Text2Geo : des données textuelles aux informations géospatiales

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

* LIRMM, UMR 5506, 161 rue Ada, 34392 Montpellier - France
{prénom.nom}@lirmm.fr
** UMR TETIS, 500 rue Jean-François Breton, 34093 Montpellier - France
{prénom.nom}@teledetection.fr
*** LIUPPA, UPPA, Avenue de l'Université, BP 576, 64012 PAU cedex - France
{prénom.nom}@univ-pau.fr

Résumé. Dans cet article, nous nous intéressons aux méthodes d'extraction d'informations spatiales dans des documents textuels. Nous présentons la méthode hybride Text2Geo qui combine une approche d'extraction d'informations, fondée sur des patrons avec une approche de classification supervisée permettant d'explorer le contexte associé. Nous discutons des résultats expérimentaux obtenus sur le jeu de données de l'étang de Thau.

1 Introduction

Au-delà de sa stricte définition d'entité administrative et politique, le territoire, selon Guy Di Méo, témoigne d'une *"appropriation à la fois économique, idéologique et politique de l'espace par des groupes qui se donnent une représentation particulière d'eux-mêmes, de leur histoire, de leur singularité"* (Di Méo (1998)). Dans ce contexte éminemment subjectif, la caractérisation et la compréhension des perceptions d'un même territoire par les différents acteurs est difficile, mais néanmoins particulièrement intéressante dans une perspective d'aménagement du territoire et de politique publique territoriale. Le travail présenté s'inscrit dans le cadre du projet Senterritoire [1], qui adopte une démarche pluridisciplinaire, initiée à partir d'une méthode automatique et visant à fournir aux géographes et aux environnementalistes, une aide à la découverte de connaissances. Nos contributions portent, dans cette publication, sur l'accès à l'information spatiale et proposent (1) d'affiner et d'enrichir les patrons d'extraction d'informations existants dans la littérature afin d'améliorer l'identification du sens de l'entité spatiale extraite et (2) de définir une approche originale utilisant différentes techniques de fouille de textes afin de distinguer une entité spatiale d'une entité d'organisation.
La suite de l'article est organisée de la façon suivante. En section 2, nous présentons les définitions préliminaires et les travaux du domaine. En section 3, nous décrivons la méthode hybride Text2Geo. En section 4, nous présentons les expérimentations réalisées sur le jeu de données du bassin de Thau et concluons dans la section 5.

1. http ://www.msh-m.fr/programmes/senterritoire/

2 Etat de l'art

De l'information géographique à l'information spatiale textuelle. L'expression de l'**information géographique** est communément définie comme la composition d'une entité spatiale, d'une entité temporelle et d'une entité thématique. Dans cet article, nous nous limitons au traitement des entités spatiales (ES) et nous nous appuyons pour cela sur le modèle Pivot (Lesbegueries (2007)) qui permet d'interpréter la plupart des ES exprimées en langage naturel dans les textes. Dans ce modèle, l'ES est constituée d'au moins une Entité Nommée (notée EN) et d'un nombre variable d'indicateurs spatiaux, précisant sa localisation. On distingue : 1) Une **entité spatiale absolue (ESA)** qui est une référence directe à un espace géo-localisable comme *la ville de Séville* ; 2) Une **entité spatiale relative (ESR)** définie à l'aide d'au moins une autre ES et d'indicateurs spatiaux d'ordre topologique (par exemple *au sud de la ville de Séville*). Dans la section suivante, nous présentons les techniques existantes pour extraire l'information spatiale dans des textes.

Les méthodes d'extraction des entités spatiales. L'extraction d'ENs consiste à rechercher des objets textuels de type ENs. Les ENs ont été définies comme les noms de Personnes, Lieux et Organisations lors des campagnes d'évaluation américaines MUC [2], et plus récemment ACE [3]. Ces classes ont depuis été enrichies (Maurel et al. (2011)). De nombreuses méthodes permettent de reconnaitre les ENs en général et les ES en particulier (Nadeau et Sekine (2007)). On trouve des approches statistiques consistant généralement à étudier les termes co-occurrents par analyse de leur distribution dans un corpus (Agirre et al. (2000)) ou par des mesures calculant la probabilité d'occurrence d'un ensemble de termes (Velardi et al. (2001)). Ces méthodes ne permettent pas toujours de qualifier des termes comme étant des ENs et notamment les ENs de type Lieu ou Organisation. On trouve également des méthodes de fouille de données fondées sur l'extraction de motifs. Ces derniers permettent de déterminer des règles de transduction utilisant des informations syntaxiques propres aux phrases pour repérer les ENs (Nouvel et al. (2011)). Pour la reconnaissance des classes d'ENs, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé comme les SVM (Joachims (1998)). Les algorithmes exploitent divers descripteurs (positions des candidats, étiquettes grammaticales, informations lexicales, etc.) et des données expertisées/étiquetées. Dans cet article, nous combinons de telles méthodes d'apprentissage supervisé associées à des patrons linguistiques.

