Using Generic Ontologies to Infer the Geographic Focus of Text*

Christos Rodosthenous[1][0000−0003−2065−9200] and Loizos Michael[1]

Open University of Cyprus, P.O. Box 12794, Nicosia, Latsia, Cyprus
christos.rodosthenous@ouc.ac.cy, loizos@ouc.ac.cy

Abstract. Certain documents are naturally associated with a country as their geographic focus. Some past work has sought to develop systems that *identify* this focus, under the assumption that the target country is explicitly mentioned in the document. When this assumption is not met, the task becomes one of *inferring* the focus based on the available context provided by the document. Although some existing work has considered this variant of the task, that work typically relies on the use of specialized geographic resources. In this work we seek to demonstrate that this inference task can be tackled by using generic ontologies, like ConceptNet and YAGO, that have been developed independently of the particular task. We describe GeoMantis, our developed system for inferring the geographic focus of a document, and we undertake a comparative evaluation against two freely-available open-source systems. Our results show that GeoMantis performs better than these two systems when the comparison is made on news stories whose target country is either not explicitly mentioned, or has been artificially obscured, in the story text.

Keywords: Information Retrieval, Geographic Focus Identification, Ontologies, Natural Language Processing, Geographic Information Systems

1 Introduction

In this work we tackle the problem of identifying the geographic focus of a text document. Humans are able to read a document and identify its geographic focus [1]. According to Silva et al. [2], “Geographic scope or focus of a document is the region, if it exists, whose readers find it more relevant than average.”. Narratives are examples of such documents, that human readers can identify the location where the story takes place, along with other properties (e.g., the protagonist, the timeline, etc.) [3].

* An earlier version of this work was presented at the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018). In this article, compared to the conference paper we give a more extensive evaluation of the GeoMantis system with different datasets and comparisons with more systems. Furthermore, we updated the GeoMantis system with new knowledge from ontologies and present details on the structure of the ontologies and the way they are used.

For a machine to perform this task, it needs to process the text, identify location mentions from the text, and then try to identify its geographic focus. The majority of systems developed in this line of research rely on gazetteers, atlases, and dictionaries with geographic-related content, that identify the geographic focus of the text. In this work, we investigate whether generic ontologies can be exploited for tackling this problem with a special focus on cases where no explicit mention of the target country exists in the text.

We present **GeoMantis**, a system developed to identify the country-level focus of a text document or a web page using knowledge from generic ontologies. In particular, the system takes as input any type of document, processes it, and it stores the contents of the document in a database. Independently of the previous process, the system retrieves triples from ontologies about countries, processes each triple, filters it using its internal mechanisms, and stores it in a database. In this workflow, a full-text search algorithm is used for matching each search text of the document against the search text of each triple in the country's knowledge base set. A number of filtering options are also available during this process.

The outcome of the above-mentioned search process is the set of country triples that are activated by the document text. This outcome is used in the query answering process to produce a list of countries in order of confidence. The ordering of this list is performed using one of the four supported by the system strategies presented in detail later in this work.

In the following sections, we present the current state in geographic focus identification, along with systems developed to perform this task. Next, the GeoMantis system is presented, followed by a detailed presentation of the generic ontologies employed by it. The penultimate section, presents the results of the parameter selection process and the comparative evaluation of the system. In the final section, new features and possible extensions to the GeoMantis system are discussed as part of our ongoing work.

2 Problem Definition and Related Work

The geographic focus of a document can be defined as the geographic location the document is related to. In this work, we limit this area to locations on earth that have administrative boundaries. For example, the text snippet "*A letter to creditors says Mr Tsipras is prepared to accept most conditions that were on the table before talks collapsed and he called a referendum.*"[1] has a geographic focus in Greece, Europe.

The task of identifying the geographic focus of text goes back to the 90's and the research in this area [4] led to the development of several systems. Many of these systems rely on geoparsers, i.e., systems for extracting places from text [5, 6], for identifying locations, disambiguating them, and finally for identifying the geographic focus of the text. These systems, perform well when documents include place mentions for geoparsers to work, but leave open the

[1] http://www.bbc.com/news/

case of documents that have none or very few place mentions. It is common for a document to also contain references to geographic locations in the form of historical dates, monuments, ethnicity, typical food, traditional dances and others [7]. These references can be used to infer the geographic focus of a text document.

In the 90's, the Geo-referenced Information Processing SYstem **GIPSY** [8] was created. This system was able to perform geocoding on documents related to the region of California. Geocoding was applied using a subset of the US Geological Survey's Geographic Names Information System (GNIS) database. GIPSY's document geocoding pipeline included three steps. First, the system extracts keywords and phrases from each document according to their spatial relatedness. Each of these phrases are weighted according to a heuristic algorithm. Second, the system identifies the spatial locations for the keywords and phrases extracted in the first step using synonyms and hierarchical containment relations. Third, geographic reasoning is applied and after extracting all the possible locations for all the terms and phrases denoting places in a given document, the final step presents the geospatial footprints as a three-dimensional polyhedron.

In the 00's, the **Web-a-Where** system [9] was introduced, which can identify a place name in a document, disambiguate it, and determine its geographic focus. This system detects mentions of places in a document or a webpage and determines the location each place name refers to. Moreover, it assigns a geographic focus to it by using a similar workflow with the GIPSY system and it also has a specific approach for disambiguating locations for both geo/non-geo and geo/geo ambiguity. When a place name has the same name as a non-place (e.g., Turkey the country and Turkey the bird), a geo/non-geo ambiguity is identified. When two or more places have the same name (e.g., Athens in Greece and Athens in the USA), a geo/geo ambiguity is identified. Furthermore, the system can assign a geographic focus to a document, even though its location is not explicitly mentioned in it, but it is inferred from other locations. The Web-a-Where system was evaluated using two different pre-annotated datasets. The authors report that their system detects a geographic focus in 75% of the documents and report a score of 91% accuracy in detecting the correct country.

A more recent attempt is the geo-referencing system developed within the **MyMose project** framework [10]. This system, performs a city-level focus identification using dictionary search and a multistage method for assigning a geographic focus to web pages, using several heuristics for toponym disambiguation and a scoring function for focus determination. The authors report an accuracy of over 70% with a city-level resolution in English and Spanish web pages.

A similar to the Web-a-Where system workflow was used in the **CLIFF-CLAVIN** system [11], which identifies the geographic focus of news stories. This system uses a three step workflow to identify the geographic focus. First, it recognizes toponyms in each story, then, it disambiguates each toponym, and finally, it determines the focus using the "most mentioned toponym" strategy.

This system relies on "CLAVIN"[2], an opensource geoparser that was modified to facilitate the specific needs of news story focus detection. The authors report an accuracy of 90-95% for detecting the geographic focus when tested on various datasets. This system is freely available under an opensource license. It is also integrated in the MediaMeter[3] suite of tools for quantitative text analysis of media coverage.

Related to this line of research, is the work on **SPIRIT** [12], a spatially aware search engine which is capable of accepting spatial queries in the form of <theme><spatial relationship><location>. Relevant research is also found in the work of Yu [13] on how the geographic focus of a named entity can be resolved at a location (e.g. city or country).

Furthermore, work done on a system called **Newstand** [14], monitors RSS feeds from online news sources, retrieves the articles in realtime and then extracts geographic content using a geotagger. These articles are grouped into story clusters and are presented on a map interface, where users can retrieve stories based on both topical significance and geographic region.

