Towards Heterogeneous Resources-based Ambiguity Reduction of Sub-typed Geographic Named Entities

Mauro Gaio, Tien Nguyen Van

Towards Heterogeneous Resources-based Ambiguity Reduction of Sub-typed Geographic Named Entities

Mauro GAIO and Van Tien NGUYEN
mauro.gaio@univ-pau.fr, vantien.nguyen@univ-pau.fr

Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex

Abstract. The aim of this work is to find sub-typed Geographic Named Entities from the analysis of relations between Place Names surrounded nominal group within a specific phrasal context in a set of textual documents. The paper presents a method involving natural language processing and heterogeneous resources like gazetteers, thesauri or ontologies. The work and the results focus a French language corpus. However, the uses of quite generic lexico-syntactic patterns in pre-selected phrasal context can be tuned for others languages.

Keyword : natural language processing, Named Entity categorization, verbal relations, lexico-syntactic patterns, finite-state transducers

1 Introduction

Many applications such as automatic ontology creation, enrichment from text, information extraction, automatic indexation for digital libraries and question answering applications rely on different types of approaches. Currently, the first task consists in finding key terms (such as named entities and associated technical terms) in text of selected repository of documents and then use them as seeds for the next process. In automatic ontology enrichment these key terms are related to concepts in the target ontology.

Linguistic processing system using rule-based grammar and lexical resources may carry out this core task. Generally, this processing is combined with the recognition and classification of Named Entities [Nan98] and traditionally this process classifies named entities into persons, organizations and places. At the same time, research on analysis of geographic references is becoming a hot topic in the research area of information retrieval. So, "... *Finding geographical references in text is a very difficult problem and there have been many papers that deal with different aspects of this problem and describe complete systems such as Web-a-where, MetaCart, and STEWARD ...*" [BLPD10].

However, in NER systems places entities are not classified in their specific sub-types and it is essentially due to the difficulty of the task [LL07]. In this paper, we focus on the identification and the geographical named entities subcategorization, gathered with two classes of additional lexical information. The

first class indicates a contextual geographic focusing and the second one indicates the sub-type e.g. "the EASTERN PART(1) of the Aspe VALLEY(2)". The second sub-class has been already proposed in [MTV07]. By subcategorizing location entities, we propose a method to reduce ambiguities, retaining these key terms as seeds for the next step of a given process.

The proposed method involves lexico-syntactic patterns and external heterogeneous resources such as gazetteers, non-specific thesaurus, ontologies and event structures [RBH10] expressed in a finite-state description. While the use of such patterns essentially is not new in itself (for example Hearst patterns [HEA92] unlike most previous work, refining Hearst initial patterns or defining patterns within a very specified domain, in these work we investigate the uses of quite generic lexico-syntactic patterns in pre-selected phrasal context.

We are currently working on a repository of documents in French from the nineteenth century, devoted to the Pyrenees (especially travel stories). A travel story is a genre in which the author describes one or more travels, people encountered, emotions, things seen and heard. These documents contain numerous geographic references to typed geographic named entities of a defined area (French Pyrenees Mountains), all our examples hereafter are extracted from this repository. In a previous work [LES07] we have explored phrasal contexts where geographical point of view is predominant.

We propose, on the one hand, an automatic data processing sequence marking the contextual geographic focusing and the key term (syntactically represented by a nominal group) candidate to be a sub-type. This context will be fetch with a lexico-syntactic pattern. The goal is to mark, if it exists, the nominal group involved in such a context as a real sub-type or a "good" candidate to be a real sub-type. The choice depends on the existence of such key terms in external heterogeneous resources.

For limiting such a heavy task, our method proposes to checks if there is a sufficiently strong relationship between nominal group's participation in a particular linguistic relation and its capacity to evoke a geographical sub-type. In a travel story genre the more interesting geographic phrasal contexts are represented by descriptions of persons' movements or landscape perceptions. We propose a full-implemented automatic process in order to localize this potential geographic phrasal context. This context will be fetch with finite-state transducers embedded in a lexico-syntactic pattern. The last step is to use various external resources to validate or reduce the ambiguity of its geographical meaning.

2 Problems and Background

The traditional named entity recognition task is a well-known problem in the natural language processing (NLP) tasks and Information Extraction and Retrieving (IE & IR). Many systems have been developed, mainly on English, to recognize and categorize the proper names appearing in the texts [DM00] but any of them are able to classify places into specific sub-types such as RIVER,

GLACIER, PEAK, MOUNTAIN... The identification of the geographic names is a well known much more complex task that simply recognizing place names or locations from others Named Entities. We are mainly interested by this category: locations and their intrinsic ambiguity as related in [VJW07], [LL07], [LEI04]. Our goal is to find an existing sub-type to reduce this intrinsic ambiguity. For example, *in Artouste lake*, or *in the peak of Artouste*, the place *Artouste* have a different semantics and different spatial representation according to the geographic type carried out by *lake* or *peak* key terms. In other words, in a task of the sub-type location geometry recovery with a known label, the called upon resources and the strategies of interrogation of these resources will be much more accurate. Hereafter in table 1 a part of a travel story extracted in our corpus from the book: "Ascension au pic de Néthou , Platon de Tchihatcheff, 80 pages, 1842".

