Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus

Ludovic Moncla
Université de Pau et des Pays de l'Adour LIUPPA
Pau, France
lmoncla@univ-pau.fr

Walter Renteria-Agualimpia
Universidad de Zaragoza
C/ María de Luna, 1
Zaragoza, Spain
walterra@unizar.es

Javier Nogueras-Iso
Universidad de Zaragoza
C/ María de Luna, 1
Zaragoza, Spain
jnog@unizar.es

Mauro Gaio
Université de Pau et des Pays de l'Adour LIUPPA
Pau, France
mauro.gaio@univ-pau.fr

ABSTRACT

Geoparsing and geocoding are two essential middleware services to facilitate final user applications such as location-aware searching or different types of location-based services. The objective of this work is to propose a method for establishing a processing chain to support the geoparsing and geocoding of text documents describing events strongly linked with space and with a frequent use of fine-grain toponyms. The geoparsing part is a Natural Language Processing approach which combines the use of part of speech and syntactico-semantic combined patterns (cascade of transducers). However, the real novelty of this work lies in the geocoding method. The geocoding algorithm is unsupervised and takes profit of clustering techniques to provide a solution for disambiguating the toponyms found in gazetteers, and at the same time estimating the spatial footprint of those other fine-grain toponyms not found in gazetteers. The feasibility of the proposal has been tested with a corpus of hiking descriptions in French, Spanish and Italian.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms, Experimentation

Keywords

Geocoding, Toponym disambiguation, Spatio-textual searching, Geoparsing, Location based services

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666386

1. INTRODUCTION

Geoparsing and geocoding are complementary services to facilitate respectively the recognition of spatial language in text documents and the mapping of such language to explicit georeferences (e.g., lat/long values) [19]. In the last decade there has already been big research efforts addressing the problems behind these two services because they are essential for providing back-end data which is exploited later by multiple applications such as location-aware searching [26, 33] or different types of location-based services (e.g. mobile applications for finding nearest points of interest or route planning for emergency response services).

The problem of geoparsing, i.e. the recognition of toponyms (place names) in text, can be seen as a particular category of named entity recognition and classification (NERC). Two categories of approaches have been proposed, those that use learning techniques and those based on natural language processing (NLP), in particular syntactico-semantic rules. Both categories use external lexical resources. These approaches can be used in a complementary manner in hybrid systems [18]. Amongst the rule-based approaches, several use transducers with a finite number of states [25], which can also be used in cascade [22].

With respect to the problem of geocoding, also known as toponym resolution, the objective is to associate a toponym with its spatial footprint. The first issue to solve for this resolution is the ambiguity contained in place names expressed in text (the problem of solving these ambiguities is known as toponym disambiguation [7]). According to Smith and Mann [30] there are three main types of ambiguities : the same name is used for several places (*referent ambiguity*) ; the same place can have several names (*reference ambiguity*) ; and the place name can be used in a non-geographical context, like organizations or names of people (*referent class ambiguity*). Another type of ambiguity is called *structural ambiguity* and it arises when the structure of the words constituting the place name in the text are ambiguous (e.g. is the word *Lake* part of the toponym *Lake Grattaleu* or not ?) [31]. We consider that this type of ambiguity can be seen as a subset of the reference ambiguity.

The second issue is to have adequate resources to associate a footprint to a disambiguated place name. The use of resources like gazetteers is thus inescapable.

In an open data context, the availability of gazetteers is expanding and we may mention global resources such as Geonames, OpenGeoData, OpenStreetMap and Wikimapia or national resources [1] such as BD NYME (France) or Nomenclátor Geográfico Básico (Spain). However, whatever the resource is selected for geocoding, the problem that usually arises is its completeness. Currently, the present Web geocoding public market is dominated by geocoding services for average users, i.e. users that can accept low Quality of Service in terms of resolution [13], i.e. geocoding of administrative units, street names in urban areas, or names of well-known touristic sites are the main needs. But there are contexts, even for public citizens or casual users, where the completeness of resources is crucial, in particular as regards the geocoding of fine-grain toponyms. For instance, in a corpus of narrative descriptions of places in a small area, it is common to find toponyms referring to geographical entities of varying size. Additionally, there are frequent occurrences of micro-toponyms, which are not usually found in gazetteers thought for broader audiences.

The objective of this work is to propose a method establishing a processing chain to support the geoparsing and geocoding of text documents describing events strongly linked with space and with a frequent use of fine-grain toponyms. The geoparsing part of the workflow is a NLP approach based on a previous work of the authors [23], which combines the use of a Part Of Speech tagger and a cascade of transducers. The real novelty of this work lies in the geocoding part of the method. The geocoding algorithm is an unsupervised algorithm that takes profit of clustering techniques to provide a solution for disambiguating those toponyms found in gazetteers, and at the same time estimating the spatial footprint of those other fine-grain toponyms not found in gazetteer resources. This additional contribution about the estimation of the spatial footprint is closely related to the step of spatial inference proposed by Leidner and Lieberman [19] in their "Reference model for processing textual geographic references".

Additionally, for this contribution the method has been evaluated for a corpus of hiking descriptions in three different languages : French, Spanish and Italian.

The remainder of this paper is structured as follows. Section 2 discusses the related work about toponym disambiguation, and the inference of location information associated to toponyms. Section 3 describes the processing chain proposed for geoparsing and geocoding. Section 4 describes our implementation and relates the early results of our experiments. Finally, Section 5 provides conclusions and outlook on future work.