More relevant work, mainly concentrated in using knowledge bases extracted from Wikipedia, is presented in work of de Alencar and Davis Jr, and Quercini et al. [15, 16]. de Alencar and Davis Jr, presented a strategy for tagging documents with place names according to the geographical context of their textual content by using a topic indexing technique that considers Wikipedia articles as a controlled vocabulary. Quercini et al., discussed techniques to automatically generate the local lexicon of a location by using the link structure of Wikipedia.

A system called **Newsmap** [17], uses a a semi-supervised machine learning classifier to label news stories without human involvement. Furthermore, the system identifies multi-word names to automatically reduce the ambiguity of the geographical traits. The authors evaluated their system's classification accuracy against 5000 human-created news summaries. Results show that the Newsmap system outperforms the geographical information extraction systems in overall accuracy, but authors report that simple keyword matching suffers from ambiguity of place names in countries with ambiguous place names.

Imani et al. [18], proposed a mechanism that utilizes the named entities for identifying potential sentences containing focus locations and then uses a supervised classification mechanism over sentence embedding to predict the primary focused geographic location. The unavailability of ground truth (i.e., whether words in a sentence is focus or non-focus) suggests a major challenge for training a classifier and an adaptation mechanism is proposed to overcome sampling bias in training data. This mechanism was evaluated against baseline approaches on datasets that contain news articles.

Silva et al. [2], presented a system for automatically identifying the geographic scope of web documents, using an ontology of geographical concepts and a component for extracting geographic information from large collections of web documents. Their approach involves a mechanism for identifying geographic

[2] https://clavin.bericotechnologies.com/

[3] http://mediameter.org/

references over the documents and a graph ranking algorithm for assigning geographic scope. Initial evaluation of the system, suggests that this is a viable approach.

A system called **TEXTOMAP** [19], aims to design the geographic window of the text, based on the notion of important toponyms. Toponym selection is based on spatial, linguistic or semantic indicators.

A relatively new system called **Mordecai** [20], performs full text geoparsing and infers the country focus of each place name in a document. The system's workflow extracts the place names from a piece of text, resolves them to the correct place, and then returns their coordinates and structured geographic information. This system utilizes a number of natural language processing techniques and neural networks to perform these tasks.

3 The GeoMantis System

GeoMantis (from the Greek words Geo that means earth and Mantis, which means oracle or guesser), is a web application designed for identifying the geographic focus of documents and web pages at a country-level.

Users can add a document to the system using a web-interface. The document enters the processing pipeline depicted in Fig. 1 and gets processed.

The system uses factual knowledge in the form of Resource Description Framework (RDF) [21] triples retrieved from ontologies (e.g., ConceptNet and YAGO). These triples are of the form <`Subject`><`Predicate`><`Object`>, where the `Subject` has a relationship `Predicate` with the `Object`. Detailed information on the RDF semantics can be found in the W3C specification document [22]. Triples are stored locally in the system's geographic knowledge database. This database can be updated at any time by querying the corresponding knowledge source online.

Retrieved triples from ontologies are used for searching in each document and generate the predicted geographic focus. Instead of returning only one prediction for the target country, the system returns a list of countries in order of confidence for each prediction. Countries in the first places have a higher confidence score.

The system can be tuned using a number of parameters such as the selected ontology, the query answering strategy (see Section 3.3), and text filtering options (e.g., stopwords and named entities).

In the next paragraphs, we present how the GeoMantis system pipeline works.

3.1 Text Input Parsing

First, users upload a text document or type a webpage URL through a web interface. This text is firstly cleaned from HTML tags (e.g., <br>, <b>, <p>, <div>) and wiki specific format (e.g., [[Link title]]). Then, the text is parsed using a Natural Language Processing (NLP) system, the Stanford CoreNLP [24]; extracted lemmas, part of speech, and named-entity labels extracted by the Named Entities Recognition (NER) process, are stored and indexed in the

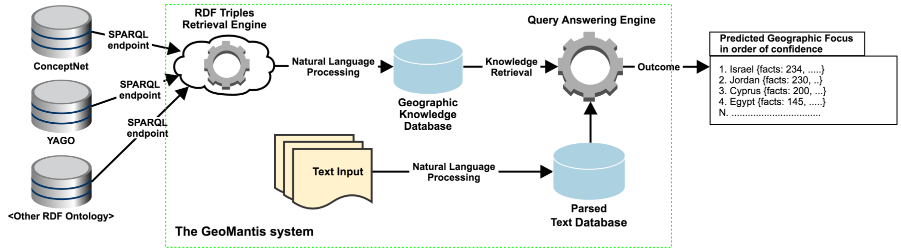

ConceptNet
SPARQL endpoint
YAGO
SPARQL endpoint
<Other RDF Ontology>
SPARQL endpoint
RDF Triples Retrieval Engine
Natural Language Processing
Geographic Knowledge Database
Knowledge Retrieval
Query Answering Engine
Outcome
Predicted Geographic Focus in order of confidence
1. Israel {facts: 234, ...}
2. Jordan {facts: 230, ..}
3. Cyprus {facts: 200, ...}
4. Egypt {facts: 145, ...}
N. ....
Text Input
Natural Language Processing
Parsed Text Database
The GeoMantis system

Fig. 1: The GeoMantis system processing workflow. The workflow includes the RDF Triples Retrieval and Processing Engine (left), the Text Processing mechanism and the Query Answering Engine. The outcome of the system appears on the right. Figure adapted from [23].

system's database. The NER system can identify named entities of type location, person, organization, money, number, percent, date and time, duration, and miscellaneous (misc).

3.2 Knowledge Retrieval

The RDF triple retrieval process starts by identifying each country's official name and alternate names from the GeoNames database[4]. Geonames is a geographical database that includes more than 10 million geographical names. It also contains over 9 million unique features where 2.8 million are populated places and 5.5 million are alternate names. The database is integrating geographical data such as names of places, alternate names in various languages, elevation, population, and others from various sources. Sources include, among others, the National Geospatial-Intelligence Agency's (NGA), the U.S. Board on Geographic Names and the Ordnance Survey OpenData.

The system retrieves triples by using an available SPARQL endpoint for every ontology integrated with the system. SPARQL [25] is a query language for RDF that can be used to express queries across diverse data sources. SPARQL contains capabilities for querying RDF graph patterns and supports extensible value testing and constraining queries by source RDF graph. The outcome of a SPARQL query can be result sets or RDF graphs. In Fig. 1 (left part), the integration of the system with a number of ontologies is presented. GeoMantis is capable of retrieving RDF triples from any ontology that exposes a SPARQL endpoint and represents factual knowledge in RDF triples.

The final step in the knowledge retrieval workflow, is the processing of the retrieved RDF triples using the CoreNLP system. The object part of the triple is tokenized and lemmatized, and common stopwords are removed. For each RDF triple in the system's geographic knowledge base, a search string is created with lemmatized words.

