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

Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, Mauro Gaio

Automatic Itinerary Reconstruction from Texts

Ludovic Moncla[1,2], Mauro Gaio[1], and Sébastien Mustière[3]

[1] LIUPPA,
Avenue de l'Université Pau, France,
{ludovic.moncla,mauro.gaio}@univ-pau.fr
[2] Computer Science and Systems Engineering Department
Universidad de Zaragoza, Spain
[3] Université Paris-Est, IGN, Laboratoire COGIT,
73 av. de Paris, 94160 Saint-Mandé, France,
sebastien.mustiere@ign.fr

Abstract. This paper proposes an approach for the reconstruction of itineraries extracted from narrative texts. This approach is divided into two main tasks. The first extracts geographical information with natural language processing. Its outputs are annotations of so called expanded entities and expressions of displacement or perception from hiking descriptions. In order to reconstruct a plausible footprint of an itinerary described in the text, the second task uses the outputs of the first task to compute a minimum spanning tree.

Keywords: NLP, itinerary reconstruction, toponym resolution, spatial named entities recognition, expression of motion

1 Introduction

In the early nineties, Frank and Mark [1] wrote *"It is conceivable that systems of the future might be able to assimilate and analyze explorer's journals such as Columbus' logs or the journals of Lewis and Clark, check them for consistency, and perhaps reach new inferences about the itineraries of their travels"*.

Since then, advances in automatic natural language processing (NLP), processing and representation of geographical information, but also the explosion of open geographical resources, have made developing such systems now possible.

In this paper, we propose a system for automatically reconstructing an itinerary from textual descriptions occurring in travelogues and guides.

The problem can be subdivided into two tasks. The first task entails annotating passages in the text that describe the various trips making up the itinerary. The second task entails creating a computational representation of the different descriptions, thus allowing the itinerary to be automatically reconstructed. Crucial to implementing such systems is the step known as toponym resolution. This essential step involves associating a non-ambiguous location with a place name[4].

[4] Either a point, or a spatial footprint, in both cases expressed as geographic coordinates.

Consider for example the following text from a true description of a hiking trail:

> "Cross Champagny-le-Haut and get around from the north of hamlet Friburge. You will see the Lac de la Plagne then walk to the refuge south of Lake Grattaleu." (1)

The proposed system proceeds as follows: the first task annotates the expanded spatial named entities *Champagny-le-Haut*, *hamlet Friburge*, *Lac de la Plagne* and *refuge south of Lake Grattaleu*. Some are associated with terms like *hamlet*, *lake*, *refuge* and/or spatial relations *from the north*, *south of*, allowing the ambiguity to be removed from the nature of the geographic objects in question. This task will annotate some others spatial relations: *cross*, *get around*, *walk to*. Once all this information has been annotated, we apply a spatial analysis algorithm, guided by various clues obtained from the text, to reconstruct the itineray. In this example, the expanded spatial named entities *Champagny-le-Haut*, *hamlet Friburge* and *refuge south of Lake Grattaleu* would be given priority while reconstructing the itinerary, as the spatial relations detected in the text imply the involvement of these entities in the itinerary to be taken. On the other hand, the spatial named entity *Lac de la Plagne* is not involved in the path, as the use of the verb *see* suggests.

This paper is set out as follows: Section 2 presents an overview of pertinent studies relating to the issue at hand; Section 3 describes our own contribution, proposal of a method of annotating expanded spatial entities, together with a method for automatic itinerary calculation; Section 4 describes our implementation and relates the early results of our experiments. Finally, in Section 5 we conclude the paper and propose future studies.

2 Relevant Work

2.1 Spatial Named Entity Recognition and Motion Expressions

Extraction and annotation of named entities is an important task in NLP, particularly in the case of automatic information extraction [2]. For named entity recognition and classification (NERC), two types of approaches have been proposed, those that use learning techniques and those based on ad-hoc rules. In the case of annotation of spatial entities in particular, there are also approaches that use external resources like gazetteers to search for and identify toponyms. These approaches can be used in a complementary manner in hybrid systems [3]. The ad-hoc approach relies on syntactic-semantic patterns developed manually with the help of experts. Amongst these rule-based approaches, several use transducers[5] with a finite number of states [2], which can also be used in cascade [4]. NERC methods automatically annotate different types of named entities: dates,

[5] Transducers are a type of finite-state machine that make insertions, replacements and deletions in a text.

people, organisations, themes, numeric values, as well as place names. There are a significant number of systems available, both proprietary and open source, such as OpenNLP[6] from Apache, OpenCalais[7] from Thomson Reuters, and CasEN [4]. More specific methods that are solely concerned with geographical data are known as geoparsing or toponym recognition [5]. The main difficulty in extracting geographical information is the ambiguity inherent in natural language. As stated in the introduction, there are actually several types of ambiguity involved in toponym resolution. In addition, a large number of spatial entity types exist: geopolitical entities (countries, administrative divisions), populated places (towns, addresses and postal codes), and natural geographical entities (parks, valleys, mountains, rivers, etc.), all of which can create ambiguities about the type of geographic object in question.