2. RELATED WORK

As mentioned in the introduction, geocoding involves two important and related issues : toponym disambiguation, and the adequate selection of a resource for assigning a footprint to a disambiguated toponym. The task of toponym disambiguation and the assignment of a footprint are two activities that are usually combined thanks to the use of a gazetteer with an appropriate coverage of the georeferences which might be associated to ambiguous toponyms. Subsection 2.1 provides an overview of approaches for toponym disambiguation.

However, sometimes there is no availability of gazetteers with appropriate coverage for the processing of text documents. In those cases, we need to explore alternative methods to infer the spatial location of toponyms. Although this last issue has received little attention in the research literature, Subsection 2.2 analyzes some approaches about how to define new toponyms or improve the spatial information associated with existing toponyms.

2.1 Related work about toponym disambiguation

Buscaldi [5] provides an overview about different ways of disambiguating toponyms. According to this work, the approaches can be classified in three categories : map-based, knowledge-based, and supervised or data-driven approaches. Map-based approaches use as the context for disambiguation other unambiguous and georeferenced toponyms found on the same document. These approaches provide a score to any of the possible locations according to the distance to the unambiguous toponym. Knowledge-based approaches make profit of knowledge sources (gazetteers, ontologies, and so on) to see if other related toponyms in this knowledge source are also referred in the document [26, 6, 20], or additional information from the toponyms (e.g., population) or document creators (e.g. documents of social media [17]) can be exploited. Finally, data-driven approaches [2, 30] are based on machine learning algorithms. The main drawback with this last category is the lack of classified collections.

As the application domain for our proposal are textual descriptions of itineraries and the texts are short, we mainly focus in this subsection about works related to map-based disambiguation, i.e. the main information to solve disambiguation is the explicit georeference of detected toponyms. Buscaldi and Rosso [7] describe a basic implementation of a map-based method that analyzes the distance from the centroid of all possible locations of toponyms cited in the text to a candidate location in order to choose the most appropriate disambiguated toponym. However, this basic implementation can be refined with multiple improvements related to the definition of the disambiguation context, the computation of distance to the candidate location or the consideration of additional physical properties.

With respect to the definition of a more refined disambiguation context, Zhao et al [33] propose, for instance, a GeoRank algorithm inspired in Google Page Rank algorithm for the disambiguation of toponyms in Web resources. The idea is that other toponyms identified on the same resource vote to the alternative locations of the ambiguous toponym according to their distance in the text and the geographical distance to the tentative location.

About the diversity of techniques for computing closeness

1. `http://unstats.un.org/unsd/geoinfo/ungegn/geonames.html`

to a candidate location, Habib and van Keulen [14] propose the use of clustering techniques to disambiguate ambiguous toponyms in holiday descriptions. The clustering approach is an unsupervised disambiguation approach based on the assumption that toponyms appearing in same document are likely to refer to locations close to each other distance-wise. Another alternative to take into account the distance closeness is the one proposed by Zhang et al [32]. They use an Exact-All-Hop Shortest Path (EAHSP) algorithm for disambiguating road names in text route descriptions. Road name disambiguation belongs to the scope of toponym disambiguation. Their proposed EAHSP algorithm aims at finding a path that maximizes the number of crossed roads in a proximity area. Besides, it must be also acknowledged that this work faces additional difficulties since databases for road names are not so popular, and their associated geometry is not point-based.

Finally, related to the consideration of additional physical properties, Derungs and Purves [8] propose a map-based disambiguation algorithm for toponyms found in landscapes descriptions as part of a mechanism to assign a general geospatial footprint to documents. Apart from using the Euclidean distance from ambiguous toponyms to already identified unambiguous toponyms, the proposed disambiguation algorithm exploits the topographic similarity between toponyms. The performance of their disambiguation algorithm is closely related to the availability of gazetteers with topographic information, which may be missing for some regions. Additionally, this work tries to identify the terms used to describe natural features in a region and compare the different vocabularies used for natural features in different regions.

2.2 Related work about the definition of new toponyms

There are some works in the literature that aim at the creation of new databases of geographic locations. For instance, Lieberman et al [21] present a method for generating a local spatial lexicon by processing a corpora from local newspapers, which is used later for the geocoding of documents. In principle, the method only aims at including in this spatial lexicon the disambiguated toponyms extracted from newspapers' articles and discovered in an existing gazetteer. The method used for disambiguating toponyms with possible interpretations is based on the definition of a convex hull covering nearby possible locations of toponyms. However, the method could be easily extended to assign the geographic extent of this convex hull to all those names recognized as place names, but not found in the gazetteer.

There are also some relevant works investigating the definition of toponyms using social media sources. For instance, Rattenbury and Naaman [27] present different methods based on burst-analysis techniques for extracting place semantics from Flickr Tags. Training these methods with Flickr data containing high-resolution location metadata (i.e., longitude and latitude), it is possible to derive automatic associations between place name tags and explicit georeferences. As the authors claim, these methods could automate the creation of place gazetteer data. In the same line, Serdyukov et al [28] propose the development of a statistical language model to predict the most likely geocoding corresponding to a set of location tags associated to a Flickr photo.

Additionally, there are some works applied in similar contexts to the ones proposed for the context of our experiments, i.e. narrative descriptions of places in small areas, whose objective is also to infer the spatial location of complex place names, not directly found in gazetteers. In this area Scheider and Purves [1] have proposed the use of semantic technologies to process narrative descriptions of mountain itineraries or historic places to infer the location of complex names in terms of spatial relations with respect to well recognized landmarks.