[4] http://www.geonames.org

Algorithm 1 Knowledge retrieval from ontologies.

```
 1: procedure RetrieveKnowledge($KB$)
    // Use the ISO two-letter country code
 2:     for each $countryCode$ in $countryCodes$ do
 3:         $countryNames \leftarrow$ RetrieveNames($countryCode$)
 4:         for each $countryName$ in $countryNames$ do
 5:             while $N \in \{subject, object\}$ do
 6:                 $SPARQLquery \leftarrow$ CreateQuery($countryName$,$N$)
 7:                 $triples \leftarrow$ RetrieveRDFTriples($SPARQLquery$)
 8:                 for each $triple$ in $triples$ do
 9:                     if N=“subject” then
10:                         $arg1 \leftarrow$ GetPart($subject$,$triple$)
11:                         $arg2 \leftarrow$ GetPart($object$,$triple$)
12:                     else
13:                         $arg1 \leftarrow$ GetPart($object$,$triple$)
14:                         $arg2 \leftarrow$ GetPart($subject$,$triple$)
15:                     end if
16:                     $relation \leftarrow$ GetPart($predicate$,$triple$)
17:                     $searchText \leftarrow arg2$
18:                 end for
    // Use NLP to tokenize and lemmatize
19:                 $searchText \leftarrow$ NLP($searchText$)
    // Use a common stopwords list
20:                 $searchText \leftarrow$ ClearStopWords($searchText$)
21:             end while
22:         end for
23:         SaveGeoDatabase($searchText$,$countryCode$)
24:     end for
25: end procedure
```

Algorithm 1 presents the knowledge retrieval process. The SPARQL query created in line 6 of Algorithm 1 is used to retrieve the RDF triples and it is of the form: `SELECT * WHERE` { <`Countryname`> `?p ?o` } when the country name is in the subject of the triple, and `SELECT * WHERE` { `?p ?o` <`Countryname`> } when the country name is in the object of the triple.

From each retrieved RDF triple, a search text is created using tokenization, lemmatization, and stopword removing techniques. The search text is stored in the GeoMantis local database.

3.3 Query Answering

For each country, a case-insensitive full-text search is executed for each unique word in the text against the search text of each triple in the country's knowledge base. A triple is activated by the text if any of the document's words matches any of the triple's search text words (excluding common stopwords). For example, a document containing the sentence “They had a really nice dish with halloumi while watching the Aegean blue." should activate the RDF triples: <`halloumi`> <`RelatedTo`> <`Cyprus`> and <`Greece`> <`linksTo`> <`Aegean_Sea`>. To maximize the search capabilities, the GeoMantis system uses lemmatized words. Full-text searching takes advantage of the MariaDB's[5] search functionality, using full-text indexing for better search performance.

The final step in the query answering process, involves the ordering of the list of countries and the generation of the predicted geographic focus. Ordering is performed using one of the following strategies:

Percentage of triples applied (PERCR): List of countries is ordered according to the fraction of each country's total number of activated triples over the total number of triples for that country that exist in the geographic knowledge bases, in descending order.

Number of triples applied (NUMR): List of countries is ordered according to each country's total number of activated triples, in descending order.

Term Frequency - Inverse Document Frequency (TF-IDF): List of countries is ordered according to the TF - IDF algorithm [26], which is applied as follows:
D_c is a document created by taking the triples of a country c
TF_t = (Number of times term t appears in D_c) / (Total number of terms in D_c)
$IDF_t = \log_e$(Total number of D_c / Number of D_c with term t in it).

Most triples per country ordering (ORDR): List of countries is ordered according to the number of triples that are retrieved for each country, in descending order.

3.4 System Implementation

The GeoMantis system is built using the PHP web scripting language and the MariaDB database for storing data. The system is designed using an extendable architecture which allows the addition of new functionality.

GeoMantis also exposes a number of its services using a REST API, based on JavaScript Object Notation (JSON)[6] for data interchange and integration with other systems. Knowledge can be updated at any time by querying the corresponding ontology SPARQL endpoint.

Furthermore, the system has a separate module for producing statistics on documents, datasets, and RDF triples and for visualizing them using a powerful graph library based on Chart.js[7]. For each processed document, a detailed log of activated triples is kept for debugging purposes and better understanding of the query answering process.

[5] https://mariadb.org/
[6] http://www.json.org/
[7] http://www.chartjs.org/

Table 1: Information on triples retrieved from ConceptNet and YAGO ontologies for UN countries. The filtered YAGO ontology (YAGO_Fil) is also depicted in this table and is described in Section 5.1.

Property	**ConceptNet**	**YAGO**	**YAGO_Fil**
Total Number of triples	51,771	2,966,765	2,903,186
Number of unique relations	33	373	300
Country with highest number of triples	China	USA	USA
Number of UN countries with triples	193	192	192

4 Empirical Material

The extended evaluation of the GeoMantis system, required three inputs: (i) a list of countries, (ii) generic knowledge from ontologies about each of these countries, and (iii) datasets where the geographic focus of the text is known.

For the first input, we chose countries which are members of the United Nations (UN). The UN is the world's largest intergovernmental organization and has 193 member states. For the other two inputs we provide information in the following sections.

4.1 Use of Generic Ontologies

A large amount of general-purpose knowledge is stored in databases in the form of ontologies. This knowledge is gathered from various sources using human workers, game players, volunteers, and contributors in general. We chose two popular ontologies: ConceptNet [27] and YAGO [28–30] which include generic knowledge for countries instead of only geographic knowledge that exist in a gazetteer. A brief overview of these ontologies is presented in the following paragraphs.

ConceptNet is a freely-available semantic network that contains data from a number of sources such as crowdsourcing projects, Games With A Purpose (GWAPs) [31], online dictionaries, and manually coded rules. In ConceptNet, data are stored in the form of edges or assertions. An edge is the basic unit of knowledge in ConceptNet and contains a relation between two nodes (or terms). Nodes represent words or short natural language phrases. ConceptNet version 5.6 includes 37 relations, such as "AtLocation", "isA", "PartOf", "Causes" etc. The following are examples of edges available in ConceptNet: <`cat`> <`RelatedTo`> <`meow`>, <`statue`> <`AtLocation`> <`museum`>. ConceptNet is not represented in an RDF format, but there is relevant work that suggests such a conversion [32]. ConceptNet's version 4 ability to answer IQ questions using simple test-answering algorithms was evaluated and the results showed that the system has the Verbal IQ of an average four-year-old child [33].

For each UN country, its name along with its alternate names are extracted and the ConceptNet 5.6 API[8] is queried for returning the proper Uniform Re-

[8] http://api.conceptnet.io/

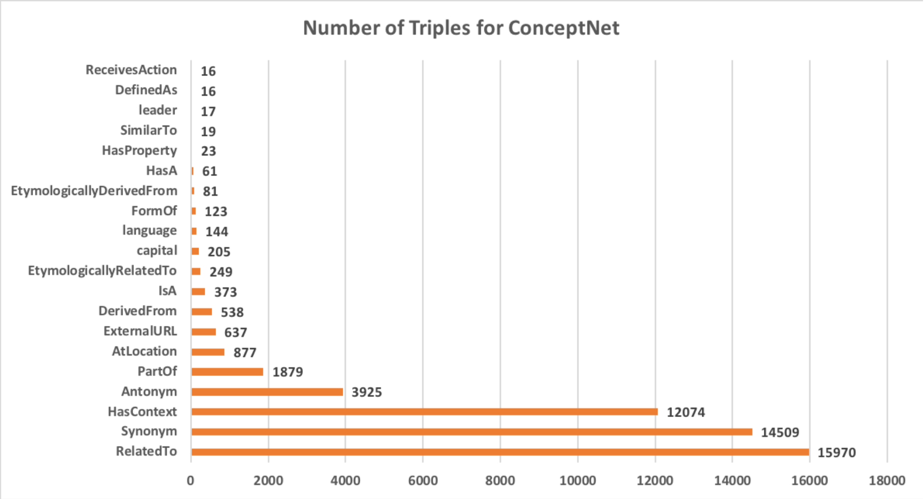


Fig. 2: The 20 most frequent relations in triples retrieved from ConceptNet ontology about UN countries.

source Identifier (URI) in the database. In ConceptNet, each URI includes the language (e.g., "en") and the term. This is an example of a complete URI: "/c/en/peru". When the term includes spaces (e.g., "United Kingdom"), these are substituted by underscores, i.e., "c/en/united_kingdom".