In itinerary analysis, it is not just spatial named entities that are important, but also their associated spatial relations. These enable the spatial named entity to be specified locally, as well as allowing the notion of movement between the different entities to be expressed. Many linguistic studies [6, 7] deal with spatial relations with a view to describing the object to be located and the point of reference used. For French in particular, we could cite Vandeloise [8] with the term pair cible (target) and site (site), and Borillo [9] with the term pair entité concrète (concrete entity) and repère spatial (spatial reference). According to Talmy [10, 11], a motion event is characterised by different conceptual components: a movement ('Motion'), a displaced object ('Figure'), a setting ('Ground'), a trajectory ('Path') and a 'Manner'. Syntactic parts of speech, in particular verbs, characterise a motion event. Many linguistic studies [12–14] have highlighted the importance of the use of motion verbs in language, especially in Romance languages. These studies suggest categorising motion verbs according to their aspectual polarity. The three polarities are initial (e.g. to leave), median (e.g. to cross) and final (e.g. to arrive). The works also show the importance of the prepositions associated with these motion verbs. Without changing the intrinsic polarity of the verb, the preposition can change what it would be called the *focus*. More specifically, the association of a motion verb with a preposition of place (e.g. *from*, *in*, *at*, *to*, *by*, etc.) can change the focus of the displacement to take on the polarity of the preposition instead of the verb. Let us take the verb *to leave* for example. Alone or in association with the preposition *from*, the focus would be considered with initial polarity, but if used with the preposition for, the focus would then be considered as having final polarity. Undeniably, *leaving from Vienna* and *leaving for Vienna* have two radically opposite focus. If we consider the role played by the name, in one case, the place name is the origin of the displacement, and in the other case the place name is the destination. In the example *leaving for Vienna* it doensn't mean *to arrive in Vienna*, because we don't know if the destination is reached or not, but we know that we are leaving a place to go to *Vienna*. In terms of place name *Vienna* is the focus, so the polarity of the whole expression may be considered as final.

[6] `http://opennlp.apache.org/`
[7] `http://www.opencalais.com/`

The use of contextual elements (other than toponyms), such as words that have a geographical denotation (downtown, valley, ridge, etc.), can be extremely important in toponym resolution and disambiguation [15]. In previous studies [16], we have put forward a method of marking non-toponymic terms associated with toponyms, especially those that have a topographical denotation (Wachau Valley, Lake Neusiedl, Saifnitzer Sattel, etc.).

2.2 Toponym Resolution

Toponym resolution [5] involves associating a non-ambiguous location with a place name. The use of resources like gazetteers is thus vital. In the last few years, we have seen a number of geographical resources emerge, such as Geonames[8], OpenGeoData[9], OpenStreetMap[10], Wikimapia[11], and BDNyme[12]. In an open data context, and with some benefiting from participative communities, these resources are expanding and being made more widely available through Web services. Some of these web-based geographic services are free, interoperable, and standardised, but the number and diversity of platforms makes using the data a complicated process. Before being able to use this mountain of data, first the most appropriate resources must be selected according to actual needs. Each resource can have different issues, for example the choice between a resource that covers a wide area but non-exhaustively and a more exhaustive resource covering a smaller area.

This resolution involves resolving the problem of ambiguities that toponyms may contain. Widely studied in recent years, the admittedly difficult task of toponym disambiguation remains a scientific problem today. According to [17] 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), the place name can be used in a non-geographical context, as in organisations or names of people (referent class ambiguity).

Using a corpus of hiking guides naturally reduce the number of ambiguities from referent class, as opposed to those used in a corpus of news articles for example. Then in this paper, we will focus on ambiguities resulting from the referent ambiguity class arised from the existence of homonyms (e.g. in a french formulation *Vienne* exists in Austria but *Vienne*, also exists in France) or arised from subtyping of toponyms [16]. Even once it has been identified that the reference is to the place named *Vienne* in France, ambiguities may remain concerning the geographic object that carries the name (*Vienne* the town, *Vienne* the county or *Vienne* the river). Another form of ambiguity arises from the presence of certain spatial expressions associated with the place name (e.g. *Paris-Nord railway station* is different from *the railway station in the north of Paris*).

[8] `http://www.geonames.org/`
[9] `http://www.opengeodata.fr/`
[10] `http://www.openstreetmap.org/`
[11] `http://wikimapia.org/`
[12] `http://www.geoportail.gouv.fr/`

A number of methods exist for disambiguating toponyms [18–21]. These methods can be classified into three categories [22]: map-based, knowledge-based, and data-driven or supervised. Many of these methods use toponyms that are geographically the closest to disambiguate the candidate toponyms. This can lead to poor results when important information is not included in the context, when the candidate toponym is not geographically close to non-ambiguous toponyms, or if it is not linked to a geopolitical entity [21]. Some studies use the notion of event to disambiguate toponyms. For example, Robert et al. [20] consider there to be three types of entities that participate in an event: people, organisations, and geopolitical entities. They use an ontology constructed from Geonames, and associate geopolitical entities with people and organisations using links from Wikipedia, but no other information or clues is used from the context. Knowledge-based methods use toponyms information extracted from gazetteers like importance, size, or population counts [23]. This kind of information is not the most suitable for a discriminating task in the case of documents describing hiking trails, because toponyms used in these documents are fine-grain toponyms or natural features such as mountains, lake, hamlet, and refuge [24].