Finally, there are some works aiming at the enrichment of existing geographic databases. An example in this area is the work of Smart et al [29]. They present a mediation framework to access and integrate distributed gazetteer resources to build a meta-gazetteer that generates augmented versions of place name information. In this mediation framework they include geofeature augmentation module. During the merging process of features from different sources, they detect matchings of these features and create more complete and consistent information, e.g. adding hierarchical information about administrative units. Another work about the enrichment of existing toponyms could be the work proposed by Hao et al [15] for the enrichment of destinations in travelogues with knowledge (e.g. local topics such as beach, mountain or other features related to the toponym) mined from a large corpus using probabilistic models.

3. OUR PROPOSAL

Our proposed approach is a hybrid solution that combines map-based disambiguation with the assignment of georeferences for new toponyms.

The first step is the geoparsing (Fig. 1a) : we annotate geographic entities thanks to a well-established process [23] based on the use of POS taggers and the application of syntactico-semantic combined patterns (cascade of transducers). Then, we start with the geocoding part of the proposal. The second step of our proposal is the georeferencing of toponyms thanks to the interrogation of well-known gazetteers (e.g., Geonames, BD NYME, OpenStreetMap)(Fig. 1b). To solve the ambiguities added by the georeferencing step we propose to use a clustering algorithm based on spatial density.

With respect to the map-based disambiguation part, our work is similar to the proposal of Habib and van Keulen [14] as we also use a clustering technique. The difference is that they do not face the problem of not finding toponyms in their gazetteer. Additionally, the granularity of the spatial footprint is higher than ours : their objective is to identify to which country a set of toponyms belong. The disambiguation part may also have some similarities with the work of Delarungs and Purves [8] as the types of input documents to be processed are similar : descriptions of natural landscapes could be considered as a superset of hiking descriptions. However, their aim is not to try to geolocate toponyms not found in a backend gazetteer. Anyway, the vocabulary for natural features that they analyze could probably intersect with the toponyms (or part of the place names) that we try to geolocate in this work.

In our proposal the clustering algorithm is also used to define the geographic extent of all geographic entities in the document used as input and to propose a geocoding for the toponyms not found in gazetteers (Fig. 1c).

With respect to the creation of new toponyms, our work also has some similarities with the one proposed by Lieberman et al [21]. Our purpose is also to create a gazetteer or spatial lexicon for an intended audience in small areas. Additionally, the proximity of locations associated with ambiguous toponyms is the main criterion to discard alternatives. Furthermore, some ideas of the works of Scheider and Purves [1] and Hao et al [15] could be used to provide additional information to discovered fine-grain toponyms. The annotation obtained after the toponym extraction process could be used to inform about a more precise geolocation, or the topics associated to this toponym.

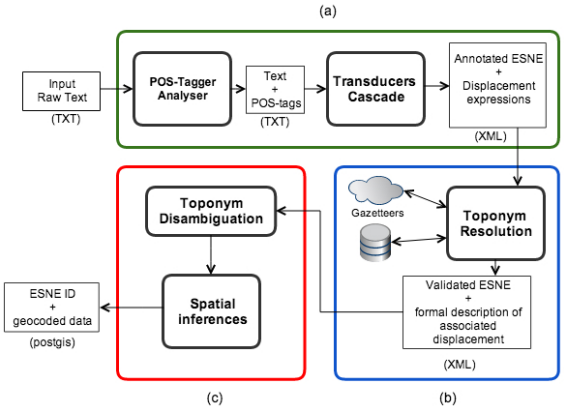


Figure 1: Block diagram of our processing chain

3.1 Corpus

In order to build a corpus of narrative descritpions of places in a small area, we chose hiking descriptions. Hiking descriptions are specific documents describing displacements using geographical information, such as toponyms, spatial relations, and natural features or landscapes. It must be noted that this corpus of documents deals with a homogeneous theme and describes specific geographic small and single areas.

To build our corpus of hiking descriptions, thousands of hiking descriptions were automatically extracted from specialized websites in French [2], Spanish [3], and Italian [4]. Each collected hiking description is associated with a GPS track useful for our experiments. Then with the aim to evaluate our proposal we built a body of reference consisting of 30 hiking descriptions manually annotated (ground truth) for each language.

Table 1 shows some features of each body of reference : number of documents, total number of words, average number of words per document, total number of toponyms, and average number of toponyms per document. These 90 documents are representative of the whole corpus, each document has an average of : 263 words (269 on the body of reference) with standard distance of 188 (242 for the body of reference) and 15 toponyms (14 on the body of reference) with standard distance of 9 (12 for the body of reference).

	French	Spanish	Italian
#Documents	30	30	30
#Words	11297	5626	6856
Avg. #Words	376	187	228
#Toponyms	583	376	409
Avg. #Toponyms	19.4	12.5	13.63

Table 1: Document sets

Table 2 shows the ten most popular terms associated with toponyms in our body of reference (French, Spanish and Italian). These terms annotated by our processing chain as sub-types and we consider that they are part of the name of the spatial entity. These spatial entities are most of the time fine-grain (hamlet, cottage, bridge, church, lake) and 46% of toponyms in our French body of reference have an associated sub-type. (24% for Spanish and 36% for Italian).

French		Spanish		Italian	
col	20	puente	17	rifugio	20
village	20	rio	17	monte	19
hameau	20	pueblo	12	villaggio	17
route	17	iglesia	10	masi	15
sentier	15	camino	9	castello	9
chalet	13	barranco	8	lago	8
refuge	11	parque	5	passo	7
pont	11	castillo	3	foce	7
lac	8	barrio	3	chiesa	6
chapelle	8	casa	2	via	5

Table 2: Most frequent terms associated with toponyms

3.2 Geoparsing

Our approach uses a geoparsing system built to extract spatial named entities, but also spatial relations. This annotation system was first designed for French corpus only and was described in a previous work [23]. For this current work we transformed our processing chain to deal with Spanish and Italian documents. The first step (Fig. 1a) of our approach is a system where spatial expressions described in textual documents are automatically annotated. The method combines the notions of marking and extraction of named entities, through the use of local grammars or external resources (lexicons).

