OPILAND : identification de la perception des territoires par la fouille de texte

Eric Kergosien, Bernard Laval, Mathieu Roche, Maguelonne Teisseire

OPILAND : identification de la perception des territoires par la fouille de texte

Éric Kergosien[*,**], Bernard Laval[**],
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

Résumé. De nombreux travaux ont été réalisés en extraction d'informations et plus particulièrement en fouille de données d'opinions dans des contextes spécifiques tels que les critiques de films, les évaluations de produits commerciaux, les discours électoraux... Dans le cadre du projet SENTERRITOIRE, nous nous posons la question de l'adéquation de ces méthodes pour des documents associés à l'aménagement des territoires. Ces documents renferment différents types d'informations se rapportant à des acteurs, des opinions, des informations géographiques, et tout autre aspect lié plus généralement à la notion de territoire. Cependant, il est extrêmement difficile d'identifier puis de lier les opinions à ces informations. Après avoir souligné les limites des propositions actuelles et les verrous soulevés par les données textuelles associées, nous proposons la méthode semi-automatique nommée OPILAND (OPinion mIning from LAND-use planning documents) combinant une chaîne de Traitement Automatique du Langage Naturel et des techniques de Fouilles de Textes pour (1) détecter les entités nommées de type lieu et organisation, (2) construire un vocabulaire d'opinions relatif au domaine d'application, et (3) identifier les opinions relatives aux entités nommées traitées. Les expérimentations sont menées sur des données du bassin de Thau (France), puis appliquées sur trois corpus relatifs à d'autres domaines afin de mettre en avant la généricité de notre approche.

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 (Derungs et Purves, 2013) et de politique publique territoriale. La notion de territoire, et plus spécifiquement d'aménagement du territoire, fait référence à différents concepts

tels que les informations spatiales et temporelles, les acteurs, les opinions, l'histoire, la politique, etc. Dans cet article, nous nous focalisons sur la détection d'opinions liées aux entités nommées EN de type lieu, que l'on nomme Entité Spatiale (ES) et à celles de type organisation (EO), et nous traitons des corpus décrivant des territoires politiques et administratifs à l'échelle des collectivités locales, voire régionales.

L'extraction d'EN ainsi que la fouille d'opinions ont été intensivement appliquées dans divers domaines tels que l'analyse automatique des critiques de films, articles politiques, tweets, etc. Les travaux présentés proposent généralement une approche statistique ou des techniques provenant du Traitement Automatique du Langage Naturel (TALN). Dans ce cadre, un lexique d'opinions polarisées est souvent utilisé. Cependant, ce type de ressources manque cruellement en langue française et celles qui existent sont souvent des traductions des lexiques anglais qui restent trop généraux pour être adaptés à un domaine spécifique. Dans le contexte de l'aménagement du territoire, même si les informations publiées sur le web (blogs, forums, etc.) et dans les médias (journaux, etc.) expriment des sentiments, les approches traditionnelles de fouille de textes ne parviennent pas à extraire des opinions de façon exhaustive et précise de ces domaines spécialisés.

Nous proposons d'aborder cette question en définissant une approche originale, appelée OPILAND (OPinion mIning from LAND-use planning documents, ce qui signifie en français « Fouille d'opinions pour les documents liés à l'aménagement du territoire »), définie pour identifier semi-automatiquement des opinions relatives aux EN dans des contextes spécialisés. Ce travail est réalisé dans le cadre du projet SENTERRITOIRE [1] qui vise à développer un environnement de prise de décision pour les documents d'aménagement du territoire. L'approche OPILAND se décompose en trois étapes (*cf.* Figure 1). Dans la première étape, OPILAND se concentre sur l'extraction automatique des ES et des EO. Ces entités doivent être enrichies par des informations territoriales liées aux opinions que les acteurs ont sur leur propre territoire mais cette tâche reste difficile. Pour ce faire, nous définissons dans une deuxième étape un vocabulaire spécialisé d'opinions sur lequel nous nous appuyons ensuite pour calculer un score de polarité pour chaque document du corpus. Cette évaluation doit nous permettre de valider le vocabulaire défini. Dans la troisième étape, nous proposons un module original capable de produire des vocabulaires d'opinions spécifiques relatifs aux ES et EO.

OPILAND est initialement défini pour un corpus spécifique, mais nous démontrons que notre approche peut être utilisée dans d'autres domaines en menant des expérimentations sur trois autres corpus français, le premier regroupant des échanges lors d'un débat politique, le second relatif aux jeux vidéo, et le dernier à des critiques de films.

Le document est structuré comme suit. Dans la section 2, un aperçu des méthodes d'extraction d'opinions est présenté. Dans la section 3, la méthode OPILAND est détaillée. La section 4 présente les expérimentations menées tout d'abord sur le corpus relatif à l'aménagement du territoire, puis sur les trois autres corpus. Nous concluons et présentons les perspectives de ces travaux en section 5.

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

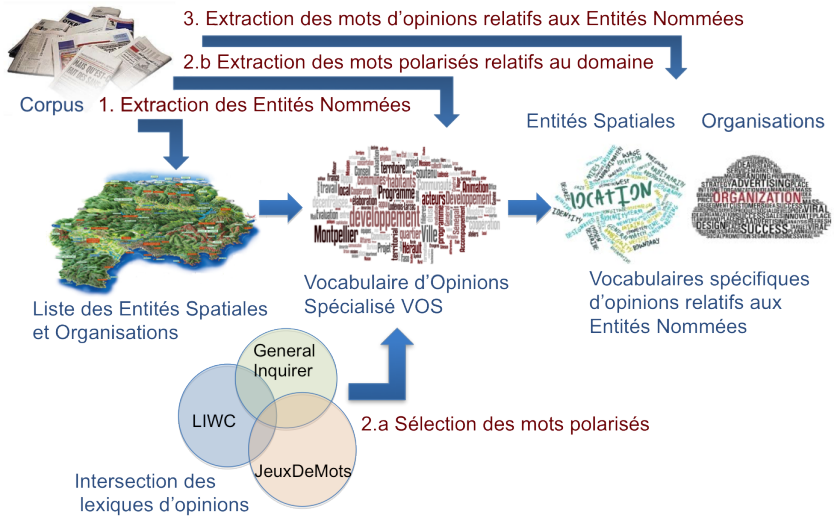

FIG. 1 – *Approche* OPILAND *pour l'extraction des opinions relatives aux Entités Nommées.*

2 État de l'art

Les **Entités Nommées** (*EN*) ont été définies comme des noms de personnes, des lieux et des organisations lors des campagnes d'évaluations américaines appelées MUC (Message Understanding Conferences), qui furent organisées dans les années 90. Dans cet article, nous nous concentrons sur les lieux et les organisations et le premier défi consiste à reconnaître ces types d'EN.

De nombreuses méthodes permettent de reconnaître les EN 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 étudient généralement 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 EN, notamment les EN de type ES ou EO. 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 EN (Maurel et al., 2011). Ces règles utilisent des informations syntaxiques propres aux phrases (Maurel et al., 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 Bellot, 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'EN, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé. Ces méthodes d'apprentissage comme les SVM (Joachims, 1998) ou encore les champs aléatoires conditionnels notés CRF (McCallum, 2003; Zidouni et al., 2009) 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 termes, 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. Nous nous appuyons notamment sur les travaux de (Lesbegueries et al., 2006) pour la définition de patrons linguistiques pour l'extraction d'ES.

Concernant la **fouille de données d'opinion** et le **TALN**, l'analyse de la subjectivité et des opinions exprimées par des personnes dans les textes (journaux, documents techniques, blogs, commentaires des internautes, lettres aux éditeurs, etc.) est connue comme l'analyse des sentiments ou des opinions (Pang et Lee, 2010; Liu, 2012). La reconnaissance de polarité tente de classer les textes selon la positivité ou la négativité des opinions qui y sont exprimées. Deux approches principales peuvent être identifiées : l'une fondée sur le recensement des termes positifs et négatifs dans chaque texte (Turney, 2002), et l'autre sur l'apprentissage automatique à partir de textes annotés (Esuli et Sebastiani, 2006). Les approches hybrides semblent proposer les meilleurs résultats (Kennedy et Inkpen, 2006). Dans toutes ces approches plusieurs descripteurs doivent être utilisés, tels que des mots, des n-grammes (Pak et Paroubek, 2010), les mots modifiés (Joshi et al., 2011)... Ces différents descripteurs peuvent être exploités par des méthodes d'apprentissage automatique en s'appuyant sur des corpus annotés. Beaucoup de corpus sont disponibles dans le cadre de challenges d'analyse de textes tels que TREC (Text Retrieval Conference), ou encore DEFT (Défi Fouille de Textes) pour la communauté française. Cependant, seuls quelques uns sont annotés en prenant en compte les opinions et la polarité. En outre, plusieurs méthodes de classification peuvent être regroupées en systèmes de vote proposés par (Plantié et al., 2008) ou en appliquant des méthodes de renforcement et / ou sacs de mots (Fan et al., 2011), ou encore des approches syntaxiques (Pak, 2012). D'autres approches s'appuient sur des méthodes incrémentales pour l'analyse des sentiments (Wiebe et Riloff, 2011), et . Hormis la classification de textes d'opinions, des travaux se sont concentrés sur la construction automatique de vocabulaires spécialisés liés à l'opinion (Husaini et al., 2012; Duthil et al., 2011). Les approches incrémentales proposées reposent sur des méthodes de fouille du web afin d'apprendre un vocabulaire d'opinion lié à un thème ou sous-thème donné.