3 Text2Geo : Vers un nouveau processus d'extraction d'information spatiale

L'extraction des entités spatiales avec Text2Geo. Nous appliquons une chaîne de Traitements Automatiques du Langage Naturel classique dans le domaine de la recherche d'information géographique (Abolhassani et al. (2003)) : (1) la **lemmatisation** (segmentation des mots et identification de leur lemme) ; (2) l'**analyse lexicale et morphologique** (identification de la catégorie grammaticale - nom, adjectif, etc. - et des paramètres de flexion - nombre, temps, etc.) ; (3) l'**analyse syntaxique** (identification du rôle des termes ou des syntagmes dans la

2. Message Understanding Conferences - années 90
3. Automatic Content Extraction - 1999-2008

phrase à partir de grammaires) ; (4) l'**analyse sémantique** (identifier de sens potentiel véhiculé par des mots ou des groupes de mots sur la base des syntagmes retenus). Cette chaîne de TALN est définie avec Linguastream [4] qui intègre notamment l'étiqueteur grammatical *TreeTagger* [5] et le langage prolog pour la définition des grammaires DCG (analyses syntaxique et sémantique).
Sur la base de la grammaire définie par (Lesbegueries (2007)) dans ces phases d'analyses syntaxique et sémantique, nous avons mis en place de nouveaux patrons dédiés à l'extraction des entités spatiales et des entités d'organisation. Cette extraction se fait selon deux étapes : **L'étape 1** extrait les ESA qui constituent les types primitifs de notre processus d'extraction. Ces types primitifs sont soit des entités nommées de lieu (Montpellier, France...), soit des indicateurs spatiaux (la région, la ville) ou alors des indicateurs de relation (Le sud...). Ceci se traduit en logique par des règles comme :
$ES \Rightarrow ESA.$; $ES \Rightarrow ESR.$
$ESA \Rightarrow IndicateurSpatial, ESA.$; $ESA \Rightarrow NomToponymique.$
Pour chaque règle définie ci dessus, "$\Rightarrow$" dans l'expression "$ESA \Rightarrow NomToponymique$" signifie que l'expression ESA est composée de l'expression NomToponymique, correspondant à un nom de lieu. Les deux premières définitions ESA sont récursives, ce qui permet de produire des patrons de tailles variables afin d'identifier des instances telles que : *Les régions rurales du sud de la France, la ville de Madrid, les communes de l'agglomération du bassin de Thau*.
L'étape 2 extrait les instances les plus complexes : les entités spatiales relatives, composées d'une ESA et précédées d'une relation d'ordre topologique suivant des règles du type :
$ESR \Rightarrow Relation, ESR.$; $ESR \Rightarrow Relation, ESA.$
$Relation \Rightarrow Adjacence \,|Orientation\,|Inclusion\,|Distance\,|Forme\ géometrique,$
$Adjacence \Rightarrow$ "$prés$" |"$lapériphérie$" $|etc.$

Dans cette chaîne de traitements, nous proposons deux contributions. Dans un premier temps, nous avons ajouté des règles à la grammaire afin d'améliorer l'identification des ESR et ESA. Dans un second temps, nous avons proposé un nouveau type de règles pour repérer de manière spécifique les entités nommées de type *Organisation*.

Définition de nouveaux patrons pour l'identification des ESA et ESR. Pour annoter les ENs spatiales, nous nous sommes appuyés sur la typologie classique du domaine qui identifie des sous-classes : les *lieux géographiques naturels* (lacs, mers, etc), les *constructions humaines* (buildings, installations, etc.), les *axes de circulations* (routes, etc.), les *adresses* (rue, code postal, etc.). Nous avons ajouté des règles (patrons) permettant d'améliorer l'identification des ESA et ESR. Par exemple, l'ajout d'un patron lié à la distribution des ES permet d'identifier le cas lié à la distribution des relations spatiales. Ainsi, dans la phrase *Les environs de Lyon, Marseille...*, nous identifions deux entités spatiales.

