GeoMantis: Inferring the Geographic Focus of Text using Knowledge Bases

Christos T. Rodosthenous and Loizos Michael
Open University of Cyprus, P.O. Box 12794, Nicosia, Latsia, Cyprus

Keywords: Information Retrieval, Geographic Focus Identification, Knowledge Bases, Natural Language Processing, Geographic Information Systems.

Abstract: We consider the problem of identifying the geographic focus of a document. Unlike some previous work on this problem, we do not expect the document to explicitly mention the target region, making our problem one of inference or prediction, rather than one of identification. Further, we seek to tackle the problem without appealing to specialized geographic information resources like gazetteers or atlases, but employ general-purpose knowledge bases and ontologies like ConceptNet and YAGO. We propose certain natural strategies towards addressing the problem, and show that the GeoMantis system that implements these strategies outperforms an existing state-of-the-art system, when compared on documents whose target region (country, in particular) is not explicitly mentioned or is obscured. Our results give evidence that using general-purpose knowledge bases and ontologies can, in certain cases, outperform even specialized tools.

1 INTRODUCTION

One of the many tasks humans can perform, is to read a document and identify its geographic focus (Tversky, 1993). This task is more evident in stories where human readers can identify the location where the story takes place, along with other properties (Bower, 1976). For the same task, a machine needs to process the document, identify location mentions from text and then try to identify its geographic focus. Relevant research in this direction has led to methods and systems that rely on gazetteers, atlases and dictionaries with geographic-related content, that identify the geographic focus of text. In this work, we investigate whether generic pre-existing knowledge bases or 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 document.

We present **GeoMantis**, a system developed to infer the country-level focus of a text document or a web page using knowledge from general-purpose knowledge bases and ontologies. In particular, the system takes as input any type of document, it processes it and it stores the contents of the document in a database. Independently of the previous process, the system retrieves facts from knowledge bases about countries, processes each fact, filters it using its internal mechanisms and stores it in a database. Moving further in this pipeline, a full-text search algorithm is in place for running each search text of the document against the search text of each fact in the country's knowledge base set. A number of filtering options are also available during this process. This search returns the set of country facts that are activated by the document text.

The outcome of the reasoning process is 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.

This work concludes with an evaluation of the GeoMantis system strategies, a comparison against other approaches, a presentation of current work on the GeoMantis system and discussion of future directions and possible extensions.

2 RELEVANT BACKGROUND

The research in the area of geographic focus identification (Andogah et al., 2012) led to the development of systems that perform such a task. Most systems share a common feature: they rely on geoparsers, i.e., systems for extracting places from text (Leidner and Lieberman, 2011; Melo and Martins, 2016), for identifying locations, disambiguating them

Rodosthenous, C. and Michael, L.
GeoMantis: Inferring the Geographic Focus of Text using Knowledge Bases.
DOI: 10.5220/0006588501110121
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 111-121
ISBN: 978-989-758-275-2

and finally identifying the geographic focus of the text. This approach performs well when documents include place mentions for geoparsers to work, but leaves open the case of documents that have none or very few place mentions. A document could also contain references to geographic locations in the form of historical dates, monuments, ethnicity, typical food, traditional dances and others (Monteiro et al., 2016).

There are several general-purpose knowledge bases and ontologies (e.g., ConceptNet, YAGO, Wikidata) available that can be used to infer the geographic focus of text since they contain broad knowledge related to geographic locations. We recognize the fact that these knowledge bases could also include geographical knowledge, but this knowledge is not inserted in any specialized form like the one entered by experts in gazetteers or atlases, and is crowd-contributed.

2.1 Existing Systems

Work on geographic focus identification goes back in the 90's with a system called **GIPSY** (Woodruff and Plaunt, 1994) for automatic geo-referencing of text. In the 00's, the **Web-a-Where** system (Amitay et al., 2004) was introduced, which can identify a place name in a document, disambiguate it and determine its geographic focus. 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 (Zubizarreta et al., 2009). 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 (D'Ignazio et al., 2014), which identifies the geographic focus on 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"[1], 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[2] suite of tools for quantitative text analysis of media coverage.

Related to this line of research, is the work on **SPIRIT** (Purves et al., 2007), a spatially aware search engine which is capable of accepting queries in the form of <theme><spatial relationship><location>. Relevant research is also found in the work of Yu (Yu, 2016) 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"** (Teitler et al., 2008), 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 Quentin et al. (de Alencar and Davis Jr, 2011; Quercini et al., 2010). de Alencar and Davis 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.

