Traitement automatique des données hétérogènes liées à l'aménagement des territoires

Mathieu Roche, Maguelonne Teisseire

Traitement automatique des données hétérogènes liées à l'aménagement des territoires

Mathieu, ROCHE & **Maguelonne, TEISSEIRE,**
UMR TETIS (Cirad, Irstea, AgroParisTech)
500 rue Jean-François Breton
34093 Montpellier, France

Contact : *prénom.nom@teledetection.fr*

Résumé

La notion 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. Aujourd'hui, avec le développement des technologies numériques (blogs, forums, réseaux sociaux, etc.), l'ensemble des acteurs impliqués s'expriment et tous les documents textuels ainsi produits constituent une source considérable d'informations qu'il est crucial d'analyser. Dans cet article, nous souhaitons poser les premières bases d'une méthode automatique d'extraction de connaissances permettant d'analyser le ressenti (opinion et/ou sentiment) des acteurs impliqués à partir d'un corpus de données totalement hétérogènes constitués spécifiquement pour un territoire. Une telle approche, qui se situe dans le domaine de la science des données, offrira aux décideurs et aux usagers d'un territoire un environnement leur permettant d'en obtenir les clefs de lecture et d'en mesurer tous les enjeux et les contours.

Mots clefs

Science des données, données massives et hétérogènes, Traitement Automatique du Langage Naturel (TALN), aménagement du territoire, entités spatiales.

1 Introduction

Traditionnellement, les géographes abordent un territoire selon une analyse fondée sur les acteurs impliqués et les usages étudiés dans un espace (Debarbieux et Vanier, 2002 ; Buléon et Di Méo, 2005). Aujourd'hui, avec le développement des technologies numériques (blogs, forums, réseaux sociaux, etc.), les acteurs comme les citoyens ou les décideurs peuvent s'exprimer de façon plus visible et souvent en temps réel. Tous les documents textuels alors accessibles deviennent une source considérable d'informations qu'il est crucial d'analyser. Pour traiter ces masses de données (c'est-à-dire l'infobésité), la problématique de recherche du Big Data (ou mégadonnées) est classiquement mise en avant avec les 3 V qui la caractérisent : volume, variété et vélocité. Mais d'autres caractéristiques ne doivent pas être négligées comme la versatilité, la véracité, la visualisation ou la valorisation des données et informations. Toutes ces problématiques ouvrent de nouvelles disciplines de recherche comme la Science des Données qui mêle mathématiques, statistiques, informatique et visualisation. Les travaux que nous proposons se situent dans ce contexte et se concentrent sur la mise en relation des données textuelles hétérogènes (critère de variété) (Adderley *et al.*, 2014).
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. La caractérisation et la compréhension des perceptions d'un même territoire par les différents acteurs sont difficiles, mais néanmoins particulièrement intéressantes dans une perspective d'aménagement du territoire (Derungs et Purves, 2013) et de politique publique territoriale.

Depuis de nombreuses années, des travaux se sont intéressés à extraire des connaissances à partir de documents textuels. Cette problématique abordée, dans le contexte des territoires numériques, est très complexe et nous confronte aux challenges suivants : comment mettre en évidence le lien existant entre les notions d'un triptyque Territoire, Acteur et Sentiment à partir de textes ?
A partir d'un corpus de données totalement hétérogènes constitués spécifiquement pour un territoire, nous souhaitons poser les premières bases d'une méthode permettant d'analyser le ressenti (opinion et/ou sentiment) des acteurs impliqués. À ce jour, malgré des besoins évidents exprimés par les professionnels du domaine, il n'existe pas d'environnement décisionnel allant de la conception de modèles d'observation/prospection jusqu'à l'analyse et la restitution de connaissances nouvelles sur la perception des dynamiques de ces territoires.
Il faut dire qu'une telle analyse n'est pas simple et trouve ses bases dans des méthodes automatiques. Celles-ci permettront d'accompagner les experts afin de répondre aux questions suivantes : Comment définir et capturer les concepts constituant la notion de Territoire (lieux géographiques, thématiques, acteurs, opinion et sentiments) ? Comment intégrer les données complexes (données potentiellement très bruitées avec des spécificités lexicales et/ou syntaxiques) et hétérogènes disponibles ? Comment identifier des ressources correspondant à des données factuelles (cartes, statistiques, rapports techniques, etc.) définies par l'administration technique et les bureaux d'étude et surtout comment les mettre en correspondance avec celles de données plus subjectives produites par les acteurs locaux des territoires sur divers médias (journaux en ligne, blogs, commentaires sur le Web, réseaux sociaux type Twitter, etc.) ?

