

# A Crowdsourcing Methodology for Improved Geographic Focus Identification of News-stories

Christos Rodosthenous<sup>1</sup> <sup>a</sup> and Loizos Michael<sup>1,2</sup>

<sup>1</sup>*Open University of Cyprus, Cyprus*

<sup>2</sup>*Research Center on Interactive Media, Smart Systems, and Emerging Technologies, Cyprus*

**Keywords:** Crowdsourcing, Story Understanding, Commonsense Knowledge.

**Abstract:** Past work on the task of identifying the geographic focus of news-stories has established that state-of-the-art performance can be achieved by using existing crowdsourced knowledge-bases. In this work we demonstrate that a further refinement of those knowledge-bases through an additional round of crowdsourcing can lead to improved performance on the aforementioned task. Our proposed methodology views existing knowledge-bases as collections of arguments in support of particular inferences in terms of the geographic focus of a given news-story. The refinement that we propose is to associate these arguments with weights — computed through crowdsourcing — in terms of how strongly they support their inference. The empirical results that we present establish the superior performance of this approach compared to the one using the original knowledge-base.

## 1 INTRODUCTION

In this work we present a crowdsourcing methodology for evaluating how strongly a given argument supports its inference. We apply and evaluate this technique on the problem of identifying the (country-level) geographic focus of news-stories when this focus is not explicitly mentioned in the news-stories.


The text snippet “. . . sitting on a balcony next to the Eiffel Tower, overlooking the City of Light . . .”, for example, is focused on France, but without this being explicitly mentioned in the text. Resolving the focus of the text is of interest both to the story-understanding community for answering the “where” question, and to the information retrieval community when seeking to retrieve documents based on a location that could be only implicit in the retrieved text.

This work builds on our previous work on **GeoMantis** (Rodosthenous and Michael, 2018; Rodosthenous and Michael, 2019), a system that identifies the country-level focus of a text document or a web page using generic crowdsourced knowledge found in popular knowledge-bases and ontologies such as YAGO (Mahdisoltani et al., 2015) and ConceptNet (Speer et al., 2017). The system treats RDF triples from ontologies that reference a particular country as arguments that support that country as being the geo-

graphic focus of a text that triggers that argument. 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. GeoMantis was evaluated on identifying the geographic focus using news-stories where the country of focus was not explicitly mentioned in the text or was obscured. The results of these experiments showed that GeoMantis outperformed two baseline metrics and two systems when tested on the same dataset.

In this work, we extend the GeoMantis methodology and architecture by adding a mechanism to evaluate the arguments retrieved from knowledge-bases using paid crowd-workers recruited from the microWorkers (Nguyen, 2014) platform. The system leverages existing GeoMantis query answering strategies such as number and percentage of applied arguments and the TF-IDF information retrieval algorithm, by adding weights to each argument applied.

In the following sections we first provide a brief overview of the available systems and then we present the updated GeoMantis system highlighting the architecture of the argument evaluation system and how this blends with the original architecture. Next, we describe the evaluation strategies and we describe the experimental setup used to test the hypothesis that arguments evaluated using crowdsourcing can yield better results compared to the results from the original system. In the final section, new features and possible

<sup>a</sup>  <https://orcid.org/0000-0003-2065-9200>

extensions to the GeoMantis system are discussed.

## 2 RELATED WORK

Attempts to identify the geographic focus of texts go back to the 1990's with a system called **GIPSY** (Woodruff and Plaunt, 1994) for automatic georeferencing of text. In the 2000's, the **Web-a-Where** system (Amitay et al., 2004) was developed, which was able to identify a place name in a document, disambiguate it and determine its geographic focus.

Recent attempts include the **CLIFF-CLAVIN** system (D'Ignazio et al., 2014), which is able to identify the geographic focus on news-stories. The system uses a similar to the Web-a-Where system workflow and is able to identify toponyms in news-stories, it disambiguates them and then it determines the focus using the "most mentioned toponym" strategy. Additionally, the "**Newstand**" system (Teitler et al., 2008), retrieves news-stories using RSS feeds and then extracts geographic content using a geotagger.

Related is also the work on the **Mordecai** system (Halterman, 2017), which 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.