The first step of this annotation tool is done by a part of speech (POS) analyser. The output is used by syntactico-semantic combined patterns in a transducers cascade to mark toponyms, spatial relations and expanded spatial named entity (ESNE). We defined ESNE as an entity built from a proper name attributed to a place (toponym), which can be associated with a sub-type and one or more concepts relating to the expression of location in the language (spatial relations).

The rules to annotate toponyms are built using proper nouns. There are also rules to describe spatial relations using POS tags like prepositions or common nouns and lexicons.

2. `http://www.visorando.com` (fr)
3. `http://senderos.turismodearagon.com` (es)
4. `http://www.parks.it/parco.alpi.marittime/` (it)

Listing 1 shows the result of the annotation for the following sentence : *Walk to the refuge south of hamlet of Fontanettes*. We can notice that in that sentence the verb of motion (*to walk*), the spatial relation (*south of*) and the toponym (*hamlet of Fontanettes*) have been annotated. Verbs of motion, or perception are considered as a special kind of spatial relations called VT [24]. This VT annotation establishes the link between a verb of motion or perception and a spatial entity (toponym or ESNE).

At the end of the geoparsing process toponyms, spatial relations, and ESNE are annotated.

```
<VT>
  <verb type="motion" polarity="median">
    <token>Walk</token>
  </verb>
  <ESNE focus="final">
    <token>to</token>
    <commonNoun>
      <token>the</token>
      <token>refuge</token>
    </commonNoun>
    <indirection type="orientation">
      <token>south</token>
      <token>of</token>
    </indirection>
    <toponym>
      <subType>
        <token>hamlet</token>
      </subType>
      <token>of</token>
      <subToponym>
        <token>Fontanettes</token>
      </subToponym>
    </toponym>
  </ESNE>
</VT>
```

Listing 1: Example of an XML result of our geoparsing process (translated from French)

3.3 Geocoding

To find a geo-coded representation for toponyms extracted from the textual descriptions (toponym resolution), we query some well-known gazetteers.
During this process a lot of ambiguities may arise. One place can have several names (*reference ambiguity*). For example, this happens when the name has changed over the time, or when the name commonly used by people is different from the official name. In order to avoid this ambiguity our processing chain looks for the toponym name in all available fields of the backend database of gazetteers containing official or alternative names. Additionally, the fact of querying different gazetteers also expands the probabilities of the method to find a matching with one of the possible names of a place. Apart from these clear cases of reference ambiguity, our method is focused on the problem of the inclusion or not inclusion of sub-types within the official name of a toponym (*structural ambiguity*).

Most of the named entities used in this corpus are spatial named entities. Thus this circumstance allows us not to take into account the *referent class ambiguity*. But as far as we know another ambiguity untreated in the literature appears : the incompleteness of the gazetteers. Not all existing toponyms are stored in databases. Thus we call this problem the *unreferenced toponyms ambiguity*.

3.3.1 Dealing with structural ambiguity

Sometimes toponyms or ESNE are stored in gazetteers with the so-called *full name*, which means that the name consists of a sub-type and a proper name. For example, the toponym *hamlet of Fontanettes* consists of the subtype *hamlet* and the proper name *Fontanettes*. But in many cases, toponyms are just stored in gazetteers using the proper name, and in such cases they are usually associated with metadata describing the type of feature (e.g., hamlet, city or stream).
The approach to know which result matches our query is to compare the metadata type of the results with the subtype of the toponym extracted from the text. In some cases, ambiguities can be solved using their sub-type [24] when it is available in the textual description.

For instance, the expression *hamlet of Fontanettes* is not found in gazetteers. Instead, we can only find it with the name *Fontanettes*. Several results exist for this query and one of them has the metadata type *hamlet*.

However, if none of the possible names of a place is stored in the gazetteers, then we have a case of unreferenced toponym ambiguity, which is considered specifically in section 3.3.3.

3.3.2 Dealing with referent ambiguity

In many cases the disambiguation approach presented in section 3.3.1 is not enough because gazetteers may contain several toponyms with the same name and type. In these cases we need a mechanism to distinguish the good group of toponyms associated with the real trajectory of the hiking description.
Similar to the works of Feuerhake and Sester [12] or Intagorn et al. [16], we propose to use clustering algorithms to find collections sharing a spatial property, and in our case, these collections enable us to find clusters of the most likely geospatial points belonging to a hiking trail. In particular, we are using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm [11]. It uses the concept of density to determine the neighborhood of a point, that is, what constitutes a cluster. DBSCAN uses two parameters to define the density concept : *Eps* and *MinPts*. *Eps* (epsilon radius) determines the area of a neighborhood and *MinPts* determines the minimum number of points that must be contained in that neighborhood to deem it a cluster. In our current methodology, the values of DBSCAN parameters have been empirically adjusted according to the features of the linking dataset used in the experiments.

DBSCAN can deal with the problems of data with noise, i.e. DBSCAN has the ability to detect outliers. In our context, an outlier is a point that does not belong to the hiking trail cluster. Additionally, since hiking trails may have many points describing different trajectory shapes, DBSCAN can find arbitrarily shaped clusters. The output of the DBSCAN is a set of clusters of toponyms whose footprints are close. Every cluster represents a possible set of points describing the hiking trail. Then we need a way to identify the best cluster matching with the set of points in the hiking trail. The heuristic is defined as follows : given a set of cluster C_1, C_2,... ,C_n generated by the clustering algorithm, the best cluster C_b is the one containing the largest number of distinct toponyms. In other words, the best cluster identifies the area with the largest co-occurrence of toponyms.

