Topographic Subtyping of Place Named Entities: a linguistic approach

Van Tien Nguyen
LIUPPA
Avenue de l'Université
Pau, France
vantien.nguyen@univ-pau.fr

Mauro Gaio
LIUPPA
Avenue de l'Université
Pau, France
mauro.gaio@univ-pau.fr

Ludovic Moncla
LIUPPA
Avenue de l'Université
Pau, France
ludovic.moncla@univ-pau.fr

Abstract

The aim of this work is to find sub-types for Place Named Entities, from the analysis of relations between Place Names and a nominal group within a specific phrasal context. The proposed method combines the use of specific intra-sentential lexico-syntactic relations and external resources like gazetteers, thesauri, or ontologies. It relies on expanded spatial named entities recognition transcribed into a symbolic representation expressed in terms of semantic features. This symbolic representation will then be associated with a geo-coded representation, depending on the available resources. Our method is completely implemented and has been tested on a corpus of travelogues.

1 Introduction

The traditional named entity recognition task is a well-known problem in natural language processing (NLP) tasks and in information extraction and retrieving (IE & IR). Many systems have been developed, mainly for the English language, to recognize and categorize the proper names appearing in textual documents. Classically, the named entities are classified into persons, organisations and places. The literature is quite poor on describing methods focusing on deep determination of sub-types for place name entities such as river, glacier, peak, mountain, etc. In [1] 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 [2], ontology is used to reduce 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.

The identification of the geographic names is a well-known much more complex task that simply recognizing place names (i.e. locations) from others Named Entities. We are mainly interested by this category: locations and their intrinsic ambiguity as related in [3], [4], [5]. Our goal is to find an existing sub-type to reduce this intrinsic ambiguity. For example in expressions like, *Artouste lake*, or *the peak of Artouste*, the place name *Artouste* have a different semantics and different spatial representation according to the geographic object carried out by *lake* or *peak* terms. In other words, in a task where you must associate the correct geo-coded representation, type will allow a more detailed search of resources.

2 Problem and background

The main problem addressed here is the construction of a topographical lexicon. This lexicon must be obtained by extracting in a given corpus, nominal groups used for their topographical denotation (eg, arid territory, south of the narrow valley, etc.). Our experimental framework consists of a textual corpus containing several hundred of travel stories in the Pyrenees. To operate automatic term extraction, our first contribution is to propose lexico-syntactic patterns to mark nominal groups having a topographical denotation in the target corpus.

The problem above may be resolved by dividing into two following sub-problems: the first one is the recognition of named entity and the second one determines the semantics of the nominal group in syntactic relation with to named entity.

For the first sub-problem, there are many tools for the automatic recognition of named entities. In fact, GATE ANNIE[1], LingPipe[2], OpenCalais[3], Stanford NER[4], OpenNLP[5] can mark several categories of named entities (person, enterprise, place, date, etc). MetaCarta[6], Yahoo! Placemarker[7] target the named entities of type place while GuTime[8], HeideTime[9] are dedicated to named entities of type dates. Regarding spatial named entities, we can mention the project CasEN [6], mainly for the French language, proposes a system based on cascades of transducers for tagging in predetermined classes place names. The approach proposed in [8] operates on document structure: for example, when working on the collaborative encyclopaedia Wikipedia, the identification of named entities is done in the title and their categorization is based on the analysis of the first sentence of the description part. The approach proposed in [7], meanwhile, aims more particularly at disambiguation of recognized place names thanks to a resource like WorldNet.

The second sub-problem may be solved thanks to the definition proposed in [18], a place is a portion of the space in which we stand, and we move. A place can be composed of

[1] http://gate.ac.uk/ie/annie.html
[2] http://alias-i.com/lingpipe
[3] http://www.opencalais.com
[4] http://nlp.stanford.edu/software/CRF-NER.shtml
[5] http://opennlp.sourceforge.net
[6] http://www.metacarta.com
[7] http://developer.yahoo.com/geo/placemaker/
[8] http://www.timeml.org/site/tarsqi/modules/gutime/index.html
[9] http://dbs.ifi.uni-heidelberg.de/index.php?id=106

several elements, where the two most important are a concrete entity, and a spatial reference. Following examples have been extracted from our corpus:

(1) Nous songeâmes bientôt à descendre sur le territoire aride de l'Aragon[10].

(2) Un torrent rapide descend de la partie orientale du glacier de la Maladetta[11].

(3) Nous arrivons au fond de la vallée d'Ossau[12].