Table 1. A paragraph from the corpus "Travel stories"

[...]Après avoir contemplé, avec une admiration mêlée d'effroi, ***la charpente altière des Monts-Maudits****, nous songeâmes bientôt à descendre* ***sur le territoire aride de l'Aragon****. Le temps était menaçant : de légers brouillards parcouraient les hauteurs, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant* ***de l'ouest des Pyrénées****, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabineros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement,* ***vers le pied de la Maladetta****, laissant à notre* ***droite les roches calcaires de la Pèna-Blanca****. Arrivés au fond de* ***la vallée du Plan-des-Etangs****, qui est plus élevée que sa voisine,* ***la vallée latérale de l'hospice de Bagnères****, de 446 mètres, nous laissâmes derrière nous une cabane habitée pendant l'été par des bergers espagnols, pour remonter,* ***par un plan rocailleux, jusqu'au gouffre de Tourmon****, qui absorbe les eaux d'un torrent rapide, descendant* ***de la partie orientale du glacier de la Maladetta[...]****.*

As we can see in the example and in agreement with ([JON94], [FM03] and [LL07]) taken on its own place name can already be of different category such as: simple pure place names (*Bagnères, Maladetta, Tourmon...*) composed of only one lexeme; complex pure place names (*Monts-Maudits, Pèna-Blanca,...*) composed of several lexemes; slightly mixed place names containing link-words (*Bagnères de Bigorre,...*). In textual document all of these categories of place name can be combined in a syntactic relation with nominal groups being able to add a specific geographical meaning and finally build a complex and heterogeneous Geographic Named Entity (GNE) where explicit sub-types play an important role (*sur le territoire aride de l'Aragon, la partie orientale du glacier de la Maladetta,...*[1]).

The GNE could be potentially ambiguous in different part of its linguistic structure. Our global goal is to propose a whole method for reduce ambiguity.

[1] On arid territory (zone) of Aragon, the eastern part of Glacier Maladetta

The method combines for each different ambiguities specific treatments. This paper focuses on the problem of sub-type ambiguity where we have defined three cases:

Proposition 1.
(1) multi-referent when two different key terms existing in an ontology of reference are associated to the same Place Name.
(2) neo-referent when a the key term associated to at least one Place Name is not directly present in ontology of reference.
(3) lacked-referent when a the key term associated to at least one Place Name is not present in ontology of reference but exists in a non-specific thesaurus.

For this last category of ambiguity it has been necessary to find a linguistic method to filter out local phrasal context suggesting a potential geographic meaning: a spatial pattern. A spatial pattern is an aspect of the identifiable speech by particular spatial characteristics. Related works have shown that these characteristics result in linguistic aspects. For instance, in [TT97] authors define, study and categorise three patterns for description, the description of way, by course of the glance, and description in over flight. In agreement with [LES07], we have retained four spatial patterns: itinerary, local description, points of, places comparison. For each of these spatial patterns the main bootstrapping linguistic mark is the use of a specific verb category.

Thus, the study of our corpus relating to travel stories showed that whenever a place name is used it could be associated with a nominal group evoking a kind of more or less sharped spatial focusing called therefore indirection. More over if the place name associated or not with an indirection is evoked in a sentence containing a verb of movement or a verb of perception, then the nominal group between the verb and the place name frequently has a geographical connotation (in our corpus approximately 50% of the terms result from an ontology of topographic domain). Thus, we put forth the assumption that such event structures [RBH10] are interesting discriminating indicators for making a first selection of nominal groups used for their geographical mining.

In this paper, due to our corpus, we have particularly explored the two first patterns:

Proposition 2.
(a) The pattern itinerary corresponds to a description of a set of the author's movements from place to place, related to a journey.
(b) The pattern local description corresponds to a description of a restricted place, the speaker being in this place. This pattern is appropriate for a description made without movement on the part of the author.

Both patterns enable us to apply filters bootstrapped by verbs. To put it in a nutshell, the pattern itinerary is characterized by verbs of displacement as defined in [M.08] and the pattern local description is characterized by verbs of perception and verbs of state. A description calls upon the five senses: sight, hearing, touch, taste and sense of smell. To evoke a feeling, the authors use above all the verbs of perception such as to see, to hear, to touch, to taste and

to feel. Due to the specificity of our corpus of travel stories we have restricted to two categories of verbs: verbs of displacement and verbs of perception and we are looking to a particular semantic relation involving a verb of displacement or a verb of perception with or without a preposition in relation with or without one or more nominal groups. These observations also collaborate research done in displacements reference in language such as that of

We essentially uphold the criteria of a verb's aspectual polarity that was introduced by Boons [BOO87] and further developed in [LAU91]. Thus, we model the particular semantic relation in the text by taking account of displacement verbs that are necessarily associated with Place Names, and that are optionally associated with spatial clauses.

In the model we propose here, the triplet is discriminating to bring out the spatial meaning of certain polysemic verbs (*to leave someone* is of little interest to us, whereas *to leave Pau* attracts our attention). The same is true for *to get out of a tough spot vs to get out of Pau*. This means that the pattern to extract the particular semantic relation defers if the verbs involved is initial (*to leave*), median (*to cross*) or final (*to arrive*). These notions meet the LRV[2] thesis of Sablayrolles [AS95] who also studied the motion verbs.

3 Method and implementation

In order to reduce different levels of ambiguity carried in different parts of GEN we use a methodology combining various lexico-syntactic patterns [HEA92], [MZB04], [MFP09], in a process taking into account phrasal context.