Among the most recent systems is **GeoTxt** (Karimzadeh et al., 2019). This is a geoparsing system that can be used for identifying and geolocating names of places in unstructured text. It exploits six named entity recognition systems for its place name recognition process, and utilizes a search engine for the indexing, ranking, and retrieval of toponyms.

## 3 THE GEOMANTIS EXPANDED SYSTEM ARCHITECTURE

GeoMantis is a web application able to identify 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 is processed through the pipeline depicted in Figure 1, with the system returning an ordered list of countries.

To identify the geographic focus, the GeoMantis system uses generic crowdsourced knowledge about countries retrieved from ontologies such as ConceptNet (Speer et al., 2017) and YAGO (Mahdisoltani et al., 2015) represented in the form of RDF triples. An RDF triple  $\langle \text{Subject} \rangle \langle \text{Relation} \rangle \langle \text{Country} \rangle$

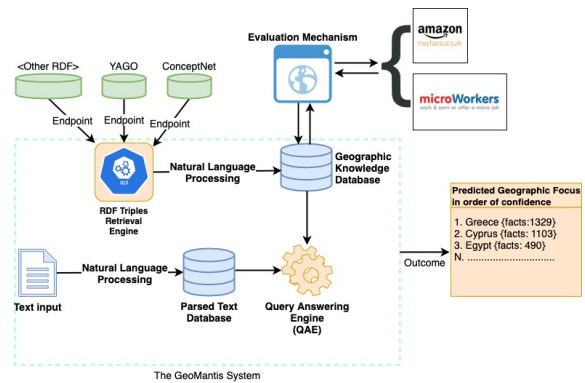


Figure 1: The GeoMantis system architecture. The diagram includes the RDF Triples Retrieval and Processing Engine (top left), the Text Processing mechanism and the Query Answering Engine (QAE). The outcome of the QAE is a predicted list of countries ranked based on confidence.

Name> comprises a Subject that has a relationship Relation with the Country Name, and represents the argument that when the text  $\langle \text{Subject} \rangle$  is included in a given document, then the document is presumably about country  $\langle \text{Country Name} \rangle$ . These arguments are stored locally in the system's geographic knowledge database and then presented to crowdworkers to evaluate them on how useful they are for identifying a specific country. After gathering the evaluations from crowd-workers, the aggregated evaluations are used to add weights to each argument.

An argument is activated when a word from the document exists in the argument's processed text using a full-text search. The argument's text is processed by removing stop words, by removing inflected forms through lemmatization, and by extracting named entities using the CoreNLP system (Manning et al., 2014). Instead of returning a single prediction for the target country, the system returns a list of countries ranked in descending order of confidence.

The system measures the Accuracy of a prediction using a number of metrics  $A_i$ , where  $i \in \{1, 2, 3, \dots, L\}$  and  $L$  is the number of countries in the dataset. The Accuracy  $A_i$  of the system is defined as  $A_i = \frac{F_i}{D}$ , where  $F_i$  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  $D$  denotes the number of available documents in the dataset.

A detailed presentation of the GeoMantis system is available in our previous work (Rodosthenous and Michael, 2018) and interested readers are directed to that paper for more information on the original GeoMantis system architecture, query answering strategies, knowledge retrieval, and processing.

### 3.1 Weighted Query Answering

The original GeoMantis system is able to perform geographic focus identification using three query strategies based on: 1) the number of activated arguments (NUMR), 2) the percentage of activated arguments over the total number of arguments for that country (PERCR), and 3) the TF-IDF algorithm (Manning et al., 2008).

This work extends the above three query strategies to include evaluations received from crowd-workers. The ordering of the list of countries and the generation of the predicted geographic focus is performed using one of the following strategies:

**Weighted Percentage of Arguments Applied ( $PERCR_w$ ):** The list of countries is ordered according to the fraction of each country's total weight of activated arguments over the total weight of arguments for that country that exist in the geographic knowledge bases, in descending order.

**Weighted Number of Arguments Applied ( $NUMR_w$ ):** The list of countries is ordered according to each country's total weight of activated arguments, in descending order.