Concernant la problématique applicative associée à l'**analyse des ressentis des acteurs** à propos de l'**aménagement des territoires**, le panorama des travaux existants soulignent l'implication de différentes communautés. En effet, depuis une trentaine d'années, le concept de territoire a été largement utilisé et discuté, sous des acceptations diverses, tant, dès l'origine, par les éthologues et écologues, que par les géographes, les sociologues (Barel, 1981), les économistes (Pecqueur, 2007) les politistes (Alliès, 1980), les agronomes (Deffontaines et al., 2001), que les philosophes (Deleuze et Guattari, 1980)... La géographie en particulier a été

particulièrement prolixe, partageant sa production entre une analyse du territoire « sujet politique » et une analyse du territoire « sujet social » (Vanier, 2009) : d'un côté, la géographie sociale a analysé la dimension identitaire du territoire, les rapports d'appartenance et d'ancrage (Buléon et Méo, 2005) ; de l'autre, une géographie plus politique s'est efforcée d'éclairer la dimension de représentation du territoire, à travers une analyse des dispositifs d'action publique (Debarbieux et Vanier, 2002), (Salles, 2009). Dans ce cadre, (Salles, 2009) propose notamment de construire une ontologie de territoire pour l'aide à la décision publique en matière de développement économique territorial.

Pour faire face à ces acceptions multiples, il convient souvent dès lors d'en préciser le sens par l'ajout d'un qualificatif : territoires biophysiques (bassin versant, grand paysage, etc.), territoires politico-administratifs (commune, Etat, Europe, etc.), grands territoires (Grand Paris, Arc Atlantique, etc.), territoires appropriés (Bretagne, Béarn, Pays Basque, Larzac, etc.), territoires mobiles des nomades traditionnels (chameaux, etc.) ou modernes (TGV, avion, etc.), territoires numériques générés par les réseaux de télécommunication, territoires virtuels sur le Web qui peuvent parfois prolonger et étendre dans le cyberespace des représentations des territoires physiques. Cependant, ces différents territoires ne constituent pas des ensembles disjoints, bien au contraire, tellement ils sont désormais étroitement imbriqués. À notre connaissance, il n'existe pas de travaux concernant la fouille de données d'opinions en lien avec l'aménagement du territoire.

Compte tenu de l'état de l'art, différents types d'approches peuvent être appliquées. Dans notre contexte, des approches supervisées ne sont pas adaptées dû au fait que nous avons à disposition des données étiquetée (données analysées par des experts) en nombre très limité. Les approches non supervisées fondées sur l'utilisation de graines de mots polarisés à enrichir sont souvent utilisés dans les travaux liés à l'analyse de sentiment (Kozareva et al., 2007; Harb et al., 2008). Généralement l'enrichissement est fondé sur l'exploitation d'informations généralistes renvoyées par des moteurs de recherche (par exemple, le « nombre de hits » retournés par Google, Yahoo, Exalead, etc.). Ces méthodes incrémentales sont difficilement exploitables pour traiter des domaines de spécialité. Cependant l'intégration de ressources génériques d'opinion qui s'appuie sur une démarche originale de combinaison peut s'avérer intéressante. Ce point sera détaillé dans nos travaux.

Dans cet article, l'approche proposée (1) combine une méthode linguistique à base de patrons et des méthodes d'apprentissage supervisé afin d'identifier les ES et EO, (2) définit un vocabulaire d'opinions spécialisé lié au domaine d'application à partir de lexiques généralistes d'opinions, et (3) propose d'identifier des vocabulaires d'opinions spécifiques à chaque type d'EN (ES et EO).

La section suivante décrit l'approche originale OPILAND et les résultats expérimentaux sont discutés dans la section 4.

3 OPILAND : extraction d'opinions relatives aux entités nommées

Dans l'objectif de définir des vocabulaires spécifiques relatifs à des EN dans le cadre de l'aménagement du territoire, nous proposons l'approche OPILAND qui se décompose en trois étapes. Afin d'illustrer ces différentes étapes dans les sections suivantes, nous nous appuierons sur un extrait du corpus relatif à l'aménagement de la région de Sète (voir Figure 2).

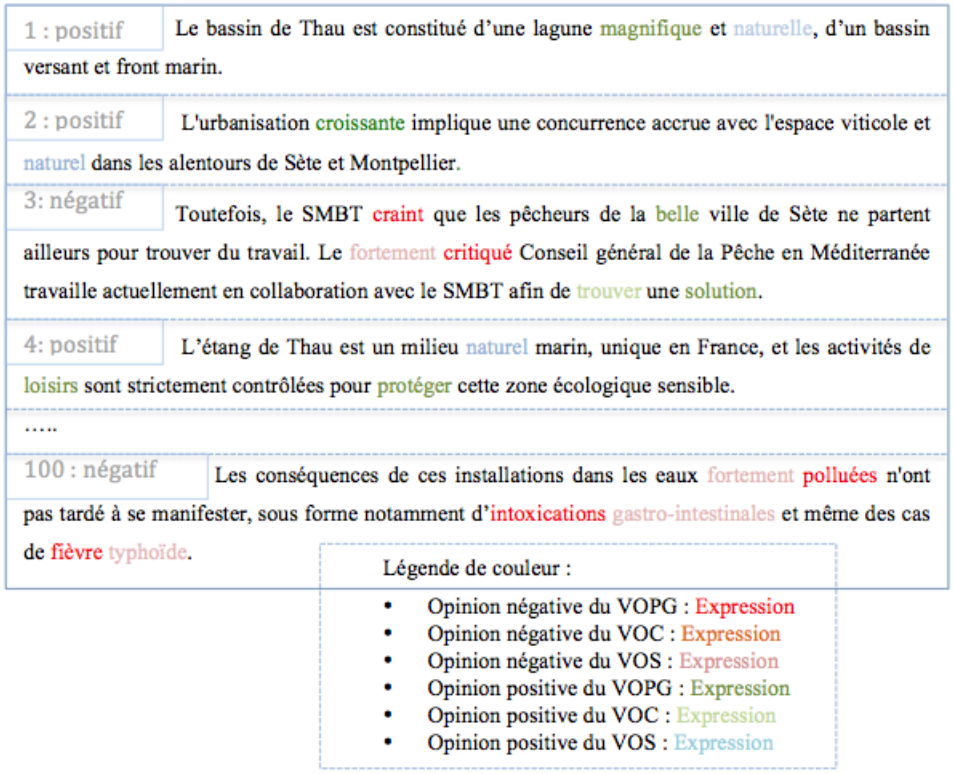

FIG. 2 – *Extraits de documents relatifs à la région de Sète.*

La première étape consiste à extraire semi-automatiquement des ES telles que : « Le bassin de Thau », « étang de Thau » , « ville de Sète », « en France » et « dans les alentours de Sète et Montpellier », ainsi que des EO telles que « le SMBT » et « Conseil Général de la Pêche en Méditerranée ». À cette fin, nous proposons de fusionner une chaîne de TALN et une méthode d'apprentissage supervisé pour améliorer l'extraction d'EN.

La deuxième étape consiste à identifier des opinions liées à un domaine telles que, par exemple, les termes « magnifique », « croissante », « belle », « fièvre », « solution », craint », « critiqué » et « polluées ». L'objectif est de constituer un vocabulaire d'opinions spécialisé lié au domaine, et nous proposons pour cela de fusionner des lexiques d'opinions généralistes produits par la communauté scientifique. Une étape intermédiaire consiste à identifier puis

polariser des nouveaux termes présents dans les documents à partir d'une analyse contextuelle. Une approche TALN classique est mise en place pour identifier les termes et leur polarité est calculée en prenant en compte les opinions provenant des lexiques généralistes.

Dans une troisième étape, notre objectif est d'identifier les opinions relatives aux ES et aux EO. Cette étape consiste à identifier les couples candidats Opinion-EN, puis identifier la taille de fenêtre contenant le plus de couples pertinents, et enfin filtrer les couples correspondants présents dans la fenêtre sélectionnée. D'après les extraits de texte présentés Figure 3, les couples concernés sont « magnifique-bassin de Thau », « belle-ville de Sète » , « SMBT-craint », « critiqué-Conseil général de la Pêche en Méditerranée » et « SMBT-solution ». Parmi les opinions identifiées dans ces extraits, les termes « magnifique » et « belle » sont liés aux ES tandis que « craint », « critiqué » et « solution » sont liés à des EO.

3.1 Extraction des Entités Spatiales et des Organisations

Pour extraire les ES et les EO, nous adoptons un module standard de TALN appliqué en Recherche d'Information Géographique (RIG) (Abolhassani et al., 2003) : (I) la lemmatisation, (ii) l'analyse morpho-lexicale, (iii) l'analyse syntaxique, (iv) l'analyse sémantique. Dans ce processus, la première étape détermine le lemme pour un mot donné. L'analyse morpho-lexicale consiste à identifier pour chaque mot la catégorie grammaticale (nom, adjectif, etc.) ainsi que les paramètres de flexion (nombre, temps, etc.). L'analyse syntaxique a pour objectif d'identifier le rôle des termes ou des syntagmes dans la phrase. Elle s'appuie pour cela sur des grammaires pour trouver les relations entre les mots. Enfin, l'analyse sémantique permet de réaliser une interprétation plus spécifique des 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 détendre la chaîne définie par (Lesbegueries et al., 2006) (patrons de base) en utilisant Linguastream [2].

Plus précisément, (Lesbegueries et al., 2006) propose un modèle cognitif pour définir une ES. Dans ce modèle, une ES est composée d'au moins une EN et un ou plusieurs indicateurs spatiaux spécifiant son emplacement. Une ES peut alors être identifiée de deux façons :

- **une ES absolue (ESA)** est une référence directe à un espace géo-localisable, une EN de lieu par exemple. Elle a la forme suivante : $< (indicateurspatial)^*$, *EN de Lieu*$>$. Par exemple, « la ville de Sète ». C'est une primitive spatiale du modèle.
- **une ES relative (ESR)** est définie à l'aide d'au moins une ESA et d'indicateurs spatiaux d'ordre topologique. Ces indicateurs spatiaux sont des relations et les deux formes possibles d'une ESR sont :

  $< (relationSpatiale)^{1..*}, ESA >$ ou $< (relationSpatiale)^{1..*}, ESR >$.