The proposed method reduces considerably the structural and referent ambiguities. Now the remaining problem is how to deal with unreferenced toponyms ambiguity, i.e. how to find spatial locations for toponyms that are not found in gazetteers.

3.3.3 Dealing with unreferenced toponym ambiguity

Our proposal is to infer locations from the locations of previously disambiguated toponyms. However, these spatial inferences cannot be as precise as points with geographical coordinates (latitude/longitude). These spatial inferences are represented by a geographical area which can be refined depending on various spatial informations contained in the textual descriptions.

For example, Figure 2 shows three different cases of inferences. In the first case (Fig.2a) there is not other explicit spatial information in the text linked with the unreferenced toponyms. In this case, when there is no information concerning the context of a toponym, we define a geographical area that contains all well located toponyms thanks to the clustering method previously described. Indeed, as we are working with hiking descriptions, toponyms are related to each other and they are located in the same area. The second case is when explicit spatial relations are associated with the unreferenced toponym. For example, if we know that the unreferenced toponym is somewhere south of *A* (Fig.2b), then we can define a new area smaller than the previous one. A third case arises when we have even more information available in the textual description. In this case we can define a much smaller area. For example, if we know that the unreferenced toponym is somewhere between two other toponyms, we can define a small area between these two toponyms (Fig.2c). Spatial relations are very important to determine the context of unreferenced toponyms. We have developed specific transducers in our cascade using lexicons or patterns to annotate different categories of spatial relations : distances, topological relations, directional relations and displacements. Currently our geoparsing process is able to annotate these categories of spatial relations for the three languages (French, Spanish and Italian).

4. EXPERIMENTS

4.1 Geoparsing

As mentioned in section 3.1, to evaluate our proposal we built a body of reference consisting of 30 hiking guides manually annotated (ground truth) for each language.

For each document set (French, Spanish, and Italian) of hiking descriptions, we evaluated the precision and recall obtained with our georpasing method (Table 3). The precision is the ratio of the number of relevant toponyms annotated (y) to the total number of toponyms annotated (z) ; and the recall is the ratio of the number of relevant toponyms annotated (y) to the number of relevant toponyms manually annotated (x).

In the context of hiking descriptions almost all named entities are spatial named entities, and the analysis of the results showed that the remaining errors come from geo/non geo ambiguities (referent class ambiguity).
It can be observed that our processing chain yields a very high precision and recall using a specific corpus dealing with spatial named entities.

	(x)	(y)	(z)	Precision	Recall
French	583	581	595	98.29%	99.42%
Spanish	376	374	376	99.26%	99.33%
Italian	409	407	410	99.10%	99.68%

Table 3: Evaluation of our geoparsing method

The first step of our annotation tool is an automatic part-of-speech annotation. The efficiency is not the same depending the language (French, Spannish or Italian) and the analyser used. To compare the same results for all languages and in order to avoid adding errors before the application of the transducers cascade, we manually corrected the output of the part-of-speech analysers. Without this manual correction, the results are between 5 to 15% lower depending on the language and the analyser used.

4.2 Geocoding

For the geocoding experiments we are using gazetteers provided by national mapping institutes : BD NYME [5](France), Nomenclátor Geográfico Básico de España [6](Spain), and Toponimi d'Italia IGM [7](Italy). The Spanish and Italian gazetteers are accessible through the Web Feature Service specification defined by the Open Geospatial Consortium [8]. Additionally we also use some well-known gazetteers : Geonames [9] and OpenStreetMap [10].

4.2.1 Dealing with structural ambiguity

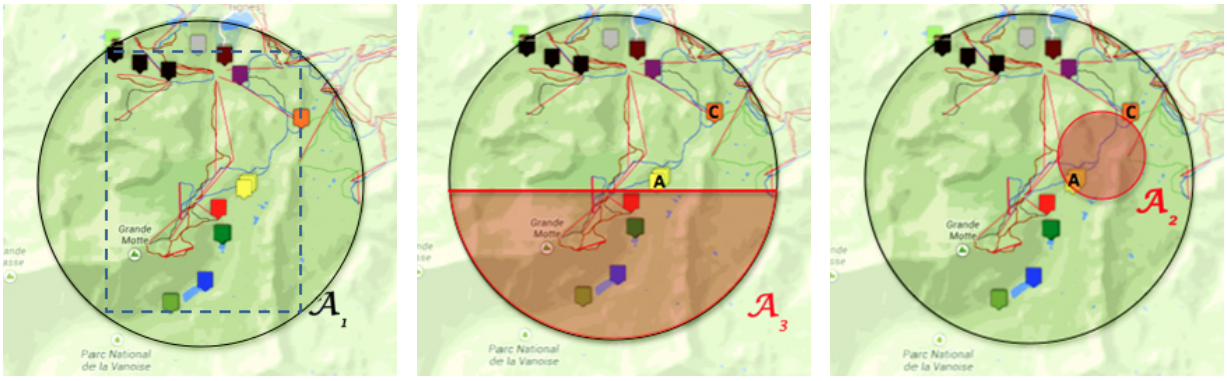
For each toponym extracted from the text we first query the databases with its full name (exact matching). If there is no result, then we query again the databases with the sub-toponym (see Section 3.3.1). Last we try to disambiguate the retrieved candidate toponyms with the metadata *type*.