**Weighted Term Frequency — Inverse Document Frequency ( $TF-IDF_w$ ):** The list of countries is ordered according to the TF-IDF metric, as follows: 1)  $D_c$  is a document created by taking the arguments of a country  $c$ , 2)  $TF_t = (\text{Sum of weights of arguments in } D_c \text{ where term } t \text{ appears}) / (\text{Sum of weights of arguments included in } D_c)$ , 3)  $IDF_t = \log_e(\text{Sum of weights of } D_c / \text{Sum of weights of } D_c \text{ with term } t \text{ in it})$ .

### 3.2 Argument Evaluation System

To evaluate arguments, we designed a mechanism that engages with crowd-workers, presents arguments for evaluation, checks the crowd-workers' confidence in evaluating an argument, and handles payments. The system<sup>1</sup> extends the GeoMantis system using the same technology stack (PHP, mariaDB, javascript).

Workers are presented with detailed instructions on how to evaluate each argument and are requested to provide their microWorkers' ID, their country of origin, and the country for which they feel confident in evaluating arguments. Next, crowd-workers are presented with arguments to evaluate. When they successfully validate all presented arguments, a unique

<sup>1</sup><https://geomantis.ouc.ac.cy/eval.php>

code is presented that each worker can copy into the microWorkers website and receive the payment.

The microWorkers platform<sup>2</sup> was chosen because it includes a large community of crowd-workers, and it offers a mechanism to integrate third-party websites. Each crowd-worker evaluates how useful each argument is on supporting the geographic focus of a specific country. Crowd-workers can choose between three options: "not useful", "I don't know", and "Useful", which are mapped into a -1, 0, and 1 integer values. Although the arguments presented to the crowd-workers are those activated on a given story, the story itself is not presented to the micro-workers so that the evaluation happens on each argument's own merit.

Each crowd-worker is expected to have a basic understanding of the English language since the arguments to be evaluated are presented in English. To improve the quality of the crowdsourced information, and when possible, the system presents arguments that infer the country of the crowd-worker's origin or the crowd-worker's chosen country of confidence.

## 4 EMPIRICAL EVALUATION

In this section we present the empirical evaluation we conducted for testing if the addition of the crowdsourcing evaluation methodology produces better results than the original GeoMantis architecture. The hypothesis that we test is whether knowledge evaluated by the crowd can yield better results in terms of accuracy compared to the results obtained from the original GeoMantis architecture.

We adopt the following high-level methodology: 1) we create a dataset of stories, 2) we identify activated arguments by using GeoMantis, 3) we evaluate those arguments using crowd-workers, and 4) we apply a weighted strategy (cf. Section 3.1) using those arguments on the selected dataset.

### 4.1 Experimental Material

First, we need to select a dataset to perform the experiments. To prepare this dataset, first we take all stories used to test the original GeoMantis architecture (EVAL\_npr dataset). This dataset includes stories in their original form chosen at random from the New York Times Annotated Corpus (Sandhaus, 2008) and the Reuters Corpus Volume 1 (Lewis et al., 2004), where the country of focus is not explicitly present in the story text and it is included in the United Nations list of countries. We then choose stories that have the

<sup>2</sup><https://www.microworkers.com/>

country of focus among the first seven in the order list of identified countries, when an original GeoMantis strategy (PERCR, NUMR, TF-IDF) is applied. Following, we identify all arguments that are activated.

Moreover, four subsets of the dataset are created as follows: 1) Dataset *Crowd\_npr\_1* includes  $N$  stories from the *EVAL\_npr* dataset, where the country of focus is correctly identified in the top position ( $A_1$ ) and  $|A_1 - A_2| > \lambda$ , 2) Dataset *Crowd\_npr\_2* includes  $N$  stories from the *EVAL\_npr* dataset, where the country of focus is correctly identified in the top position ( $A_1$ ) and  $|A_1 - A_2| < \lambda$ , 3) Dataset *Crowd\_npr\_3* includes  $N$  stories from the *EVAL\_npr* dataset, where the country of focus is not correctly identified in the top position ( $A_1$ ) and  $|A_1 - A_2| > \lambda$ , 4) Dataset *Crowd\_npr\_4* includes  $N$  stories from the *EVAL\_npr* dataset, where the country of focus is not correctly identified in the top position ( $A_1$ ) and  $|A_1 - A_2| < \lambda$ , where  $\lambda$  is a threshold.