The core of our method is based on a lexico-syntactic pattern called from hereafter *VPT*. It's a triplet which is composed of the following elements : the first one being the verb of displacement, or verb of perception, the second one being preposition, and last one toponym. In some cases, a preposition can lack in the triplet VPT. In each triplet, a `toponym` is represented by the following non-ordered items list, `[sub-type candidate*, indirection?, place name]` where:

- `?` for specifying that there must not be more than one occurrence —the item is optional— ;
- `*` for specifying that any number (zero or more) of occurrences is allowed —the content of each occurrence may be different and are generally disconnected and the item is optional—;

Some of VPT' different instances are summarised in Table 2 and the fully implemented chain including the VPT' principles is illustrated in the figure 1. In this figure (a) represents major steps of our processing sequence; (b) explains the output of each step with the input sentence: *Nous songeâmes bientôt à descendre sur le territoire aride de l'Aragon*, which comes from the paragraph in table 1. Our chain is designed for processing a corpus in french, but it can be tuned for

[2] Lieu de Référence Verbal (Verbal Reference Location)

texts in other language. This adaptation will be discussed later. In the figure 1, to make paper more understandable to non-french speakers the input and outputs in each step have been translated in English.

Table 2. Some examples of the triplet VPT

Verb	Preposition	Toponym		
		Sub-type candidate	Indirection	Place name
arriver (arrive)	à (to)	ville (city)	au sud de (in the south of)	Pau
aller (go)	dans (into)	vallée (valley)	au nord de (in the north of)	Paris
venir (come)	de (from)	rivière (river)	au centre de (in the centre of)	Aragon
voir (see)	vers (toward)	village (village)	à coté de (near from)	Azun

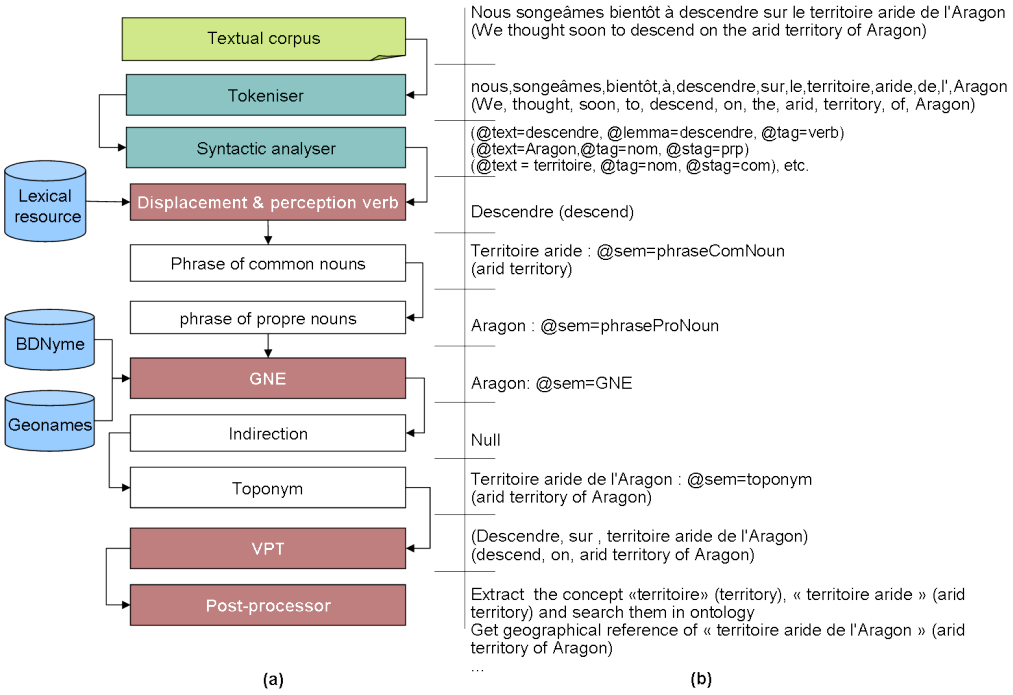


Fig. 1. Our processing main steps

3.1 Marking verbs of movement|perception

Firstly, the text is tokenized before being processed by a syntactic analyser (i.e TreeTagger[3]) which associates each token to a grammatical category (i.e. verb, noun, preposition, etc.). Then, thanks to our lexical resource, verbs of movement and verbs of perception are marked. In accordance with the retained concept of aspectual polarity, the verbs of movement are also marker as: "initial verbs" or of initial polarity, for verbs like *quitter, partir, sortir, s'échapper, s'éloigner, etc*[4]. "Final verbs" of final polarity, for verbs like *arriver à, atteindre, entrer dans, regagner*[5]. "Median verbs" or of median polarity, for verbs like *traverser, descendre, franchir, parcourir, passer par, se déplacer dans, etc* [6].

An analysis of verbs of perception enabled us to conclude that all these verbs are transitive verbs, and thus they are never followed by a preposition i.e. all the verbs of perception has a sub-behavior of median verbs of movement (median polarity). Then, we have added to our resource a lexicon of about fifty verbs of perception.

3.2 Marking toponyms

Basing on the output of the syntactic analyser, words or group of words are marked as *common nouns*, or as *proper nouns*. A single common noun could be, *vallée, village, territoire, etc*) and a complex one could be, *territoire aride, marché d'intérêt régional, etc*. Recursively, we separate the adjective(s) from the noun. This is done with rules expressed in a DCG (Definite Clause Grammar) formalism. This formalism allows to implement context-free grammar. In our case it consists of a set of rules to replace a sequence of speech (noun, adjective, verb, etc.) by a new unique identifier (noun phrase, verb phrase, etc.). Our rules marking the words or group of words as common nouns, presented in table 3, are expressed in Prolog[7]. In this table, line 3 shows how if a sequence of tokens contains an adjective which is located before a common noun (or a group of common noun, recursively), all the sequence will be represented as a common nouns. This kind of marked sequence will be retained to be candidate for sub-typing the place name.

Similarly the phrase of proper nouns can be a single proper noun (i.e Aragon, Pau, etc) or a group of proper nouns connecting by some others words (*I.e. Mont-de-Marsant, Saint-Jean-Pied de Port, etc*). The DCG rules marking theses phrases are presented in table 4.

