Table of Contents

International Journal of Information Retrieval Research

Volume 7 • Issue 4 • October-December-2017 • ISSN: 2155-6377 • eISSN: 2155-6385

An official publication of the Information Resources Management Association

Editorial Preface

Research Articles

COPYRIGHT

The *International Journal of Information Retrieval Research* is indexed or listed in the following: ACM Digital Library; Bacon's Media Directory; Cabell's Directories; DBLP; Google Scholar; INSPEC; JournalTOCs; MediaFinder; ProQuest Advanced Technologies & Aerospace Journals; ProQuest Computer Science Journals; ProQuest Illustrata: Technology; ProQuest SciTech Journals; ProQuest Technology Journals; Research Library; The Standard Periodical Directory; Ulrich's Periodicals Directory; Web of Science; Web of Science Emerging Sources Citation Index (ESCI)

International Journal of Information Retrieval Research
Volume 7 • Issue 4 • October-December 2017

Geo-Tagging News Stories Using Contextual Modelling

Md Sadek Ferdous, University of Southampton, Southampton, United Kingdom

Soumyadeb Chowdhury, Singapore Institute of Technology, Singapore, Singapore

Joemon M Jose, University of Glasgow, Scotland, United Kingdom

ABSTRACT

With the ever-increasing popularity of Location-based Services, geo-tagging a document - the process of identifying geographic locations (toponyms) in the document - has gained much attention in recent years. There have been several approaches proposed in this regard and some of them have reported to achieve higher level of accuracy. The existing approaches perform well at the city or country level, unfortunately, the performance degrades during geo-tagging at the street/locality level for a specific city. Moreover, these geo-tagging approaches fail completely in the absence of a place mentioned in a document. In this paper, an algorithm is presented to address these two limitations by introducing a model of contexts with respect to a news story. The algorithm evolves around the idea that a news story can be geo-tagged not only using place(s) found in the news, but also using certain aspects of its context. An implementation of the proposed approach is presented and its performance is evaluated on a unique data set where findings suggest an improvement over existing approaches.

KEYWORDS

Contextual Modelling, Evaluation, Geo-Tagging, Information Retrieval, Text Mining

INTRODUCTION

With the ever-increasing popularity of Location-based Services, geo-tagging a document - the process of identifying geographic locations (toponyms) in the document - has gained much attention in recent years. In such services, geographic locations act as the glue that bind together disparate document sets (such as textual contents, images and videos) from multiple data sources. Devices that produce multimedia documents such as images and videos are equipped with the capability to have additional sensors (GPS sensors) that can geo-tag the related document with geographic information such as latitude and longitude and the respective information is stored in a metadata along with the corresponding document. Web services that accumulate such documents (e.g. YouTube and Flickr) can retrieve such information automatically. In addition, such services allow any user to manually tag any multimedia document with geographic locations in cases the documents are not geo-tagged by their capturing devices. Unfortunately, the geo-tagging procedure is rather cumbersome for textual documents and generally relies on manual human input. There have been several works to address

DOI: 10.4018/IJIRR.2017100104

this limitation and some of them have reported to achieve high level of accuracy as reported in (Ding, 2000), (Amitay, 2004), (Garbin, 2005), (Lieberman, 2007), (Andogah, 2012) and (Ignazio, 2014).

As part of a large-scale project, we have been collecting news stories about a country from the country-specific RSS feed of different online news websites on a daily basis for around a year. The main idea is to aggregate this data set with other modes of public data such as social media posts from Twitter; multimedia data from image sharing websites such as Flickr and data from wearable sensors such as lifeloggers and GPS trackers to create a unique multi-modal (textual as well as multimedia) set of data about a particular geographic location. This will encode experiences from multiple user perspectives and has enormous potential in exploiting for public benefit. One of the core challenges for dealing with such heterogeneous set of data is to define the parameters that can be used to link them together for different use-case scenarios. Among several parameters, the spatio-temporal attribute pair is the simplest of choices due to their omni-presence in all our data sets except in news stories.

News stories, mostly textual, are equipped with a temporal attribute (in the form of a timestamp) to highlight the time and date of publication, however, lack any accompanying metadata to publicise the spatial attribute, even though every news generally has a geographic focus in it (Andogah, 2012). The lack of any spatial attribute makes it a challenging task to geo-tag a news story in an automatic fashion. To geo-tag our collection of news stories, we have been looking for publicly available geo-tagging APIs. CLAVIN (CLAVIN, 2016) and CLIFF (Ignazio, 2014) and (CLIFF, 2015) are two such APIs.

After utilising CLAVIN and CLIFF over a subset of our news data set, we have noticed the following shortcomings:

- They fail to geo-tag a document in the absence of direct mentions of a location; and
- They fail to create an association between a fine-grained location and a city in cases a Textual document has been geo-tagged.

What we mean by a fine-grained location is at the granularity of a street or a locality in a city. An example of a locality is Chelsea in London and an example of a street is King’s Road in London, UK. Without a proper association between such fine-grained locations and a city, it opens up the door for disambiguity, since many cities may share the same name for a locality or a street. The reason for our interest in such fine-grained locations is that it allows us to link such news with other data sets, especially lifelogs and GPS trails, which are supplemented with such fine-grained geo-information.

In this paper we investigate the ways the above mentioned problems can be rectified. Especially, we investigate how a mathematical model of context with respect to a news story can be developed and how such a model can be related with a mathematical model of geo-tagging and its algorithmic implementation to rectify such problems.

In particular, we seek answers to the following research questions:

[RQ-1]: Can we develop a mathematical model of context related to news stories and relate with a mathematical model of geo-tagging?
[RQ-2]: Can we exploit the news context model to geo-tag news in the absence of direct mentions?
[RQ-3]: Can we exploit the news context model to geo-tag news at street granularity level considering the disambiguity that may occur?

With this introduction, the paper is organised as follows. We describe the related work in Section: RELATED WORK. Section: GEO-TAGGING MODELLING introduces our mathematical model of context and geo-tagging for a news story. In Section: IMPLEMENTATION, we discuss our implementation along with the algorithm that utilises our model of context for geo-tagging a news story. We describe our evaluation procedure in Section: EVALUATION and present the results in Section: RESULT. We answer our research questions, discuss the advantages and highlight the limitations of the proposed approach in Section: DISCUSSION. We conclude in Section: CONCLUSION.

RELATED WORK

One of the earliest works on geo-tagging textual web resources was reported in (Ding, 2000) where the authors introduced heuristic techniques for automatically detecting the geographical scope(s) within the resource. The techniques relied on the analysis of textual contents and examining the geographical distribution of hyperlinks within the resources. An evaluation of their report was carried out over 150 web resources and more than 75% precision and recall was reported. Finally, a geo-aware search engine was developed using their proposed approach to show the suitability of their approach. The authors mainly focused on the city level granularity and it was not investigated if the approach would be suitable for street/locality level granularity.

An influential work for geo-tagging web documents was presented in (Amitay, 2004). The paper described a data mining approach utilising a gazetteer (an atlas enlisting the names of all places) to locate places mentioned within the document as well as to determine the geographic focus, representing the broader locality such as cities or states, of the document. The authors also discussed mechanisms to resolve two types of ambiguities: geo/non-geo and geo/geo. The first ambiguity depicts the scenarios when a location name is similar to any non-geographic name, e.g. Turkey, whereas the second ambiguity (geo/geo) illustrates the scenarios when places in different countries share the same name, e.g. London, England and London, Canada. Based on the evaluation over 600 web pages, the authors reported a precision of 82% for individual geo-tags and a precision of 91% in determining the geographic focus of the news. Their paper also did not investigate if the approach would be suitable for street/locality level granularity.