$A_1$  and  $A_2$  represent the accuracy at position one and two. The above subsets are used for testing if the argument weighting strategy can change the accuracy in both clear and borderline cases of identifying correctly the geographic focus of a story. For example the *Crowd\_npr\_1* subset is characterized by the number of confusing stories it includes, since the threshold for the top identified countries is small.

## 4.2 Preliminary Experiment 1

Before proceeding with the experiment we decided to run a short first experiment to verify our workflow and validate the argument evaluation system. As a first test, we process the *EVAL\_npr* dataset using the PERCR strategy and proceeded to generate the four *Crowd\_X* subsets. For identifying an appropriate threshold  $\lambda$  which will allow a representation of stories from all four subsets we executed a simulation where  $\lambda \in (0.1 - 7.0)$  (heuristically identified) and we selected the  $\lambda$  that allows a maximum inclusion of stories from all 4 datasets. A value of  $\lambda = 1.3$  is selected and hence we have 210 stories for *Crowd\_npr\_1*, 211 stories for *Crowd\_npr\_2*, 74 stories for *Crowd\_npr\_3* and 83 stories for *Crowd\_npr\_4*.

Further analysis reveals that 1,203,518 arguments were activated for 138 countries in the *EVAL\_npr* dataset when processed using the PERCR strategy. For all four *Crowd\_npr* subsets, 1,021,290 arguments were activated from 138 countries.

Since we need only the  $A_1$  and  $A_2$  metrics, we limited the number of activate arguments to a subset of arguments that identify only the countries for  $A_1$  and  $A_2$ . Even then, the amount of arguments was too large to use paid-crowd-workers and we limited the number

of stories to 30, chosen randomly.

We executed a short test experiment to check if the proposed workflow is valid and can be applied in evaluating arguments on all stories. A story from the *Crowd\_npr\_2* subset was selected with all arguments applied to it (total of 357 arguments). An amount of \$0.50 USD was paid to each worker who successfully completed the task.

We launched the evaluation system, where we presented 357 arguments to each worker for evaluation. 30% of the arguments were selected from the country that the worker was confident in contributing in, 40% were selected from the worker's country of origin, 20% were selected from arguments that have at least one evaluation, and 10% were selected in a random order. In case any of the former three categories had no arguments, we then retrieved arguments using random selection. From the total number of presented arguments, 10% is repeated as test (gold) questions used to evaluate the worker's evaluations. This percentage could vary from 10% to 30% (Bragg et al., 2016). More specifically, each worker is required to provide same answers for 10% of the test questions. The contributions of workers who achieved a percentage of less than the defined threshold were not accepted as valid. The threshold could vary from 50% to 70%.

In terms of validity of the worker results, we examined the contributions of the two workers that successfully completed the task. The first worker achieved a score of 16 out of 36 (44.44%) and the second worker achieved a score of 33 out of 36 (91.67%) for the validation questions. We also examined the order of validating the presented arguments. Both workers followed the instructions provided. On average, workers completed the test after 55 minutes and needed 10 seconds per evaluation. The fact that only 2 out of 10 workers completed the task and the amount of time needed to complete the task showed us that we needed to reduce the amount of arguments presented to workers.

At this point there was no need to test the performance of our methodology on the dataset as the purpose of this short test was just to verify that the workflow is valid and identify possible problems with the argument evaluation system.

## 4.3 Preliminary Experiment 2

After testing the argument evaluation mechanism, we expanded the preliminary experiment with 10 stories, taking 3 from subset *Crowd\_npr\_1*, 3 from subset *Crowd\_npr\_2*, 2 from subset *Crowd\_npr\_3*, and 2 from subset *Crowd\_npr\_4*. A total of 5,980 unique argu-

ments were activated for identifying  $A_1$  and  $A_2$ .

We set the following requirements for acceptance of a worker's contribution: 1) A total of 100 arguments should be evaluated, and 2) At least a score of 50% at the validation test should be achieved.

In Table 1 we present information on the experiment and the crowd-workers' contributions. The majority of crowd-workers are in the age group of 26-35, followed by workers in age group 18-25.