For each obtained URI, all facts are retrieved in the form of triples <Arg1> <Relation> <Arg2> and are stored in the GeoMantis geographic knowledge database. In ConceptNet, the country name can appear either in <Arg1> or <Arg2> and an additional check is needed to capture the appropriate search string. For example, when a search for "Greece" is performed, facts like the ones presented in Fig. 3 are returned, which after processing (see Algorithm 1) result to the search strings: europe and ithaka. In Fig. 2, the 20 most frequent relations in the retrieved knowledge are depicted.

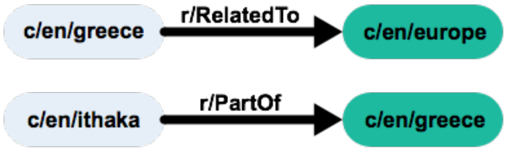


Fig. 3: Examples of facts retrieved from ConceptNet when the search term "Greece" is used.

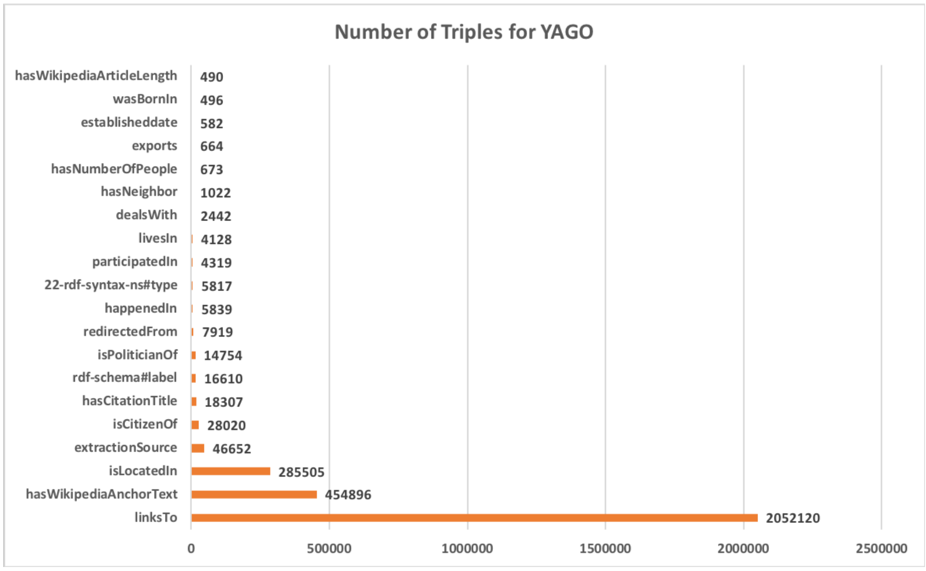


Fig. 4: The 20 most frequent relations in triples retrieved from YAGO ontology about UN countries.

YAGO (Yet Another Great Ontology) is a semantic knowledge base built from sources like Wikipedia, WordNet [34] and GeoNames[9]. More specifically, information from Wikipedia is extracted from categories, redirects and infoboxes available in each wikipedia page. Also, there is a number of relations between facts that are described in detail in the work of Hoffart et al. [28]. Currently, YAGO contains 447 million facts and about 9,800,000 entities. Facts in YAGO were evaluated by humans, reporting an accuracy of 95%.

Relations in YAGO are both semantic (e.g., "`wasBornOnDate`", "`locatedIn`" and "`hasPopulation`") and technical oriented (e.g., "`hasWikipediaAnchorText`", "`hasCitationTitle`"). A search for "Greece" in YAGO returns facts like the ones presented in Fig. 5.

Moreover, YAGO has a number of spatial relations that place an object in a specific location (i.e., country, city, administrative region, etc.). For example, relations "`wasBornIn`", "`diedIn`", "`worksAt`" place an entity of type `Person` in a location, e.g., <`Isaac_Asimov`> <`wasBornIn`> <`Petrovichi`>.

For retrieving facts, the YAGO SPARQL endpoint[10] was queried for each UN country name along with its alternate names.

[9] http://www.geonames.org
[10] https://linkeddata1.calcul.u-psud.fr/sparql

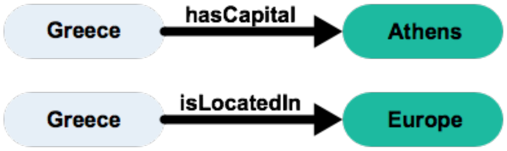


Fig. 5: Examples of facts retrieved from YAGO when the search term “Greece” is used.

4.2 Corpora and Datasets

The last of the inputs needed for the evaluation process are the pre-tagged text corpora. These are collections of texts whose geographic focus is known and available for machine reading.

To evaluate the GeoMantis system in a challenging setting, we processed a number of documents from popular corpora by removing any reference to the country of focus for that document and its alternate names, i.e., a document with geographic focus in “Greece” will not have the word “Greece” or “Hellas” or “Hellenic Republic” in its text after the processing.

There are two commonly used corpora for conducting experiments in this line of research; the Reuters Corpus Volume 1 (RCV) and the New York Times Annotated Corpus (NYT). The available content is tagged with location metadata at country-level. Moreover, they contain a plethora of documents for experimentation from different news topics and about various countries.

The Reuters Corpus Volume 1 (RCV) comprises 810,000 Reuters, English language news stories that were made available in 2000 by Reuters Ltd. Each story is in English and the corpus contains stories from 20/08/1996 to 19/08/1997, tagged with information on where the story is geographically located [35]. Tagging was performed by a combination of automatic categorizing techniques, manual editing, and manual correction.

The New York Times Annotated Corpus (NYT) has in its collection over 1,800,000 articles, written and published by the New York Times between 1987 and 2007. Most articles are tagged with location metadata [36]. The NYT corpus categorization allows a news story to be tagged with more than one locations. Tagging was performed by humans.

From the above two corpora we created six datasets to use in the evaluation of the GeoMantis system. These datasets had either the target country and its alternate names obscured, i.e., substituted with the word “unknown” or not present at all. To the best of our knowledge, there is no corpus that guarantees that there is no mention of the target country inside the document. For that reason, we used corpora that are frequently used in this line of research and we constructed datasets either by obscuring or by selecting texts that do not have a mention of the target country to evaluate GeoMantis. The alternate names of the countries were retrieved from the GeoNames database and were limited to english alternate names only.

From the RCV corpus, two datasets were created using 1000 documents, uniformly randomly selected, without replacement, from the set of news stories in the dataset: the `RCV_obs`, where the target country and its alternate names are obscured and the `RCV_npr`, where the target country and its alternate names are not present in the document's text.

From the NYT corpus, two datasets were created using 1000 news stories, uniformly randomly selected, without replacement, from the set of news stories in the dataset that belong to the "Top/News/World/ Countries and Territories/" category with a single country tag: the `NYT_obs`, where the target country and its alternate names are obscured, and the `NYT_npr`, where the target country and its alternate names are not present in the document's text.