2.2 Knowledge Sources

Currently, a large amount of general-purpose knowledge is gathered from various sources using human workers, players and volunteers. This knowledge is stored in the form of facts or rules in conceptual knowledge bases like ConceptNet (Speer and Havasi, 2013) and YAGO (Hoffart et al., 2011; Suchanek et al., 2007; Suchanek et al., 2008). A brief overview of these knowledge bases is presented in the following paragraphs.

ConceptNet is a freely-available semantic network that contains data from a number of sources like crowdsourcing projects, Games With A Purpose (GWAPs) (von Ahn and Dabbish, 2008), 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

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

[2] http://mediameter.org/

or short natural language phrases. Currently ConceptNet (version 5) includes more than 25 relations like "AtLocation", "isA", "PartOf", "Causes" etc. The following are examples of edges available in ConceptNet: <dog> <CapableOf> <bark>, <mount_olympus> <AtLocation> <greece>.

An earlier version of ConceptNet (version 4) was evaluated for its ability to answer IQ questions using simple test-answering algorithms. The results of this evaluation showed that the system has the Verbal IQ of an average four-year-old child (Ohlsson et al., 2013).

YAGO (Yet Another Great Ontology) is a semantic knowledge base, built from sources like Wikipedia, WordNet (Fellbaum, 2010) and GeoNames[3]. More specifically, information from Wikipedia is extracted from categories, redirects and infoboxes available in each wikipedia page. Also, there are a number of relations between facts that are described in detail in the work of Hoffart et al. (Hoffart et al., 2011). Currently, YAGO contains 447 million facts and about 9,800,000 entities. Facts in YAGO were evaluated by humans, reporting an accuracy of 95%.

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, these relations wasBornIn, diedIn, worksAt place an entity of type Person in a location, e.g., <LeonardCohen> <wasBornIn> <Montreal>.

3 THE GeoMantis SYSTEM

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

First, users select a document and upload it to the system. The document enters the processing pipeline, depicted in Figure 1, and gets processed.

The system uses general-purpose knowledge in the form of facts retrieved from knowledge bases. The GeoMantis knowledge database is populated with facts from ConceptNet and YAGO. These facts, 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 facts from the knowledge bases 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.

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

3.1 Document Input and Processing

The uploaded document is cleaned from HTML tags, wiki specific format and then it is parsed using a Natural Language Processing (NLP) system, the Stanford CoreNLP (Manning et al., 2014). Extracted lemmas, part of speech and named-entity labels extracted by the Named Entities Recognition (NER) process, are stored and indexed in the system's database.

3.2 Knowledge Retrieval

The knowledge retrieval process starts by identifying each country's official name and alternate names from the GeoNames database. Both ConceptNet and YAGO allow knowledge retrieval directly, without the need to download and deploy data locally. ConceptNet 5.5 uses a web accessible API and YAGO 2 can be queried using SPARQL, a query language for RDF (Quilitz and Leser, 2008). The GeoMantis system is capable of integrating with any knowledge base as long as an accessible API is available for retrieving facts.

ConceptNet: For each country name, the ConceptNet API is queried for returning the proper Uniform Resource Identifier (URI) in the database. In ConceptNet, each URI includes the language (e.g., "en") and the term. When the term includes spaces (e.g., "United Kingdom"), these are substituted by underscores.

For each URI, all facts are retrieved in the form of triplets <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 "China" is performed, facts like the ones presented in Figure 2 are returned, which after processing (see Algorithm 1) result to the search strings: pagoda and fungus.

YAGO: YAGO follows a similar format to represent triples e.g., <Arg1> <Relation> <Arg2> and includes both "semantic" (e.g., "wasBornOnDate", "locatedIn" and "hasPopulation") and "technical" relations (e.g., "hasWikipediaAnchorText", "hasCitationTitle").

A similar search for "China" in YAGO returns facts like the ones presented in Figure 3.

