Text2Geo: from textual data to geospatial information

Sabiha Tahrat, Eric Kergosien, Sandra Bringay, Mathieu Roche, Maguelonne Teisseire

Text2Geo: from textual data to geospatial information

Sabiha Tahrat
LIRMM
sabiha.tahrat@lirmm.fr

Eric Kergosien
LIRMM, TETIS
eric.kergosien@lirmm.fr

Sandra Bringay [*]
LIRMM, MIAp
bringay@lirmm.fr

Mathieu Roche [†]
LIRMM
mroche@lirmm.fr

Maguelonne Teisseire [‡]
Irstea, UMR TETIS
teisseire@teledetection.fr

ABSTRACT

In this paper, we focus on methods for extracting spatial information in text documents. After presenting textual description of space and manual annotation of named entities, mainly location and organization, we present our proposal Text2Geo. It is a hybrid method which combines information extraction approach based on patterns with a supervised classification approach to explore context. We discuss some results obtained on the dataset of *Thau* lagoon.

1. INTRODUCTION

Extracting spatial information from Web documents is still challenging. In this context, we aim at providing geographers and environmentalists with automatic tool for knowledge discovery. In this paper, we focus on retrieving spatial information contained in textual corpora from the Web. In [10], a linguistic method based on patterns called PIV is define to extract Spatial Entities from texts. To do this, a cognitive model, called "Pivot" , is proposed to define the spatial feature (SF). In this model, SF is composed of at least one Named Entity (denoted NE) and one variable number of spatial indicators specifying its location. SF can then be identified in two ways:

1. an **absolute spatial feature (A_SF)** one NE allowing a geo-localization, such as $<(spatialIndicator)^*$, *NE of Location*$>$ (ex: *the city of Sevilla*).
2. a **relative spatial feature (R_SF)** one spatial relationship (topological or Euclidean) with at least one SF (ex: *in the south of Madrid*). An R_SF is defined as $<(spatialrelation)^{1..*}, A\_SF>$ or $<(spatialrelation)^{1..*}, R\_SF>$. Five spatial relation types are considered: orientation, distance, adjacency, inclusion, and geometric which defines union or intersection linking two SFs.

In our proposal, we focus on data mining techniques to qualify terms as NEs, including Places or Organizations. For example, let us consider the two sentences "The marriage happens at the City Hall" and "The City Hall finances this project". In the first one, "City Hall" refers to a Place and in the second one to an Organization. Our aim is to define patterns in order to discover these NEs [11]. Recent web based approaches establish links between features and their types (or categories) [3]. To recognize classes of NEs, many approaches rely on supervised learning methods. Such algorithms exploit various features and labeled data. For instance, types of features are positions of the candidates, grammatical labels, lexical information [4], etc. In our approach, we combine such methods of supervised learning with linguistic patterns.

Our contribution is twofold: (1) we refine and enrich the information extraction patterns existing in the literature by improvement of the semantic extraction of spatial features and (2) we define an original approach using various mining techniques from text in order to distinguish between SF and Organization.

2. STATE OF ART ON EXPTRACTION METHODES OF SPATIAL FEATURES

Named entities (NEs) were defined as name of persons, places and organizations in american campaign assessments called MUC (Message Understanding Conferences), organized in the 90s. The main issue was to extract information such as ENs from documents (U.S. Navy messages, stories of terrorist attacks, etc.). As indicated by [5], more elements can be brought to these classes. For instance, [13] define new classes as Document (software, materials, machines) and Scientific (illness, medications, etc).

Many methods are used to recognize ENs in general and ES in particular [12]. Among of extraction methods based on texts, statistical approaches usually involve a study of co-occurring terms by analyzing their distribution in a corpus [2] or measures calculated the probability of occurrence of

[*]MIAp group, Univ. Paul-Valéry, Montpellier France

[†]LIRMM - CNRS, Univ. Montpellier 2, 161 rue Ada 34095 Montpellier Cedex 5 France

[‡]Irstea, UMR TETIS, 500, rue J.F.Breton 34093 MONTPELLIER Cedex 5 France

a set of terms [14]. These approaches have some drawbacks as they do not always allow to qualify terms as being ENs, especially ENs that qualify Place or Organization. Pattern mining methods extract rules (called rules transduction) in order to spot the ENs [11]. These rules use syntactic information specific to sentences [11]. Recent studies rely on the Web to establish links between entities and their type (or category). For instance, in [3], the approach is based on the probabilities of word occurrences in the associated pages for a given entity compared to similar distributions on types. Overall, relations can be identified by similarities between their syntactic contexts [7], a prediction by using Bayesian networks [15], by techniques of text mining [8] or by inference of knowledge by means of learning algorithms [6]. These methods are efficient, but they do not always identify the semantic relationships.