The majority of stories in the NYT corpus are geographically focused on the United States of America and Russia, and the majority of stories in the RCV1 corpus are geographically focused on the United States of America and the United Kingdom. For each of the four datasets, we tried to have a balanced distribution of news stories per target country of focus, hence five news stories were uniformly randomly selected, without replacement (if they were available), for each UN member country from the respective corpus. The remaining documents were uniformly randomly selected, without replacement, from the whole pool of documents of that corpus.

We also created two new datasets for the comparison of GeoMantis with other systems and two baseline metrics, the `EVA_obs` and the `EVA_npr`.

The `EVA_obs` dataset included 500 uniformly randomly selected without replacement news stories from the RCV corpus and 500 uniformly randomly selected without replacement news stories from the NYT corpus categorized under the "Top/News/World/ Countries and Territories/" category with a single country tag, in a similar way as with the rest of the datasets. Every occurrence of the target country was substituted with the word "unknown". For the `EVA_npr` dataset the same procedure was followed, but each story in the dataset did not have any occurrence of the target country or its alternate names.

For uniformity, from each of the two corpora, two documents were uniformly randomly selected without replacement (if they were available) for each UN member country. The remaining documents were uniformly randomly selected without replacement from the whole pool of documents. As before, this process allowed a balanced distribution of stories per country in the dataset.

5 Evaluation and Analysis

The GeoMantis system is evaluated on whether it can identify the geographic focus of a text document, when the country name in that text is obscured or does not exist, using only knowledge from generic ontologies. The process followed, the metrics, and the results of the evaluation are presented in this section.

Table 2: Characteristics of the six datasets, including number of documents, number of tagged countries, total and mean number of words and the percentage of the NER labels. Details on the identified named entities are presented as the percentage of words tagged with NER labels in each dataset along with the five labels used in our experiments which are presented as the fraction of the words tagged with each label over the total number of NER labels, converted to a percentage .

Dataset	**RCV_obs**	**RCV_npr**	**NYT_obs**	**NYT_npr**	**EVA_obs**	**EVA_npr**
Number of documents in dataset	1000	1000	1000	1000	1000	1000
Number of countries in dataset	180	125	171	117	186	138
Number of words in dataset	174347	166373	393531	362228	283896	216014
Mean number of words per document	174	166	394	362	284	216
Percentage of Named Entities	23.19%	31.76%	29.36%	24.37%	25.51%	27.86%
[location]	10.97%	9.83%	15.14%	14.68%	14.25%	12.66%
[organization]	21.78%	19.40%	15.08%	17.44%	17.16%	17.49%
[money]	2.63%	2.62%	1.49%	1.83%	1.69%	1.86%
[person]	20.25%	18.88%	23.59%	24.36%	22.31%	22.63%
[misc]	6.39%	6.36%	10.88%	9.93%	9.28%	8.69%

A two phase evaluation was conducted: the 1st phase measured the system's performance for each of the parameters (parameter selection) in identifying the geographic focus of a document at a country-level, and the 2nd phase compared the GeoMantis system using the prevailing strategy from the 1st phase, with two opensource freely available systems and two common baseline metrics (comparative evaluation). For these experiments, general-purpose knowledge was retrieved for countries that are members of the United Nations (UN)[11] as described in Section 4.1.

[11] http://www.un.org

5.1 Parameter Selection

The 1st phase of the evaluation was conducted using the four datasets described in Section 4.2. We evaluated every combination of values for the ontology, and the PERC and TF-IDF query answering strategies.

A similar evaluation was conducted and described in detail in our previous work [23]. That evaluation included three datasets (two from the same sources as with this evaluation and one manually created from the WikiTravel[12] website) and knowledge from Conceptnet and YAGO. The results of that evaluation suggested that the best performing parameters were the YAGO ontology, the application of NER filtering, and the PERC query answering strategy, even though the TF-IDF strategy was also performing very well. Those datasets were processed by just obscuring the reference country name from the document, as opposed to the extensive filtering of both the name and alternate names we performed in this evaluation.

Parameters like NER filtering, were tested thoroughly in the previous evaluation of GeoMantis and found to increase the performance of the system when used, hence it was always enabled in this evaluation. NER filtering includes the use of words that were labeled as location, person, organization, and money by the NER process. Although not reported here, the application of the NER filter also significantly reduces the processing time. Furthermore, the *Number of triples activated (NUMR)* and *Most triples per country ordering (ORDC)* query answering strategies, were found not to perform well and were not tested in this evaluation.

For the evaluation process, the datasets were imported to the GeoMantis database and processed with the Stanford CoreNLP. Then, the system's knowledge retrieval engine was directed to ConceptNet and YAGO ontologies to retrieve RDF triples. These triples were processed using the NLP system. Table 1 depicts the properties for the ontologies used.

The performance of each combination of parameters, was evaluated using the mean position metric and the accuracy. The mean position ($\bar{P}$) denotes the position of the target country in the ordered list of countries over the number of countries available in the dataset. For comparison purposes, this number is converted to a percentage.

The accuracy(A_i) of the system is defined as $A_i = \frac{N_i}{C}$, where $i \in \{1, 2, 3, ..., M\}$ and M is the number of countries in the dataset, N_i denotes the number of correct assignments of the target country when the target country's position is $\leq i$ in the ordered list of countries and C denotes the number of available documents in the dataset.

The parameter selection process was applied on the RCV_obs, RCV_npr, NYT_obs and NYT_npr datasets.

In Table 3, we present the results of the parameter selection process after the chosen ontology and the query answering strategy followed (see Section 3.3) are tested. These results are also depicted graphically in Fig. 6.

[12] https://wikitravel.org

Table 3: Results from the parameter selection phase of the GeoMantis system evaluation. The query answering strategies and ontologies, when the NER filtering option is used, were evaluated. Rows highlighted in light blue, identify the best performing set of parameters in terms of minimum value for $\bar{P}$ and maximum value for A_1 and A_2.

#	**Dataset**	**Ontology**	**Strategy**	A_1	A_2	$\bar{P}$
YP1	RCV_obs	YAGO	PERCR	23.70	39.80	8
YT1	RCV_obs	YAGO	TF-IDF	41.10	61.60	6
CP1	RCV_obs	ConceptNet	PERCR	18.70	27.7	16
CT1	RCV_obs	ConceptNet	TF-IDF	19.80	29.30	16
YP2	RCV_npr	YAGO	PERC	36.30	48.80	8
YT2	RCV_npr	YAGO	TF-IDF	45.40	58.60	8
CP2	RCV_npr	ConceptNet	PERCR	29.40	42.80	12
CT2	RCV_npr	ConceptNet	TF-IDF	27.50	37.90	13
YP3	NYT_obs	YAGO	PERCR	18.60	31.20	11
YT3	NYT_obs	YAGO	TF-IDF	34.00	52.40	7
CP3	NYT_obs	ConceptNet	PERCR	11.60	22.20	14
CT3	NYT_obs	ConceptNet	TF-IDF	15.10	27.00	13
YP4	NYT_npr	YAGO	PERCR	36.40	50.70	10
YT4	NYT_npr	YAGO	TF-IDF	49.80	65.50	7
CP4	NYT_npr	ConceptNet	PERCR	26.50	44.00	11
CT4	NYT_npr	ConceptNet	TF-IDF	28.80	43.70	11

Comparing the results in terms of ontology used, knowledge from YAGO yields better results than that of ConceptNet. Further analysis of the two ontologies, shows a huge gap in the amount of facts retrieved for each country. In particular, YAGO includes 2,966,765 triples against 51,771 triples in ConceptNet.