The final step in the retrieval workflow, is the processing of facts using the CoreNLP system. Facts are tokenized and lemmatized and common stop words

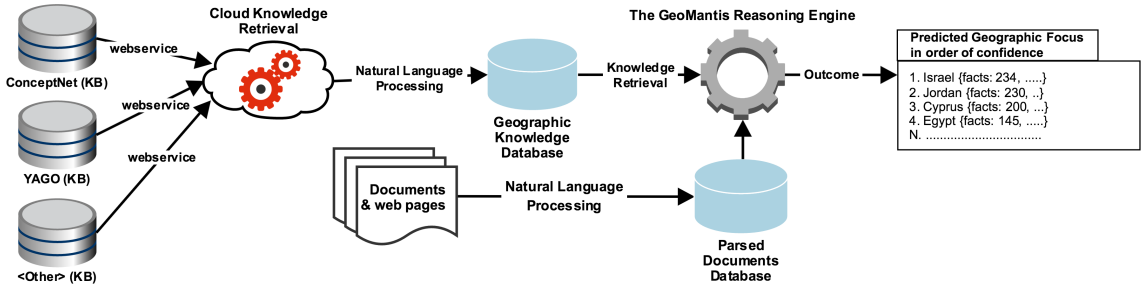

Figure 1: The GeoMantis system processing workflow. The workflow includes the Knowledge Retrieval and Processing mechanism, the Document Processing mechanism and the Reasoning Engine. The outcome of the system appears on the right.

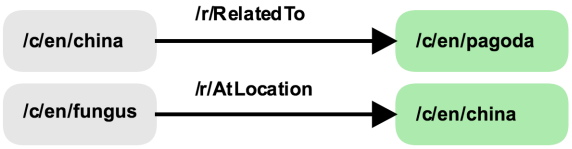

Figure 2: Examples of facts retrieved from ConceptNet using the search term "China".

are removed. For each fact in the system's geographic knowledge base, a search string is created with lemmatized words.

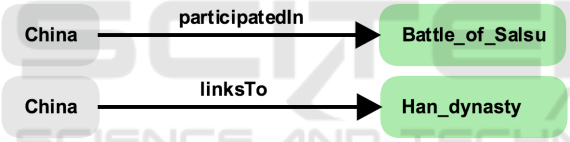

Figure 3: Examples of facts retrieved from YAGO using the search term "China".

3.3 The Reasoning Engine

For each country, a case-insensitive full-text search is executed for each unique word in the document against the search text of each fact in the country's knowledge base. A fact is activated by the text if any of the document's words matches any of the fact's search text words (excluding common stopwords). For example, a document containing the text "They had a really nice dish with halloumi" should activate the rule `<halloumi> <RelatedTo> <Cyprus>`. To maximize the search capabilities, the GeoMantis system uses lemmatized words. Full-text searching takes advantage of the MariaDB's[4] search functionality, using full-text indexing for better search performance.

The final step in the reasoning 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 Facts Applied (PERCR): List of countries is ordered according to the fraction of each country's total number of activated facts over the total number of facts for that country that exist in the geographic knowledge bases, in descending order.

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

Term Frequency - Inverse Document Frequency (TFIDF): List of countries is ordered according to the TF - IDF algorithm (Manning et al., 2008) which is applied as follows:
D_c is a document created by taking the facts 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 Facts per Country Ordering (ORDR): List of countries is ordered according to the number of facts that are retrieved for each country, in descending order.

3.4 Technical 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 that allows the addition of new functionality.

The system, exposes a number of its services using a REST API, based on JavaScript Object Notation (JSON)[5] for data interchange and integration with other systems. Knowledge can be updated at

[4]https://mariadb.org/

[5]http://www.json.org/

Algorithm 1: Knowledge retrieval from knowledge bases.

```
1: procedure RETRIEVEKNOWLEDGE(KB)
   // Use the ISO two-letter country code
2:     for each countryCode do
3:         countryNames ← RetrieveNames(countryCode)
4:         for each countryName do
5:             if (KB = ConceptNet) then
6:                 uri ← RetrieveURI(countryName)
7:                 facts ← RetrieveFacts(uri)
8:                 for each fact do
9:                     arg1 ← GetPart(arg1,fact)
10:                    relation ← GetPart(relation,fact)
11:                    arg2 ← GetPart(arg2,fact)
12:                    if (arg1 = countryName) then
13:                        searchText ← arg2
14:                    else
15:                        searchText ← arg1
16:                    end if
17:                end for
18:            else if (KB = YAGO) then
19:                searchStr ← ValidString(countryName)
20:                facts ← RetrieveFacts(searchStr)
21:                for each fact do
22:                    arg1 ← GetPart(arg1,fact)
23:                    relation ← GetPart(relation,fact)
24:                    arg2 ← GetPart(arg2,fact)
25:                    searchText ← arg2
26:                end for
27:            end if
   // Use NLP to tokenize and lemmatize
28:            searchText ← NLP(searchText)
   // Use a common stopwords list
29:            searchText ← ClearStopWords(searchText)
30:        end for
31:    end for
32:    return SaveGeoDatabase(searchText)
33: end procedure
```

any time by querying the corresponding knowledge source online.