La notion de sentiment, même si elle a été largement étudiée dans le contexte des forums, des critiques de films, de livres… n'a pas été très étudiée dans le contexte complexe des territoires. Il est nécessaire de procéder à une analyse plus fine des sentiments que celles proposées dans la littérature qui se cantonnent généralement à qualifier un sentiment sur une phrase, un paragraphe, un document sans réellement identifier finement des relations entre sentiments et informations liées à un territoire. De manière complémentaire, la notion d'émotion a encore été très peu étudiée. Elle trouve toute sa place lorsque les différents acteurs expriment différentes formes d'émotion (peur, joie, colère, etc.).

Dans cet article, nous nous intéressons plus particulièrement à l'étape permettant d'extraire, de façon automatique, à partir de données textuelles hétérogènes, les entités nécessaires à une telle analyse (Sentiments, Thématiques, Lieux et Acteurs). Analyse qui, même si elle est ambitieuse, devient nécessaire face aux développements des médias, supports de transmission actuels et au volume de données qui en résulte.

L'article est organisé de la façon suivante. La section 2 décrit la méthodologie globale de fouille de textes hétérogènes que nous proposons d'appliquer au domaine de l'aménagement du territoire. Nous détaillons l'extraction automatique de la terminologie qui est la base de notre approche en Section 3. Nous proposons ensuite, en Section 4, une discussion à partir des premiers résultats obtenus. Nous brossons ensuite quelques pistes d'intérêt, en Section 5, avant de conclure sur l'ensemble des travaux à mener pour fournir aux décideurs et aux usagers d'un territoire un environnement leur permettant d'obtenir les clefs de lecture et d'en mesurer tous les enjeux et les contours.

2 Fouiller les textes hétérogènes liés à l'aménagement du territoire

La démarche méthodologique que nous proposons d'adopter pour fouiller et rapprocher les données hétérogènes est itérative, c'est-à-dire que l'expert peut être guidé ou non par les données et/ou le résultat d'un traitement. Dans ce cadre, trois voies de mise en relation des données hétérogènes sont proposées : (1) liaison thématique, (2) liaison spatiale, (3) liaison temporelle. L'objectif est de proposer une démarche générique qui combine ces trois mises en relation qui sont tout à fait complémentaires. Ces dernières reposent, en partie, sur la présence dans les textes de descripteurs linguistiques associés à ces trois concepts (thème, entité spatiale, entité temporelle). Pour extraire ces descripteurs, des méthodes de Traitement Automatique du langage Naturel (TALN) sont mises en œuvre et évaluées à partir d'un corpus de 300 articles de presse issus de Midi Libre décrivant l'aménagement du territoire de l'étang de Thau (Kergosien *et al.*, 2015).
Outre l'extraction de descripteurs linguistiques liés aux trois axes définis, l'approche décrite en Section 3 permet de mettre en exergue le vocabulaire véhiculant un sentiment. L'approche automatique et générique pour extraire ces descripteurs, en particulier les lieux, personnes, thèmes ainsi que les sentiments, est détaillée dans la section suivante.

3 Terminologie liés et aménagement du territoire

Cet article décrit une approche de Fouille de Textes (FT) fondée sur l'extraction de la terminologie. Dans un tel processus, il est nécessaire, au préalable, d'étiqueter les textes. Ce processus est détaillé dans la section suivante.