  Cinq types de relations spatiales sont considérés dans ces travaux : l'orientation, la distance, l'adjacence, l'inclusion et la figure géométrique qui définit l'union ou l'intersection liant au moins deux ES. Un exemple de ce type d'ES est « dans les alentours de Sète ».

Dans notre proposition, nous nous concentrons sur les techniques de fouille de textes pour qualifier les EN en tant que ES ou EO. Tout d'abord, nous ajoutons des patrons linguistiques dans l'étape d'analyse sémantique afin d'améliorer l'identification automatique des ESA et

2. http ://www.linguastream.org/

ESR. Ensuite, nous proposons un nouveau type de patrons pour identifier spécifiquement les EO.

De nouveaux patrons pour l'identification des ESA et ESR. L'annotation d'ES est fondée sur la typologie du domaine couramment utilisée sur les sous-types de lieux. Les EN de type Lieu peuvent être polysémiques : constructions humaines (bâtiments, etc.) et adresses (rues, etc.). Dans ce contexte, des règles (patrons) ont été ajoutées pour améliorer l'identification des ESA et ESR. Nous avons notamment défini des patrons permettant d'identifier la distribution des relations spatiales (par exemple, l'expression de l'extrait de document présenté Figure 2 : « dans les alentours de Sète et Montpellier » $\rightarrow$ « dans les alentours de Sète » + « dans les alentours de Montpellier »). Ici , l'ESR est enrichie par la règle $< (ESR)^{1..*}, SepSpatial, ESA >$ et $SepSpatial$ est défini de la façon suivante : $<$ ", "|";"|"et"|"ou". $>$. D'autres types de règles ont été ajoutés, ce qui améliore la qualité des ES extraites (Voir section 4.2) .

Les ES identifiées à partir des extraits de documents présentés Figure 2 sont listées ci-dessous :

- le bassin de Thau ;
- la ville of Sète ;
- dans les alentours de Montpellier.
- l'étang de Thau ;
- dans les alentours de Sète ;

De nouveaux patrons pour l'identification des Organisations. L'ajout de règles spécifiques facilite l'identification des EO et permet dans le même temps de désambiguïser une partie des ES. Ces règles sont les suivantes : (1) une EO est suivie par un verbe d'action, (2) une EO est précédée par des prépositions telles que : *avec*, *par*, *pour*, *de la part de*, etc. Nous nous appuyons à cette étape sur une liste de verbes d'action [3].

Les EO identifiées dans les extraits de documents présentés Figure 2 sont listées ci-dessous :

- Le SMBT ;
- Le Conseil Général de la Pêche en Méditerranée.

Ces règles prennent en compte un contexte local réduit. Nous faisons l'hypothèse que l'utilisation d'un contexte plus large améliorerait encore la désambiguïsation des ES et EO. Ainsi, dans la section 3.1.1, nous proposons une approche hybride combinant nos patrons et un processus d'apprentissage supervisé.

3.1.1 Approche hybride s'appuyant sur l'apprentissage supervisé

Afin de distinguer les ES des EO, nous proposons une approche fondée sur de l'apprentissage supervisé qui se décompose en 4 étapes :

- **Étape 1 : Construction d'un corpus d'apprentissage.** Elle consiste à acquérir un corpus d'apprentissage composé de phrases. Chaque phrase est étiquetée manuellement comme contenant une ou plusieurs ES ou une ou plusieurs EO. Notez que les phrases ambiguës contenant les deux types d'EN ne sont pas prises en compte dans notre corpus d'apprentissage.
- **Étape 2 : Représentation des données textuelles.** Chaque phrase est décrite par un vecteur signature. Les colonnes représentent les mots (*cf.* descripteurs) et les lignes les

3. http ://www.unamur.be/services/euraxess/cellule-prodoc/documents-mission-accompagnement-prodoc-AL/verbes-actions/view

phrases. Chaque cellule contient une valeur booléenne pour un descripteur donné dans une phrase donnée, indiquant qu'il est présent ou non. En outre, chaque phrase est associée à une classe (c-à-d. ES ou EO) afin de préparer l'étape suivante.

- **Étape 3 : Processus d'apprentissage.** Elle consiste à entraîner un classificateur (c-à-d. apprentissage supervisé) afin d'identifier les phrases contenant une ES et celles contenant une EO.
- **Étape 4 : Prédiction.** Le modèle appris est appliqué sur des données textuelles non étiquetées afin de prédire le type d'entités présent dans les phrases.

Le processus d'apprentissage (*cf.* étape 3) s'appuie sur deux méthodes classiques utilisées en fouille de données qui sont Naive Bayes et SVM. Les descripteurs utilisés sont les mots des phrases (approche dite « sac de mots »). L'originalité de notre approche hybride est de considérer les patrons proposés comme descripteurs dans le modèle d'apprentissage. Ainsi, ces descripteurs booléens sont paramétrés à 1 quand une phrase contient un motif de type de <*ConceptOrg*, *EN*> (formalisation d'une EO) ou <*ConceptSpa*, *EN*> (formalisation d'une ES). *ConceptOrg* représente les prépositions typiques précédant une EO (avec, par, etc.). *ConceptSpa* est divisé en trois sous-concepts (précédant une ES) : les prépositions spatiales (dans, sur, etc.), les indicateurs de relation (sud, etc.), et les indicateurs spatiaux (ville, etc.).

Dans nos expérimentations, chaque type de descripteurs est évalué indépendamment. Cela donne plus de poids aux mots relatifs aux ES (les prépositions, les indicateurs spatiaux et de relation définis dans un dictionnaire). Par exemple, les mots outils sont pondérés plus faiblement, voire éliminés en utilisant une liste de mots vides. De plus, l'approche *sac de mots* standard ne tient pas compte de l'ordre des mots. Or, dans notre modèle d'apprentissage, nous sommes en mesure d'intégrer un ordre partiel des descripteurs linguistiques grâce à nos patrons.

Les EN identifiées dans les extraits de documents présentés Figure 2 sont listées ci-dessous :

- le bassin de Thau : intégrant l'indicateur spatial « le bassin de » ;
- Sète : intégrant l'indicateur spatial « la ville de » ;
- Montpellier et Sète : précédé par l'indicateur de relation « aux alentours de ».

À la fin de cette première étape, nous obtenons une liste contenant des ES et une seconde liste composée des EO. La deuxième étape consiste maintenant à construire un vocabulaire d'opinions spécialisé à partir de lexiques généralistes d'opinions disponibles dans la communauté.

3.2 Extraction d'un vocabulaire d'opinions spécialisé relatif à l'aménagement du territoire

Notre objectif est d'identifier automatiquement les opinions dans des corpus relatifs à un domaine d'application spécifique. Le projet SENTERRITOIRE traite un corpus de 300 articles de pesse du Midi Libre décrivant l'aménagement du territoire de l'étang de Thau. Les méthodes d'extraction d'opinions classiques échouent lorsqu'elles sont appliquées sur ce type de corpus peu volumineux et contenant un vocabulaire complexe. Aussi, les lexiques d'opinions généralistes ne sont pas adaptés à ce domaine spécifique (voir la section 4.3, Tableau 4). Nous avons également testé une approche « sac de mots » en appliquant des algorithmes de classification classiques pour l'apprentissage supervisé. Malheureusement, en raison du petit volume

de données traitées et de la diversité des sujets abordés, les résultats sont restés peu satisfaisants (entre 50% et 55% de documents bien classés, voir la section 4.4.3). Nous avons ainsi défini une nouvelle approche plus générique que nous détaillons dans les paragraphes suivants.

L'approche non supervisée proposée repose sur l'identification d'un vocabulaire spécialisé afin d'évaluer la polarité d'un document. Cette polarité peut être positive ou négative. Nous n'avons pas besoin d'un corpus annoté contrairement aux méthodes classiques de fouille d'opinions (Torres-Moreno et al., 2009). Toutefois, dans un objectif d'évaluation de la méthode proposée, une version du corpus a été étiquetée (positif/négatif). Aussi, contrairement aux travaux s'appuyant sur un lexique d'opinions (Turney, 2002), nous cherchons à construire un lexique d'opinions à partir de plusieurs lexiques d'opinions généralistes nous permettant de prédire la polarité véhiculée dans les textes relatifs à l'aménagement du territoire.

Pour identifier un vocabulaire d'opinion spécialisé VOS, nous proposons une méthode en trois étapes :

1. Vocabulaire d'Opinions Pivots généraliste (VOPG) : identification de la liste d'Opinions Pivots (OP) présentes dans le corpus. Une OP est un mot extrait du corpus et présent dans au moins l'un des trois lexiques d'opinions utilisés ;
2. Vocabulaire d'Opinions Contextualisé (VOC) : enrichissement du vocabulaire VOPG par une liste de mots qui seront polarisés au regard du contexte dans lequel ils se trouvent ;
3. Vocabulaire d'Opinions Spécialisé (VOS) : sélection dans le vocabulaire VOC de la liste de mots polarisés liés au domaine. Ce vocabulaire est validé par les géographes qui travaillent sur le projet SENTERRITOIRE.

Dans les sections suivantes, nous décrivons tout d'abord comment construire les différents vocabulaires et détaillons la méthode d'attribution d'un score d'opinion aux documents. Ce score d'opinion est fondé sur une méthode mixte, à la fois statistique et linguistique.

3.2.1 Construction du vocabulaire d'opinions spécialisé

Vocabulaire pivot d'opinion généraliste. Le vocabulaire dit général dispose de nombreux mots ayant une polarité clairement définie. Par exemple, les mots « aimer », « beau », « intéressant » sont clairement associés à une opinion positive tandis que les mots « mauvais », « laid » et « insuffisant » ont une polarité négative. Une « Opinion Pivot » *OP* est un mot extrait du corpus et présent dans au moins l'un des trois lexiques d'opinion utilisés. Les lexiques sont construits manuellement, semi-automatiquement ou de manière contributive. Ainsi, leur richesse et leur « fiabilité » sont très dépendantes de leur mode d'acquisition. Leur utilisation conjointe ou séparée peut être délicate. Il s'agit alors de trouver la meilleure combinaison afin d'obtenir une liste de descripteurs linguistiques d'opinions pertinente et adaptée à la problématique étudiée. Les différentes combinaisons proposées seront évaluées sur le corpus en section 4.3. Il existe peu de lexiques d'opinion spécifiques au français et nous en avons sélectionné trois :

- **Lexique 1** : le lexique *GeneralInquirer* français (Bestgen, 2008), version traduite du GeneralInquirer [4], contient des informations syntaxiques, sémantiques et pragmatiques sur une liste de mots polarisés. La polarité permet d'expliciter pour chaque mot s'il est positif ou négatif. Cette liste est disponible en version française après traduction, lemmatisation et vérification par deux juges (Bestgen, 2008). Au final, le lexique contient 1246 mots positifs et 1527 mots négatifs.