Gazetteers (BNNyme, Geonames, etc.) are used to validate phrases of proper nouns as existing Place Names, after each validation the phrase of proper noun

[3] TreeTagger is a language independent part-of-speech tagger. It was developed by Helmut Schmid in the TC project http://www.ims.uni-stuttgart.de/projekte/tc/) (at the Institute for Computational Linguistics of the University of Stuttgart.

[4] to quit, to leave, to go out, to escape, to get away, etc.

[5] to arrive, to reach, to get in, to go back, etc.

[6] to cross, to go down, to overcome, to cover, to go by, to move in, etc.

[7] Prolog is a general purpose logic programming language

Table 3. DCG marking the phrase of common noun

```
root(commonNoun:X) --> group(X).
%case 1 : ex : belle ville
group(adjectif:A..nom:N) --> adjectif(A), group(N).
%case 2 :ex : territoire aride
group(nom:N..adjectif:A) --> commonNoun(N), adjectif(A).
%case 3 :hotel de ville (recursively)
group(nom1:N1..nom2:N2) --> commonNoun(N1), %hotel
        (ls_token('de');ls_token('d\'');ls_token('des')), %de
        group(N2). %ville
%cas4 : territoire
group(X) --> commonNoun(X).
commonNoun(adjectif:' '..nom:lemma:X) --> ls_token(_,
lemma:X..stag:com, token).
adjectif(A) --> A@tag:adj.
```

Table 4. 3 among 14 DCG rules for marking the phrase of proper noun

```
root(lemma:X) --> group(X).
%ex1 : Mont de Marsant (recursively)
group(X) --> N1@stag:pro, ls_token('de'), group(N2),
{string_concat(N1, ' de ', S)},
{string_concat(S, N2, X)}.
%ex2: Saint Jean
group(X) --> N1@stag:pro, group(N2),
     {string_concat(N1, ' ', S)},
     {string_concat(S, N2, X)}.
%ex3 : Aragon
group(X) --> X@stag:pro.
...
```

is marked as a GNE. In next step, the indirection (i.e. au sud de, au centre de, etc) will be marked thanks to a specific lexical resource. Finally the toponym is defined as a composition of the elements marked in previous steps : the phrase of common nouns, the indirection, the GNE. This is realized thanks to a set of DCG rules (table 5). Note that we categorize toponyms into two main classes : *absolute and relative*. An absolute one is illustrated by an example in lines 1 to 8, and a relative one in lines 10 to 19.

Next task marks the linguistic structure VPT: Verb of movement (Vmov) or Verb of perception (Vperc), preposition and toponym.

3.3 Extracting the structure VPT (Vmov|Vperc, preposition, toponym)

In agreement with [Lou08], this is done by the transducers. This section presents the analysis obtained thanks to these transducers. Two sentences, each one been a membership of a spatial pattern as previously presented in proposition 2.

Table 5. DCG for marking toponyms

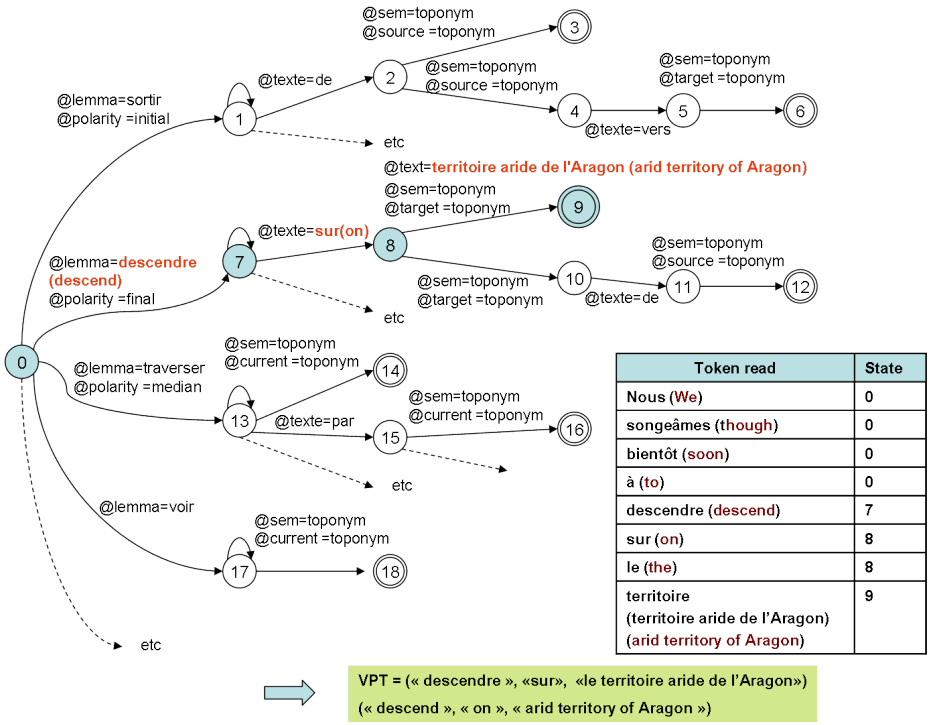
```
%Absolute toponym : territoire aride de l'Aragon (arid territory of
    Aragon)
toponym(esa:X..type:a) --> esa1(X).
%Define absolute toponym
esa1(subType:X..placeName:Y) --> subType(X), %territoire aride
   de,     %de
   placeName(Y). %Aragon
subType(X) --> ls_token(_, X, commonNoun). %territoire aride
placeName(X) --> ls_token(_, X, placeName). %Aragon

%Relative toponym : territoire aride au sud de la ville de Pau (arid
    territory in the south of Pau city)
toponym(esr:X..type:r) --> esr1(X);
%Define the relative toponym
esr1(subType:X..indirection:Y..esa:Z) -->
   subType(X), %territoire aride
   indirection(Y), %au sud de
   article, %la
   esa(Z). %ville de Pau
esa(Z) --> esa1(Z); esa2(Z).
indirection(X) --> ls_token(_, lemma:X, indirection).
...
```

Spatial pattern: itinerary Verb of movement is the core of this pattern.Transducers are illustrated in figure 2. The processing of the pattern *itinerary* is explained hrough the sentence already given above. Tokens and transitions are given in the table as well as different state transitions of the transducer.