3.1 Etiquetage grammatical

L'étiquetage est le processus qui consiste à associer aux mots d'un texte une fonction grammaticale (nom, verbe, etc.) en exploitant des informations lexicales et/ou contextuelles. Nos travaux s'appuient sur l'utilisation du Tree Tagger (Schmid, 1994). Un tel étiqueteur estime la probabilité qu'un mot ait une étiquette grammaticale (nom, adjectif, déterminant, etc.) en s'appuyant sur des arbres de décision binaires (Quinlan, 1986). Ces derniers sont construits récursivement à partir d'un ensemble de trigrammes connus (suites de trois étiquettes grammaticales consécutives constituant un ensemble d'apprentissage). Le processus complet de construction des arbres de décision est détaillé dans (Schmid, 1994). Un exemple d'étiquetage est donné ci-dessous.

- Exemple *avant* étiquetage :
L'agriculture itinérante reste un système de production partagé par une grande majorité de groupes…

- Exemple *après* étiquetage (la première colonne indique la phrase originale, la deuxième colonne donne les étiquettes grammaticales associées aux mots et la dernière colonne décrit les mots lemmatisés) :

L'	*DET:ART*	*le*
agriculture	*NOM*	*agriculture*
itinérante	*ADJ*	*itinérant*
reste	*VER:pres*	*rester*
un	*DET:ART*	*un*
système	*NOM*	*système*
de	*PRP*	*de*
production	*NOM*	*production*
partagé	*VER:pper*	*partager*
par	*PRP*	*par*
une	*DET:ART*	*un*
grande	*ADJ*	*grand*
majorité	*NOM*	*majorité*
de	*PRP*	*de*
groupes	*NOM*	*groupe*
…		

Une fois l'étiquetage grammatical effectué, l'étape suivante consiste à extraire la terminologie. Dans ce contexte, nous nous sommes concentrés sur l'extraction de la terminologie nominale, c'est-à-dire les groupes de mots faisant intervenir des noms et qui sont pertinents au regard de la thématique abordée dans nos travaux liés à l'aménagement du territoire.

3.2 Extraction de termes issus de corpus spécialisés

Afin d'extraire la terminologie liée aux thématiques, acteurs et sentiments, nous avons adapté un système de fouille de textes qui exploite à la fois des informations statistiques et linguistiques pour extraire la terminologie à partir de textes libres. Les informations statistiques apportent une pondération des termes candidats extraits. Cependant, la fréquence d'un terme n'est pas nécessairement un critère de sélection adapté. À titre d'exemple, le mot *Montpellier* présent dans de très nombreux textes liés à notre jeux de données (cf. section 4) se révèle très général et pas suffisamment discriminant. Ainsi, des mesures de discriminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes, peuvent être appliquées.

Pour effectuer une telle sélection, nos travaux s'appuient sur la mesure TF-IDF. Cette dernière donne un poids plus important aux termes caractéristiques d'un document (Salton et McGill, 1983). Ainsi, pour attribuer un poids de TF-IDF, il est nécessaire, dans un premier temps, de calculer la fréquence d'un terme (*Term Frequency*). Celle-ci correspond au nombre d'occurrences de ce terme dans le document considéré. Ainsi, pour le document d_j et le terme t_i, la fréquence du terme dans le document est donnée par l'équation suivante :

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

où n_{ij} est le nombre d'occurrences du terme t_i dans d_j. Le dénominateur correspond au nombre d'occurrences de tous les termes dans le document d_j.
La fréquence inverse de document (*Inverse Document Frequency*, IDF) mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme et est définie de la manière suivante :

$$IDF_i = \log_2 \frac{|D|}{|d_j : t_i \in d_j|}$$

où $|D|$ représente le nombre total de documents dans le corpus et $|d_j : t_i \in d_j|$ représente le nombre de documents où le terme t_i apparaît. Enfin, la pondération finale s'obtient en multipliant les deux mesures :

$$TF-IDF_{ij} = TF_{ij} \times IDF_i$$

Ces mesures sont intégrées dans une suite opérationnelle, nommée BioTex[1], développée au sein de notre équipe de recherche (Lossio Ventura *et al.*, 2014). BioTex prend en compte deux facteurs pour extraire la terminologie. Tout d'abord, l'approche extrait des termes selon des patrons syntaxiques définis (nom-adjectif, adjectif-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un autre filtrage statistique est appliqué. Celui-ci mesure

[1] http://tubo.lirmm.fr/biotex/

l'association entre les mots composant un terme (par exemple, *zone inondable*) en utilisant une mesure appelée C-value (Frantzi *et al.*, 2000) tout en intégrant la pondération TF-IDF. Le but de C-value est d'améliorer l'extraction des termes complexes.