These various methods are often applied to corpora of news articles [18, 19, 21] in which toponyms are associated with events, well-known figures or geopolitical entities and not with spatial relations. In this type of discourse, toponyms are not necessarily related to each other, and are not for example linked by motion events. Speriosu and Baldridge [21] show that toponym disambiguation methods that are based on the text (context extraction and interpretation of spatial relations) are more effective than methods based on metadata or heuristics that use distance calculations.

2.3 Wayfinding

With the rise of new needs and behaviours (e.g. route planning and tourist applications), the democratisation of devices equipped with GPS and the wide availability of geographical information, the notion of itinerary is being studied more and more. Hao et al. [25] put forward a probabilistic model to identify place elements taken from travelogues. The aim of this work is to improve the tourist experience, providing them with information or recommendations about the places they are visiting. Zhang et al. [26] use these learning methods to extract the three elements they consider to be the most important in an itinerary: origin, destination, and the path taken (instructions). They work from a corpus of webpages where instructions giving directions can be found.

Other studies [27] have focused on ancient documents with the aim of finding and modelling historical roads that no longer exist. A large number of studies have looked into trajectories [28–30], with a focus on the movement of mobile objects (animal migrations, aeroplanes, ships, pedestrians, etc.). The concepts explored in these studies can be considered similar to those applicable to itineraries.

The notion of itinerary has already been a focus of research for our team. In previous studies [31], Loustau proposed a definition of the concept of itinerary, but more importantly contributed to the proposal of an initial approach to extracting constitutive information about itineraries.

3 Contribution

3.1 Problem elucidation

Our aim is to identify geographical information in a text, as well as any textual clues that allow us to link some and exclude other information that should not be taken into consideration and then map the most likely route. In this study, we chose to test our approach on a french corpus of 1,295 descriptions of mountain bike and hiking trails in France.

As mentioned in the literature review (Section 2), spatial relations are an important factor in the disambiguation of toponyms. Particularly in forms of discourse, as can be found in our corpus and those of same categories where spatial relations exist on several levels of granularity and can be applied to the discourse at different scales. In this paper, we will examine two levels of granularity. The first involves local spatial relations that are part of a spatial named entity. To illustrate our discourse, let's take the next example (1) page 2. In the spatial named entity *south of Lake Grattaleu* the spatial relationship contained *south of* needs to be interpreted in order to solve the ambiguity of the referent. The second level involves spatial relations that associate various spatial named entities or a participant with a spatial named entity relative to another. In case of a hiking trail, the object under consideration is the participant in the motion event. In *Cross Champagny-le-Haut and get around from the north of hamlet Friburge* the spatial relations are *the north* and *around*, but also the motion verb *to cross* and *to get*. This is a description of a motion event relative to spatial named entities. Finally the toponym *Lac de la Plagne* is associated with a perception verb *to see*, which means that the toponym is not really part of the itinerary taken but a visual landmark. Moreover, the term *lake* serves to precisely identify the toponym, removing all ambiguity from the geographic object being referred to, i.e. it is most definitely a lake, and not a town for example.

3.2 Solution adopted

The first step in our approach is a system whereby 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[13] [33] or external resources (lexicons, gazetteers, etc.). The second step in our approach is a system capable of interpreting and linking information in order to automatically reconstruct an itinerary.

[13] These grammars are lexicalised graphs that make use of dictionaries and have the advantage of being able to be applied directly to texts. [32]

Expanded spatial named entities We define an *expanded spatial named entity* (*ESNE*) as an entity built from a proper name attributed to a place. This proper name can be associated with one or more ontological concepts with a geographical sense, and with one or more concepts relating to the expression of location in the language (spatial relations). For example *the north of hamlet Friburge* and *the refuge south of Lake Grattaleu* are two *ESNE* built from proper names (*Friburge* and *Grattaleu*), associated with ontological concept having a geographical sense (*hamlet*, *refuge* and *lake*), and spatial relations (*the north of* and *south of*).

We integrate spatial relations into a more generic concept that we have called *indirections* allowing a geographic object to be addressed indirectly. Indirections can be a part of the concrete entity, their role being to specify location, and grammatically they can belong to different word classes (prepositions, adverbs or adjectives). For example, in the *ESNE the north of hamlet Friburge*, as the toponymic name is the word *Friburge*, the concrete entity is *the north of hamlet*, which contains the indirection *the north*. Other parts of speech are annotated to reveal spatio-temporal sequences in the discourse, such as spatial adverbs of location [34]. These are prepositions of place, which occur frequently in hiking guides (e.g. here, there, near of, from the left, etc.). These prepositions structure the discourse by describing a spatial sequence (a step in a journey) and/or a temporal sequence (a succession of events).