Définition de nouveaux patrons pour l'identification des Organisations. Nous avons ajouté des règles afin d'identifier un autre type d'Entité : les Organisations. Par exemple, la chaîne de traitements classique relève que les entités ci-dessous sont des ES et ne permet pas de les distinguer des organisations : *Le projet défendu par Montpellier Agglomération...* et *La France a autorisé un quota de...* Ci-dessous deux exemples de règles utiles pour

4. http ://www.linguastream.org/whitepaper.html
5. lien vers le projet : http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

repérer les Organisations : une entité d'organisation est suivie par un *verbe d'action* ; une entité d'organisation est précédée par certaines prépositions : *avec, par, pour...* Ces règles s'appuient sur des grammaires qui exploitent un contexte local réduit. Dans la section suivante, nous prenons en compte un contexte plus important pour distinguer une ES d'une organisation.

Vers une méthode hybride. Nous proposons d'apprendre un modèle permettant de distinguer une entité de type Organisation et une ES. Pour cela, nous avons étiqueté manuellement un ensemble de phrases en deux classes correspondant aux deux types d'entités. Nous n'avons pas considéré les phrases dites ambiguës, c'est-à-dire présentant une ES et une Organisation. Pour l'apprentissage supervisé, nous avons utilisé la méthode classique SVM (Joachims (1998)). Les descripteurs utilisés sont les mots des phrases qui représentent un "sac de mots". Nous avons complété ces descripteurs en considérant les patrons définis dans les sections précédentes comme des descripteurs à part entière. Pour cela, dans la représentation vectorielle de nos textes, nous avons ajouté des attributs de type *booléen* signifiant qu'une phrase peut contenir un motif de type <*ConceptOrg*, *Entité*> (motif propre à une organisation) ou <*ConceptSpa*, *Entité*> (motif propre à une Entité Spatiale). *ConceptOrg* représente les prépositions typiques précédant une Organisation (*avec, par*, etc). *ConceptSpa* se décline en trois sous-concepts précédant, en général, une Entité Spatiale : *Préposition spatiale* : *en, sur*, etc. ; *Indicateur de relation* : *sud, vers*, etc. ; *Indicateur spatial* : *ville, région*, etc. Cette représentation a deux avantages : 1) Elle donne plus de poids à certains mots propres au domaine de la Recherche d'Information Géographique (prépositions spatiales et d'organisation, indicateurs spatiaux et de relation). Dans un contexte plus général, de tels mots peu porteurs de sens sont souvent moins pris en compte voire supprimés ; 2) contrairement à l'approche *sac de mots* classique qui ne prend pas en considération l'ordre des mots, les nouveaux descripteurs prennent en compte un ordre partiel et se révèlent déterminants comme le montrent les expérimentations.

4 Experimentations

Le corpus utilisé représente un ensemble d'articles sélectionnés depuis 2006 dans le quotidien *Le Midi Libre* et diffusé dans la région Languedoc-Roussillon. Les articles traitent les questions de réaménagement communautaire de l'étang de Thau ainsi que son développement économique et environnemental. Ce jeu de données est particulier dans la mesure où le journaliste, dans ses procédés de reprise, évite la répétition d'un seul terme pour désigner les sujets de l'article. Par ailleurs, il diversifie son vocabulaire et utilise des formulations différentes pour désigner une entité déjà citée. Ce procédé de reprise est très fréquent par exemple dans l'utilisation d'indicateurs spatiaux comme : *région, ville, département...* pour désigner une EN de lieu comme : *Languedoc-Roussillon, Montpellier, Hérault....*

Évaluation de la nouvelle chaîne de traitements. Nous avons évalué manuellement les différentes chaînes de traitements (patrons de base vs. patrons de Text2Geo) à partir d'un sous-ensemble du corpus constitué de 20 articles journalistiques (8141 mots). Nous avons mesuré les résultats retournés en terme de précision (proportion d'entités correctes retournées par le système), rappel (proportion d'entités pertinentes retournées au regard de toutes les entités pertinentes attendues) et F-mesure (moyenne harmonique du rappel et de la précision).

Les résultats de la Table 1 montrent que les patrons de base fondés sur le modèle pivot retournent des résultats corrects (précision acceptable) mais avec un silence important (rappel très faible) pour les ESR. Du fait de l'utilisation de patrons de nature différente (cf. section 3), le résultat est symétrique concernant les ESA (rappel acceptable mais précision très faible). L'enrichissement des patrons initiaux avec notre approche Text2Geo améliore significativement les résultats de la précision et du rappel. Le taux de F-mesure est plus que doublé. De plus, nos patrons permettent d'identifier les organisations. Notons que les Organisations identifiées par notre système sont de bonne qualité (précision à 92%).