One of the major challenges in geo-tagging a document is to handle disambiguity. In this regard, the authors in (Garbin, 2005) presented an approach based on unsupervised machine learning by aggregating two publicly available gazetteers. At first, ambiguous locations were disambiguated automatically by applying preference heuristics which acted as a training data set for the machine learner. Next, the machine learner was used to disambiguate ambiguous locations from other data. Their result of their approach was compared with a human-annotated news corpus containing 7,739 documents with 78.5% precision.

Lieberman et. al. presented a Spatio-Textual search engine called STEWARD which is a system for geo-tagging, determining geographic focus, querying, and visualising geographic locations in text documents in (Lieberman, 2007). The authors utilised a document tagger to extract potential references of locations which are then resolved to geographic locations using two gazetteers. To resolve disambiguity, they formulated an algorithm called Pair Strength Algorithm. The algorithm utilised the frequency count of ambiguous locations, their distance within the document, their geodesic distance and their populations. The document distance is a measurement of offset between a pair of locations from the start of the document. An algorithm called Context-Aware Relevancy Determination (CARD) was formulated to determine the focus of the document. Then, the proposed approach was applied to design and develop a system that visualises the document onto a map allowing document retrieval based on spatio-textual queries. To resolve disambiguity, we have adopted a similar approach as document distance for calculating the distance between a city name and street name (called vicinity score in our approach, see Section: IMPLEMENTATION) mentioned in the news.

With the assumption that every single document in an IR (Information Retrieval) System and the query to retrieve such documents from the system has a geographic focus, Andogah et. al. presented an approach for determining geographic scope of a document (Andogah, 2012). Then, the scope had been utilised for toponym resolution, relevance ranking and query expansion. The geographic scope was determined by extracting named entities using a Named Entity Recognizer (NER) tool. Especially, they exploited the hierarchical structure of the named places and people, particularly political and government leaders, to determine the scope of the document. Then, they employed a heuristic algorithm for resolving ambiguity. Finally, the authors described how their approach could be used for query expansion and relevance ranking. An evaluation was carried over a data set called TR-CoNLL containing 946 documents (Leidner, 2006). They reported an accuracy of 79% over manually annotated articles from the data set for geographic resolution as well as an accuracy of 71% - 80% over manually annotated articles for toponym resolution.

A recent work on geo-tagging the news article was presented in (Ignazio, 2014) where the authors extended an existing geo-tagging API called CLAVIN (CLAVIN, 2016) by applying a few heuristics based on the method described in (Amitay, 2004). Their approach determined the focus of a news article as well as all places mentioned in the article. They reported 95% accuracy over a small manually annotated data set of 75 news and 90% - 91% accuracy in determining focus at the country level using separate 10,000 samples from the New York Times Annotated Corpus (Sandhaus, 2008) and Reuters RCV-1 Corpus respectively (Sandhaus, 2004).

There are several commercial APIs available for geo-tagging textual documents such as news stories, e.g. OpenCalais (OpenCalais, 2016), Placespotter (Placespotter, 2016) and GeoTag (GeoTag, 2016), however, we have been looking for publicly available APIs so that they can be extended to meet our requirements. We have found two such APIs, namely, CLAVIN (CLAVIN, 2016) and CLIFF (CLIFF, 2016). Between these two, CLIFF is based on CLAVIN and has extended CLAVIN's capability. Moreover, it has been reported to achieve better performance than CLAVIN in (Ignazio, 2014). Therefore, we have selected CLIFF for our experiment. CLIFF utilises Stanford NER (StanfordNER, 2016) to extract named entities and then applies a few heuristics to geo-tag the news and to determine its geographic focus. Even though CLIFF can identify a street/locality, it does not associate a street with a city. In addition, all the works discussed above mainly focused either on a country or a city level and did not investigate if their approach would be suitable for street or locality level geo-tagging. Furthermore, their approach would fail in the absence of direct mentions of locations.

Having been inspired by the work of (Lieberman, 2007) and (Ignazio, 2014), we would like to investigate if the mentioned shortcomings can be handled and the effectiveness of geo-tagging can be improved by introducing and exploiting a mathematical model of context and geo-tagging in relation to a news story. To our knowledge, this is the first attempt to formalise a context with respect to a news story and then to apply that context for geo-tagging news stories.

GEO-TAGGING MODELLING

In this section, we define our mathematical model of context of a news story. At first, we define the term Context. Next, we define a model of spatial location information. Finally, we relate contextual information with spatial location information by formally defining the process of geocoding and geo-tagging.

Contextual Information

The term "Context" (also known as contextual information) has been popularised from the domain of Context-aware services. Interestingly, what the term Context means in highly debated and it has been defined in numerous ways. Schilit and Theimer used the term context-aware for the first time in (Schilit, 1994) where they described contexts as locations, identities of nearby people, objects and changes to those objects. Similarly, Ryan et al. regarded contexts as the user's location, environment, identity and time (Ryan, 1997). Hull et al. represent contexts as different aspects of the current situation (Hull, 1997). One of the most accurate and widely-used definitions is given by Abowd et al. where a context has been described as "any information that can be used to characterize the situation of entities (i.e., whether a person, place or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves" (Abowd, 1999).

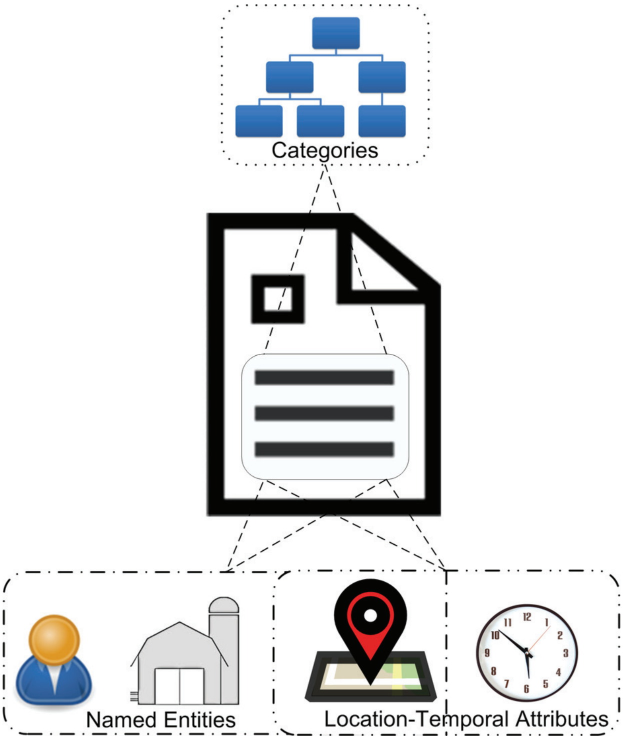
In the domain of Information Retrieval, a context is used to define the state of a user (including the user's spatio-temporal attributes, interests, previous retrieval histories and so on) and such information can be used as a knowledge base in order to achieve higher accuracy during information retrieval (Gross, 2002). Interestingly, in an information retrieval system which mainly deals with news stories (as in ours), how a news story is structured can offer valuable insights, which can be supplemented with the context of a user to achieve highly accurate retrieval results.

A news story outlines a factual story by answering 5 core questions of who, what, where, when and why (Errico, 1997). Of these answers, the answer of who highlights the named entities such as people or organisations mentioned in the news whereas the answers of when and where highlight the temporal and aspatial location information respectively regarding the news. The answers of what and why represent the contents of the news and can be used to classify the news. The combined answers of these questions essentially define the state of a news, much like the way a context defines the state of a user. This motivates us to define the context of a news in the following way:

Definition 1: The context of a news story consists of the named entities and the temporal and aspatial location information mentioned in the news as well as the categories depicting the classification of the news (see Figure 1).

