Text2Geo : des données textuelles aux informations géospatiales

Résumé. Dans cet article, nous nous intéressons aux méthodes d'extraction d'informations spatiales dans des documents textuels. Cette information est parfois polysémique et peut être associée, selon le contexte, au concept de lieu ou d'organisation. Après avoir décrit l'étude et l'analyse des formes textuelles de description de l'espace ainsi que les conventions adoptées pour l'annotation d'entités nommées de type lieu et organisation, nous présentons la méthode hybride Text2Geo. Cette méthode combine une approche d'extraction d'informations, basée sur des patrons avec une approche de classification supervisée permettant d'explorer le contexte associé. Nous discutons ensuite des résultats expérimentaux obtenus sur le jeu de données du Bassin de Thau.

1 Introduction

Au-delà de sa stricte définition d'entité administrative et politique, le territoire, selon Guy Di Méo Meo (1998) 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"*. 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. La recherche d'informations associées incluant les groupes d'acteurs porteurs d'un même discours territorial, représente un verrou scientifique multidisciplinaire. Deux types d'acteurs peuvent être pris en compte (personnes et organisations) et nous nous limitons dans ces travaux aux acteurs de type "Organisation".
Le travail présenté s'inscrit dans le projet XXXX [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. La problématique associée se

1. Nom du projet anonyme dans le cadre de cette soumission à EGC'2013

décompose en deux phases : accéder à l'information spatiale contenue dans un corpus textuel puis l'interpréter dans une phase de détection du sentiment associé. Nos contributions portent, dans cette publication, sur l'accès à l'information spatiale et plus précisément proposent (1) d'affiner et d'enrichir les patrons d'extraction d'information existant 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 du modèle sur lesquelles se fondent notre étude ainsi que les travaux du domaine. Ensuite, en section 3.1, nous décrivons notre proposition par l'intermédiaire de la chaîne de traitements adoptée et de la méthode Text2Geo, méthode hybride permettant de faciliter le processus de désambiguïsation des entités spatiales découvertes. En section 4, les expérimentations réalisées sur le jeu de données du bassin de Thau sont présentées. Enfin, la section 5 dresse le bilan de la première phase du projet et liste les perspectives associées.

2 Etat de l'art

2.1 De l'information géographique à l'information spatiale textuelle

L'expression de l'**information géographique** est abordée en abondance dans différents domaines et les travaux en géomatique, appliqués au contenu des documents textuels. Ces travaux s'appuient sur les avancées des linguistes. Parmi ces derniers, Vandeloise (1986) et Borillo (1998) centrent leur approche sur l'identification de relations spatiales exprimées par des marqueurs (prépositions, verbes, etc.) mettant en évidence un lien entre une entité à localiser et une entité de référence. Nous retenons pour nos travaux la définition de A. Borillo Borillo (1998) indiquant que, lorsqu'il est invoqué dans un texte, un lieu est une portion de l'espace matériel dans lequel nous nous situons et, dans le cas précis des lieux géographiques, il peut être rattaché à une catégorie (montagnes, lacs, etc.). À partir des travaux de Vandeloise (1986) qui définit le concept cible/site, Borillo (1998) met en avant le fait qu'une référence à un lieu correspond à une relation entre une entité concrète (l'objet cible décrit dans le texte) et une localisation (localisation de l'objet cible). L'ensemble de ces travaux montre l'importance de la composante spatiale pour identifier et définir une information géographique. À partir de ces travaux, Usery (2000) définit l'information géographique comme une molécule formée d'une composante spatiale, d'une composante temporelle et d'une composante thématique (ou phénomène). Quelque soit le type de document traité, l'information géographique apparaît donc sous forme d'Entités Géographiques (EGs), chacune étant composée d'une entité thématique ou phénomène (EP), d'une entité spatiale (ES) et d'une entité temporelle (ET). Généralement l'ES correspond à la portion de l'espace géographique dont traite l'EG. Dans (Lesbegueries (2007)), un modèle cognitif dit "Pivot" est défini (cf. figure 1) qui met en avant l'ES telle qu'elle peut être présente dans la molécule géographique. Dans cet article, nous nous focalisons sur les ES uniquement.