We can then consider that in the expression *territoire aride de l'Aragon* (the arid territory of Aragon) concrete entity is denoted by the topographical term *territoire aride* (arid territory) and the spatial reference can be derived from the place name Aragon. Same goes for the expression *la partie orientale du glacier de la Maladetta (*the eastern part of the glacier of Maladetta*)*, the spatial reference can be derived in the same way as in the previous example, the place name *Maladetta* and concrete entity is represented here by the nominal group *partie orientale du glacier (*eastern part of the glacier*)*. In the last expression there is no sub-type but precision is given through a spatial relation. Then the interpretation of this relation requires a spatial reasoning. [9] proposes a classification of spatial relations into three categories: topological, projective, and metric. These classes are respectively based on the properties of space: topological, projective, and Euclidean. Topological relations were the most studied, and among the first models proposed, the RCC-8 [10] became the basis of many other proposals. In the book edited by [11] a synthesis is proposed around these models. The other two categories have been less explored. The main interest of projective relations is that they can be described by projective properties without using metric properties [12]. Projective relations attempt to formalize relations expressed in natural language by expressions such as: right of, in front of, between, along, in the suburbs of north of, etc. Although specific models have been proposed for some of these relations, for orientation relations by [13] and [14], and cardinal directions by [15]. As topological relation, projective relations can be considered as qualitative, because they don't need to rely on a Euclidean representation when involved in a reasoning process. This problem of logical representation and algorithmic processing of space is fully described in [16]. Metric relations, such as the distance between two points, are generally considered to be quantitative nature, such as in the approach proposed by [17].

Let us return now to the relation (concrete entity, spatial reference) mentioned above. Despite its obvious interest its expressiveness is too small to be able to make a difference between a geographical expression and an expression with a different meaning, as revealed by the following example:

(4) Je parlai de mes intentions à plusieurs guides de Luchon.
(I spoke about my intentions to several guides of Luchon.)

Guide is a concrete entity but not topographical. It means that *guide* cannot be used for topographically sub-typing *Luchon*. The question that arises here is how to filter such terms. The study of our corpus shows that the pair relation (concrete entity, spatial reference) is frequently in relation, in the same sentence with verb expressing movement (*to come down on the arid area of Aragon, arrive at the bottom of Ossau valley*). According to [18], in Romance languages at least, movement is characterized by the verb. In travelogues corpus, according to some experimental studies, the expression of movement is essential. Furthermore many authors, such as [19], [20] and [21], have show the role of motion verbs in natural language. These authors have proposed a categorization of verbs of movement by their polarity. Therefore in our approach we have adopted the polarity of a verb can be initial (eg *leave*), medial (eg *cross*) or final (eg *arrive*). On the other hand, according to our own observations on our travelogues corpus, the verbs of perception (eg *see*) are also important in some context of evocation, especially when the narrator during his trip wants to report certain situations or feelings. Now, the problem is: how mark and formalise the relation between place name, topographical term, spatial relation, and the verb of movement or of perception?

As a solution we propose a process named *VT* described in the following section.

3 Method and implementation

In order to reduce different levels of ambiguity carried in different parts of GEN we use a methodology combining linguistic patterns [22], [23], [24], in a process taking into account phrasal context. The core of our method is based on a cascade of lexico-syntactic patterns called from hereafter *VT*.

Formally, let *V, I, T, G* respectively a set of verbs (which contains only verbs of movement and verbs of perception), a set of *indirections* (or spatial relations), a set of topographical terms and a set of place names. We define $VT=(v, t)$ with $v \subset V$, and $t = (te, i, nt|t)$ where $te \subset T$, $I \subset I$ and $nt \subset G$ with condition that *te* and *i* can be an empty set. We can release that the *toponym* (i.e. extended place name) *t* is recursively defined as its third element (*nt*|*t*) shows.

Consider the example 1 mentioned above (*Nous songeâmes bientôt à descendre sur le territoire aride de l'Aragon*). We have VT=(v, ,t) with v=*[descendre*[13]*]*, t=*[sur le territoire aride de l'Aragon*[14]*]*. In this case, $t = (i, t')$ where *i = [sur*[15]*], t'=(te, nt) = ([territoire aride*[16]*], [Aragon])*.

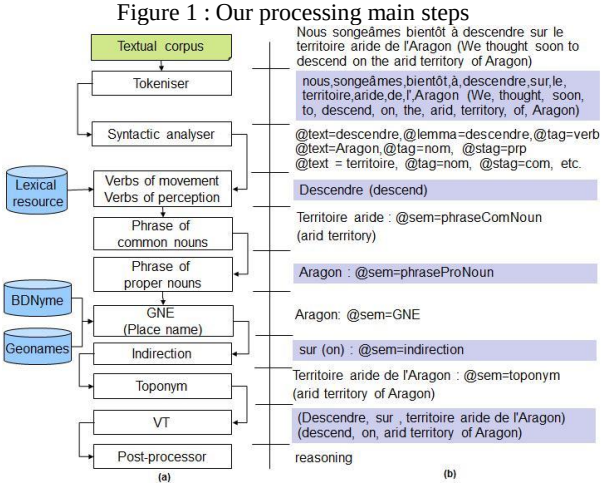
Figure 1 shows our fully implemented chain including the *VT* principles. In figure (a) represents major steps of our processing sequence; (b) explains the output of each step with above example sentence (for non-french speakers the input and outputs in each step have been translated in English but actually the process deals exclusively French language).

