other. For example, if an article contains the ambiguous location names Rome, USA, Texas, this approach will select Rome in USA instead of the Rome in Italy based on the distance between the location names mentioned in the article. [8] Even though great advances in toponym disambiguation have been made this decade, it is still difficult to decide which approach is the best.

A different problem emerges from non-standard practices of the *gazetteers* (Geo-spatial lexicon) when they allocate coordinates to location names. According to [3] Cambridge has just two locations in Wordnet, 38 in Yahoo! Geoplanet,

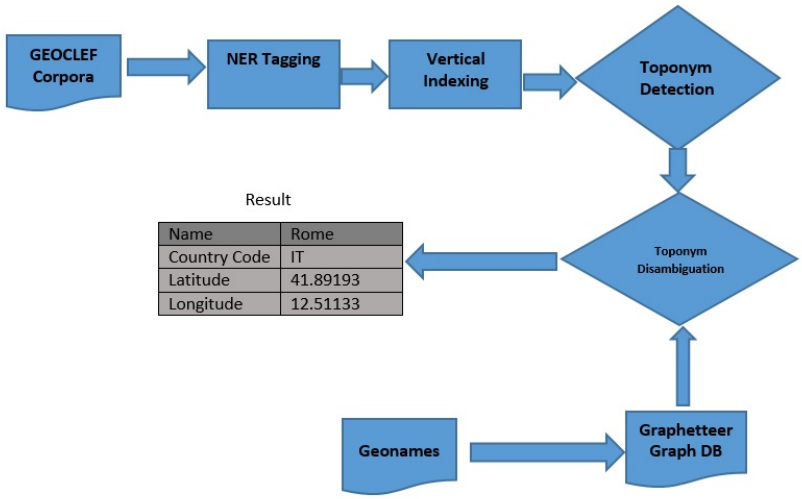

Fig. 1. An overview of our toponym disambiguation work-flow setup

and 40 in GeoNames. The scope may vary extensively from one resource to another, and the latitude and longitude allocated for the same area may fluctuate too, bringing about an unjustified disambiguation when scoring the frameworks.

Another problem arises with the use of different corpora, because the variations in the corpora that are used for the evaluation make it complicated to find the best approach [13]. A new approach that is said to have higher precision considering one corpora will have variations if used on the other corpora.

3 Experiment Setup

Successful disambiguation will be hugely beneficial to systems that utilize place names, as the names themselves cannot always be used for disambiguating them. Let us consider the toponym Springfield, for instance. Around the world, there are no less than 33 settlements with that name – it is also part of college and University names. At the point when the system is given an article about Springfield College, these lines not straightforwardly clear to which Springfield College it indicates. If the article happens to mention other places, this situation changes. For instance, if Springfield, MA (Springfield, Massachusetts) is specified, the message likely alludes to Springfield in Massachusetts state of USA. In this case, toponym was identified in the interesting way based on the State name given with the toponym. We have designed the system in such a manner that it consists of two phases. In the first phase of our experiment, toponyms are extracted from the articles. The second phase reports the disambiguation of each toponym recognized by the first phase, hence all the toponyms are assigned with their respective coordinates based on the gazetteer.

Toponym Extraction and Entity Look-up The first phase for toponym disambiguation requires the labeling of toponyms in the article. To assign the location labels in the article, any named entity recognizer or gazetteer can be connected. The first step involves matching the extracted toponyms with a candidate location name in the location network. We utilize Stanford's Named Entity Recognition system [5] to both coordinate the toponyms in the article with those in the system, and to sort out the toponyms within the articles. As a result of this step, we receive more than one toponym in each article but there are other possible cases: **i)** no match is found, **ii)** one match found (un-ambiguous location), or **iii)** more than one match is found (ambiguous location).

Toponym Disambiguation As a result of the first stage in toponym recognition, we have list of toponyms for each article. This will be used for the disambiguation of toponyms. Resolving the cases mentioned above would be as follows. In the first case, there is no match for the toponym in the area system and therefore it cannot be connected to any area. In the second case, the area specified is unambiguous, and the assignment of connecting it to the network in the system is clear. In the third case, we find toponym with ambiguity, which can be resolved in a variety of ways. Our system takes the result of a previous phase as an input and assigns coordinates in latitude and longitude. To carry out this disambiguation, we have executed different approaches, two of which proved to be effective in studies proposed by [7]. The first is the population based approach and second approach is based on the distance. Moreover we have developed a novel approach that is called the edge-based approach, because we have used the graph database for the computations and would like to introduce a new approach based on graphs to enhance this procedure.