It is important to understand that the location information in such a context merely represents an aspatial locationally descriptive text, usually identified by a Named Entity Recognizer (NER) and hence, is not a valid spatial representation of a location.

Figure 1. Modelling contexts of a news



Mathematically, let P denote the set of people, LN denote the set of aspatial location names, T denote the singleton set depicting a timestamp and O denote the set of organisations. Then, the context of a news can be defined using the following set:

$$INFO_{CONTEXT} = P \cup LN \cup T \cup O$$

where, $P \subseteq P$, $LN \subseteq LN$ and $O \subseteq O$.

Spatial Location Modelling

A spatial location describes the physical location of a location name and is represented using geospatial coordinates such as latitude and longitude (Research, 2016). To model a spatial location, we assume a tree structure where the world is the root of the tree and all countries in the world is its first level children. A country is assumed to be divided in different cities consisting of roads, localities and postcodes. The model is presented below.

Let C denote the set of all countries in the world and $CITY_c$ denote the set of all cities in a country $c \in C$. We define the set of all roads in $city \in CITY_c$ of a country c with R_{city}. Each country defines its own format of a postcode. Without specifying what that format is, we assume that a postcode is assigned to a collection of one or more roads within a specific city. We denote the set of postcodes with PC_{city} for $city \in CITY_c$. To relate a postcode with the corresponding roads, we define the following function.

Definition 2: Let $pcToRoads : PC_{city} \rightarrow P\left(R_{city}\right)$ be the function that returns the set of roads which are part of that postcode in $city \in CITY_c$.

Inversely, we can define another function called, $roadToPC$ which given an input of a road will return the postcode of that road. Formally:

Definition 3: Let $roadToPC : R_{city} \rightarrow PC_{city}$ be the function that returns the postcode of that road in $city \in CITY_c$.

For example, if we assume that the postcode $p \in PC_{city}$ is assigned for roads r_1, r_2 and r_3 in the $city \in CITY_c$, then:

$$pcToRoads\left(p\right) = \left\{r_1, r_2, r_3\right\}$$
$$roadToPC\left(r_1\right) = p,\ roadToPC\left(r_2\right) = p\ and\ roadToPC\left(r_3\right) = p$$

An example of a road along with its postcode is 176 King's Road (depicting the road), SW3 4UP (depicting the postcode) in city London, UK.

Next, we define a locality as the collection of several postcodes within a city and denote the set of localities within a city as LOC_{city}. Like above, we define the following functions to relate a postcode with a locality within a city.

Definition 4: Let $locToPC : LOC_{city} \to P(PC_{city})$ be the function that returns the set of postcodes which are part of that locality in $city \in CITY_c$.

Definition 5: Let $pcToLoc : PC_{city} \to LOC_{city}$ be the function that returns the locality of that postcode in $city \in CITY_c$.

For example, if we assume that a locality $loc \in LOC_{city}$ consists of p_1, p_2 and p_3 postcodes in $city \in CITY_c$, then:

$$locToPC\left(loc\right) = \left\{p_1, p_2, p_3\right\}$$

$$pcToLoc\left(p_1\right) = loc,\ pcToLoc\left(p_2\right) = loc\ and\ pcToLoc\left(p_3\right) = loc$$

An example of a locality in London, UK is Chelsea consisting of several postcodes and one of its postcodes is SW3 4UP. It may happen that for a country or even for a city in a country there is no defined locality. In such cases, the locality set for that city $\left(LOC_{city}\right)$ will represent an empty set.

Finally, we define a location as an ordered pair consisting of a road, a locality, a postcode, a city and a country. Formally, the set of locations for a city $city \in CITY_c$ in country $c \in C$ is denoted as L_{city} and is defined as:

$$L_{city} = \left\{\left(r, loc, pc, city, c\right)\right\}$$

where:

$$r \in R_{city},\ loc \in LOC_{city},\ pc \in PC_{city},\ city \in CITY_c\ and\ c \in C$$

An example of an element of the location set can be given as follows:

$$l = \left(r_1, pcToLoc\left(roadToPC\left(r_1\right)\right), roadToPC\left(r_1\right), city, c\right)$$

where $l \in L_{city}$, $roadToPC\left(r_1\right)$ resolves the rode to the corresponding postcode and $pcToLoc\left(roadToPC\left(r_1\right)\right)$ resolves the road to the postcode which is then resolved to the corresponding locality. An example of a location where a locality is defined is: 176 King’s Road, Chelsea, SW3 4UP, London, UK. Another example of a location where a locality is not defined is: 29 Ethelbert Road, CT1 3NF, Canterbury, UK where Canterbury is a city, not a locality, in the UK.

Then, we can define the set of locations for a country $c \in C$ as the union of the location set for all cities in that country and the universal location set (denoted as L) can be defined as the union of the location set of all countries in the world. Formally:

$$L_C = \cup\{L_{city} \mid city \in CITY_c\}\ and\ L = \cup\{L_C \mid c \in C\}$$

There may exist different ways a spatial information can be represented. In this paper, we assume that a spatial information is represented using a location as defined above and a geographic coordinate as defined next.

A geographic coordinate consists of a latitude and longitude. We denote the set of latitudes as LAT and the set of longitudes as LON. Formally, a geographic coordinate is denoted as $COORD$ and is defined as an ordered pair of latitude and longitude:

$$COORD = \{(lat, lan) \mid lat \in LAT \ and \ lon \in LON\}$$

We denote the set of spatial information as $INFO_{SPATIAL}$ and define it as an ordered pair of location and coordinates in the following way:

$$INFO_{SPATIAL} = \{(B, coord) \mid B \in L \ and \ coord \in COORD\}$$

Geocoding and Geo-Tagging Modelling

To relate contextual information $(INFO_{CONTEXT})$ with spatial information $(INFO_{SPATIAL})$, we define the concept of geocoding. In (Goldberg, 2008), Geocoding is defined as: "the act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process". Formally, we define the geocoding process as a function as follows:

Definition 6: Let $geocoding : C \times (INFO_{CONTEXT} \setminus T) \rightarrow INFO_{SPATIAL}$ be the function that, given the inputs of a country and any contextual information except a timestamp, returns the spatial information for that context within that country.

The country in the input of the $geocoding$ function is used to restrict the geographic focus onto a specific country and is utilised by the geocoding services such as Geocode.Farm API (Geocode, 2016).

As mentioned in Section 1, a geo-tagging process identifies locations within a document. This is merely an informal ambiguous definition as a location information can be aspatial (as defined in $INFO_{SPATIAL}$ or spatial (as defined in $INFO_{CONTEXT}$). A rigorous formal definition, hence, should eliminate such ambiguity. Furthermore, to resonate with the scope of this paper, we mainly focus on geo-tagging news stories. With this goal in mind, we define the geo-tagging process for a news story as a function in the following way where N is the set of news stories:

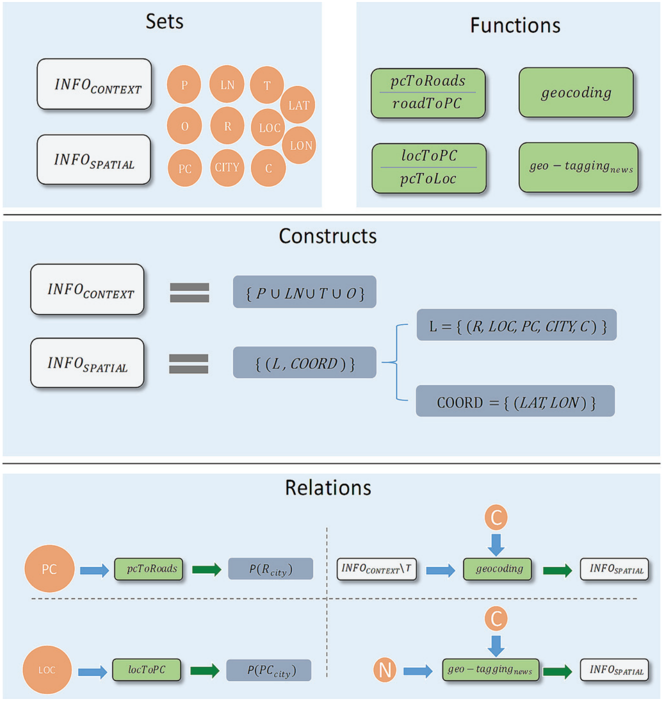
Definition 7: Let $geo-tagging_{news} : N \times C \rightarrow P(INFO_{SPATIAL})$ be the function that, given the inputs of a news and a country, returns the set of spatial location information which are identified in that news and located within that country.