4. http ://www.wjh.harvard.edu/∼inquirer/

- **Lexique 2** : le lexique *LIWC* français (Piolat et al., 2011) est la traduction du dictionnaire anglais du Linguistic Inquiry and Word Count [5] (LIWC). On trouve notamment dans les catégories des mots proposés, ceux ayant trait aux émotions positives et négatives.
- **Lexique 3** : le lexique *JeuxDeMots* (Lafourcade, 2007) est un lexique français étendu à toutes les parties du discours (nom, verbe, adjectif, adverbe) mais également à un grand nombre d'EN (personnes, lieux, marques, évènements). Défini sur la base d'un jeu sérieux [6], permettant de capturer les informations de polarité (opinion positive, négative ou neutre), plus de 250000 termes sont ainsi potentiellement polarisables. À l'heure actuelle, 27529 termes ont été polarisés avec plus de 218739 opinions exprimées.

Afin de consolider ces données d'opinions, nous proposons dans un premier temps de fusionner les différents lexiques en un vocabulaire d'opinion généraliste VOPG.

Ensuite, un score de fiabilité est attribué à chaque type de fusion. Par exemple, si un mot est présent dans les trois lexiques positifs (ou les trois lexiques négatifs), une fiabilité élevée pourra être accordée à ce mot. Ainsi, nous proposons d'accorder un score de fiabilité S_i selon les configurations suivantes :

- $GeneralInquirer \cap LIWC \cap JeuxDeMots = S_1$;
- $GeneralInquirer \cap LIWC = S_2$;
- $GeneralInquirer \cap JeuxDeMots = S_3$;
- $LIWC \cap JeuxDeMots = S_4$;
- $GeneralInquirer = S_5$: mots de GeneralInquirer distincts des autres lexiques ;
- $LIWC = S_6$: mots de LIWC distincts des autres lexiques ;
- $JeuxDeMots = S_7$: mots de JeuxDeMots distincts des autres lexiques.

De manière globale nous pouvons déterminer trois types de scores de fiabilité : un score élevé (S_1) aux mots présents dans les trois lexiques, un score moyen (S_2,S_3,S_4) lorsqu'un mot est présent dans deux lexiques et un score faible (S_5, S_6, S_7) lorsqu'un mot est présent dans un seul des lexiques. Ces scores seront rigoureusement évalués dans la partie expérimentale. Notons qu'un pré-traitement initial a consisté à supprimer les mots présents à la fois dans les lexiques positifs et négatifs (862 mots supprimés).

À partir des extraits de documents présentés Figure 2, les mots suivants sont des *OP*, c'est-à-dire présents au moins dans l'un des trois lexiques d'opinions utilisés :

- magnifique ; - belle ; - solution ;
- loisirs ; - protéger ; - polluées ;
- critiqué ; - craint.

Vocabulaire d'Opinions Contextualisé. Les mots issus du vocabulaire VOPG constitueront des « opinions pivots » (*OP*). Les mots situés dans le même contexte, a priori neutres, seront alors polarisés. Concrètement, si un mot *M* de type adjectif, nom, verbe ou adverbe se situe « proche » d'une *OP* c'est-à-dire selon une fenêtre donnée, pour une phrase donnée), il sera à son tour polarisé. Le score de polarité qui lui sera attribué dépendra de deux facteurs : le score du mot pivot *OP* provenant du vocabulaire d'opinion généraliste et la distance *d* indiquant le

5. http ://www.liwc.net/
6. http ://www.jeuxdemots.org/likeit.php

nombre de mots entre le mot courant *M* et une *OP* (*cf.* Formule (1)).

$$ScoreMotVoisin(M) = \frac{\sum \frac{ScoreMot(OP)}{d}}{\sum OP} \quad (1)$$

Pour chaque mot non pivot candidat à intégrer le vocabulaire d'opinion *MotC*, on identifie l'ensemble des *OP* voisines (c'est-à-dire dans une fenêtre de mots voisins) et pour chaque *OP* sélectionnée, un score de polarité *SPVoisin* est calculé en divisant son score de polarité par la distance avec *MotC*. Le score *ScoreMotVoisin* de *MotC* correspond alors à la moyenne de ces scores *SPVoisin*. Le vocabulaire VOC obtenu à cette étape correspond aux *OP* constituant le vocabulaire VOPG et les mots *MotC* présents dans le même contexte que les *OP*.

L'exemple présenté Figure 3 illustre cette étape. Prenons le mot courant « fortement », en choisissant une fenêtre de quatre mots voisins, seule l'*OP* « critiqué » est retenue. Le score *SPVoisin* du mot « fortement » est égal à -1/1, signifiant que ce mot est à polarité négative.

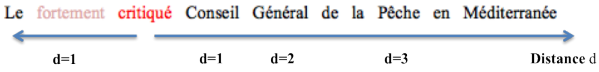

FIG. 3 – *Phrase extraite d'un document du corpus décrivant la région de Sète.*

À partir des extraits de documents présentés Figure 2, les mots suivants sont un extrait des mots ajoutés au vocabulaire VOC :

- naturel ;
- fortement ;
- trouver ;
- gastro-intestinale.

Vocabulaire d'Opinions Spécialisé. Des méthodes de fouille de textes ont permis de mettre en exergue du vocabulaire d'opinions spécialisé VOS à notre domaine. Pour déterminer un tel vocabulaire, nous avons mis en place un module permettant l'identification de descripteurs linguistiques représentatifs sur la base d'un corpus expertisé et du vocabulaire VOC défini en section 3.2.1. Pour chaque descripteur d'opinion *O*, nous comptabilisons le nombre de documents positifs (*nbPos*) et négatifs (*nbNeg*) dans lesquels il est présent. Un premier critère de sélection est défini par la Formule (2) pour supprimer des descripteurs (très peu présents au regard du nombre de documents *nbTDocs*, c'est-à-dire la taille du corpus).

Nous avons testé empiriquement sur le corpus expertisé les mesures couramment utilisées (support, logarithme népérien, logarithme en base 10, tf-idf, etc.) pour filtrer les descripteurs, et nous avons choisi d'utiliser la mesure logarithme en base 10. En effet, nos résultats indiquent que cette mesure est moins restrictive que les autres lorsqu'elle est utilisée sur des corpus de petite ou moyenne taille.

$$nbPos(O) + nbNeg(O) \leq \log(nbTDocs) \quad (2)$$

À partir des extraits de documents présentés en Figure 2, les mots « naturel » et « fortement » sont, par exemple, candidats à intégrer le vocabulaire VOS en utilisant ce critère comme

décrit ci-après :

$$\underbrace{\begin{array}{c} naturel \\ \hline nbPos = 3 \\ nbNeg = 0 \\ log(100) = 2 \end{array}}_{\mathbf{3+0 \geq 2}} \quad \underbrace{\begin{array}{c} fortement \\ \hline nbPos = 0 \\ nbNeg = 2 \\ log(100) = 2 \end{array}}_{\mathbf{2+0 \geq 2}} \quad \underbrace{\begin{array}{c} loisirs \\ \hline nbPos = 1 \\ nbNeg = 0 \\ log(100) = 2 \end{array}}_{\mathbf{1+0 \leq 2}}$$

À noter que le mot « loisirs » n'est pas retenu.

Nous attribuons aux autres descripteurs *O* un Score de Pondération *SPond* (voir Formule (3)) en fonction de leur facteur discriminant et de la proportion de documents positifs et négatifs dans le corpus, *nbTotalDocNeg* et *nbTotalDocPos* étant respectivement le nombre total de documents positifs et négatifs. La fonction *max* avec le deuxième paramètre 1 est utilisée pour éviter que le dénominateur ou le numérateur soit égal à 0, signifiant que le descripteur *O* n'est pas présent dans les documents positifs ou négatifs.

$$SPond(O) = \frac{max(nbPos(O), 1)}{max(nbNeg(O), 1)} \times \frac{nbTDocNeg}{nbTDocPos} \tag{3}$$

À partir des extraits de documents présentés en Figure 2, le score *SPond* pour l'exemple « naturel » serait :

$$SPond(naturel) = \frac{max(3, 1)}{max(0, 1)} \times \frac{40}{60} = \frac{3}{1} \times \frac{2}{3} = 2$$

De la même façon, le score *SPond* pour le mot « fortement » serait 1/3 :

$$SPond(fortement) = \frac{max(0, 1)}{max(2, 1)} \times \frac{40}{60} = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

Cette fonction permet de prendre en compte des corpus déséquilibrés en termes de nombre de documents polarisés. Un descripteur est jugé représentatif si sa présence dans une classe de documents (positive ou négative) est plus importante que dans l'autre. Par conséquent, les descripteurs ambigus, qui ne sont pas représentatifs d'une classe, sont supprimés.