Node-Based Approach This approach is figured in light of the population property of the GeoNames database. It will continuously pick the location which has the highest population for all ambiguous locations. Therefore Rome in Italy will dependably be favored over Rome in any of 15 states of the USA. One limitation of this approach is the incorrect population data of continents that sometimes appears in the GeoNames database, as well as a few different toponyms. We have made changes to the different states and continents population manually where the value of population was mentioned 0. This approach is mostly taken into consideration: if only one location name is mentioned in the article and there is no other reference to be considered for the ambiguous location.

Geographic Distance-Based Approach This approach is figured based on the shortest distance between the toponyms. The distance between the locations is computed by the *haversine formula* that takes latitude and longitude of the locations as an input and results in the distance between the candidates. All ambiguous toponyms present in the article would be used as separate pairs and

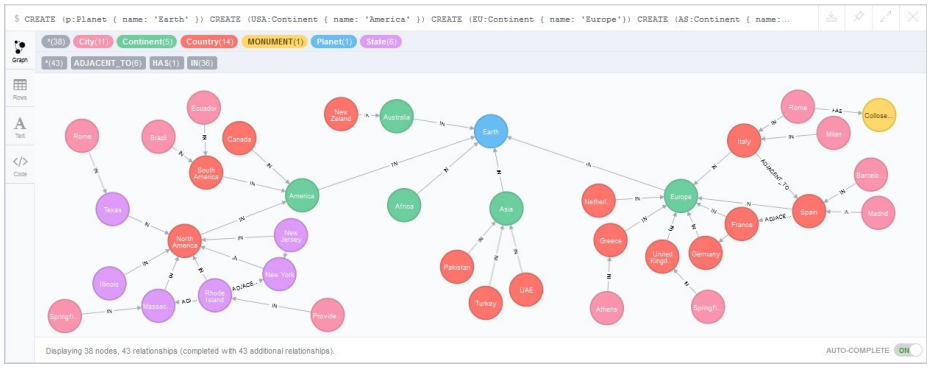

Fig. 2. Sample diagram of graphetteer designed on NEO4

the shortest distance between the toponyms would be selected as the resulting toponym.

Edge-Based Approach We are introducing the edge-based approach for our computation. This approach could only be computed with the graph based databases. It computes the distance based on the number of edges and nodes between the pair of toponyms. But with this approach, we have one constraint. For example, if we have set of toponyms Italy and Rome in the article and we gave Italy and Rome in the Graphetteer as an input to compute the edge-based approach. The resulting output for the edge-based approach might give Italy and Rome in USA based on the smaller number of edges. But our required nodes were to get the Italy and Rome that belongs to Europe. To achieve satisfactory results, we have attached the query of considering the population property of the node to improve the precision.

4 Methodology, Evaluation Metrics

To conduct our experiments, we decided to work on the graph database instead of a traditional relational one.

Graphetteer Previously, researchers have used the gazetteer based on the traditional RDBMS for toponym disambiguation. In this research, the Gazetteer that is used for the location database is taken from GeoNames, US National Geospatial Intelligence Agency and US board on Geographic Names. Graphetteer is the name given to the Gazetteer based on Graph Database, and the conceptual model for Graphetteer was proposed by [2]. Graph database has several new features and algorithms to perform the evaluation. GeoNames database was cleaned based on our research requirements (for example: columns with values of the modification date).

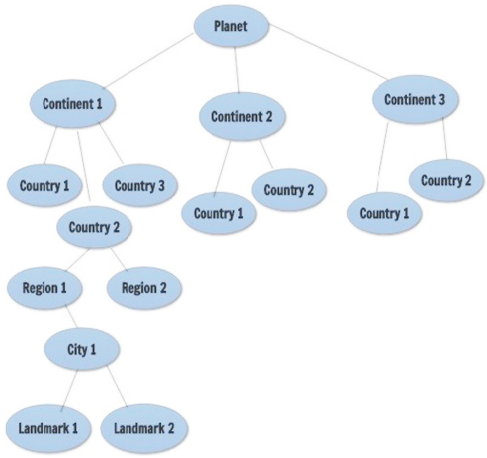

Fig. 3. Sample diagram for our Geographic Ontology

After cleaning the data, it was arranged according to the continents and then NEO4J was used to convert all the database into the Graph database. Cypher Query Language is the main language that is used to work on the NEO4J database. In this database, we have used Continents, Countries, Cities, Regions, Administration Divisions, Counties and Monuments as Nodes. The relationships between them are the edges. Two types of relationship edges were used:

IN edge to show the relationship between towns to cities, cities to countries, countries to continents;

HAS edge to show the famous places, tourist attractions and monuments within a city.