Furthermore, the system has a separate module for producing statistics on documents, datasets, facts and visualizing them using a powerful graph library based on Chart.js library[6]. For each document processed, a detailed log of activated facts is kept for debugging purposes and better understanding of the reasoning process.

4 EXPERIMENTAL EVALUATION

A two phase evaluation was conducted: phase one measured the system's accuracy for each of the strategies in identifying the geographic focus of a document at a country-level, and phase two compared the GeoMantis system using the prevailing strategy from phase one with a freely available opensource system and two common baseline metrics. For these experiments, general-purpose knowledge was retrieved for countries that are members of the United Nations (UN)[7].

Phase one evaluation was conducted using three

[6]http://www.chartjs.org/

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

datasets that were created and used to select the prevailing strategy. For phase two, a fourth independent dataset (see Section 4.4) was used comprising previously unseen documents from the same sources used for phase one.

4.1 Evaluation Datasets

First, pre-tagged document corpora were selected, with metadata of each document's target location. Since all available corpora had explicit mentions of the tagged country, the Reuters Corpus Volume 1 (Lewis et al., 2004) and the New York Times Annotated Corpus (Sandhaus, 2008) were selected and the target country from each document was obscured. Also a smaller dataset with crowd-contributed travel guides was created. In each of the selected datasets, documents were selected from the pool of countries that are members of the UN (193 countries).

The selected datasets (Reuters and New York Times) were chosen because they are among the prevailing datasets for conducting experiments in this line of research. Each story is tagged with location metadata. Moreover, they contain a plethora of stories for experimentation. To the best of our knowledge, there is no dataset that guarantees that there is no mention of the tagged country inside the document. For that reason, we constructed such a dataset to evaluate GeoMantis.

4.1.1 WikiTravel Dataset (WiTr)

This dataset comprises 193 articles about each UN country, retrieved from the Wikitravel website[8] on 25/11/2016. This website uses crowd contributions for building a travel guide for each country. Currently, the site hosts 109,820 pages in English, showcasing numerous places for traveling. All articles were included in the WiTr dataset.

4.1.2 Reuters Corpus Volume 1 (RCV1)

This corpus comprises 810,000 Reuters, English language news stories that were made available in 2000 by Reuters Ltd. Each news story is in English and contains stories from 20/08/1996 to 19/08/1997, tagged with information on where it is geographically located (Lewis et al., 2004). 1000 news stories were chosen at random to create the RCV1 dataset.

4.1.3 The New York Times Annotated Corpus (NYT)

The New York Times Annotated Corpus contains over 1,800,000 articles, written and published by the New York Times between 1987 and 2007. Most articles are tagged with location metadata (Sandhaus, 2008). 1000 news stories from the "Top/News/World/ Countries and Territories/" category with a single country tag were randomly selected to create the NYT dataset. NYT categorization allows a news story to be tagged with more than one locations, but only news stories with a single tag were selected in this case.

[8]http://wikitravel.org

Table 1: Characteristics of the three datasets, including number of documents, number of tagged countries, total and average number of words and NER labels.

Corpus	WiTr	RCV1	NYT
# of documents	193	1000	1000
# of countries	193	110	171
# of words	1,164,783	187,551	378,701
AVG # of words per document	6035	188	379
Named Entities	23.14%	31.61%	25.02%
[location]	8.74%	4.03%	4.84%
[organization]	2.47%	5.92%	3.55%
[money]	0.42%	0.98%	0.37%
[person]	1.31%	6.28%	5.36%

4.2 Evaluation Metrics

For evaluation purposes, two metrics were introduced: the accuracy and the average position. 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 average 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.

4.3 Evaluation of the GeoMantis System Strategies

In this section the results of the first phase of experiments are presented per dataset. Tables 2,3,4 depict the chosen knowledge base (KB), the strategy followed (see Section 3.3), and the applied filtering. Moreover, in Figures 4, 5, 6 we provide a graphical representation of the experiment results. Each graph shows the values of i on the x-axis moving from 1 to 7 and the values on the y-axis, presenting A_i, that is the per-

cent of the correct assignments of the target country in the first i responses of the system.

Filtering options include the use of all words in the document (excluding stopwords) or only words that were labeled as location, person, organization, money by the NER process.