As in Definition 6, the country input in the $geo-tagging_{news}$ function is used to restrict the geographic focus onto a specific country.

SUMMARY

Our model of geo-tagging a news story is summarised in Figure 2. In essence, our model consists of different atomic sets (P, LN, O, R, PC and so on) and combined sets (L_{city}, $INFO_{CONTEXT}$,

Figure 2. Geo-tagging model



$INFO_{SPATIAL}$ and so on) which are used to represent different entities, attributes and geographical properties. For example, $INFO_{CONTEXT}$ encodes the notion of contextual information with respect to a news story whereas $INFO_{SPATIAL}$ represents a spatial location of a city within a country along with its coordinates. The model also consists of different functions that define the inter-relation between the specified sets using mathematical functions. Among all defined functions, $geocoding$ and $geo-tagging_{news}$ are of particular interest. The $geocoding$ function illustrates the concept of converting a context into a spatial locations consisting of the corresponding coordinates. On the other hand the $geo-tagging_{news}$ function illustrates the concept of tagging a news with a spatial information and provides a powerful abstraction that hides away the internal geocoding process.

These two functions, in reality, provide the blue-print for developing algorithms that can be utilised to implement an improved geo-tagger. In the next section, we elaborate how we have achieved these goals.

IMPLEMENTATION

To utilise the model of context, an application, called Geo-Tagger, has been designed and implemented. The application utilises an algorithm, called Geo-Tagging algorithm (see below), which revolves around the idea that there are other aspects of a context, apart from a location, that can be used to geo-tag a news. Especially, we seek to exploit information regarding organisations found in a news story for geo-tagging the story. In such, our idea goes beyond the techniques utilised by existing geo-tagging tools such as CLAVIN and CLIFF where only location information was exploited for geo-tagging. Other aspects of a context such as people and timestamp have not been considered for our current implementation. This is because a timestamp has no geographic location associated with it. Furthermore, even though people, particularly political leaders, have been exploited in geo-tagging a news in (Andogah, 2012), we argue that this is quite tricky and can be susceptible to errors, since the location of a person is not stationary.

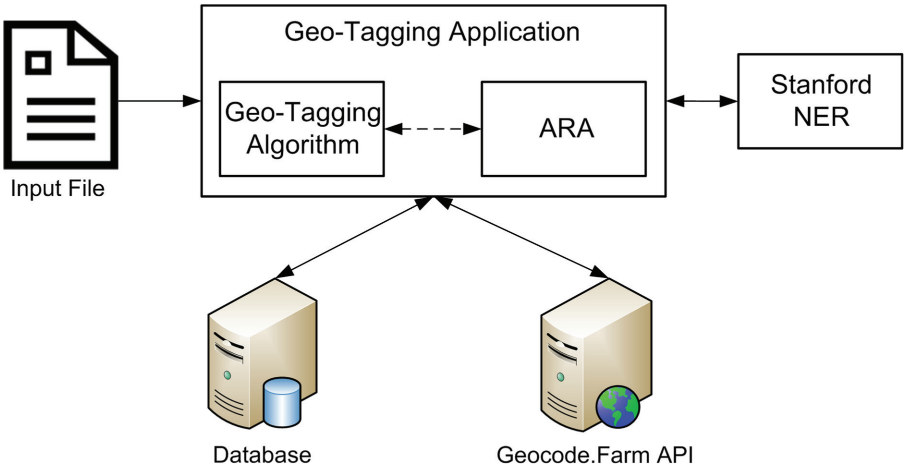
The architecture of the Geo-Tagger is illustrated in Figure 3. The application relies on the following components:

- The Stanford Named Entity Recognizer (NER), which is used to extract named entities and aspatial location information such as locations, persons and organisations within a news (StanfordNER, 2016) representing the set of contextual information $\left(INFO_{CONTEXT}\right)$;
- Geocode.Farm REST API (Geocode, 2016), which has been used to geocode a named entity (considered as an implementation of the $geocoding$ function);
- A database from where a new story is retrieved and to where the result after geo-tagging the news is stored;
- An input file containing bounding box coordinates of the corresponding geographic location; and
- Two algorithms called Geo-Tagging Algorithm and Ambiguity Resolution Algorithm (ARA in short) described below.

Since our news collection (denoted as N) primarily consists of news stories from a specific country, the main focus of the Geo-Tagger is to identify locations within that country. The bounding box coordinates specified in the input file is used to filter out any other locations outside of this bounding box. In this way, the bounding box coordinates in the input file acts as the geographic focus specified by the user and represents a country $c \in C$ c for the $geo-tagging_{news}$ function. This is in contrast with any existing approach where the geographical focus is detected automatically. The advantage of our approach will be discussed in Section: Advantages.

The flow for geo-tagging a news story is as follows. The user of the Geo-Tagger inputs the required bounding box coordinates into the input file. A news is fetched from the database and is passed into the Geo-Tagging algorithm (Algorithm 1) along with the input file. The algorithm processes the file and outputs the locations along with their coordinates. This information is then stored back into the database with a reference of the news, indicating that the corresponding news has been geo-tagged.

Figure 3. Architecture of geo-tagger application



The geo-tagging algorithm (Algorithm 1) essentially is an implementation of the $geo-tagging_{news}$ function. It takes a news $(n \in N)$ and a bounding box of a country (representing $c \in C$). Internally, it exploits the contextual information extracted from named entities using the Stanford NER representing the $INFO_{CONTEXT}$ set. At the first phase, all aspatial locations $(LN \subseteq INFO_{CONTEXT})$ are processed one by one. If such a location presents a city having the coordinates (retrieved using the Geocode.Farm API) within the specified bounding box, the document is geo-tagged with the location. If the location represents either a locality or a street, then the location is fed into the ARA (Algorithm 2). This is because such a location may exist in more than one city within the specified bounding box. The ARA may return a properly disambiguated location and if so, the news is geo-tagged with the location. The ARA may also return a list of locations indicating the locations in the news are still ambiguous. In such cases, the news is geo-tagged with the location along with this ambiguity tag and all matched cities and is stored in the database. At the second phase, all organisations $(O \subseteq INFO_{CONTEXT})$ are extracted from the news using the Stanford NER. The coordinates of each organisation is then retrieved using Geocode.Farm API and if they are within the bounding box, the news is geo-tagged with the location of the organisation.