4 Expérimentations

4.1 Protocole expérimental

Dans nos expérimentations, nous avons constitué un corpus de données textuelles liées à l'aménagement de la nouvelle gare TGV de Montpellier. Ce corpus composé d'une vingtaine[2] de textes très hétérogènes est issu d'articles de journaux, blogs, Facebook, description officielle du projet, etc.

La nouvelle gare de Montpellier devrait être mise en place en 2017 sur un site éloigné du centre de la ville de Montpellier. Complémentaire à la gare Saint-Roch, elle doit accompagner le développement du transport de passagers et marchandises à grande vitesse, sur de longues distances, via un nouveau tracé contournant Nîmes et Montpellier. La nouvelle gare a pour objectif de favoriser le développement économique de Montpellier en permettant le passage plus fréquent des trains à grande vitesse. Cependant, la construction de la gare suscite des débats très riches et contradictoires dans les divers médias et notamment à travers les réseaux sociaux. La section suivante met en exergue la terminologie associée à l'aménagement du territoire liée à cette nouvelle gare.

4.2 Analyse des résultats

A partir du corpus précédemment présenté, nous avons étudié qualitativement les premiers termes obtenus avec le logiciel BioTex. Ces termes sélectionnés sont ceux obtenant les meilleures pondérations statistiques selon les mesures précédemment présentées. Le lecteur intéressé peut se reporter à (Lossio Ventura *et al.*, 2014) pour obtenir plus de détails sur les définitions correspondantes.

Les tableaux 1 et 2 mettent en avant les termes composés extraits automatiquement (sur les 100 premiers termes classés par une mesure statistique). Nous avons choisi d'extraire des termes composés qui sont davantage porteurs de sens que les termes simples (par exemple, *gare*, *projet*, *zone*, *enquête*, *accès*, etc.). Les termes composés avec notre système de fouille de textes véhiculent des informations associées au triptyque Territoire (entités spatiales et thèmes), Acteurs et Sentiments.

Dans un premier temps, nous remarquons que les termes propres aux sentiments (cf. Tableau 1) peuvent être répertoriés en 2 catégories : **termes généraux** (*avis favorable*, *avis négatif, etc.*) ou **termes spécifiques liés à la problématique de l'aménagement du territoire** (*infrastructures inutiles*, *projet inutile, etc.*) et plus particulièrement au projet de construction de la nouvelle gare de Montpellier (*gare fantôme*, *gares excentrées, etc.*).

[2] Le thème étant très spécifique, le nombre d'articles disponible est assez réduit. L'objectif de cet article est avant tout une preuve de faisabilité de la méthodologie.

Sentiments positifs	avis favorable, coût moindre, gare innovante, gares complémentaires
Sentiments négatifs	avis défavorable ; avis négatif ; gares excentrées, gare fantôme ; projet inutile ; aveuglement coupable ; choix étonnant ; complètement perdus ; craintes légitimes ; gares absurdes ; graves inconvénients ; infrastructures inutiles

Tableau 1 : Sentiments liés à l'aménagement de territoires extraits (mots composés).