The principle is quite simple, the transducer starts in state 0, and the text is processed token by token. Depending on the semantic of each token, the transducer passes into a new state or not. For example, there is no state changes when tokens "nous", "songeâmes", "bientôt", "à" are read. But when the token "descendre" (go down to) is read its semantic mark *verb of movement* allows the transducer to change its state from 0 to 7. So the first element of the triplet VPT (i.e. the verb "descendre") is marked. When the token containing the preposition "sur" (on) is reached the transducer passes into state 8. Finally, the toponym is marked when the token containing "territoire" (territory) is reached. In fact, this token belongs to a group of words (the toponym "territoire aride of Aragon") whose semantic mark GNE allows the transducer to pass into the state 9. So the toponym "territoire aride of Aragon" is marked. The transducer pass in its final state and returns the structure VPT ("descendre","sur", "territoire aride of Aragon") figure 3.

Spatial pattern: local description This pattern may be characterised by a verb of perception. Consider an example sentence "J'ai vu la vallée au sud du village d'Azun" (I saw the valley in the south of Azun village). The figure 4 presents the state transition of our transducer for marking the triplet VPT. When the transducer passes in a final state: the structure VPT ("voir"," ", "vallée au sud du village d'Azun") is obtained.



Token read	State
Nous (We)	0
songeâmes (though)	0
bientôt (soon)	0
à (to)	0
descendre (descend)	7
sur (on)	8
le (the)	8
territoire (territoire aride de l'Aragon) (arid territory of Aragon)	9

Fig. 2. Examples of transducers involving verbs of movement like: *sortir, arriver, descendre* and by extension a verb of perception like *voir*

The figure 5 represents the output exported by our processing chain for input as the paragraph in the table 1.

Thus, thanks to the transducers, we are able to analyze both types of verbs: verbs of movement and verbs of perception. In next step, the structure VPT will be useful for helping to reduce ambiguity in the GNE.

3.4 How the structure extracted is useful in the three cases of ambiguities?

Our previous example, "*nous songeâmes bientôt à descendre sur le* ***territoire aride de l'Aragon***" illustrates a lacked-referent case and how the method could not fully disambiguate the sub-type:

"le territoire aride de l'Aragon" has been marked as a toponym, but, neither key term territoire aride nor key term territoire are present in the domain-specific ontology of reference [8]. But this toponym is involved in the structure VPT, so the nominal group "territoire aride" (arid territory) could be considered as a

[8] In this work the domain-specific ontology has been established in collaboration with the COGIT a research group of IGN ([AM10]