The contributed evaluations were used to add weights to all evaluated arguments. More specifically, for each argument we counted the number of positive, negative and neutral feedback. When the sum of negative and neutral feedback was smaller than the sum of positive feedback then we added an integer weight of 600. When they were equal then we added a weight of 0 and when larger we added a weight of 0. We used PERCR<sub>w</sub> and NUMR<sub>w</sub> strategies and the results showed an increase of 20% for both strategies on the accuracy when compared to the original strategies. The TF-IDF strategy was not tested at that time, since it required all arguments to be evaluated, even the ones that were not activated.

#### 4.3.1 Weighting Strategy

The argument weighting strategy used in the preliminary experiment 2 is just one possible strategy of the many that could be used. In this section we present other possible weighting strategies that could be employed, relying on the results of the preliminary experiment. Weights ( $W$ ) are assigned to each of the arguments in the following manner: 1) We assign an a priori weight ( $W$ ) of 1 to each argument, 2) We count all positive feedback, i.e., "Very Confident (1)" ( $F_{pos}$ ), 3) We count all negative feedback, i.e., "Not Very Confident (-1)" ( $F_{neg}$ ), 4) We count all neutral feedback, i.e., "Somewhat Confident (0)" ( $F_{neu}$ )

Nine different strategies were identified based on the observations we made from the preliminary experiments and we present them in the list below:

1) **Strategy  $S_{X,1}$ :** if  $F_{pos} > F_{neg} + F_{neu}$  then  $W=W_p$ , if  $F_{pos} < F_{neg} + F_{neu}$  then  $W=W_n$ , if  $F_{pos} = F_{neg} + F_{neu}$  then  $W=W_{ne}$ ,

2) **Strategy  $S_{X,2}$ :** if  $F_{pos} > F_{neg}$  then  $W=W_p$ , if  $F_{pos} < F_{neg}$  then  $W=W_n$ , if  $F_{pos} = F_{neg}$  then  $W=W_{ne}$ ,

3) **Strategy  $S_{X,3}$ :** if  $F_{pos} + F_{neu} > F_{neg}$  then  $W=W_p$ , if  $F_{pos} + F_{neu} < F_{neg}$  then  $W=W_n$ , if  $F_{pos} + F_{neu} = F_{neg}$  then  $W=W_{ne}$ . Where  $X \in \{1, 2, 3\}$ .

For strategies  $S_{1,1}$ ,  $S_{1,2}$ , and  $S_{1,3}$  we assign both positive ( $W_p = 600$ ) and negative weights ( $W_n =$

$-600$ ) in a symmetrical way and for neutral evaluations the weight of the argument remains intact ( $W_{ne} = 1$ ).

For strategies  $S_{2,1}$ ,  $S_{2,2}$ , and  $S_{2,3}$  we assign positive integer weights ( $W_p = 600$ ) to positive evaluations, for negative evaluations the weight of the argument remains intact ( $W_n = 1$ ) and for neutral evaluations we assign a positive integer weight, less than the one assigned to positive evaluations ( $W_{ne} = 100$ ).

For strategies  $S_{3,1}$ ,  $S_{3,2}$ , and  $S_{3,3}$  we assign positive integer weights ( $W_p = 600$ ) to positive evaluations, negative evaluations are assigned a zero weight ( $W_n = 0$ ) and for neutral evaluations the weight of the argument remains intact ( $W_{ne} = 1$ ).

The value of 600 and 100 were identified heuristically by applying different values of weights in the various strategies and were tested using the query answering strategies during the preliminary experiments.

An additional set of weighting strategies ( $SC_{X,X}$ ) are generated from the selection of arguments that were evaluated by workers who stated in their profile that they originate or are confident in contributing for the same country as the one the argument supports. These weighting strategies follow the same rules as the ones presented earlier and differ only on the source of the arguments.

## 4.4 Experimental Setup

The next step is to launch an experiment to test our hypothesis. We took under consideration the three original GeoMantis strategies and created a broad coverage dataset using the four Crowd<sub>npr</sub> subsets (cf. Section 4.1). More specifically, we selected stories from each of the four subsets of Crowd<sub>npr</sub> using each time one of the 3 strategies, i.e., NUMR, PERCR and TF-IDF to calculate the accuracy. These resulted to the generation of 12 subsets. For each of these 12 subsets, we retrieved stories based on a  $\lambda$  per strategy which maximizes the number of included stories.