For the recognition of ENs classes, many approaches rely on supervised learning methods. These learning methods such as SVM [9] are often used in the challenge Conference on Natural Language Learning (CoNLL). The algorithms exploit various features as well as data labeled by experts. Types of features are for example the candidate's positions, grammatical labels, lexical information (e.g. upper /lower) affixes, all words closed to the candidate [4]. In this paper, the proposed approach combines supervised learning methods and linguistic patterns.

3. TEXT2GEO: TOWARDS A NEW PROCESS OF EXTRACTING SPATIAL INFORMATION

3.1 Text2Geo and the addition of new rules

We adopt a classical Natural Language Processing (NLP) process in the field of Geographic Information Retrieval (GIR) [1]: (i) lemmatization, (ii) morpho-lexical analysis, (iii) syntactic analysis, (iv) semantic analysis. In this process, the *lemmatization* step determines the lemma for a given word. The *morpho-lexical analysis* is the identification of the structure of a given language's linguistic units defined in a lexicon. The *syntactic analysis* is the formal analysis of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relations to each other. At the end, the *semantic analysis* allows to extract more specific interpretation of the chosen sentences in order to identify the potential meaning conveyed by a word or a set of words. We extend the NLP sequence defined by [10] (basic patterns) by using Linguastream[1] (Cf. Figure 1).

First, we add patterns in the semantic analysis step (Cf. DCG Marker step in our process) in order to improve the automatic identification of the A_SF and R_SF. And secondly, we propose a new type of patterns to identify specifically NE such as *Organization*.

New patterns to identify A_SF and R_SF. The SF annotation is based on the classical typology of the domain and more precisely on the sub-types of locations. Locations can be polysemous: human constructions (e.g. buildings) and addresses (e.g. streets). In this context, rules (patterns) has been added to improve the identification of A_SF and R_SF. For example, added patterns identify the distribution of spatial relationships (e.g. *near Lyon and Marseille* $\rightarrow$ *near Lyon* + *near Marseille*). Here, R_SF is improved by the rule $< (R\_SF)^{1..*}, SpatialSep, A\_SF >$ and $SpatialSep$ is defined as $< ","|";"|"and"|"or". >$. Other types of rules have been added, which increased the number of extracted SFs and improved their quality (See section 4).

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

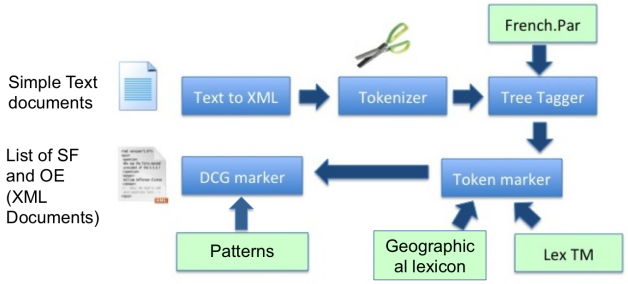

Figure 1: Text2Geo, a process of extracting spatial information

New patterns to identify Organization. New rules identify another NE type: Organizations (OE). The addition of specific rules allows to identify Organizations which could be confused with SF. Such rules are: (1) an OE is followed by an *action verb*; (2) an OE is preceeded by prepositions: with, by, for, on behalf of, etc.

These rules take into account a reduced local context. The use of a larger context to distinguish SFs and OEs can be relevant. Hence, we propose in section 3.2 a hybrid approach combining a methodology related to information retrieval and Text2Geo patterns.

3.2 Towards a hybrid method

We propose to learn a model that can distinguish between Organization and SF. For this task, we apply an Information Retrieval (IR) process based on four steps (Cf. Figure 2):

- **Step 1: Construction of a learning corpus**. The first step consists in acquiring a learning corpus. It is composed of sentences containing an entity (SF or Organization). The aim is building a model in order to predict which type of entity is present in the sentences.
  Each sentence is manually labeled (SF or Organization). Note that ambiguous sentences with both entities are not taken into account in our learning corpus.