Algorithm 1. Geo-tagging algorithm

```
Input: a news story $(n \in N)$, input file (representing $c \in C$)
Output: spatial locations identified in the news (a subset of $INFO_{SPATIAL}$)
→ use Stanford NER to extract the set $(INFO_{CONTEXT})$ of named
  entities within the news
→ extract locations $LN \subseteq INFO_{CONTEXT}$
loop for $ln \in LN$
    if $ln$ represents a $city \in CITY_c$ then
        → utilise the Geocode.Farm API by passing $ln$ and $c$
          and retrieve coordinates $(i \in INFO_{SPATIAL})$ of $ln$ within $c$.
        → geotag $n$ with $i$
    else if $\ln \in LOC_{city}$ or $\ln \in R_{city}$ (i.e. $ln$ representing a
     locality or a road in a city) then
        → call ARA passing the location $(ln)$ and the named
          entities $(INFO_{CONTEXT})$ as inputs and retrieve
          coordinates $(i \in INFO_{SPATIAL})$
          if $i \neq null$ then
              → geotag $n$ with $i$
end loop
→ extract the set of organisations locations $O \subseteq INFO_{CONTEXT}$
loop for $o \in O$
    →utilise the Geocode.Farm API by passing $o$ and $c$ and
      retrieve coordinates $(i \in INFO_{SPATIAL})$ of $o$ within $c$
      if $i \neq null$ then
          → geotag $n$ with $i$
end loop
```

Algorithm 2. Ambiguity resolution algorithm

```
Input: an aspatial location $(ln)$, a news $(n \in N)$, named entities
       from the news $(INFO_{CONTEXT})$ and a country $c \in C$
Output: a set of spatial locations $i \subseteq INFO_{SPATIAL}$ containing either
        one or multiple locations
  → retrieve $i' \subseteq INFO_{SPATIAL}$ the set of coordinates for $ln$ in $c$
    using the Geocode.Farm API
loop for $i \in i'$
     if coordinates in $i$ belong to a $city \in CITY_c$ then
          → return $i$ associating $ln$ with $city$
   else if coordinates in $i$ belong to different cities within
    country $c$ then
          → extract city locations $LN \subseteq INFO_{CONTEXT}$
              if only one city is found $(|LN| = 1)$ then
                 → return $i$ associating $ln$ with $city$
              else if more than one match is found $(|LN| > 1)$ then
                 → match between cities extracted in $ln$ and $i$
                    if only a single match is found then
                       → return $i$ associating $ln$ with $city$
                    else if more than one match is found then
                       → determine the vicinity between $ln$ and
                        the matched cities
                       → choose $city$ having the lowest vicinity score
                       → return $i$ associating $ln$ with $city$
                       → if more than one cities having the
                        same vicinity score then
                          → store all cities in a list $(l1)$
              else if no city is found or list $l1$ exists then
                 → extract $O \subseteq INFO_{CONTEXT}$
                 → retrieve locations for each organization
                  using similar heuristic and return i
                  associating $ln$ with $city$
end loop
→return $l1$
```

The Ambiguity Resolution Algorithm (ARA) retrieves the coordinates of a location using the Geocode.Farm API. If the location ($ln \in LN$, mainly a locality or a street) is resolved to only one city having the coordinates within the bounding box, the location is associated with the city and the news is geo-tagged with the location along with its coordinates. If the location is resolved to more than one city, cities residing within the bounding box are selected. Then, aspatial city locations $(LN \subseteq INFO_{CONTEXT})$ from the news are extracted using the Stanford NER. Two lists of cities are matched. If only one match is found, the location is associated with the matched city. If more than one match is found, this means that the location may reside in any of the matched cities. In the next step, a vicinity score between the location and matched cities is calculated. A vicinity score measures the distance between the location and the matched cities mentioned within the news. The distance itself is measured by counting the number of words between the location and the city. The intuition is that a locality or a street is generally mentioned along with its city in the same sentence and/or in the same paragraph. The location is associated with the city having the lowest vicinity score. If more than one city has the same or reasonably close vicinity score, all of them are stored in a list (called $l1$). As the last resort for ambiguity resolution, all organisations $O \subseteq INFO_{CONTEXT}$ are extracted from the news using the Stanford NER. Coordinates for each organisation are retrieved using the Geocode.Farm API. Cities having the coordinates within the bounding box are chosen and are matched against the cities listed in $l1$. If only one match is found, the location is associated with the city and the news is geo-tagged with the location. The reason for this is that often a news having mentioned a locality or a street may discuss about an organisation from the same city. However, this may not be always true and thus may result in more than one match. This means that the location is still ambiguous and hence, a list of ambiguous locations is returned by the algorithm.

EVALUATION

The study with the highest number evaluated documents was reported in (Ignazio, 2014) which utilised New York Times Annotated Corpus (Sandhaus, 2008) and Reuters RCV-1 Corpus respectively (Sandhaus, 2004), both annotated at country and city level. However, the evaluation was not conducted for finer granularity (e.g. street/locality level). To our knowledge, there is no publicly available data set which is annotated at street or locality level. Hence, a comparative evaluation with such data sets was not performed. Instead, we have designed a user study with a smaller data set further discussed in Section: Our Data Set below. This data set was geo-tagged by the Geo-Tagger application and used for the user study. The main goal of the study is to demonstrate the effectiveness of the algorithm as well as to identify its limitations. The data set generation procedure, the user-study and its protocols are presented below.

Reuter Data Set

For a large scale comparison, we compared the country level result identified by Geo-Tagger with the reported result by CLIFF using the Reuter RCV-1 Corpus (Sandhaus, 2004). The RCV-1 corpus consists of over 800,000 news where each news includes a country tag. From this collection, a sample of 10,000 news was randomly selected representing our Reuter data set which were then geo-tagged using CLIFF and the Geo-Tagger.

Our Data Set

Our collection (n) of news stories consists of more than 11, 000 news retrieved from different news websites for around a period of one year, as discussed in Section: INTRODUCTION. At first, we have utilised the CLIFF API to geo-tag all news stories in n generating two data sets: the first set, S_1, consisting of news for which CLIFF has failed to generate any location and the second set, S_2 consisting of news that have been geo-tagged by CLIFF. Then, from S_1, we have chosen 100 random news representing our first evaluation data set (denoted as $EVAL-SET_{100}$) and from S_2 we have selected 200 random news representing our second evaluation data set (denoted as $EVAL-SET_{200}$). Next, our application has been utilised to geo-tag these 300 news in $EVAL-SET_{100}$ and $EVAL-SET_{200}$ which are then used to carry out the following user study.

User Study

The web-based user study consisted of 6 subjects (Female: 2; Male: 4; age range: 25-35 years). The subjects were recruited by sending an email to 6 different research groups. Since $EVAL-SET_{100}$ and $EVAL-SET_{200}$ primarily consisted of news from a specific geographic location, one of the recruitment conditions was that the subject needed to have good familiarity regarding that geographic location. This was to ensure that the subject can identify the location appropriately within that specific region and can associate the street with a city with higher confidence. We received 8 responses who agreed to voluntarily take part in our study. Among them, we chose 6, since the other two subjects have not lived in that geographic location for more than a year. The chosen subjects lived into that geographic location for more than three years, which ensured that they have higher familiarity with the location. Moreover, only 2 subjects were experts in the field of IR (subject area researched in this paper), 2 were Computing Science researchers and 2 were conducting their research in Engineering. We did not choose a crowd sourced evaluation since it would be difficult to ensure the geographical knowledge of the evaluators, which is significant to fulfil the objective of the evaluation.

In order to evaluate the evaluation data sets, we assigned 100 news stories to each subject, which ensured that each story was evaluated by 2 subjects. In our study, each subject was given a link to the web-based study, and an access key. Upon accessing the study, a list of 100 stories was displayed to them. Upon clicking a news story, the title and the news story was displayed, with three questions. The subjects were asked to first read the whole content. Then, they were required to complete the following tasks (i.e. to answer the questions):

T1: If any sort of location information (i.e. name of a city, country, etc.) relevant to the news is present in the news story? T1 helped us to gather statistics related to the number of stories having no location information. These results will be further used to evaluate the effectiveness of our approach, i.e. ability to find a geo-location in the absence of direction mentions in a story.