Unlike traditional named entity annotation tools, we only annotate spatial named entities. We follow the proposal of Gaio et al. [35] for the recognition of spatial references and spatial relations in language, as well as using a hybrid approach [36] combining the three main categories of spatial relations: topological relations [37], distance relations, and directional relations [38]. In order to establish the steps of an emerging route, itinerary is defined as being a special type of spatial relation. It is a spatio-temporal sequence of steps moving between different places. An itinerary could thus be thought of as a succession of spatial relations.

Expression of motion Based on preposition polarity and classification of verbs, we are able to establish simple linguistic rules in order to extract the source or target named entities in a motion event. We classified verbs into categories: motion verbs (*to go*, *to leave*, etc.), verbs of visual perception (*to glimpse*, *to see*, etc.), verbs we refer to as topographic (*to converge*, *to overhang*, etc.), topographic verbs are used when the narrator is describing a place using its topographical features. And finally location verbs (*to locate*, *to be*, etc.) [39]. The use of this last class of verbs, from a syntactic point of view, is very similar to that of motion verbs as previously described. They are often associated with prepositions of location or place names in hiking guides. They can be used, for example, to describe a step or stop in a journey. They also allow for better spatial representation and facilitate the location of the different events relative to each other. In order to formalise the relationship between expanded spatial named entities and verbs in travelogues, we use *VTo structures* [35, 16]. *VTo*

structures are formally defined as *V, I, T, G*: groups of classified verbs, indirections, geographical terms, and place name, respectively. $VTo = (v, t)$ where v is an instance of V and the t set is defined as $t = (te, i, g|t)$, in such a way that te is an instance of T, i is an instance of I and g is an instance of G. The symbol $g|t$ indicates that the third t group can be made up of either t (recursion) or g.

Here is a *VTo structure* from example (1):
{{get around,.v}{{from the north of,.i}{{hamlet,.te}{Friburge,.g},.t},.t},.VTo}

Itinerary reconstruction This final step entails using the information annotated in previous steps to reconstruct a plausible footprint of the itinerary.

Toponym Resolution The information gathered from the marking process during the annotation step includes candidate spatial entities and candidate *VTo structures*. We use gazetteer-style external resources to verify the existence of toponyms and obtain their geographic coordinates. When toponyms exist in one of the resources they are validated. If the validated toponyms are part of a candidate expanded spatial entity, the entity is automatically validated. However, ambiguities still occur, for example when an entity is made up of several words and only one of those is validated. Toponyms may also be associated with several locations when the name occurs several times in the resources with different locations, e.g. *Pau* occurs nine times in BDNyme and three times in Geonames. In addition to those ambiguities, one may notice that every spatial named entity mentioned is not necessarily part of the itinerary. Some clues extracted from the text, such as negations, descriptions of places and scenic lookouts (e.g. with the use of perception verbs), allow us to conclude that some places should probably not be included in the representation of the itinerary, or only in a certain way.

Itinerary calculation This step combines contextual information extracted from the text with geometric information extracted from gazetteers. This combined spatial and textual analysis aims at resolving some ambiguities and reconstructing the footprint of the itinerary. The approach proposed here is based on the idea that the most probable itinerary linking a set of places is the route linking all places and with a minimal length, to "optimise" the displacement. Finding this optimal itinerary should help to remove ambiguities or places appearing in the text but not actually crossed. This naturally leads to the notion of 'minimal weighted spanning tree'. The minimum weight spanning tree of a set of points is the tree connecting all the points together with the minimum weight, this weight being the sum of the weights of the edges linking points (e.g. equals to the distance between points). The implementation of the continuity detection with the notion of minimum spanning tree is not a new idea. It has already been developed for example by Zahn [40] for detecting clusters of points.

4 Implementation

As described in Section 3, we developed a two-step solution. The first step (Fig. 1a) entails annotating toponyms and the various spatial relations (indirections,

expressions of motion). The second step (Fig. 1b) entails calculating an itinerary between the different toponyms previously annotated. We will now explain the implementation of our method in more detail.