Table 4 shows the percentage of toponyms (or sub-toponyms) found in gazetteers. For French texts, 13.78% of toponyms are not found (5.59% in the case of Spanish texts and 23.17% in the case of Italian texts). Furthermore, we can notice that the gazetteers complement each other.

	French	Spanish	Italian
Full name query	**54.79%**	**73.14%**	**43.90%**
National Gazetteer	37.82%	62.77%	17.32%
Geonames	29.24%	53.99%	29.02%
OpenStreetMap	40.17%	63.56%	34.63%
Sub-toponym query	**31.43%**	**21.28%**	**32.93%**
National Gazetteer	23.03%	18.88%	17.32%
Geonames	23.19%	18.09%	22.20%
OpenStreetMap	24.87%	19.41%	25.85%
Total	**86.22%**	**94.41%**	**76.83%**

Table 4: Percentage of toponyms found in gazetteers

5. http://www.geoportail.gouv.fr
6. http://www.ign.es
7. http://www.pcn.minambiente.it/GN/
8. http://www.opengeospatial.org/standards/wfs
9. http://www.geonames.org
10. http://www.openstreetmap.org



(a) (b) (c)

Figure 2: Refining spatial inferences according to the context

The way we query the gazetteers is not flexible and it is possible to improve these results. For example, some toponyms are not retrieved because their names contain articles and determiners, and these articles and determiners are not specified in the text. For example, the toponym *Lac de Rocheure* extracted from a hiking description is stored in gazetteers as *Lac de la Rocheure*. Applying more flexible queries would obtain better results in terms of retrieved toponyms, but on the other hand the number of toponyms ambiguous would increase. These results can also be improved because of misspellings. We can imagine the use of algorithms like the "edit distance" (commonly used in natural language processing) to find results in gazetteers for the misspelled toponyms.

4.2.2 *Dealing with referent ambiguity*

As we said before, even if toponyms are found in gazetteers, there are still some remaining ambiguities. For instance, in the case of French hiking descriptions and depending on the used gazetteer, between 45 to 70 % of found toponyms are ambiguous. This means that for these toponyms the gazetteers give more than one result. For technical reasons and in order to compare the results coming from the three gazetteers, we set a maximum limit of results per query. The limit is set to 100 results to have less than 2% of toponyms having a limited number of results. Table 5 shows that gazetteers give an average of 15-20 locations for each toponym depending on the language.

	French	Spanish	Italian
Total # of toponyms	595	376	409
# Retrieved toponyms	513	355	315
# Results	6850	6359	5722
Avg. # Results	13.35	17.91	18.16

Table 5: Number of toponyms (and results) found in gazetteers

In order to avoid this ambiguity the clustering method explained in section 3.3.2 was applied. The validity of the best cluster for every document proposed by our algorithm was verified by comparing the similarity between the point set of each generated cluster and the original point set of the trajectory described in a GPX file. To measure the similarity we computed the convex polygon of the original point set of the trajectory and every cluster with the ST_ConvexHull PostGIS function, and then, we calculated the distance between these point sets using the ST_Distance PostGIS function [9, 3].

In 29 out of the 30 French cases, 30 out of the 30 Spanish cases and 29 out of the 30 Italian cases the best cluster suggested by our method is the best one, that is, the cluster with the best matching with respect to the real points in the trajectory. In general, for French cases, 65.63% of the toponyms found in a gazetteer are in the best cluster (65.35% for Spanish and 59.68% for Italian cases). Additionally, thanks to the comparison with respect to the real trajectory, we analyzed the missing points in every best cluster and we found that all missing points were not included within the set of points associated with the toponyms that were retrieved from gazetteers. This means that only the points included in the best clusters are well located. Table 6 shows the comparison between the number of toponyms retrieved from gazetteers and the number of well-located toponyms .

	French	Spanish	Italian
Well-located BC	29 of 30	30 of 30	29 of 30
Retrieved toponyms	513 (86.22%)	355 (94.41%)	315 (77.01%)
Well-located toponyms	340 (57.14%)	232 (61.7%)	188 (45.96%)

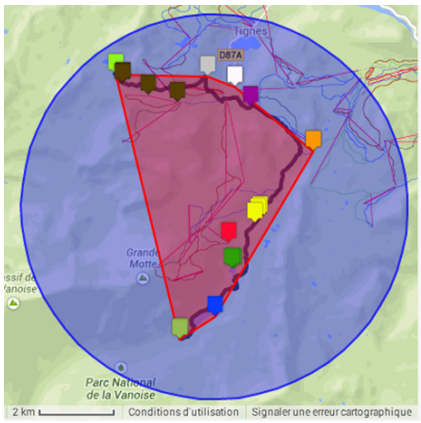
Table 6: Number of best clusters (BC) and toponyms well located

This points out the problems derived from the lack of coverage in gazetteers and the need to assign a geographic reference to those toponyms not found. In each case analyzed, the missing points were associated with fine-grain toponyms.

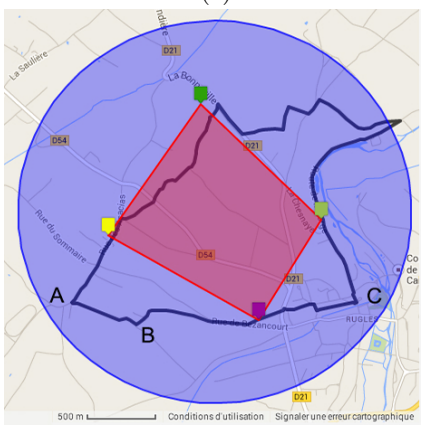
Our approach failed in 2 out of 90 cases. In these cases the problem was caused by the few number of results retrieved. For example, in one of these cases only two toponyms were retrieved from the gazetteer and the rest were not found (they were missing points). Moreover, the retrieved toponyms were referencing places not related with to real hiking description, (i.e., bigger location in other countries or regions).