Firstly, the text is tokenized before being processed by a syntactic analyser (i.e TreeTagger[17]), 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" for verbs like *quitter, partir*

[10] Soon we were preparing to come down on he arid territory of Aragon.
[11] A rapid torrent descends from the eastern part of the glacier of Maladetta
[12] We arrived at the bottom of Ossau valley.
[13] to descend, to come down
[14] on the arid territory of Aragon
[15] on
[16] arid territory
[17] 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).

de, sortir, s'échapper, s'éloigner, etc., (quit, leave from, exit, escape, get away) "final verbs" for verbs like *arriver à, atteindre, entrer dans, accéder, etc.* (arrive at, reach, enter, access), "medial verbs" for verbs like *traverser, descendre, parcourir, passer par, se déplacer dans, etc*. (across, walk down, go, pass, move in). 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. Then, we have added to our resource a lexicon of about fifty verbs of perception.

Figure 1 : Our processing main steps



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*. (valley, village, territory, etc.) and a complex one could be, *territoire aride, marché d'intérêt regional, etc (arid territory, market of regional interest).* Recursively, the adjective(s) is separated from the noun. This is done thanks to rules expressed in a DCG (Definite Clause Grammar) formalism. This formalism allows the implementation of context-free grammars. In our case it consists of replacing, thanks of a set of rules, a text sequence (noun, adjective, verb, etc.) with a unique symbol. For example, the following rule allows defining a symbol named term containing an adjective and a noun.

```
term( adjectif :A . . noun:N) −−> adjectif(A) , noun(N) .
```

In this way we have defined 4 cases for phrases of common noun and 14 cases for phrases of proper nouns. In a second time gazetteers (BNNyme, Geonames, etc.) are used to validate proper nouns as existing Place Names. After each validation the proper noun is marked with a new symbol GNE. Next step marks the, so-called, indirection (i.e. *au sud de*[18]*, au centre de*[19]*, etc)* thanks to a specific lexical resource. After that the toponyms are defined as a composition of the elements marked in previous steps: the phrase of common nouns, the indirection and the GNE. This is done thanks to a cascade of DCG rules as show table 1. As mentioned in comments of the DCG code, we distinguish two cases of toponyms : absolute and relative.

Table 1 : DCG rules marking toponyms

```
% Case 1 - 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

%Case 2 - 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).
...
```

Finally, the *VT* structure is marked.

```
VT(verb: V..toponym:T) −−> verb(V), toponym(T)
```

After this step nominal groups contained in the structure are examined. If the nominal group, or at part least part of it, matches with a concept or a label of concept of a topographical ontology, it is marked as the subtype the place name. Consider the example above: *Nous arrivons au fond de la vallée d'Ossau*[20]. The term *vallée* (valley) will be tagged as the subtype of the place name *Ossau* because *vallée* matches with a concept of a topographical ontology[21]. In case there would have been no match, a second search is triggered on a generic thesaurus to try to deduce the meaning of the term. In this case, the term will not be directly used to sub-type the place name, but it can be used to enrich the topographical ontology, after validation by a human. For the example in , *je traversais le plus vaste territoire karstique de l'Aragon* (I've crossed the largest karstic territory of Aragon) *le territoire karsitique de l'Aragon (*karstic territory of Aragon*)* has been marked as a toponym, but, neither the term *territoire karstique (*karstic territory*)* nor the term *territoire* (territory) is present in the domain-specific ontology. However this toponym is involved in a *VT* structure, so the term *territoire karstique* could be considered as a "good" candidate to be a topographic label of a concept. So, a second external resource (a generic thesaurus for francophone libraries called RAMEAU) is queried and the term territoire *karstique* is found as a key-concept. The term is therefore proposed to an expert for possible addition to the ontology.

4 Some experimentations

[18] in the south of
[19] in the center of
[20] We arrived at the bottom of Ossau valley.
[21] In this work the domain-specific ontology has been established in collaboration with the COGIT a research group of IGN

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

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

The experiments show that the verbs of perception reveal new geographical information. We also have false positive response as expressions like, *voir la duchesse d'Albe* (see Duchess of Alba).

We finally get 214 distinct terms that are connected to verbs of movement, and 68 connected to verbs of perception. On the travelogue corpus, 30% of terms appear only with verbs of perception.