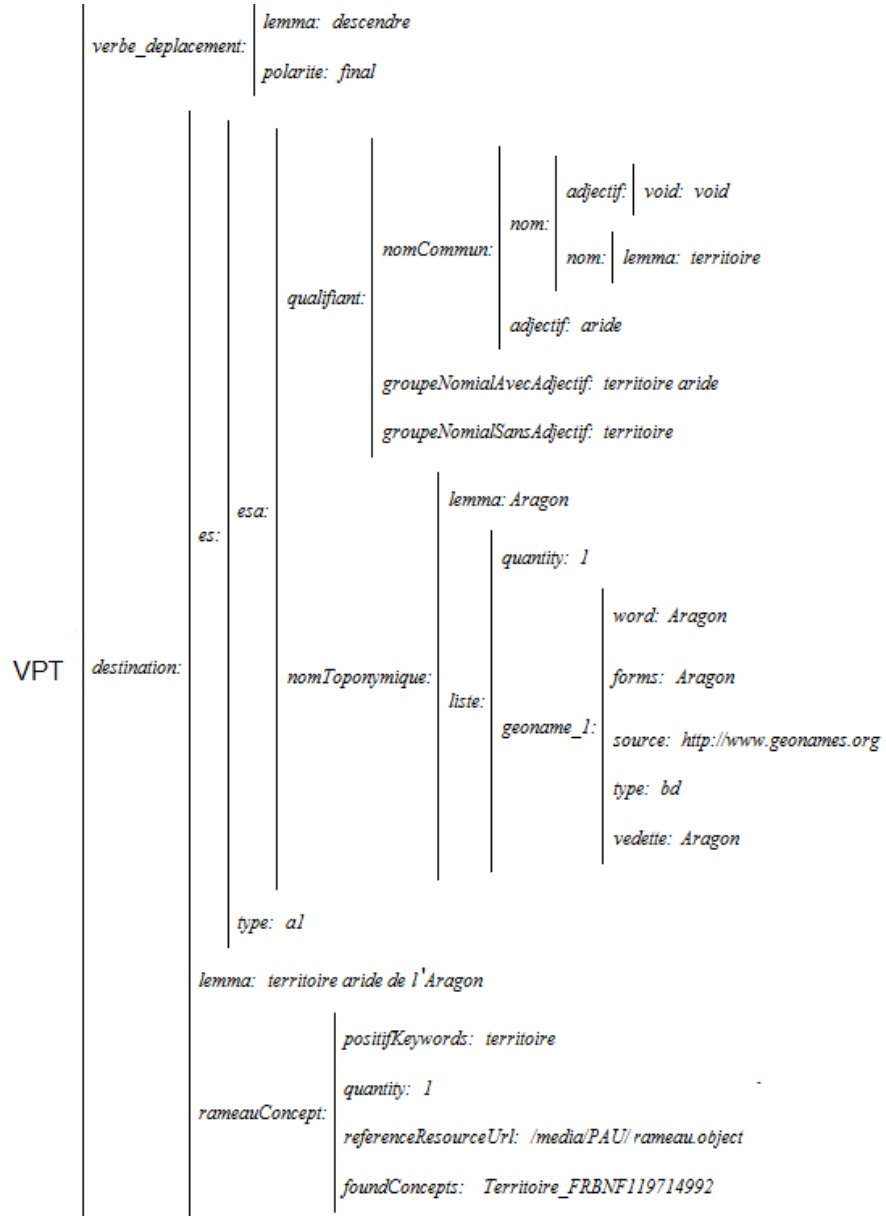


Fig. 3. Visualization of the triplet VPT marked

"good" candidate to be geographic sub-type. So, the second external resource (the generic thesaurus RAMEAU) is queried and this time the concept *territoire* is found. Unfortunately its relations with others concepts do not permit to fully validated the GNE as a toponym.

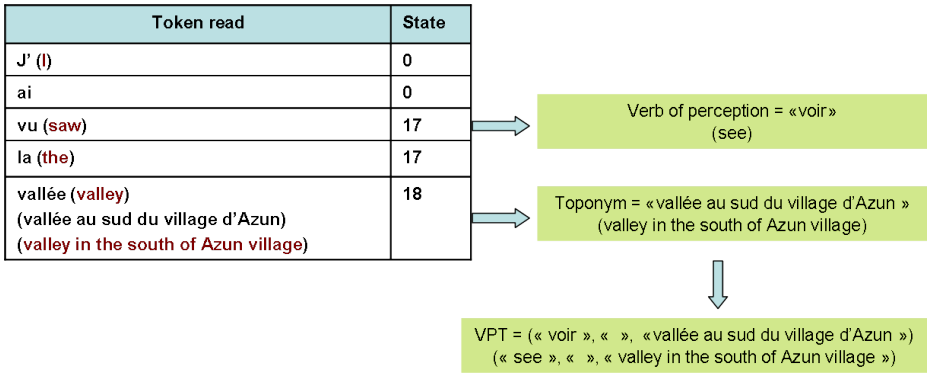


Fig. 4. The state transition of transducer for the second example sentence

Après avoir contemplé, avec une admiration mêlée d'effroi, la charpente altière de la partie centrale des Monts-Maudits, nous songeâmes bientôt à descendre sur le territoire aride de l'Aragon. Le temps était menaçant : de légers brouillards parcouraient les hauteurs, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant de l'ouest des Pyrénées, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabineros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement, vers le pied de la Maladetta, laissant à notre droite les roches calcaires de la Pèna-Blanca. Arrivés au fond de la vallée du Plan-des-Etangs, qui est plus élevée que sa voisine, la vallée latérale de l'hospice de Bagnères, de 446 mètres, nous laissâmes derrière nous une cabane habitée pendant l'été par des bergers espagnols, pour remonter, par un plan rocailleux, jusqu'au gouffre de Tourmon, qui absorbe les eaux d'un torrent rapide, descendant de la partie orientale du glacier de l'Aragon.

Legende

- Verbe of displacement or verbe of perception
- Preposition
- Phrase of the proper noun
- Phrase of common noun
- Indirection

Fig. 5. Automatic output colored text result of various lexico-syntactic processing

Table 6 presents examples of key terms extracted in our corpus of travel stories not present in the domain-specific ontology but found in the generic thesaurus RAMEAU. The table also illustrate how conceptual relations can add or reduce ambiguities.

Table 6. Some example with the term having instances in Rameau

Text	**Term**	**Instance in Rameau**	**Fathers**
Ayant atteint **le col ou le port** d'Albe, nous aperçûmes au-dessous de nous un petit lac de forme ovaloïde	port	Ports, Villes portuaires, Ports maritimes, Installations portuaires, Génie portuaire, Équipements portuaires...	Transports_maritimes, Terminaux_(transport, Ouvrages_hydrauliques, Canaux_(génie_hydraulique)
Nous prîmes le chemin du port de la Picade, en passant devant le **trou du Toro**	trou	Cavités, Orifices, Ouvertures (trous), Perforations	Surfaces_(mathématiques)
Nous ne regagnâmes nos **logements respectifs à Bagnères-de-Luchon** qu'après avoir été trempés jusqu'aux os	logement	Logements, Logement en milieu urbain, Hébergement, Habitat humain, Habitat (logement), Conditions d'habitation, etc.	Urbanisme
Nous songeâmes bientôt à descendre sur le **territoire aride de l'Aragon**	territoire	Territoires, Acquisition de territoire	Territoire_national