2.2 Le modèle Pivot

Dans ce modèle, l'ES exprimée dans un texte, est constituée d'au moins une Entité Nommée (notée ENs) et d'un nombre variable d'indicateurs spatiaux, précisant sa localisation. Très souvent, on évoque une ES en se référant à une EN de lieu dont on connait la position. Voici quelques exemples d'ESs : "Séville", "au nord de Madrid", "dans les zones montagneuses". La relation cible/site est clairement apparente dans les deux premières mais est plus complexe dans la dernière car trop dépendante du contexte pour être située. L'ES est décrite dans le modèle Pivot selon le format UML illustré figure 1 :

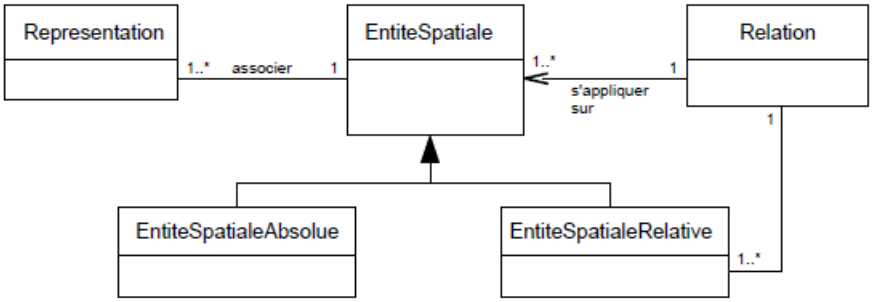

Fig. 1: L'entité Spatiale dans le modèle Pivot tiré de Lesbegueries (2007)

- Une **entité spatiale absolue (ESA)** est une référence directe à un espace géo-localisable, une EN de lieu par exemple. Elle a la forme suivante : $< (indicateur spatial)^*$, *Entité nommée de lieu*$>$. Par exemple, *Séville* ou *la ville de Séville*. C'est une primitive spatiale du modèle.
- Une **entité spatiale relative (ESR)** est définie à l'aide d'au moins une autre ES et d'indicateurs spatiaux d'ordre topologique. Par exemple, *près de Madrid, au sud de Madrid, à une heure de marche de la ville de Madrid, entre Séville et Madrid*. Ces indicateurs spatiaux sont des relations et les deux formes possibles d'une ESR sont : $< (relation\ spatiale)^{1..*}, ESA >$ ou $< (relation\ spatiale)^{1..*}, ESR >$.

Cinq types de relations spatiales sont identifiées dans Lesbegueries (2007) : l'**orientation** *(au sud de)*, la **distance** *(à 1 heure de marche de, à 20 km de)*, l' **adjacence** *(près de, loin de, la périphérie de)*, l'**inclusion** *(le quartier de, la frontière entre, le sommet de)* et la **figure géométrique** qui définit l'union ou l'intersection liant au moins deux ES (*entre A et B, le triangle A, B, C, à l'intersection de A et B, la frontière A-B, etc.*).

Le modèle Pivot permet d'interpréter la plupart des informations exprimées en langage naturel dans les documents textuels. Dans la section suivante, nous présentons les techniques existantes pour extraire l'information spatiale dans des documents texte.

2.3 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'évalua-

tion américaines MUC (Message Understanding Conferences), organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les ENs dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Comme le précisent Daille et al. (2000), ces classes peuvent être enrichies. Par exemple, Paik et al. (1996) définissent de nouvelles classes comme Document (logiciels, matériels, machines) et Scientifique (maladie, médicaments, etc).

De nombreuses méthodes permettent de reconnaitre les ENs en général et les ES en particulier Nadeau et Sekine (2007). Parmi les méthodes d'extraction d'informations s'appuyant sur des textes, les approches statistiques consistent 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 posent des problèmes car elles ne permettent pas toujours de qualifier des termes comme étant des ENs, notamment les ENs de type Lieu ou Organisation comme nous le souhaitons dans cet article.