4.2.3 *Dealing with unreferenced toponyms ambiguity*

As mentioned in section 3.3.3, our proposal is to infer locations for the unreferenced toponyms. We implemented two approaches to define a geographic area where the unreferenced toponyms are supposed to be in order to evaluate which one is the best. The first approach takes into account the geometric outline of the displacement by implementing the convex hull (in red in figure 3) computed with all the toponyms included in the best cluster. The second approach implements the circumscribed circle (in blue in figure 3) around the rectangle of the bounding box and does not take into account the geometric outline of the displacement.



(a)



(b)

Figure 3: Refining spatial inferences according to the context (Source : GoogleMaps)

We manually reviewed all the unreferenced toponyms of each document in the corpus. We sought into different resources like web pages or detailed geographical maps to find the real locations of the unreferenced toponyms. When the toponyms were impossible to find, we also used the GPS track available with each document of our corpus.

After running the experiments, we identified different cases of spatial inference. The first one was the perfect case, that is, when all toponyms cited in the textual description were found and were all well located (Fig.3a). The second case was when there are some unreferenced toponyms, but when their real locations are located inside the convex hull. Then the third case happened when there are several unreferenced toponyms and when the real locations of these toponyms are not included in the convex hull but are included in the circumscribed circle (see points A,B, and C in Fig.3b).

Last there are also some cases when the real locations of unreferenced toponyms are located neither in the convex hull nor in the circumscribed circle.

Table 7 shows the number of unreferenced toponyms that are actually located inside the area defined by the convex hull or the circumscribed circle. We can notice that using the circumscribed circle we are able to propose a good location for more unreferenced toponyms than with the convex hull.
We removed from the total number of automatically annotated toponyms the ones which were not toponyms (referent class ambiguity) and also the ones which were associated with an expression of perception. These toponyms can be far from the real trajectory described and our method is not adapted to locate them.

	French	Spanish	Italian
unreferenced toponyms	200	123	162
Convex hull	140 (70.0%)	75 (53.65%)	52 (32.1%)
Circumscribed circle	180 (90.0%)	102 (77.23%)	120 (74.07%)

Table 7: Numbers of unreferenced toponyms found in the convex hull or in the circumscribed circle

Furthermore our experiments show that we need at least 4 or 5 well located toponyms in order to find the best cluster and to propose a good geographic area for unreferenced toponyms.

To summarize the experiments of the full processing chain, Table 8 shows some global results : the initial numbers of toponyms manually and automatically annotated (excluding toponyms associated with expression of perception or errors) ; the number of toponyms located by gazetteers after the cluster-based disambiguation ; the number of toponyms located by our spatial inference method ; and the number of toponyms still unlocated at the end of the process. Additionally, the table shows the percentages of located (unlocated) toponyms for each body of reference.

Toponyms	French	Spanish	Italian
manually annotated	542	362	350
automatically annotated	540	360	349
located by gazetteers	340 (62.96%)	232 (64.44%)	188 (53.86%)
located by inferences	180 (33.33%)	102 (28.33%)	120 (34.38%)
unlocated	20 (3.71%)	26 (7.22%)	41 (11.74%)

Table 8: Global results of our processing chain

Finally, Table 9 shows the global accuracy of our proposal to locate toponyms. The accuracy takes into account the total number of toponyms manually annotated, including those associated with expressions of perception. We can no-

tice that adding the step of spatial inference for the location of toponyms increase significantly the accurary. On average and considering the whole corpus, our method is able to identify and correctly locate 55% of toponyms without spatial inference. In contrast, using the spatial inference method, this percentage increases up to 84% of toponyms correctly identified and located.

	French	Spanish	Italian
Total # of toponyms	595	376	409
Accuracy without SI	57.14%	61.70%	45.96%
Accuracy with SI	87.39%	88.82%	75.3%

Table 9: Accuracy without and with spatial inference (SI)

5. CONCLUSIONS

This work has proposed a processing chain for the geoparsing and geocoding of texts containing travel descriptions, making a special emphasis on two main problems related to the geocoding part of the method : the existence of ambiguous toponyms, and the lack of gazetteers with enough coverage for fine-grain toponyms. The solution proposed for addressing these two problems has been based on the use of clustering techniques. On the one hand, the definition of clusters provides a map-based disambiguation approach to identify the clusters with the highest number of candidate toponyms in terms of distance. On the other hand, the bounding polygon of these clusters can be used as an estimate to define the location of those fine-grain toponyms not found in gazetteers.
Additionally, the feasibility of the proposed method has been tested for the geoparsing and geocoding of toponyms in a corpus of hiking descriptions obtaining good results in terms of accuracy, precision and recall. Moreover, it must be noted that the method performs well in a multilingual environment. The results obtained for the three different tested languages (French, Italian and Spanish) are comparable. The only requirement for the application of the proposed method to a set of travel descriptions in a new language is the customization of the geoparsing part and the availability of gazetteers for the geographic area covered by the texts in this new language.

However, some refinements could be included in the proposed method to increase its performance. On the one hand, additional heuristics could be taken into account to address the problem of reference ambiguity (i.e., several place names for the same place). In the current work we have considered the problem of structural ambiguity (i.e., finding or not finding sub-types within the toponym names in gazetteers). But other problems could be taken into account : variants of toponym names and types in other languages, abbreviations, etc. On the other hand, with respect to the adjustment of parameters in the application of the DBSCAN clustering method to deal with the problem of referent ambiguity, a possible refinement would be the automatic definition of parameter values by means of machine learning techniques as it is proposed in other works using clustering techniques [4, 10].