3.5 Can our method be reused to process other languages ?

The processing schema figure 1(a) is designed to be natural language independent in the present work, due to the corpus, the process as been fully tested for French. For it use with another language, a tuned phase is necessary for taking into account the specificity of this new language. For example, for English text, after the syntactic analysis also realized with TreeTagger. Concerning the marking process of verbs of movement and of perception, it lies on a specific lexical resource. And this lexical base is very specific for each natural language. According ([Tal00]), for the Romance languages like French or Spanish, there are a large number of verb that indicates the direction of movement (eg "entrer", "sortir", "monter", etc..). In Contrary, for the Germanic languages like English or German, the direction is indicated by a particle (expressed by preposition) associated with the verb (e.g "go in ", "go out" "go up ", "go down"). This characteristic plays an important role to determine the construction of the transducer marking the triplet VPT. Finally the DCG rules to mark common noun, proper noun and toponym, must be rewrote. For example the case 3 of the phrase of common noun(from line 6 to line 9 in the table 3) can be rewrote for English text as in the table 7

Table 7. Example marking the phrase of common noun for english text

```
%case 3 :(recursively)
group(nom1:N1..nom2:N2) --> commonNoun(N1), %hotel
        ls_token('of'), %of
        group(N2). %ville
commonNoun(adjectif:' '..nom:lemma:X) --> ls_token(_,
   lemma:X..stag:com, token).
...
```

4 Some experimentations

Some global statistics We tried out our data processing sequence on a corpus of 14 books, in a nutshell we have:

- for 10555 occurrences of motion verbs found 1390 are involved in a VPT pattern.
- 560 VPT patterns containing candidates for sub-typing Place Name.
  - 44 of them already exist in the domain-specific ontology,
  - and 49 of them have matched with a key-concept in the RAMEAU thesaurus.

Verbs of perception effects Thanks to the verbs of perception, we collect sentences such as those given in the example in table 8. It reveals new geographical information, which we could not take into account with the verbs of movement: *lac de Fachon, tour carrée de Vidalos, lac de Suyen.* We also have false positive response as with expressions like, *voir* ***la duchesse d'Albe***.

Table 8. A paragraph from the corpus "Travel stories" illustrating the use of verbs of perception

Par 2.300 mètres, Wallon appuie à droite pour **admirer l'encadrement étrange et chaotique du petit lac de Fachon**. Nous sommes sur le meilleur observatoire pour **contempler l'énorme architecture du cirque de Troumouse**. **Contemplez la merveilleuse transparence des eaux du petit lac de Suyen ! s'écrie Russell**. Je suis allé **voir la duchesse d'Albe**. Le donjon de Lourdes **voyait les trois tourelles du château de Pau** qui **apercevait la tour carrée de Vidalos**.

We have counted 62 different occurrences of verbs. Among these verbs, the verb *voir* (to see) is the most used in our corpus. We created 7 classes of occurrences of the verbs (Figure 6) in relation to substantives. We notice that 7 verbs are in the frequencies of relations section **f7** *higher than 50* and that on the other hand, 23 verbs are in the frequencies of relations section **f1** *lower than 5 times*.

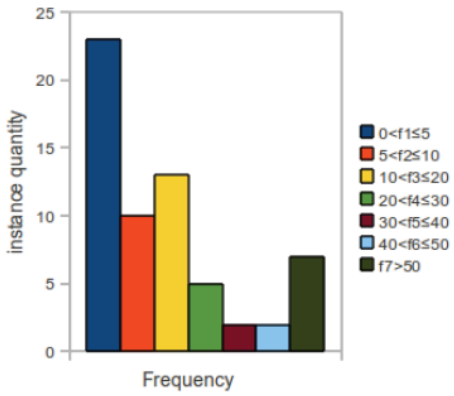


Fig. 6. Frequency of verbs operating in the linguistic structure: Verb of movement (Vmov) or Verb of perception (Vperc), preposition and toponym.

Table 9. Distribution of verbs in our corpus.

f1	f2	f3	f4	f5	f6	f7
abandonner	admirer	contourner	apercevoir	partir	atteindre	aller
approcher	contempler	diriger	entrer	passer	traverser	arriver
appuyer	dépasser	engager	revenir			conduire
border	entendre	franchir	venir			descendre
charmer	fixer	gagner	visiter			monter
dévorer	parvenir	observer				suivre
écouter	pénétrer	parcourir				voir
éloigner	redescendre	quiter				
envahir	regarder	rejoindre				
examiner	rentrer	remonter				
goûter		rendre				
grimper		retrouver				
longer		sortir				
marcher						
précipiter						
promener						
regagner						
réjouir						
repasser						
retourner						
rôder						
sentir						
toucher						

The occurrences of verbs are distributed according to Table 9. Among these verbs, 16 are verbs of perception.

We finally get 214 distinct terms that are connected to verbs of movement, and 68 connected to verbs of perception. On these collections, 30% of terms appear only with verbs of perception, thus enabling us to widen the list of the potential candidates to the enrichment of ontology.

5 Conclusion

We have presented a global method for reducing ambiguity in complex geographic named entities. This method improves the task of geographic named entity annotation, both on identification and on sub-categorization. Thanks to a particular linguistic relationship, our main objective is to reduce the different opportunities that we can handle in the task of querying in huge external resources like generic thesaurus.

The assumption that we presumed on the presence of verbs of movement | verbs of perception as indicating a geographical connotation of the nominal group candidates linked to the place names is checked. The methodology suggested enables us to extract from our corpus of travel stories a lexicon of semantic labels.

The literature is quite poor in methods which focus on a deep determination of the sub-typing of place name. In [MTV07] when the named entities are classified and disambiguated, place name is assigned to the type "Location" or "Organisation". These place names have no details of their nature. In [RBH10], an ontology is used to reduced the ambiguity. However, This core ontology only defines a simple tree structure with four levels: a root (i.e., Earth), countries, states, and localities. Moreover, the pattern used to identify the sub-type candidate is very simple : for the cities in U.S, the pattern `[city-name, state-name]` is used; for all others `[name, country-name]` is used.

