Extraction d'information spatiale à partir de données textuelles non-standards

Sarah Zenasni

THESE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Informatique

École doctorale I2S

Unité de recherche TETIS

Extraction d'information spatiale à partir de données textuelles non-standards

Présentée par Sarah ZENASNI
Le 05/01/2018

Sous la direction de Maguelonne TEISSEIRE,
Mathieu ROCHE et Eric KERGOSIEN

Devant le jury composé de

Cédrick FAIRON, PR, CENTAL, Université Catholique de Louvain	Rapporteur
Christian SALLABERRY, MCF HDR, LIUPPA, Université de Pau	Rapporteur
Carmen GERVET, PR, Espace-Dev, Université de Montpellier	Examinatrice
Maguelonne TEISSEIRE, DR, TETIS, Irstea	Directrice
Mathieu ROCHE, Chercheur HDR, TETIS, Cirad	Co-Directeur
Eric KERGOSIEN, MCF, GERIICO, Université Lille 3	Encadrant

UNIVERSITÉ DE MONTPELLIER

Résumé

L'extraction d'information spatiale à partir de données textuelles est désormais un sujet de recherche important dans le domaine du Traitement Automatique du Langage Naturel (TALN). Elle répond à un besoin devenu incontournable dans la société de l'information, en particulier pour améliorer l'efficacité des systèmes de Recherche d'Information (RI) pour différentes applications (tourisme, aménagement du territoire, analyse d'opinion, etc.). De tels systèmes demandent une analyse fine des informations spatiales contenues dans les données textuelles disponibles (pages web, courriels, tweets, SMS, etc.). Cependant, la multitude et la variété de ces données ainsi que l'émergence régulière de nouvelles formes d'écriture rendent difficile l'extraction automatique d'information à partir de corpus souvent peu standards d'un point de vue lexical voire syntaxique.

Afin de relever ces défis, nous proposons, dans cette thèse, des approches originales de fouille de textes permettant l'identification automatique de nouvelles variantes d'entités et relations spatiales à partir de données textuelles issues de la communication médiée. Ces approches sont fondées sur trois principales contributions qui sont cruciales pour fournir des méthodes de navigation intelligente. Notre première contribution se concentre sur la problématique de reconnaissance et d'extraction des entités spatiales à partir de corpus de messages courts (SMS, tweets) marqués par une écriture peu standard. La deuxième contribution est dédiée à l'identification de nouvelles formes/variantes de relations spatiales à partir de ces corpus spécifiques. Enfin, la troisième contribution concerne l'identification des relations sémantiques associées à l'information spatiale contenue dans les textes. Les évaluations menées sur des corpus réels, principalement en français (SMS, tweets, presse), soulignent l'intérêt de ces contributions. Ces dernières permettent d'enrichir la typologie des relations spatiales définies dans la communauté scientifique et, plus largement, de décrire finement l'information spatiale véhiculée dans les données textuelles non standards issues d'une communication médiée aujourd'hui foisonnante.

Remerciements

En premier lieu, je tiens à remercier ma directrice de thèse Maguelonne TEISSEIRE d'avoir accepté de diriger cette thèse. Elle m'a fourni un cadre de travail idéal qui m'a permis de développer mes idées et de mener à bien mes travaux.

Je remercie très sincèrement mon co-directeur de thèse Mathieu ROCHE pour son soutien, son grand investissement et sa constante disponibilité durant toute cette période. Nos échanges m'ont aidé à cerner le domaine de fouille de textes et les défis relatifs à l'extraction d'information spatiale à partir de textes.

Je remercie également mon encadrant Eric KERGOSIEN pour sa contribution importante dans l'encadrement de cette thèse. Ses encouragements m'ont donné la motivation nécessaire pour surmonter mes jours de doute. Il a toujours su m'orienter vers des pistes de recherches pertinentes.

Je voudrais remercier les membres du jury Prof. Cédric FAIRON, Dr. Christian SALLABERRY et Prof. Carmen GERVET pour l'intérêt qu'ils ont porté à mon travail et d'avoir accepté d'évaluer ma thèse, pour le temps qu'il ont consacré à la lecteur de mon manuscrit et pour leurs commentaires perspicaces qui m'ont permis d'améliorer ma recherche de différentes perspectives.

Merci à tous mes amis et collègues de TETIS et de LIRMM, en particulier l'équipe ADVANSE, pour leur accueil et la très bonne ambiance au travail.

Un grand Merci également à Lynda Khiali pour son soutien et son encouragement durant toutes ces années.

Finalement, je tiens à remercier du plus profond de mon cœur mes parents Ahmed ZENASNI et Khadidja NEDJADI et toute ma famille. Merci de votre infaillible soutien. Je vous aime.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction et État de l'art

Chapitre

1 Introduction

Contents

1.1 Contexte et Motivations

L'extraction d'information est désormais un sujet de recherche important dans le domaine du Traitement Automatique des Langues Naturelles (TALN). Elle connaît ces dernières années un intérêt grandissant car elle répond à un besoin devenu incontournable dans la société de l'information.
Ce phénomène est encore plus notable ces dernières années avec le développement d'Internet, des communications par courriers électroniques et des communications par messages courts. En effet, les diverses avancées technologiques ont rendu la création, l'acquisition, le stockage et le partage de documents numériques de plus en plus accessibles pour les usagers. Dans ce sens, la communication médiée via des messages courts (SMS, tweets, messages de forums, etc.) est devenue un phénomène social dépassant les frontières. Des millions de messages sont échangés chaque jour pour communiquer, participer à des concours, obtenir des informations (localisation, opinion, etc.).
Face aux coûts élevés que peut représenter le traitement manuel d'un corpus important de textes même courts, les analyses automatiques tendent à rendre la tâche bien plus abordable, tant au niveau de l'effort, que du temps passé et des coûts financiers engendrés.

1.1.1 Contexte

Au fil des dernières décennies les données textuelles provenant des réseaux sociaux et des téléphones mobiles sont devenues une source cruciale d'information et cela dans de nombreux domaines, notamment le tourisme, l'économie, la culture, etc. Les services culturels des collectivités par exemple, qui organisent de nombreux évènements durant l'année (journée du patrimoine, festivals multi-sites, évènements sportifs dans plusieurs infrastructures, etc.) souhaiteraient vivement connaître de façon précise les dynamiques de déplacement durant ces évènements pour en améliorer l'organisation. Et cela passe par une analyse fine des messages courts exprimés par la population. Les professionnels de la sécurité sont également demandeurs de nouvelles méthodes pour identifier les informations spatiales exprimées dans les messages courts car cela pourrait permettre d'être informé très rapidement de catastrophes naturelles ou non naturelles avec une localisation précise du lieu sur lequel intervenir. Dans le secteur du tourisme, une analyse fine des informations spatiales exprimées sur les réseaux sociaux aiderait les professionnels du secteurs à identifier les lieux d'intérêts des visiteurs, autres que les lieux touristiques pour lesquels ils peuvent comptabiliser le nombre d'entrées, ou à l'inverse les lieux peu fréquentés qui nécessiteraient un effort supplémentaire pour les rendre attractifs.

Pour répondre à ce type de besoins, nous nous intéressons dans cette thèse à identifier et extraire les informations spatiales exprimées dans les corpus de messages courts.

Dans la communauté scientifique, de nombreux travaux liés à l'analyse de

l'information spatiale s'appuient sur l'exploitation de ces données textuelles à partir d'un ou plusieurs types de sources (articles de presse, documents techniques, revues, romans, etc.), éventuellement combinés à d'autres types de documents (images satellites, données GPS, etc.), afin d'améliorer l'efficacité des systèmes de recherche d'information, etc.

Toutefois, l'avènement des nouvelles technologies (Internet 2.0 et 3.0, téléphones portables) a fait évoluer les pratiques et notamment les modalités d'écriture dans les corpus de messages courts, et notamment les réseaux sociaux et les SMS. Et en outre, à notre connaissance, très peu de ces approches prend en compte les nouvelles formes d'expression présentes dans les messages courts de types SMS et tweets en langue française.

Nous distinguons ainsi ici deux types de texte : (i) le texte bien formé ou standard tels que les articles de presse et (ii) le texte bruité ou non-standard tels que les messages courts issus des SMS et des réseaux sociaux, type de textes sur lequel nous nous concentrons dans le cadre de ce travail de recherche. Dans ce contexte, les nouvelles formes d'expression des informations spatiales doivent être identifiées et converties en données plus structurées afin de faciliter leur analyse.

1.1.2 Problématique

Les corpus non-standards se caractérisent par l'utilisation de mots plutôt courts en raison du besoin d'exprimer le plus d'information possible dans un texte de taille limitée et/ou le plus rapidement possible. La multitude et la variété des corpus non-standards ainsi que l'émergence régulière de nouvelles formes d'expression (nouveau vocabulaire, présence de fautes d'orthographe, etc.) rendent difficile la reconnaissance et l'extraction automatique d'information. De ce fait, les méthodes classiques d'extraction d'information ne s'avèrent pas adaptées au traitement et à l'analyse de corpus de messages courts. En effet, en raison des particularités lexicales et syntaxiques de ce type de corpus, les performances dans la reconnaissance d'information et particulièrement dans la reconnaissance des informations spatiales diminuent nettement quelle que soit l'approche utilisée. De plus l'absence de ressources linguistiques spécialisées (lexiques, corpus d'apprentissage) pour ce type de données ne permet pas d'adapter les outils existants à ces particularités.

En effet, parmi les approches existantes, les approches par apprentissage supervisé pour l'extraction d'Entités Nommées (EN) donnent généralement de bons résultats mais celles-ci s'appuient sur des corpus annotés. Or, à notre connaissance, il n'existe pas de corpus de messages courts annotés, que ce soit en français ou en anglais.

Dans cette thèse, nous nous concentrons plus précisément sur les messages courts que sont les SMS et les tweets. À notre connaissance, il n'existe pas d'études traitant

l'extraction fine de l'information spatiale à partir de corpus de SMS spécifiquement en langue française.

1.1.3 Présentation des données

Dans le cadre de nos travaux, nous nous appuyons sur trois corpus distincts pour mener à bien nos expérimentations. Le premier corpus est un corpus de SMS nommé 88milSMS [1]. Ce corpus est composé de plus de 88,000 SMS authentiques en français qui ont été recueillis par une équipe pluridisciplinaire de linguistes et d'informaticiens issus de Montpellier en 2011 dans le cadre du projet Sud4Science Languedoc Roussillon (Panckhurst et al., 2014). Dans le but de garantir la mise à disposition de ce corpus, celui-ci a été entièrement anonymisé afin de masquer l'identité des individus et de supprimer les indications personnelles qui permettraient de les reconnaître (Patel et al., 2013). Les noms des personnes sont remplacés par les balises <PRE> qui représentent les prénoms, <NOM> qui représentent les noms, et <SUR> qui représentent les surnoms.

Le second corpus est un corpus de tweets qui inclut des données récupérées à l'aide de l'outil de Digital Methods Initiative appelé DMI-TCAT (DMI Twitter Capture and Analysis Toolset), destiné à capturer et à analyser les données de Twitter (Borra and Rieder, 2014) [2]. Le corpus de tweets contient plus d'un million de tweets, dont 811,871 sont en français. Les tweets ont été collectés en 2014 et sont liés à la zone géographique de la métropole de Lille. Nous nous sommes concentrés dans cette étude sur les tweets en langue française. Une première étude fut menée sur ce corpus pour l'analyse spatiale (hashtags et métadonnées de géolocalisation) dans (Severo et al., 2015). Nous nous appuyons sur ces deux premiers corpus non-standards de messages courts pour tester nos différentes contributions visant à extraire de nouvelles formes d'entités et relations spatiales.

Le troisième corpus est un corpus de 9,800 articles de presse mis à disposition par le Journal Midi Libre. Les documents de Midi Libre ont été collectés pendant la période $2010 - 2013$ et traitent de l'actualité sur le territoire du bassin de Thau (sud de la France). Ce troisième corpus sera utilisé dans nos expérimentations pour valider l'ensemble de nos approches sur un corpus standard.

Pour étayer l'ensemble des étapes de nos contributions dans les différents chapitres de ce mémoire de thèse, nous présentons ci-dessous un ensemble de 35 SMS et 20 tweets provenant des corpus de SMS (88milSMS) et tweets respectivement. Ils sont représentatifs de la diversité des formes d'expression des informations spatiales que l'on peut trouver dans les messages courts. Les termes en gras sont les entités et les relations sémantiques (spatiales et non-spatiales), exprimées de façon standard ou via des variantes.

(SMS 1) *Est ce que l'apéro est au frais ? Je suis* ***au niveau de Valence en route vers Montpellier****. Bonne continuation pour les vacances.*

1. http ://88milsms.huma-num.fr
2. Le corpus est récolté grâce au financement du GIS-CIST (axe médias et territoire).

(SMS 2) *Hey, tu arrives quand ? Tu restes* ***à Béziers*** *?*

(SMS 3) *Jeudi soir on a le concert de Dick Annegarn* ***à Lattes*** *! Je t'M plus que tout le* ***Massachussets*** *! < PRE_3/ >*[3]

(SMS 4) *Oui mais la je suis dans le tram et le temps d'arrive y sera 6 h ch ui* ***a odysseum***

(SMS 5) *Pffff ... Je sais pas ! Le temps de venir* ***vers frontignan*** *... Euh 21 heures 15 ...*

(SMS 6) *T'as le reçu la CAF ? Ben moi je rentre aujourd'hui* ***à sète****.*

(SMS 7) *Si* ***place 2 leurope****, é voui y va* ***a latte***

(SMS 8) *Je rentre* ***à Monptel'*** *finalement, j'ai eu accès au site pour les cours et y a une réunion de regroupement de lundi à jeudi.*

(SMS 9) *Je viens d'arriver. Je suis du côté de la sortie* ***vers Ste Eustache****. A toute*

(SMS 10) *Ya < PRE*$_4$[4]*/ > qui vient chez nous vers 14h15. Rentre, on ira* ***a odyseum*** *en voiture*

(SMS 11) *Salut ! Pas grave je suis avec < PRE*$_5$*/ > il part demain. Ce w-end je reste* ***a monpellier*** *mais le w-end prochain je serais* ***a Leucate*** *...*

(SMS 12) *Est ce que tu es dispo pour samedi 1 octobre pour gardez des enfants pour le comite de quartier de 15h a 18h ? Puis il y as < PRE*$_5$*/ > qui et* ***sur motpellier***

(SMS 13) *Ouii j'viendrais* ***a Bezier*** *faire tes petites soirees etudiantes !*

(SMS 14) *On s'est garé un peu à l'extérieur avant les déviations, on file vers le centre à* ***pattes****.*

(SMS 15) *Mais nan nan on part* ***à tahiti*** *il parait que c'est le paradis du ski !*

(SMS 16) *Quand je Prend une douche avec le gel douche* ***tahiti*** *J'ai l'impression d'avoir*

(SMS 17) *Y'a le* ***marrakech*** *du rire ce soir sur m6 ! Tu veux venir regarder ça à la maison ?*

(SMS 18) *Ah non c nul* ***Tlse*** *!Viens vivre* ***a Mtp*** *fait plus beau et y a la mer.*

(SMS 19) *Pu...* ***Dvant belevile*** *ya une riviere avc les ècler sa fai flipè*

(SMS 20) *Nous avons laissé à son stage de musique. et moi poursuivons* ***sur Montpellier****. Nous avons passé le 1er bouchon à* ***Valence*** *et roulons vers les prochains :* ***Montelimar*** *et* ***avant Montpellier !!!***

(SMS 21) *Nous sommes arretés sur la voie, juste* ***avant la gare***

3. Étiquettes PRE représente le prénom utilisé lors de la phase d'anonymisation.