The results indicate that the common prevailing strategy for all four datasets is **TF-IDF** when the **YAGO** knowledge base is used. These results are inline with the results from our previous experiments, since the TF-IDF strategy performed almost equally well with the PERC startegy in that evaluation. Furthermore, we speculate that the increase in the amount of triples from the YAGO ontology required a more refined method of selecting the activated triple than the simple PERC strategy.

The results propose that further tuning of the selected parameters could increase the accuracy and minimize the mean position. Instead of using the "money" NER tag, we chose the "misc" tag that actually contains named entities that do not exist in any other tags. The "money" tag included words like "billion", "4,678,909" that do not offer much in the query answering process.

Furthermore, we created a filtered version of the YAGO ontology (`YAGO_Fil`), by removing triples with relations that identify and contain technical information

Table 4: Results from fine-tuning the parameter selection phase of the GeoMantis system evaluation. We examined the performance when using the "misc" NER tag instead of "money" and the use of the filtered YAGO ontology (`YAGO_Fil`).

#	**Dataset**	**Ontology**	**Strategy**	A_1	A_2	$\bar{P}$
YFT1	RCV_obs	YAGO_Fil	TF-IDF	42.80	61.60	5
YFT2	RCV_npr	YAGO_Fil	TF-IDF	49.60	62.20	6
YFT3	NYT_obs	YAGO_Fil	TF-IDF	36.60	55.20	5
YFT4	NYT_npr	YAGO_Fil	TF-IDF	52.90	67.90	5

(e.g., "`owl#sameAs`", "`extractionSource`","`hasWikipediaArticleLength`") and relations like "`imageflag`" and "`populationestimaterank`", that do not include useful information.

Results presented in Table 4, suggest that the usage of the **YAGO_Fil** ontology with the "**misc**" tag, minimize $\bar{P}$ and maximize the accuracy of both A_1 and A_2 for all four datasets. In fact, the $\bar{P}$ is decreased by two positions in three out of four datasets and A_1 and A_2 were increased for all datasets.

5.2 Comparative Evaluation

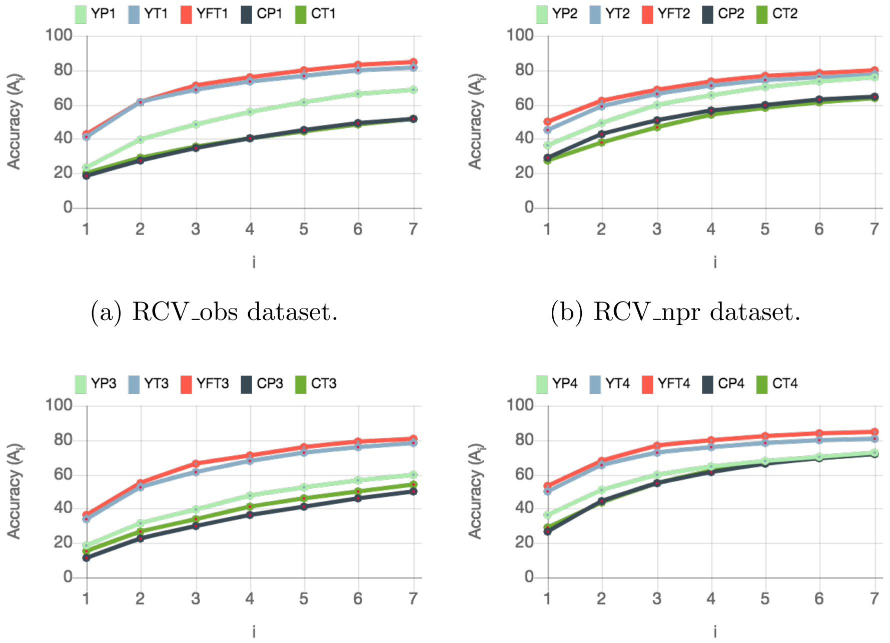
In the 2nd phase of the evaluation, the GeoMantis system, using the prevailing strategy identified in the 1st phase of the evaluation, was compared with two freely available opensource systems, CLIFF-CLAVIN and Mordecai, and two common baseline metrics. These metrics included the random selection of countries (RAND) and the ordering of countries based on their frequency of appearance in the dataset (ORDC) for ordering the list of countries.

Two additional independent datasets were used comprising previously unseen documents from the same sources used for the 1st phase.

For the comparative evaluation, we used the accuracy metric and the unanswered metric. The unanswered metric U denotes the percentage of the number of documents processed without the system returning a result.

To conduct the comparative evaluation, the CLIFF-CLAVIN geolocation service was set up and a script was used to read the JSON output of the system. More specifically, the "places/focus/countries" array of the JSON results was used.

Results returned from the CLIFF-CLAVIN system are not ordered, so for comparison reasons with the GeoMantis system, the A_1 and A_7 metrics are used, where A_1 is the accuracy of the system when only one result is returned and it is the correct target country assignment and A_7 is the accuracy of the system when up to 7 results are returned and the correct target country assignment is in this set. The reason 7 was chosen is that it corresponds to the maximum number of predicted countries CLIFF-CLAVIN returns when executed on both the `EVAL_obs` and the `EVAL_npr` datasets and the target country is identified by



(a) RCV_obs dataset.

(b) RCV_npr dataset.

(c) NYT_obs dataset.

(d) NYT_npr dataset.

Fig. 6: Graphical representation of the results when the four datasets are used. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

any one of them. This weakness of the CLIFF-CLAVIN system is also stressed by other researchers [18] who used this system for comparison purposes.

For Mordecai, a webservice was not available, hence we set up the system locally, following the instructions[13] given by its developer. More specifically, this system requires Python version 3, spaCy NLP model and the GeoNames database. In order to work, Mordecai needs access to a Geonames gazetteer running in Elasticsearch[14]. We created a python script that can take a folder of documents and parse them using the Mordecai API using the `geo.infer_country` function.

The results are stored in a new file and are filtered so that only the returned tag "`country_predicted`" is stored in the output file. Mordecai returns the predicted country for each place name in ISO3 country code format (e.g., GRC, BGR). To be able to compare this system, we created a script that converts ISO3

[13] https://github.com/openeventdata/mordecai
[14] https://www.elastic.co/

to ISO country code format and suggests a geographic focus for the document according to a frequency-based approach, i.e., the returned countries are ordered according to their frequency of appearance. The comparative evaluation was applied on the EVA_obs and EVA_npr datasets.

In Table 5, rows highlighted in light green identify the best results in terms of A_1 and A_7 for each of the two datasets. In Fig. 7 these results are presented graphically, illustrating all comparative evaluation experiments.

Results from the 2nd phase evaluation for the GeoMantis system are comparable to that of CLIFF-CLAVIN, Mordecai and that of the two baseline metrics. In cases where the target country is obscured or not present in the dataset, the GeoMantis system outperforms both CLIFF-CLAVIN and Mordecai, and the two baseline metrics.

The `EVA_npr` dataset presents better results in terms of accuracy, since the information present in this dataset is unaffected by the obscuring process. The way stories are written probably includes other type of information to identify the country without an explicit mention of it in the text. On the other hand, stories in the `EVA_obs` dataset have an explicit mention of the target country in the document text that was obscured. This led to fewer references left in the story text and hence, made it more difficult to identify the target country.