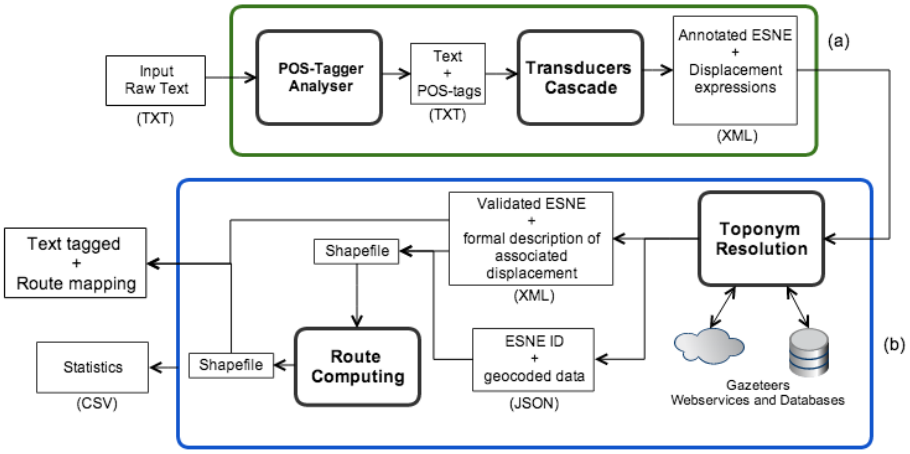


Fig. 1: Block diagram of our processing chain

4.1 Spatial annotation for itinerary calculation

We developed an annotation processing chain[14] (Fig. 1(a)) that takes a raw text (written in natural language) as input. The main annotation module was designed using the finite-state transducer cascade creation program CaSys [4] available on the Unitex platform[15]. Transducers are represented by graphs on the Unitex platform, which simplifies both writing and maintenance. The cascade allows all or some of the elements labelled by preceding transducers to be used in those that follow. The cascade designed for annotation comprises six main transducers that mark the text in the following order: indirections, candidate toponyms, classified verbs, candidat terms with a geographical denotation, *ESNE* and *VTo structures*.

Let us illustrate the execution of this cascade with an example. Take the sentence from example (1): "walk to the refuge south of Lake Grattaleu." The **indirection** transducer will be applied first. Indirections (*at*, *nearby*, *south*, etc.) are sought out with the help of lexicons and linguistic patterns (*the southern part*, *at the center of*, etc.) that make use of these lexicons. Our first main transducer gives the following output:

[14] Demonstration tool available online: `http://erig.univ-pau.fr/PERDIDO/`
[15] `http://www-igm.univ-mlv.fr/~unitex/`

Walk to the refuge **{{**south**,.directional}** of**,.indirection}** Lake Grattaleu.

The second transducer of the cascade marks **candidate toponyms**. This transducer is designed to recognise the various complex forms of toponym construction. There even exist toponymic guidelines published by IGN[16]. For example, toponyms can be composed and formed of several distinct terms that may be accompanied by an article (e.g. *Champagny-le-Haut*). In our example the toponym is tagged like this: **{**Lake Grattaleu**,.toponymCandidat}**

The third transducer annotates **classified verbs**. It relies on subgraphs designed to label verbs according to different categories (motion verbs, position verbs, perception verbs, and topographic verbs). For motion verbs this transducer also specify the polarity of the verb (initial, median, or final). In our example the verb is tagged like this: **{**Walk**,.motionVerb+median}**. The fourth transducer annotates **common nouns** or common noun phrases. These will then be identified or not as candidate terms with a geographical denotation, thanks to *VTo structures*. Execution of this transducers tag common noun as follow: **{**the refuge**,.commonNoun}** The fifth transducer of the cascade annotates **candidate expanded spatial named entities**. The indirections, candidate toponyms, and candidate common nouns with a geographical denotation have already been annotated by the preceding tranducers. This transducer uses these previously carried out annotations. Execution of this transducer finds this *ESNE*: **{**{the refuge,.commonNoun} {{south,.directional} of,.indirection} {Lake Grattaleu,.toponymCandidat}**,. ESNE}**.

The final transducer in our cascade annotates **VTo structures**. It behaves in the same way as the previous transducer, in that it uses previously made annotations.

In our example, the observed structure is composed of a motion verb and an expanded spatial named entity. Execution of our cascade of transducers gives the following final output:
{{Walk,.motionVerb+median} to {{the refuge,.commonNoun} {{south,. directional} of,.indirection} {Lake Grattaleu,.toponymCandidat} ,.ESNE}**,.VTo}**.

The different annotations (toponyms, spatial named entities, *VTo structures*) are candidate annotations, i.e. they require a further step of validation. This is done next during toponym resolution.

4.2 Experimental results for the first step

Our first experiment was to run our automatic tagging chain for french texts on a body of reference in order to compare the results obtained automatically with those obtained manually.

For the moment we have chosen the resources Geonames and BDNyme as they complement each other. BDNyme is the French benchmark toponymic database provided by IGN. It lists 1,500,000 place names resulting from toponyms and georeferenced activities and points of interest. Geonames may only

[16] `http://www.ign.fr/sites/all/files/charte_toponymie_ign.pdf`

list around 135,000 names on French soil, but it is especially useful for cross-border trail descriptions where part of the descriptions may refer to locations outside France.

To evaluate our method, we wanted to know the rate of correct recognition of spatial named entities and if they are present or not in a *VTo structure*. The aim is also to identify errors and causes in order to subsequently improve our processing chain.

Currently our body of reference is composed of 24 randomly selected hiking guides manually annotated (ground truth) over 1295 available[17]. Each document of our corpus has an average of:

- 263.3 words (269.2 on the body of reference) with a standard distance of 188.5 (242.9 for the body of reference) ;
- 12.12 candidate toponyms (13.87 on the body of reference) with a standard distance of 9.88 (12.71 for the body of reference) ;
- 4.72 candidate toponyms included in a *VTo structure* (5.46 on the body of reference) with a stadard distance of 4.14 (4.31 for the body of reference).

The corpus has an average of 5.09% of candidate toponyms for 100 words (5.44% on the body of reference). Furthermore 39.33% of candidate toponyms are in a *VTo structure*. Figure 2 shows for each document of the body of reference, the number of candidate toponyms that are included in a *VTo structure* over the total number of candidate toponym.