Des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles (appelées règles de transduction) afin de repérer les ENs (Nouvel et Soulet (2011)). Ces règles utilisent des informations syntaxiques propres aux phrases (Nouvel et Soulet (2011)). Des approches récentes s'appuient sur le Web pour établir des liens entre des entités et leur type (ou catégorie). Par exemple, l'approche de Bonnefoy et al. (2011) repose sur le principe que les distributions de probabilités d'apparition des mots dans les pages associées à une entité donnée sont proches des distributions relatives aux types. Globalement, les relations peuvent être identifiées par des calculs de similarité entre leurs contextes syntaxiques (Grefenstette (1994)), par prédiction à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko (2007)), par des techniques de fouille de textes (Grčar et al. (2009)) ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage (Giuliano et al. (2006)). Ces méthodes sont efficaces, mais elles n'identifient pas toujours la sémantique de la relation.

Pour la reconnaissance des classes d'ENs, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé. Ces méthodes d'apprentissage comme les SVM (Joachims (1998)) sont souvent utilisées dans le challenge Conference on Natural Language Learning (CoNLL). Les algorithmes exploitent divers descripteurs ainsi que des données expertisées/étiquetées. Les types de descripteurs utilisés sont par exemple les positions des candidats, les étiquettes grammaticales, les informations lexicales (par exemple, majuscules/minuscules), les affixes, l'ensemble des mots dans une fenêtre autour du candidat, etc. Carreras et al. (2003). Dans l'approche proposée 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

3.1 Text2Geo et l'ajout de nouvelles règles d'extraction

3.1.1 L'extraction des entités spatiales avec Text2Geo

L'approche Text2Geo que nous proposons suit une chaîne de traitements qui est assez classique dans le domaine de la recherche d'information géographique (Abolhassani et al.

(2003)) :

1. la **lemmatisation** pour segmenter les mots et identifier leur lemme ;
2. l'**analyse lexicale et morphologique** pour la reconnaissance des mots. Cette étape consiste pour chaque mot à identifier la catégorie grammaticale (nom, adjectif, etc.) ainsi que les paramètres de flexion (nombre, temps, etc.) ;
3. l'**analyse syntaxique**, basée sur des grammaires, afin de trouver les relations entre les mots. Cette étape permet d'identifier le rôle des termes ou des syntagmes dans la phrase ;
4. l'**analyse sémantique** pour réaliser une interprétation plus spécifique sur les syntagmes retenus. L'objectif est ici d'identifier le sens potentiel véhiculé par un mot ou un groupe de mots.

Pour appliquer cette méthode, nous proposons d'utiliser puis étendre la chaîne de traitements de Linguastream (Bilhaut (2006)), en partie illustrée sur la figure 2. Cette chaîne se compose de sept modules Linguastream :

- cinq modules de traitement (Text to XML, Tokeniser, Tree Tagger, Token Marker, DCG Marker) ;
- deux modules de visualisation (Web Viewer, Open With) ;
- trois fichiers en entrée.

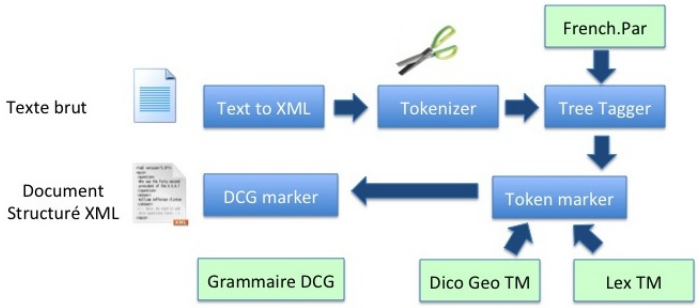

Fig. 2: la nouvelle chaine de traitement basée sur Linguastream

Cette chaîne de traitement utilise à différents niveaux des méthodes et outils de Traitement Automatique du Langage Naturel tel que l'étiqueteur grammatical *Tree-Tagger*[2].