Furthermore, the comparison of C1 with M1 and C2 with M2 shows that CLIFF-CLAVIN performs marginally better than Mordecai, when the target country is obscured or not present in the document. This was also tested in work of Imani et al. [18], on sentences without the target country obscured and the results show that the CLIFF-CLAVIN system outperformed Mordecai in terms of accuracy.

In terms of the U metric, CLIFF-CLAVIN and Mordecai have a relatively high percentage of unanswered documents. More specifically, CLIFF-CLAVIN was not able to identify the geographic focus of 179 documents in the `EVA_npr` dataset and 107 documents in the `EVA_obs`.

6 Discussion

The evaluation process results, show that the methodology chosen, i.e., using general purpose ontologies, is applicable and well suited for the problem of identifying the geographic focus of documents that do not explicitly mention the target country. In this work, a number of strategies were tested and the one that presents better results, is the ordering of the list of countries according to the TF-IDF algorithm, in descending order (TF-IDF). In terms of knowledge source, the YAGO ontology results present a greater accuracy than the ConceptNet ontology results. Moreover, the usage of named entities filtering on the document text increases the performance and the accuracy of target country identification.

The field of text comprehension can benefit from the recent advances in Artificial Intelligence [37]. Researchers started growing concern in algorithm transparency and accountability, since most newly developed “intelligent” systems and algorithms are opaque black boxes where you give an input and the output

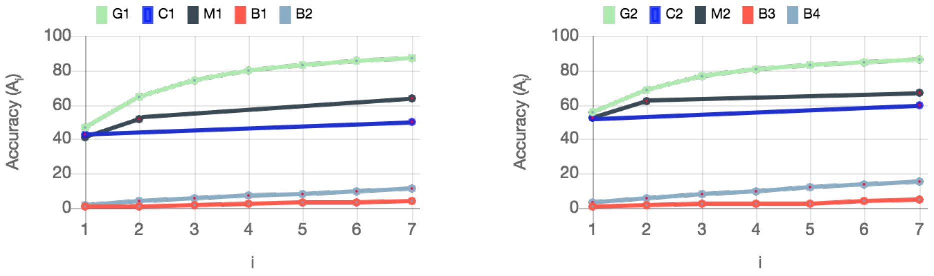
Table 5: Comparison of the GeoMantis system with CLIFF-CLAVIN, Mordecai and the Baseline. Rows highlighted in light green identify the results that are comparable.

#	Dataset	System	Parameters	A_1(%)	A_2(%)	A_7(%)	U(%)
G1	EVA_obs	Geomantis	YAGO_Fil, TF-IDF	46.60	64.60	87.02	0
C1	EVA_obs	CLIFF-CLAVIN	default	42.50	-	50.00	10.70
M1	EVA_obs	Mordecai	default	41.10	51.50	64.00	7.20
B1	EVA_obs	Baseline	RAND	0.50	1.10	3.90	0
B2	EVA_obs	Baseline	ORDC	2.00	3.80	11.00	0
G2	EVA_npr	Geomantis	YAGO_Fil, TF-IDF	55.40	68.20	86.10	0
C2	EVA_npr	CLIFF-CLAVIN	default	52.70	-	59.50	17.90
M2	EVA_npr	Mordecai	default	52.10	62.20	66.90	14.80
B3	EVA_npr	Baseline	RAND	0.80	1.30	5.10	0
B4	EVA_npr	Baseline	ORDC	3.30	5.50	15.70	0

is presented without actually presenting their “thinking” process. Algorithms should provide transparency [38] on their methods, results, and explanations. The system we designed is inline with that direction, since it exposes its query answering strategy and can provide explanations on why a specific geographic focus of a document was chosen, i.e., the facts that were activated from the ontology. The explanatory role of such systems, with respect to the target natural cognitive systems they take as source of inspiration, is highlighted in work of Lieto and Radicioni [39].

Currently, there are not many systems dedicated for the task of identifying the geographic focus of a text document. The majority of the available systems are basically geoparsers that offer focus identification as an additional feature of their primary purpose and they rely on text that has a good amount of place mentions in it. When these systems are tested on documents that have few place mentions, they perform poorly in terms of accuracy, as opposed to the high accuracy they present when tested on datasets that have mentions of locations. This limitation is waived in GeoMantis, which does not rely exclusively on place mentions to work, but uses any type of general-purpose knowledge that can be found in generic ontologies. Comparative evaluation was only possible with CLIFF-CLAVIN and Mordecai, since the other systems presented in Section 2 were not accessible or they were not freely available for local deployment and testing.

GeoMantis is currently able to identify country-level geographic focus, but it can be expanded to handle other levels (e.g., administrative area, city), as long as the relevant knowledge triples exist in the selected ontologies. The tech-



(a) EVA_obs dataset.

(b) EVA_npr dataset.

Fig. 7: Graphical representation of the comparative evaluation results when the `EVA_obs` and `EVA_npr` datasets are used. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

niques used for news stories, could also apply to other types of documents such as myths, novels, legal documents, etc. This line of research can also find applications for document classification and geographic knowledge extraction from text. Moreover, it can be used with techniques for linking image and text-based contents together, for document management tasks [40].

7 Conclusion and Future Work

In this work we tried to tackle the problem of identifying the geographic focus of text that does not explicitly mention the target country, making our problem one of inference or prediction, rather than one of identification. General-purpose ontologies were used, instead of gazetteers, atlases or other purposed built geographic bases, to address this problem. More specifically, we demonstrated a methodology that retrieves general-purpose knowledge in the form of RDF triples, processes it and identifies the geographic focus of a document. This methodology and the GeoMantis system, were evaluated in various scenarios using "gold standard" annotated datasets and metrics, and results showed that the GeoMantis system outperforms the other two systems tested and the two baseline metrics, when certain conditions apply.

GeoMantis can be extended to utilize paths of various lengths between a geographical entity (e.g., country) and other entities. An example of a length 2 relation path is depicted in Fig. 8. In such a scenario, if a document contains the word "Florence", facts related to `Greece` will be activated. Results from this approach will be compared with results from using direct connections between the entities (length 1 relation path). Early experiments suggest that this will decrease the performance of the system, as it ends up connecting countries to entities, spatial or not, that are conceptually remote (see Fig. 8).

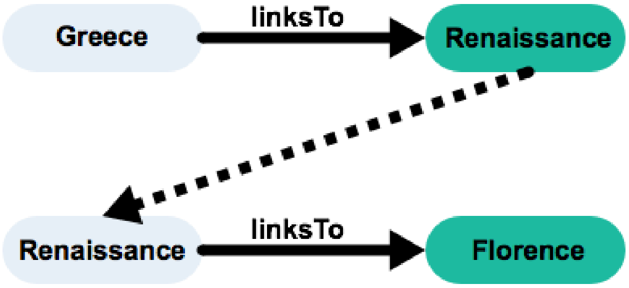


Fig. 8: An example of a length 2 relation path from YAGO.

Crowdsourcing approaches like GWAPs or hybrid solutions [41], could also be applied in future versions of the system for fact disambiguation. The integration of other ontologies or knowledge bases with GeoMantis, like the one generated from the Never Ending Language Learner [42], DBpedia [43], Wikidata [44] or their combination, could also be explored.