- **Step 2: Representation of textual data**. Each sentence is described by a vector. Rows represent words (i.e features) and columns represent sentences. Each cell contains the weight (e.g. boolean weight) of the corresponding word in the corresponding sentence. Moreover, each sentence is associated to a class (i.e. SF or Organization) in order to learn the model of the next step.

- **Step 3: Learning process**. This step consists in training a classifier (i.e. supervised learning) in order to decide which sentence contains a SF or an Organization. The built prediction model will be apply on new data.

- **Step 4: Prediction**. This learnt model is applied on unlabeled textual data in order to predict the type of entity present in sentences.

The learning process of Text2Geo approach is based on two conventional methods in data mining, i.e. Naive Bayes and SVM [9], and the used features are words of sentences ("bag of words"). The originality of our hybrid approach is to consider proposed patterns as features in the learning model. In this way, these boolean features are at 1 when a sentence contains a pattern of type $< ConceptOrg, NE >$ (design for an OE) or $< ConceptSpa, NE >$ (design for a SF). *ConceptOrg* represents typical prepositions preceding Organization (*with, by*, etc.). *ConceptSpa* is divided into three subconcepts (preceding SF): spatial prepositions (in, on, etc.), relationship indicators (south, towards, etc.), and spatial indicators (city, area, etc.).

In our experiments, each type of features are independently evaluated. It gives more weight to words in relation with Geographic Information topics (prepositions and spatial organization, spatial and relationship indicators defined in a dictionary). For instance, such words (i.e. stop words) are less considered or removed with a basic IR process. In addition, the classical *bag of words* approach does not take into account the order of words. By using, in our learning model, a partial order for some linguistic features makes it possible.

4. EXPERIMENTS

The corpus is a collection of articles selected since 2006 in the French daily newspaper *Midi Libre*. The articles deal with issues of community redevelopment of the *Thau* lagoon and its economic and environmental issue. At first, we evaluated both NLP sequences (basic patterns vs. Text2Geo patterns) from a subset of the corpus of 20 articles (8141 words). The evaluation is based on precision (i.e. proportion of relevant entities extracted), recall (i.e. proportion of relevant entities extracted regarding all relevant entities), and F-measure (i.e. combination of precision and recall). Table 1 shows that enrichment of initial patterns, based on the Pivot model, significantly improves the precision and recall results. The rate of F-measure is more than doubled. Moreover, our patterns allow to identify Organizations with precision at 92 %. Adding rules in future work should improve recall.

In our experiments, the training set consists of 272 sentences: 138 sentences containing spatial features and 134 sentences containing organizations. Each sentence is then lemmatized and represented by a binary vector. The best results were obtained with SVM and Naive Bayes (by using Weka[2]). Table 2 shows the associated confusion matrix for both classes (Spatial Features and Organizations) for cross

[2]http://www.cs.waikato.ac.nz/ml/weka/

Table 1: Text2Geo patterns evaluation

Basic patterns			Text2Geo paterns			
	A_SF	R_SF		R_SF	R_SF	OE
Precision	20%	**48%**	Precision	53%	84%	**92%**
Recall	**63%**	27%	Recall	**94%**	66%	35%
F-mesure	30%	**34%**	F-mesure	67%	**74%**	50%

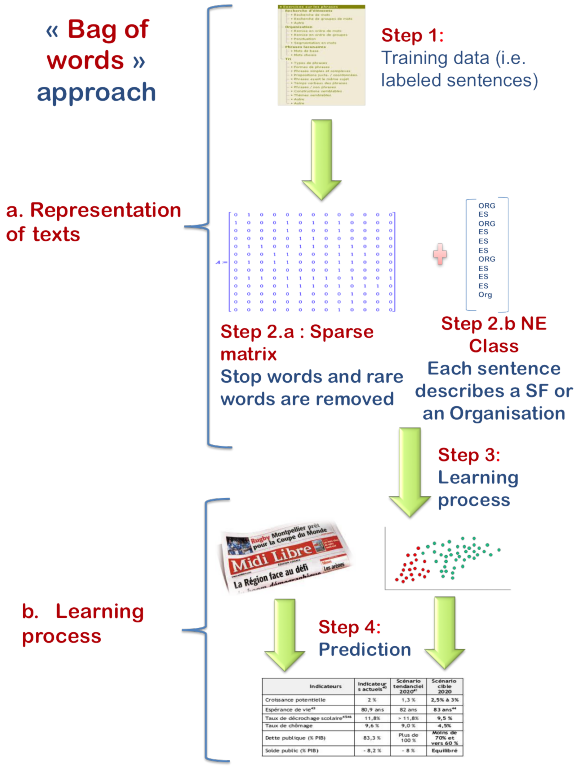

Figure 2: Hybrid method of Text2Geo

validation evaluation. The accuracy rate corresponds to the proportion of well classified examples. The hybrid approach improves the results in terms of accuracy (Cf. Table 3). This shows significant improvement with the use of specific features to spatial features (ConceptSpa) particularly suitable within the context of the hybrid model Text2Geo.