4. Le chiffre 4 renvoie au nombre de caractères du prénom dans le SMS brut (e.g. Sarah <PRE_5/>).

(SMS 22) *Oui je dors bd voltaire et oui on passe* ***par st lazar****, pk ? Tu y va vers quelle heure toi ?*

(SMS 23) *Ça marche pour 19 heures ou plus tard. Rdv devant la grosse tête* ***près de l'église Ste Eustache*** *?*

(SMS 24) *Fin c'est pas sur car* ***collioure*** *c'est plus* ***pres de montpellier*** *:) t'es deja allée la bas ?*

(SMS 25) *C sur ! Je lui ai demandé ou il habitait exactement je le pensais pas* ***loin de gueret****. Il est* ***a coté de masgot a 20min de gueret*** *ms de l autre coté ! Je lui ai dit que ca faisait loin du coup si jms il devait venir me chercher. Il a dit qu un de ses amis etait encore plus loin et n hesitait pas a aller voir sa "chérie" s il fallait*

(SMS 26) *Normalement je vais l'avoir a 41. J'espère ! Je suis juste* ***a cote de la gare****, mais j'attend qu'on me fasse ma prise de sangg*

(SMS 27) *Tu crois tu pourrai me rejoindre au creps* ***à coté du stade philippides***

(SMS 28) *Ce soir y à une white party que des teissiers* ***à coté de l'église saint roch***

(SMS 29) *j'suis deg c l'anniv de mariage de <PRE_5/> en* ***suisse****, c prevu depuis 6mois ! ! !*

(SMS 30) *je* ***suis ds le parc***

(SMS 31) *Cern* ***à 731 km de distance en Suisse***

(SMS 32) *j'ai vu que <PRE_6/>* ***est originaire de Beaumont****, qui se trouve a 20 min de chez moi*

(SMS 33) *le <PRE_4/> soient rentabilisés en le faisant payer des impots s'il* ***bosse en France***

(SMS 34) *je dois recup mon frère sur la route il* ***bosse a portel***

(SMS 35) *Je suis* ***a 5 min de la gare***

(Tweet 1) *ptetre jpourrais aller* ***a montpelier*** *? ?*

(Tweet 2) *@calant42 et vous* ***du cote de st etienne*** *?*

(Tweet 3) *@Valentino_Greci c'était une salle* ***à st andre****, yavait 4-5 vigiles mais jsp c'etait quoi le bail*

(Tweet 4) *@Pokosphere62 moi je vais sûrement descendre en train* ***sur la gare de Lilli Flandre***

(Tweet 5) *Jv aller fr un tour* ***dvant gambetta*** *lundi*

(Tweet 6) *donc du coup on a du attendre facile 25min* ***dvant le palais de justice*** *mais on avait pas ldroit de sortir du bus*

(Tweet 7) *pecnologic puis mes amis qui ont ét'e dans d'autres villes* ***avant tokyo*** *ont vraiment détesté tokyo je suis pas des pires tbh ça va*

(Tweet 8) @*davmarouani @Bontemps78 si tu passes* ***par Lille*** *je te promets un donjon au top Et en exclu*

(Tweet 9) *Et il dit rien a ceux qui passent* ***par Lille*** *pr aller* ***a Neufchatel*** *la direction opposee bien sur Bref la sncf t'es pas mon amie aujourd'hui*

(Tweet 10) @*EddySS7 C'est Quand que tu viens* ***sur Lille***

(Tweet 11) @*Othigui nan je ne peux pas jr n'y habite plus je suis* ***a coté de Lille*** *maintenant c t la ville de mon enfance et coucou*

(Tweet 12) *t'habites dans une ville inconnue* ***a coté de Lille*** *kestu parles*

(Tweet 13) *je vais conduire de Lille à Aulnoye dans 5min je panique déjà à l'idée de voir des abrutis déboîter n'importe comment à la sortie de Lille*

(Tweet 14) *Distribution de bonbon à la sortie de la gare cooool*

(Tweet 15) *Demain je vais aller faire un ptit tour au nouveau magasin Nike* ***a cote du grand stade***

(Tweet 16) *Qui serais me depaner paypal qui habite* ***pres de valencienne*** *que je lui rendre l'argent en IRL SVP c'est urgent*

(Tweet 17) *J'ai vue un clown* ***A wasquehal pres de la gare*** *oú y a le tunnel Je rigole pas C'est la première fois j'en vois un*

(Tweet 18) *Malik il avait etait monstrueux* ***àu Sébastopol***

(Tweet 19) *Ark degouté il pleut* ***àu nord de la Françe***

(Tweet 20) *Et ce moment ou il a consolé Victoria* ***au bord de l'etang*** *enfin du lac ce tableau magnifique il est peint dans ma tête*

Les exemples **(SMS)** et **(Tweet)** représentent respectivement des extraits de SMS et de tweets provenant des corpus utilisés dans cette thèse. Nous distinguons dans ces exemples la présence des entités spatiales telles que « Montpellier, motpellier, la gare de Lilli Flandre » et « l'etang » (cf. (SMS 1), (SMS 12), (Tweet 4) et (Tweet 20)), des relations spatiales telles que « sur, par, dvant » et « avant » (cf. (SMS 12), (SMS 22), (Tweet 1) et (Tweet 3)), ainsi que des relations sémantiques telles que « est originaire de, bosse a » et « a 5 min de » (cf. (SMS 32), (SMS 34) et (SMS 35)).
Les corpus utilisés pour nos expérimentations étant décrits, nous allons maintenant présenter nos contributions que nous détaillons dans les chapitres suivants.

1.1.4 Contributions

Dans ce contexte, nous proposons une approche générique aux différents corpus traités (SMS, tweets et articles de presse).

Notre première contribution consiste à extraire de nouvelles formes d'expression d'entités spatiales absolues (ESA) dans des corpus de messages courts (par exemple, « motpellier » et « odyseum », voir exemples (SMS 12) et (SMS 10)).

L'identification de nouvelles variantes et/ou standards d'ESA ne permet pas toujours de découvrir et désambiguïser l'information spatiale. Par exemple, dans les SMS des exemples (SMS 19) et (SMS 20), l'identification de relations spatiales, respectivement « Dvant » et « sur » permet de préciser la localisation spatiale exprimée dans les messages. Dans ce sens, les relations spatiales sont importantes à prendre en compte pour compléter et valider les ESA. De ce fait, dans la deuxième partie de cette thèse (cf. Chapitre 4), nous proposons une approche permettant d'identifier de nouvelles formes d'expression de relations spatiales, écrites dans le langage des messages courts (par exemple, « Dvant » et « sur », voir exemples (SMS 19) à (SMS 20)).

En dépit des résultats encourageants, certaines relations ne sont pas encore identifiées (cf. Exemples (SMS 30) et (SMS 31)). En effet, les métriques que nous définissons dans la contribution présentée en chapitre 4 (cf. Section 4.2.1) s'appuient notamment sur la fréquence d'apparition des relations candidates présentes dans le corpus de messages courts. Or, par exemple, la relation « ds » (cf. Exemple (SMS 30)) apparaît une seule fois et nous verrons que notre approche ne permet pas de l'identifier et la valider comme information spatiale pertinente. Aussi, d'autres relations faisant intervenir une entité spatiale existent mais ne sont pas prises en compte dans cette contribution.

Afin de résoudre ce problème, dans la troisième contribution de cette thèse (cf. Chapitre 5), nous proposons une méthodologie pour identifier les relations sémantiques entre deux entités nommées avec la présence d'au moins une entité spatiale (par exemple, « est originaire de », « bosse a » voir les exemples (SMS 32) et (SMS 34)). Nous faisons l'hypothèse dans cette contribution que le fait d'utiliser des contenus externes volumineux provenant du Web nous permet de résoudre les problèmes rencontrés dans le chapitre 4, et facilite par la même occasion l'identification de nouvelles formes de relations exprimées autour des entités spatiales.

1.2 Organisation du mémoire

Outre la présente introduction qui expose le contexte, la problématique et les objectifs de cette thèse, le reste du mémoire est organisé de la façon suivante. Tout d'abord, le chapitre 2 présente l'état de l'art. La première partie de ce chapitre aborde les approches existantes qui traitent le problème de l'extraction et de l'identification d'information spatiale, i.e, des entités et des relations spatiales, à partir des documents textuels standards. Nous exposons également les limites de ces approches sur les différents types de corpus (non-standards). Nous détaillons ensuite dans la deuxième partie du chapitre 2 les travaux existants menés sur des corpus non-standards et nous présentons ensuite les limites de ces approches sur l'anglais, et plus spécifiquement sur le français. Finalement, nous concluons ce chapitre en présentant l'originalité de nos contributions permettant de prendre en compte les particularités des corpus de messages courts.

Les trois chapitres suivants (cf. Chapitres 3, 4 et 5) regroupent le détail de nos contributions visant à identifier de nouvelles entités et relations sémantiques (spatiales et non-spatiales) à partir des messages courts.

Le troisième chapitre (Zenasni et al., 2016a), (Zenasni et al., 2016b) et (Zenasni et al., 2018) (cf. Chapitre 3) présente notre première contribution qui consiste à extraire de nouvelles formes d'expression d'entités spatiales dans des corpus de messages courts. Nous y introduisons une nouvelle méthode exploitant différentes approches de traitement automatique du langage naturel, incluant l'analyse statistique et les caractéristiques lexicales. Nous montrons que la combinaison de ces deux dernières approches permet d'enrichir le dictionnaire d'entités spatiales avec un nouveau vocabulaire. Nous présentons, dans la suite du chapitre, les expérimentations menées pour évaluer notre approche. Les résultats obtenus montrent une amélioration notable des performances par rapport à celles existantes dans l'état de l'art.

Le quatrième chapitre (Zenasni et al., 2015), (Zenasni et al., 2016b) et (Zenasni et al., 2018) (cf. Chapitre 4) est constitué de deux parties détaillant notre deuxième contribution : (i) la prédiction du type de relations spatiales et (ii) l'identification de nouvelles variantes/formes de relations spatiales. La première partie présente une typologie de relations entre entités spatiales permettant d'identifier, finement et de manière automatique, les types des relations spatiales exprimés dans les textes. À cet effet, nous proposons une méthode hybride, combinant des informations lexicales et contextuelles à une approche de fouille de textes pour prédire le type de relation spatiale. Tandis que, dans la deuxième partie, nous proposons d'étudier la modalité de l'écriture des relations spatiales dans différents types de corpus tels que les corpus de messages courts. Ainsi, nous proposons une chaîne de traitements qui combine une analyse grammaticale et une approche de fouille de textes afin d'identifier les différentes variantes de relations spatiales à partir de corpus non-standards. Nous proposons également des patrons généraux pour identifier ces relations. Nous mesurons ensuite les performances de ces deux approches à l'aide des différents benchmark.

Le cinquième chapitre (cf. Chapitre 5) détaille la troisième de nos propositions ayant pour objectif d'identifier les relations sémantiques entre entités nommées dont une, au moins, est une entité spatiale. À cet effet, nous proposons, comme première étape, d'extraire l'ensemble des relations candidates associant les entités nommées et les entités spatiales. Puis, nous proposons, dans une deuxième étape, de les valider en exploitant les données issues du Web.

Finalement, le dernier chapitre du manuscrit (cf. Chapitre 6) résume les différentes propositions de nos travaux de thèse. Il synthétise nos contributions et discute les possibles améliorations qui peuvent y être apportées.

1.3 Publications

Ces différentes contributions ont donné lieu à des publications listées ci-dessous.

Revue internationale avec comité de lecture

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2018). *Spatial Information Extraction from Short Messages*. Expert Systems With Applications, Elsevier, Vol : 95, pp : 351- 367, IF 3.93.

Chapitre d'ouvrage international avec comité de lecture

- Cédric Lopez, Sarah Zenasni, Eric Kergosien, Ioannis Partalas, Mathieu Roche, Maguelonne Teisseire and Rachel Panckhurst. (2018). *Extracting Absolute Spatial Entities from SMS : Comparing a Supervised and an Unsupervised Approach*. In CMC and Language, special issue in the Cahiers du Cental (UCL).

Actes de conférences internationales avec comité de lecture

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2016). *Extracting new spatial entities and relations from short messages*. In Proceedings of the 8th International Conference on Management of Digital EcoSystems MEDES'16, ACM, pages 189–196
- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2015). *Discovering Types of Spatial Relations with a Text Mining Approach*. In Proceedings of 22nd International Symposium on Methodologies for Intelligent Systems ISMIS'15, LNCS, Springer, pages 442–451

Actes de conférences nationales avec comité de lecture

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2016). *Découverte de nouvelles entités et relations spatiales à partir d'un corpus de SMS*. Actes de la conférence TALN'2016, pages 403-410

Atelier

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2016). *Identification automatique des types de relations spatiales dans les textes*. Atelier Gestion et Analyse des données Spatiales et Temporelles GAST'16, organisé dans le cadre de la conférence EGC'16.

Communications orales

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2015). *Une approche de fouille de textes pour l'identification automatique de relations spatiales*. Big Data Mining and Visualization (journées communes aux groupes de travail EGC et AFIHM)
- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2015). *Fouille de relations spatiales : Comment les descripteurs linguistiques caractérisent les relations entre entités spatiales*. Session Poster au École thématique CNRS FOCOLISE.
- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2015). *Fouille de relations spatiales : Comment les descripteurs linguistiques caractérisent les relations entre entités spatiales*. Doctoriales -TETIS - La journée des doctorants de la Maison de la Télédétection.

Données produites

- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2017). *A corpus of 1000 authentic SMS in French with spatial labels*, doi :10.18167/DVN1/0ZGJRC - Dataverse
- Sarah Zenasni, Eric Kergosien, Mathieu Roche and Maguelonne Teisseire. (2017). *Dic-ES : Liste d'entités spatiales en français*, doi :10.18167/DVN1/LPY080 - Dataverse

Chapitre

2

État de l'art

Contents

2.1 Contexte

Dans ce chapitre, nous détaillons la problématique liée à l'extraction d'informations spatiales à partir de données textuelles. Nous présentons dans cette section les motivations propres à l'extraction d'informations spatiales ainsi que les concepts associés. Puis, en Sections 2.2 et 2.3, nous évoquons les approches mises en œuvre pour résoudre le processus d'identification automatique des éléments caractérisant une information spatiale (entités et relations). Ensuite, nous décrivons en Section 2.4 les méthodes liées à l'extraction d'information à partir de corpus complexes et non-standards issus des communications médiées aujourd'hui foisonnantes.

Au fil des dernières décennies, les données textuelles ont constitué une source cruciale d'information dans de nombreux domaines (notamment le tourisme, l'économie, l'histoire, etc.). Dans ce contexte, de nombreux travaux propres à l'analyse de l'information spatiale s'appuient sur l'exploitation de ces données textuelles à partir d'un ou plusieurs types de sources (articles de presse, documents techniques, revues, romans, etc.), éventuellement combinés à d'autres types de documents (images satellites (Salas et al., 2014), données GPS (Rikitianskii et al., 2014), etc.). En outre, de nombreuses approches s'intéressent à l'amélioration des systèmes de recherche d'information (Kergosien et al., 2017 ; Lesbegueries, 2007).

Dans la communauté scientifique, l'information géographique est définie comme la composition d'une information spatiale, temporelle et thématique (Gaio, 2001 ; Usery, 1996). L'idée principale est que la combinaison de ces trois informations permet de décrire un évènement qui se déroule ou s'est déroulé dans un lieu donné, à un moment donné (Nguyen, 2012). Parmi les trois composantes de l'information géographique, nous nous intéressons dans cette thèse au traitement de l'information spatiale.

De nombreux travaux ont décrit la manière particulière d'exprimer l'information spatiale en langage naturel. Lesbegueries (2007) a défini l'information spatiale, exprimée dans un texte, par au moins une entité spatiale (ES), qui consiste en une ou plusieurs entités nommées (EN) de type lieu (par exemple, « Montpellier », « Église Saint-Paul », etc.), et d'un nombre variable d'indicateurs spatiaux (par exemple, « près de », « au nord de », etc.). Ces indicateurs spatiaux sont également connus sous le nom de relations spatiales (RS).
Les ES simples telles que « Montpellier », « Église Saint-Paul » sont définies, dans le modèle Pivot généré par Lesbegueries (2007), sous le nom d'entités spatiales absolues (ESA), tandis que les ES complexes intégrant les RS telles que « au nord de Montpellier », « près de l'église Saint-Paul » sont appelées entités spatiales relatives (ESR).

Notons que les travaux de Jones et al. (2002) ont également présenté l'information spatiale, dans le cadre du projet **SPIRIT** [1], par un nom de lieu (i.e. *place*

1. Spatially-Aware Information Retrieval on the Internet (http ://www.geo-spirit.org).

name) et une relation spatiale (par exemple, « close to », « North of », etc.).

Dans le cadre du projet **GéoSem**, Bilhaut et al. (2007) ont défini l'information spatiale par des noms d'Entités Géographiques (*villes, régions, fleuves et autres entités géoréférencées*). Par exemple, « Paris » et « Bretagne » ainsi que les entités géoréférencées « la région Rhône-Alpes », « les départements de la grande banlieue parisienne » sont considérés. Cependant, comme l'information spatiale ne se réduit pas à des entités géographiques, les auteurs ont également intégré des relations spatiales (e.g. « près de », « au nord de », « dans », etc.) afin de préciser plus finement les localisations.

Nous détaillons par la suite les travaux liés à la détection des entités spatiales (cf. Section 2.2) et des relations spatiales (cf. Section 2.3) à partir de corpus standards (e.g. articles de presse, articles Wikipédia, etc.) et non-standards (e.g. SMS, tweets, etc.) (cf. Section 2.4).

2.2 Les Entités Spatiales (ES) et leur identification

Cette section explore les recherches concernant le processus d'extraction d'ES contenues dans un texte standard. Cette problématique intéresse de nombreux chercheurs dans divers domaines, notamment en recherche d'information, en reconnaissance d'entités nommées (REN), etc.

2.2.1 Les entités nommées (EN) et les ES

Le concept d'entité nommée (EN) a été introduit dans les années 90, dans le cadre des conférences MUC (Message Understanding Conference) (Grishman and Sundheim, 1996) afin d'encourager le développement de nouvelles méthodes d'extraction d'information. La communauté MUC a développé la tâche EN, qui consiste essentiellement à identifier des noms propres dans les textes (noms de personnes, noms d'organisations ou noms de lieux), des expressions temporelles et des expressions numériques (monétaires ou pourcentages).

Rappelons qu'une EN de type lieu peut se référer à des concepts topographiques (par exemple, *rivière, rue*, etc.) et un nom toponymique (par exemple, « Paris », « gare Saint-Roch », etc). Dans la littérature, le vocabulaire utilisé est diversifié pour identifier les noms de lieux, par exemple, on parle *de toponymes, de noms de lieux, de lieux géographiques, de geo-entités, d'entités spatiales* et *d'entités spatiales absolues*. Dans ce manuscrit de thèse, le terme ESA sera privilégié (Lesbegueries and Loustau, 2006).

Dans la section qui suit, outre la présentation des techniques existantes d'extractions d'ES, nous détaillons également les techniques relatives à la REN à partir des

documents textuels.

2.2.2 Les méthodes d'identification des ES à partir de textes

La reconnaissance des ES à partir des documents textuels peut être catégorisée en trois principales familles d'approches : des méthodes à base de règles, des méthodes d'apprentissage et des méthodes hybrides (Wu et al., 2006).

Méthodes à base de règles

Les méthodes à base de règles se concentrent sur l'extraction d'information à l'aide d'un ensemble de règles (par exemple des règles morpho-syntaxiques) en combinant avec des éléments issus de dictionnaires plus ou moins spécialisés (par exemple, une liste de *noms de pays, de villes*, etc.) préalablement définis par des experts (Mansouri et al., 2008) ou de gazetteers. Un gazetteer est un dictionnaire de toponymes (*noms de lieux*) qui sont liés à des concepts et la plupart du temps à leur empreinte géographique (Hill, 2000). De nombreux gazetteers sont disponibles en ligne, par exemple : Geonames [2], Getty Thesaurus of Geographic Names [3], etc.