Dans nos travaux, nous nous sommes focalisés sur une extension du module DCGMarker. Ce module s'appuie sur des grammaires DCG (implémenté à l'aide du langage Prolog dans les fichiers .pro). Ces grammaires permettent de s'appuyer sur les mécanismes d'inférence et d'unification de Prolog à l'aide de règles simples. Dans nos travaux, le module DCGMarker nous permet de générer des instances des nouveaux patrons que nous avons définis pour l'extraction des entités spatiales et des entités d'organisation. Cette extraction se fait selon deux étapes s'appuyant sur le modèle pivot :

- La première extrait les ESA qui constituent les types primitifs de notre processus d'extraction. Ces types primitifs sont soit des entités nommées de lieu (Montpellier,

2. lien vers le projet : http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

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 par exemple :
$ES \Rightarrow ESA$,
$ESA \Rightarrow Indicateurderelation, ESA$,
$ESA \Rightarrow Indicateurspatial, ESA$.
$ESA \Rightarrow NomToponymique$.
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*. La figure 3 présente un extrait de la sortie intérmédiaire au format XML résultante du marquage de l'expression "la ville de Madrid" via notre chaine de traitements.

```
<lss:sem id="120" type="token">
    <lss:text>ville</lss:text>
    <lss:value>
        <tag>nom</tag><stag>com</stag>
        <lemma>ville</lemma>
        <egn>non</egn>
        <val>non</val>
        <intro><type>geo</type><stype>commune</stype>
        </intro>
    </lss:value>
</lss:sem>
<lss:sem id="121" type="token">
    <lss:text>de</lss:text>
    <lss:value>
        <tag>det</tag><stag/>
        <lemma>de</lemma>
        <egn>non</egn>
        <val>non</val>
        <intro><type>non</type><stype>non</stype>
        </intro>
    </lss:value>
</lss:sem>
<lss:sem id="122" type="token">
    <lss:text>Madrid</lss:text>
    <lss:value>
        <tag>nom</tag><stag>pro</stag>
        <lemma>Madrid</lemma>
        <egn>oui</egn>
        <val>oui</val>
        <intro><type>non</type><stype>non</stype>
        </intro>
    </lss:value>
</lss:sem>
```

Fig. 3: Un indicateur spatial "Commune" et une entité nommée de lieu "Madrid" tels qu'ils sont marqués par notre chaîne d'extraction

– La deuxième extrait les instances les plus complexes qui sont les entités spatiales relatives, celles composées déjà d'une ESA et précédées d'une relation d'ordre topologique suivant des règles du type :
$Relation \Rightarrow Adjacence \,|Orientation\,|Inclusion\,|Distance\,|Forme\ géometrique$,

$Adjacence \Rightarrow$ "$prés$" |"$lapériphérie$" |etc,
$Orientation \Rightarrow$ "$au\ sud$" |"$au\ nord$" |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*.

3.1.2 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, montagnes, etc), les *constructions humaines* (buildings, installations, etc.), les *axes de circulations* (routes, autoroutes, etc.), les *adresses* (rue, code postal, etc.). Le sous-type *axe de circulation* est identifiable par le modèle Pivot et ne représente pas un cas de polysémie. De même, les lieux géographiques naturels, même s'ils sont introduits par des modificateurs de sens (introducteurs spatiaux : rivière, lac, montagne, etc.), ne peuvent être polysémiques. Les sous-types *adresse* et *construction humaine* n'ont pas été traités dans cette étude.

Dans nos travaux, nous nous nous sommes uniquement concentrés sur les sous-types de lieux qui peuvent se révéler polysémiques. Dans ce cadre, nous avons ajouté des règles (patrons) permettant d'amélioration 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 pouvons identifier deux entités spatiales.

D'autres types de règles ont été ajoutées qui ont permis non seulement d'augmenter le nombre d'ES extraites mais aussi d'améliorer la qualité des ES extraites. Ce point sera discuté dans la section 4 de cet article.

3.1.3 Définition de nouveaux patrons pour l'identification des Organisations

Dans notre chaîne de traitement, nous avons également ajouté des règles afin d'identifier un autre type d'Entité : les Organisations. Ainsi, notre approche et les règles associées permettent de distinguer une ES d'une Organisation.

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...*
- *La France a autorisé un quota de...*

L'ajout de règles spécifiques permet de pallier ce problème. À titre d'exemple, les règles ci-dessous sont utiles pour repérer avec précision 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, de la part de...*

Les nouvelles règles présentées dans cette section s'appuient sur des grammaires qui exploitent un contexte local assez réduit. Il semble intéressant de prendre en compte un contexte plus important permettant de distinguer une ES d'une organisation. Ce type de travaux liés à la désambiguisation est détaillé dans la section 3.2. L'approche que nous proposons est hybride

car elles combine une méthodologie liée à la recherche d'information tout en exploitant les patrons de Text2Geo.