We believe that the GeoMantis system can be used in several application scenarios, such as document searching and tagging, games (e.g., taboo game challenges), and news categorization. Its extendable architecture enables the addition of new functionality and new sources of knowledge and also the integration with other systems. GeoMantis could also be used in conjunction with other systems to return results in cases where the other systems are not able to return any.

References

1. Tversky, B.: Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In Frank, A.U., Campari, I., eds.: Spatial Information Theory A Theoretical Basis for GIS: European Conference, COSIT'93 Marciana Marina, Elba Island, Italy September 19–22, 1993 Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (1993) 14–24
2. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N.: Adding Geographic Scopes to Web Resources. Computers, Environment and Urban Systems **30**(4) (2006) 378–399
3. Bower, G.H.: Experiments on Story Understanding and Recall. Quarterly Journal of Experimental Psychology **28**(4) (1976) 511–534
4. Andogah, G., Bouma, G., Nerbonne, J.: Every Document has a Geographical Scope. Data and Knowledge Engineering **81-82** (2012) 1–20
5. Leidner, J.L., Lieberman, M.D.: Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. SIGSPATIAL Special **3** (2011) 5–11
6. Melo, F., Martins, B.: Automated Geocoding of Textual Documents: A Survey of Current Approaches. Transactions in GIS **21**(1) (2016) 3–38

7. Monteiro, B.R., Davis, C.A., Fonseca, F.: A survey on the Geographic Scope of Textual Documents. Computers and Geosciences **96** (2016) 23–34
8. Woodruff, A.G., Plaunt, C.: GIPSY: Georeferenced Information Processing SYstem. Journal of the American Society for Information Science **45** (1994) 645–655
9. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web Content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2004) 273–280
10. Zubizarreta, Á., de La, Cantera, J., Arias, M., Cabrero, J., García, G., Llamas, C., Vegas, J., Garc, G.: Extracting Geographic Context from the Web: GeoReferencing in MyMoSe. Advances in Information Retrieval (2009) 554–561
11. D'Ignazio, C., Bhargava, R., Zuckerman, E., Beck, L.: CLIFF-CLAVIN: Determining Geographic Focus for News Articles. In: Proceedings of the NewsKDD: Data Science for News Publishing. (2014)
12. Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B.: The Design and Implementation of SPIRIT: A Spatially Aware Search Engine for Information Retrieval on the Internet. International Journal of Geographical Information Science **21**(7) (2007) 717–745
13. Yu, J.: Geotagging Named Entities in News and Online Documents. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management. (2016) 1321–1330
14. Teitler, B.E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: NewsStand: A New View on News. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. (2008) 1–18
15. de Alencar, R.O., Davis Jr, C.A. In: Geotagging Aided by Topic Detection with Wikipedia. Springer Berlin Heidelberg, Berlin, Heidelberg (2011) 461–477
16. Quercini, G., Samet, H., Sankaranarayanan, J., Lieberman, M.D.: Determining the Spatial Reader Scopes of News Sources Using Local Lexicons. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10. (2010) 43–52
17. Watanabe, K.: Newsmap. Digital Journalism **6**(3) (2018) 294–309
18. Imani, M.B., Chandra, S., Ma, S., Khan, L., Thuraisingham, B.: Focus Location Extraction From Political News Reports With Bias Correction. In: 2017 IEEE International Conference on Big Data (Big Data). (2017) 1956–1964
19. Brun, G., Dominguès, C., Paris-est, U.: TEXTOMAP : Determining Geographical Window for Texts. In: Proceedings of the 9th Workshop on Geographic Information Retrieval. GIR '15, New York, NY, USA, ACM (2015) 7–8
20. Halterman, A.: Mordecai: Full Text Geoparsing and Event Geocoding. The Journal of Open Source Software **2**(9) (2017)
21. Lassila, O., Swick, R.R.: Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, 1999 (1999)
22. Hayes, P., McBride, B.: RDF Semantics. W3C Recommendation. World Wide Web Consortium (2004)
23. Rodosthenous, C.T., Michael, L.: GeoMantis: Inferring the Geographic Focus of Text using Knowledge Bases. In: Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,, INSTICC, SciTePress (2018) 111–121
24. Manning, C.D., Bauer, J., Finkel, J., Bethard, S.J., Surdeanu, M., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of

the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. (2014) 55–60

25. Quilitz, B., Leser, U. In: Querying Distributed RDF Data Sources with SPARQL. Springer Berlin Heidelberg, Berlin, Heidelberg (2008) 524–538
26. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Volume 1. Cambridge University Press (2008)
27. Speer, R., Havasi, C. In: ConceptNet 5: A Large Semantic Network for Relational Knowledge. Springer Berlin Heidelberg, Berlin, Heidelberg (2013) 161–176
28. Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-kelham, E., Melo, G.D., Weikum, G.: YAGO2 : Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In: Proceedings of the 20th International Conference on World Wide Web. (2011) 229–232
29. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: Proceedings of the 16th International Conference on World Wide Web. (2007) 697–706
30. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web **6**(3) (2008) 203–217
31. von Ahn, L., Dabbish, L.: Designing Games With a Purpose. Communications of the ACM **51**(8) (2008) 57
32. Najmi, E., Malik, Z., Hashmi, K., Rezgui, A.: ConceptRDF: An RDF Presentation of ConceptNet Knowledge Base. In: 2016 7th International Conference on Information and Communication Systems (ICICS). (2016) 145–150
33. Ohlsson, S., Sloan, R.H., Turán, G., Urasky, A.: Verbal IQ of a Four-Year Old Achieved by an AI System. In: Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence. (2013) 89–91
34. Fellbaum, C.: WordNet. In Poli, R., Healy, M., Kameas, A., eds.: Theory and Applications of Ontology: Computer Applications. Springer Netherlands (2010) 231–243
35. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research **5** (2004) 361–397
36. Sandhaus, E.: The New York Times Annotated Corpus LDC2008T19. DVD. Linguistic Data Consortium, Philadelphia (2008)
37. Hermann, K.M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching Machines to Read and Comprehend. In: Advances in Neural Information Processing Systems 28 (NIPS 2015). (2015) 1–13
38. Dignum, V.: Responsible Autonomy. In: Proceedings of the Twenty -Sixth International Joint Conference on Artificial Intelligence (IJCAI2017). (2017) 4698–4704
39. Lieto, A., Radicioni, D.P.: From Human to Artificial Cognition and Back: New Perspectives on Cognitively Inspired AI Systems. Cognitive Systems Research **39** (2016) 1–3
40. Cristani, M., Tomazzoli, C.: A Multimodal Approach to Relevance and Pertinence of Documents. In Fujita, H., Ali, M., Selamat, A., Sasaki, J., Kurematsu, M., eds.: Trends in Applied Knowledge-Based Systems and Data Science: 29th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2016, Morioka, Japan, August 2-4, 2016, Proceedings. Springer International Publishing, Cham (2016) 157–168
41. Rodosthenous, C., Michael, L.: A Hybrid Approach to Commonsense Knowledge Acquisition. In: Proceedings of the 8th European Starting AI Researcher Symposium. (2016) 111–122

42. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Mishra, B.D., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., Welling, J.: Never-Ending Learning. In: AAAI Conference on Artificial Intelligence. (2015) 2302–2310
43. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., Others: DBpedia–a Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web **6**(2) (2015) 167–195
44. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D. In: Introducing Wikidata to the Linked Data Web. Springer International Publishing, Cham (2014) 50–65