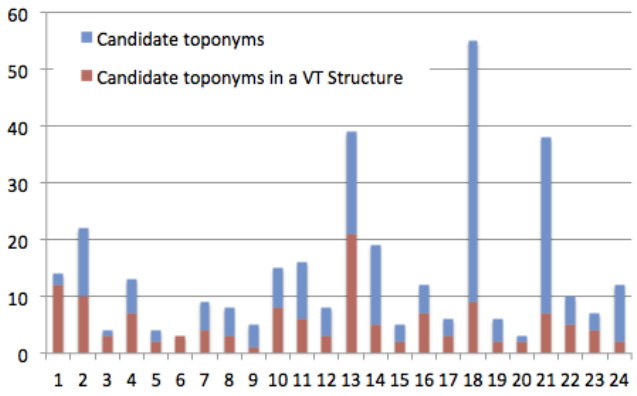


Fig. 2: Number of candidate toponyms in *VTo structure* over the total number of candidate toponyms per document

[17] We are now developing a tool for a controlled manual tagging in order to easily enrich the body of reference.

Of the 366 spatial expanded named entities present in the body of reference, 325 are correct recognitions and we get 49 false recognition which gives us a precision[18] of 85.37% and a recall[19] of 88.80%.

41 toponyms (11.20%) were not detected by the processing chain. Detection errors (false recognition or non-detection) are due to several problems. The main problem is the morpho-syntactic analyser that tag proper names as common names. In this case if a proper name is considered as a common name, the annotation transducers of *ESNE* is not triggered. Some bad recognitions are also due to errors in the linguistic patterns described in the transducers, for example: missing rules for road names (like *GR55, A10, M25, etc*), this last kind of errors where corrected during the experiments.

49 toponyms (14.63%) are false detection, this problem is mainly a problem of ambiguities (eg, toponyms recognized instead of names of people or organizations). 46.97% of the toponyms are in a *VTo structure*. These figures are partly explained by the fact that many cases are not listed as *VTo structures* (examples: *Continue to...*, *Take direction...*, etc). In addition, 59.33% of *VT structures* are not composed of toponyms. But they are composed of common nouns referencing toponyms, we call this *VTr structures*. This important amount of *VTr structures* can be explained by the type of corpus studied. Indeed a number of descriptions of hiking use less of toponyms and more of landscape features or spatial relationship (for example: "walk along the trail far as the church"). Finally, 69.93% of well localized toponyms are detected by at least one resource. Among these toponyms localized, 63.50% are located in the two resources, 19.00% are geo-coded only by BDNyme and 17.50% are geo-coded only by Geonames. The figures about the toponyms resolution take no account of ambiguities. These first results concerning the annotation of expanded spatial named entities in a corpus of real descriptions of hikes, are very encouraging. Following this first experiment our objective is to correct and enrich some transducers.

4.3 Itinerary calculation

The second step (Fig. 1(b)) in our method entails reconstructing an itinerary from the elements annotated in the first step. It can be divided into two tasks. The first task is toponym resolution, where candidate toponyms are validated. The second task is the true itinerary calculation step.

For toponym resolution, we developed a module that consults external resources to validate the existence of toponyms and obtains their geographic coordinates. This module takes as input the output of our transducer cascade. This step further improves the content of our annotated document (using XML markup), and thanks to unique identifiers links it to the validated and geocoded spatial named entities (in a separate document in JSON format). The toponym validation program also generates a file in shape format containing information

[18] fraction of retrieved instances that are relevant
[19] the fraction of relevant instances that are retrieved

on the spatial entities, such as the name, geographic coordinates, associated classified verb and its polarity. In this way, all additional information can be easily added to one of the three documents according to its purpose.

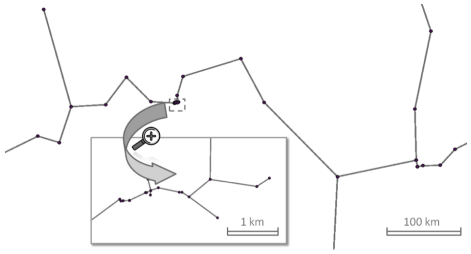
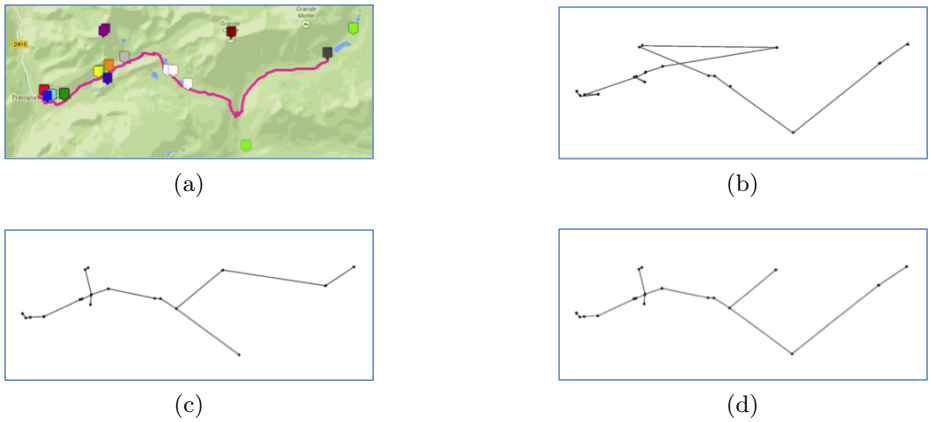