Next, we needed to choose a number of news-stories from the Eval<sub>npr</sub> dataset that are unique per subset. For that purpose we designed an automated process that randomly chooses news-stories per strategy that follow the four subsets constraints. The selection process was repeated until all 12 chosen subsets were unique in terms of stories and where that was not possible, we would choose the maximum possible subset. 71 unique stories were chosen and formed the Crowd<sub>npr\_diverse</sub> dataset.

To calculate the arguments activated, we applied the 3 strategies on the Crowd<sub>npr\_diverse</sub> dataset. The amount of arguments activated for these 71 sto-

ries to calculate  $A_1$  and  $A_2$  is 434,562 (178,469 unique). 63% of these arguments was selected and loaded to the crowdsourcing module for evaluation. The acceptance threshold was raised to 70% for the validation test meaning that all contributions below that threshold were not accepted.

#### 4.4.1 Microworkers Platform Setup

In this section we provide insights on the microWorkers platform campaign setup. To start a crowdsourcing task at the microWorkers platform a user first needs to create a campaign. Then we need to choose a group of workers and assign tasks to that particular Group. For our case, we chose the “All International workers” group which included 1,346,882 crowd-workers.

Next we needed to set the TTF (Time-To-Finish), which is the amount of time expected for a worker to complete the task. Based on the results we received from the preliminary experiment 1, it was set at 6 minutes. During the experiment setup we also needed to state the TTR (Time-To-Rate), i.e., the number of days allowed to rate tasks. Choosing a low value is a good incentive for a crowd-worker to perform the task as their payment will be processed earlier than tasks with higher TTR. We set that to 2 days, while the proposed maximum is 7. Next, we set the *Available positions* for the task to 7,180, as this is an estimate of the number of crowd-workers needed to complete this task. Additionally, we added the amount each worker will earn when they successfully complete the task. We chose to pay \$0.20 for each completed task.

The last part of the information needed before launching the campaign is the category of the crowdsourcing task. For our experiment the chosen category is “Survey/Research Study/Experiment” which allows crowd-workers to visit an external site and complete the task. A template also needs to be created with instructions and a placeholder for crowd-workers to enter a verification code when they successfully complete the task.

#### 4.4.2 Crowd-workers Analysis

The experiment was conducted for a total of 24 days, through the microWorkers platform. A total of 8,341 crowd-workers contributed, of which 6,112 (73.28%) provided accepted contributions, i.e., contributions that passed the threshold of 70% at the validation test. For one of the crowd-workers, the contribution time exceeded 23 hours and we removed both the worker and the contributions from the accepted data list, leaving a total of 6,111 crowd-workers with accepted contributions.

The majority of contributors are from Asia. In total, crowd-workers come from 133 countries. Additionally, crowd-workers were confident in contributing for 154 countries. Similar to the country of origin the majority of crowd-workers are confident in contributing in Asian countries and the US. From 6,111 crowd-workers, 4,396 (71.94%) are confident in contributing for their country of origin and 1,715 (28.06%) were confident in contributing to a country other than their country of origin.

Crowd-workers completed a session, i.e., 100 argument evaluations, on average in 6 minutes with  $\sigma = 8$  minutes, and for each argument evaluation they spent on average 4 seconds with  $\sigma = 25$  seconds. Table 1 summarizes the crowd-workers contributions. In terms of age range, 76% of crowd-workers are between 18 and 35 years old.

## 4.5 Experimental Results

After recording the evaluations of the arguments we added weights to each argument using one of the proposed weighting strategies presented in Section 4.3.1. We then used the GeoMantis QAE to predict the geographic focus for the stories in the *Crowd\_npr\_diverse* dataset using the weighted query answering strategies (cf. Section 3.1).

It is important to note that the  $S_{X,X}$  weighting

Table 1: Information and crowd-worker statistics for the 2nd preliminary experiment (2nd column) and for evaluating GeoMantis on the *Crowd\_npr\_diverse* dataset (3rd column).