Finally, it must be noted that the proposed processing chain for geoparsing and geocoding could be applied to other types of text corpora. The proposed method is general enough to be applied for any kind of narrative descriptions in a small area. In the future we plan to test this method with different corpora referring to travel descriptions in other countries and with a higher level of granularity (e.g., travel tours across different countries).

Acknowledgments

This work has been partially supported by : the Communauté d'Agglomeration Pau Pyrénées (CDAPP) and the National Geographic Institute of France through the PERDIDO project ; the Spanish Ministry of Education through the International Excellence Campus Program (Campus Iberus mobility grants) ; the Aragon Government through the grant ref. B181/11 and the Pyrenees Working Community mobility grant ref. CTPM2/13 ; and the Keystone COST Action IC1302.

6. REFERENCES

[1] Semantic place localization from narratives. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, COMP '13, pages 16 :16–16 :19, New York, NY, USA, 2013. ACM.

[2] R. J. Agrawal and J. G. Shanahan. Location disambiguation in local searches using gradient boosted decision trees. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 129–136, 2010.

[3] A. Aji, X. Sun, H. Vo, Q. Liu, R. Lee, X. Zhang, J. H. Saltz, and F. Wang. Demonstration of Hadoop-GIS : a spatial data warehousing system over MapReduce. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 518–521, 2013.

[4] K.-H. Anders and M. Sester. Parameter-free cluster detection in spatial databases and its application to typification. *International Archives of Photogrammetry and Remote Sensing*, 33(B4/1 ; PART 4) :75–83, 2000.

[5] D. Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2) :16–19, jul 2011.

[6] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22(3) :301–313, Jan. 2008.

[7] D. Buscaldi and P. Rosso. Map-based vs. knowledge-based toponym disambiguation. In *Proceedings of the 2Nd International Workshop on Geographic Information Retrieval*, GIR '08, pages 19–22, New York, NY, USA, 2008. ACM.

[8] C. Derungs and R. S. Purves. From text to landscape : Locating, identifying and mapping the use of landscape features in a swiss alpine corpus. *International Journal of Geographical Information Science*, 28(6) :1272–1293, 2013.

[9] A. Eldawy, Y. Li, M. F. Mokbel, and R. Janardan. Cg hadoop : computational geometry in mapreduce. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 284–293, 2013.

[10] M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer, and X. Xu. Incremental clustering for mining in a data warehousing environment. In *VLDB*, volume 98, pages

323–333, 1998.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[12] U. Feuerhake and M. Sester. Mining group movement patterns. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 510–513, 2013.

[13] A. J. Florczyk, F. J. Lopez-Pellicer, P. R. Muro-Medrano, J. Nogueras-Iso, and F. J. Zarazaga-Soria. Semantic selection of georeferencing services for urban management. *Journal of Information Technology in Construction*, 15 (Special Issue Bringing urban ontologies into practice) :111–121, 2010.

[14] M. Habib and M. Van Keulen. Improving toponym disambiguation by iteratively enhancing certainty of extraction. In *KDIR 2012 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2012.

[15] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*, pages 401–410. ACM, 2010.

[16] S. Intagorn and K. Lerman. Learning boundaries of vague places from noisy annotations. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 425–428. ACM, 2011.

[17] N. Ireson and F. Ciravegna. Toponym resolution in social media. In *The Semantic Web–ISWC 2010*, pages 370–385. Springer, 2010.

[18] J. L. Leidner. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal-Publishers, Jan. 2008.

[19] J. L. Leidner and M. D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2) :5–11, July 2011.

[20] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740. ACM, 2012.

[21] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 201–212. IEEE, 2010.

[22] D. Maurel and N. Friburger. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313 :93–104, 2004.

[23] L. Moncla, M. Gaio, and S. Mustière. Automatic itinerary reconstruction from texts. In *Eighth International Conference on Geographic Information Science, GIScience 2014*, Vienna, September, 23-26.

[24] V. T. Nguyen, M. Gaio, and L. Moncla. Topographic subtyping of place named entities : a linguistic approach. In *The 16th AGILE International Conference on Geographic Information Science*, Leuven, Belgium, 2013.

[25] T. Poibeau. Extraction automatique d'information(du texte brut au web sémantique). 2003.

[26] T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang. An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 53–60, 2010.

[27] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1) :1, 2009.

[28] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[29] P. D. Smart, C. Jones, and F. Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In S. Fabrikant, T. Reichenbacher, M. Kreveld, and C. Schlieder, editors, *Geographic Information Science*, volume 6292 of *Lecture Notes in Computer Science*, pages 234–248. Springer Berlin Heidelberg, 2010.

[30] D. Smith and G. Mann. Bootstrapping toponym classifiers. *Association for Computational Linguistics*, Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1 :45–49, 2003.

[31] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Proceedings of the fifth conference on Applied natural language processing*, pages 202–208. Association for Computational Linguistics, 1997.

[32] X. Zhang, B. Qiu, P. Mitra, S. Xu, A. Klippel, and A. M. MacEachren. Disambiguating Road Names in Text Route Descriptions using Exact-All-Hop Shortest Path Algorithm. In *ECAI*, pages 876–881, 2012.

[33] J. Zhao, P. Jin, Q. Zhang, and R. Wen. Exploiting location information for web search. *Computers in Human Behavior*, 30 :378–388, 2014.