Parmi les approches à base de règles, nous pouvons citer plusieurs approches caractéristiques de la littérature.
Wakao et al. (1996) ont décrit un système d'extraction d'information (LaSIE « LArge Scale Information Extraction ») dans lequel quatre catégories d'entités (organisation, personne, nom de lieu et expression temporelle) sont reconnues et classées. LaSIE est basé sur un étiquetage morpho-syntaxique du texte et une grammaire locale axée sur les noms propres. Cette dernière exploite la connaissance graphique, syntaxique, sémantique ainsi que le niveau de discours pour la reconnaissance et la classification des noms propres. L'approche a été testée sur 30 textes en anglais du Wall Street Journal. Les scores globaux de précision et du rappel sont respectivement 93% et 91%.

Stern and Sagot (2010) ont décrit NP (du Noms Propres français), un système pour REN qui suit deux étapes : une pour la détection et le typage, et l'autre pour la désambiguïsation et la résolution des EN. Une grammaire indépendante du contexte comprenant 130 règles a été développée pour détecter et typer des EN. Ensuite, des heuristiques de désambiguïsation fondées sur des informations quantitatives et qualitatives ont été appliquées afin de réduire l'ambiguïté. Le système NP a atteint un score de F-mesure de 77% sur un corpus de fil d'actualité français annoté manuellement auprès de l'Agence France-Presse.

Maurel et al. (2011) ont présenté le système open-source CasEN dédié au traitement des textes en français. La cascade CasEN utilise des ressources lexicales et des grammaires locales (transducteurs), exécutées dans un ordre précis, agissant sur un

2. http ://www.geonames.org
3. http ://www.getty.edu/research/tools/vocabularies/tgn/index.html

ensemble de textes pour la REN. Le principe d'une cascade est de pouvoir utiliser dans les descriptions suivantes les motifs déjà détectés ou, au contraire, d'éviter un étiquetage non souhaité pour un motif déjà reconnu. CasEN permet de construire les grammaires locales selon le contexte en utilisant le système CasSys [4] (Friburger and Maurel, 2004) intégré à la plate-forme Unitex (Paumier, 2003). La plate-forme Unitex permet une écriture aisée des transducteurs sous forme de graphes. Unitex est un ensemble de logiciels permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de lexiques (Paumier, 2003).

Notons que la plupart des systèmes à base de règles utilisent des règles qui prennent en compte le contexte morpho-syntaxique associé aux EN candidates (Maynard, 2003 ; Sallaberry et al., 2009) selon les corpus et les langues dédiées (anglais (Wakao et al., 1996), français (Maurel et al., 2009), polonais (Savary and Piskorski, 2010), malais (Alfred et al., 2014), arabe (Btoush et al., 2016), etc.). Les résultats peuvent différer selon les langues étudiées. Par exemple, sur le français (Maurel et al., 2009), l'anglais (Wakao et al., 1996) et le malais (Alfred et al., 2014), les auteurs ont obtenu des scores de F-mesure de 90%, 92% et 89% respectivement.

Ainsi, les approches à base de règles de la littérature sont des tâches qui prennent du temps car les règles sont principalement générées manuellement par des experts en sciences du langage (Pooja Pandey, 2016). Ce type d'approche peut aboutir à des résultats pertinents, bien qu'ils soient souvent dépendants des langues et des domaines (Mansouri et al., 2008). De manière générale, les approches à base de règles permettent d'obtenir une bonne précision en raison de leur spécificité. Cependant, si la généralisation échoue, on obtient un faible rappel, ce qui ne permet pas d'avoir une extraction exhaustive d'EN.

Méthodes d'apprentissage

Les méthodes d'apprentissage utilisent des modèles de classification afin d'identifier les EN (supervisés ou non-supervisés). Les méthodes supervisées reposent sur l'utilisation d'un échantillon de données étiquetées pour apprendre un modèle qui sera par la suite utilisé pour déduire la classe d'appartenance de nouvelles données non étiquetées.

- **Méthodes d'apprentissage supervisé**

  Au cours des dernières années, plusieurs méthodes statistiques basées sur l'apprentissage supervisé ont été proposées pour les tâches de REN. Elles utilisent des méthodes d'apprentissage de l'état de l'art, par exemple le modèle de Markov Caché (HMM) (Bikel et al., 1997), le modèle de l'entropie maximale (EM)

4. CaSys est un programme de création de cascades de transducteurs à états finis.

(Borthwick et al., 1998), les arbres de décision (Borthwick et al., 1998), les machines à vecteurs de support (SVM) (Isozaki and Kazawa, 2002) et l'approche CRF (Conditional random field) (Béchet and Charton, 2010 ; Dinarelli and Rosset, 2011 ; Hatmi et al., 2013 ; Raymond, 2013).

Florian et al. (2003) ont proposé un cadre expérimental de combinaison de classifieurs pour la reconnaissance d'EN dans lequel quatre différents classifieurs statistiques d'EN (classifieur linéaire, entropie maximale, apprentissage basé sur la transformation et modèle de Markov Caché) ont été combinés dans des conditions différentes. Sur les données en anglais, le système combiné atteint une performance de F-mesure de 92%.

Stanford NER (Finkel et al., 2005) identifie trois classes de base d'entités nommées (personne, organisation et nom de lieu) en utilisant des modèles de séquences CRF pour les identifier.
Finkel et al. (2005) ont proposé une approche qui exploite des structures non locales [5] pour améliorer la REN. Les auteurs ont utilisé Gibbs sampling (Geman and Geman, 1984) pour étendre un système d'extraction d'informations basé sur des CRF avec des modèles de LDA (Long-Distance Dependencies). Les modèles non locaux ont obtenu une valeur de F-mesure de 86% et 92% pour les deux corpus anglais, CoNLL-2003 [6] et CMU Seminar Announcements [7] respectivement.
Mansouri et al. (2008) ont proposé une méthode fondée sur l'apprentissage supervisé en utilisant l'algorithme de machine à vecteurs de support (SVM) appelé Fuzzy Support Vector Machine (FSVM) pour la REN. Tout d'abord, les ensembles de données (tests et entraînement) ont été segmentés. Pour les données d'entraînement, les auteurs ont utilisé une classification SVM linéaire pour prédire la classe de chaque EN (noms de personnes, noms de lieux, etc.). Puis, FSVM a été appliqué pour reconnaître la classe exacte des EN.
Raymond and Fayolle (2010) ont utilisé différents algorithmes d'apprentissage automatique (CRF, SVM, et transducteurs à états finis (FST)) pour la reconnaissance d'entités nommées dans les transcriptions de la parole. L'approche a obtenu des résultats pertinents sur les données d'évaluation de la campagne ESTER 2 [8] (Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques) (Galliano et al., 2009).

5. En prenant uniquement en compte les structures locales, il n'est pas clair si l'entité « Tanjug » dans les expressions « airport , **Tanjug** said » et « the news agency **Tanjug** reported » est une personne ou une organisation.

6. http ://www.cnts.ua.ac.be/conll2003/ner/

7. http ://www.cs.cmu.edu/ dayne/SeminarAnnouncements/__Source__.html

8. La campagne d'évaluation ESTER 2 visait à évaluer les performances des systèmes de transcription de la parole, les performances des systèmes de segmentation en tours de paroles et la capacité à extraire automatiquement des informations, en particulier les entités nommées (EN) (Brun and Ehrmann, 2010).

Tkachenko and Simanovsky (2012) ont construit un système basé sur les CRF, qui a atteint une F-mesure de 91% sur le corpus de CoNLL-2003 et une F-mesure de 81% sur l'ensemble des données de CNN OntoNotes version 4 (Hovy et al., 2006) (corpus en anglais).
Arulanandam et al. (2014) ont présenté une méthodologie pour extraire des phrases de localisation de crime à partir d'articles de journaux en ligne. Les auteurs ont utilisé des algorithmes de REN pour identifier les localisations dans les phrases et un modèle CRF a été appliqué pour identifier les phrases propres à une localisation de crime. Le système proposé a été évalué sur quatre journaux anglais différents provenant de trois pays. Les résultats obtenus pour les quatre ensembles d'articles varient de 73% à 90% en terme d'exactitude (accuracy).

Les méthodes de classification représentent les principales approches du domaine de l'apprentissage supervisé. Ces approches possèdent, cependant, l'inconvénient majeur de rendre obligatoire l'obtention préalable d'un large échantillon de données étiquetées pour la phase d'apprentissage. C'est pourquoi les approches non-supervisées et semi-supervisées sont également largement mobilisées. Elles représentent des alternatives intéressantes pour le traitement automatique de grandes masses de données.

- **Méthodes d'apprentissage non- et semi-supervisé**

Les principales méthodes d'apprentissage non-supervisé utilisées dans les systèmes d'Extraction d'Information cherchent à apprendre directement à partir de textes bruts ou se fondent sur une petite base d'exemples dans le cas d'apprentissage semi-supervisé.
Par exemple, Lample et al. (2016) ont proposé deux modèles LSTM-CRF et S-LSTM (stack LSTMs) pour la tâche de REN. LSTM-CRF apprend un LSTM bidirectionnel sur une séquence de mots donnée (représentée en tant que vecteur de mots « word embedding [9] »). Puis, un modèle CRF, qui utilise les unités cachées comme caractéristiques, est utilisé pour prédire la séquence d'étiquettes de sortie. S-LSTM construit et marque des segments en utilisant une approche fondée sur la transition. Leur approche a été expérimentée sur les jeux de données CoNLL-2002 [11] et CoNLL-2003 [12] pour les langues anglaise, espagnole, allemande et néerlandaise. Par exemple, pour la langue anglaise, les EN détectées à partir de la phrase « John Smith went to Pittsburgh » sont « John_Smith$_{[PER]}$ went$_{[0]}$ to$_{[0]}$ Pittsburgh$_{[LOC]}$ ».

9. Word embedding est la représentation vectorielle des mots d'un texte. Elle capture la similarité syntaxique et sémantique entre les mots. Un exemple de word embedding est word2vec [10] (Mikolov et al., 2013).
11. http ://www.cnts.ua.ac.be/conll2002/ner/
12. Http ://www.cnts.ua.ac.be/conll2003/ner/

Etzioni et al. (2005) ont décrit KnowItAll [13], un système non-supervisé indépendant du domaine qui extrait des EN à partir du Web. KnowItAll prend comme entrée un ensemble de noms de prédicats (e.g. *City, Country, capitalOf (City,Country)*, etc.) et amorce son processus d'extraction à partir d'un ensemble de motifs d'hyponymes [14] (Hearst, 1992). Finalement, KnowItAll teste automatiquement la plausibilité des faits candidats qu'il extrait en utilisant des statistiques d'informations mutuelles (PMI) (Turney, 2001) calculées en traitant le Web comme un corpus de texte massif. Par exemple, en soumettant à un moteur de recherche la requête « and other cities » provenant du patron généré « NP and other cities », les auteurs ont détecté la ville « Fes » issue de la phrase « Short flights connect Casablanca with Fes and other cities ».

Al-Rfou et al. (2015) ont présenté un outil nommé Polyglot qui propose d'annoter les EN sur 40 langues principales. Ils ont proposé une approche semi-supervisée qui utilise du word embeddings (Polyglot embeddings [15] (Al-Rfou et al., 2013)) comme descripteur pour chaque langue étudiée. Les auteurs ont proposé un modèle d'apprentissage discriminant (Discriminative Learning) en utilisant un réseau de neurones (une couche cachée) qui prend en entrée la représentation distribuée du word embeddings et qui donne en sortie une distribution de probabilité d'appartenance à une classe (e.g. *location*, i.e. lieu en français). Pour l'apprentissage du système, les auteurs ont généré automatiquement un corpus à partir de Wikipédia. Pour cela, ils ont utilisé Freebase [16] (Bollacker et al., 2008) pour associer une classe (e.g. *person, location*, etc.) à chaque page Wikipédia. Ensuite, ils ont annoté chacune des pages de Wikipédia en utilisant leurs liens internes. Cependant, en raison des règles d'écriture imposées par Wikipédia, seule la première apparition d'une EN est associée à un lien interne. Par conséquent, les auteurs ont proposé des méthodes afin de résoudre le problème des annotations manquantes. Par exemple, après avoir appliqué le système Polyglot sur la phrase « La *France* veut satisfaire à ses engagements envers l'*Union européenne* », les auteurs ont détecté « France » comme un lieu et « Union européenne » comme une organisation.

Contrairement à l'apprentissage non-supervisé, l'apprentissage supervisé peut fournir des résultats plus précis car les modèles sont appris à partir de corpus spécifiques et adaptés aux domaines et types de textes étudiés.

13. http ://projectsweb.cs.washington.edu/research/knowitall/

14. Par exemple, le motif générique « NP1 such as NPList2 » indique que la tête de chaque expression nominale simple « NP City, Country, capitalOf (City,Country) » dans la liste « NPList2 » est un membre de la classe nommée dans « NP1 ».

15. Les vecteurs Polyglot embeddings sont entraînés sur Wikipédia. Le vocabulaire de chaque langue est composé de $100K$ mots les plus fréquents, chaque mot est associé à un vecteur de 64 dimensions.

16. Freebase est une base de données ouverte qui couvre un grand nombre et une diversité de données (Bollacker et al., 2007).

Méthodes hybrides

Les approches hybrides combinent les méthodes à base de règles et les méthodes d'apprentissage. Mikheev et al. (1999) ont proposé un système de REN qui combine des grammaires locales avec des modèles statistiques. La méthode proposée applique des règles grammaticales qui se fondent sur des désignateurs d'entreprises connues (par exemple, *Ltd.*[17], *Inc.*, etc.), les titres de la personne (par exemple, *Mr., Dr., Sen.*), et des contextes précis (par exemple, « in Washington » au lieu de seulement « Washington »). Ensuite, le système collecte toutes les EN identifiées dans le document. Puis, pour chacune des EN qui se compose de plusieurs mots, il génère toutes les combinaisons possibles des mots auxquels sont affectés une même classe (e.g. personne, organisation, etc.). Par exemple, les combinaisons « Kluver Ltd », « Adam Ltd » et « Adam Kluver » sont étiquetées comme des organisations si à l'étape précédente « Adam Kluver Ltd. » est identifié comme une organisation. Une méthode de classification d'entropie maximale a été appliquée ensuite pour que le système effectue une affectation définitive. Finalement, des règles plus simples, fondées sur les données annotées pendant la phase précédente, ont été appliquées, suivies d'un deuxième passage de classification d'entropie maximum. Le système, qui a été construit pour la tâche MUC-7[18], a obtenu un score de 93% en terme de F-mesure.

Nagesh et al. (2012) ont présenté une approche qui facilite le processus de construction de règles personnalisables pour la tâche de REN via l'induction de règles dans l'AQL (Annotation Query Language). La contribution principale de leur travail est de générer un ensemble de règles qui peuvent être personnalisées par les humains. Le système s'appuie alors sur un ensemble de règles de base comprenant des dictionnaires et des expressions régulières et un corpus de documents annotés. Finalement, les auteurs ont introduit une mesure quantitative pour calculer la complexité de leur système. Les auteurs ont expérimenté leur approche sur CoNLL-2003. Le système a obtenu un score de 69% en terme de F-mesure.

2.2.3 Discussion

Comme abordé dans cette section, la reconnaissance des entités nommées, en particulier la reconnaissance des entités spatiales, joue un rôle important dans le traitement automatique du langage naturel et également dans l'extraction automatique d'information.
Plusieurs méthodes ont fait leurs preuves pour la phase de reconnaissance d'entités spatiales ou dans le contexte plus global d'identification des entités nommées. Parmi ces travaux, des plateformes dédiées au développement d'architectures visant à extraire les EN et les ES ont été développées, telles que GATE[19], LinguaStream

17. Par exemple, si une séquence de mots en majuscules se termine par le mot « *Ltd.* », ce dernier est qualifié d'être un nom d'organisation.
18. https ://catalog.ldc.upenn.edu/LDC2001T02
19. https ://gate.ac.uk/projects.html

(Widlöcher and Bilhaut, 2005), ou encore UMIA [20]. En complément de ces méthodes, certains autres outils ont été développés pour extraire les EN et les ES à partir de ressources publiées sur le Web. Ces outils ont été transformés en services Web tels que AlchemyAPI [21], DBpedia Spotlight [22], OpenCalais [23], NERD [24], etc. Ils fournissent généralement une sortie similaire composée d'un ensemble d'entités nommées extraites et leur type (Rizzo and Troncy, 2011).
Ces méthodes sont efficaces mais elles ne permettent pas toujours d'identifier l'information spatiale présente dans le texte. En effet, la limite de ces approches consiste en leur incapacité à analyser finement les interactions spatiales présentes dans des documents textuels. À titre d'exemple, lorsqu'un texte évoque « les villes au sud de Montpellier » l'information est clairement spatiale. A contrario, avec l'expression « la mégalopole de Rio est jumelée avec la ville Montpellier », l'information véhiculée est économique et/ou politique sans information spatiale évidente. Par ailleurs, ces méthodes ne permettent pas de résoudre des problèmes liés aux tâches d'extraction et de raisonnement intégrant la spatialité.

Aussi, il semble crucial de proposer une vue plus complète de l'information spatiale contenue dans le texte en prenant en compte également les relations spatiales présentes dans le texte.
Ces relations servent à définir l'entité spatiale selon qu'elle soit à l'intérieur, à l'extérieur ou près de l'entité nommée (Lesbegueries, 2007).
L'extraction des relations spatiales à partir des documents textuels peut jouer un rôle important pour enrichir l'information spatiale extraite et pour permettre une meilleure appréhension du contenu véhiculé dans les documents textuels.

Dans la section suivante, nous présentons les travaux liés à l'identification des relations spatiales ainsi que la classification des relations sémantiques (spatiales et non-spatiales) qui sont susceptibles de relier les entités nommées entre elles.