Comparing the results in terms of knowledge base used, knowledge from YAGO presents better results than that of ConceptNet. Further analysis of the two knowledge bases shows a huge gap in the amount of facts retrieved for each country. In particular, YAGO includes 587,458 facts against 99,051 facts in ConceptNet.

The results indicate that the prevailing strategy for all three datasets is **PERCR** when the YAGO knowledge base is used. Furthermore, the experiments show that when only named-entity labeled words are used (**NER** filter), the results are better than when not. Although not reported here, the application of the NER filter also significantly reduces the processing time.

In Tables 2,3,4, rows highlighted in light blue identify the best performing experiments in terms of minimum value for $\bar{P}$.

4.4 Prevailing Strategy Against Other Approaches

In phase two of the evaluation, the GeoMantis system, using the prevailing strategy identified in the first phase of the evaluation (i.e., YAGO, PERCR, NER), is compared with a freely available opensource system, CLIFF-CLAVIN and two common baseline metrics. These metrics include 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.

The comparison was made on a new dataset (EVAL) to avoid possible biases with the datasets used for identifying the prevailing strategy. This dataset comprises 1000 never used before documents, chosen at random from the RCV1 and NYT corpora. In particular, 500 documents where chosen from the RCV1 corpus and 500 from the NYT corpus. For uniformity, from each of the two corpora, two documents were retrieved (if available) for each UN member country at random. The remaining documents were chosen at random from the whole pool of documents.

For conducting the comparison, the CLIFF-CLAVIN geolocation service was set up and a script was used to read the "places/focus/countries" array of the JSON results. Each document in the EVAL dataset was processed as is and with the target country obscured. In particular, the country name was substituted with the word "Unknown", so that the text structure was maintained. For CLIFF-CLAVIN, a new metric U (unanswered) was introduced that denotes the percentage of the number of documents processed without CLIFF-CLAVIN returning a result.

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. 7 was chosen as it corresponds to the maximum number of predicted countries CLIFF-CLAVIN returns when executed on the EVAL dataset and the target country is identified by any one of them.

Results from the second phase evaluation for the GeoMantis system are comparable to that of CLIFF-CLAVIN and that of the two baseline metrics. In cases where the target country is obscured, the GeoMantis system outperforms CLIFF-CLAVIN.

In Table 5, the row highlighted in light green identifies the best results in terms of A_1 and A_7 when the country is obscured. In Figure 7 these results are presented graphically, illustrating all phase two evaluation experiments. By comparing E1 against E2 one can see that the GeoMantis system matches the accuracy of CLIFF-CLAVIN in A_7 eventhough CLIFF-CLAVIN benefits from the visibility of the country in the documents in this experiment.

5 DISCUSSION

Recent advances in Artificial Intelligence bring great promises in the field of text comprehension (Hermann et al., 2015). One of the concerns though, is that artificial intelligence algorithms should provide transparency (Dignum, 2017) on their methods, results and explanations, instead of just leading to opaque black boxes. Following this direction, we focused on designing a system that can be tuned to provide explanations on why a specific geographic focus of a document was chosen, by listing the knowledge facts that led to this result and allowing users to investigate further the reasoning process of the algorithm. This is important for highlighting the explanatory role played by such systems, with respect to the target natural cognitive systems they take as source of inspiration (Lieto and Radicioni, 2016).

Table 2: Results from evaluating the GeoMantis system strategies, KB and filtering options using the WiTr dataset.

#	KB	Strategy	Filter	A_1	A_2	$\bar{P}$
W1	**YAGO**	**PERCR**	**NER**	**71.50**	**90.67**	**1**
W2	YAGO	PERCR	none	56.99	74.09	3
W3	YAGO	NUMR	NER	2.59	9.33	12
W4	YAGO	NUMR	none	0.52	1.55	24
W5	**YAGO**	**TFIDF**	**NER**	**58.55**	**82.90**	**1**
W6	YAGO	TFIDF	none	58.55	76.68	2
W7	YAGO	ORDR	-	0.52	1.04	50
W8	ConceptNet	PERCR	NER	5.70	8.81	12
W9	ConceptNet	PERCR	none	1.55	5.18	14
W10	ConceptNet	NUMR	NER	0.52	2.07	29
W11	ConceptNet	NUMR	none	0.52	1.04	37
W12	ConceptNet	TFIDF	NER	29.02	44.04	4
W13	ConceptNet	TFIDF	none	18.65	27.46	5
W14	ConceptNet	ORDR	-	0.52	1.04	50

Table 3: Results from evaluating the GeoMantis system strategies, KB and filtering options using the RCV1 dataset.