Fig. 3: Disambiguation based on the length of edges of the tree

As suggested before, we compute a minimum spanning tree algorithm in order to link the various located spatial entities. A first approach is to directly weight edges of the tree with the Euclidian distance between places. As exemplified in Figure 3, this tree can be use to disambiguate toponyms: in the actual area of interest (zoomed area in the figure), actual places are close to each other, and a simple selection of places linked by short edges in the tree is useful to focus on the area of interest and reject other toponyms.



(a) (b) (c) (d)

Fig. 4: Actual displacement recorded by GPS and spatial entities detected in the associated text (up), and itineraries built with different strategies (b,c,d)

Once this focus on the area of interest is done, the tree may be used to reconstruct an approximation of the actual itinerary. Figure 4 illustrates this idea applied to the description of hiking route, associated with a 'ground truth' of the itinerary collected by GPS. Reconstructing the itinerary by simply ordering places as they appear in the text is of course inefficient because the discourse is rarely so linear (Fig. 4(b)). The result of the automatic creation of a minimum spanning tree on these locations is more efficient: the longest line of this tree is a first approximation of the itinerary (Fig. 4(c)). However, the built itinerary goes through places seen but not passed through. If one computes the tree minimising the length weighted by information automatically extracted from the text, the result can be improved: one may under-weight edges linking places associated to a displacement verb, and over-weight edges linking places associated to a perception verb (Fig. 4(d)). In this example, the built itinerary is close to the actual one.

Those first experiments illustrate that neither language analysis alone, nor spatial analysis alone, may be sufficient. However a combination of those analyses is promising and should be explored more in depth to build itineraries or other spatial configurations from texts. Such a method could be enriched to consider other information issued from the text analysis, such as the polarity of verbs, negative forms, and so on.

5 Discussion and outlook

This paper proposes an approach to reconstruct a plausible footprint of an itinerary extracted from narrative texts. This approach was divided into two main tasks. The first annotates expanded spatial named entities and expression of displacement or perception from textual hiking descriptions. The second task is a method to reconstruct itineraries as from the elements marked in the first stage. This method computes a minimum spanning tree weighted using information extracted from the text.

According to preliminary tests, the results are encouraging. The use of evidence extracted from the text (eg. verbs of displacement, verbs of perception, spatial relationship etc) associated with the names of places improves the results of the minimum spanning tree. Furthermore our first observations show that disambiguation of names can be partly solved by the route calculation in the context of a body of descriptions of hiking. However, the first evaluation of our tagging method has identified some errors. A first short-term perspective is to increase the accuracy of the first task by refining transducers.

But our main goal is to improve the second task. Unlike many studies that use only metadata from geographic resources (e.g. number of inhabitants, area, etc.), we believe that information extracted from texts allows a better interpretation. As a first experimental approach we used the Euclidian distance to compute distances between places but it could be more efficient to use other methods like travel effort or a combination of information like the affordance. In addition to the concept of continuity already used in our calculation method of route, we

plan to introduce in the method the two others laws of the gestalt theory (the similarity/complementarity and the proximity). Furthermore, the possibilities of interaction that a human could have with a named spatial object change the interest addressed to that object. This is why, like [41], we plan to integrate in our method the notion of affordance. According to the authors the affordance of a place consists of the following aspects: physical features, actions, narrative, symbolic representations, or socioeconomic and cultural factoring typologies. All these aspects are strongly present in the itineraries descriptions.