3.2 Vers une méthode hybride

Dans la suite de notre travail, nous proposons d'apprendre un modèle qui permet de distinguer une entité de type Organisation et une ES. Pour cela, nous avons étiqueté manuellement un ensemble de phrases en deux classes relativement aux deux types d'entités. Un ensemble constitué de plus de 250 exemples est utilisé dans nos expérimentations. Notons que durant cette phase d'étiquetage, nous n'avons pas considéré les phrases dites ambiguës, c'est-à-dire présentant deux différents types d'entités. L'apprentissage repose sur deux méthodes classiques dans le domaine de la fouille de données qui sont les SVM et les Naives Bayes (Joachims (1998)). Dans le cadre du processus d'apprentissage supervisé, les descripteurs utilisés sont les mots des phrases qui représentent un "sac de mots". Les résultats de prédiction avec cette approche sont discutés en section 4.

L'originalité de l'approche hybride que nous proposons va consister à considérer les patrons définis dans les sections précédentes comme descripteurs à part entière dans notre modèle d'apprentissage. Cette intégration va notamment permettre de mettre en exergue les types de descripteurs (c'est-à-dire les patrons) les plus pertinents à utiliser afin d'identifier une organisation et/ou une entité spatiale.

De manière concrète, 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, nord, vers*, etc. ;
- *Indicateur spatial* : *ville, région*, etc.

Dans nos expérimentations, chaque type de descripteur sera évalué indépendamment afin de mesurer l'influence de chaque proposition.

Notons que notre représentation a deux avantages. Elle permet dans un premier temps de donner plus de poids à certains mots propres au domaine de la Recherche d'Informations Géographiques (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. D'autre part, rappelons que l'approche *sac de mots* classique ne prend pas en considération l'ordre des mots. Le fait de prendre en compte dans notre modèle d'apprentissage un ordre partiel pour certains descripteurs linguistiques permet d'ajouter des informations qui se révèlent déterminantes comme le montrent les expérimentations décrites dans la section suivante.

4 Experimentations

4.1 Description du corpus

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 : *Langue-docRoussillon, Montpellier, Hérault....*

4.2 Évaluation de la nouvelle chaîne de traitements

Dans un premier temps, nous avons évalué manuellement les différentes chaînes de traitements (partons de base vs. patrons de Text2Geo) à partir d'un sous-ensemble du corpus constitué de 20 articles journalistiques (8141 mots).

Dans ce cadre, nous avons mesuré les résultats retournés en terme de précision, rappel et F-mesure. Dans notre contexte, la précision calcule la proportion d'entités correctes retournées par le système. Le rappel détermine quelle est la proportion d'entités pertinentes retournées au regard de toutes les entités pertinentes attendues. La F-mesure combine la précision et le rappel selon la moyenne harmonique.

Les résultats donnés dans 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 3.1), 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%). L'ajout de règles dans nos futurs travaux devrait permettre d'améliorer le rappel.

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

4.3 Évaluation de la méthode hybride

La méthode hybride que nous proposons repose sur une méthode d'apprentissage supervisé. Dans nos expérimentations, notre 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 alors représentée par un vecteur binaire.

Nous avons appliqué deux algorithmes de classification issus de Weka [3] qui retournent les meilleurs résultats : SVM et Naive Bayes. Les évaluations données dans la suite utilisent le principe de validation croisée.

La Table 2 montre la qualité des résultats selon les deux approches en terme de matrice de confusion par rapport aux deux classes (Entités Spatiales et Organisations). Le taux d'exactitude (*Accuracy*) correspondant à la proportion d'exemples bien classés. Les résultats sont du même ordre pour les deux algorithmes SVM (70%) et Naive Bayes (69%).

SVM			**Naive Bayes**		
	ES	Orga		ES	Orga
ES	103	35	ES	98	40
Orga	98	40	Orga	44	90

Tab. 2: Classification des phrases sans utiliser les descripteurs de Text2Geo

L'approche hybride améliore globalement les résultats en terme d'exactitude (cf. Table 3). Cette amélioration se révèle significative avec l'utilisation des descripteurs spécifiques aux entités spatiales (ConceptSpa) particulièrement adaptés dans le cadre du modèle hybride Text2Geo.