Toponym Recognition An alternative to NLTK's NER classifier is provided by the Stanford NER tagger[4] [5]. This tagger uses an advanced statistical learning algorithm it's more computationally expensive than the option provided by NLTK. It labels sequence of words in an article based on the 7 classes (Location, Organization, Person, Money, Percent, Date and Time). Our requirement was to extract the location names from the article. For this purpose, we have used 3 class model to label the location names in the articles. Our approach consists of the following steps:

1. Run the Stanford NER tagger on the GEOCLEF database. This would label the article based on three classes (`Location`, `Name` and `Organization`). One drawback of the Stanford NER tagger is that it labels United Kingdom as United `/Location` Kingdom `/Location`.
2. Use Python script to merge the multiple named location names that has more than single `/Location` label into single label.

[4] `https://nlp.stanford.edu/software/CRF-NER.shtml`

3. Convert the article into "vertical" or "word-per-line (WPL)" format, as defined at the University of Stuttgart in the 1990s. This method allows us to find the distance between the words in the article.
4. Extract every location name that have `/Location` along with the assigned index.

Toponym Resolution Once we have all the location names listed in the article, We will be using the script to assign the location coordinates, latitude and longitude based on the approaches that we have considered for the evaluation of this research. We created three different scripts to evaluate the approaches based on population, distance and edge-based. All location names files are run through the graphetteer and the resulting locations are achieved with their coordinates for our evaluation.

Metrics We have used typical metrics to evaluate the approaches: Precision, Recall and *F*-measure F_1 to evaluate the performance of our toponym disambiguation experiments:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where *TP* is the count of correctly disambiguated toponyms by the system, *FP* is the count of incorrectly disambiguated toponyms by the system, *FN* is the count of toponyms not identified by the system. Since our geographical database and the annotated corpus is based on the GeoNames DB, the toponyms effectively distinguished are known by a basic match between the place IDs retrieved by the framework and with the annotated corpus.

5 Dataset, Data Preprocessing

Corpus We used the GEOCLEF corpus to carry out our evaluation. This stands for Geographic Cross Language Evaluation Forum. This corpus consists of articles from English and German sources. For evaluation, we have used the corpus Information Retrieval in English that consist of the 169,477 articles from the Glasgow Herald (British) 1995 and LA Times (American) 1994. These articles are associated with (number of toponyms) tokens. This corpus is widely used to conduct the research based on the Geographic Information Retrieval. Further details can be found at GEOCLEF 2008[5] link. Toponym statistics are to be found in Table 1.

Gazetteer For toponym disambiguation, we needed a gazetteer to specify locations for each toponym in the articles. To acquire a gazetteer that secured overall data, we used the GeoNames database for locations. It is a completely

[5] `https://www.uni-hildesheim.de/geoclef/`

Table 1. Numbers of 10 most frequent toponym in the corpus

Toponym	Frequency	Ambiguous
Los Angeles	139,881	Yes
Glasgow	73,402	Yes
United States	65,993	No
Scotland	34,835	Yes
Washington	24,135	Yes
California	23,812	Yes
United Kingdom	16,180	No
New York	12,079	Yes
London	11,790	Yes
England	11,240	Yes

accessible gazetteer containing more than 10 million entries worldwide. Every location entry contains a name, alternative names, administrative level, country codes, latitude/longitude coordinates and elevation. Each location has their respective coordinates and geonames ids that make them unique from every other location entry. We have processed the data according to our needs and all the properties of the data are included in the Graphetteer except for the alternative names with special characters in it.

6 Results

We have compared three approaches: node-based approach, geographic distance-based approach and edge-based approach. Results are summarized in the Table 2. For the node based approach, where we have considered population as the main property for disambiguation. We computed all approaches on 169,450 articles from GEOCLEF and the toponym frequency is also given in the Table 1. There are 1,238,686 toponyms occurrencies in all articles together.

All of the data and code is available for download for reproducibility and comparison of approaches. Our evaluation framework is available on the project web page `https://nlp.fi.muni.cz/projekty/toponym_disambiguation`. Since our graphetteer and the annotated article corpus is based on GeoNames database. Toponyms that are identified as positive candidates are

Table 2. Toponym Disambiguation on GEOCLEF data based on different approaches

Approach	Precision	Recall	F_1
Node-based	0.70	0.89	0.78
Geographic distance-based	0.39	0.89	0.54
Edge-based	0.74	0.89	0.80

referred by the `GeoNames ID` resulted by our system and the experts annotated the corpus.