Dans un second temps, nous avons analysé les thèmes qui sont mis en avant par notre outil d'extraction de la terminologie (cf. Tableau 2). Nous pouvons identifier 5 catégories : (1) les aspects **politiques et administratifs liés à l'aménagement** (*enquête publique, commission départementale, etc.*), (2) les aspects **socio-économiques** (*analyse socio-économique, aspect financier, etc.*), (3) les concepts liés aux **nouvelles infrastructures urbaines** (*mosaïque urbaine, campus créatif, etc.*), (4) la thématique des **transports et de l'accessibilité** (*circulation routière, accès piéton, etc.*), (5) les aspects **environnementaux** (*développement durable ; espèces protégées, etc.*). Notons que certains termes identifiés manuellement comme des instances de thèmes peuvent également être associés aux acteurs au sens large (par exemple, *tribunal administratif*). Les termes correspondant aux acteurs qui sont extraits sont plutôt liés aux **décideurs** (par exemple, *monsieur saurel*) ou aux **animateurs de la politique locale** (par exemple, *élus locaux ; député vert*). Enfin, les entités spatiales extraites identifient des **localisations précises** (par exemple, *embranchement saint-brès*) ou des **zones un peu plus diffuses** (par exemple, *zone inondable*).

Thèmes	enquête publique ; utilité publique ; campagne électorale ; commission départementale ; commission nationale ; conseil municipal ; tribunal administratif ; autorités locales analyse socio-économique ; aspect financier ; aménagement commercial ; activité commerciale ; zone commerciale campus créatif ; mosaïque urbaine ; espace public ; complexité technique transports publics ; accès piéton ; accès routier ; circulation routière ; contournement ferroviaire terrain naturel ; coulées vertes ; développement durable ; espèces protégées ; espaces verts
Lieux	zone inondable ; embranchement saint-brès ; emplacement ouest
Acteurs	commissaire enquêteur ; élus locaux ; député vert ; monsieur saurel ; communes voisines

Tableau 2 : Informations liées aux thèmes, lieux et acteurs avec extraits avec un système d'extraction de la terminologie (mots composés).

Le protocole de validation qualitatif des différents termes extraits par BioTex a été réalisé en double aveugle par des utilisateurs. Il est intéressant de constater que certains termes ont obtenu un double étiquetage. Par exemple, *campus créatif* a été annoté par un des experts à la fois comme un thème et un sentiment positif. Il est important de souligner l'aspect cognitif indéniable qu'il existe dans la perception individuelle des différents termes car celui-ci augmente la complexité d'une analyse systématique des ressentis et opinions des acteurs et individus impliqués.

5 Discussion

Les premiers résultats soulignent la pertinence de l'approche proposée. Il s'agit maintenant de poser la question de méthodes plus spécifiques pour chacune des entités explorées (Sentiments, Thèmes, Acteurs et Lieux). Nous proposons d'explorer plus particulièrement la problématique de l'extraction des Entités Nommées (EN) en soulignant leurs spécificités et complexités. Les EN sont classiquement définies comme les noms de Personnes, Lieux et Organisations. Initialement, une telle définition est issue des campagnes d'évaluation américaines MUC (Message Understanding Conferences) qui furent organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les EN dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Aujourd'hui, de telles campagnes d'évaluation couvrent des tâches très variées sur la base de textes de différents domaines (textes spécialisés en biologie, dépêches d'actualités, blogs, etc). Nous pouvons, entre autres, citer les challenges TREC - Text REtrieval Conference (international) et DEFT - DEfi Fouille de Textes (francophone) qui sont aujourd'hui très actifs dans la communauté « fouille de textes ». De nombreuses méthodes permettent d'identifier les EN (Nadeau et Sekine, 2007). Par exemple, 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 (Nouvel et Soulet, 2011). Ce type de règles utilise des informations syntaxiques propres aux phrases (Brun et Hagège, 2004 ; Nouvel et Soulet, 2011). Par ailleurs, pour identifier les EN, de nombreux systèmes s'appuient sur la présence de majuscules (Daille *et al.*, 2000). Cependant ceci peut se révéler peu efficace dans le cas d'EN non capitalisées et pour le traitement de textes non normalisés (mails, blogs, textes ou fragments de textes inégalement en majuscule ou minuscule, etc.). 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. Ce même principe est utilisé pour le traitement des EN à partir de données textuelles complexes comme les tweets (Ritter *et al.*, 2011).