Pour les descripteurs sélectionnés *O*, nous définissons un Score de Représentativité *SR* lié à la classe représentée (voir formule (4)). Une polarité positive est affectée au descripteur si le score de pondération *SPond(O)* est supérieur ou égal à T_r, et négative si la valeur de pondération est inférieure ou égale à 1-T_r.

$$\begin{array}{llll} Si\ SPond(O) \geq T_r & Alors & SR_{pos}(O) = SPond(O) \\ Si\ SPond(O) \leq (1 - T_r) & Alors & SR_{neg}(O) = 1 - SPond(O) \end{array} \tag{4}$$

Le seuil utilisé dans nos expérimentations est T_r = 65%, signifiant que tous les éléments, qui n'ont pas au moins une distribution de 65% dans l'une des deux classes, sont enlevés. En effet, nous considérons qu'un descripteur est représentatif si son nombre d'occurrences dans le corpus est d'au moins 65% en faveur d'une des deux classes (positive ou négative).

Ce seuil (près de 2/3) discuté avec les experts au regard des résultats obtenus nous permet de sélectionner les descripteurs les plus discriminants. Une étape de validation manuelle de ces descripteurs réalisée par les experts géographes permet de produire un vocabulaire d'opinions spécialisé *VOS*.

Par exemple, le mot « naturel » est associé au vocabulaire spécialisé d'éléments positifs lié à l'aménagement du territoire. En effet, comme indiqué ci-dessous, le score *SR* est positif.

$$SPond(naturel) \geq 0,65\ Alors\ SR_{pos}(O) = 2$$

Par contre, le score *SR* de l'opinion « fortement » est ajouté au vocabulaire VOS négatif :

$$SPond(fortement) \leq 0,35\ Alors\ SR_{neg}(O) = \frac{2}{3}$$

3.2.2 Construction du vocabulaire d'opinions spécialisé

Une fois les différents vocabulaires définis, l'étape suivante consiste à attribuer un score de polarité à chaque document. Deux types de pré-traitements sont réalisés : (1) pré-traitement statistique : suppression des mots non discriminants en utilisant leur score IDF (Inverse Document Frequency), (2) pré-traitement linguistique : prise en compte de la négation et attribution de poids aux mots selon leur catégorie grammaticale (*cf. Cat*) identifiée via Tree-Tagger. En ce qui concerne la négation, un premier module de pré-traitement inverse la polarité des *OP* présentes dans les formulations négatives (ne...pas, etc.), et dans une fenêtre de cinq mots après la négation.

Attribution d'un score global pour chaque document. Sur la base des pré-traitements réalisés, un score global de polarité est attribué à chaque objet textuel (mot, phrase et document). Tout d'abord, le score d'opinions est calculé en fonction de leur présence dans les lexiques positif ou négatif (voir Formule (5), avec la polarité de l'opinion *polarity (O)* $\in$ {-1, 1}). Par la suite, cette polarité est pondérée en utilisant un score de fiabilité S_i (Voir la section 3.2.1).

$$ScoreMot(O) = S_i \times polarity(O) \tag{5}$$

À partir des extraits de documents présentés en Figure 2, le score *ScoreMot* des opinions « croissante » et « belle » sont tous les deux égaux à 1 car ils sont extraits de lexiques positifs. Ensuite, le mot « naturel », contenu dans le VOC a un *ScoreMot* égal à 1/3 (voir la section 3.2.1 pour la description de *ScoreMotVoisin*). Dans cet exemple, S_i= 1.

Le score *ScorePhrase* d'une phrase *S* est obtenu en calculant la moyenne des scores attribués à chaque opinion *O* la composant (voir Formule (6)). Un poids noté *Cat* est attribué aux mots en fonction de leur catégorie grammaticale.

$$ScorePhrase(S) = \frac{\sum Cat \times ScoreMot(O)}{\sum Cat \times O} \tag{6}$$

À partir de l'extrait de document 1 présenté Figure 2, le *ScorePhrase* est d'environ 1,70/2 (score de l'Opinion Pivot « magnifique » à 1 et score du mot « naturel » à environ 0,7), signifiant qu'il s'agit d'une phrase positive. Dans cet exemple, *Cat* = 1. Enfin, le score global du document analysé est défini en calculant la moyenne des scores des phrases qui le composent.

3.3 Identification de vocabulaires d'opinions spécifiques aux Entités Spatiales et Organisations

Sur la base des premiers résultats obtenus (listes des EN et le vocabulaire d'opinions VOS), nous proposons dans une troisième étape de construire des vocabulaires d'opinions spécifiques aux ES et aux EO. La difficulté ici est d'identifier des relations entre les opinions et les EN.

L'approche, fondée sur des traitements statistiques, se décompose en quatre étapes : (1) indexation des documents, (2) identification des couples correspondant Opinion-EN, (3) sélection de la taille d'une fenêtre pertinente pour l'identification des couples, et (4) filtrage des opinions pertinentes :

1. **Indexation des documents** : Tout d'abord (Voir la section 3.1), on extrait deux listes d'EN (ES et EO) liées au domaine d'application (*l'aménagement du territoire du bassin de Thau*), puis (voir la section 3.2) nous définissons un vocabulaire d'opinions spécialisé lié au domaine.

2. **Identification des couples Opinion-EN** : Nous investiguons la combinatoire de tous les types de couples de *O-EN* possibles :
   - $O - ES$: une Opinion *O* provenant du vocabulaire spécialisé VOS suivie d'une *ES* ;
   - $ES - O$: une *ES* suivie d'une Opinion *O*. À partir des extraits de documents présentés en Figure 2, les paires « Bassin de Thau, magnifique » et « Sète, craint » sont candidats à constituer des couples *ES-O* ;
   - $O - EO$: une Opinion *O* suivie d'une *EO*, comme par exemple la paire « critiqué, Conseil Général de la Pêche en Méditerranée » ;
   - $EO - O$: une *EO* suivie d'une Opinion *O*, par exemple « SMBT, craint ».

   Afin d'analyser les couples extraits du corpus, nous attribuons un score à chaque couple. Ce score correspond au nombre d'occurrences de couples présents dans les documents selon sept tailles de fenêtres F_i : le document F_1, le paragraphe F_2, deux phrases consécutives F_3, la phrase F_4, N-grammes avec n allant de quatre à deux (*cf.* F_5, F_6, F_7).

3. **Sélection de la fenêtre pertinente** : pour identifier la pertinence d'un couple *O-EN*, nous proposons d'évaluer sa pertinence à différentes échelles en utilisant une fonction de classement. Cette évaluation doit d'abord mettre en évidence la fenêtre F_i contenant les couples les plus pertinents. Dans nos travaux, la pertinence est évaluée sur les possibles associations existant entre opinions et EN dans le corpus traité. Par exemple, nous faisons l'hypothèse de pouvoir déterminer si les paires candidates « croissante, dans les alentours de Sète » et « bassin de Thau, magnifique » sont pertinentes à devenir des couples *O-EN*. Cependant, il est difficile d'identifier les associations de la taille de la fenêtre du corpus. Par conséquent, nous proposons d'évaluer la pertinence de couples à différentes échelles F_1, F_2, F_3, F_4, F_5, F_6 et F_7 en utilisant une fonction de classement s'appuyant sur le nombre d'occurrences des couples dans l'ensemble du corpus.

   Pour illustrer cette étape, nous prenons en exemple les couples présents dans les extraits de documents présentés Figure 2 :
   - bassin de Thau-magnifique, *pertinent* ;
   - naturel-dans les alentours de Montpellier, *pertinent* ;

– craint-Sète, *non-pertinent* ;
– croissante-dans les alentours de Montpellier, *non-pertinent* ;
– critiqué-CGPM, *pertinent*.

Ces couples font partie des candidats à devenir des couples *O-EN* dans la fenêtre F_4, c'est-à-dire la phrase. Les géographes, travaillant sur le projet SENTERRITOIRE, ont validé manuellement la pertinence du vocabulaire spécifique extrait. À noter que les expérimentations présentées en Section 4 sont réalisées sur l'ensemble du corpus traité. Afin de sélectionner les meilleurs couples via la fonction *ClassementOpiland*, nous ordonnons les couples en fonction de leur nombre d'occurrences dans l'ensemble du corpus :

1. bassin de Thau-magnifique, *pertinent*
$nOccurrence = 6$, $ScoreMot(magnifique) = 5.2$;
2. critiqué-CGPM, *pertinent*
$nOccurrence = 5$, $ScoreMot(\textit{critiqué}) = 3$;
3. croissante-dans les alentours de Montpellier, *non-pertinent*
$nOccurrence = 3$, $ScoreMot(croissante) = 5$;
4. naturel-dans les alentours de Montpellier, *pertinent*
$nOccurrence = 2$, $ScoreMot(naturel) = 5$;
5. craint-Sète, *non-pertinent*
$nOccurrence = 1$, $ScoreMot(craint) = 3$.

L'évaluation de la fonction *ClassementOpiland* est discutée en section 4.5.

4. **Filtrage des couples pertinents** : Dans une quatrième étape, les deux listes de couples (une relative aux ES et une deuxième aux EO) sont filtrées en sélectionnant uniquement les couples qui apparaissent dans la fenêtre F_i choisie. Pour chacune des deux listes de couples sélectionnés, les opinions sont extraites pour construire les vocabulaires d'opinions spécifiques à chaque type d'EN.

4 Expérimentations

Dans cette section, nous présentons les résultats de l'approche OPILAND que nous appliquons tout d'abord sur un ensemble de données spécifiques à l'aménagement du territoire du bassin de Thau, puis sur trois autres corpus traitant de domaines différents. Notre objectif est de souligner la pertinence des trois étapes de notre approche au regard de l'état de l'art, et de montrer également la généricité de notre proposition.

4.1 Jeux de données

Nous avons sélectionné 300 articles de journaux relatifs à l'aménagement du territoire décrivant l'étang de Thau situé au Sud de la France. À partir de ce corpus, 100 documents (SENT_100) ont été divisés en deux classes d'opinions : positifs et négatifs. L'opinion exprimée est liée à la formation d'une agglomération regroupant plusieurs communes (Montpellier Agglomération). Le corpus a été validé par des experts géographes travaillant sur le projet SENTERRITOIRE. Le corpus ne contient pas de textes polarisés neutres. Afin d'étudier la généricité de l'approche, nous avons évalué notre proposition sur trois corpus français constitués dans le cadre du défi DEFT'07 (voir la section 4.4).