One of the advantages of our method is to use the resources with a large number of hierarchical concepts to reduce the ambiguity of sub-type for place name : the domain-specific ontology consists of more than 700 topographic concepts; the generic thesaurus RAMEAU is composed of more than 170 000 concepts in various domain. This allow to reduce the ambiguity of place name at various semantic level. Moreover, we use a generic pattern to identify the sub-type candidate of the place name : it's the pattern toponym (`[sub-type-candidate, indirection, place-name]`). Recursively, this pattern allows not only to extract the sub-type associated directly to the place-name (i.e., city in *Pau city*), but also determined the sub-type associated indirectly to it at different levels (i.e., hill in *argillaceous hill in the south of Pau city*).

The first objective of our work is to exploit key terms in several options:

- Either a term specifying the type during the geometry of the location recovery (for instance in resources like geographical databases or gazetteers) because a correspondence was found via geographical ontology.
- Or a term constitutes a proposal to the domain-specific ontology enrichment if it can't be found.

– Further option will be to use the pattern VPT to make validation *a posteriori*. Indeed, in a toponym we can have the place name not validated and whereas the sub-type is known: this can lead to a strong presumption that the place name is a real geographical named entity.

Acknowledgements

This work is realized in the framework of GeOnto project "Constitution, alignement et exploitation d'ontologies géographiques" (http://geonto.lri.fr/), partly funded by the French Research Agency (ANR-O7-MDCO-005).

References

[AM10] N. ABADIE and S. MUSTIERE. Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*, 20(2):145–174, juin 2010.

[AS95] N. ASHER and P. SABLAYROLLES. A typology and discourse semantics for motion verbs and spatial pps in french. *Journal of Semantics*, 2(12):163–209, 1995.

[BLPD10] Nieves R. BRISABOA, Miguel R. LUACES, Angeles S. PLACES, and SECO Diego. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based inde. *Special Issue: Semantic and Conceptual Issues in Geographic Information Systems, GeoInformatica*, 14(3):307–331, 2010.

[BOO87] J.-P. BOONS. La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *LANGUE FRANÇAISE*, pages 5–40, 1987.

[DM00] B. DAILLE and E. MORIN. Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. *Traitement automatique des langues*, 41(3):601–621, 2000.

[FM03] Nordine FOUROUR and Emmanuel MORIN. Apport du web dans la reconnaissance des entités nommées. *Revue québécoise de linguistique*, 32(1):41–60, 2003.

[HEA92] M. HEARST. Automatic acquisition of hyponyms from large text corpora. In *the Fourteenth International Conference on Computational Linguistics*, Nantes, France, 1992.

[JON94] K. JONASSON. Le nom propre, constructions et interprétations, 1994. Duculot, Champs linguistiques.

[LAU91] D. LAUR. *Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple.* PhD thesis, Université de Toulouse II, 1991.

[LEI04] Jochen L. LEIDNER. Toponym resolution in text: "which sheffield is it?". In *the 27th, Annual International ACM SIGIR Conference (SIGIR 2004)*, pages 602–606, Sheffield, UK, 2004. ACM Press.

[LES07] J. LESBEGUERIE. *Plate-forme pour l'indexation spatial multi-niveaux d'un corpus territorialisé.* PhD thesis, Université de Pau et des Pays de l'Adour, 2007.

[LL07] Seungwoo LEE and Gary Geunbae LEE. Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation. *Information systems*, 32(4):575–592, 2007. ISSN 0306-4379.

[Lou08] Pierre Loustau. *Interprétation automatique d'itinéraires dans des recits de voyage*. type, Université de Pau et des Pays de l'Adour, address, month 2008. note.

[M.08] AURNAGUE M. Qu'est-ce qu'un verbe de déplacement ? : Critères spatiaux pour une classification des verbes de déplacement intransitifs du français. In *CONGRES MONDIAL DE LINGUISTIQUE FRANÇAISE*, PARIS, FRANCE, 2008. DOI: 10.1051/CMLF08041.

[MFP09] Diana MAYNARD, Adam FUNK, and Wim PETERS. Using lexico-syntactic ontology design patterns for ontology creation and population. In *WOP2009 collocated with ISWC2009*, 2009.

[MTV07] Claude MARTINEAU, Elsa TOLONE, and Stavroula VOYATZI. Les entités nommées : usage et degrée de precision et de désambiguïsation. In *the 26th conference on Lexis and Grammar*, France, 2007.

[MZB04] Véronique MALAISE, Pierre ZWEIGENBAUM, and Bruno BACHIMONT. Detecting semantic relations between terms in definitions. In *Ananadiou and Zweigenbaum*, pages 55–62, 2004.

[Nan98] CHINCHOR Nancy. Named entity task definition (version 3.5). In *the 7th Message Understanding Conference (MUC-7)*, Fairfax, VA, 1998.

[RBH10] Kirk ROBERTS, Adrian BEJAN, Cosmin, and Sanda HARABAGIU. Toponym disambiguation using events. In *the 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS'10), Applied Natural Language Processing track*, Daytona Beach, FL, USA, 2010.

[Tal00] L. Talmy. *Toward a Cognitive Semantics*, chapter How language structures space. The MIT Press, 2000.

[TT97] B TVERSKY and H.A TAYLOR. *Langage et perspective spatiale*, chapter chapter 2. Sciences cognitives. Masson, 1997.

[VJW07] R VOLZ, KLEB J., and MUELLER W. Towards ontology-based disambiguation of geographical identifiers. In *WWW2007 Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web*, pages 1–7, Banff, Canada, May 8–12 2007. WWW2007.