Extraction of Geospatial Information from Documents

Sabiha Tahrat
Lirmm, UM2
Montpellier, France
tahrat@lirmm.fr

Mathieu Roche
Lirmm, UM2, CNRS
Montpellier, France
mroche@lirmm.fr

Maguelonne Tesseire
UMR TETIS
Montpellier, France
teisseire@teledetection.fr

1. INTRODUCTION

Geographical or spatial information is now included in most of exchanged data. Sometimes, it is directly provided through metadata, but it is very often hidden and it becomes crucial to automatically discover it.

Natural Language Processing (NLP) and Data Mining communities have thus merged their efforts in order to extract geospatial information from textual documents, web pages, field data, and so forth. In this way, recent researches take into account the content of documents (e.g. terms) to identify geospatial data or to predict its geographic location.

Nevertheless, spatial information has some specificities that make discovering spatial information and/or spatial correlations from large amount of data still challenging. In this context, some proposals have been focused on the formalization of geospatial concepts and relationships, on the extraction of geospatial relations (e.g. rivers/body of water, town/suburb) in free texts to offer to the database community a unified framework for geodata discovery. Our work is part of the SENTERRITOIRE[1] project dedicated to a decision-making environment based on an automatic analysis of texts related to land planning use. The first step of this project focuses on the automatic extraction of geospatial descriptors. In this paper we describe the methodology we have adopted.

2. SPATIAL INFORMATION EXTRACTION

In this section, we describe a workflow supporting automatic tagging and interpretation of spatial information in document. Firstly, we present the spatial model on which SENTERRITOIRE spatial information process flow relies. Finally, we explain the main stages of the process.

2.1 Spatial Model

The Pivot basic model of J. Lesbegueries [3] (See figure 1) is based on the linguistic hypothesis that a spatial feature (**SF**) is defined from landmarks Named Entities (**NE**) [5] and spatial relationships. This model supports absolute and relative SFs. Named SFs such as *the city of Selles-sur-Cher* are well-known named places, also called Absolute SFs (**ASF**). Complex SFs such as *near Selles-sur-Cher* or *North of Selles-sur-Cher* need some linguistic and spatial reasoning processes. Such features are called Relative SFs (**RSF**). We associate each RSF to one or more spatial relationships (adjacency, inclusion, distance, orientation), derived from the qualitative spatial reasoning area, for a recursive definition [3].

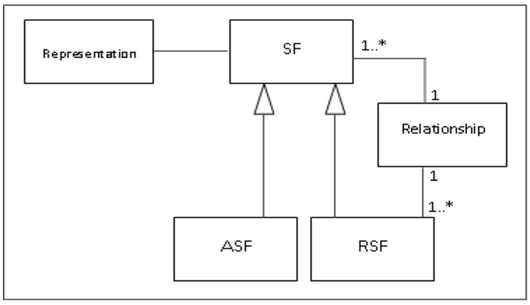


Figure 1: the spatial entity in the Pivot model

2.2 Spatial Information Processing

In this context, the extraction of spatial information is relied on the model already defined and composed of the following steps (See figure 2): (1) Tokenization divides the document into smallest blocks of text; (2) Lexical analysis carries out transformation of these blocks into lexeme and the extraction of NE; (3) Morphosyntactic analysis which allows to retreive words type; (4) Semantic analysis marks ASFs and RSFs using a recovering DCG [2] grammar of indicators, located around these entities. Then, the process produces instances of the Pivot model (ASFs, RSFs). These ASFs are validated using external geographical ressources as Gazetteers.

This Information Extraction (**IE**) process has been validated for textual documents through the development of linguistic processing chains using the platform Linguastream[3].

[1] http://msh-m.fr/programmes-2012/senterritoire.

[2] Definite Clause Grammars.

[3] http://www.linguastream.org.

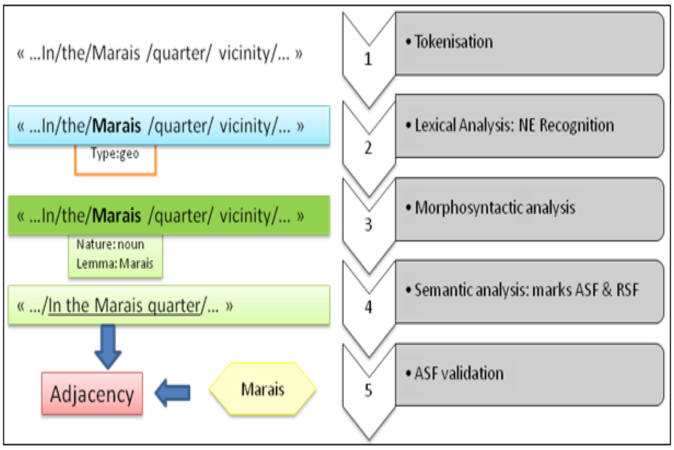

« ...In/the/Marais /quarter/ vicinity/... »
« ...In/the/Marais /quarter/ vicinity/... »
Type:geo
« ...In/the/Marais /quarter/ vicinity/... »
Nature:noun
Lemma:Marais
« .../In the Marais quarter/... »
Adjacency
Marais
1
• Tokenisation
2
• Lexical Analysis: NE Recognition
3
• Morphosyntactic analysis
4
• Semantic analysis: marks ASF & RSF
5
• ASF validation

Figure 2: The linguistic processing phases

2.3 Linguastream chain

LinguaStream is a generic platform used for automatic natural language processing. It allows the design and the evaluation of complex data processing workflow, by assembling various analysis modules.

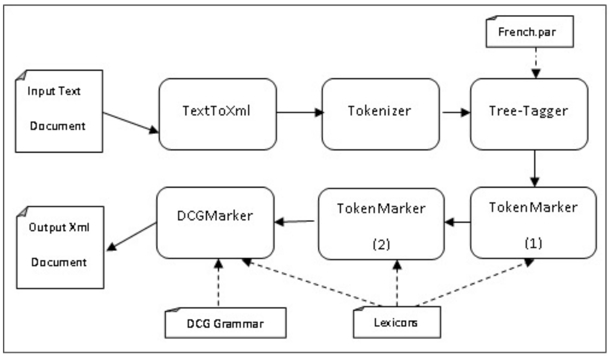


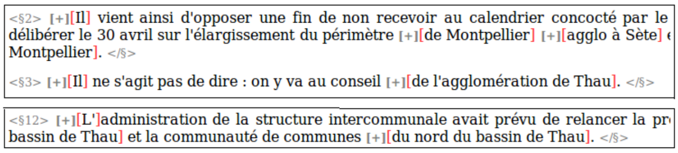
Figure 3: The linguastream processing chain

In our case we distinguish six processing chains listed as follows (See figure 3):

TextToXml converts a plain text document in a file format expected by Linguastream. **Tokenizer** splits the text into lexical units. **TreeTagger** [4] realizes morphosyntactic marking of each lexical unit using an external analyzer: Tree-tagger and a parameter file. **TokenMarker** tags lexical units using a basic regular expressions, in terms of character (words with capital letter, words belonging to a lexicon file). **DCGMarker** realizes a sematic analysis based on the file containing the DCG grammar written in Prolog[4].

3. APPLICATION

This section reports our experimental results to validate the effectiveness and efficency over the Pivot model on texts retrieved from web sites. In particular, local newspapers, association fora addressing the redevelopment issues of the Thau lagoon. According to this model and as shown in figure 4, we highlight the following feedbacks:



<§2> [+][Il] vient ainsi d'opposer une fin de non recevoir au calendrier concocté par le délibérer le 30 avril sur l'élargissement du périmètre [+][de Montpellier] [+][agglo à Sète] Montpellier]. </§>

<§3> [+][Il] ne s'agit pas de dire : on y va au conseil [+][de l'agglomération de Thau]. </§>

<§12> [+][L']administration de la structure intercommunale avait prévu de relancer la pr bassin de Thau] et la communauté de communes [+][du nord du bassin de Thau]. </§>

Figure 4: Result of the Linguastream workflow chain on a text

Each word with capital letter is considered as a spatial entity. The resulting noises (e.g. Le, Il) will be removed using Geonames[5] plus a probabilistic unsupervised approach of L. Bonnefoy [2] at the validation step.
Some SFs intruducters may not be detected due to the non exaustivity of the lexicon file. This latter can be enriched using a pattern-based approach of MN. Bessagnet [1].
DCG grammar does not detect Spatial relationships and introductors when these ones are located after the absolute spatial entity (e.g. Montpellier centre, Montpellier sud). To solve this problem, we have extended the grammar.

4. CONCLUSIONS

This paper studies the relevance of the Pivot model over our dataset, and how to tackle some weaknesses related to resources, and the richness of the language. Prospects for this work consist in realizing the proposed solutions for the extraction step of the SFs from documents. In the second stage, we will proceed with the extraction of semantic relations between theses SFs. Finally the SF extraction methods will be rigorously evaluated with TETIS[6] experts.

5. ACKNOWLEDGMENTS

A special thanks to Julien Lesbegueries and Patrice Enjalbert for their help in this project.

6. REFERENCES

[1] M.-N. Bessagnet, M. Gaio, E. Kergosien, and C. Sallaberry. Extraction automatique d'un lexique à connotation géographique à des fins ontologiques dans un corpus de récits de voyage. In *Conférences sur le TALN, Montréal, 19-23/07/2010*, page 10 pages, Montréal, Canada, July 2010.

[2] L. Bonnefoy, P. Bellot, and M. Benoit. Une approche non supervisée pour le typage et la validation d'une réponse à une question en langage naturel : application à la tâche entity de trec 2010. In *CORIA*, pages 191–206, 2011.

[3] J. Lesbegueries, M. Gaio, and P. Loustau. Geographical information access for non-structured data. In *SAC*, pages 83–89, 2006.

[4] H. Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. 1994.

[5] C. Vandeloise. *L'espace en français*. Seuil, Oct. 1986.

[4] http://www.swi-prolog.org

[5] http://www.geonames.org: services available on the web and allowing named entities to be georeferenced

[6] http://tetis.teledetection.fr