5 Conclusion

This paper present a global method for adding sub-types to place named entities, thanks to particular linguistic relations. This method can be used effectively to reduce ambiguities. It also could be used to improve the formulation of queries for searching in very large resources. The experimentation verifies the assumption consists in saying that the nominal group, which lies between a verb and a place name, has a very high probability of having a geographical sense. The methodology suggested enables us to extract from our corpus of travel stories a lexicon of topographic labels.

One of the advantages of our method is the possibility to use resources with a large number of hierarchical concepts, eg. domain-specific ontology used for experiments consists of more than 700 topographic concepts. Furthermore the generic thesaurus RAMEAU is composed of more than 170000 concepts in various domains. This possibility allows reducing the ambiguity of place name at various semantic levels. Moreover, we use local lexico-syntactic patterns allowing inexpensive adaptation to other Indo-European languages. Recursively, this method allows not only to extract the sub-type associated directly to the place-name, but also determined the sub-type associated indirectly to it at different levels (i.e., hill in argillaceous hill in the south of Pyreneans mountains).

References

[1] Martineau, C., Tolone, E., Voyatzi, S.: Les entités nommées: usage et degree de precision et de désambiguïsation. In: The 26th Conference on Lexis and Grammar, France (2007).

[2] Roberts, K., Bejan, C.A., Harabagiu, S.: Toponym disambiguation using events. In: The 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS 2010), Applied Natural Language Processing track, Daytona Beach, FL, USA (2010)

[3] Volz, R., Kleb, J., Mueller, W.: Towards ontology-based disambiguation of geographical identifiers. In: WWW 2007 Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8-12, pp. 1–7 (2007)

[4] Lee, S., Lee, G.G.: Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation. Information Systems 32(4), 306–4379 (2007) ISSN 0306-4379

[5] Leidner, J.L.: Toponym resolution in text: "which sheffield is it?". In: The 27th, Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, UK, pp. 602–606. ACM Press, New York (2004)

[6] Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D., « Cascades de transducteurs autour de la reconnaissance des entités nommées », TAL, vol. 52, n° 1, p. 69-96, 2011.

[7] Buscaldi D., Rosso P., « Using GeoWordNet for Geographical Information Retrieval », CLEF, p. 863-866, 2008.

[8] Bouamor H., «Extraction des connaissances à partir du Web pour la recherche des images géoréférencées », CORIA, p. 519-526, 2009.

[9] Clementini E., A Conceptual Framework for Modelling Spatial Relations, PhD thesis, Institut National des Sciences Appliquées de Lyon, Lyon, Juin, 2009.

[10] Randell D. A., Cui Z., Cohn A. G., « A spatial logic based on regions and connection », 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, p. 165-176, 1992.

[11] Aiello M., Pratt-Hartmann I., Van Benthem J. F., (eds), « Handbook of spatial logics », edn, Springer, 2007.

[12] Billen R., Clementini E., « Étude des caractéristiques projectives des objets spatiaux et de leurs relations », Revue Internationale de Géomatique, vol. 14, n° 2, p. 145-165, 2004.

[13] Freksa C., « Using orientation information for qualitative spatial reasoning », in A. U. Frank, I. Campari, U. Formentini (eds), Theories and methods of spatio-temporal reasoning in geographic space, vol. 639 of LNCS, Springer, Berlin, p. 162-178, 1992.

[14] Hernández D., « Maintaining Qualitative Spatial Knowledge », in A. U. Frank, I. Campari (eds), COSIT'93, vol. 761 of LNCS, Springer-Verlag, p. 33-53, 1993.

[15] Ligozat G., « Reasoning about Cardinal Directions », J. Vis. Lang. Comput., vol. 9, n° 1, p. 23-44, 1998

[16] Balbiani P., Muller P., « Le raisonnement spatial », edn, Cepadues Editions, 2000.

[17] Berretti S., Del Bimbo A., Enrico V., «Weighted walkthroughs between extended entities for retrieval by spatial arrangement », IEEE Transactions on Multimedia, vol. 5, n° 1, p. 52-70, 2003.

[18] Talmy L., « Toward a Cognitive Semantics », edn, The MIT Press, chapter How language structures space, 2000.

[19] Boons J.-P., « La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs », Langue Française, n° 76, p. 5-40, 1987.

[20] Laur D., 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.

[21] Sarda L., « L'expression du déplacement dans la construction transitive directe », Syntaxe et Sémantique, p. 121-137, 2000.
[22] Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: The Fourteenth International Conference on Computational Linguistics, Nantes, France (1992)
[23] Malaise, V., Zweigenbaum, P., Bachimont, B.: Detecting semantic relations between terms in definitions. Ananadiou and Zweigenbaum, 55–62 (2004)
[24]Maynard, D., Funk, A., Peters, W.: Using lexico-syntactic ontology design patterns for ontology creation and population. In: WOP 2009 Collocated with ISWC 2009 (2009)