T2: If a street name (e.g. Jamaica Street) or a locality information (e.g. Chelsea) relevant to the news is present explicitly in the news? T2 gathered statistics related to the number of stories having granular location information (street name, locality etc.). These results will be further used to demonstrate the effectiveness of our approach in absence of direct mentions of a street/locality in a story.

T3: The subjects were shown a list of locations geo-tagged by Geo-Tagger, and asked to choose the irrelevant ones. T3 helped to find erroneous geo-tagged locations. These results will be further analysed to identify the limitation of the proposed algorithm.

Upon completing all tasks, the subjects were contacted to voice their opinions on the relevance of the locations presented to them during the study. The results obtained during the study and their analysis along with the collected user opinions are discussed next.

RESULT

In this section, we present our evaluation results. At first, we present the results of our study where the independent variables are user responses, CLIFF and Geo-Tagger application whereas the dependant variable represents the effectiveness for each independent variable. Then, we present a comparative result between CLIFF and Geo-Tagger over a sample of 5,000 news stories from RCV-1 corpus.

T1 and T2

At first, we analyse the results with respect to T1 and T2. The agreement level (i.e. both subjects having the same opinion) in relation to T1 was 100 and for T2 was 98. The disagreement level in the case of T2 is attributed to the fact that some subjects considered a highway zone (e.g. A390) as a part of the street information, but others did not. We also found slight disagreement with respect to locality. The high agreement level is attributed to our subject demographics (knowledge about location), which indeed made the results obtained from the evaluation valid.

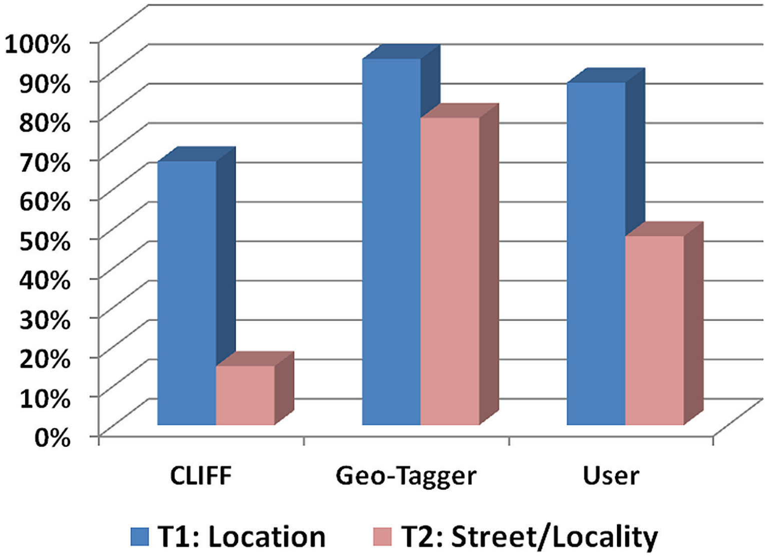
The comparison of CLIFF, Geo-Tagger and User evaluation over the evaluation data set with respect to T1 and T2 is presented in Table 1 along with the corresponding standard deviation (SD column in Table 1) and illustrated in Figure 4. From $EVAL-SET_{100}$, the Geo-Tagger is able to geo-tag 78% of the news stories. This means that for the evaluation data set comprising of 300 news, the success ratio (indicating the total number of news successfully geo-tagged with respect to 300 news) for CLIFF is 67 whereas for the Geo-Tagger, it is 93% - an improvement of 26%.

The results from the Geo-Tagger and the user response with respect to T1 and T2 are also compared. The subjects identified locations in 260 news out of 300 (87%). This is due to the fact that subjects have been able to identify locations from the name of organisations because of their local knowledge. However, they failed to identify as many locations as the Geo-Tagger (87% vs. 93%) since they could not identify the locations of some organisations, e.g. primary schools, pubs, etc. On the other hand, the subjects identified streets/localities in 143 news out of 300 (48%) and hence performed better than CLIFF (15% vs 48%). This is attributed to their local geographical knowledge allowing them to identify many streets or localities. However, the performance of the subjects was still inferior compared to the Geo-Tagger which identified streets/localities in 233 news out of 300 (78%). The reason for this is that the Geo-Tagger, in the absence of a street/locality name, used the names of the organisations to identify streets or localities. Even though the subjects had local knowledge, they could not resolve the street/locality information for some organisations. In the words of one subject:

Table 1. CLIFF vs. geo-tagger vs. user response

	CLIFF	Geo-Tagger	User	SD
T1: Location	67%	93%	87%	14%
T2: Street/Locality	15%	78%	48%	32%

Figure 4. Result plot for T1 and T2



"I did not find any location in the news. I understood that the system identified the location using X Primary School mentioned in the news. However, I have no idea where this school is". Another subject said: "Even if the news did not have a location, I was presented with location names, which seemed relevant to the news, for example, contained the location of the organisation sheriff court presented in the news." Another interesting observation is the comparison of T1 and T2 with respect to Geo-Tagger. Geo-Tagger was able find locations in 93% of news whereas it could find streets/localities in only 78% of news. The reason is that there were many news in which there was no mention of street or localities. In many such scenarios Geo-Tagger was able to identify streets/localities using the locations of the identified organisations in the news. However, there were situations where the Stanford NER could not find any organisation which resulted with news not geo-tagged with streets or localities.

A non-parametric Kruskal-Wallis test is used to examine the significance of the results obtained from the three independent groups (User Responses, Geo-Tagger and CLIFF), in relation to the number of news stories having location information in them (T1). The test results showed significant differences $\left(\chi^2 = 75.172, df = 2, p < 0.001\right)$. A Mann-Whitney test (post-hoc) was conducted to follow up the findings by applying a Bonferroni correction, to report all the effects at a 0.016 level of significance. This correction is used to reduce the chances of obtaining false-positive results (type I errors), when multiple pair-wise tests are performed on a single set of data (especially, when a non-parametric test is used as a post-hoc test). We performed a Bonferroni correction, by dividing the critical p value $\left(\alpha = 0.05\right)$ by the number of comparisons being made (i.e. 3). The post-hoc test results showed significant differences between all pairs $\left(p < 0.001\right)$, except, User Responses and Geo-Tagger ($\left(p = 0.018\right)$). Hence, the statistical test also suggests that the Geo-Tagger is significantly effective (its ability to find locations in a news story) than CLIFF.

A non-parametric Kruskal-Wallis test is used to examine the statistical significance of the results obtained from the three independent groups, in relation to the number of news stories having street locations in them (T2). The test results showed significant differences $\left(\chi^2 = 233.20, df = 2, p < 0.001\right)$. A Mann-Whitney test (post-hoc) was conducted to follow up the findings by applying a Bonferroni correction, to report all the effects at a 0.016 level of significance. The post-hoc test results showed significant differences between all pairs $\left(p < 0.001\right)$. These results suggest that the Geo-Tagger is significantly effective (in finding streets/localities in a news story) than CLIFF.

In summary, the Geo-Tagger has effectively found spatial locations better than CLIFF even in the absence of such information in a news story using our contextual model (Section: GEO-TAGGING MODELLING).

T3

Next, we analyse the results with respect to T3. The agreement level of the subjects in relation to T3 was 95%. The difference in agreement level is attributed to the fact that some subjects chose subset locations as non-relevant, even if this accurately represented the news. For example, locations in a news may include: i) City X and ii) Road A, City X. Our approach presented such locations because we wanted to show the city level scope for a news to help the subjects in identifying irrelevant street/locality information. Moreover, in the case of sports news (e.g. football match between Chelsea Vs Arsenal, both in London, UK), some subjects considered the name of the teams as relevant locations, but the rest did not. Furthermore, some subjects chose all irrelevant options and others did not. This variation on the agreement, we believe, is a common phenomenon since different users will have different subjective opinions regarding how a location can be inferred even in the absence of a direct mention. In addition their opinion will be influenced depending on their familiarity with a particular location.