#	KB	Strategy	Filter	A_1	A_2	$\bar{P}$
R1	**YAGO**	**PERCR**	**NER**	**61.80**	**73.60**	**4**
R2	YAGO	PERCR	none	44.70	55.90	8
R3	**YAGO**	**NUMR**	**NER**	**59.20**	**74.20**	**4**
R4	YAGO	NUMR	none	38.40	49.70	8
R5	YAGO	TFIDF	NER	46.00	59.60	7
R6	YAGO	TFIDF	none	31.10	47.00	7
R7	YAGO	ORDR	-	17.60	21.50	23
R8	ConceptNet	PERCR	NER	28.00	39.40	10
R9	ConceptNet	PERCR	none	14.60	21.50	14
R10	ConceptNet	NUMR	NER	30.30	47.80	8
R11	ConceptNet	NUMR	none	21.00	33.80	13
R12	ConceptNet	TFIDF	NER	26.40	37.00	14
R13	ConceptNet	TFIDF	none	14.50	25.10	17
R14	ConceptNet	ORDR	-	17.60	27.90	24

Table 4: Results from evaluating the GeoMantis system strategies, KB and filtering options using the NYT dataset.

#	KB	Strategy	Filter	A_1	A_2	$\bar{P}$
N1	**YAGO**	**PERCR**	**NER**	**37.40**	**58.40**	**7**
N2	YAGO	PERCR	none	24.30	36.30	13
N3	YAGO	NUMR	NER	15.20	29.90	12
N4	YAGO	NUMR	none	3.30	7.80	23
N5	**YAGO**	**TFIDF**	**NER**	**41.30**	**59.30**	**7**
N6	**YAGO**	**TFIDF**	**none**	**22.20**	**39.50**	**7**
N7	YAGO	ORDR	-	2.20	3.90	37
N8	ConceptNet	PERCR	NER	8.70	14.70	14
N9	ConceptNet	PERCR	none	3.80	6.10	20
N10	ConceptNet	NUMR	NER	2.40	8.10	20
N11	ConceptNet	NUMR	none	0.60	4.10	29
N12	ConceptNet	TFIDF	NER	15.60	25.80	15
N13	ConceptNet	TFIDF	none	9.40	15.10	15
N14	ConceptNet	ORDR	-	0.50	1.60	44

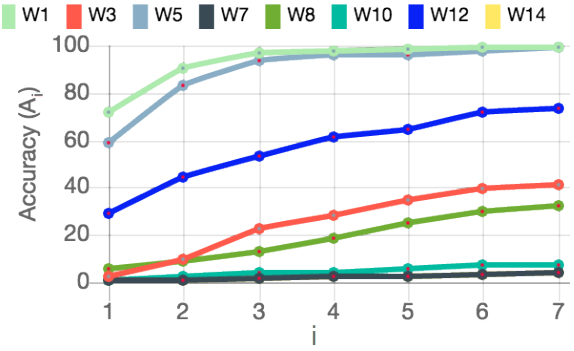

Figure 4: Graphical representation of the results when the WiTr dataset is used. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

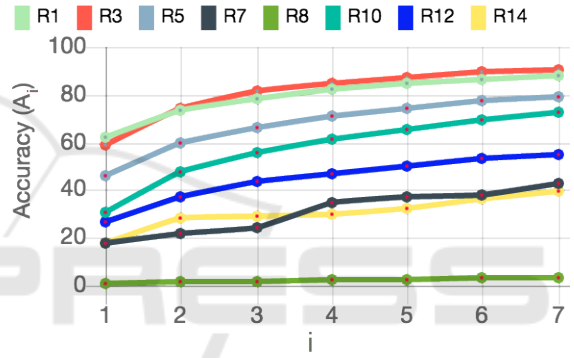

Figure 5: Graphical representation of the results when the RCV1 dataset is used. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

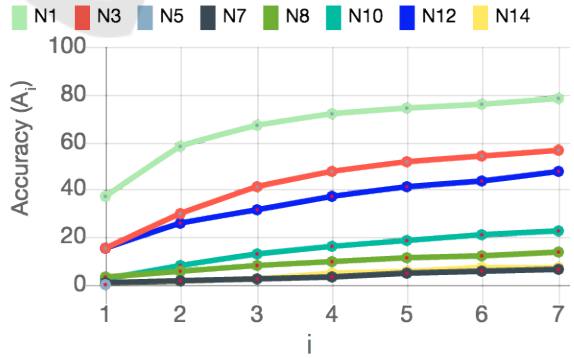

Figure 6: Graphical representation of the results when the NYT dataset is used. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

Table 5: Comparison of the GeoMantis system with CLIFF-CLAVIN and the Baseline. Rows in bold text identify the results that are comparable.