4.2 Extraction des Entités Spatiales et des Organisations

Concernant la **première étape de l'approche** OPILAND (voir la section 3.1), dans des travaux antérieurs (Tahrat et al., 2013), nous avons expérimenté deux chaînes de TALN pour comparer les patrons définis dans le modèle Pivot et les patrons définis dans OPILAND à partir d'un sous-ensemble du corpus (300 phrases, 8141 mots). L'évaluation est réalisée selon les critères de précision (proportion d'entités pertinentes extraites), de rappel (proportion d'entités pertinentes extraites au regard de l'ensemble des entités pertinentes), et de F-mesure (moyenne harmonique de la précision et du rappel). Le Tableau 1 montre que l'enrichissement des patrons initiaux, fondé sur le modèle Pivot, améliore considérablement la précision et le rappel. La F-mesure est ainsi plus que doublée. De plus, nos modèles nous permettent d'obtenir une précision de 92% pour l'identification des EO. À l'avenir, le rappel pourra être amélioré par l'ajout de nouvelles règles.

Patrons basiques			Patrons OPILAND			
	ESA	ESR		ESA	ESR	EO
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 – *Évaluation des patrons* OPILAND.

Dans nos expérimentations, l'ensemble d'apprentissage se compose de 272 phrases : 138 contenant des ES et 134 contenant des EO. Chaque phrase est alors lemmatisée et représentée par un vecteur binaire. Les meilleurs résultats ont été obtenus avec SVM (Platt, 1999) et Naive Bayes (en utilisant Weka [7]). Le Tableau 2 présente la matrice de confusion associée aux deux classes (ES et EO) pour une évaluation par validation croisée (10 échantillons). Le taux d'exactitude correspond à la proportion d'exemples correctement classés. La méthode hybride améliore le taux d'exactitude (*cf.* Tableau 3). Cela montre que les patrons définis sont particulièrement adaptés à l'identification des ES et des EO (Tahrat et al., 2013).

4.3 Construction du vocabulaire d'opinions spécialisé

La section suivante présente les résultats d'expérimentations concernant la deuxième étape de l'approche OPILAND (voir la section 3.2) sur le corpus SENT_100.

4.3.1 Première base d'évaluation.

La première étape propose une base d'évaluation qui prend en considération l'ensemble des mots pré-sélectionnés selon leur catégorie grammaticale. Le score de classification des documents polarisés est obtenu en utilisant les trois lexiques d'opinions sélectionnés (*GeneralInquirer*, *LIWC* et *JeuxDeMots*). Ces lexiques sont testés indépendamment (voir le Tableau 4).

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

SVM			**Naive Bayes**		
Réel \ Prédit	ES	EO	Réel \ Prédit	ES	EO
ES	103	35	ES	98	40
EO	44	90	EO	44	90
Taux d'exactitude	**70.96%**		*Taux d'exactitude*	*69.12%*	

TAB. 2 – *Classification des phrases sans utilisation de descripteurs.*

Descripteurs avec ConceptOrg			**Descripteurs avec ConceptSpa**			**Les deux types de descripteurs**		
Réel\Prédit	ES	EO	Réel\Prédit	ES	EO	Réel\Prédit	ES	EO
ES	108	30	ES	112	26	ES	113	25
EO	47	87	EO	19	115	EO	19	115
Exactitude	*71.69%*		*Exactitude*	*83.45%*		*Exactitude*	**83.82%**	

TAB. 3 – *Classification des phrases avec utilisation de descripteurs.*

Lexiques	Scores de classification correcte
General Inquirer	**57,5%**
LIWC	54,5%
JeuxDeMots	51,5%

TAB. 4 – *Calcul du score global de classification à partir des lexiques d'opinions généralistes.*

Notons que *GeneralInquirer*, plus complet que *LIWC*, donne les meilleurs résultats (57,5%). Le lexique *JeuxDeMots* apparaît comme le moins efficace. Ces faibles résultats s'expliquent par le fait que ces lexiques ne sont pas directement liés à des domaines spécifiques et ne sont pas adaptés pour détecter des opinions liées à l'aménagement du territoire.

Dans les paragraphes suivants, nous décrivons les expérimentations menées sur le corpus SENTERRITOIRE pour construire un vocabulaire spécialisé pour le marquage des opinions relatives à notre domaine d'application, l'aménagement du territoire.

Construction du Vocabulaire d'Opinions Pivots Généraliste. Nous combinons les trois lexiques afin d'identifier un vocabulaire général contenant les Opinions Pivots *OP*. Pour appliquer la combinaison optimale, nous faisons varier les scores de fiabilité de 0 à 3 [8] (voir la section 3.2.1). La fusion de tous les lexiques fait baisser le score de classification correcte à 52,5%. Cependant, le score est de 63,6% en utilisant un lexique d'opinion général résultant

8. Une plus grande amplitude en pondérant les scores de 0 à 10 s'est avérée non pertinente expérimentalement.

des combinaisons sans prendre en compte le vaste lexique *JeuxDeMots*. Nous observons que les scores sont améliorés en :

- (i) ne conservant que l'intersection des trois lexiques, l'intersection des deux lexiques *GeneralInquirer* et *LIWC* ainsi que le reste du lexique *GeneralInquirer*,
- (ii) en pondérant les Opinions Pivots en fonction du degré d'intersection des lexiques sélectionnés. En utilisant le Vocabulaire d'Opinions Pivots Généraliste VOPG (voir le Tableau 5), l'approche OPILAND fournit un score de classification de 64,6%, comparativement à 57,5% en utilisant uniquement le lexique *GeneralInquirer*.

Construction du Vocabulaire d'Opinions Contextualisé. Dans une troisième étape, nous enrichissons le vocabulaire généraliste VOPG en prenant en compte le contexte des *OP*. Pour ce faire, nous attribuons un score de polarité aux mots voisins des *OP* comme discuté dans la section 3.2.1. Ensuite, un score de polarité est attribué à chaque mot de type nom, verbe, adverbe ou adjectif ayant une ou plusieurs *OP* voisine(s). Ce score correspond à la moyenne des scores des *OP* identifiées dans la fenêtre de taille choisie. Les expérimentations sont effectuées avec une taille de fenêtre de 1 à 4 mots. En effet, 65,6% des documents sont correctement classés lorsque nous étendons la fenêtre de contexte à 4 mots de chaque côté de la cible (voir VOC dans le Tableau 5).

Construction du Vocabulaire d'Opinions Spécialisé. La quatrième étape consiste à définir un Vocabulaire d'Opinions Spécialisé VOS à partir du vocabulaire contextualisé VOC et des descripteurs représentatifs identifiés par validation croisée. Cette méthode est fondée sur une partition des documents en deux échantillons de façon aléatoire. La première partition (appelée échantillon d'apprentissage) est utilisée pour définir un modèle. Ensuite, avec les données restantes (l'échantillon de validation), la qualité de chaque estimation est évaluée en les comparant aux valeurs observées.

Pour rappel, les éléments pertinents sont les mots utilisés pour distinguer les documents positifs des documents négatifs. Tout d'abord, tous les noms, verbes, adverbes et adjectifs sont candidats à devenir descripteurs. Ensuite, nous appliquons deux filtres décrits dans la section 3.2.1 (voir les Formules (2) et (4)) afin de sélectionner les descripteurs pertinents. À partir des 527 descripteurs obtenus par validation croisée, pondérés par défaut à 1, nous obtenons un vocabulaire d'opinions qui améliore considérablement l'identification de la polarité des documents du corpus (81,8%). Dans ces expérimentations, attribuer une pondération de 1 signifie que le descripteur est considéré comme un mot appartenant au Vocabulaire d'Opinions Contextualisé. Ensuite, un score de représentativité *SR* est attribué à chaque descripteur sélectionné en fonction de sa capacité à distinguer les deux classes de documents. Le score d'identification de la polarité est alors nettement amélioré (voir Tableau 5, colonne VOS).

Corpus	Scores corrects		
	VOPG	VOC	VOS
SENT_100	64,6%	65,6%	**91,9%**

TAB. 5 – *Construction des vocabulaires d'opinions relatifs à l'aménagement du territoire.*

4.3.2 Application de pré-traitements linguistiques

Sur la base du vocabulaire spécialisé VOS, nous avons appliqué plusieurs règles linguistiques constituées de (i) la négation, (ii) la polarité (en pondérant les mots d'opinion du vocabulaire général VOPG), et (iii) en fonction de leur catégorie grammaticale. Une analyse préalable du corpus révèle que 18% des phrases contiennent des constructions négatives, et qu'il y a 2,5 fois plus de mots positifs que de mots négatifs. Cependant, les expérimentations ne soulignent pas une amélioration des résultats de classification. Nous souhaitons modifier les règles de marquage des *OP* à l'intérieur des structures négatives en vue d'améliorer ces résultats préliminaires. Un dernier prétraitement consiste à pondérer les *OP* selon leur catégorie grammaticale *Cat* (Voir Formule 6). À noter que ce module ne permet pas d'améliorer significativement les scores de classification.

4.4 Généricité de l'approche

4.4.1 Utilisation de corpus de domaines différents

Afin de valider la généricité de notre approche, nous avons testé la méthode OPILAND sur trois corpus français définis dans le cadre de la campagne d'évaluation DEFT'07. Le premier corpus nommé *CorpusP* est un extrait de 300 interventions anonymisées (Hommes et partis politiques) sur des projets de lois relatifs à l'énergie provenant d'un corpus de débats parlementaires. Le deuxième corpus de test nommé *CorpusV* est constitué de 994 critiques de jeux vidéos. Chaque critique comporte une analyse des différents aspects du jeu - graphisme, jouabilité, durée, son, scénario - et une synthèse globale du jugement. Le troisième corpus nommé *CorpusM* est composé de 3000 critiques de films, livres, spectacles et bandes dessinées.