The Geo-Tagger identified 1118 streets/localities in 300 news including repetitive entries in many locations, meaning a street/locality was found in more than one news. Out of these, 79 streets/localities were tagged as irrelevant by the users which is around 7% of 1118, which was statistically insignificant according to Mann-Whitney test ($p < 0.001$), i.e. demonstrated the effectiveness of the algorithm, indicating an accuracy of around 93% (see Figure 5).

For each story with erroneous streets/localities identified by a subject in T3, the results were further explored to find the nature of errors by the two experts. Most of the errors are attributed to the way Geocode.Farm API geocoded a location name. For example, when America/Holland or any other countries were passed to Geocode.Farm API with a focus of a city, the API resolved the name to a street (e.g. America Street or Holland Street) of that city. We believe this can be corrected by applying an exclusion policy that will exclude the name of a country outside the chosen scope unless the name is succeeded by a reference of street (e.g. America Street) in a news. In addition, the API also could not geocode a few organisations properly and resolved the organisation name into a completely another organisation having a similar name in another city. An example of such an organisation is a supermarket having multiple branches in a city or branches in multiple cities. In the words of another subject: "Sometimes the location was not correct for the organisation. Because I could see the news about city X and contained the reference of organisation A which was completely resolved to a location belonging to city Z". One way to handle this is to restrict the scope during organisation resolution to the city mainly focused in the news. We plan to rectify this problem in future.

Reuter RCV-1

The comparative result is presented in Table 2. As evident from Table 2, Geo-Tagger performed slightly better than CLIFF. This improvement is attributed to the fact that Geo-Tagger exploited

Figure 5. Result plot for T3

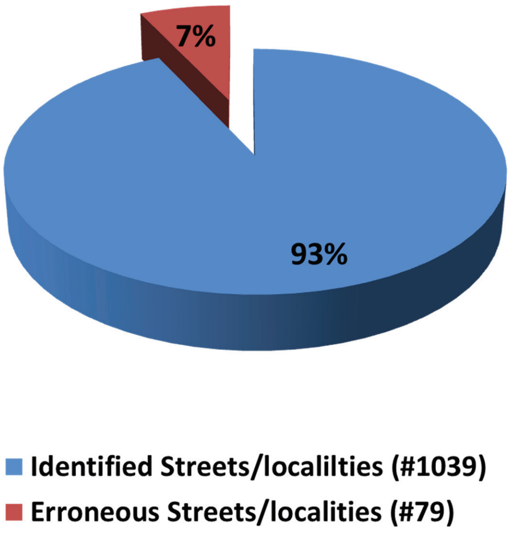


Table 2. CLIFF vs. geo-tagger using RCV-1

Method	Sample Size	Accuracy
CLIFF	10,000	90%
Geo-Tagger	10,000	92%

organisations, in addition to aspatial locations and this helped Geo-Tagger to geo-tag news even in the absence of direct mentions of aspatial locations. The accuracy of Geo-Tagger almost resonate with the results obtained using our data set. This demonstrates the effectiveness of our approach over different data sets of news stories. The accuracy of CLIFF using our Reuter data set is slightly lower than what was reported in (Ignazio, 2014). The exact reason is difficult to establish since the sample distribution in (Ignazio, 2014) and used in this paper will most likely be different even though they originated from the same data set. We could not compare the effectiveness of our approach in finding the street/locality level granularity over this 10,000 data set since the news stories in RCV-1 are not geo-tagged at this granularity and hence it is not reported in this paper. Furthermore, it is important to realise that, even though a smaller data-set was used for carrying out this comparison, it does not invalidate the comparison whatsoever as the same data set has been used for carrying out the comparison between CLLIFF and our approach.

DISCUSSION

In this section, we explore the research questions mentioned in Section: INTRODUCTION, highlight the advantages of our approach and discuss its limitations.

Research Questions

With respect to RQ-1, we have developed a mathematical model of context for a news story (Section: GEO-TAGGING MODELLING) consisting of information that answers 5 core questions of who, what, where, when and why. In this way, the context of a news essentially characterises the news, just like the way a context of a user characterises his/her situation. This is a simple yet powerful model as several of its components such as people, organisations and locations can be utilised for geo-tagging news stories. Moreover, some of its components such as location and category can be easily expanded depending on the granularity of our choice. We have also formalised a model of spatial information and have shown how a contextual information can be related with a spatial information by developing a mathematical model of geocoding and geo-tagging. Our model, in essence, provides a blue-print for developing algorithms to be utilised for the geo-tagging process. We have discussed how we have developed such algorithms utilising our model that encode the geo-tagging process in the implementation section.

The effectiveness of our model becomes apparent when we answer the RQ-2. We have evaluated our geo-tagging procedure over two data sets. The first data set consists of 10,000 news which have been randomly collected from the Reuter RCV-1 Corpus consisting of 800,000 news. The second data set consisting of around 11,000 news retrieved from different news websites for around a period of one year in the UK. The results of our evaluation (Section: RESULT) clearly show that our model of context can be exploited to geo-tag news stories even in the absence of direct mentions of locations (see Figure 1). This is achieved by geo-tagging news stories not only with locations, but also with other aspects of the context model.

As for RQ-3, the results of the evaluation also demonstrate the applicability of exploiting different aspects of the context for geo-tagging news at the street level. The main challenge regarding this is to detect and resolve ambiguity that may occur at this granularity (street/locality). Our approach suggests that it is even possible to resolve locations at the street level even if there is no mention of a street. In addition, our approach has also been able to associate streets with a city both mentioned within a news (see Figure 5).

Advantages

A number of advantages of our approach are highlighted below:

- Our approach enables geo-tagging news stories even in the absence of direct mentions of locations by utilising the contextual information, especially organisations;
- Having the ability to geo-tag news at the street/locality level opens up the opportunity to integrate any news collections with other multi-modal data which are already equipped with such fine-grained location information. The whole data set can then be utilised to create a unique fine-grained spatio-temporal retrieval system;
- Allowing a user of the Geo-Tagger to fixate geographical scopes with the help of an external input file will enable him/her to geo-tag news over the same data set for different scopes at different times. Unlike any existing approach that determines such scopes automatically, this will enable a user to geo-tag news by "zooming in", or "zooming out", within a geographic location for different scenarios. For example, a user can restrict the focus to a particular city within a country by using a particular bounding box coordinates at one time or can restrict the focus over the whole country in another time using another bounding box coordinates.

Limitations

The main limitation of the approach is the inaccuracy of some of the street/locality locations. As it turns out, most of such inaccuracies are attributed to the external geocoding API that is used to resolve location names into geographic coordinates. One possible way to rectify this problem is to combine the results from other geocoding APIs such as Google Place (Google, 2016) with the result of Geocode.Farm API and then apply a policy to resolve the correct location. We plan to incorporate this into our application in future.

Another limitation is the way the user study was carried with a smaller data set and with a small number of subjects. Both are attributed to the following reasons:

- There is no publicly available data set geo-tagged with the granularity of street/locality;
- We had to ascertain that the subjects have good local knowledge within the geographical scope. This ruled out the possibility for a large-scale crowd-sourced evaluation.

CONCLUSION

In this paper, we have developed a mathematical model of context and geo-tagging with respect to news stories and have exploited that model to geo-tag news stories even in the absence of direct mentions of locations as well as at the granularity of street/locality level. For this, we have incorporated our model with a geo-tagging algorithm and utilised off-the-shelf tools and existing Geocoding APIs. The data set generated after applying our approach has been evaluated with 6 users. In addition, we have evaluated our approach over 10,000 news from the Reuter data set. The results demonstrate the effectiveness of our approach against existing publicly available APIs.