Patrons de base			Patrons de Text2Geo			
	ESA	ESR		ESA	ESR	ORG
Précision	20%	**48%**	Précision	53%	84%	**92%**
Rappel	**63%**	27%	Rappel	**94%**	66%	35%
F-mesure	30%	**34%**	F-mesure	67%	**74%**	50%

TAB. 1 – *Evaluation des patrons de Text2Geo.*

Évaluation de la méthode hybride. L'ensemble d'apprentissage est composé de 138 phrases contenant des ENs de type Lieu et 134 phrases contenant des ENs de type Organisation. Chaque phrase lemmatisée est représentée par un vecteur binaire. Nous avons appliqué l'algorithme de classification issus de Weka [6] qui retourne les meilleurs résultats : SVM. Les évaluations données dans la suite utilisent le principe de validation croisée. La Table 2 montre la qualité des résultats selon les différentes approches en terme de matrice de confusion par rapport aux deux classes (ES et Organisations). Le taux d'exactitude TE correspond à la proportion d'exemples bien classés. L'approche hybride améliore de façon significative les résultats, en terme d'exactitude, avec l'utilisation des descripteurs spécifiques aux ES (ConceptSpa) particulièrement adaptés dans le cadre du modèle hybride Text2Geo.

Descripteurs de base			**Descripteurs avec ConceptOrg**			**Descripteurs avec ConceptSpa**			**Les deux types de descripteurs**		
	ES	Orga		ES	Orga		ES	Orga		ES	Orga
ES	103	35	ES	108	30	ES	112	26	ES	113	25
Orga	98	40	Orga	47	87	Orga	19	115	Orga	19	115
TE 70%			*TE 71,69%*			*TE 83,45%*			*TE 83,82%*		

TAB. 2 – *Classification des phrases.*

5 Conclusion et perspectives

Dans le cadre du projet Senterritoire, nous avons proposé une méthode hybride qui permet l'extraction d'informations spatiales et la recherche d'informations. Ces approches exploitent

6. http ://www.cs.waikato.ac.nz/ml/weka/

un contexte et permettent la désambiguïsation d'entités nommées (EN) de type Lieu et Organisation. Sur la base des travaux de Lesbegueries (2007), notre première contribution est un ensemble de patrons morpho-syntaxiques, intégrant notre chaîne de traitement linguistique Text2Geo. Cette dernière permet d'affiner l'identification d'EN spatiales. Nous nous appuyons ensuite sur une méthode d'apprentissage supervisé classique dans le domaine de la fouille de données pour typer les EN (de type Lieu et Organisation) et éliminer les possibles ambiguïtés.

Dans les perspectives à ce travail, nous envisageons d'appliquer le processus d'apprentissage supervisé à trois classes (Organisation, ESR, ESA) afin d'afiner le procesus l'étiquetage.

Remerciements : Les auteurs remercient Pierre Maurel (IRSTEA, UMR TETIS), expert du corpus utilisé pour nos expérimentations, ainsi que la Maison des Sciences de l'Homme de Montpellier (MSH-M) pour son soutien.

Références

Abolhassani, M., N. Fuhr, et N. Gövert (2003). Information extraction and automatic markup for xml documents. In *Intelligent Search on XML Data*, pp. 159–178.

Agirre, E., O. Ansa, E. H. Hovy, et D. Martínez (2000). Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*.

Di Méo, G. (1998). *Extrait de Géographie sociale et territoire*. Nathan.

Joachims, T. (1998). Text categorization with suport vector machines : Learning with many relevant features. In C. Nedellec et C. Rouveirol (Eds.), *ECML*, Volume 1398 of *Lecture Notes in Computer Science*, pp. 137–142. Springer.

Lesbegueries, J. (2007). *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé*. Ph. D. thesis, Université de Pau et des Pays de l'Adour.

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

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

Nouvel, D., J.-Y. Antoine, N. Friburger, et A. Soulet (2011). Recognizing named entities using automatically extracted transduction rules. In *(LTC'2011)*.

Velardi, P., P. Fabriani, et M. Missikoff (2001). Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pp. 270–284.

Summary

We focus on the evaluation of methods for extracting spatial information from texts. After describing the study and analyzing all possibles forms of textual description of space and all the conventions adopted for the manual annotation of named entities of Location and Organization types, we propose to develop a hybrid method. This method combines information extraction approach based on patterns, with a supervised classification approach, to explore the context. We then discuss the different results obtained on the dataset of the Thau lagoon.