4.4.2 Résultats des expérimentations

La première étape consiste à évaluer indépendamment les trois lexiques sélectionnés (*GeneralInquirer GI* , *LIWC* et *JeuxDeMots*) sur les trois corpus de domaine à disposition. Les résultats en terme de classification sont présentés Tableau 6.

Comme pour les expérimentations réalisées sur le corpus lié à l'aménagement du territoire, nous notons que le lexique *GI* offre de meilleurs résultats sur le marquage des opinions que les deux autres lexiques. Par exemple, sur le corpus *CorpusM*, le score de classification est de 67,4% en utilisant le lexique *GI*, contre 64,2% en utilisant *LIWC* et 62,2 % en utilisant *JeuxDeMots*. Hormis pour le corpus *CorpusP*, ces premiers résultats sont meilleurs que ceux obtenus sur le corpus lié à l'aménagement du territoire. Cela peut s'expliquer par le fait que ces corpus sont de plus grande taille et ils traitent de moins de sujets.

Ensuite, nous avons testé l'approche OPILAND (i) en générant, un par un , les trois vocabulaires (VOPG , VOC , et VOS) pour chaque corpus (*cf. CorpusP*, *CorpusV*, et *CorpusM*), et (ii) en calculant le score global de polarité. Les résultats sont donnés Tableau 6. Pour chaque corpus, nous notons que l'enrichissement des lexiques traditionnels d'opinions améliore l'identification de la polarité dans les documents. Par exemple, le score obtenu avec *CorpusM* est de 82,6% (voir Tableau 6, colonne VOS) contre 67,4% (voir Tableau 6, colonne GI) à l'aide seulement du lexique *GI*.

Corpus	Scores corrects					
	JeuxDeMots	LIWC	GI	VOPG	VOC	VOS
CorpusP	48,6%	52%	53,6%	54,0%	55,0%	**69,8%**
CorpusV	50%	54,4%	65,5%	67,3%	68,4%	**73,7%**
CorpusM	62,2%	64,2%	67,4%	77,6%	78,8%	**82,6%**

TAB. 6 – *Experimentations de l'approche* OPILAND *sur trois corpus.*

4.4.3 Discussion

Afin de tester la validité de notre travail, nous avons comparé notre approche à des méthodes d'apprentissage supervisé. Notez que de telles approches sont très exigeantes car elles nécessitent un volume important de données étiquetées manuellement. Le Tableau 7 présente les résultats obtenus avec la méthode Naive Bayes (validation croisée avec 10 échantillons). Notons enfin que la méthode SVM donne des résultats du même ordre. Des approches de la littérature proposent également d'intégrer un lexique d'opinion pour la classification en appliquant des algorithmes supervisés. (Sista et Srinivasan, 2004) mettent en avant un score de classification de 76.70% en prenant en compte et en pondérant les mots d'opinions du lexique GI dans une approche SVM classique. L'approche est expérimentée sur un corpus de critique de films et nous pouvons donc les comparer à ceux obtenus sur *corpusM* (82,6% avec notre approche et 88,9% avec une approche classique de type sac de mots).

Méthode	SENT_100	*CorpusP*	*CorpusV*	*CorpusM*
Naive Bayes	51,5%	**78,6%**	**90,5%**	**88,9%**
OPILAND	**91,9%**	69,8%	73,7%	82,6%

TAB. 7 – *Résultats* OPILAND *vs. classifieur Naive Bayes.*

OPILAND vs. approches supervisées. et non-supervisées. Les résultats indiquent que les méthodes supervisées sont inefficaces pour le corpus SENT_100 en raison de sa spécificité, la complexité et l'hétérogéité du vocabulaire utilisé, et sa petite taille. En effet, les méthodes d'apprentissage classiques trouvent leurs limites pour de petits jeux de données. A contrario, les résultats montrent que l'approche OPILAND est plus adaptée pour ce type de corpus.

Enfin, concernant les approches non-supervisées, les travaux s'appuient généralement sur un lexique d'opinions de la communauté et donnent des résultats moins satisfaisants également. (Ohana et Tierney, 2009) s'appuient sur SentiWordNet [9] pour améliorer les résultats de classification des documents d'un corpus en langue anglaise. Bien qu'intéressante par l'aspect automatique, l'approche, utilisant une ressource conséquente en langue anglaise, propose des résultats (67.4%) un peu en deçà de ceux présentés précédemment. En effet, l'approche OPILAND permet d'améliorer ces scores en proposant de combiner différents lexiques d'opinions de la communauté, puis en prenant en compte le contexte dans les documents du corpus traité.

9. http ://sentiwordnet.isti.cnr.it/

4.5 Construction de vocabulaires d'opinions spécifiques liés aux Entités Nommées

4.5.1 Protocole expérimental

Afin d'évaluer la fonction *ClassementOpiland* (voir la section 3.3), nous avons calculé la somme des rangs des couples jugés comme pertinents par les experts. La somme obtenue est appelée $ScoreClassement$. La meilleure fenêtre F_i est celle qui a le score $ScoreClassement$ le plus faible. Cela est équivalent à des approches fondées sur la courbe ROC (Receiver Operating Characteristics) et au calcul de l'aire sous la courbe associée (Ferri et al., 2002).

Notez qu'une fonction parfaite de classement place tous les couples pertinents en tête de la liste. Ce score de classement est appelé $ScoreClassementReference$. Le nombre de couples pertinents change en fonction de la taille de la fenêtre F_i, et nous avons normalisé le score pour chaque fenêtre afin de les comparer et ainsi sélectionner la plus pertinente (voir Formule (7)).

$$NormeScoreClassement(F_i) = \frac{ScoreClassementReference(F_i)}{ScoreClassement(F_i)} \quad (7)$$

Les résultats s'appuyant sur ces critères (*cf.* $ScoreClassement$ et $NormeScoreClassement$) sont présentés en section 4.5.2.

À partir du vocabulaire spécialisé VOS et des deux listes d'EN (ES et EO), nous avons donc effectué des expérimentations pour produire des vocabulaires spécifiques à chacun des deux types d'EN. À cette fin, nous proposons d'évaluer la pertinence de couples *Opinion-EN* aux différentes échelles en utilisant une fonction de classement afin de mettre en évidence la fenêtre F_i contenant les couples les plus pertinents. Un couple *Opinion-EN* est considéré comme pertinent s'il existe une association entre les deux éléments, indiquant ainsi que l'opinion est exprimée sur l'EN. L'ensemble des couples contenus dans cette fenêtre F_i sont ensuite sélectionnés pour constituer les listes d'opinions relatives aux ES et aux EO.

Pour identifier la fenêtre F_i avec les couples les plus pertinents, nous définissons tout d'abord l'ensemble d'apprentissage. Pour ne pas surcharger le travail de l'expert avec des dizaines de milliers de couples apparaissant dans différentes tailles de fenêtres pour les 300 documents qui constituent le corpus de l'étang de Thau, nous avons sélectionné 50 couples *Opinion-EN* liés à chaque type d'EN, soit 100 couples au total.

4.5.2 Résultats

Les résultats d'évaluation sont présentés dans le Tableau 8 (Opinions relatives aux ES) et 9 (opinions relatives aux EO).

L'évaluation de la fonction de classement des résultats présentée dans le Tableau 8 montre qu'il est préférable d'analyser les couples contenant des ES dans la fenêtre de quatre mots F_5, car elle donne le meilleur score normalisé de 0,41. Le Tableau 9 montre qu'il est préférable d'analyser les couples contenant des EO dans la fenêtre de deux mots F_7 avec un score normalisé de 1, mais ces résultats sont discutables car la taille de cette fenêtre permet seulement de détecter deux couples. Par conséquent, nous avons choisi de sélectionner la fenêtre F_5 également qui donne le deuxième score de classement normalisé *NormeScoreClassement* avec dix fois plus de couples pertinents (21).

Fenêtre	F_1	F_2	F_3	F_4	F_5	F_6	F_7
Nombre de paires pertinentes	0	2	7	15	18	13	8
Nombre de paires traitées	50	50	50	50	30	20	10
$ScoreClassementReference$	0	3	28	120	171	91	36
$ScoreClassement$	0	77	211	362	422	274	122
$NormeScoreClassement$	0	0,04	0,13	0,33	**0,41**	0,33	0,26

TAB. 8 – *Evaluation des couples ES - Opinion.*

Fenêtre	F_1	F_2	F_3	F_4	F_5	F_6	F_7
Nombre de paires pertinentes	8	14	18	23	21	13	2
Nombre de paires traitées	50	50	50	50	22	14	2
$ScoreClassementReference$	36	105	171	276	231	91	3
$ScoreClassement$	296	435	538	574	266	134	3
$NormeScoreClassement$	0,12	0,24	0,32	0,48	**0,87**	0,68	**1**

TAB. 9 – *Evaluation des couples Organisation - Opinion.*

À partir des extraits de documents présentés en Figure 2, les couples filtrés sont :

- Pour les ES :
  - bassin de Thau-magnifique ;
  - bassin de Thau-naturel ;
  - étang de Thau-naturel ;
  - naturel-dans les alentours de Sète ;
  - naturel-dans les alentours de Montpellier.
- Pour les EO :
  - critiqué-CGPM.

En filtrant les couples contenus dans la fenêtre sélectionnée F_5, notre méthode génère deux vocabulaires d'opinions spécifiques liés aux ES et aux EO. À noter que le score *ScoreClassement* n'est pas proche du score de référence *ScoreClassementReference* pour les ES et le traitement des occurrences statistiques pourrait être adapté afin d'améliorer le score normalisé *NormeScoreClassement*.

Parmi les 431 opinions constituant le vocabulaire spécialisé VOS, 412 sont liées à des ES ou des EO dans la fenêtre de quatre mots F_5. Parmi ces 412 opinions, 95 sont liées aux ES (parmi lesquelles nous pouvons citer les termes « magnifique » et « naturel », 282 aux EO (comme par exemple le terme « critiqué »), et 35 sont communes aux deux types d'entités. Ces résultats mettent en avant le fait que l'identification de vocabulaires d'opinions spécifiques est également utile pour désambiguïser les EN de type ES et EO.