References

1. Frank, A.U., Mark, D.M.: Language issues for GIS. In: Geographical Information Systems: principles and applications. Essex: Longman Scientific & Technical (1991) 147–163
2. Poibeau, T.: Extraction automatique d'information: du texte brut au web sémantique. In: Extraction automatique d'information: du texte brut au web sémantique. Hermès Lavoisier (2003)
3. Béchet, F., Sagot, B., Stern, R.: Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In: TALN'2011, Montpellier, France (2011)
4. Maurel, D., Friburger, N.: Finite-state transducer cascades to extract named entities in texts. Theoretical Computer Science **313** (2004) 93–104
5. Leidner, J.L.: Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal-Publishers (January 2008)
6. O'Keefe, J.: The spatial prepositions in english, vector grammar, and the cognitive map theory. Language and space (1996) 277–316
7. Bloom, P.: Language and space. MIT press (1999)
8. Vandeloise, C.: L'espace en français. Seuil, Paris (1986)
9. Borillo, A.: L'espace et son expression en français, l'essentiel. In: L'espace et son expression en français, L'essentiel. Orphrys (1998)
10. Talmy, L.: Lexicalization patterns: Semantic structure in lexical forms. Language typology and syntactic description, vol. 3, Grammatical categories and the lexicon, ed. by Timothy Shopen, 57–149. Cambridge: Cambridge University Press (1985)
11. Talmy, L.: Toward a cognitive semantics. In: Toward a Cognitive Semantics. The MIT Press (2000)
12. Boons, J.P.: La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. Langue Française (76) (1987) 5–40
13. Slobin, D.I.: Two ways to travel: Verbs of motion in english and spanish. Grammatical constructions: Their form and meaning (1996) 195–219
14. Aurnague, M.: How motion verbs are spatial: The spatial foundations of intransitive motion verbs in french. Lingvisticae Investigationes **34**(1) (2011) 1–34
15. Hollenstein, L., Purves, R.: Exploring place through user-generated content: using flickr todescribe city cores. Journal of Spatial Information Science (1) (2010)
16. Nguyen, V.T., Gaio, M., Moncla, L.: Topographic subtyping of place named entities: a linguistic approach. In: The 15th AGILE International Conference on Geographic Information Science, Louvain, Belgique (2013)
17. Smith, D.A., Mann, G.S.: Bootstrapping toponym classifiers. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 45–49

18. Garbin, E., Mani, I.: Disambiguating toponyms in news. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 363–370
19. Buscaldi, D., Magnini, B.: Grounding toponyms in an italian local news corpus. In: Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR '10, New York, NY, USA, ACM (2010) 15:1–15:5
20. Roberts, K., Adrian Bejan, C., Harabagiu, S.: Toponym disambiguation using events. In: Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010). (2010) 271–276
21. Speriosu, M., Baldridge, J.: Text-driven toponym resolution using indirect supervision, Sofia, Bulgaria (August 2013)
22. Buscaldi, D.: Approaches to disambiguating toponyms. SIGSPATIAL Special **3**(2) (July 2011) 16–19
23. Overell, S., Rüger, S.: Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science **22**(3) 265–287
24. Derungs, C., Purves, R.S.: From text to landscape: locating, identifying and mapping the use of landscape features in a swiss alpine corpus. International Journal of Geographical Information Science 1–22
25. Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: Proceedings of the 19th international conference on World wide web. WWW '10, New York, NY, USA, ACM (2010) 401–410
26. Zhang, X., Mitra, P., Klippel, A., MacEachren, A.: Automatic extraction of destinations, origins and route parts from human generated route directions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **6292 LNCS** (2010) 279–294
27. Breier, M.: The way is the Goal–Modelling of historical roads. In: 26th International Cartographic Conference. (August 2013)
28. Lee, J.G., Han, J., Li, X.: Trajectory outlier detection: A partition-and-detect framework. In Alonso, G., Blakeley, J.A., Chen, A.L.P., eds.: ICDE, IEEE (2008) 140–149
29. Kim, J., Sridhara, V., Bohacek, S.: Realistic mobility simulation of urban mesh networks. **7**(2) 411–430
30. Yuan, Y., Raubal, M.: Extracting dynamic urban mobility patterns from mobile phone data. In Xiao, N., Kwan, M.P., Goodchild, M.F., Shekhar, S., eds.: Geographic Information Science. Number 7478 in Lecture Notes in Computer Science. Springer Berlin Heidelberg 354–367
31. Loustau, P., Nodenot, T., Gaio, M.: Spatial decision support in the pedagogical area: Processing travel stories to discover itineraries hidden beneath the surface. In: AGILE Conf.
32. Constant, M.: Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion. PhD thesis, Université Paris-Est (2003)
33. Gross, M.: The Construction of Local Grammars. In Schabès, E.R..Y., ed.: Finite-State Language Processing. MIT Press (1997) 329–354
34. Borillo, A.: Quand les adverbiaux de localisation spatiale constituent des facteurs d'enchaînement spatio-temporel dans le discours. In: Information Temporelle, Procédures Et Ordre Discursif, Genève (2004) 123–138
35. Gaio, M., Sallaberry, C., Nguyen, V.T.: Typage de noms toponymiques à des fins d'indexation géographique. TAL **53** (2012) 1–35

36. Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaà, C., Lesbegueries, J.: A global process to access documents' contents from a geographical point of view. Journal of Visual Languages & Computing **19**(1) (2008) 03–23
37. Egenhofer, M., Franzosa, R.: Point-set topological spatial relations. International journal for Geographical Information Systems **5**(2) (1991) 161–174
38. Frank, A.U.: Qualitative reasoning about distances and directions in geographic space. Journal of Visual Languages and Computing **3**(4) (1992) 343–371
39. Borillo, A.: A propos de la localisation spatiale. Langue française **86**(1) (1990) 75–84
40. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. Computers, IEEE Transactions on **100**(1) (1971) 68–86
41. Abdalla, A., Frank, A.U.: Combining trip and task planning: How to get from a to passport. In Xiao, N., Kwan, M.P., Goodchild, M.F., Shekhar, S., eds.: Geographic Information Science. Number 7478 in Lecture Notes in Computer Science. Springer Berlin Heidelberg 1–14