#	Options	$A_1(\%)$	$A_2(\%)$	$A_7(\%)$	$U(\%)$
GeoMantis					
E1	**YAGO, PERCR, NER**	**40.80**	**60.90**	**79.20**	0
CLIFF-CLAVIN					
E2	country visible	72.00	-	80.80	5.60
E3	**country obscured**	**34.40**	**-**	**42.30**	**13.70**
Baseline					
E4	**RAND**	**0.50**	**1.40**	**3.40**	0
E5	**ORDC**	**3.40**	**5.70**	**13.60**	0

Results from the experimental evaluations suggest that the proposed methodology, i.e., using general purpose knowledge bases, is well suited for the problem of inferring geographic focus of documents that do not explicitly mention the target country. The strategy that presents better results is the ordering of the list of countries according to the fraction of each country's total number of activated facts over the total number of facts for that country that exist in the geographic knowledge bases, in descending order (PERCR). Moreover, the usage of the YAGO knowledge base results in greater accuracy than when using the ConceptNet knowledge base.

Despite the existence of other systems for identifying geographic focus (cf. Section 2.1), these systems rely on a geoparser to work and hence the existence of place mentions. This is clearly presented in the experimental evaluation of the CLIFF-CLAVIN system, which in 13.7% of the documents tested did not return a result when the country was obscured. This limitation is waived in GeoMantis that does not rely on place mentions. Comparisons with the other systems were not possible as 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 facts exist in the selected knowledge base. Similar techniques used for news stories could also apply to other types of documents like novels, myths, 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 (Cristani and Tomazzoli, 2016).

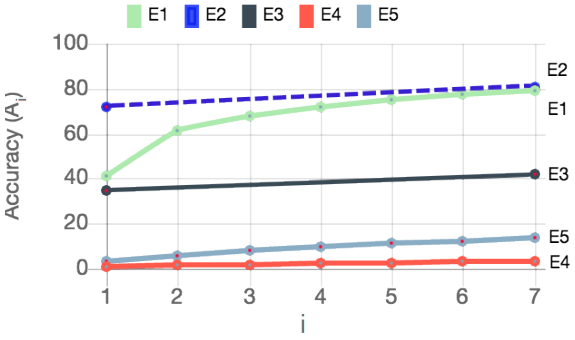

Figure 7: Accuracy comparison graphs between the GeoMantis system, CLIFF-CLAVIN and two baseline metrics. On the x-axis, i gets values from 1 to 7 and the values on the y-axis present A_i, that is the percent of the correct assignments of the target country in the first i responses of the system.

6 CONCLUSION AND FUTURE WORK

This work considered 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. We used general-purpose knowledge bases, instead of gazetteers, atlases or other purposed built geographic bases, to tackle this problem. More specifically, we demonstrated a methodology that retrieves general-purpose knowledge, processes it and infers 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 system tested and the two baseline metrics.

Currently, we are considering extending GeoMantis to utilize paths of various lengths between a geographical entity (e.g., country) and other entities. Figure 8 depicts an example of a length 2 relation path. In such a scenario, if a document contains the word "alps", facts related to `Cyprus` 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 that are conceptually remote (see Figure 8). Safe conclusions can be drawn only after the completion of a systematic experimental evaluation.

Future versions of the system could also benefit from crowdsourcing approaches like GWAPs or hybrid solutions (Rodosthenous and Michael, 2016) for

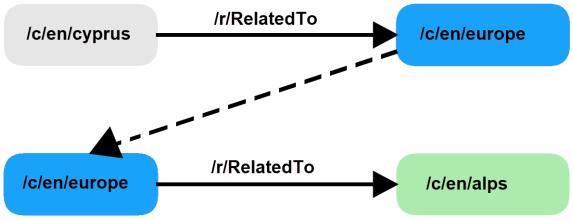

Figure 8: An example of a length 2 relation path from ConceptNet.

fact disambiguation. The integration of other knowledge bases with GeoMantis, like the one generated from the Never Ending Language Learner (Mitchell et al., 2015), DBpedia (Lehmann et al., 2015), Wikidata (Erxleben et al., 2014) or their combination could also be explored.