2.3 Les Relations Spatiales (RS) et leur identification

Comme nous l'avons vu précédemment, une relation spatiale est un indicateur spatial qui fait référence aux relations entre les entités spatiales. Ces indicateurs sont des adverbes spatiaux, des verbes de déplacement, etc. utilisés pour localiser par exemple les évènements (Nguyen, 2012) (e.g. « près de », « au nord de », etc.). La relation spatiale a été étudiée dans la littérature, en étant formulée de différentes manières, à savoir *les indicateurs spatiaux, les prépositions spatiales, les relations spatiales et les entités spatiales relatives*. Un état de l'art des travaux associés est

20. https ://uima.apache.org/
21. http ://www.alchemyapi.com
22. http ://demo.dbpedia-spotlight.org
23. http ://www.opencalais.com
24. http ://nerd.eurecom.fr

présenté dans cette section.

2.3.1 Les relations sémantiques et spatiales

L'extraction de relations consiste, généralement, à détecter et typer une relation spatiale. Ceci permet de redéfinir (restriction ou élargissement) la zone spatiale désignée par l'ESA associée. D'autres travaux s'appliquent à détecter dans les textes les liens exprimés entre deux entités (Serrano, 2014) afin d'améliorer la compréhension des textes notamment. Dans ce sens, les relations sémantiques entre paires de mots sont intéressantes pour l'interprétation d'une phrase, l'analyse du discours, etc. (Hendrickx et al., 2009).
Des tâches d'extraction de relations ont notamment été proposées lors de la campagne MUC-7. Trois classes de relations ont été intégrées à MUC-7 : la classe *location of* qui présente les relations entre un lieu et une organisation, la classe *employee of* qui présente les relations entre une personne et une organisation et la classe *product of* qui présente les relations entre un artefact et une organisation (Chinchor and Marsh, 1998).

À partir des années 2002, le programme ACE [25] propose d'extraire les relations entre deux entités nommées. Ces relations sont portées sur cinq classes. Par exemple, la classe « At » présente les relations qui indiquent l'emplacement d'une personne ou d'une organisation à un endroit donné, la classe « Near » présente les relations qui indiquent la proximité d'un endroit à l'autre, la classe « Role » présente les relations qui lient une personne à une organisation, etc. Certaines des cinq classes sont encore subdivisées ce qui donne 24 sous-classes, notamment la classe « Role » qui est subdivisée en plusieurs sous-classes telles que « Membre », « Owner », « Founder », « Client », etc. (Doddington et al., 2004).

Par ailleurs, dans SemEval-2010 (Hendrickx et al., 2009), les auteurs ont introduit la tâche de reconnaissance des relations sémantiques entre deux expressions nominales (par exemple, les entités nommées). Dans cette tâche, les auteurs considèrent un ensemble de neuf classes de relations sémantiques. Parmi elles, la classe *Entity-Destination* présente les relations de déplacement d'une entité vers une destination, la classe *Content-Container* présente les relations du stockage ou d'inclusion physique d'un objet dans une zone délimitée de l'espace, etc.

Dans SemEval-2015 (Pustejovsky et al., 2015), une tâche SpaceEval a été introduite dans le but d'identifier et de classer les informations à partir d'un inventaire de concepts spatiaux. La tâche SpacEval se compose de six sous-tâches, trois d'entre elles sont liées aux relations spatiales. Ces trois sous-tâches sont : (i) *Motion Relation Identification (MoveLink)* sont des relations spatiales définies sur des objets

25. Automatic Content Extraction (ACE) est un programme de recherche pour le développement de technologies avancées d'extraction d'information. En général, le programme ACE est motivé par les mêmes problèmes que le programme MUC. Ce programme permet d'extraire automatiquement les entités, les relations entre ces entités et les événements auxquels ces entités participent en langage naturel (Doddington et al., 2004).

en mouvement ; (ii) *Spatial Configuration Identification (QSLink)* sont des relations spatiales définies entre des éléments spatiaux stationnaires avec une connexion régionale ; et (ii) *Spatial Orientation Identification (OLink)* sont des relations spatiales définies entre des éléments spatiaux stationnaires exprimant leurs orientations relatives ou absolues.

Dans ce contexte, une relation spatiale (e.g. *location of*, *Near*, etc.) est une relation sémantique spécifique entre deux EN. Les relations spatiales spécifient la manière dont certaines entités spatiales sont liées dans l'espace à d'autres. Il inclut les abstractions des entités spatiales physiques et la structure de l'espace (Chen et al., 2015). Selon le modèle pivot (Egenhofer, 1991 ; Lesbegueries, 2007), les relations spatiales entres les entités peuvent être classifiées en cinq catégories, à savoir l'inclusion (« dans », etc.), la direction (« au nord », etc.), l'adjacence (« à côté de », etc.), la distance (« à X km de », etc.) et la relation géométrique (« entre X et Y », etc.).

L'objectif de l'identification des relations spatiales est de résoudre de nombreuses tâches comme la réduction de l'ambiguïté, l'extraction plus fine d'informations géographiques, les systèmes de navigation, la gestion du trafic, le raisonnement spatial, la réponse aux requêtes, etc.

2.3.2 Les méthodes d'identification des RS à partir de textes

L'identification de relations spatiales a été étudiée dans plusieurs travaux. Nous citons Nguyen et al. (2010) qui ont proposé une méthode basée sur une approche lexico-syntaxique. La méthode s'appuie sur l'identification des relations n-aires pour l'enrichissement d'une ontologie géographique à partir de l'analyse automatique d'un corpus textuel. À partir de ces relations et d'un lexique de termes géographiques, des relations spatio-temporelles sont identifiées dans un contexte de descriptions d'itinéraires (e.g. *Le frère de mon ami a quitté Pau, pour une ville près de Lyon, depuis deux semaine*). L'idée proposée est de traiter l'ensemble des phrases qui comportent une relation n-aire afin d'identifier un verbe de déplacement ainsi que les entités spatiales et temporelles. Dans ce cas, une relation spatio-temporelle entre ces entités est alors constituée.

Blessing and Schütze (2010) ont proposé une méthode pour annoter automatiquement les relations entre les entités spatiales en s'appuyant sur des sources de données infobox de Wikipédia - Allemagne. Les auteurs ont utilisé une approche d'apprentissage supervisé en y intégrant différents descripteurs linguistiques. Notons que ce type d'approche dépend fortement de la qualité et de la quantité de données annotées dans la phase d'apprentissage. Les auteurs ont proposé une nouvelle approche (auto-annotation). La tâche d'extraction de relations est effectuée par un classifieur supervisé (ClearTK (Ogren et al., 2008)). Ce classifieur est implémenté comme composant UIMA (Unstructured Information Management Architecture) (Hahn et al., 2008).

Loglisci et al. (2012) ont proposé d'identifier les relations spatiales de type topologique, directionnel ou distance entre deux géo-entités à partir d'un fragment textuel d'un document, où les deux géo-entités ont été reconnues (par exemple, « The Thames flows through London », où « Thames » et « London » sont les deux geo-entités). Plus précisément, les auteurs ont proposé de récolter des faits sur des lieux géographiques grâce à une approche non supervisée basée sur l'utilisation combinée d'une ontologie spatiale et d'un classifieur (Nearest Prototype-based Classification). Les auteurs ont identifié les noms de lieux géographiques en associant les termes marqués comme nom propre avec les noms géographiques disponibles dans le gazetteer Geonames [26] (e.g. « New York »). Puis, les auteurs ont extrait le graphe de dépendance (Dependency Graph) de chaque phrase afin de construire le chemin de dépendance (Dependency Path « DP [27] »). Finalement, en utilisant l'ontologie SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001), les auteurs ont prédit, pour chaque relation de dépendance DP, la classe (c'est-à-dire le type de relation) à l'aide d'un classifieur et une mesure de similarité sémantique.

Dans (Roberts et al., 2013), une approche pour reconnaître les relations spatiales entre deux événements spatiaux est définie. Par exemple, dans la phrase « The [bombing] victim [died] immediately », l'événement « died » est spatialement lié à l'événement « bombing ». Les auteurs proposent une méthode par apprentissage supervisé : (1) SVM binaire (binary support vector machine) pour la reconnaissance des relations spatiales, explicitement et implicitement exprimées, entre deux événements mentionnés en utilisant des descripteurs lexicaux, syntaxiques et l'identification des événements sémantiquement liés ; et (2) SVM multi-classes pour déterminer le type de relation selon l'un des cinq types de relations spatiales. Par exemple, le type « SAME » caractérise deux événements qui ont des frontières spatiales indiscernables, le type « NEAR » spécifie que deux événements ne partagent pas de frontières spatiales, mais ils sont proches l'un de l'autre, etc. Les auteurs ont évalué leur système sur un corpus de 162 documents de presse, un sous-ensemble du corpus SpatialML [28] (Mani et al., 2008).

D'Souza and Ng (2015) ont proposé une méthode pour l'extraction de relations spatiales. L'approche a obtenu les meilleurs résultats sur la tâche 3a [29] du défi SpaceEval [30]. Elle traite la détection de trois types de relations, deux dites statiques (par exemple, « The flower is in the vase ») et une troisième dynamique (par exemple, « He biked from Cambridge to Maine »). La relation dynamique est la relation de

26. http ://www.geonames.org

27. Les relations de dépendance correspondent à des relations grammaticales qui relient deux géo-entités dans un texte.

28. SpatialML (Mani et al., 2008) est un schéma d'annotation utilisé pour marquer les lieux mentionnés dans le texte ainsi qu'un ensemble de relations spatiales entre ces lieux.

29. La tâche 3a se concentre uniquement sur l'extraction de relations spatiales en utilisant un ensemble spécifié d'éléments spatiaux pour une phrase donnée.

30. http ://alt.qcri.org/semeval2015/task8/

déplacement (MoveLink), les relations statiques sont les relations qualitatives (QSLink) et d'orientation (OLink). Les auteurs ont proposé d'entraîner un classifieur (Link classifier) pour l'extraction des relations QSLink et OLink, qui sont représentées via un triplet « LINK(trajector, landmark, trigger) ». Trente-et-un descripteurs ont été utilisés pour entraîner LINK classifier en utilisant l'algorithme d'apprentissage SVM. Pour ces travaux, sept classifieurs ont été mobilisés pour l'identification de la relation MoveLink représentée sous le format suivant : « MoveLink(trigger, mover, source, mid-point, goal, landmark, path, motion-signal) ». Les auteurs ont décomposé la relation MoveLink en sept tuples (une paire et six triplets). Ensuite, sept classifieurs distincts sont construits pour identifier les sept tuples MoveLink. Finalement, les auteurs ont entraîné chacun des classifieurs MoveLink à l'aide de SVM^{light} (Joachims, 1999).

2.3.3 Discussion

Cette section relate les différents travaux qui existent sur l'analyse et l'identification d'entités et de relations spatiales à partir des documents textuels standards. L'état de l'art met en avant que la prise en compte des relations spatiales augmente la précision et la finesse du marquage des informations spatiales.
Bien que ces travaux soient intéressants, dans notre contexte d'étude, ils ne permettent pas de prendre en compte la diversité des formes pour exprimer les entités et les relations spatiales dans les messages courts tels que les SMS et les tweets.

En effet, les méthodes classiques d'extraction d'information ne s'avèrent pas adaptées au traitement et à l'analyse de corpus de messages courts. Les spécificités lexicales et syntaxiques de ces corpus diminuent nettement les performances des approches présentées dans cette section et la section 2.2 (Even, 2005 ; Ritter et al., 2011). En effet, si nous prenons les expressions « sur motpellier », « Dvant belevile » et « st lazar » présentent dans les exemples (SMS 12), (SMS 19) et (SMS 22), les méthodes classiques présentées dans les sections précédentes ne permettent pas d'identifier ces nouvelles formes d'expression.

Nous allons présenter dans la section qui suit, quelques travaux effectués sur l'analyse et l'extraction d'information (spatiale et non-spatiale) à partir des messages courts.

2.4 Les ES et RS dans les textes non-standards

L'accès à l'information spatiale est un problème essentiel pour la plupart des entreprises et des institutions. Ces dernières étudient les textes afin d'en extraire des informations intéressantes et pertinentes d'un point de vue commercial, scientifique, industriel, etc. De nos jours, il existe une quantité considérable de nouvelles sources d'informations qu'il faut intégrer, organiser et utiliser au mieux pour pouvoir traiter

l'information géographique.

2.4.1 Spécificités de l'information dans les textes non-standards

Les corpus non-standards se caractérisent par l'utilisation de mots plutôt courts en raison du besoin d'exprimer le plus d'informations possibles dans un texte de taille limitée et/ou le plus rapidement possible. La variété du contenu textuel des corpus non-standards ainsi que l'émergence régulière de nouveaux types de textes (nouveau vocabulaire, présence de fautes d'orthographe, etc.) rendent difficiles la reconnaissance et l'extraction automatique d'information par les outils de TALN (Tarrade and Lopez, 2017), en particulier les ES qui peuvent prendre des formes multiples.

2.4.2 Les méthodes liées aux corpus non-standards

Les formes d'écriture non standards ont fait l'objet de plusieurs études, en particulier en sciences du langage et en TALN (Anis, 2001 ; Fairon et al., 2006). Nous pouvons par exemple citer Kobus et al. (2008) qui ont présenté une étude comparative des systèmes visant à normaliser l'orthographe des messages SMS en français. Cooper et al. (2005) ont décrit un système pour extraire des informations à partir des messages courts (e-mail et texte). Le système comprend trois phases : (i) la génération d'une collection de modèles de structures de phrases (par exemple, « The year of <titleValue> is <yearValue > ») ; (ii) l'utilisation de cette collection pour localiser de nouvelles informations. En effet, les messages sont passés par une étape de tokenisation afin d'identifier les phrases. Puis, chaque phrase est comparée aux modèles. Si la phrase correspond à un modèle, les données sont extraites et le contexte est mis à jour ; et (iii) la mise à jour de base de données en utilisant les informations identifiées.

Han and Baldwin (2011) ont proposé une méthode pour identifier et normaliser des mots non-standards (fautes de frappe, abréviations ad hoc, etc.) dans un texte anglais bruité (e.g. « shuld » « should »). Pour atteindre cet objectif, les auteurs ont (i) mené une étude pilote sur la distribution de mots non-standards de Twitter comparativement à d'autres types de textes (ceci permet de comparer les formes scripturales selon les sources) ; (ii) généré un ensemble de données issue de Twitter ; (iii) proposé une nouvelle approche de normalisation qui exploite la similarité des mots et leur contexte sans nécessité de données annotées.

Han et al. (2013) ont décrit une méthode pour identifier et normaliser les variantes lexicales issues de tweets et de SMS anglais (e.g. « tmrw », « tomorrow »). La méthode proposée utilise un classifieur (SVM) pour détecter les variantes lexicales et pour générer des corrections. L'approche est basée sur la similarité morphophonémique [31] en utilisant à la fois des informations contextuelles et des similarités de

31. morphophonémique est basée sur la variation morphologique et phonétique des mots.

chaînes de caractères. Ces deux dernières sont ensuite exploitées pour sélectionner le candidat le plus probable pour la tâche de normalisation.

Ces travaux traitent donc de l'analyse de différents types d'information en tenant compte des spécificités des corpus de messages courts, plus particulièrement les SMS. Certaines de ces approches visent essentiellement à normaliser les termes non-standards présents dans ce type de corpus. Dans la section suivante, nous présentons quelques travaux à partir de textes non-standards, plus particulièrement la reconnaissance d'entités nommées dont ceux de type lieu à partir des tweets.

2.4.3 Les méthodes d'identification des ES et RS

Nous présentons, dans cette section, quelques travaux existants sur la REN (contenant une entité de type lieu) (Liu et al., 2013 ; Ritter et al., 2011 ; Sileo et al., 2017), ainsi que des travaux visant à analyser et identifier les informations spatiales (Li and Sun, 2014 ; Liu et al., 2014) à partir de messages courts.

Ritter et al. (2011) ont développé un pipeline NLP appelé TwitterNLP [32] pour la REN dans les tweets. Ce pipeline repose sur une approche supervisée qui s'appuie sur des nouveaux outils de NLP spécifiques aux tweets (étiquetage grammatical (T-POS), chunking (T-CHUNK), classifieur de capitalisation (T-CAP)) en raison des résultats peu satisfaisants obtenus avec les outils standards. T-POS et T-CHUNK sont entraînés en utilisant des CRF avec des descripteurs classiques et spécifiques aux tweets. Les descripteurs spécifiques aux tweets sont : retweets, @usernames, #hashtags, URLs. Ensuite, afin d'identifier les entités nommées dans les tweets, les auteurs ont proposé de segmenter les entités nommées (T-SEG) puis de les classifier (T-CLASS). Les sorties des différents outils sont utilisées comme des descripteurs pour T-SEG en utilisant des CRF. Finalement, afin de classifier les entités nommées, le système T-CLASS utilise Latent Dirichlet Allocation (LDA [33]) (Ramage et al., 2009) pour exploiter des tweets non étiquetés en utilisant les dictionnaires d'entités recueillies à partir de Freebase. L'évaluation de ce système montre que T-NER double le score de F-mesure par rapport à Stanford NER (Finkel et al., 2005) sur un corpus de test composé de 2400 tweets.

En outre, Liu et al. (2013) ont proposé une approche d'apprentissage semi-supervisé pour la tâche de REN pour les tweets anglais. Cette approche combine le classifieur K-plus proches voisins (KPPV) avec un modèle CRF pour résoudre le problème du manque d'informations dans les tweets et l'indisponibilité des données d'apprentissage. Les KPPV et le modèle CRF sont entraînés sur différents ensembles d'apprentissage. Ces ensembles sont basés sur le même ensemble de départ, progressivement enrichi avec de nouveaux tweets étiquetés automatiquement.

Li and Sun (2014) ont proposé une méthode pour extraire le nom de chaque point d'intérêt (POI) (par exemple, des restaurant, des centres commerciaux, librai-

32. https ://github.com/aritter/twitter_nlp

33. Labeled LDA est un modèle graphique probabiliste qui décrit un processus pour générer une collection de documents étiquetés.

ries, etc.) dans des tweets anglais et les relier aux valeurs temporelles relatives à des visites. Pour atteindre cet objectif, la méthode Peter qui a été mise en œuvre est constituée de deux composants principaux : (1) *L'inventaire POI* (POI inventory) qui est une collection de mots et de phrases, dont chacun est un nom de POI ou une partie d'un nom de POI. Le but est d'enrichir l'inventaire POI par des formes courtes et des abréviations informelles. Les auteurs ont construit l'inventaire en exploitant le Foursquare check-ins [34] ; (2) *Le tagger de POI* (time-aware POI tagger) basé sur des CRF qui modélisent les classes de sortie en tant que séquences. Le tagger considère l'inventaire POI comme base de connaissances et utilise quatre types de caractéristiques (lexicales, grammaticales, géographiques et caractéristiques BILOU du schéma [35]). Ces derniers permettent de désambiguiser un nom de POI candidat mentionné dans un tweet (par exemple, le mot *mac* peut se référer aux produits de « Apple » et du restaurant « McDonald's »). Les expérimentations montrent que la méthode Peter donne les meilleurs résultats avec une F-mesure à 87% (en comparaison, Standford NER obtient 79%).

Dans (Liu et al., 2014), une méthode d'identification automatique des expressions de localisation (LEs) dans les textes des réseaux sociaux a été proposée. Les auteurs ont annoté manuellement 3500 phrases sélectionnées aléatoirement à partir des corpus collectés auprès des réseaux sociaux populaires (Twitter, Forums, Blogs et Wikipedia) et un corpus anglais (British National Corpus). Ensuite, les auteurs ont évalué les géoparsers (Stanford NER, GeoLocator, Unlock, etc.) sur ces données annotées. Dans ce contexte, Stanford NER a le meilleur comportement avec une F-mesure pourtant assez faible ($\sim 31\%$).

Dittrich et al. (2015) ont analysé plus de soixante prépositions dans les messages courts en anglais (tweets) afin d'identifier parmi ces dernières celles qui sont des relations spatiales. L'analyse montre qu'il existe peu de termes prépositionnels qui sont utilisés dans les descriptions spatiales verbales [36] dans les messages courts tels que « at, in, to, on, from, out (of) ». De plus, les auteurs ont calculé une probabilité basée sur la fréquence pour indiquer si un terme est utilisé dans un contexte spatial.

Enfin, Sileo et al. (2017) ont proposé un système de REN pour le challenge CAp

34. Foursquare est un réseau social qui permet à chaque utilisateur de « checker » (pointer) dans un lieu et de signaler son passage grâce à un système de géolocalisation. Il permettait également de consulter et de recommander les lieux « checkés » (check-ins) par les utilisateurs pour repérer des points d'intérêts (restaurants, cafés, magasins, etc.). Foursquare contient non seulement les noms formels des points d'intérêts, mais aussi les abréviations informelles (https ://foursquare.com/).

35. Le schéma BILOU (Ratinov and Roth, 2009) identifie le début (**B**eginning), l'intérieur (**I**nside) et le dernier (**L**ast) mot d'un nom de POI à plusieurs mots (multi-mots) ainsi que le nom de POI de longueur unitaire (**U**nit-length). Les mots qui n'apparaissent pas dans les noms de POI sont identifiés par l'étiquette externe (**O**utside). Par exemple, le schéma BILOU, pour le tweet « We're all for Asian delights ! **Thai express** $[POI_z]$ today, **suki sushi** $[POI_f]$ tomorrow », est *We're\O all\O for\O Asian\O delights\O !\O Thai*B_z *express*L_z *today\O ,\O suki*B_f *sushi*L_f *tomorrow\O*, où $_z$ (présent) et $_f$ (future) indiquent le temps du POI

36. La forme syntaxique d'une description spatiale verbale incluant une préposition est */ [subject verb] preposition] NP (phrase nominale).*

2017 (Lopez et al., 2017) sur des tweets français. Le challenge traite 13 types d'entités dont l'entité « Geoloc » caractérisant un emplacement (par exemple « Gaule », « mont de Tallac », « France », « NYC », etc.). Ce système utilise les CRF comme modèle d'apprentissage en appliquant des descripteurs morpho-syntaxiques (étiquetage grammatical, longueur des mots, etc.), word embedding (Fasttext [37] (Bojanowski et al., 2016)) ainsi que des clusters de mots basés sur ces représentations en utilisant un modèle de mélange gaussien. Ce système a remporté le défi CAp 2017 avec un score de F-mesure de 58%.

Dans cette section, nous avons présenté des travaux liés à l'identification et l'analyse de l'information spatiale à partir des corpus de messages courts et spécifiquement les tweets. Ces travaux sont intéressants cependant ils présentent des lacunes quant à la précision de l'information spatiale. En effet, la plupart de ces travaux traitent les corpus non-standards anglais et très peu traitent les messages courts en français. Ces derniers visent à analyser les EN de façon classique (notamment pour les lieux en traitant les noms toponymiques uniquement sans tenir compte des RS (Sileo et al., 2017)).

2.5 Conclusion

Les méthodes classiques d'extraction d'information ne s'avèrent pas adaptées au traitement et à l'analyse de corpus de messages courts. En effet, en raison des particularités lexicales et syntaxiques de ce type de corpus, les performances diminuent nettement quelle que soit l'approche utilisée. De plus, la disponibilité limitée (en particulier en français) de ressources linguistiques spécialisées (lexiques, corpus annotés) ne permet pas d'adapter les outils existant à ce type de textes.

La différence de notre approche par rapport à d'autres comme celles de Stanford NER (Finkel et al., 2005) ou Polyglot (Al-Rfou et al., 2015) est illustrée par deux points majeurs : (1) Stanford NER et Polyglot marquent essentiellement les ESA classiques ; (2) La spécificité de nos corpus nécessitent des traitements souvent plus fins. En effet, les approches de la littérature produisent des résultats intéressants à partir des corpus standards mais elles ne s'avèrent pas adaptées aux corpus de messages courts qui peuvent être atypiques. Pour cette raison, les méthodes classiques ne sont pas nécessairement en adéquation avec à notre problématique.

Dans cette thèse, nous nous concentrons plus précisément sur l'étude des messages courts comme les SMS et les tweets en français. En raison de la spécificité de leur écriture, le traitement de ce type de données représente un défi pour la communauté scientifique en TALN et RIG (Recherche d'Information Géographique). En

37. FastText est une bibliothèque créée par l'équipe de recherche Facebook pour l'apprentissage efficace des représentations de mots et de la classification des phrases.

effet, de nombreux travaux sont dédiés à l'extraction et l'identification d'informations spatiales à partir des corpus standards. En revanche, et jusqu'à une période récente, très peu de travaux se sont intéressés à cette problématique à partir de corpus non-standards en langue française. Les travaux qui existent sur les corpus français visent à traiter les EN de façon classique, notamment pour les lieux en traitant les noms toponymiques uniquement sans tenir compte des RS (Sileo et al., 2017).

Dans ce contexte, une contribution principale de cette thèse est l'identification de nouvelles entités spatiales absolues et relatives à partir de messages courts. Notons que l'absence de corpus de messages courts et peu standards préalablement annotés et la difficulté d'en constituer pour ce type de textes rendent très difficile le recours aux méthodes éprouvées de la littérature qui reposent sur des approches d'apprentissage supervisé.

Deuxième partie

Contributions méthodologiques

Chapitre 3

Extraction des entités spatiales à partir de données textuelles non-standards

Contents

3.1 Introduction

Les corpus de messages courts tels que les SMS et les tweets se caractérisent par l'utilisation de mots plutôt courts en raison du besoin d'exprimer le plus d'informations possibles dans un texte de taille limitée et/ou le plus rapidement possible. En effet, outre le volume de données à traiter, ce type de corpus contient une multitude d'expressions différentes ou variantes pour exprimer une même entité nommée, et plus particulièrement une même entité spatiale absolue (ESA) (e.g. « montpelie », « montpel » pour « Montpellier »).

L'identification et la reconnaissance des ESA constituent un champ important du domaine de l'extraction d'information dans des corpus documentaires. Il s'agit de pouvoir retrouver dans le texte l'ensemble des informations permettant de décrire précisément l'entité spatiale afin de déterminer sa représentation spatiale la plus précise possible. Ce sont des tâches liées au domaine de la recherche d'information géographique notamment, ou encore à celui de la construction des ressources sémantiques.

Notre première contribution consiste à extraire de nouvelles formes d'expression d'ESA dans des corpus de messages courts. Dans ces travaux, nous nous concentrons plus précisément sur les messages courts de type SMS et tweets en français.

Pour rappel, une ESA est une entité nommée (EN) simple de type « lieu » dans la catégorisation des entités nommées (Grishman and Sundheim, 1996). C'est une référence directe à un espace géo-localisable comme « la ville de Paris », « la gare Saint-Roch », « Montpellier », etc. (Lesbegueries and Loustau, 2006 ; Tahrat et al., 2013). Une ESA définit le vocabulaire utilisé pour nommer des concepts topographiques lorsqu'il est nécessaire de le préciser (par exemple, *ville*, *gare*, etc.) et un nom toponymique (par exemple, « Paris », « Saint-Roch », « Montpellier », etc).

Dans le chapitre précédent, nous avons présenté l'état de l'art pour l'extraction automatique d'EN, et plus spécifiquement des ESA. De nombreux travaux ont été effectués dans ce champ de recherche et des outils d'extraction automatique ont été mis en place, en particulier, pour des corpus standards tels que les articles de presse [1] (Finkel et al., 2005), les articles Wikipédia (Al-Rfou et al., 2015), les articles scientifiques, les corpus fournis par les organisateurs de défis scientifiques de Traitement Automatique du Langage Naturel (TALN) (Florian et al., 2003), etc.

Les méthodes classiques d'extraction produisent des résultats intéressants mais elles ne sont pas adaptées aux messages courts en raison de leurs spécificités, notamment, lexicales et syntaxiques. En effet, la multitude et la variété des corpus non-standards se caractérisant par l'émergence régulière de nouvelles formes de textes (nouveau vocabulaire, présence de fautes d'orthographe, etc.), la reconnaissance et l'extraction automatique d'information sont rendues plus difficiles. Et comme le montre l'état de l'art présenté dans le chapitre 2, à notre connaissance, peu de travaux traitant l'extraction des informations spatiales ont été réalisés sur les corpus de

1. Europeana Newspapers (http ://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/).

messages courts, en particulier les SMS et les tweets. En comparaison, et jusqu'à une période récente, très peu de travaux ont été menés sur ce type de corpus en langue française. Cette tâche permet de bénéficier de la grande quantité de connaissances géographiques exprimées dans divers textes en langage naturel afin d'obtenir une vision plus complète de la connaissance spatiale.

Parmi les méthodes existantes, nous ne pouvons nous appuyer sur des approches d'apprentissage supervisé qui nécessitent un corpus annoté d'apprentissage volumineux. Ces ressources n'existent pas à ce jour dans notre contexte d'étude. À noter également que la production de ce type de corpus via l'annotation d'experts est une tâche complexe car le langage SMS/tweet n'est pas clairement standardisé et il évolue très vite.

Notre approche exploite des connaissances linguistiques combinées à des dictionnaires (liste de pays, villes, etc.) qui sont pré-définis manuellement par des experts. Dans ce cadre, notre première contribution combine plusieurs approches de TALN, à savoir une analyse statistique (mesures de similarité) ainsi qu'une analyse lexicale pour identifier et extraire de nouvelles formes d'expression d'ESA apparaissant dans des corpus de messages courts.

L'avantage principal de notre approche est qu'elle peut être adaptée à la grammaire locale et à des styles de textes spécifiques. En outre, elle peut obtenir une précision élevée pour des textes particuliers. Nous verrons dans ce chapitre que les résultats obtenus par notre approche sont encourageants. Les expérimentations ont été menées sur les trois corpus en français 88milSMS, tweets et Midi Libre présentés en Section 3.4.1. Afin de situer notre méthode par rapport à l'état de l'art, nous avons comparé nos résultats avec ceux produits par deux approches existantes, à savoir Stanford NER (Finkel et al., 2005) et Polyglot (Al-Rfou et al., 2015) (cf. Chapitre 2 Section 2.2.2).

Nous reprenons, dans ce qui suit, des extraits du corpus de SMS *88milSMS* et de tweets, contenant des ESA standards et des variantes d'ESA, que nous utilisons tout au long de ce chapitre pour bien comprendre le domaine d'application de nos travaux (pour plus de détails voir Chapitre 1 Section 1.1.2).

(SMS 1) *Est ce que l'apéro est au frais ? Je suis* ***au niveau de Valence en route vers Montpellier****. Bonne continuation pour les vacances.*

(SMS 2) *Hey, tu arrives quand ? Tu restes* ***à Béziers*** *?*

(SMS 3) *Jeudi soir on a le concert de Dick Annegarn* ***à Lattes*** *! Je t'M plus que tout le* ***Massachussets*** *! < PRE_3/ >* [2]

(SMS 4) *Oui mais la je suis dans le tram et le temps d'arrive y sera 6 h ch ui* ***a odysseum***

(SMS 5) *Pffff ... Je sais pas ! Le temps de venir* ***vers frontignan*** *... Euh 21 heures 15 ...*

(SMS 6) *T'as le reçu la CAF ? Ben moi je rentre aujourd'hui* ***à sète****.*

2. Étiquettes PRE représente le prénom utilisé lors de la phase d'anonymisation.

(SMS 7) *Si **place 2 leurope**, é voui y va **a latte***

(SMS 8) *Je rentre **à Monptel'** finalement, j'ai eu accès au site pour les cours et y a une réunion de regroupement de lundi à jeudi.*

(SMS 9) *Je viens d'arriver. Je suis du côté de la sortie **vers Ste Eustache**. A toute*

(SMS 10) *Ya < PRE_4[3]/ > qui vient chez nous vers 14h15. Rentre, on ira **a odyseum** en voiture*

(SMS 11) *Salut ! Pas grave je suis avec < PRE_5/ > il part demain. Ce w-end je reste **a monpellier** mais le w-end prochain je serais **a Leucate** ...*

(SMS 12) *Est ce que tu es dispo pour samedi 1 octobre pour gardez des enfants pour le comite de quartier de 15h a 18h ? Puis il y as < PRE_5/ > qui et **sur motpellier***

(SMS 13) *Ouii j'viendrais **a Bezier** faire tes petites soirees etudiantes !*

(SMS 14) *On s'est garé un peu à l'extérieur avant les déviations, on file vers le centre à **pattes**.*

(SMS 15) *Mais nan nan on part **à tahiti** il parait que c'est le paradis du ski !*

(SMS 16) *Quand je Prend une douche avec le gel douche **tahiti** J'ai l'impression d'avoir*

(Tweet 1) *ptetre jpourrais aller **a montpelier** ? ?*

(Tweet 2) *@calant42 et vous **du cote de st etienne** ?*

(Tweet 3) *@Valentino_ Greci c'était une salle **à st andre**, yavait 4-5 vigiles mais jsp c'etait quoi le bail*

(Tweet 4) *@Pokosphere62 moi je vais sûrement descendre en train **sur la gare de Lilli Flandre***

(Tweet 5) *Jv aller fr un tour **dvant gambetta** lundi*

(Tweet 6) *donc du coup on a du attendre facile 25min **dvant le palais de justice** mais on avait pas ldroit de sortir du bus*

(Tweet 7) *pecnologic puis mes amis qui ont ét'e dans d'autres villes **avant tokyo** ont vraiment détesté tokyo je suis pas des pires tbh ça va*

(Tweet 8) *@davmarouani @Bontemps78 si tu passes **par Lille** je te promets un donjon au top Et en exclu*

La suite de ce chapitre est organisée de la façon suivante. La section 3.2 détaille une sélection de mesures classiques de pondération. La section 3.3 présente l'approche proposée. Puis, nous décrivons en section 3.4 le protocole expérimental

3. Le chiffre 4 renvoie au nombre de caractères du prénom dans le SMS brut (e.g. Sarah <PRE_5/>).

et les résultats obtenus en les comparant à deux approches connues de la littérature. La section 3.5 présente une conclusion liée à notre première contribution pour l'identification d'ESA.

3.2 Mesures de pondération existantes

Dans cette section, nous présentons différentes approches, en rapport avec notre recherche, qui permettent de comparer des textes afin de déterminer dans quelle mesure nous pouvons nous en servir dans le cadre de notre problématique. En effet, différentes études sur l'extraction d'entités nommées se concentrent sur des mesures de similarité entre deux chaînes de caractères qui permettent de calculer la proximité lexicale entre ces deux chaînes. De nombreuses mesures de similarité ont été proposées comme String Matching (Maedche and Staab, 2002), Jaro (Jaro, 1989), Jaro-Winkler (Winkler, 1999) et Lin (Lin, 1998) que nous allons détailler dans les sections suivantes.

3.2.1 String Matching

La mesure String Matching, que nous nommons dans cette thèse String Matching de base (SM_B), est une mesure de similarité lexicale fondée sur la distance de Levenshtein (Levenshtein, 1966) qui compare deux chaînes de caractères.

La distance de Levenshtein (Levenshtein, 1966), aussi connue sous le nom de distance d'édition (E), est une mesure formulée par Vladimir Levenshtein qui permet de pondérer la différence entre deux chaînes de caractères. La distance est mesurée en fonction du nombre minimal d'opérations d'insertion, de suppression, ou de substitution nécessaires pour transformer une chaîne en une autre. La distance égale à 0 indique que les deux chaînes de caractères sont identiques.

$$\mathrm{E}(\mathrm{Ch}_1, \mathrm{y}) = \begin{cases} n & si\ m = 0 \\ m & si\ n = 0 \\ \min\{E(d[i-1,j]+1,\ d[i,j-1]+1, & \\ d[i-1,j-1]+coût\} & sinon \end{cases} \tag{3.1}$$

Dans la formule (4.1), n représente la longueur de la chaîne Ch_1 ; m est la longueur de la chaîne Ch_2 ; d est une matrice de longueur $d[n+1, m+1]$; $d[i-1, j]$ est l'élément directement au-dessus $d[i, j]$; $d[i, j-1]$ est l'élément directement avant $d[i, j]$; $d[i-1, j-1]$ est la diagonale qui précède $d[i, j]$; $coût = 0$ si $Ch_1 = Ch_2$, sinon $coût = 1$.

À partir de l'exemple présenté dans la Figure 3.1 dans lequel nous comparons les chaînes « Bezier » (cf. Exemple (SMS 13) Section 3.1) et « Béziers » (cf. Exemple (SMS 2) Section 3.1), nous obtenons E(Béziers, Bezier) = 2. En effet, il y a deux opérations permettant de passer de la chaîne « Bezier » à « Béziers ».

FIGURE 3.1 – Distance d'édition Levenshtein (E) pour les ES « Bezier » et « Béziers ».

La mesure String Matching SM_B considère le nombre de modifications qui doivent être apportées pour changer une chaîne en une autre puis elle pondère le nombre de ces modifications par rapport à la longueur de la chaîne la plus courte (Maedche and Staab, 2002) (Équation 3.2).

SM_B retourne un score de similarité entre 0 et 1 où 1 représente une correspondance parfaite (les deux termes sont identiques) et 0 pour une mauvaise correspondance (les deux termes sont distincts).

$$\mathrm{SM_B}(Ch_1, Ch_2) = \max[0, \frac{(\min(|Ch_1|, |Ch_2|) - \mathrm{E}(Ch_1, Ch_2))}{\min(|Ch_1|, |Ch_2|)}] \tag{3.2}$$

Dans la formule (3.2), $|Ch_1|$ est la longueur de la chaîne Ch_1 ; $|Ch_2|$ est la longueur de la chaîne Ch_2 ; E présente la distance de Levenshtein.

À partir de l'exemple de la Figure 3.1, nous calculons SM_B comme suit (cf. Exemple 3.2.1) :

Exemple 3.2.1 $SM_B(Béziers, Bezier) = \max[0,\ (6-2)/6] = 0.66.$

3.2.2 Jaro

La distance de Jaro se base sur le nombre et l'ordre des caractères communs entre deux chaînes de caractères (Équation 3.3). En effet, si un caractère commun aux deux mots n'est pas à la même position, la valeur de la similarité diminue. La distance de Jaro pour les deux chaînes Ch_1 et Ch_2 est définie comme suit :

$$Jaro(Ch_1, Ch_2) = \frac{1}{3}\left(\frac{|m|}{|Ch_1|} + \frac{|m|}{|Ch_2|} + \frac{|m| - t}{|m|}\right) \tag{3.3}$$

Dans la formule (3.3), m représente le nombre de caractères en commun entre Ch_1 et Ch_2 ; Ch_i est la longueur de la chaîne i ; t est le nombre de caractères différents présents à des positions identiques.

En prenant l'exemple (3.2.2), nous obtenons une distance de Jaro égale à 0.71, entre les deux chaînes de caractères « Béziers » et « Bezier ».

Exemple 3.2.2 $Jaro(Béziers, Bezier) = \frac{1}{3}\left(\frac{|5|}{|7|} + \frac{|5|}{|6|} + \frac{|5|-2}{|5|}\right) = 0.71$

3.2.3 Jaro-Winkler

La distance de Jaro-Winkler est une amélioration de la mesure de Jaro. Elle favorise les correspondances entre les chaînes avec des préfixes communs. Ainsi, plus les préfixes identiques de deux mots sont importants, plus la similarité entre les deux mots sera grande. Le résultat est normalisé de façon à avoir une mesure entre 0 et 1 où 0 représente l'absence de similarité et 1 l'égalité des chaînes comparées. L'équation (3.4) présente la formule de distance Jaro-Winkler.

$$JW(Ch_1, Ch_2) = Jaro(Ch_1, Ch_2) + (lp(1 - Jaro(Ch_1, Ch_2))) \tag{3.4}$$

Dans la formule (3.4), l représente la longueur du plus grand préfixe en commun entre Ch_1 et Ch_2 ; p est un coefficient fixé à la valeur 0.1.

Pour le couple (Béziers, Bezier), nous obtenons un score de similarité égal à 0.73 (cf. Exemple 3.2.3), tandis que nous avons obtenu pour le même couple un score égal à 0.71 avec la mesure de Jaro (cf. Exemple 3.2.2).

Exemple 3.2.3 $JW(Béziers, Bezier) = 0.71 + (1(0.1)(1 - 0.71)) = 0.73$

3.2.4 Lin

La mesure proposée par Lin, que nous nommons dans cette thèse *Lin de base* (Lin_B), est basée sur le calcul du nombre de caractères tri-grammes (noté T) en commun entre deux chaînes Ch_1 et Ch_2. La technique des n-grammes (Shannon, 1948) est utilisée pour calculer le nombre de n [4] caractères identiques consécutifs entre deux chaînes de caractères (Roche, 2011). Lin_B retourne un score normalisé entre 0 et 1 où 1 indique que les deux termes sont identiques et 0 que les deux termes sont distincts.

L'équation 3.5 présente la mesure de similarité Lin_B.

$$\mathrm{Lin_B}(Ch_1, Ch_2) = \frac{1}{[1 + |\mathrm{T}(Ch_1)| + |\mathrm{T}(Ch_2)| - 2 \times |\mathrm{T}(Ch_1) \cap \mathrm{T}(Ch_2)|]} \tag{3.5}$$

Dans la formule (3.5), $|\mathrm{T}(Ch_1)|$ est l'ensemble de tri-grammes de Ch_1 ; $|\mathrm{T}(Ch_1) \cap \mathrm{T}(Ch_2)|$ est l'ensemble de tri-grammes en commun entre Ch_1 et Ch_2.

Par exemple, les tri-grammes des deux chaînes de caractères « Béziers » et « Bezier » ainsi que leur intersection sont respectivement :
|**T** (Béziers)| = {Béz,ézi,zie,ier,ers} = 5
|**T** (Bezier)| = {Bez,ezi,zie,ier}= 4
|**T** (Béziers) ∩ **T** (Bezier)| = {zie,ier} = 2

Pour calculer la similarité entre « Béziers » et « Bezier », nous obtenons le résultat ci-dessous en reprenant la formule 3.5.

Exemple 3.2.4 $\mathrm{Lin_B}(Béziers, Bezier) = \frac{1}{[(1+5+4)-(2\times2)]} = 0.16$

4. Généralement, la valeur de n varie entre 2 et 5

3.2.5 Quelle(s) mesure(s) utiliser ?

Les mesures de traitement de chaînes peuvent généralement être divisées en distance d'édition et distance basée sur les n-grammes. Le choix entre une métrique d'édition ou une distance basée sur les n-grammes est dépendant de la longueur de la chaîne (Van der Loo, 2014).
Pour déterminer la distance d'édition, nous calculons le nombre d'opérations fondamentales nécessaires pour transformer une chaîne en une autre. La mesure basée sur les n-grammes est obtenue en comparant les occurrences des séquences de n-caractères entre deux chaînes.

Parmi les nombreuses mesures de similarité existantes, nous avons choisi d'adopter la mesure SM_B fondée sur la distance de Levenshtein ainsi que celle de Lin_B fondée sur les n-grammes de caractères. Ces mesures sont particulièrement adaptées au traitement de chaînes de caractères.

Concernant les distances d'édition, la distance Jaro et Jaro-Winkler sont très proches de la distance de Levenshtein. La seule différence tient au fait que Jaro-Winkler favorise les chaînes qui ont le plus grand préfixe en commun. Dans notre étude, nous proposons de favoriser l'élimination des chaînes qui ne commencent pas par le même préfixe (par exemple, « Lattes » et « pattes » voir les exemples (SMS 3) et (SMS 14)). Nous n'avons donc pas retenu l'utilisation de ces distances dans nos approches.

Ainsi, nous avons privilégié l'utilisation de métriques de base comme la distance de Levenshtein intégrée dans la mesure SM_B qui donne des résultats tout à fait satisfaisants pour la détection de variantes de chaînes identifiables par les opérations de suppression, ajout et substitution de caractères (Bilenko et al., 2003 ; He et al., 2008 ; Sehgal et al., 2006). En effet, la distance de Levenshtein est classiquement utilisée dans la littérature pour répondre à ce besoin et cela dans de nombreux domaines, comme l'alignement d'ontologies (Stoilos et al., 2005), la désambiguïsation des entités (Song et al., 2007), l'extraction d'information (Reul et al., 2016), etc.

Cependant, comme le montrent les exemples 3.2.1 et 3.2.4, ces mesures ne suffisent pas à identifier des similarités entre deux termes en générant des résultats insatisfaisants liés aux écrits non-standards qui sont traités dans cette thèse. Dès lors, nous avons ajouté des règles structurelles décrites dans la Section 3.3.

3.3 Notre approche

Dans l'objectif d'extraire des ESA présentes dans des corpus textuels constitués de messages courts, nous proposons l'approche comprenant les deux étapes suivantes (cf. Figure 3.2) : (1) Identification des ESA standards, (2) Découverte de nouvelles formes d'expression d'ESA.

Nous identifions dans un premier temps, les ESA à partir de corpus de messages courts et de listes d'ESA connues (cf. Section 3.3.1).

Dans un second temps, nous enrichissons le dictionnaire initial des ESA avec de nouvelles formes d'expression de ces ESA en combinant une mesure de similarité et une analyse lexicale (cf. Section 3.3.2). Les phrases (SMS 7) à (SMS 13) illustrent les extraits de SMS contenant de nouvelles formes d'expression d'ESA. Nous détaillons par la suite chacune des étapes.

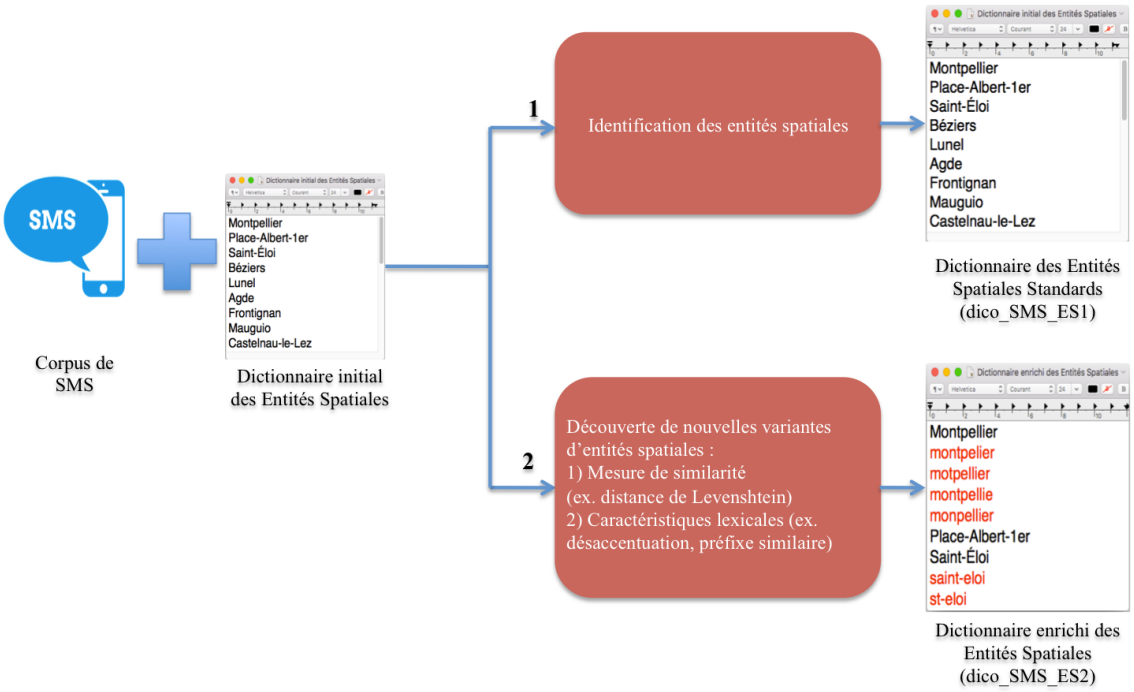

FIGURE 3.2 – Processus d'identification de nouvelles formes d'expression d'ESA. Les entités notées en rouge dans « dico_SMS_ES2 » correspondent aux variantes ajoutées avec notre approche.

3.3.1 Identification des entités spatiales absolues

Une première étape consiste à identifier les ESA standards dans les corpus de messages courts. Pour cela, une approche lexicale classique s'appuyant sur différentes sources de données est appliquée.

Nous nous appuyons notamment sur les bases ©BD Topo [5] et ©BD Carto [6] de l'IGN [7] ainsi que sur un ensemble de listes de noms de lieux liés aux zones considérées (Plus de détails sont donnés en Section 3.4). À noter qu'une approche classique par patrons intégrant des règles lexicales telles que l'identification de mots capitalisés est inefficace en raison du style d'écriture utilisé dans les messages courts. En effet, ce type d'information textuelle est peu standardisé et n'utilise pas de capitalisation de manière cohérente (cf. Exemples (SMS 5) et (SMS 6)).

Le travail présenté dans cette section permet d'identifier les ESA standards telles que « Montpellier », « Paris », etc. Si nous considérons les exemples présentés en Section 3.1, les ESA présentes dans les exemples (SMS 1) et (SMS 2) sont alors identifiées. Les ESA reconnues sont respectivement « Valence », « Montpellier » et « Béziers ». Les ESA présentes dans les exemples (SMS 12) et (SMS 13) ne sont pas identifiées (« motpellier », « bezier »). Pour résoudre ce problème, nous proposons, dans la section suivante, une approche pour identifier de nouvelles formes d'expression d'ESA.

3.3.2 Découverte de nouvelles formes d'expression d'entités spatiales absolues

Pour identifier et extraire de nouvelles formes d'expression d'ESA, correspondant à des formulations différentes des ESA existantes, nous proposons une méthode calculant la similarité entre le dictionnaire initial des ESA standards et les mots issus du corpus de messages courts. Chaque ESA candidate (chaque terme du corpus) est associée à un score de similarité pour chacune des ESA standards afin de mesurer leur pertinence (cf. Exemples 3.2.1 et 3.2.4 Section 3.2). Ensuite, nous déterminons les mots les plus proches à partir des nouvelles ESA candidates en fixant un seuil $S1$ correspondant au score de similarité à partir duquel une variante est estimée valide.

Parmi les nombreuses mesures de similarité existantes, nous avons intégré la mesure SM_B ainsi que la mesure Lin_B. Ces deux méthodes permettent de calculer la proximité lexicale entre deux chaînes de caractères. Elles sont classiquement utilisées dans la littérature car elles donnent des résultats pertinents (Duchateau et al., 2008

5. BD Topo est une description vectorielle 3D (structurée en objets) des éléments du territoire et de ses infrastructures, de précision métrique. Elle couvre l'ensemble des départements français ainsi que les collectivités d'Outre-Mer de Saint-Martin, Saint-Barthélemy et Saint-Pierre-et-Miquelon (http ://professionnels.ign.fr/doc/DC-BDTOPO-2-2.pdf).

6. BD Carto est structurée en thèmes regroupant des objets partageant une même fonctionnalité sur le terrain ou dans la base. Elle couvre l'ensemble des départements français, ainsi que les départements et les régions d'Outre-Mer, pour lesquels les spécifications de contenu ont parfois dû être adaptées pour certains thèmes (https ://www.ppige-npdc.fr/portail/sites/default/files/dc_bdcarto_3_1.pdf).

7. IGN (Institut national de l'information géographique et forestière) a pour vocation de décrire la surface du territoire national et l'occupation de son sol, d'élaborer et de mettre à jour l'inventaire permanent des ressources forestières nationales (http ://www.ign.fr/institut/).

; Zenasni et al., 2015).

L'écriture SMS ne tient pas compte de nombreuses règles orthographiques et grammaticales (Oliva et al., 2011). Bien que l'écriture de messages courts SMS et tweets ne soit pas normalisé, des linguistes ont identifié les spécificités lexicales de ce type de corpus (e.g. mots orthographiés de manière non-standard, mots écrits sans les accents, prénoms avec ou sans majuscules, abréviations, etc.) (Cougnon and Ledegen, 2008 ; Panckhurst, 2009 ; Panckhurst et al., 2013).

Afin de réduire l'ambiguïté identifiée dans la première étape (la comparaison des chaînes de caractères à l'aide de la mesure de similarité SM_B et Lin_B), nous proposons une approche plus robuste en prenant en compte dans la mesure de similarité la désaccentuation. En effet, plusieurs utilisateurs francophones évitent systématiquement l'utilisation de caractères accentués dans les messages courts, au moins dans la communication informelle (Simard and Deslauriers, 2001). Ainsi que les préfixes similaires lorsque deux termes commencent avec les mêmes caractères (Liao and Wu, 2012). Ce principe adopté est appelé *Similarité + Caractéristiques Lexicales* (SM_{CL}, Lin_{CL}).

- **Désaccentuation :** tout d'abord, nous calculons la similarité entre les ESA standards provenant des lexiques d'entités spatiales et tous les mots issus du corpus de messages courts sans prendre en considération les caractères accentués. L'utilisation principale des lettres accentuées ou des signes diacritiques [8] est de guider la prononciation. Ce type d'approches est utilisé dans différentes tâches en TALN, notamment l'analyse du langage, la classification de documents et l'extraction d'information (Collin et al., 2013 ; Dini et al., 2013).
  À noter que l'on peut classer les erreurs d'accentuation en trois catégories : accent manquant (e.g. « Bezier »), accent erroné (e.g. « château ») ou accent mal placé (e.g. « odysséum »). Quatre types d'accents sont utilisés en français : accent aigu, accent grave, accent circonflexe et accent tréma.