Pour l'identification des EN et/ou leur catégorie, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé (Tjong Kim Sang et De Meulder, 2003). Ces méthodes d'apprentissage comme les SVM ou les arbres de décision 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 la suite de nos travaux, nous souhaitons combiner de telles méthodes d'apprentissage supervisé associées à des patrons linguistiques et nous appuyer notamment sur les travaux de (Lesbegueries *et al.*, 2006) pour la définition de patrons linguistiques pour l'extraction d'ES permettant d'obtenir des résultats plus pertinents (Kergosien *et al.*, 2015).

Outre les EN, d'autres approches consistent à identifier les sentiments dans les textes à partir de méthodes de fouille de textes. Les principaux travaux de recherche considèrent que l'orientation sémantique d'une opinion est exprimée par l'intermédiaire des adjectifs (Turney, 2002 ; Taboada *et al.*, 2006 ; Kamps *et al.*, 2004) bien que les verbes puissent également véhiculer un sentiment (Sokolova et Lapalme, 2011). Des approches ont enrichi l'apprentissage des adjectifs à l'aide de ressources existantes, par exemple WordNet (Miller, 1995). Dans ce cadre, il s'agit d'intégrer automatiquement les synonymes et les antonymes (Andreevskaia et Bergler, 2006) ou d'acquérir des mots porteurs d'opinions (Hu et Liu, 2004). Ces dictionnaires représentent la base essentielle pour déterminer la polarité générale véhiculée par un document (Taboada *et al.*, 2011). Notons que des traitements complémentaires, comme la prise en compte de la négation pour le changement de polarité, sont souvent déterminants (Wiegand *et al.*, 2010 ; Taboada *et al.*, 2011).

6 Conclusion et perspectives

Une fois les données collectées, structurées et les entités extraites, il s'agit de s'intéresser à l'étape d'analyse en elle-même, étape orientée vers les décideurs. L'idée est ici de pouvoir répondre à de nombreuses questions sociétales qui pourraient être posées afin de mieux appréhender les usages et la perception du territoire par les acteurs.

Voici quelques exemples de questions posées : le sentiment est-il associé à l'appropriation du territoire ou de parties du territoire par les usagers ? L'opinion ou l'émotion ne concerne-t-elle que les projets d'aménagement soumis à consultation ? Que pense et que ressent chaque catégorie d'acteurs de la dynamique de leur territoire ? Comment les supports médiatiques, les rôles et statuts des acteurs qui s'expriment, contraignent-ils les formes et modalités d'expression des sentiments et des opinions dans les documents numériques ? Existe-t-il des événements particuliers qui suscitent l'expression de sentiments ou d'émotions jusque-là inexprimés ? Quels effets déformants sur l'expression des sentiments ont les médiateurs humains (ex : journaliste, porte-parole d'un collectif) ou non humains (par exemple, blog vs. registre papier d'enquête publique) ? Cette forme d'analyse implique de mettre à disposition l'ensemble des données ainsi constituées pour enrichir la prise de décisions territoriales et améliorer leur acceptation sociale.

Les travaux à mener sont nombreux et nous avons souhaité en poser les premières bases en soulignant l'intérêt d'adjoindre, à toute analyse sur la dynamique territoriale et les perceptions des acteurs sur ces territoires, des procédés automatiques d'extraction de connaissances.

Remerciements

Nous souhaitons remercier tout particulièrement Juan Antonio Lossio pour ses travaux de recherche et la mise à disposition de l'outil BioTex ainsi que le projet **Tectoniq**[3] qui ont été sources de soutien et d'inspiration pour cet article.

Références bibliographiques

ADDERLEY R., SEIDLER P., BADII A., TIEMANN M., NERI F., RAFFAELLI M. (2014) *Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights*. Proceedings of The Fourth International Conference on Advances in Information Mining and Management, IARIA, p.36-40,

ANDREEVSKAIA A. AND BERGLER S. (2006) Semantic tag extraction from wordnet glosses. In Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation.

