

KLocator: An Ontology-Based Framework for Scenario-Driven Geographical Scope Resolution

Panos Alexopoulos, Carlos Ruiz, Boris Villazon-Terrazas, and José-Manuel Gómez-Pérez
iSOCO, Intelligent Software Components S.A.
Av. del Partenon, 16-18, 28042, Madrid, Spain
Email: {palexopoulos, cruiz, bvillazon, jmgomez}@isoco.com

Abstract—The automatic extraction of geographical information from textual pieces of information is a challenging task that has been getting increasing attention from application and research areas that need to incorporate location-awareness in their methods and services. In this paper, we present KLocator, a novel ontology-based system for correctly identifying geographical entity references within texts and mapping them to knowledge sources, as well as determining the geographical scope of texts, namely the areas and regions to which the texts are geographically relevant. Compared to other similar approaches, KLocator has two important novelties: i) It does not utilize only background geographical information for performing the above tasks but allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. ii) It is highly customizable, allowing users to define and apply custom geographical resolution models that best fit to the domain(s) and expected content of the texts to be analyzed. Both these features, according to our experiments, manage to substantially improve the effectiveness of the geographical entity and scope resolution tasks, especially in scenarios where explicit geographical information is scarce.

Keywords-Geographical Entity Resolution; Geographical Scope Resolution; Ontologies; Semantic Data.

I. INTRODUCTION

In this paper, we present KLocator a novel ontology-based framework for performing geographical semantic analysis of textual information, in the form of geographical entity and scope resolution. An initial version of the framework has already been presented in [1]; in this paper we extend this work by providing i) a more comprehensive positioning of it in the semantic information processing research landscape, ii) a detailed technical description of the system's implementation and way of use and iii) enhanced experiments with more input data.

In general, our work is related to Geographical Intention Retrieval [2], an area that covers techniques related to the retrieval of information involving some kind of spatial awareness. The goal is to improve services and applications that rely on geographical information, ranging from its quite straightforward use in map services, to more advanced personalization techniques. The main idea is that a text or a query has a geographic scope. For instance, a query for cheap flights from London to Paris would include both

London and Paris in the geographic scope, but not locations in between. Similarly, a text describing the Eiffel tower will have the geographic scope of Paris, rather than of France.

Current geo-location services retrieve likely geographical locations for given keywords or text [3] by mostly applying data mining and statistical techniques on large-scale Web data. Nevertheless, the analysis they perform is primarily a syntactic one, without any exploitation of the text's semantics. The result of this are problems like ambiguity where locations with the same name (Paris, France vs. Paris, Texas) or locations named somehow similar to non-geographic concepts (such as Reading, UK) are not correctly resolved. Thus, semantic analysis, either built on top of statistical analysis or as a standalone approach, can improve current approaches by extracting not only geographical entities from a text, but also other types of entities (people, companies, etc.) that can, via reasoning or inference techniques, improve the accuracy and completeness of the extracted geographical information.

Of course, a bottleneck in applying semantic approaches is the need for geographical knowledge bases as input to the system. Previous approaches have tried to build geographic knowledge on top of different kind of resources, including ad hoc ontologies, geo-gazetteers or more generic knowledge hubs such as Wikipedia. A more promising approach, however, for avoiding or at least limiting the initial entry barriers for geographical semantic analysis is the reuse of Open Data. In particular, the Linked Data initiative [4] provides a crucial starting point for building a large and reliable geographical centered knowledge base, with enough information from other type of entities to allow for a comprehensive coverage of most domains. Moreover, there are some Linked Data initiatives, such as GeoLinkedData [5] and LinkedGeoData [6], that aim to enrich the Web of Data with geographical data.

Given the above, in this paper, we focus on geographical analysis of textual information and we present KLocator, a novel ontology-based framework that focuses on tackling two problems:

- 1) The problem of **geographical entity resolution**, namely the detection within a text of geographical entity references and their correct mapping to ontological

uris that represent them.

- 2) The problem of **geographical scope resolution**, namely the determination of areas and regions to which the text is geographically relevant.

The proposed framework has two distinguishing characteristics. First, unlike other ontology-based approaches which utilize only geographical information for performing the above tasks, it allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. In that way, it manages to significantly improve the accuracy of the above tasks in domains and scenarios where explicit geographical information is scarce.

Second, it is highly customizable as it allows users to define and apply **Geographical Resolution Evidence Models**, based on their knowledge about the domain(s) and expected content of the texts to be analyzed. This allows KLocator to adapt to the particular characteristics of different domains and scenarios and be more effective than other similar systems primarily designed to work in open domain and unconstrained scenarios.

The rest of the paper is as follows. Section II presents related works. Section III describes in detail KLocator's geographical resolution framework while Section IV provides implementation details of the system as well as guidelines on how to use it in practice. Section V presents and discusses experimental results regarding the evaluation of the framework's effectiveness. Finally, a critical discussion of the overall framework is given in Section VI, while conclusions and lines of future work are outlined in Section VII.

II. RELATED WORK

In this section, we describe relevant to our work approaches, both from the area of geographical information processing and from the semantic content analysis one.

A. Geographical-Specific Approaches

The majority of related approaches to our work are found in the area of geographical information retrieval [2], where several approaches based on information retrieval, machine learning or semantic techniques attempt to resolve geographic entities and scope.

Andogah et al. [7] describe an approach to place ambiguity resolution in text consisting of three components; a geographical tagger, a geographical scope resolver, and a placename referent resolver. The same authors, in [8], also propose determining the geographical scope as means to improve the accuracy in relevance ranking and query expansion in search applications. However, these processes only rely on limited geographical information rather than using some other data available.

Following a strict semantic approach, Kauppinen et al. [9] present an approach using two ontologies (SUO - a large Finnish place ontology, and SAPO - a historical and

geographical ontology) and logic rules to deal with heritage information where modern and historical information is available (e.g., new name for a place, new borders in a country). This method is combined with some faceted search functionalities, but they do not propose any method for disambiguating texts.

More related to the fact that the disambiguation of a location depends on the context (such as in "London, England" vs. "London, Ontario"), Peng et al. [10] propose an ontology-based method based on local context and sense profiles combining evidence (location sense context in training documents, local neighbor context, and the popularity of individual location sense) for such disambiguation.

B. Generic Entity and Scope Resolution Approaches

Since geographical entities are just a subset of named entities that are typically considered in the Information Extraction literature (persons and organizations are other examples), more generic named entity resolution approaches may be applied to them.

A recent ontology-based entity resolution approach is described in [11] where an algorithm for entity reference resolution via Spreading Activation on RDF Graphs is proposed. The algorithm takes as input a set of terms associated with one or more ontology elements and uses the ontology graph and spreading activation in order to compute Steiner graphs, namely graphs that contain at least one ontology element for each entity. These graphs are then ranked according to some quality measures and the highest ranking graph is expected to contain the elements that correctly correspond to the entities.

Several approaches utilize Wikipedia as a highly structured knowledge source that combines annotated text information (articles) and semantic knowledge (through the DBpedia [12] and YAGO [13] ontologies). For example, DBpedia Spotlight [14] is a tool for automatically annotating mentions of DBpedia resources in text by using i) a lexicon that associates multiples resources to an ambiguous label and which is constructed from the graph of labels, redirects and disambiguations that DBpedia ontology has and ii) a set of textual references to DBpedia resources in the form of Wikilinks. These references are used to gather textual contexts for the candidate entities from wikipedia articles and use them as disambiguation evidence.

A similar approach that uses the YAGO ontology is the AIDA system [15], which combines three entity disambiguation measures: the prior probability of an entity being mentioned, the similarity between the contexts of a mention and a candidate entity, and the semantic coherence among candidate entities for all mentions together. The latter is calculated based on the distance between two entities in terms of type and subclass of edges as well as the number of incoming links that their Wikipedia articles share.

Other related approaches utilize ontological information for semantically characterizing documents [16] [17] [18]. The first two frameworks assume a categorized ontology, i.e., an ontology whose concepts belong to particular predefined categories (e.g., education, sports, politics, etc.) and, based on the entities found in the document, they compute the categories it belong to through graph similarity measures. On the other hand, the framework of [18] annotates particular segments of the documents with entities derived from a database.

The difference between the above approaches and KLocator is detected in the way they treat the available semantic data. For example, Spotlight uses the DBpedia ontology only as an entity lexicon without really utilizing any of its relations, apart from the redirect and disambiguation ones. Thus, it is more text-based than ontology-based. On the other hand, AIDA builds an entity relation graph by considering only the type and subclass of relations as well as “assumed” relations inferred by the links within the articles. The problem with this approach is that important semantic relations that are available in the ontology are not utilized and, of course, there is no control over which edges of the derived ontology graph should be utilized in the given scenario. Such control is not provided either in [11] or any of the rest aforementioned approaches.

III. GEOGRAPHICAL SCOPE RESOLUTION FRAMEWORK

KLocator facilitates geographical entity and scope resolution in application scenarios where:

- The documents’ domain(s) and content nature are a priori known or can be predicted.
- Comprehensive ontologies covering these domain(s) are available (either purposely built or from existing sources such as Linked Data).

By content nature, we practically mean the types of semantic entities and relations that are expected to be found in the documents. For example, in film reviews one can expect to find films along with directors and actors that have directed them or played in them, respectively. Similarly, in texts describing historical events one will probably find, among others, military conflicts, locations where these conflicts took place and people and groups that participated in them. Documents with known content nature, like the above, can be found in many application scenarios where content is specialized and focused (e.g., reviews, scientific publications, textbooks, reports, etc).

Given such scenarios, our proposed framework targets the two tasks of geographical entity and scope resolution based on a common assumption: that the existence of both geographical and non-geographical entities within a text may be used as **evidence** that indicate which is the most probable meaning of an ambiguous location term as well as which locations constitute the geographical scope of the whole text.

To see why this assumption makes sense, assume a historical text containing the term “Tripoli”. If this term is collocated with terms like “*Siege of Tripolitsa*” and “*Theodoros Kolokotronis*” (the commander of the Greeks in this siege) then it is fair to assume that this term refers to the city of Tripoli in Greece rather than the capital of Libya. Also, in a historical text like “*The victory of Greece in the Siege of Tripolitsa under the command of Kolokotronis was decisive for the liberation from Turkey*”, the evidence provided by “*Siege of Tripolitsa*” and “*Kolokotronis*” and “*Greece*” indicates that Tripoli is more likely to be the location the text is about rather than Turkey.

Now, which entities and to what extent are potential evidence in a given application scenario depends on the domain and expected content of the texts that are to be analyzed. For example, in the case of historical texts we expect to use as evidence historical events and persons that have participated in them. For that reason, our approach is based on the a priori determination and acquisition of the optimal evidential knowledge for the scenario in hand. This knowledge is expected to be available in the form of an ontological knowledge base and it is used within the framework to perform geographical entity and scope resolution. The framework components that enable this are the following:

- A **Geographical Resolution Evidence Model** that contains both geographical and non-geographical semantic entities that may serve as location-related evidence for the application scenario and domain at hand. Each entity is assigned evidential power degrees, which denote its usefulness as evidence for the two resolution tasks.
- A **Geographical Entity Resolution Process** that uses the evidence model to detect and extract from a given text terms that refer to locations. Each term is linked to one or more possible location uris along with a confidence score calculated for each of them. The uri with the highest confidence should be the correct location the term refers to.
- A **Geographical Scope Resolution Process** that uses the evidence model to determine, for a given text, the location uris that potentially fall within its geographical scope. A confidence score for each uri is used to denote the most probable locations.

In the following paragraphs, we elaborate on each of the above components.

A. Geographical Resolution Evidence Model

For the purpose of this paper, we define an ontology as a tuple $O = \{C, R, I, i_C, i_R\}$ where

- C is a set of concepts.
- I is a set of instances.
- R is a set of binary relations that may link pairs of concept instances.

- i_C is a concept instantiation function $C \rightarrow I$.
- i_R is a relation instantiation function $R \rightarrow I \times I$.

Given an ontology, the **Geographical Resolution Evidence Model** defines which ontological instances and to what extent should be used as evidence towards i) the correct meaning interpretation of a location term to be found within the text and ii) the correct geographical scope resolution of the whole text. More formally, given a domain ontology O and a set of locations $L \subseteq I$, a geographical resolution evidence model consists of two functions:

- A **location meaning evidence function** $lmef : L \times I \rightarrow [0, 1]$. If $l \in L$ and $i \in I$ then $lmef(l, i)$ is the degree to which the existence, within the text, of i should be considered an indication that l is the correct meaning of any text term that has l within its possible interpretations.
- A **geographical scope evidence function** $gsef : L \times I \rightarrow [0, 1]$. If $l \in L$ and $i \in I$ then $gsef(l, i)$ is the degree to which the existence, within the text, of i should be considered an indication that l represents the geographical scope of the text.

It is important to note that, though similar in form, these two functions have different meaning and use which, as we show in subsequent sections, is reflected in the way they are calculated and applied. Function $lmef$ is to be used for disambiguation purposes and its values depend primarily on the ambiguity of the evidential entities. On the other hand, $gsef$ is to be used for geographical scope resolution and its values have mostly to do with the number and of the evidential entities.

Both functions are expected to be constructed prior to the execution of the resolution process through a semi-automatic process. To do that, for a given domain and scenario, we need to consider the concepts whose instances are directly or indirectly related to locations and which are expected to be present in the texts to be analyzed. The more domain specific the texts are, the smaller the ontology needs to be and the more effective and efficient the whole resolution process is expected to be. In fact, it might be that using a larger ontology than necessary could reduce the effectiveness of the resolution process.

For example, assume that the texts to be analyzed are about American History. This would mean that the locations mentioned within these texts are normally related to events that are part of this history and, consequently, locations that had nothing to do with these events need not be considered. In that way, the range of possible meanings for location terms within the texts as well as the latter's potential scope is considerably reduced. Therefore, a strategy for selecting the minimum required instances that should be included in the location evidence model would be the following:

- First, identify the concepts whose instances may act as location evidence in the given domain and texts.

- Then, identify the subset of these concepts, which constitute the central meaning of the texts and thus “determine” mostly their location scope.
- Finally, use these concepts in order to limit the number of possible locations that may appear within the text as well as the number of instances of the other evidential concepts.

The result of the above process should be a location evidence mapping function $lem : C \rightarrow R^n$ which given an evidential concept $c \in C$ returns the relations $\{r_1, r_2, \dots, r_n\} \in R^n$ whose composition links c 's instances to locations.

Table I shows such a mapping for the history domain and in particular about that of military conflicts where, for instance, military conflicts provide scope related evidence for the locations they have taken place in and military persons provide evidence for locations they have fought in. The latter mapping, shown in the third row of the table, is facilitated by the chain of two relations: i) the inverse of the relation **dbpprop:commander** that relates persons with battles they have commanded and ii) the relation **dbpprop:place** that relates battles to their locations). In a similar way, one may define a location evidence mapping for the same scenario by, for example, considering the military conflicts mentioned in the text as evidence for the disambiguation of the military persons.

Table I
LOCATION EVIDENCE MAPPING FUNCTION FOR MILITARY CONFLICTS DOMAIN

Evidence Concept	Location Linking Relation(s)
Military Conflict	<i>dbpprop:place</i>
Military Conflict	<i>dbpprop:place, dbpedia-owl:isPartOf</i>
Military Person	<i>is dbpprop:commander of, dbpprop:place</i>
Location	<i>dbpedia-owl:isPartOf</i>

Using this mapping function, we can calculate the location meaning evidence function $lmef$ as follows. Given a location $l \in L$ and an instance $i \in I$, which belongs to some concept $c \in C$ and is related to l through the composition of relations $\{r_1, r_2, \dots, r_n\} \in lem(c)$, we derive the set of locations $L_{amb} \subseteq L$ which share common names with l and are also related to i through $\{r_1, r_2, \dots, r_n\} \in lem(c)$. Then the value of the function $lmef$ for this location and this instance is:

$$lmef(l, i) = \frac{1}{|L_{amb}|} \tag{1}$$

The intuition behind this formula is that the evidential power of a given instance is inversely proportional to the number of different target locations it provides evidence for. If, for example, a given military person has fought in 2 different locations with the same name, then its evidential power for this name is 0.5.

Using the same equation we can also calculate the geographical scope evidence function $gsef$, the only difference being that we consider the set L'_{amb} that contains all the locations related to i , not just the ones with the same name as l :

$$gsef(l, i) = \frac{1}{|L'_{amb}|} \quad (2)$$

Again, the intuition here is that the geographical scope-related evidential power of a given instance is inversely proportional to the number of different locations it is related to. Thus, if the military person of the above example has fought battles in 4 locations in total (independently of whether they share the same name), then its scope-related evidential power would be 0.25.

B. Geographical Entity Resolution

Given a text document and a location meaning evidence function, the detection and disambiguation of the text's locations is performed as follows. First, we extract from the text the set of terms T that match to some $i \in I$ along with a term-meaning mapping function $m : T \rightarrow I$ that returns for a given term $t \in T$ the instances it may refer to. We also consider I_{text} to be the superset of these instances.

Then, we consider the set of potential locations found within the text $L_{text} \subseteq I_{text}$ and for each $l \in L_{text}$ we derive all the instances from I_{text} that belong to some concept $c \in C$ for which $lem(c) \neq \emptyset$. Subsequently, by combining the location evidence model function $lmef$ with the term meaning function m we are able to derive a location-term meaning support function $sup_m : L_{text} \times T \rightarrow [0, 1]$ that returns for a location $l \in L_{text}$ and a term $t \in T$ the degree to which t supports l . If $l \in L_{text}$, $t \in T$ then

$$sup_m(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} lmef(l, i) \quad (3)$$

Using this function, we are able to calculate for a given term $t \in T$ in the text the confidence that it refers to location $l \in m(t)$:

$$cref(t, l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in m(t)} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_m(l, t_j) \quad (4)$$

where $K(l, t) = 1$ if $sup_m(l, t) > 0$ and 0 otherwise.

In other words, the overall support score for a given candidate location is equal to the sum of the location's partial supports (i.e., function sup_m) weighted by the relative number of terms that support it. It should be noted that in the above process, we adopt the one referent per discourse approach, which assumes one and only one meaning for a location in a discourse.

C. Geographical Scope Resolution

The process of geographical scope resolution is similar to the entity resolution one, the difference being that we consider as candidate scope locations not only those found within the text but practically all those that are related to instances of the evidential concepts in the ontology. In that way, even if a location is not explicitly mentioned within the text, it still can be part of the latter's scope.

More specifically, given a text document and a geographical scope evidence function $gsef$ we first consider as candidate locations all those for which there is evidence within the text, that is all those for which $gsef(l, i) > 0$, $l \in L$, $i \in I_{text}$. We call this set L_{cand} . Then, for a given $l \in L_{cand}$ we compute the scope related support it receives from the terms found within the text as follows:

$$sup_s(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} gsef(l, i) \quad (5)$$

Finally, we compute the confidence that l belongs to the geographical scope of the text in the same way as Equation (4) but with sup_s substituting sup_m :

$$c_{scope}(l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in L_{cand}} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_s(l, t_j) \quad (6)$$

where $K(l, t) = 1$ if $sup_s(l, t) > 0$ and 0 otherwise.

IV. SYSTEM IMPLEMENTATION AND USAGE

In this section, we provide details on the technical realization of KLocator and illustrate the way it is meant to be used.

A. System Architecture

The main components of KLocator's architecture, depicted in Figure 1, are the following.

- **Geographical Resolution User Interface:** This interface, depicted in Figure 2 allows users to define and manage their own geographical resolution evidence models and use them to geographically resolve texts.
- **Geographical Resolution Service:** This service layer implements and exposes the required functionality for performing the geographical entity and scope resolution tasks, as described in Section III.
- **Evidence Model Management Service:** This service layer implements and exposes the required functionality for defining, storing and editing geographical resolution evidence models.
- **Evidence Model Repository:** This repository stores all the created evidence models.
- **Evidence Model Manager** This is a low level API for retrieving and manipulating information from the Evidence Model repository.

- **Semantic Data Repository:** This repository stores all the domain and scenario-related ontologies and semantic data that are meant to be used by the system.
- **Semantic Data Manager:** This is a low level API for retrieving and manipulating information from the available ontologies and semantic data that are to be used by the Geographical Resolution Service. At the moment, it is designed to work with locally stored data but, in the future, it will be able to query directly the Linked Open Data Cloud [19].

B. System Usage

The definition and usage of geographical resolution evidence models is performed through the user interface of KLocator. The whole process comprises three steps:

- 1) The user (manually) defines the scenarios's location evidence mapping function by determining the location-related concepts whose instances may serve as contextual disambiguation and scope evidence within his/her scenario's texts.
- 2) The system automatically generates the functions $lmef$ and $gsef$ and stores them for future use.
- 3) The user is then able to apply the model to relevant texts and perform geographical entity and scope resolution.

The execution of the first step starts by pressing the "Create New Evidence Model" button to reveal the model creation form (Figure 3). Then, a name should be given for the new model (e.g., "Locations in Military Conflict Texts") and the table form to be filled with information like that of Table I. In doing that, the user first selects the target concept (e.g., "PopulatedPlace"), then the one to be used as evidence (e.g., "MilitaryConflict") and then the (automatically calculated) relation path between them that we want to consider.

When the model is complete the "Generate Model" button is used to store the model in the server and generate location-evidence entity pairs as in Table II. Depending on the size of the underlying ontology, the generation of these pairs can take a while but it is a process that will need to be performed only once.

When the generation process is finished, the new model appears as an option in the list of defined evidence models and can be used to perform location disambiguation and scope resolution. To do that one needs to select the model and then use the "Input Text" form and the "Perform Geographical Resolution" button to analyze texts relevant to the scenario the model has been defined for. Figure 4 shows the results of executing this process on an example text.

V. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of KLocator, we used it to perform geographical entity and scope resolution on

historical texts describing military conflicts. In particular, we performed two experiments. In the first, we focused on correctly resolving ambiguous location references within the texts while in the second, on correctly determining the texts' geographical scope.

In both cases, we considered DBpedia as a source of semantic information, utilizing a subset of it comprising about 4120 military conflicts, 1660 military persons, 4270 locations and, of course, the relations between them (conflicts with locations, conflicts with persons, etc.). Using this semantic data, we defined the location evidence mapping function of Table I and we used it to automatically calculate the evidential functions $lmef$ and $gsef$ for all pairs of locations and evidential entities (other locations, conflicts and persons).

Table II shows a small sample of these pairs where, for example, James Montgomery acts as evidence for the disambiguation of Beaufort County, South Carolina because he has fought a battle there. Moreover, his evidential power for that location is 0.5, practically because he has fought a battle in another location called Beaufort County. Similarly, Pancho Villa acts as evidence for the consideration of Columbus, New Mexico as the scope of a text (because he has fought a battle there) and his evidential power for that is 0.2 since, according to the ontology, has fought battles in 4 other locations as well.

Table II
EXAMPLES OF LOCATION EVIDENTIAL ENTITIES

Location	Evidential Entity	lmef	gsef
Columbus, Georgia	James H. Wilson	1.0	0.17
Columbus, New Mexico	Pancho Villa	1.0	0.2
Beaufort County, South Carolina	James Montgomery	0.5	0.5

Using this model, we first applied our proposed geographic entity resolution process in a dataset of 150 short texts describing military conflicts like the following: "The Siege of Augusta was a significant battle of the American Revolution. Fought for control of Fort Cornwallis, a British fort near Augusta, the battle was a major victory for the Patriot forces of Lighthorse Harry Lee and a stunning reverse to the British and Loyalist forces in the South". The choice of this domain and scenario was driven from the fact that it has the key characteristics of the application scenarios our framework is designed for, namely predictability of text content and available background ontological knowledge.

The texts were manually compiled from web resources, including Wikipedia and other history-related pages. They were, in average, 2-4 sentences long, all contained ambiguous location entities but little other geographical information and, in average, each ambiguous location reference had 2.5 possible interpretations. For each such reference, 2 human

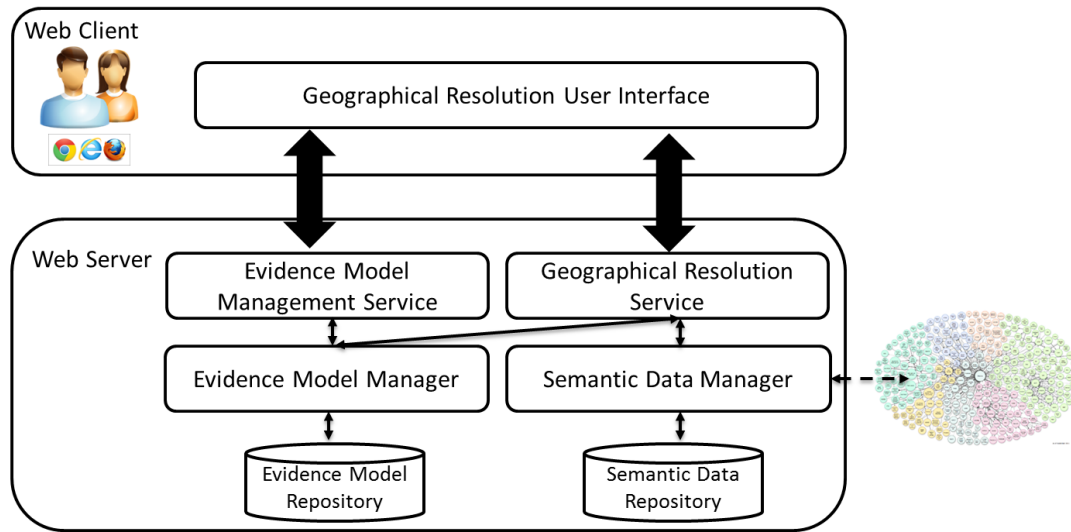


Figure 1. High Level Overview of KLocator Architecture

judges identified the correct location it referred to, with a very high inter-agreement of 0.9.

Then, we used KLocator to perform the same task automatically and we measured the precision and recall of the process. Precision was determined by the fraction of correctly interpreted locations (i.e., locations for which the interpretation with the highest confidence was the correct one) to the total number of interpreted locations (i.e., terms with at least one interpretation). Recall was determined by the fraction of correctly interpreted locations to the total number of annotated locations in the input texts. It should be noted that all target locations for disambiguation in the input texts were known to the knowledge base (i.e., DBPedia).

Table III shows results achieved by our approach compared to those achieved by some well-known publicly available semantic annotation and disambiguation services, namely DBPedia Spotlight [20], AIDA [21] [15], Wikimeta [22], Zemanta [23], AlchemyAPI [24] and Yahoo! [25]. As one can see, the consideration of non-geographical semantic information that our approach enables, manages to significantly improve the effectiveness of the geographical entity resolution task.

Of particular significance is the improvement achieved over DBPedia Spotlight and AIDA as these two systems i) also use DBPedia as a knowledge source and ii) they provide some basic mechanisms for constraining the types of entities to be disambiguated, though not in the same methodical way as our framework does. Practically, the two systems merely provide the users the capability to select the classes whose instances are to be included in the process.

In all cases, it should be made clear that the goal of this comparison was not to disprove the effectiveness and value of these systems as tools for open domain and unconstrained situations but rather to illustrate the importance of

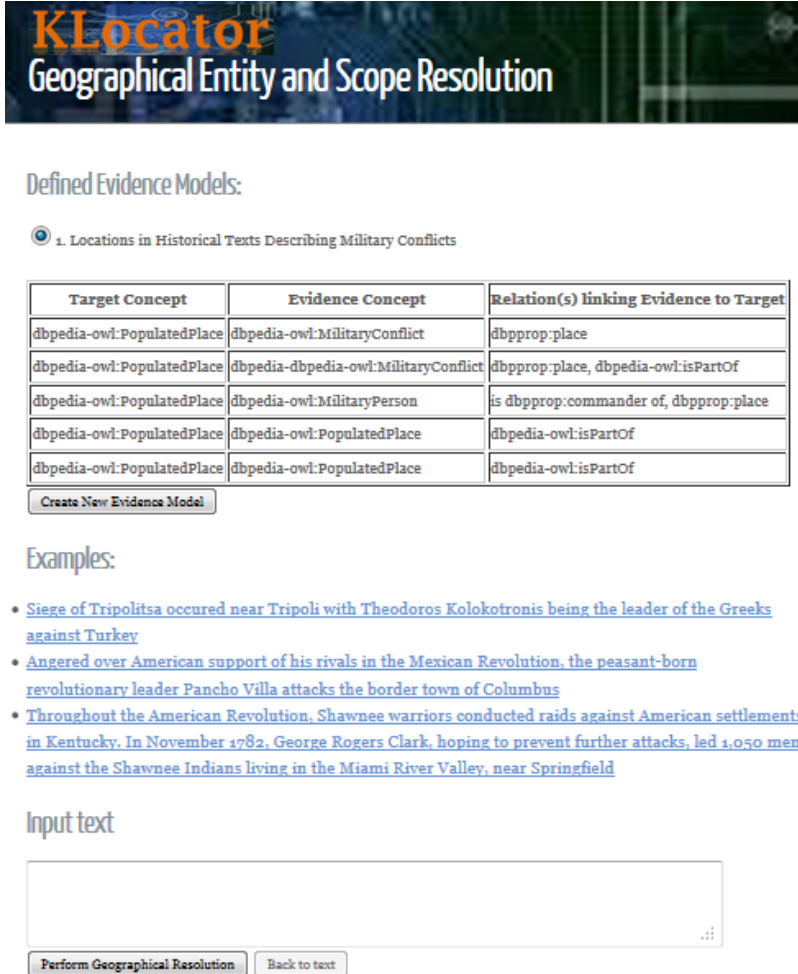
customization and verify our claim that our approach is more appropriate for disambiguation in “controlled” scenarios, i.e., scenarios in which a priori knowledge about what entities and relations are expected to be present in the text is available. Of course, the availability of comprehensive background semantic knowledge about the domain is also an important effectiveness factor, but this is a requirement for any relevant system that follows a knowledge-based approach. A useful evaluation of popular semantic entity recognition systems for open scenarios may be found at [26].

Table III
GEOGRAPHICAL ENTITY RESOLUTION EVALUATION RESULTS

System/Approach	Precision	Recall	F_1 Measure
Proposed Approach	88%	83%	85%
DBPedia Spotlight	71%	69%	70%
AIDA	44%	40%	42%
Wikimeta	33%	30%	31%
Zemanta	26%	30%	28%
AlchemyAPI	26%	28%	27%
Yahoo!	24%	26%	25%

As a second experiment, we applied our proposed geographic scope resolution process in two different datasets, all comprising 150 short military conflict related texts but with different characteristics. The first dataset comprised texts whose geographical scope was not explicitly mentioned within them and which contained little other geographical information. The second dataset comprised texts whose geographical scope related locations were explicitly and unambiguously mentioned within them but along with other geographical entities that were not part of this scope.

In both cases, we used again 2 human judges who decided the scope location of the texts, with an inter-agreement of 0.85. The we used KLocator to automatically determined for each text the possible locations that comprised its



KLocator
Geographical Entity and Scope Resolution

Defined Evidence Models:

1. Locations in Historical Texts Describing Military Conflicts

Target Concept	Evidence Concept	Relation(s) linking Evidence to Target
dbpedia-owl:PopulatedPlace	dbpedia-owl:MilitaryConflict	dbpprop:place
dbpedia-owl:PopulatedPlace	dbpedia-dbpedia-owl:MilitaryConflict	dbpprop:place, dbpedia-owl:isPartOf
dbpedia-owl:PopulatedPlace	dbpedia-owl:MilitaryPerson	is dbpprop:commander of, dbpprop:place
dbpedia-owl:PopulatedPlace	dbpedia-owl:PopulatedPlace	dbpedia-owl:isPartOf
dbpedia-owl:PopulatedPlace	dbpedia-owl:PopulatedPlace	dbpedia-owl:isPartOf

Create New Evidence Model

Examples:

- [Siege of Tripolitsa occurred near Tripoli with Theodoros Kolokotronis being the leader of the Greeks against Turkey](#)
- [Angered over American support of his rivals in the Mexican Revolution, the peasant-born revolutionary leader Pancho Villa attacks the border town of Columbus](#)
- [Throughout the American Revolution, Shawnee warriors conducted raids against American settlements in Kentucky. In November 1782, George Rogers Clark, hoping to prevent further attacks, led 1,050 men against the Shawnee Indians living in the Miami River Valley, near Springfield](#)

Input text

Perform Geographical Resolution Back to text

Figure 2. KLocator User Interface

geographical scope and ranked them using the confidence score derived from Equation (6). We then measured the effectiveness of the process by determining the number of correctly scope resolved texts, namely texts whose highest ranked scope locations were the correct ones. As a baseline, we compared our results to the ones derived from Yahoo! Placemaker [27] geoparsing web service.

The results of the above process are shown in Table IV. As one can see, the improvement our method achieves in the effectiveness of the scope resolution task is quite significant in both datasets and especially in the first one where the scope-related locations are not explicitly mentioned within the texts. This verifies the central idea of our approach that non-geographical semantic information can significantly improve the geographical scope resolution process and in particular the subtasks of:

- 1) Inferring relevant to the text's geographical scope locations even in the absence of explicit reference of them within the text (first dataset).
- 2) Distinguishing between relevant and non-relevant to the text's geographical scope locations, even in the presence of non-relevant location references within the text (second dataset).

Table IV
GEOGRAPHICAL SCOPE RESOLUTION EVALUATION RESULTS

System/Approach	Dataset 1	Dataset 2
Proposed Approach	70%	85%
Yahoo! Placemaker	18%	30%

VI. DISCUSSION

It should have been made clear from the previous that our KLocator is not independent of the content or domain of the input texts but rather adaptable to them. That is exactly its main differentiating feature as our purpose was not to build another generic geographical resolution system but rather a reusable framework that can i) be relatively

New Evidence Model Creation

Evidence Model Name:

Target Concept	Evidence Concept	Relation(s) linking Evidence to Target	
http://dbpedia.org/ontology/PopulatedPlace	http://dbpedia.org/ontology/MilitaryConflict	http://dbpedia.org/ontology/place	
http://dbpedia.org/ontology/PopulatedPlace	http://dbpedia.org/ontology/MilitaryConflict	http://dbpedia.org/ontology/place , http://dbpedia.org/ontology/isPartOf	<input type="button" value="Delete row"/>
http://dbpedia.org/ontology/PopulatedPlace	http://dbpedia.org/ontology/MilitaryPerson	http://dbpedia.org/ontology/commander (inverse) , http://dbpedia.org/ontology/place	<input type="button" value="Delete row"/>
http://dbpedia.org/ontology/PopulatedPlace	http://dbpedia.org/ontology/PopulatedPlace	http://dbpedia.org/ontology/isPartOf	<input type="button" value="Delete row"/>

Figure 3. New Evidence Model Creation Form

Examples:

- [Siege of Tripolitsa occurred near Tripoli with Theodoros Kolokotronis being the leader of the Greeks against Turkey](#)
- [Angered over American support of his rivals in the Mexican Revolution, the peasant-born revolutionary leader Pancho Villa attacks the border town of Columbus](#)
- [Throughout the American Revolution, the British established Indian settlements in Kentucky. In November 1782, George Rogers Clark, hoping to prevent further attacks on the settlements, burned Indian settlements in Kentucky. In November 1782, George Rogers Clark, hoping to prevent further attacks on the settlements, burned Indian settlements in the Miami River Valley, near Springfield](#)

Input text

Siege of Tripolitsa occurred near [Tripoli](#) with Theodoros Kolokotronis being the leader of the Greeks against [Turkey](#)

Figure 4. Semantic Entity Resolution Example

easily adapted to the particular characteristics of the domain and application scenario at hand and ii) exploit these characteristics in order to increase the effectiveness of the disambiguation process. Our motivation for that was that, as the comparative evaluation showed, the scenario adaptation capabilities of existing generic systems can be inadequate in certain scenarios (like the ones described in this paper), thus limiting their applicability and effectiveness.

In that sense, our proposed framework is not meant as a substitute or rival of other geographical resolution approaches (that operate in open domains, use geographical information and relevant heuristics and apply machine learning and statistical methods) but rather as a complement of them in application scenarios where text domain and content are a priori known and comprehensive domain ontological knowledge is available (as in the case of historical texts used in our experiments).

Of course, the usability and effectiveness of our approach is directly proportional to the content specificity of the texts to be disambiguated and the availability of a priori knowledge about their content. The greater these two parameters are, the more applicable is our approach and the more effective the disambiguation is expected to be. The opposite is true as the texts become more generic and the information we have out about them more scarce. A method that could

a priori assess how suitable is our framework for a given scenario would be useful, but it falls outside the scope of this paper.

Also, the framework's approach is not completely automatic as it requires some knowledge engineer or domain expert to manually define the scenario's geographical resolution evidence mapping function. Nevertheless, this function is defined at the schema level thus making the number of required mappings for most scenarios rather small and manageable.

As far as the scalability of our approach is concerned, the main computational burden of the process is the building of the evidence index which takes place offline. In our experiments with the history knowledge base, the index building took about 2 minutes, in a standard server. On the other hand, the online location identification process took 1-2 seconds, depending of course on the size of the text. More generally, although we have not yet formally evaluated scalability, the fact that our framework is based on the constraining of the semantic data to be used makes us expect that it will perform faster than traditional approaches that use the whole amount of data. Furthermore, as the resolution evidence model is constructed offline and stored in some index, the most probable bottleneck of the process will be the phase of determining the candidate entities for

the extracted terms rather than the resolution process.

Finally, the typical errors our system is prone to, are related to two steps of the process, namely text entity detection and entity and scope resolution. In the text entity detection step, it can be the case that tokens are matched to wrong entities for reasons having to do with the linguistic analysis subsystem and/or the quality of the semantic data (entity coverage labeling, etc). On the other hand, entity and scope resolution typically fails when available evidence in the text is either too poor or too ambiguous.

VII. CONCLUSION

In this paper, we proposed KLocator, a novel framework for optimizing geographical entity and scope resolution in texts by means of domain and application scenario specific non-geographical semantic information. First, we described how, given a priori knowledge about the domain(s) and expected content of the texts that are to be analyzed, one can define a model that defines which and to what extent semantic entities (especially non-geographical ones) can be used as contextual evidence indicating two things:

- Which is the most probable meaning of an ambiguous location reference within a text (geographical entity resolution task).
- Which locations constitute the geographical scope of the whole text (geographical scope resolution task).

Then, we described how such a model can be used for the two tasks of geographical entity and scope resolution by providing corresponding processes. The effectiveness of these processes was experimentally evaluated in a comprehensive and comparative to other systems way. The evaluation results verified the ability of our framework to significantly improve the effectiveness of the two resolution tasks by exploiting non-geographical semantic information.

Given the semi-automatic nature of our framework and its dependence on the availability of comprehensive semantic data, future work will focus on investigating how statistical and machine learning approaches may be used, in conjunction with our approach, in order to i) automatically build geographical resolution evidence models based on text corpora and ii) deal with cases where available domain semantic information is incomplete.

ACKNOWLEDGMENT

This work was supported by the European Commission under contracts FP7- 248984 GLOCAL and FP7-287615 PARLANCE.

REFERENCES

- [1] P. Alexopoulos, C. Ruiz, and J. M. Gomez-Perez, "Optimizing geographical entity and scope resolution in texts using non-geographical semantic information." in *Proceedings of the Sixth International Conference on Advances in Semantic Processing (SEMAPRO)*, 2012, pp. 65–70.
- [2] C. B. Jones and R. S. Purves, "Geographical information retrieval," *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 219–228, Jan. 2008.
- [3] J. Raper, G. Gartner, H. Karimi, and C. Rizos, "Applications of location-based services: a selected review," *J. Locat. Based Serv.*, vol. 1, no. 2, pp. 89–111, Jun. 2007.
- [4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, p. 122, 2009.
- [5] L. M. V. Blázquez, B. Villazón-Terrazas, V. Saquicela, A. de León, Ó. Corcho, and A. Gómez-Pérez, "Geolinked data and inspire through an application case," in *GIS*, 2010, pp. 446–449.
- [6] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "Linked-geodata: A core for a web of spatial open data," *Semantic Web Journal*, vol. 3, no. 4, pp. 333–354, 2012.
- [7] G. Andogah, G. Bouma, J. Nerbonne, and E. Koster, "Place-name ambiguity resolution," in *Methodologies and Resources for Processing Spatial Language (Workshop at LREC 2008)*, 2008.
- [8] G. Andogah, G. Bouma, and J. Nerbonne, "Every document has a geographical scope," *Data and Knowledge Engineering*, 2012.
- [9] T. Kauppinen, R. Henriksson, R. Sinkkilä, R. Lindroos, J. Vtinen, and E. Hyvnen, "Ontology-based disambiguation of spatiotemporal locations," in *Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*. Tenerife, Spain: CEUR Workshop Proceedings, ISSN 1613-0073, June 1-5 2008.
- [10] Y. Peng, D. He, and M. Mao, "Geographic named entity disambiguation with automatic profile generation," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 522–525.
- [11] J. Kleb and A. Abecker, "Entity reference resolution via spreading activation on rdf-graphs," in *Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part I*, ser. ESWC'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 152–166.
- [12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: a nucleus for a web of open data," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 722–735.
- [13] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," in *16th international World Wide Web conference (WWW 2007)*. New York, NY, USA: ACM Press, 2007.
- [14] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems*, ser. I-Semantics '11. New York, NY, USA: ACM, 2011, pp. 1–8.

- [15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenaу, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 782–792.
- [16] M. Janik and K. Kochut, "Wikipedia in action: Ontological knowledge in text categorization." in *ICSC*. IEEE Computer Society, 2008, pp. 268–275.
- [17] M. Wallace, P. Mylonas, G. Akrivas, Y. Avrithis, and S. Kollias, *Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra*. Studies in Fuzziness and Soft Computing, Soft Computing in Ontologies and Semantic Web, Springer, Ma, Z. (Ed.), Vol. 204, 2006.
- [18] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania, "Efficiently linking text documents with relevant structured information." in *VLDB*, U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, Eds. ACM, 2006, pp. 667–678.
- [19] "Linked open data cloud," <http://www.lod-cloud.net/>, accessed: 15/12/2013.
- [20] "Dbpedia spotlight," <http://spotlight.dbpedia.org/>, accessed: 15/12/2013.
- [21] "Aida," <https://gate.d5.mpi-inf.mpg.de/webaida/>, accessed: 15/12/2013.
- [22] "Wikimeta," <http://www.wikimeta.com>, accessed: 15/12/2013.
- [23] "Zemanta," <http://www.zemanta.com>, accessed: 13/07/2013.
- [24] "Alchemy api," <http://www.alchemyapi.com>, accessed: 15/12/2013.
- [25] "Yahoo!" <http://developer.yahoo.com/search/content/V2/contentAnalysis.html>, accessed: 15/12/2013.
- [26] G. Rizzo and R. Troncy, "NERD: A framework for evaluating named entity recognition tools in the Web of data," in *ISWC 2011, 10th International Semantic Web Conference, October 23-27, Bonn, Germany, 2011*.
- [27] "Yahoo! placemaker," <http://developer.yahoo.com/boss/geo/>, accessed: 15/12/2013.