5. CONCLUSION AND FUTURE WORK

In order to extract spatial information from documents, we proposed to use context to distinguish Location and Organization NE types. Our contribution is a set of morpho-syntactic patterns, integrating our NLP sequence Text2Geo. Two supervised learning methods, SVM and Bayes Naives, were used to classify NE (such as Location and Organization) and eliminate possible ambiguities. Experiments show that the enrichment in Text2Geo of initial patterns significantly improves results in terms of precision and recall.

Our prospects aim at applying the supervised learning process to three classes: Organization, A_SF, R_SF. Thus, we will be able to check if the use of a more important local context (the sentence) makes it possible to precisely distin-

Table 2: Classification of the sentences without using the Text2Geo features

SVM	SF	OE	**Naive Bayes**	SF	OE
SF	103	35	SF	98	40
OE	44	90	OE	44	90
Accuracy	*70.96%*		*Accuracy*	*69.12%*	

Table 3: Classification of sentences with constraints

Features with ConceptOrg	SF	OE	**Features with ConceptSpa**	SF	OE	**Both types of features**	SF	OE
SF	108	30	SF	112	26	SF	113	25
OE	47	87	OE	19	115	OE	19	115
Accuracy	*71.69%*		*Accuracy*	*83.45%*		*Accuracy*	*83.82%*	

guish two specific SF types (A_SF and R_SF).
We also plan to compare our approch with the existing tools AlchemyAPI and OpenCalais in order to extract named entity such as Organisations.

6. ACKNOWLEDGMENT

The authors thank Pierre Maurel (IRSTEA, UMR TETIS) for its expertise on the corpus.
This work was partially funded by the labex *NUMEV* and the *Maison des Sciences de l'Homme de Montpellier (MSH-M)*.

7. REFERENCES

[1] M. Abolhassani, N., Fuhr, and N. Gövert. Information extraction and automatic markup for xml documents. In *Intelligent Search on XML Data*, pages 159–178, 2003.

[2] E. Agirre, O. Ansa, E.-H. Hovy, and D. Martìnez. Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*, 2000.

[3] L. Bonnefoy and P. Bellot. Lia-ismart at the trec 2011 entity track: Entity list completion using contextual unsupervised scores for candidate entities ranking. In *TREC*, 2011.

[4] X. Carreras, L. S M Arque, and L. S Padro. A simple named entity extractor using adaboost. In *In Proceedings of CoNLL-2003*, pages 152–155, 2003.

[5] B. Daille, N. Fourour, and N. Morin. Catégorisation des noms propres : une étude en corpus. volume 25, chapter Cahiers de Grammaire, pages 115–129. 2000.

[6] C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, April 2006.

[7] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[8] M. Grčar, E. Klien, and B. Novak. Using Term-Matching Algorithms for the Annotation of Geo-services. In Bettina Berendt, Dunja Mladenič, Marco Gemmis, Giovanni Semeraro, Myra Spiliopoulou, Gerd Stumme, Vojtěch Svátek, and Filip Železný, editors, *Knowledge Discovery Enhanced with Semantic and Social Information*, volume 220, chapter 8, pages 127–143. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[9] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, pages 137–142, 1998.

[10] J. Lesbegueries, C. Sallaberry, and M. Gaio. Associating spatial patterns to text-units for summarizing geographic information. In *Proceedings of ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop*, pages 40–43. LIUPPA, Aug 2006.

[11] D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella, and D. Nouvel. Casen: a transducer cascade to recognize french named entities. *TAL*, 52(1):69–96, 2011.

[12] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January 2007.

[13] W. Paik, E.-D. Liddy, E. Yu, and M. Mckenna. Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition*, pages 61–73. MIT Press, 1996.

[14] P. Velardi, P. Fabriani, and M. Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pages 270–284, 2001.

[15] D. Weissenbacher and A. Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents: l'intérêt de la classification bayésienne. In *Proceedings of Traitement Automatique des Langues Naturelles*, pages 145–155, France, 06 2007. ATALA.