BONNEFOY L., BELLOT P., MICHEL B. (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*, Proceedings of Conférence en Recherche d'Informations et Applications (CORIA)

BRUN C. AND HAGÈGE C. (2004). *Intertwining deep syntactic processing and named entity detection.* In Proceedings of Advances in Natural Language Processing, 4th International Conference (EsTAL), p.195-206

BULÉON P. et DI MÉO G. (2005) *L'espace social*, Armand Colin.

CARRERAS X., MÀRQUEZ L., AND PADRO L. (2003) *A Simple Named Entity Extractor using AdaBoost*. Proceedings of Conference on Natural Language Learning.

DAILLE B., FOUROUR N., AND MORIN E. (2000) Catégorisation des noms propres : une étude en corpus, Cahiers de Grammaire, Vol 25, p.115-129.

DEBARBIEUX B, VANIER M. (2002) *Ces territorialités qui se dessinent*. Editions de l'Aube, Datar, 267p.

DERUNGS C. et PURVES R. S. (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.

FRANTZI K., ANANIADOU S., MIMA H. (2000) *Automatic recognition of multi-word terms: the C-value/NC-value method.* International Journal on Digital Libraries, 3(2), p. 115-130.

HU M. AND LIU B. (2004) Mining and summarizing customer reviews. In Proceedings of KDD'04, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA.

[3] http://tectoniq.meshs.fr/

KAMPS J., MARX M., MOKKEN R., AND DE RIJKE M. (2004) *Using wordnet to measure semantic orientation of adjectives.* In Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation, IV :174-181.

KERGOSIEN E., MAUREL P., ROCHE M., TEISSEIRE M. (2015) *SENTERRITOIRE pour la détection d'opinions liées à l'aménagement d'un territoire.* Revue Internationale de Géomatique 25(1): 11-34.

LESBEGUERIES, J., SALLABERRY, C., GAIO, M. (2006) *Associating spatial patterns to text- units for summarizing geographic information.* Proceedings of ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop, 40-43.

LOSSIO VENTURA J.-A., JONQUET C., ROCHE M., TEISSEIRE M. (2014) *Integration of linguistic and web information to improve biomedical terminology extraction*. Proceedings of the International Database Engineering and Applications Symposium (IDEAS), p. 265-269.

MILLER G. (1995) *Wordnet : A lexical database for english.* In Communications of the ACM.

NADEAU D. AND SEKINE S. (2007). A survey of named entity recognition and classification, Linguisticae

Investigationes, 30(1), p. 3-26

NOUVEL D. AND SOULET A. (2011) *Annotation d'entités nommées par extraction de règles de transduction.* Proceedings of Extraction et Gestion des Connaissances (EGC), p.119-130.

QUINLAN J.R. (1986*) Induction of decision trees*, Machine Learning, Vol.1, p. 81-106.

RITTER A., CLARK S., MAUSAM, ETZIONI O. (2011) *Named Entity Recognition in Tweets: An Experimental Study.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). p.1524-1534.

SALTON G., MCGILL M.J. (1983*) Introduction to Modern Information Retrieval*. McGraw-Hill.

SCHMID H. (1994) *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the Int. Conf. on New Methods in Language Processing, p. 44-49.

SOKOLOVA M. AND LAPALME G. (2011) *Learning opinions in user-generated web content.* Natural Language Engineering, 17(4) :541-567.

TABOADA M., ANTHONY C., AND VOLL K. (2006). *Creating semantic orientation dictionaries.* In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).

TABOADA M., BROOKE J., TOLOSKI M., VOLL K., AND STEDE M. (2011) *Lexicon-based methods for sentiment analysis.* Computational Linguistics, 37(2) :267-307.

TJONG KIM SANG E.F., DE MEULDER F. (2003) *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.* Proceedings of Conference on Natural Language Learning

TURNEY P. (2002) *Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews.* In Proceedings of 40th Meeting of the Association for Computational Linguistics, pages 417-424.

WIEGAND M., BALAHUR A., ROTH B., KLAKOW, D., AND MONTOYO A. (2010*) A survey on the role of negation in sentiment analysis.* In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing.