Extended Named Entity Recognition Using Finite-State Transducers: An Application To Place Names

Conference Paper · March 2017

CITATIONS
4

READS
38

2 authors, including:

Ludovic Moncla
Ecole Navale de Lanvéoc-Poulmic
14 PUBLICATIONS **61** CITATIONS

SEE PROFILE

Extended Named Entity Recognition Using Finite-State Transducers: An Application To Place Names

Mauro Gaio, Ludovic Moncla

Laboratoire d'Informatique, Université de Pau et des Pays de l'Adour (LIUPPA)
64000 Pau, France
Email: mauro.gaio@univ-pau.fr
ludovic.moncla@univ-pau.fr

***Abstract*—The textual geographical information is frequently organized around spatial named entities. Such entities have intrinsic ambiguities and Named Entity Recognition and Classification methods should be improved in order to handle this problem. This article describes a knowledge-based method implementing a full process with the aim of annotating in a more precise way the spatial information in the textual documents. This gain in accuracy guarantees a better analysis of the spatial information and a better disambiguation of places. The backbone of our proposal is a construction grammar and a cascaded finite-state transducers. The evaluation shows that the introduced concept of hierarchical overlapping, is very helpful to detect a local context associated with Named Entities.**

Keywords–Geo-information processing; Geo-spatial Web Services and processing; Geo-spatial data mining.

I. INTRODUCTION

Different from other forms of geographical data, text-based spatial descriptions are subject to all sorts of ambiguities that prevent effective use [1]. 'Geocoding' textual documents refers to the function of creating unambiguous representation (i.e., footprint) of those text-based spatial descriptions. Significant efforts have been invested in geocoding, however, in order to achieve such a function it is clear that one must first correctly annotate spatial descriptions in the text. Such process is commonly known as 'geoparsing'.

The purpose of this article is to describe a method for implementing a full geoparsing process. A formal grammar describes the concept of extended spatial named entity and their relations with movement verbs. A cascaded finite-state transducers implements a parser with respect for the grammar rules. The parser annotates places and their spatial and verbal relations in order to produce an output including the more detailed description as possible.

We introduce the concept of 'Extended Named Entity' as an entity built with both categories of proper names [17] (i.e., pure and descriptive), and that can be composed of one or more other concepts. Whereas most NERC systems, such as OpenCalais [2], DBpedia Spotlight [2], OpenNER [3], CasEN [4] and Stanford NER [5], usually only consider pure proper names.

We argue that for a fine-grained task, such as marking, classifying and disambiguating spatial named entities, it is essential to consider more accurately the spatial information in relation with named entities [6].

The paper is structured as follows. In Section II, we present the theoretical background and some work useful to implement an automatic process on extracting valuable spatial information in texts. In Sections III-A and III-B, we describe the core principles for establishing our construction grammar. In Section III-C, we briefly describe the implemented solution and we assess the results of a series of evaluations, and Section IV concludes this paper.

II. BACKGROUND AND RELATED WORK

In computational linguistics, parsing is the process of analysing natural language data in accordance with the rules of a formal grammar. In order to automatically parse such data, it is initially necessary to agree on the grammar to be used. Syntactic parsing, then, is the task of recognizing a sentence and assigning a syntactic structure to it. Parsers can be viewed as searching through the space of possible parse trees to find the correct representation for a given input, using two basic search strategies: top-down search and bottom-up search. The top-down strategy tries to build the correct tree from the root node to the leaves, whereas in the bottom-up strategy the parser starts with the words of the input, and tries to build trees from the leaves to the root node, by applying, one by one the rules of the grammar.

Alongside the development of these parsers the notion of construction grammar emerged. This kind of grammar evolved out of work initiated by [7]–[9] and assigns a major role to the concept of construction as a theoretical entity. As specified by [10] the elements of the grammar are constructions: a construction is a pattern used to generate the elements of a language, or to extract these elements from an instance produced from a language. Construction grammars may specify a semantics that differs from the sum of the lexical meanings of its components. Construction grammars can reuse concepts already employed in other linguistic theoretical frameworks, such as Noun Phrase (NP) or Verb Phrase (VP), or Prepositional Phrase (PP). In this kind of construction, a feature structure is usually used to represent the elements of the language. A feature structure is a set of attribute-value pairs; the value can be atomic or another feature structure. A feature structure can be represented as a directed acyclic graph (DAG), with the nodes corresponding to the variable values and the paths to the variable names. Often however, feature structures are written as follows:

$$\begin{bmatrix} \text{role} & \textit{target} \\ \text{named entity} & \begin{bmatrix} \text{component} & \textit{noun phrase} \\ \text{category} & \textit{descriptive} \\ \text{type} & \textit{location} \end{bmatrix} \end{bmatrix}$$

Local and global ambiguities are perhaps the trickiest problem that parsers have to tackle. This problem is particularly

 ISBN: 978-1-61208-539-5

important when the parser is based on a complex grammar. In the literature, many strategies have been proposed to remove as many ambiguous cases as possible. Currently, in tasks known as Named Entity Recognition and Classification (NERC) the problem of ambiguities remains unresolved for some contexts. The notion of 'Named Entity' (NE) was formally established at the Sixth Message Understanding Conference (MUC-6, 1995). From the beginning the notion included names of persons, locations and organisations, but also numerical expressions of time, date, money, etc.

A considerable amount of work in NERC research takes the language factor as a parameter and in this body of work a significant proportion is devoted to the study of English, but French is also considered [11] [12], as well as other languages. The impact of literary genre (narrative, memoir, journalism, etc.) and domain (supply of raw materials, market or economic intelligence, politics, etc.) is a problem that has been more recently addressed in the NERC literature. Globally approaches for named entity parsing cover a huge variety of strategies, methods and representations. These approaches are generally classified in two main categories, data-driven approaches and knowledge-based approaches. One of the earliest research papers in the field of NERC was written by [13]. Her approach was based on heuristics and handcrafted rules, in other words was knowledge-based, this is also the case of our proposal. This kind of approach do not require a complete parse for all the input. A shallow parse of input sentences may be sufficient; as it is usually the case in information extraction systems that focus on the segments in a text that are likely to contain valuable information. Many different methods can be used, but, it should be mentioned that a finite-state automaton is probably the most widely used mathematical device to implement shallow parsers. Some implementations make use of cascaded finite-state transducers to produce tree-like representations. Because regular languages and relations can be encoded as finite automata they can be more easily manipulated than more complex languages; cascaded finite-state transducer principle have therefore turned out to be very useful for linguistic applications, in particular for shallow parsing. Generally, there are different finite-state transducers at different stages. Each stage bundles a set of items in a package that will be considered as a single element in the next stage.

Parsing that is solely concerned with geographical data is known as geoparsing and aims at extracting keywords and keyphrases describing geographical references from unstructured text. There are currently several types of specific ambiguity involved in geoparsing and more specifically with the problem of toponym recognition.

Toponym disambiguation is defined as a subtask of toponym resolution and is complementary to the subtask of toponym recognition. It involves associating a non-ambiguous location with a place name [14]. According to [15], the approaches for disambiguating toponyms can be classified in three categories: supervised or data-driven approaches, map-based approaches and knowledge-based approaches. Data-driven approaches are based on machine learning algorithms and exploit non-geographical content and events to build probabilistic models using spatial relationships between entities (i.e., persons, organisations) and places. As pointed out by [16], a place is more likely to be located near other places mentioned around it. Knowledge-based approaches aim at considering semantic relations between named entities, concepts or key terms such as social, associative or lexical relatedness and not only co-occurrence statistics of terms. These methods use knowledge sources (gazetteers, ontologies, etc.) to determine whether other related toponyms in the knowledge source are also referred to the document, or exploit additional information from the toponyms, such as importance, size or population counts. Finally, map-based disambiguation approaches use other unambiguous and georeferenced toponyms found on the same document as context for disambiguation.

As previously mentioned, NERC approaches are classified in two main categories, data-driven approaches and knowledge-based approaches. The main drawback of data-driven approaches is the lack of classified collections and the need for large corpora of annotated ground truths. Knowledge-based methods are more suitable for approaches based on domain-specific corpus analysis and rules are described in a readable way and are easy to modify and maintain. This is the case in our proposal, where the goal is to design and implement a parser, based on a bottom-up stategy, for recognising and classifying places in a dynamic space context mentioned in French, Spanish or Italian texts.

III. Annotating Spatial descriptions

A. Extended Named Entity (ENE) structure

According to [17] there are two categories of proper names: pure and descriptive. Pure proper names can be simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and are composed of proper names only. Descriptive proper names refer to a composition of proper names and common names (i.e., expansion). In other words, descriptive proper names overlap pure proper names. Descriptive proper names refer to a NE built with a pure proper name and a descriptive expansion. This expansion can change the implicit type (e.g., location, person, etc.) of the initial pure proper name.

We define several levels of overlapping (0, 1, 2, etc.) for the representation of ENE. Each level is encapsulated in the previous level.

Level 0: refers to pure proper names. It can be seen as the core component of an ENE. Thus, we consider NE as a special kind of ENE. Examples (1) illustrate level 0 entities:

Aragón → one entity (location)
Greenpeace → one entity (organisation) (1)
Charles de Gaulle → one entity (person)

Level 1: refers to descriptive proper names composed of a pure proper name (i.e., an entity of level 0) and a common noun (i.e., expansion). The following examples (2) show the representation of ENE. In these cases, and according to a same given ontology, descriptive expansions may not change the implicit or default nature of the object described by the proper name; they just specify the nature or the feature type.

comunidad autónoma de *Aragón*
'autonomous community of Aragón'
ufficio italianno di *Greenpeace* (2)
'Greenpeace Italian office'

However, when the associated term has not the same type of the intrinsic or default type of the pure proper name, it defines a new entity that overlaps the pure proper name one. The following examples (3) show that an entity may contain the name of another entity, and that the new entity may have a different type, examples (3b. - 3c.).

a. *Delta del* *Ebro*

'Ebro Delta'

→ two entities, ***Ebro*** (location) and ***Delta del Ebro*** (location)

b. *siège social de l'* *Unicef*

'Unicef headquarters' (3)

→ two entities, ***Unicef*** (organisation) and ***siège social de l'Unicef*** (location)

c. *sindaco di* *Genova*

'mayor of Genoa'

→ two entities, ***Genova*** (location) and ***sindaco di Genova*** (person)

Level >1 : refers to a descriptive proper name composed of another descriptive proper name. ENE of level >1 are built with ENE of level 1 and with one or more descriptive expansions, as shown on the examples (4a. - 4b.). The behavior is the same as for the previous level, i.e., the expansion can change the type of the object described by the ENE of level 1. In fact, there is not really a limit to the overlapping. However, it is quite uncommon to find an ENE of a level greater than 3. The example (4c.) show an ENE of Level 3.

a. *portavoce della* *Villa Médicis a* *Roma*

'spokesperson of the Villa Médicis in Rome'

b. *club de pelota vasca de la* *ciudad de* *San Sebastián*

'Basque pelota club of San Sebastian city' (4)

c. *dépression karstique dans le* territoire aride au sud de la *région* d'Aragon

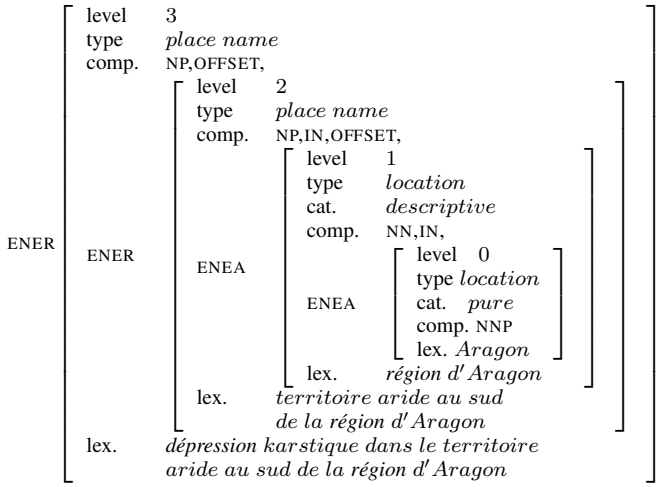
'karstic depression in the arid land south of the Aragon Region'

We have considered the annotation of ENE as a shallow parsing and the grammar to be used as a specific construction. The core of the grammar is given in 5.

With this kind of grammar and with a parser based on a bottom-up strategy each level of the ENE can be marked, from the pure proper name to the whole ENE and it can distinguish between two types of ENE, 'absolute' refering to standard spatial ENE and 'relative' refering to spatial ENE associated with spatial relations (i.e., 'offset' and 'measure').

$$\begin{array}{rcl} S & \rightarrow & ENE \\ ENE & \rightarrow & ENEA \mid (Term)\ ENER \\ ENER & \rightarrow & Offset\ ENEA \mid Offset\ ENER \\ ENEA & \rightarrow & (Term)\ ProperNoun \mid Term\ ENEA \\ Term & \rightarrow & Nominal\ Det \\ Nominal & \rightarrow & Noun \mid Nominal\ Noun \end{array} \quad (5)$$

$Offset$ can be seen as an adverbial clause. For instance, taking example (4c.), using the proposed NERC process it produces the results represented in Feature Structure form in Fig. 1.



NP=Noun phrase, NN=Noun, IN=Preposition, NNP=Proper noun, singular

Figure 1. The annotation of an ENE.

With respect to the specific problem of the NERC category of place names, this makes it possible, in particular, to move beyond reducing a place to a name and then geocoded with a single set of coordinates, a model that is still predominant in Geographic Information Science [18].

Standard NER tools consider only the entity 'Aragon Region', and therefore lead to inaccuracies in classification and/or disambiguation. Of course this must be consistent with the discourse context.

B. Movement verbs and Extended Spatial Named Entity structures

In view of our aim of automatically producing the more detailed description to achieve a better disambiguation, we propose to use additional information in relation with NE. For a better understanding of the spatial context of a NE, the linguists have highlighted the importance of the use of motion verbs, especially in Romance languages [19]. That is why we opted for taking into account movement verbs in the parsing process.

The core of the 'VT' grammar proposed hereafter can be seen both as a specialisation and as an extension of the ENE construction grammar. The symbol V represents a set of movement verbs and the symbol T a set of n-tuples, i.e., a composition of elements belonging respectively to three sets: SO a set of spatial offsets, TG a set of geographical noun phrases and E a set of ENE.

 ISBN: 978-1-61208-539-5

Consider the following sentence 6:

descendre sur le territoire aride au sud de la région d'Aragon.
'go down onto the arid land south of the Aragon region.' (6)

Example (6) has the following VT structure $= (v, t)$, with: $v =$ *descendre*, $t =$ *sur le territoire aride au sud de la région d'Aragon*. With t respectively composed of: $tg_3 = \emptyset$, $so_3 =$ *sur*, $\text{ENE}_2 =$ *territoire aride au sud de la région d'Aragon*, $tg_2 =$ *territoire aride*, $so_2 =$ *au sud de*, $\text{ENE}_1 =$ *la région d'Aragon*, $tg_1 =$ *région*, $so_1 = \emptyset$, $\text{ENE}_0 =$ *Aragon*. The set SO of spatial offsets is composed of locative phrases in which, at least in verb-framed languages such as French, the role of prepositions is central. A large number of studies have shown that prepositions are involved in the operation of spatial tracking, or location. With respect to the location concept, following the conclusions of work conducted according to Talmy's [20] and Vandeloise's [21] proposals, prepositions contribute significantly to bringing together two entities: a locator and a localised entity (i.e., a landmark and a target in Vandeloise's terms). The phrase used as locator must have spatial properties that facilitate its identification and the explanation of the spatial relationship in which it is involved. Linguistically, there are three kinds of phrases that can serve as locators: noun phrases including a name with spatial properties, noun phrases indicating distance, e.g., *le refuge se trouve à trois kilomètres ou à une heure de marche* ('the refuge is three kilometers or an hour's walk away') or orientation, e.g., *prendre la bretelle de droite* ('take the exit on the right') and noun phrases evoking an activity that may be associated with a place, e.g., *je me rendais au cours de natation* ('I was on my way to my swimming lesson').

The first category of phrases used as locator is the most common one and it can be associated to the greatest number of prepositions. The proposed VT construction grammar relies only on this category. In this category, the included name can be of two types: place names and the names of concrete objects (objects that can be located in the same place at the same time), in other words the ENE elements contained in the set E. Frequently, a particular sub-group consisting of noun phrases referring to specific parts of a locator (the peak, the bottom, the slope, the interior) is considered separately. They are unique in that they are considered suggestive of spatial properties only if they are in relation with ENE via prepositions such as *de* (from) and *à* (to, at). In the VT structure, the set TG represents this sub-group of noun phrases.

What can be retained from the literature is that the same prepositional phrase can be used to describe a variety of spatial situations, and that the discriminating factors are at the level of modalities of action. As pointed out by various studies, location is a static principle unless a dynamic component related to the verb also operates. Moreover, languages are not fully part of one category or the other [22]. For instance, in English, which is mostly a satellite-framed language, there are many verbs, such as: 'enter, exit, ascend, descend', that refer both to Motion and Path. Conversely, in verb-framed languages there are also some satellite-framed expressions such as *partir de* ('to leave'), *partir à* ('to go') where the path is encoded in the French prepositions *de* and *à*.

[23] proposed to classify motion verbs according to the 'aspectual' properties of movement called hereafter 'polarity'. The three polarities are initial (e.g., to leave), median (e.g., to cross) and final (e.g., to arrive). Without changing the intrinsic polarity of the verb, the preposition can change what could be called the focus of the displacement. More specifically, the association of a motion verb with a spatial preposition can change the focus of the displacement and take on the polarity of the preposition instead of that of the verb. Undeniably, 'leaving from Paris' and 'leaving for Paris' are two expressions with radically opposite focus of the displacement. 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. The place name *Paris* is the target, so the polarity of the whole expression may be considered as final.

The VT construction grammar aims to be a computational synthesis of the work on the means used by the language to express displacement, the work on the functioning of movement verbs in a sentence, and the work on the combinatorial principles of these verbs with different prepositions. The core of the grammar is given in 7.

$$\begin{array}{rcl} S & \rightarrow & V\ T \\ V & \rightarrow & Verb \mid Verb\ SO \\ C & \rightarrow & Conjonction \mid , \\ LT & \rightarrow & ENE\ C\ T \\ T & \rightarrow & (SO)\ (det)\ ENE \mid \\ & & (SO \mid ENE)\ T \mid (SO)\ LT \end{array} \quad (7)$$

SO can be seen as a spatial adverbial clause.

Of course, in order to take into account the combinations, which by their structure are inconsistent with French, the real grammar is more complex. The VT construction grammar reuse a sub-set of the concepts employed in a traditional parts-of-speech (POS) grammar.

The bottom-up parser, based on the real grammar and implemented with a cascade of transducers, can be viewed as searching through the space of possible parse trees to find the correct parse tree for a given 'VT' phrase. Then if a correct parse tree is found the ENE becomes a candidate to be an Extended Spatial Named Entity (ESNE).

Finally, consider the following sentences :

Emprunter successivement ***rue des Capucins*** *et* ***rue de Compostelle***.
'Walk down **Capucins Street** and then **Compostelle Street**.'
Prendre à gauche après l'entrée de ***l'usine de Fontanille***.
'Turn left after the entry to the **Fontanille factory**.'
Suivre la route depuis le ***hameau Lic*** *jusqu'à la* ***Chapelle Saint-Roche***.
'Follow the road from the **hamlet Lic** to the **Chapelle Saint-Roche**.' (8)

These sentences are extracted from a French hiking description. For each of them the cascade of transducers found a correct parse tree. So each marked ENE becomes a potential ESNE but first all the ambiguities must be removed. In fact, most of the descriptive proper names used to build the ENE in these sentences are very common proper nouns and moreover

refer to small localised objects. These are specific aspects that may cause ambiguity.

For our experiments we used the multilingual Perdido corpus [24], which is a TEI [25] compliant gold-standard corpus containing 90 hiking descriptions (French, Spanish and Italian) manually annotated. Hiking descriptions are a specific type of document describing displacements using geographical information, such as toponyms, spatial and motion relations, and natural features or landscapes. The corpus analysis shows that only 2% of ENE are not referring to spatial entities. Furthermore, 53% of the occurrences of ESNE are contained within a VT structure and 47% are associated with feature types (i.e., 53% of ESNE belong to the level 0) and a very few number of ESNE (3%) are built with more than one expansion (level $>$1). Additionally, about 59% of verbs are motion verbs. Median and final motion verbs are the most frequent ones and only 3% of verbs belonging to a VT structure refer to verbs of perception.

C. NERC Processing and Evaluation

As we saw above, we consider only two types of named entities: spatial and non-spatial, and ENE and ESNE are considered as described in the previous sections. With respect to the NERC task, we implemented the construction grammars previously described using an hybrid solution combining a pre-processing POS analysis, a cascaded finite-state transducers for annotating the segments in a text containing valuable information, and external resources for the named entity classification task. The pre-processing component of the Perdido processing chain (PPC) transforms and pre-annotates raw texts with different process: sentence splitting, tokenisation, lemmatisation, and POS tagging. These shallow linguistic tasks are language dependent and are done by standards POS taggers. We propose to integrate different POS taggers in order to solve language or performance issues. Thus, we developed an integration framework designed to handle the output provided by different POS taggers, which use various tag-sets to assign grammatical categories of words. The integration framework implements a generic transformation to standardise tag-sets in order to turn the POS taggers output into a compliant input format for the next component of the PPC. The main component of the PPC deals with the automatic annotation of ENE and geospatial information such as VT structures. The proposed cascaded finite-state transducers, which annotates spatial information and ENE was developed using the CasSys program available in the Unitex platform [26]. For the development of the PPC, we have followed the principles introduced for the development of the CasEN system [27], which implements a combination of two cascaded finite-state transducers. The first one called *analysis cascade* is the core of the annotation process, it executes a sequence of transducers which annotate elements in a specific order. The second cascade called *synthesis cascade* transforms the output of the first cascade (XML-CasSys) into the TEI-compliant XML markup language described in [28]. We have designed web services [29] for the POS and NERC components of the PPC. The Perdido POS web service returns the result of the POS processing using the Unitex compliant input format and the Perdido NERC web service returns the TEI-compliant XML results.

The NERC task was evaluated using both manual POS processed texts (POS 100% corrected) and a fully automatic process (automatic POS processed texts) in order to show the percentage of errors introduced during the pre-processing step of our method. The configuration for experiments done with manually corrected POS is called Perdido I hereafter and the configuration for experiments done with POS automatically processed Perdido II.

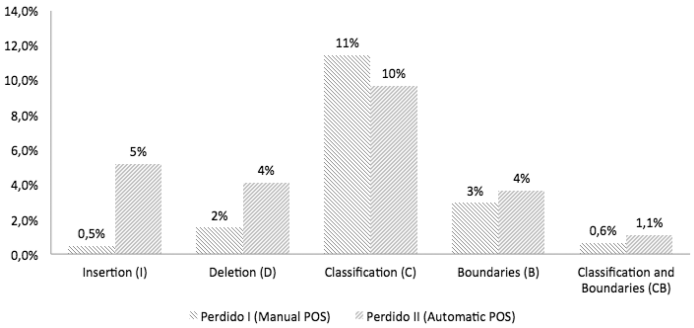


Figure 2. Comparison of the percentage of slot errors of Perdido I and Perdido II (French).

TABLE I. NUMBER OF CORRECTLY DETECTED ENE WITH PERDIDO I AND II (FRENCH).

	N	*Perdido I*		*Perdido II*	
level 0	304	235	77%	244	80%
level 1	332	302	91%	280	84%
level 2	20	16	80%	17	85%
level 3	4	0	0%	1	25%
total	660	553	84%	542	82%

Table I shows the number of ENE that were correctly detected by Perdido I and Perdido II without any errors and the column 'N' shows the reference number of ENE in the French Perdido gold-standard corpus. The evaluation of the automatic NERC task gives a number of correct recognition (i.e., true positives) of 553 ENE with Perdido I and 542 ENE with Perdido II over a total number of 660 ENE, which represents 84% and 82% respectively with Perdido I and Perdido II. For further details concerning the evaluation of the results, we used the SER metric [30] which represents the total slot error rate taking into account the different types of errors related with the NERC task (i.e., insertion, deletion, classification, boundary detection and both classification and boundary detection).

Fig. 2 shows the comparison of the percentage of the different slot errors used in the SER metric. Each bar on this chart refers to the percentage of errors, thus, the lower the percentages are, the better the results are. Concerning errors of insertion (i.e., false positives), it can be seen that Perdido II (5%) makes more errors than Perdido I (0.5%). This can be explained by the fact that as Perdido I is based on a manually corrected POS pre-processing, there is no ambiguity or mistake concerning which words are proper names or not. This can explain also errors of deletion, 4% with Perdido II and only 2% with Perdido I. Then the difference of 1% between Perdido I and Perdido II concerning classification errors is not significant. Indeed, the percentage of classification errors refers to the number of errors over the number of detected entities (i.e., deletion errors are not taken into account in the calculation). The evaluation process gives a total SER of 10% with Perdido I and 17% with Perdido II. As expected, the

 ISBN: 978-1-61208-539-5

Perdido I configuration, which is based on a manual POS analysis, obtains better results than the Perdido II configuration. Approximately seven percent of the errors are introduced by the POS pre-processing step of our method. However, considering the different levels of encapsulation (ENE) and all the different types of errors, 17% of SER and 82% of correct recognition of ENE is a good score.

IV. Conclusion and Future Work

With respect to the annotation of spatial information used to extract geographical data from text-based spatial descriptions, we have proposed a geoparser based on construction grammars implemented with two cascaded finite-state transducers. As a computational synthesis of the work on the expression of space and motion in natural languages, we described the construction grammar VT which aims to mark and formalise the relations between ENE, geographical terms, spatial relations and movement verbs. We have shown that the hierarchical overlapping introduced by the concept of ENE is very helpful to detect a local context associated with NE. For instance, the local context contained within ESNE, such as feature types, helps to produce a detailed description that can be used for a better analysis of the spatial information and a better disambiguation of places. The feasibility of our proposal has been evaluated using a corpus of hiking descriptions and obtains an overall SER score of 17%.

To our knowledge NERC is an important pre-processing step for most of these tasks and automatic NERC process for Indo-European languages might be more or less challenging (e.g., for German it is especially challenging). Our proposal relies on the TEI standard which is widely used in digital humanities and linguistics for Indo-European languages. Thus, the work in progress is to define several other specific finite-state transducers, each one adapted to the specific needs of a given Indo-European language but all based on the same generic core layer. The proposed generic core layer may be used to create and share pre-processed corpus.

Acknowledgment

This work has been partially supported by: the Communauté d'Agglomération Pau Pyrénées (CDAPP) and the Institut National de l'Information Géographique et Forestiére (IGN) through the PERDIDO project; the Spanish Government (project TIN2012-37826-C02-01); and the Aragon and Aquitaine Regional Governments cooperation programme through the YACA project.

References

[1] R. R. Larson, "Geographic Information Retrieval and Spatial Browsing," GIS and Libraries: Patrons, Maps and Spatial Information, Apr. 1996, pp. 81–124.

[2] OpenCalais, http://www.opencalais.com/, 2017, [accessed 2017-01-12].

[3] OpenNER, http://opennlp.apache.org/, [accessed 2017-01-12].

[4] CasEN, http://tln.li.univ-tours.fr/Tln_CasEN_eng.html, [accessed 2017-01-12].

[5] Stanford-NER, http://nlp.stanford.edu/ner/, [accessed 2017-01-12].

[6] L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio, "Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus," in Proceedings of the 22Nd ACM SIGSPATIAL. Dallas, TX, USA: ACM, 2014, pp. 183–192.

[7] C. J. Fillmore, "Syntactic Intrusions and The Notion of Grammatical Construction," Annual Meeting of the Berkeley Linguistics Society, vol. 11, no. 0, Jun. 1985, pp. 73–86.

[8] G. Lakoff, Women, fire, and dangerous thinks – What categories reveal about the mind. University of Chicago Press, 1987.

[9] R. W. Langacker, Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites. Stanford, CA: Stanford University Press, 1987.

[10] Y. Yannick-Mathieu, "La Grammaire de Construction," Approches syntaxiques contemporaines, no. 48, 2003, pp. 43–56.

[11] T. Poibeau, "Extraction automatique d'information: du texte brut au web sémantique," in Extraction automatique d'information: du texte brut au web sémantique. Hermès Lavoisier, 2003.

[12] N. Friburger and D. Maurel, "Finite-state transducer cascades to extract named entities in texts," Theoretical Computer Science, vol. 313, no. 1, Feb. 2004, pp. 93–104.

[13] L. F. Rau, "Extracting Company Names from Text," in Artificial Intelligence Applications. Miami Beach: IEEE, 1991, pp. 29–32.

[14] J. L. Leidner, Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal-Publishers, Jan. 2007.

[15] D. Buscaldi and P. Rosso, "A conceptual density-based approach for the disambiguation of toponyms," Int. J. Geogr. Inf. Sci., vol. 22, no. 3, Jan. 2008, pp. 301–313.

[16] D. A. Smith and G. Crane, "Disambiguating Geographic Names in a Historical Digital Library," in Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ser. ECDL '01. London, UK: Springer-Verlag, 2001, pp. 127–136.

[17] K. Jonasson, Le nom propre. Duculot, Louvain-la-Neuve, Belgium, 1994.

[18] R. S. Purves and C. Derungs, "From Space to Place: Place-Based Explorations of Text," International Journal of Humanities and Arts Computing, vol. 9, no. 1, Mar. 2015, pp. 74–94.

[19] M. Aurnague, "How motion verbs are spatial: The spatial foundations of intransitive motion verbs in French," Lingvisticae Investigationes, vol. 34, no. 1, 2011, pp. 1–34.

[20] L. Talmy, How language structures space, ser. Berkeley cognitive science report. Berkeley, CA, Etats-Unis: Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley, 1983, no. 4.

[21] C. Vandeloise, L'Espace en français. Sémantique des prépositions spatiales. Editions du Seuil, 1986.

[22] S. Pourcel and A. Kopecka, "Motion expression in French: typological diversity," Durham & Newcastle working papers in linguistics, vol. 11, 2005, pp. 139–153.

[23] J.-P. Boons, "La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs," Langue Française, no. 76, 1987, pp. 5–40.

[24] L. Moncla, M. Gaio, J. Nogueras-Iso, and S. Mustière, "Reconstruction of itineraries from annotated text with an informed spanning tree algorithm," International Journal of Geographical Information Science, vol. 30, no. 6, 2016, pp. 1137–1160.

[25] TEI, "Text encoding initiative," http://www.tei-c.org/, 2017, [accessed 2017-01-12].

[26] Unitex, "Unitex/gramlab: an open source, cross-platform, multilingual, lexicon- and grammar-based corpus processing suite," http://www-igm.univ-mlv.fr/~unitex/, 2017, [accessed 2017-01-12].

[27] D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella, and D. Nouvel, "Cascades de transducteurs autour de la reconnaissance des entités nommées," TAL, vol. 52, no. 1, 2011, pp. 69–96.

[28] L. Moncla and M. Gaio, "A Multi-layer Markup Language for Geospatial Semantic Annotations," in Proceedings of the 9th Workshop on Geographic Information Retrieval, ser. GIR '15. Paris, France: ACM, 2015, pp. 5:1–5:10.

[29] PERDIDO, "Expanded named entity annotation service," http://erig.univ-pau.fr/PERDIDO/api.jsp, 2017, [accessed 2017-01-12].

[30] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in In Proceedings of DARPA Broadcast News Workshop, 1999, pp. 249–252.

 ISBN: 978-1-61208-539-5