Number of Stories	10	71
Total number of Arguments to evaluate	5,980	178,469
Number of arguments evaluated	5,980 (100%)	113,219 (63.44%)
Minimum number of evaluators per argument	3	3
Test acceptance percentage	>50%	>70%
Time needed for evaluation	3.8 days	24 days
Number of Crowd-workers (completed contributions)	280	6,796
Number of Crowd-workers (accepted contributions)	217	6,111
Avg time per contribution (completed contributions)	36 min.	6 min.
Avg time per contribution (accepted contributions)	7 min.	6 min.
Avg time per evaluation (completed contributions)	21 sec.	4 sec.
Avg time per evaluation (accepted contributions)	5 sec.	4 sec.
Amount paid per worker	\$0.10	\$0.20

Table 2: Accuracy at  $A_1$  and  $A_2$  when tested on each of the 3 strategies (NUMR<sub>w</sub>, PERCR<sub>w</sub>, TF-IDF<sub>w</sub>) for the Crowd\_npr\_diverse dataset. The left column depicts the system and weighting strategies presented in Section 4.3.1. On the top rows of the table we present the original strategies (GeoMantis v1) and the results when these are applied on the EVAL\_npr dataset and the Crowd\_npr\_diverse dataset, comprising 1000 and 71 stories respectively, for comparison with their weighted versions (GeoMantis v2). In the last 2 rows we present results from two widely used systems when applied on the Crowd\_npr\_diverse dataset.

System	Dataset	NUMR		PERCR		TF-IDF	
		$A_1$	$A_2$	$A_1$	$A_2$	$A_1$	$A_2$
GeoMantis (v1)	EVAL_npr	30.50%	47.30%	43.40%	58.60%	55.40%	68.20%
GeoMantis (v1)	Crowd_npr_diverse	33.80%	56.34%	50.70%	83.10%	84.51%	92.96%
		NUMR <sub>w</sub>		PERCR <sub>w</sub>		TF-IDF <sub>w</sub>	
GeoMantis (v2, $S_{1,1}$ )	Crowd_npr_diverse	43.66%	64.79%	61.97%	84.51%	94.37%	95.77%
GeoMantis (v2, $S_{1,2}$ )	Crowd_npr_diverse	45.07%	64.79%	59.15%	88.73%	94.37%	97.18%
GeoMantis (v2, $S_{1,3}$ )	Crowd_npr_diverse	43.66%	63.38%	53.52%	84.51%	94.37%	97.18%
GeoMantis (v2, $S_{2,1}$ )	Crowd_npr_diverse	42.25%	64.79%	67.61%	90.14%	95.77%	98.59%
GeoMantis (v2, $S_{2,2}$ )	Crowd_npr_diverse	42.25%	63.38%	61.97%	87.32%	95.77%	98.59%
GeoMantis (v2, $S_{2,3}$ )	Crowd_npr_diverse	42.25%	64.79%	60.56%	87.32%	95.77%	98.59%
GeoMantis (v2, $S_{3,1}$ )	Crowd_npr_diverse	42.25%	64.79%	67.61%	90.14%	95.77%	98.59%
GeoMantis (v2, $S_{3,2}$ )	Crowd_npr_diverse	42.25%	64.79%	63.38%	88.73%	95.77%	98.59%
GeoMantis (v2, $S_{3,3}$ )	Crowd_npr_diverse	42.25%	64.79%	60.56%	87.32%	95.77%	98.59%
		$A_1$		$A_2$		$A_7$	
CLIFF-CLAVIN	Crowd_npr_diverse	61.97%		-		74.65%	
Mordecai	Crowd_npr_diverse	56.33%		70.42%		76.06%	

strategies yields better or the same results to the  $SC_{X,X}$  strategies. The  $SC_{X,X}$  strategies use evaluations only from crowd-workers who stated that the originate or are confident in contributing to arguments supporting their stated country.

Furthermore, we applied two widely used systems CLIFF-CLAVIN and Mordecai to identify the geographic focus of news-stories in the Crowd\_npr\_diverse dataset.