5 Conclusion et perspectives

L'application de l'approche OPILAND sur un corpus de l'aménagement du territoire de l'étang de Thau a fourni des résultats prometteurs. Dans un premier temps, nous avons proposé une méthode hybride s'appuyant sur le contexte pour extraire des entités nommées de

type entité spatiale et organisation. Notre première contribution est un ensemble de patrons morpho-syntaxiques intégrés à la chaîne de TALN OPILAND. Deux méthodes d'apprentissage supervisé, SVM et Naives Bayes, ont été utilisées pour identifier et désambiguïser les entités spatiales et organisations. Les expérimentations montrent que l'ajout des nouveaux patrons améliore considérablement les résultats en termes de précision et de rappel. Dans une deuxième étape, nous avons construit un vocabulaire d'opinions spécialisé propre au corpus traitant de l'aménagement du territoire. Nous avons montré qu'en utilisant un vocabulaire d'opinions généraliste formé sur la base de trois lexiques traditionnels d'opinions définis par la communauté scientifique, nous avons considérablement amélioré l'identification de la polarité des documents (57,5% en utilisant uniquement le lexique *GeneralInquirer* (le plus efficace des trois lexiques), contre 64,6% en utilisant notre vocabulaire d'opinions généraliste défini). Ensuite, l'identification, par validation croisée, des opinions discriminantes a considérablement amélioré les résultats donnant un score de 91,9% de documents classés correctement. Ainsi, nous pouvons dire que le vocabulaire construit est à la fois spécialisé et adapté pour le marquage des opinions dans le domaine applicatif ciblé. Dans la troisième et dernière étape, nous avons proposé une méthode supervisée pour identifier les vocabulaires d'opinions spécifiques aux entités nommées extraites dans la première étape.

Les travaux à venir seront tout d'abord consacrés à valider le vocabulaire d'opinions spécialisé. Nous prévoyons également d'expérimenter l'approche OPILAND sur différents types de contenu textuel tels que les blogs et les sites Web qui contiennent des opinions exprimées par les acteurs impliqués dans le processus de planification territoriale. Aussi, nous souhaitons expérimenter notre approche sur des corpus multilingues. Enfin, l'identification et la représentation de l'évolution dans le temps de l'opinion des acteurs est un objectif que nous allons aborder à plus long terme.

Remerciements

Les auteurs remercient Pierre Maurel (UMR TETIS) pour son expertise sur le corpus de référence, ainsi que Cédric Lopez (VISEO, France) et Sabiha Tahrat (LIRMM, Montpellier) pour leur participation à ces travaux. Ce travail est soutenu par le labex *NUMEV* et la *Maison des Sciences de l'Homme de Montpellier*.

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. Martìnez (2000). Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*.

Alliès, P. (1980). *L'Invention du territoire*. 184 pages.

Barel, Y. (1981). Modernité, code, territoire, annales de la recherche urbaine. Volume 10-11, pp. 3–21.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In *Proceedings of LREC*, Trento, Italy, pp. 496–500.

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

Buléon, P. et G. D. Méo (2005). *L'espace social*. Annales de la recherche urbaine : Armand Colin.

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

Debarbieux, B. et M. Vanier (2002). *Ces territorialités qui se dessinent*. Datar : Editions de l'Aube, 267 pages.

Deffontaines, J., E. Marcelpoil, et P. Moquay (2001). *Le développement territorial : une diversité d'interprétations*. Maurel, P. Lardon, S., Piveteau, V. (Eds) 184 pages, Paris, Hermès Sciences Publications.

Deleuze, G. et F. Guattari (1980). *Mille plateaux*. Paris : Editions de Minuit. coll. Capitalisme et schizophrénie, 645 pages.

Derungs, C. et R. S. Purves (2013). From text to landscape : locating, identifying and mapping the use of landscape features in a swiss alpine corpus. *International Journal of Geographical Information Science 0*(0), 1–22.

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

Duthil, B., F. Trousset, M. Roche, G. Dray, M. Plantié, J. Montmain, et P. Poncelet (2011). Towards an automatic characterization of criteria. In *International Conference on Database and Expert Systems Applications (DEXA'11)*, Volume 1, Toulouse, France, pp. 457–465. Springer-Verlag, LNCS.

Esuli, A. et F. Sebastiani (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *5th Conference on Language Resources and Evaluation*, pp. 417–422.

Fan, W., S. Sun, et G. Song (2011). Sentiment classification for chinese netnews comments based on multiple classifiers integration. In *Proc. of the Int. Joint Conf. on Comp. Sciences and Optimization*, pp. 829–834.

Ferri, C., P. Flach, et J. Hernandez-Orallo (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of 9th International Conference on Machine Learning, ICML'02*, pp. 139–146.

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.

Harb, A., M. Plantié, G. Dray, M., Roche, F. Trousset, et P. Poncelet (2008). Web opinion mining : How to extract opinions from blogs ? In *Proceedings of the 5th International*

Conference on Soft Computing As Transdisciplinary Science and Technology, CSTST '08, New York, NY, USA, pp. 211–217. ACM.

Husaini, M., A. Kocyigit, D. Tapucu, B. Yanikoglu, et Y. Saygin (2012). An aspect-lexicon creation and evaluation tool for sentiment analysis researchers. In *Proc. of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases*, Volume Part II, pp. 804–807.

Joachims, T. (1998). Text categorization with suport vector machines : Learning with many relevant features. In *ECML*, pp. 137–142.

Joshi, A., P. Balamurali, P. Bhattacharyya, et R. Mohanty (2011). C-feel-it : a sentiment analyzer for microblogs. In *Proc. of HLT*, pp. 127–132.

Kennedy, A. et D. Inkpen (2006). Sentiment classification of movie reviews using contextual valence shifters. In *Computational Intelligence*, Volume 22(2), pp. 110–125.

Kozareva, Z., B. Navarro, S. Vazquez, et A. Montoyo (2007). UA-ZBSA : a headline emotion classification through web information. In *4th International Workshop on Semantic Evaluations*, Stroudsburg, PA, USA, pp. 334–337. ACL.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proc. 7th Symposium on Natural Language Processing (SNLP 2007)*, pp. 13–15.

Lesbegueries, J., C. Sallaberry, et M. Gaio (2006). Associating spatial patterns to text-units for summarizing geographic information. In *Proceedings of ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop*, pp. 40–43. LIUPPA.

Liu, B. (2012). Sentiment analysis and opinion mining. pp. 167. Morgan and Claypool Publishers.

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.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03)*.

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

Ohana, B. et B. Tierney (2009). Sentiment classification of reviews using sentiwordnet.

Pak, A. (2012). *Automatic, Adaptive, and Applicative Sentiment Analysis*. Ph. D. thesis, Thèse de lÕÉcole Doctorale d'Informatique de l'Université Paris-Sud, Orsay.

Pak, A. et P. Paroubek (2010). Microblogging for micro sentiment analysis and opinion mining. In *TAL*, Volume 51(3), pp. 75–100.

Pang, B. et L. Lee (2010). Opinion mining and sentiment analysis. In *Found. and Trends in IR*.

Pecqueur, B. (2007). L'économie territoriale : une autre analyse de la globalisation. In *Alternatives économiques, lÕéconomie politique*, Volume 1 (33), pp. 41–52.

Piolat, A., R. Booth, C. Chung, M. Davids, et J. Pennebaker (2011). La version française du dictionnaire pour le liwc : modalités de construction et exemples d'utilisation. In *Psychologie Française*, Volume 56(3), pp. 145–159.

Plantié, M., M. Roche, G. Dray, et P. Poncelet (2008). Is a voting approach accurate for opinion

mining ? In *Proc. of DatawWrehousing and Knwoledge discovery (DaWaK'08)*, pp. 413–422.

Platt, J. C. (1999). Advances in kernel methods. Chapter Fast training of support vector machines using sequential minimal optimization, pp. 185–208. Cambridge, MA, USA : MIT Press.

Salles, M. (2009). In F. L. Gandon (Ed.), *Actes d'IC*, pp. 109–120.

Sista, S. et S. Srinivasan (2004). Polarized lexicon for review classification. In *MLMTA'04 : Proceedings of the International Conference on Machine Learning ; Models, Technologies et Applications, CSREA*. Press.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teisseire (2013). Text2geo : from textual data to geospatial information. In *The 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS'13)*.

Torres-Moreno, J., M. El-Bèze, F. Béchet, et N. Camelin (2009). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pp. 417–424.

Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pp. 417–424.

Vanier, M. (2009). Territoires, territorialité, territorialisation - controverses et perspectives. In *PUR*, pp. 417–424.

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.

Wiebe, J. et E. Riloff (2011). Finding mutual benefit between subjectivity analysis and information extraction. In *IEEE Transactions on Affective Computing*, Volume 2 (4), pp. 175–191.

Zidouni, A., M. Quafafou, et H. Glotin (2009). Structured named entity retrieval in audio broadcast news. In S. D. Kollias et Y. S. Avrithis (Eds.), *CBMI*, pp. 126–131. IEEE Computer Society.

Summary

A great deal of research on information extraction from textual datasets has been performed in specific data contexts, such as movie reviews, commercial product evaluations, campaign speeches, etc. In the SENTERRITOIRE project, we raise the question on how appropriate these methods are for documents related to land-use planning. The kind of information sought concerns the stakeholders, opinions, geographic information, and everything else related more generally to the territory. However, it is extremely challenging to link opinions to these kinds of informations. After highlighting the limitations of existing proposals and discussing issues related to textual data, we present a semi-automatic method called Opiland (OPinion mIning

from LAND-use planning documents) combining a NLP process and Text Mining tools in order (1) to extract named-entities (Locations and Organizations), (2) to build a vocabulary of opinions related to a domain, and (3) to mine opinions related to the extracted named-entities. Experiments are conducted on a Thau lagoon dataset (France), and then applied on three datasets that are related to diverse areas in order to highlight the relevance and the genericity of our proposal.