We believe that the GeoMantis system can be used in several application scenarios, like 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 (like CLIFF-CLAVIN) to return results in cases where the other systems are not able to return any.

REFERENCES

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-Where: Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280.

Andogah, G., Bouma, G., and Nerbonne, J. (2012). Every Document has a Geographical Scope. *Data and Knowledge Engineering*, 81-82:1–20.

Bower, G. H. (1976). Experiments on Story Understanding and Recall. *Quarterly Journal of Experimental Psychology*, 28(4):511–534.

Cristani, M. and Tomazzoli, C. (2016). *A Multimodal Approach to Relevance and Pertinence of Documents*, pages 157–168. Springer International Publishing, Cham.

de Alencar, R. O. and Davis Jr, C. A. (2011). Geotagging Aided by Topic Detection with Wikipedia. In Geertman, S., Reinhardt, W., and Toppen, F., editors, *Advancing Geoinformation Science for a Changing World*, pages 461–477. Springer Berlin Heidelberg, Berlin, Heidelberg.

D'Ignazio, C., Bhargava, R., Zuckerman, E., and Beck, L. (2014). CLIFF-CLAVIN: Determining Geographic Focus for News Articles. In *Proceedings of the NewsKDD: Data Science for News Publishing*.

Dignum, V. (2017). Responsible Autonomy. In *Proceedings of the Twenty -Sixth International Joint Conference on Artificial Intelligence (IJCAI2017)*, pages 4698–4704.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. In *Proceedings of the 13th International Semantic Web Conference*, pages 50–65, Cham. Springer International Publishing.

Fellbaum, C. (2010). WordNet. In Poli, R., Healy, M., and Kameas, A., editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Proceedings of the Neural Information Processing Systems (NIPS 2015)*, pages 1–13.

Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-kelham, E., Melo, G. D., and Weikum, G. (2011). YAGO2 : Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *Proceedings of the 20th International Conference on World Wide Web*, pages 229–232.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., and Others (2015). DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3:5–11.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397.

Lieto, A. and Radicioni, D. P. (2016). From Human to Artificial Cognition and Back: New Perspectives on Cognitively Inspired AI Systems. *Cognitive Systems Research*, 39:1–3.

Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., Surdeanu, M., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *An Introduction to Information Retrieval*, volume 1. Cambridge University Press.

Melo, F. and Martins, B. (2016). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38.

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., and Welling, J. (2015). Never-Ending Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2302–2310.

Monteiro, B. R., Davis, C. A., and Fonseca, F. (2016). A survey on the Geographic Scope of Textual Documents. *Computers and Geosciences*, 96:23–34.

Ohlsson, S., Sloan, R. H., Turán, G., and Urasky, A. (2013). 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*, pages 89–91.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B. (2007). 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):717–745.

Quercini, G., Samet, H., Sankaranarayanan, J., and Lieberman, M. D. (2010). 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*, pages 43–52.

Quilitz, B. and Leser, U. (2008). Querying Distributed RDF Data Sources with SPARQL. In *Proceedings of the 5th European Semantic Web Conference, The Semantic Web: Research and Applications*, pages 524–538, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rodosthenous, C. and Michael, L. (2016). A Hybrid Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium*, pages 111–122.

Sandhaus, E. (2008). The New York Times Annotated Corpus LDC2008T19. DVD. *Linguistic Data Consortium, Philadelphia*.

Speer, R. and Havasi, C. (2013). ConceptNet 5: A Large Semantic Network for Relational Knowledge. In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, pages 161–176. Springer Berlin Heidelberg, Berlin, Heidelberg.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. (2008). NewsStand: A New View on News. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–18.

Tversky, B. (1993). *Cognitive Maps, Cognitive Collages, and Spatial Mental Models*, pages 14–24. Springer Berlin Heidelberg, Berlin, Heidelberg.

von Ahn, L. and Dabbish, L. (2008). Designing Games With a Purpose. *Communications of the ACM*, 51(8):57.

Woodruff, A. G. and Plaunt, C. (1994). GIPSY: Georeferenced Information Processing SYstem. *Journal of the American Society for Information Science*, 45:645–655.

Yu, J. (2016). Geotagging Named Entities in News and Online Documents. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 1321–1330.

Zubizarreta, Á., de La, Cantera, J., Arias, M., Cabrero, J., García, G., Llamas, C., Vegas, J., and Garc, G. (2009). Extracting Geographic Context from the Web: GeoReferencing in MyMoSe. *Advances in Information Retrieval*, pages 554–561.