The results of the experiment show an improvement on all query answering strategies for the  $S_{X,X}$  weighting strategies, when compared to the original strategies. In Table 2 we present the results of the experiments per weighting strategy and per query answering strategy. The highlighted rows show the best performing weighting strategies, i.e,  $S_{2,1}$  and  $S_{3,1}$ . In terms of the query answering strategies, the best performing strategy is the TF-IDF<sub>w</sub>, followed by PERCR<sub>w</sub> and then NUMR<sub>w</sub>.

When we also compare the results from the updated GeoMantis architecture using the  $S_{2,1}$  and  $S_{3,1}$  strategies and the TF-IDF<sub>w</sub> and PERCR<sub>w</sub> query answering strategies, to that of CLIFF-CLAVIN and Mordecai we observe that our system outperforms both of them. The CLIFF-CLAVIN system, returned an unidentified geographic focus for 16.90% of the news-stories. Moreover, due to the fact that CLIFF-CLAVIN does not order the results and for compar-

ison reasons we calculated  $A_1$  when only one country was returned and it was the correct one and  $A_7$  (where 7 is the maximum number of countries returned for that dataset) when more than one country was returned and the correct one was among them.

## 5 CONCLUSION

We have presented a crowdsourcing methodology for evaluating the strengths of arguments used by the GeoMantis geographic focus identification system, and have shown that this processing of arguments leads to an improved performance.

In a more general context, our work shows that even crowdsourced knowledge, like the one used by the original GeoMantis system, can benefit by a further crowdsourced processing step that seeks to evaluate the validity of the original knowledge.

Our approach is inspired by the work on computational argumentation, where each piece of knowledge is treated as an argument in support of a given inference (Dung, 1995). Adding weights on these arguments is one of several extensions that have been considered in the computational argumentation literature (Dunne et al., 2011). Future versions of the system could benefit further by borrowing ideas from computational argumentation techniques on how conflicting

arguments can be reasoned with, giving rise to additional strategies.

Finally, indirect crowdsourcing approaches, such as the use of Games With A Purpose (Rodosthenous and Michael, 2016), could be used to evaluate existing, or even acquire novel, arguments instead of appealing to paid crowdsourcing (Habernal et al., 2017).

## ACKNOWLEDGEMENTS

This work was supported by funding from the EU's Horizon2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development.

## 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.
- Bragg, J., Mausam, and Weld, D. S. (2016). Optimal Testing for Crowd Workers. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 966–974, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- 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*.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.
- Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., and Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486.
- Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12.
- Halterman, A. (2017). Mordecai: Full Text Geoparsing and Event Geocoding. *The Journal of Open Source Software*, 2(9).
- Karimzadeh, M., Pezanowski, S., MacEachren, A. M., and Wallgrün, J. O. (2019). GeoTxt: A Scalable Geoparsing System for Unstructured Text Geolocation. *Transactions in GIS*, 23(1):118–136.
- 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.
- Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2015). YAGO3: A Knowledge Base from Multilingual Wikipedias. *Proceedings of CIDR*, pages 1–11.
- 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.
- Nguyen, N. (2014). Microworkers Crowdsourcing Approach, Challenges and Solutions. In *Proceedings of the 3rd International ACM Workshop on Crowdsourcing for Multimedia*, page 1, Orlando, Florida, USA. Association for Computing Machinery (ACM).
- Rodosthenous, C. and Michael, L. (2016). A Hybrid Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium*, volume 284, pages 111–122. IOS Press.
- Rodosthenous, C. and Michael, L. (2018). Geomantis: Inferring the geographic focus of text using knowledge bases. In Rocha, A. P. and van den Herik, J., editors, *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, pages 111–121.
- Rodosthenous, C. and Michael, L. (2019). Using generic ontologies to infer the geographic focus of text. In van den Herik, J. and Rocha, A. P., editors, *Agents and Artificial Intelligence*, pages 223–246, Cham. Springer International Publishing.
- Sandhaus, E. (2008). The New York Times Annotated Corpus LDC2008T19. DVD. *Linguistic Data Consortium, Philadelphia*.
- Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, California.
- 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.
- Woodruff, A. G. and Plaunt, C. (1994). GIPSY: Georeferenced Information Processing System. *Journal of the American Society for Information Science*, 45:645–655.