> GIS is waking up the world to the power of geography, this science of integration, and has the framework for creating a better future. (Jack Dangermond)

7 Conclusion and Future Work

We have compared three approaches to toponym disambiguation. We have proposed a new approach based on the edges between the pairs of toponyms in an ontology, taking a population attribute into account. According to our comparison between the most commonly used heuristics (population and geographic distance), the best results were achieved using the edge-based approach.

Using a graph database is efficient and as new features could be used to compute like centrality measures, it brings new opportunities for further improvements, e.g. matching nodes based on the relationships and their specific properties.

Several toponym disambiguation approaches could be supported by our framework in the future:

vector representations Taking the context in which a toponym was used is the key for a further increase of precision. Vector space word representations and their similarity computed by word2vec [11,6] or similar system is yet another way to be tested in the future.

weighting Experiments with weighting based on the level of Ontology, e.g.: Continents is on the top level followed by Countries and so on and lowest level is considered as the street or landmark in a City. Starting from the top level higher weights and lower weights for the bottom level ontology.

metadata We can also improve the result by using the metadata of article news, and a knowledge base about the location names.

alternate toponym names One can handle the alternative names for the locations with special characters or letters from other languages than English. To disambiguate toponyms with location names in different languages, corpora based on other languages would also be required.

voting Using different approaches to disambiguation to vote on the right toponym disambiguation. Hybrid approaches are giving excellent results [1].

geonames similarity It is important to quantify similarity of geographical names for the purpose of information retrieval, alerting systems and other uses of disambiguated toponyms.

Different approaches will be compared and evaluated on the same data.

Acknowledgments Funding of the TA ČR Omega grant TD03000295 is gratefully acknowledged.

References

1. Badieh Habib Morgan, M., van Keulen, M.: Named entity extraction and disambiguation: The missing link. In: Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval. pp. 37–40. ESAIR '13, ACM, New York, NY, USA (2013), https://doi.org/10.1145/2513204.2513217
2. Bär, M.: Graphetteer – A conceptual model for a graph driven gazetteer (Jan 2016), http://www.geonet.ch/graphetteer-a-conceptual-model-for-a-graph-driven-gazetteer/
3. Buscaldi, D.: Approaches to Disambiguating Toponyms. SIGSPATIAL Special 3(2), 16–19 (2011), https://doi.org/10.1145/2047296.2047300
4. Buscaldi, D., Rosso, P.: Map-based vs. knowledge-based toponym disambiguation. In: Proc. of the 2nd International workshop on Geographic IR. pp. 19–22. ACM, Napa Valley, CA, USA (2008), https://doi.org/10.1145/1460007.1460011
5. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting of ACL. pp. 363–370. ACL '05, ACL, Stroudsburg, PA, USA (2005), https://doi.org/10.3115/1219840.1219885
6. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR abs/1405.4053 (2014), http://arxiv.org/abs/1405.4053
7. Leidner, J.L.: Toponym Resolution in Text (Annotation, Evaluation and Applications of Spatial Grounding). Dissertation Abstract. ACM SIGIR Forum 41(2), 124–126 (2007), https://doi.org/10.1145/1328964.1328989
8. Lieberman, M.D., Samet, H.: Multifaceted Toponym Recognition for Streaming News. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information – SIGIR '11. pp. 843–852 (Jul 2011), https://doi.org/10.1145/2009916.2010029
9. Lieberman, M.D., Samet, H.: Adaptive context features for toponym resolution in streaming news. In: Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval – SIGIR '12. pp. 731–740 (Aug 2012), https://doi.org/10.1145/2348283.2348381
10. Lieberman, M.D., Samet, H.: Supporting Rapid Processing and Interactive Map-Based Exploration of Streaming News. In: International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2012) (Nov 2012), https://doi.org/10.1145/2424321.2424345
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013), http://arxiv.org/abs/1310.4546
12. Piskorski, J., Yangarber, R.: Information Extraction: Past, Present and Future. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization. pp. 23–49. Springer (2013), https://doi.org/10.1007/978-3-642-28569-1_2
13. Roberts, K., Bejan, C.A., Harabagiu, S.: Toponym Disambiguation Using Events. In: Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010). pp. 271–276 (2010), http://www.aaai.org/ocs/index.php/FLAIRS/2010/paper/viewFile/1291/1754
14. Sagcan, M., Karagoz, P.: Toponym Recognition in Social Media for Estimating the Location of Events. In: Proc. of 15th IEEE International Conference on Data Mining Workshop, ICDMW '15. pp. 33–39 (2016), https://doi.org/10.1109/ICDMW.2015.167