Descripteurs avec ConceptOrg			**Descripteurs avec ConceptSpa**			**Les deux types de descripteurs**		
	ES	Orga		ES	Orga		ES	Orga
ES	108	30	ES	112	26	ES	113	25
Orga	47	87	Orga	19	115	Orga	19	115
Tx exactitude	*71,69%*		*Tx exactitude*	*83,45%*		*Tx exactitude*	*83,82%*	

Tab. 3: Classification des phrases avec contraintes

5 Conclusion et perspectives

Dans le cadre du projet XXXX, nous avons proposé dans cet article une méthode hybride qui permet l'extraction d'informations spatiales et la recherche d'informations. Ces approches exploitent un contexte et permettent la désambiguïsation d'entité nommées de type Lieu et Organisation.

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

Sur la base du modèle pivot définit par 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'entités nommées spatiales. Nous nous appuyons ensuite sur deux méthodes d'apprentissage supervisé classiques dans le domaine de la fouille de données, les SVM et les Naives Bayes, pour typer les entités nommées (de type Lieu et Organisation) et éliminer les possibles ambiguïtés. Les expérimentations, menées sur un ensemble d'articles sélectionnés depuis 2006 dans un quotidien ont permis d'évaluer notre approche de repérage des entités spatiales, puis le prototype final d'identification des organisations. L'enrichissement des patrons initiaux dans Text2Geo améliore significativement les résultats en terme de précision et rappel. De plus, notre approche à base de patrons et/ou d'apprentissage supervisé permet d'identifier spécifiquement les organisations.

Dans les perspectives à ce travail, nous envisageons d'appliquer le processus d'apprentissage supervisé à trois classes : Organisation, ESR, ESA. Ainsi, nous pourrons vérifier si l'utilisation d'un contexte local plus important (la phrase) permet de distinguer deux types d'ES fines (ESR et ESA).

L'approche hybride que nous proposons consiste à prendre en compte de nouveaux descripteurs, fondés sur des patrons, dans le processus d'apprentissage. Nous pourrions envisager d'autres types d'hybridation qui reposent sur l'utilisation de patrons comme pre- et/ou post-filtrage ou les combiner via un système de votes.

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*.

Bilhaut, F. (2006). *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*. Ph. D. thesis, Université de Caen.

Bonnefoy, L., P. Bellot, et M. Benoit (2011). 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*, pp. 191–206.

Borillo, A. (1998). *L'espace et son expression en français*. L'essentiel. Ophrys.

Carreras, X., , X. Carreras, L. S. M. Arquez, et L. S. P. O (2003). A simple named entity extractor using adaboost. In *In Proceedings of CoNLL-2003*, pp. 152–155.

Daille, B., N. Fourour, et E. Morin (2000). Catégorisation des noms propres : une étude en corpus. Volume 25, Chapter Cahiers de Grammaire, pp. 115–129.

Giuliano, C., A. Lavelli, et L. Romano (2006). 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.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA : Kluwer Academic Publishers.

Grčar, M., E. Klien, et B. Novak (2009). Using Term-Matching Algorithms for the Annotation of Geo-services. In B. Berendt, D. Mladenič, M. Gemmis, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svátek, et F. Železný (Eds.), *Knowledge Discovery Enhanced with Semantic and Social Information*, Volume 220, Chapter 8, pp. 127–143. Berlin, Heidelberg : Springer Berlin Heidelberg.

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.

Meo, G. D. (1998). *Extrait de Géographie sociale et territoire*. Nathan.

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

Nouvel, D. et A. Soulet (2011). Annotation d'Entités Nommées par Extraction de Règles de Transduction. In *EGC'2011*, Brest, France, pp. 119.

Paik, W., E. D. Liddy, E. Yu, et M. Mckenna (1996). Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition*, pp. 61–73. MIT Press.

Usery, E. L. (2000). Multidimensional Representation of Geographic Features. pp. 240–247.

Vandeloise, C. (1986). *L'espace en français*. Seuil.

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

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

Summary

In this article, we focus on the evaluation of methods for extracting spatial information in text documents. After describing the study and analysis of all forms of textual description of space and all the conventions adopted for the manual annotation of named entities of type location and organization, 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.