In future, we plan to extend the evaluation with a larger sample and subjects, once the approach is improved and then to conduct an experiment with independent style design. At the end, we aim to release the data set containing locations at the granularity of street/localities to the research community for their research. As we have found that different subjects have different opinions regarding a few specific locations (e.g. locations regarding an organisation), it will be interesting to incorporate a confidence level while the users evaluate the approach. Then, results containing a very low confidence level can be specially treated or even excluded during the overall evaluation.

The ultimate goal of our approach is to integrate our news data collection with an array of multi-modal data from different social media such as Flickr, Twitter, YouTube as well as from different wearable sensors such as lifeloggers and GPS trackers and to develop a location-based information fusion system where locations, among other factors, will act as the glue to bind together all data sets. In this regard, the proposed approach will be an essential ingredient of the system. However, we believe that the approach can be adapted for geo-tagging any news stories in any other scenarios with little or no further modifications.

REFERENCES

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999, September). Towards a better understanding of context and context-awareness. *Proceedings of the International Symposium on Handheld and Ubiquitous Computing* (pp. 304-307). Springer. doi:10.1007/3-540-48157-5_29

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004, July). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 273-280). ACM.

Andogah, G., Bouma, G., & Nerbonne, J. (2012). Every document has a geographical scope. *Data & Knowledge Engineering*, *81*, 1–20. doi:10.1016/j.datak.2012.07.002

Berico Technologies. (n. d.). CLAVIN. Retrieved from https://clavin.bericotechnologies.com/

MIT. (n. d.). CLIFF. Center for Civic Media. Retrieved from http://cliff.mediameter.org/

D'Ignazio, C., Bhargava, R., Zuckerman, E., & Beck, L. (2014). Cliff-clavin: Determining geographic focus for news. *Proceedings of KDD '14*.

Ding, J., Gravano, L., & Shivakumar, N. (1999). Computing geographical scopes of web resources.

Errico, M., April, J., Asch, A., Khalfani, L., Smith, M., & Ybarra, X. (1997). *The evolution of the summary news lead*. Media History Monographs-On Line Journal of Media History.

Garbin, E., & Mani, I. (2005, October). Disambiguating toponyms in news. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 363-370). Association for Computational Linguistics. doi:10.3115/1220575.1220621

Geocode.Farm API. (n. d.). Retrieved from https://www.geocode.farm/

Goldberg, D. W. (2008). *A geocoding best practices guide*. Springfield, IL: North American Association of Central Cancer Registries.

Google Place, A. P. I. Retrieved 25 April, 2016, from https://developers.google.com/places/

Gross, T., & Klemke, R. (2002). Context Modelling for Information Retrieval-Requirements and Approaches. Proceedings of ICWI (pp. 247-254).

Hull, R., Neaves, P., & Bedford-Roberts, J. (1997, October). Towards situated computing. Proceedings of the First International Symposium on Wearable Computers (pp. 146-153). IEEE. doi:10.1109/ISWC.1997.629931

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, *30*(4), 400–417. doi:10.1016/j.compenvurbsys.2005.07.003

Lieberman, M. D., Samet, H., Sankaranarayanan, J., & Sperling, J. (2007, November). STEWARD: architecture of a spatio-textual search engine. *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* (p. 25). ACM. doi:10.1145/1341012.1341045

MetaCarta. (n. d.). GeoTag. Retrieved from http://www.metacarta.com/products-platform-geotag.htm

OPEN CALAIS. Thomson Reuters. Retrieved from http://new.opencalais.com

Placespotter. Yahoo. Retrieved from https://developer.yahoo.com/boss/geo/docs/key-concepts.html

Research Data Australia Content Providers Guide. What is spatial location? Retrieved from http://guides.ands.org.au/rda-cpg/spatiallocation

Ryan, N., Pascoe, J., & Morse, D. (1999). Enhanced reality fieldwork: The context aware archaeological assistant. *Bar International Series*, *750*, 269–274.

Sandhaus, E. (2004). Reuters Corpora (RCV1, RCV2, TRC2). Retrieved from http://trec.nist.gov/data/reuters/reuters.html

Sandhaus, E. (2008). The New York Times annotated corpus LDC2008T19. Retrieved from https://catalog.ldc.upenn.edu/LDC2008T19

Schilit, B. N., & Theimer, M. M. (1994). Disseminating active map information to mobile hosts. *IEEE Network*, *8*(5), 22–32. doi:10.1109/65.313011

Stanford Named Entity Recognizer (NER), & the Stanford Natural Language Processing Group. (n. d.). Retrieved from http://nlp.stanford.edu/software/CRF-NER.shtml

Md Sadek Ferdous is a Research Fellow at the University of Southampton, UK. He received his PhD degree in Computing Science from the University of Glasgow, UK in 2015. His research interests include Information Retrieval, Security, Privacy, Identity Management, Trust Management and Blockchain technologies.

Soumyadeb Chowdhury completed his Masters and PhD in Computing Science from University of Glasgow. Employed as a Research Assistant thereafter working in the Integrated Multimedia City Data project in Urban Big Data Center (funded by EPSRC). Currently, lecturer of InfoComm Technology in Singapore Institute of Technology. Research interests include Information Retrieval, Information Security and Privacy, Human Computer Interaction, Pervasive Sensing Technologies and Visualisation Metaphors and Usable Security.

Call for Articles

International Journal of Information Retrieval Research

Volume 7 • Issue 4 • October-December 2017 • ISSN: 2155-6377 • eISSN: 2155-6385
An official publication of the Information Resources Management Association

MISSION

The mission of the **International Journal of Information Retrieval Research (IJIRR)** is to provide an outlet for researchers to present their research and obtain inspiration in the areas of information retrieval, computer science, and information science. Focusing on theories, methods, technologies, and tools, IJIRR is aimed towards information engineers, scientists, and related professionals. This journal exhibits expert experiences and state-of-the-art technologies in search and storage of texts, images, videos, and other data to stimulate innovation and exploration of improved approaches for conquering industry problems.

COVERAGE AND MAJOR TOPICS

The topics of interest in this journal include, but are not limited to:

Advanced software development related information retrieval issues • Classification • Clustering approaches • Content and context awareness and environment awareness • Filtering system • Index techniques • Information mining • Information retrieval in cloud computing issues • Information retrieval in education • Information retrieval in healthcare • Information retrieval in science, engineering, and technologies • Information retrieval in social science, social behaviors • Information retrieval with business, commerce, etc. • Information retrieval with Internet of Things • Knowledge mining • Link analysis • Machine learning on documents • Message passing • Metadata and XML retrieval • Mobile computing related information retrieval issues • Multimedia retrieval • Performance measures • Query languages and optimization • Retrieval architecture • Retrieval evaluation • Retrieval languages and operations • Retrieval strategies • Retrieval systems • Retrieval theories • Retrieval with big data technologies • Scalability • Search algorithms • Search engine • Social media related information retrieval issues • Taxonomy theory and applications • Text mining • Text, document, and image retrieval • Web mining

ALL INQUIRIES REGARDING IJIRR SHOULD BE DIRECTED TO THE ATTENTION OF:

Zhongyu (Joan) Lu, Editor-in-Chief • IJIRR@igi-global.com

ALL MANUSCRIPT SUBMISSIONS TO IJIRR SHOULD BE SENT THROUGH THE ONLINE SUBMISSION SYSTEM:

http://www.igi-global.com/authorseditors/titlesubmission/newproject.aspx

IDEAS FOR SPECIAL THEME ISSUES MAY BE SUBMITTED TO THE EDITOR(S)-IN-CHIEF

PLEASE RECOMMEND THIS PUBLICATION TO YOUR LIBRARIAN

For a convenient easy-to-use library recommendation form, please visit:
http://www.igi-global.com/IJIRR