  Après l'application d'une méthode de désaccentuation, nous obtenons, à partir de l'exemple présenté dans la Figure 3.3, E(Béziers, Bezier) = 1. En effet l'utilisation de la désaccentuation permet de réduire le nombre d'opérations à une seule opération pour passer la chaîne « Bezier » à « Béziers ». Ainsi, ce type de normalisation permet d'améliorer de manière significative la reconnaissance des ESA dans les messages courts.

  Dans notre exemple, l'ESA standard « Béziers » et l'ESA variante « Bezier », présentes dans les phrases (SMS 2) et (SMS 13) (cf. Section 3.1), sont considérées comme très proches avec $SM_{CL} = 0.83$ (cf. Exemple 3.3.1), alors que la valeur de comparaison initiale de la mesure SM est de 0.66 (cf. Exemple 3.2.1,

8. Les signes diacritiques sont des symboles graphiques (accents, cédilles) combinés à des lettres déjà existantes afin d'en modifier la phonétique ou d'éviter la confusion entre des mots homographes (e.g. cote, côte et côté) (Gotti and Lapalme, 2014).

FIGURE 3.3 – Distance d'édition Levenshtein (E) pour les ESA « Bezier » et « Béziers » en utilisant la désaccentuation.

Section 3.2.1), indiquant deux chaînes bien distinctes.

Exemple 3.3.1 $SM_{CL}(Béziers, Bezier) = \max[0,\ (6-1)/6] = 0.83$

- **Préfixes similaires :** dans un deuxième temps, nous appliquons une analyse lexicale pour vérifier si deux termes ont le même préfixe. Dans le cadre du traitement des variantes d'une ESA, nous avons remarqué qu'en général, les utilisateurs utilisent les mêmes préfixes que les ESA standards afin qu'elles restent compréhensibles par d'autres utilisateurs (par exemple, pour « Montpellier » : « montpelie », « montpel », « monpellier », etc.). Dans ce sens, nous faisons l'hypothèse que deux mots qui n'ont pas le même préfixe, c'est-à-dire qui ne commencent pas avec les mêmes caractères, ne représentent pas un même lieu. En effet, il est possible malgré tout qu'ils soient proches selon la mesure de similarité mais avec une signification très différente. En prenant les exemples (SMS 3) et (SMS 14), les deux mots « lattes » et « pattes » sont considérés comme proches selon la mesure SM. Cependant le mot « lattes » représente une ESA et « pattes » est sémantiquement très éloigné.

  Nous pouvons également remarquer que le nombre de caractères (taille) du préfixe dépend généralement de la taille de l'ESA. Les utilisateurs abrègent les mots pour gagner du temps et de l'espace. En nous appuyant sur ce constat, nous classons les ESA simples (comprenant un seul mot) dans l'une des deux catégories *long* et *court* en fonction de leur taille. Dans le cas des ESA dites courtes (e.g. « Lattes », « Sète », « Béziers », etc.), un mot est raccourci à sa première syllabe, la plupart des modifications sont faites à la fin du mot. En revanche, pour les ESA dites longues (e.g. « Montpellier », « Odysseum », « Frontignan », etc.), un mot est raccourci à ses premières lettres.
  Sur la base d'une analyse qualitative, qui vise à donner sens et à comprendre les résultats obtenus par notre système, nous constatons qu'une ESA courte contient au maximum 7 caractères, et généralement les modifications sont apportées aux derniers caractères (e.g. « lattes » et « latte » issus des exemples (SMS 3) et (SMS 7)). Par ailleurs, dans une ESA longue supérieure ou égale à 7 caractères, les modifications sont apportées sur l'ensemble du mot hormis

les deux premières lettres généralement (e.g. « montpellier » et « motpellier » voir exemples (SMS 1) et (SMS 12)).

Au regard de cette analyse préliminaire réalisée avec nos jeux de données, nous considérons que si la taille de l'ESA est inférieure à 7 caractères, nous fixons la taille du préfixe à 4 caractères. Sinon, nous la fixons à 2 caractères.

En ce qui concerne les ESA composées standards (par exemple, « Saint Éloi, Saint Jean, Saint Eustache » qui comprennent plusieurs mots), nous appliquons le même processus que les ESA simples. Cependant, nous vérifions la taille du premier mot de l'ESA candidate (par exemple, les mots « st », « Ste » et « sain » présents comme premiers mots dans les ESA candidates « st eloi », « Ste Eustache » et « sain jean »). Si la taille du premier mot contient moins de 5 caractères et qu'il commence par la même lettre que l'ESA composée standard [9] (par exemple, « st » et « Saint »), nous calculons la similarité entre le reste de l'ESA candidate et l'ESA standard (par exemple, le reste de l'ESA standard « Saint Eloi » est « Eloi » et le reste de l'ESA candidate « st eloi » est « eloi »).

Nous verrons dans les expérimentations menées et rapportées en Section 3.4 que ces traitements améliorent significativement la qualité de l'extraction des ESA.

3.4 Expérimentations

Dans cette section, nous détaillons une série d'expérimentations permettant d'évaluer les méthodes d'identification automatique des ESA et de leurs nouvelles formes d'expression.

3.4.1 Présentation des données

Deux corpus constitués de messages courts en français ainsi qu'un corpus d'article de presse ont été choisis (cf. Chapitre 3, Section 3.4.1). Dans ce chapitre, nous évaluons notre approche, identification de nouvelles formes d'expression d'ESA, sur un échantillon de 1000 SMS et 1000 tweets, sélectionnés de façon aléatoire. Puis, nous mesurons la qualité de nos résultats en nous appuyant sur un extrait d'articles de journaux, choisi au hasard, fourni par la presse française. Ce corpus, contenant 173 ESA distinctes, est lié à la région du bassin de Thau (sud de la France)

Un lexique initial d'ESA nommé *Dic-ES* (Zenasni et al., 2017b) est utilisé pour la tâche d'extraction. Le dictionnaire contient un ensemble de noms de lieux à partir des listes fournies par (1) la métropole de Montpellier (rues, quartiers, etc.), car le corpus SMS est lié à la zone géographique autour de Montpellier ; (2) la métropole européenne de Lille, car le corpus de tweets est lié à la zone géographique autour de Lille ; (3) les noms de pays et les capitales de chaque pays. Le dictionnaire

9. Sans prendre en compte les majuscules/minuscules.

contient également une description complète de plus de 8000 éléments géographiques en France concernant les unités administratives, les rivières, les noms des villes et communes, les quartiers, les lieux culturels répertoriés, les stations de transport en commun, etc.

Nous présentons, dans la sous-section suivante (cf. Section 3.4.3), le protocole expérimental adopté ainsi que les résultats obtenus pour la phase d'extraction d'ESA standards et de nouvelles formes d'expression.

3.4.2 Présentation du protocole expérimental

Les résultats sont évalués en termes de macro-moyenne et micro-moyenne, toutes deux faisons intervenir les mesures de précision, rappel et F-mesure (Asch, 2013). Chaque ESA standard est considérée comme une classe (e.g. Montpellier ⇒ motpellier, montpelie, monptellier, etc.).
Le rappel et la précision ont été introduits par Kent et al. (1955). Ces mesures évaluent la capacité du système à identifier les informations pertinentes. Afin de faciliter la compréhension de ces mesures, nous commençons par définir les quatre cas possibles pour identifier une ESA :

1. Vrais positifs (True Positive « TP ») : les ESA pertinentes correctement identifiées par le système ;
2. Faux positifs (False Positive « FP ») : les ESA non pertinentes identifiées par le système ;
3. Vrais négatifs (True Negative « TN ») : les ESA non pertinentes correctement non identifiées par le système ;
4. Faux négatifs (Flase Negative « FN ») : les ESA pertinentes non identifiées par le système.

Nous pouvons ainsi définir le rappel et la précision :

La précision mesure la capacité du système à refuser les ESA non-pertinentes. Elle permet de calculer le pourcentage des ESA pertinentes correctement identifiées par le système (TP) par rapport à toutes les ESA identifiées par le système ($TP + FP$).

$$Précision = \frac{TP}{TP + FP} \tag{3.6}$$

Le rappel mesure la capacité du système à identifier toutes les ESA pertinentes. Il permet de calculer le pourcentage des ESA pertinentes correctement identifiées par le système (TP) par rapport à toutes les ESA pertinentes identifiées par les experts ($TP + FN$).

$$Rappel = \frac{TP}{TP + FN} \tag{3.7}$$

L'utilisation du rappel sans la précision ou de la précision sans le rappel peut fausser l'évaluation des résultats. En effet, nous pouvons obtenir un rappel maximal en extrayant toutes les ESA candidates, cependant, cette étape ajoute beaucoup de

bruit qui abaisse les résultats de précision. Afin d'éviter ce problème, la F-mesure combine ces deux métriques et introduit une nouvelle mesure d'évaluation :

$$F - Mesure = \frac{2.Précision.Rappel}{Précision + Rappel} \tag{3.8}$$

La macro-moyenne consiste à calculer la moyenne de la précision et du rappel dans la phase de marquage des entités. Concrètement, nous calculons la précision et le rappel pour chaque ESA standard [10], puis nous calculons la moyenne pour chaque mesure. La macro-moyenne est utilisée pour donner un poids égal à chaque classe.

$$\begin{aligned} Macro - Précision &= \frac{1}{n} \sum_{ESA=ESA_1}^{ESA_n} Précision \\ Macro - Rappel &= \frac{1}{n} \sum_{ESA=ESA_1}^{ESA_n} Rappel \\ Macro - F - Mesure &= \frac{1}{n} \sum_{ESA=ESA_1}^{ESA_n} F - Mesure \end{aligned} \tag{3.9}$$

Où n est le nombre des ESA distinctes présentes dans le corpus.

La micro-moyenne calcule la précision et le rappel sur l'ensemble des instances d'entités. Concrètement, nous calculons *TP*, *FP* et *FN* pour toutes les ESA standards, puis nous calculons la précision et le rappel. La micro-moyenne est utilisée pour traiter les classes déséquilibrées.

$$\begin{aligned} Micro - Précision &= \frac{\sum_{ESA=ESA_1}^{ESA_n} TP_{ESA}}{\sum_{ESA=ESA_1}^{ESA_n} (TP_{ESA} + FP_{ESA})} \\ Micro - Rappel &= \frac{\sum_{ESA=ESA_1}^{ESA_n} TP_{ESA}}{\sum_{ESA=ESA_1}^{ESA_n} (TP_{ESA} + FN_{ESA})} \\ Micro - F - Mesure &= \frac{2 \times Précision \times Rappel}{Précision + Rappel} \end{aligned} \tag{3.10}$$

3.4.3 Découverte de nouvelles formes d'expression d'entités spatiales absolues

Pour mener à bien ces expérimentations, nous évaluons la capacité du système à identifier de nouvelles formes d'expression d'ESA.
La première ligne des Tableaux 3.1, 3.2, 3.3 et 3.4 (*Similarité de base*) montre les

10. Par exemple, le rappel d'une ESA donnée peut être égal à 1 si toutes les nouvelles expressions de l'ESA sont identifiées.

résultats obtenus lors de la première série d'expérimentations en appliquant respectivement les algorithmes SM et Lin uniquement. Les expérimentations intitulées *Similarité + caractéristiques lexicales* montrent les résultats obtenus en appliquant notre approche (suppression des accents et identification de préfixes similaires).

Trois utilisateurs[11] ont manuellement annoté toutes les ESA à partir du même échantillon de 1000 SMS. Chacun d'entre eux a identifié les mots susceptibles d'être une ESA en fonction des éléments contextuels du corpus.
À la suite du processus d'annotation, 67 ESA distinctes ont été identifiées manuellement dans le corpus de SMS. Ce résultat est obtenu en combinant les annotations des trois utilisateurs qui ont respectivement identifié 63, 61 et 59 ESA manuellement. En effet, les 67 ESA ont été identifiées en rassemblant les résultats obtenus de chaque utilisateur. Par exemple, un utilisateur parmi les trois a pu identifier l'expression « tahiti » (cf. Exemple (SMS 15) dans Section 3.1) comme ESA. Ceci nous a permis d'évaluer la qualité de notre méthode d'extraction des ESA variantes et d'intégrer les nouveaux éléments dans notre dictionnaire pour les contributions suivantes visant à extraire des relations entre entités nommées.

Une approche expérimentale a été adoptée pour sélectionner le seuil $S1$ le plus approprié, i.e. 0.60, 0.70, 0.80 et 0.90 (voir les Tableaux 3.1, 3.2, 3.3 et 3.4).

String Matching			
	Précision	Rappel	F-mesure
Similarité de base (S1>0.80)	0.32	0.77	0.44
Similarité + Caractéristiques Lexicales (S1>0.60)	0.20	**0.88**	0.33
Similarité + Caractéristiques Lexicales (S1>0.70)	0.51	0.87	0.64
Similarité + Caractéristiques Lexicales (S1>0.80)	**0.83**	**0.86**	**0.84**
Similarité + Caractéristiques Lexicales (S1>0.90)	**1**	0.73	0.84

TABLE 3.1 – Résultats liés à l'extraction des ESA sur un échantillon de 1000 SMS en terme de Micro-moyenne.

L'analyse de nos résultats montre que les seuils 0.80 et 0.90 donnent les meilleurs scores avec la mesure SM. Les seuils de 0.50 à 0.90 donnent les meilleurs résultats sur la mesure Lin avec une micro-F-mesure égale à 0.84. Cependant, la meilleure F-mesure est obtenue avec une valeur de seuil de 0.80 en utilisant la mesure SM. Nous pouvons remarquer que l'ajout des caractéristiques lexicales à la mesure de similarité augmente significativement la macro- et micro-F-mesure. En utilisant une mesure de similarité de base, le système donne une macro-F-mesure à 46% pour l'extraction de l'ESA. En revanche, en ajoutant les caractéristiques lexicales, ce pourcentage

11. Doctorants impliqués dans ce projet de recherche.

String Matching			
	Précision	Rappel	F-mesure
Similarité de base (S1>0.80)	0.32	0.86	0.46
Similarité + Caractéristiques Lexicales (S1>0.60)	0.18	**0.91**	0.30
Similarité + Caractéristiques Lexicales (S1>0.70)	0.46	**0.91**	0.61
Similarité + Caractéristiques Lexicales (S1>0.80)	**0.86**	**0.90**	**0.87**
Similarité + Caractéristiques Lexicales (S1>0.90)	**1**	0.77	0.87

TABLE 3.2 – Résultats liés à l'extraction des ESA sur un échantillon de 1000 SMS en terme de Macro-moyenne.

Lin			
	Précision	Rappel	F-mesure
Similarité de base (S1>0.80)	0.96	0.56	0.70
Similarité + Caractéristiques Lexicales (S1>0.40)	0.72	**0.79**	0.74
Similarité + Caractéristiques Lexicales (S1>0.50)	**1**	0.73	0.84
Similarité + Caractéristiques Lexicales (S1>0.80)	**1**	0.73	0.84
Similarité + Caractéristiques Lexicales (S1>0.90)	**1**	0.73	0.84

TABLE 3.3 – Résultats liés à l'extraction des ESA sur un échantillon de 1000 SMS en terme de Micro-moyenne.

Lin			
	Précision	Rappel	F-mesure
Similarité de base (S1>0.80)	0.96	0.68	0.79
Similarité + Caractéristiques Lexicales (S1>0.40)	0.81	**0.83**	0.81
Similarité + Caractéristiques Lexicales (S1>0.50)	**1**	0.77	0.87
Similarité + Caractéristiques Lexicales (S1>0.80)	**1**	0.77	0.87
Similarité + Caractéristiques Lexicales (S1>0.90)	**1**	0.77	0.87

TABLE 3.4 – Résultats liés à l'extraction des ESA sur un échantillon de 1000 SMS en terme de Macro-moyenne.

augmente jusqu'à 87% d'ESA correctement identifiées.

Par exemple, l'algorithme de base associe le toponyme « lattes » aux mots « pattes, « mattes, « battes, « nattes, « flattes, « luttes » et « latte », tandis que « Béziers » est associé à « beziers ». Après l'application de nos propositions, « lattes » est seulement associé à « latte » et « Béziers » est associé à « beziers » mais aussi à « bezier » et « bezies » qui sont des variantes tout à fait pertinentes.
Les résultats présentés dans les Tableaux 3.1, 3.2 3.3 et 3.4 confirment l'hypothèse suivante : combiner l'analyse statistique (mesures de similarités) et l'analyse lexicale (dictionnaires d'ESA, désaccentuation, préfixes similaires) permet d'améliorer les performances du système en terme de précision et de rappel.

Dans les Tableaux 3.5 et 3.6, nous évaluons l'approche sur un échantillon de 1000 tweets et nous comparons les résultats avec ceux obtenus sur le corpus SMS en utilisant la mesure SM. Notons que nous nous intéressons à identifier de nouvelles formes d'expression d'ESA, c'est-à-dire que nous cherchons à obtenir de bons résultats pour le rappel sans trop pénaliser la F-mesure. Nous choisissons la mesure SM parce qu'elle offre le meilleur score en terme de F-mesure avec un bon score du rappel (cf. Tableaux 3.1 et 3.2).
Sur l'échantillon de 1000 SMS, notre système a été capable d'identifier 46 ESA standards (par exemple, « Montpellier », « béziers », « saint éloi » ...) et 21 nouvelles formes d'expression d'ESA (par exemple, « montpelie », « bezier », « st-eloi », etc.) avec $S1 > 0.80$ (voir Section 3.3.2). Sur l'échantillon de 1000 tweets, notre système identifie 67 ESA standards et 23 ESA variantes avec le même seuil $S1 > 0.80$. Les résultats présentés dans le tableau 3.5, avec des scores de 0.84 et 0.86 en terme de micro-F-mesure pour les deux corpus SMS et tweets respectivement, montrent que l'utilisation de notre approche permet d'améliorer, de manière significative, l'identification des ESA dans les messages courts par rapport à l'utilisation simple des mesures de similarité.

		Précision	Rappel	F-mesure
Corpus de SMS	Similarité de base	0.32	0.77	0.44
	Similarité + Caractéristiques Lexicales	**0.83**	**0.86**	**0.84**
Corpus de tweets	Similarité de base	0.33	0.83	0.47
	Similarité + Caractéristiques Lexicales	**0.78**	**0.96**	**0.86**

TABLE 3.5 – Comparaison de l'extraction d'ESA entre les deux corpus SMS et tweets en utilisant SM avec $S1 > 0.80$ en terme de Micro-moyenne.

Nous menons des expérimentations approfondies pour évaluer la méthode proposée en la comparant à deux approches communément validées par la communauté

		Précision	Rappel	F-mesure
Corpus de SMS	Similarité de base	0.32	0.86	0.46
	Similarité + Caractéristiques Lexicales	**0.86**	**0.90**	**0.87**
Corpus de tweets	Similarité de base	0.36	0.91	0.51
	Similarité + Caractéristiques Lexicales	**0.78**	**0.98**	**0.86**

TABLE 3.6 – Comparaison de l'extraction d'ESA entre les deux corpus SMS et tweets en utilisant SM avec $S1 > 0.80$ en terme de Macro-moyenne.

scientifique : Stanford NER et Polyglot. Ces outils libres d'utilisation sont adaptés à la langue française.

Stanford NER (Finkel et al., 2005) est une bibliothèque Java open source pour la reconnaissance d'entités nommées (REN) de différents types (personne, organisation et lieu), l'un des outils de REN les plus populaires qui réalisent des performances robustes dans différents domaines sur différents types de corpus (e.g. articles de presse, articles de Wikipédia, etc.) (Luo et al., 2015 ; Yang et al., 2011). Cette bibliothèque est également connue sous le nom de CRF Classifier. Le logiciel Stanford NER utilise une approche statistique qui nécessite un apprentissage sur un corpus manuellement annoté pour fournir les résultats de reconnaissance optimaux pour une collection de textes donnée (Neudecker, 2016). Nous avons utilisé le classifieur formé pour Stanford NLP Europeana Newspapers [12]. Le projet Europeana Newspapers a produit un corpus de 400 pages de journaux en néerlandais, français et allemand sélectionnés et annotés manuellement avec des entités nommées.

Polyglot (Al-Rfou et al., 2015) est un pipeline de chaînes de TALN qui supporte de multiples applications multilingues. C'est un système qui créé des annoteurs de REN pour 40 langues principales en utilisant Wikipédia et Freebase [13].

Les Tableaux 3.7 et 3.8 et les Figures 3.4 et 3.5 mettent en avant les études comparatives menées sur les corpus SMS et tweets. Nous montrons que notre méthode donne les meilleures performances pour l'extraction d'ESA en terme de F-mesure pour les deux corpus, avec un score de 0.84 pour les SMS et 0.86 pour les tweets, alors que Polyglot donne des scores respectifs de 0.33 et 0.54, et Stanford NER donne un score de 0.26 pour les SMS et 0.31 pour les tweets.

Les figures 3.4 et 3.5 montrent que notre approche, nommée *Notre méthode-Similarité + Caractéristiques Lexicales*, fournit les meilleurs résultats en termes de précision et rappel et permet de détecter une information pertinente sur les deux corpus SMS et tweets en comparaison des autres approches. Nous constatons également que Stanford NER est plus performant concernant le rappel que Polyglot,

12. http ://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/
13. https ://sites.google.com/site/rmyeid/projects/polylgot-ner

	Micro-moyenne		
	Précision	Rappel	F-mesure
Stanford NER	0.19	0.40	0.26
Polyglot	0.56	0.23	0.33
Notre méthode-Similarité de base	0.32	0.77	0.44
Notre méthode-Similarité + Caractéristiques Lexicales	**0.83**	**0.86**	**0.84**

TABLE 3.7 – Comparaison des différentes approches de REN sur le corpus de SMS.

	Micro-moyenne		
	Précision	Rappel	F-mesure
Stanford NER	0.22	0.51	0.31
Polyglot	0.76	0.43	0.54
Notre méthode-Similarité de base	0.33	0.83	0.47
Notre méthode-Similarité + Caractéristiques Lexicales	**0.78**	**0.96**	**0.86**

TABLE 3.8 – Comparaison des différentes approches de REN sur le corpus de tweets.

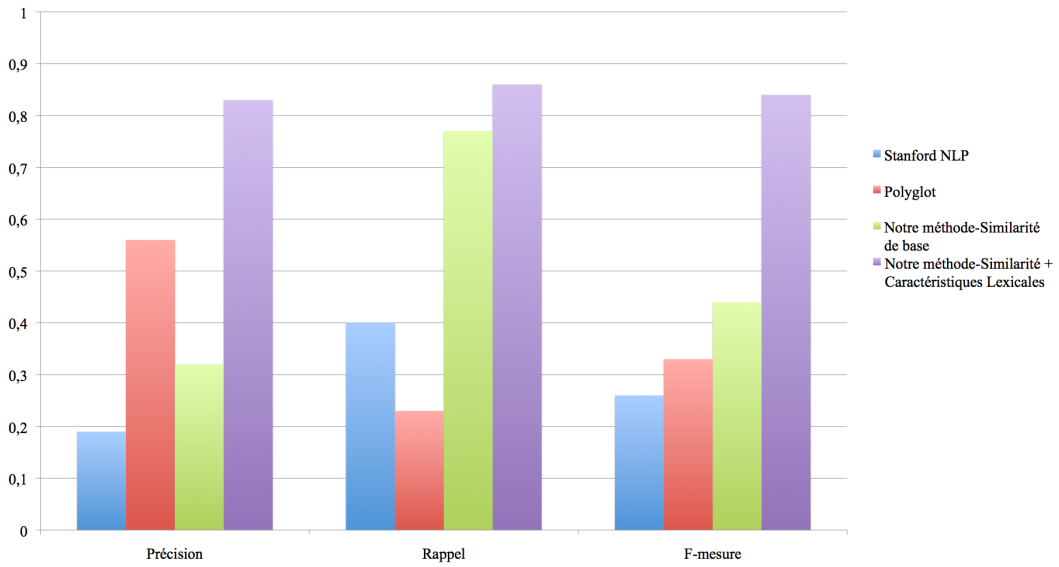

FIGURE 3.4 – Résultats liés à l'extraction d'ESA sur un échantillon de 1000 SMS.

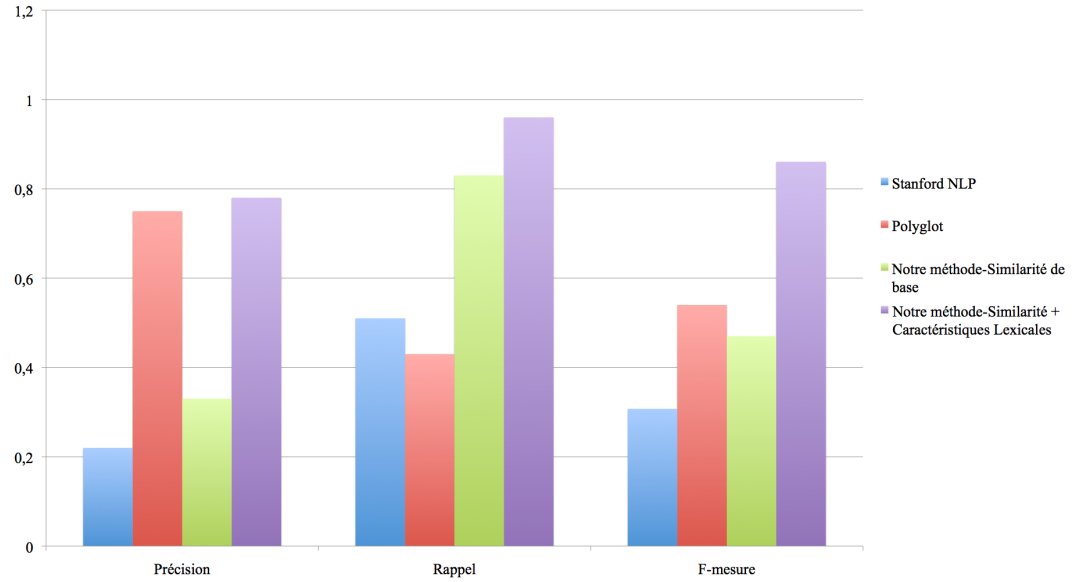

FIGURE 3.5 – Résultats liés à l'extraction d'ESA sur un échantillon de 1000 tweets.

tandis que ce dernier donne des meilleurs résultats en terme de précision.

Les résultats obtenus avec Stanford NER et Polyglot peuvent s'expliquer par la spécificité de nos corpus. Par conséquent, nous proposons de comparer et discuter les résultats à partir d'un corpus d'articles de la presse locale (journal Midi Libre).

Les résultats des expérimentations effectuées sur ce corpus standard, sur un échantillon contenant 173 ESA distinctes, sont présentés dans le Tableau 3.9. L'annexe .1 présente un extrait de cet échantillon. Notre système obtient 0.86 en terme de micro-précision en utilisant la mesure SM avec le seuil $S1 > 0.80$. En appliquant ce seuil, notre approche permet d'identifier des variantes d'ESA plus ou moins complexes (e.g. « Le Cap d'Agde » et « Le Grau d'Agde » (cf. Exemple Midi Libre (10) Annexe .1) à l'inverse des outils Stanford NER et Polyglot. La plupart des erreurs détectées sont liées aux mots très similaires qui sont considérés comme de nouvelles formes d'expression d'ESA (e.g. « teyron » qui est incorrectement associée à l'ESA standard « teyran » et « gigea » qui est incorrectement associée à l'ESA standard « gigean » (cf. Exemple Midi Libre (11) Annexe .1)). À noter que le rappel obtenu via notre approche est de 1, c'est-à-dire que les 173 entités spatiales pertinentes ont été extraites, à l'instar de Stanford NER et Polyglot qui retournent respectivement les scores 0.82 et 0.61.

Nous notons que Polyglot et Stanford NER donnent une performance plus élevée sur les corpus standards que sur les corpus de messages courts avec respectivement un score de micro-F-mesure de 0.61 et 0.68. Ceci est dû au fait que les modèles utilisés par Stanford NER et Polyglot sont construits sur des corpus standards (articles de journaux, articles de wikipedia, etc.). Ces résultats signifient que les modèles proposés ont un bon comportement sur les ensembles de données textuelles standards

mais pas sur des données textuelles moins standardisées. Comme le montre le Tableau 3.9, la meilleure performance en terme de F-mesure est de 0.93 obtenue avec notre méthode en combinant donc une approche de calcul de similarité de chaînes de caractères à une approche lexicale.

	Micro-moyenne		
	Précision	Rappel	F-mesure
Stanford NER	0.48	0.82	0.61
Polyglot	0.76	0.61	0.68
Notre méthode-Similarité de base	0.34	1	0.50
Notre méthode-Similarité + Caractéristiques Lexicales	**0.86**	**1**	**0.93**

TABLE 3.9 – Comparaison des différentes approches de REN sur le corpus Midi Libre.

En comparant les résultats obtenus par notre approche à ceux obtenus par Stanford NER et Polyglot, nous pouvons remarquer une amélioration de la précision et du rappel sur les différents extraits, ce qui tend à prouver que notre approche améliore globalement la qualité d'identification et d'extraction des ESA.

3.5 Conclusion

Dans cette étude, nous avons examiné différentes approches pour aborder la problématique de l'identification des entités spatiales absolues à partir des corpus de SMS et de tweets en combinant plusieurs approches de TALN.

À l'aide de dictionnaires d'ESA standards, nous avons combiné l'analyse statistique avec l'analyse lexicale pour extraire de nouvelles formes d'expression d'ESA. Nos expérimentations mettent en avant des scores d'identification d'ESA intéressants.

Nos résultats montrent que l'utilisation de notre approche permet d'améliorer de manière significative l'identification des ESA, avec un score de macro-F-mesure de 0.87 obtenu sur un corpus de 1000 SMS et de 0.86 sur un corpus de 1000 tweets. L'approche classique SM donne respectivement un score de macro-F-mesure de 0.46 et 0.51 (cf. Tableau 3.6).
En comparaison à l'état de l'art, notre méthode a obtenu la meilleure performance pour l'extraction des entités spatiales absolues en terme de F-mesure avec un score de 0.84 pour le corpus de SMS et 0.86 pour le corpus de tweets, alors que Polyglot obtient respectivement 0.33 et 0.54 et Stanford NER 0.26 et 0.31.

Nous remarquons enfin que les méthodes classiques de REN donnent de bonnes performances sur des corpus standards. Notre approche, combinant un calcul de similarité entre chaînes de caractères et une approche lexicale, donne également un

meilleur score de F-mesure sur un corpus d'articles de presse de Midi Libre annoté manuellement par des utilisateurs (0.93 contre 0.61 pour Stanford NER et 0.68 pour Polyglot). Les performances des outils de l'état de l'art se voient dégradées lorsqu'il s'agit de textes bruités (SMS et tweets). Cette dégradation est due notamment à la spécificité de ces corpus qui sont, par plusieurs aspects, différents des autres types de corpus (absence de majuscule, absence de ponctuation, répétition de caractères, suppression de caractères, la multitude d'expressions différentes pour exprimer une même entité spatiale absolue, etc.).

Malgré les bons résultats obtenus par notre système, le problème reste néanmoins loin d'être résolu. Notre meilleur système a toujours fait au moins quelques erreurs dans l'identification de nouvelles formes d'expression d'ESA. La plupart de ces erreurs étaient liées à :

1. L'abréviation très courte de l'ESA standard, ce qui complique l'identification (e.g. « monpel », « mtp » sont associées à l'ESA standard « Montpellier », « brdx » est associée à l'ESA standard « bordeaux »). Une solution, qui pourrait être intéressante serait de calculer la similarité entre les ESA et les mots du corpus sans considérer les voyelles (par exemple, calculer la similarité entre « brdx » après avoir supprimé les voyelles de l'ESA « bordeaux » et le mot « brdx » du corpus SMS) ;
2. les mots très similaires qui ont été considérés comme une nouvelle variante d'ESA (par exemple, « bassin » qui a été associé de manière incorrecte à l'ESA standard « Bassan »).

Ainsi, l'identification de la nouvelle variante et/ou standard d'ESA ne permet pas toujours de découvrir et désambiguïser l'information spatiale. Par exemple, dans les exemples donnés en section 3.1, les messages (SMS 15) et (SMS 16) contiennent l'expression « tahiti » mais elles ne font pas toutes deux références à une ESA, nous pouvons remarquer que l'EN « tahiti » dans l'expression « on part à tahiti » représente bien une ESA (cf. Exemple (SMS 15)). En revanche, la même EN dans l'expression « gel douche tahiti » représente un objet et non une ESA (cf. Exemple (SMS 16)). Enfin, cette première contribution ne permet pas toujours d'identifier de façon précise la sémantique de l'ESA, et par conséquent la bonne représentation spatiale associée. En effet, si nous prenons les exemples (SMS 12) et (SMS 19) listés en chapitre 1, l'ESA est bien identifiée par notre approche mais l'information extraite reste incomplète car nous ne prenons pas en compte les relations « sur » et « Dvant ». Ces relations spatiales apportent pourtant du sens qui permet de préciser les lieux mentionnés dans les messages. Dans le chapitre suivant (cf. Chapitre 4), nous proposons une approche pour identifier et extraire les relations spatiales précédant les ESA. Ces entités sont appelées Entités Spatiales Relatives (ESR).

Chapitre

4 Extraction des relations spatiales à partir de textes

Contents

4.1 Introduction

Dans le chapitre précédent, nous avons proposé une méthode qui permet de repérer et extraire de nouvelles variantes d'entités spatiales absolues (ESA) contenues dans des corpus de messages courts. Ce type d'information est constitué d'un nom toponymique et éventuellement d'un nom commun ayant un sens géographique qui le précède (par exemple, « Montpellier », « église Saint-Paul », etc.). L'identification des ESA joue un rôle important dans de nombreux traitements automatiques mais cela ne permet pas toujours de découvrir de façon précise et/ou désambiguïser l'information spatiale exprimée. De ce fait, de nombreux travaux, ont décrit la manière particulière d'exprimer l'information spatiale dans le langage naturel, comme ESA et entités spatiales relatives (ESR) (Egenhofer and Franzosa, 1991 ; Lesbegueries et al., 2006) . Une ESR est définie par une RS et une ou plusieurs ESA, comme par exemple « au nord de Paris » et « près de l'église Saint-Paul ». L'identification des RS fait l'objet de différents axes de recherche (cf. Chapitre 2, Section 2.3), la plupart d'entre eux se basent sur l'analyse de corpus standards. Exprimées en langage naturel, ces ESR nécessitent un traitement particulier.

Dans la communauté scientifique, une RS définit les indicateurs spatiaux d'ordre topologique, à proximité et directionnel (par exemple, « près de », « au sud de », etc.). Dans le chapitre 2, nous avons présenté les travaux définissant les différents types de RS, à savoir l'inclusion (« dans », etc.), la direction (« au nord », etc.), l'adjacence (« à côté de », etc.), la distance (« à X km de », etc.) et la relation géométrique (« entre X et Y », etc.). L'analyse de ces relations et l'identification de leur type qui structurent les différentes entités spatiales aident à la compréhension de l'information spatiale présente dans le texte, permettant ainsi d'identifier une représentation plus précise de l'entité spatiale exprimée.

À notre connaissance il n'existe pas de travaux visant à prendre en compte les particularités des corpus de messages courts. Par exemple dans l'expression « J'aime le parc au nord-ouest de Montpel », on ne fait pas référence à un parc se situant au centre ou au sud de la ville.

Dans ce chapitre, nous proposons deux principales contributions liées aux problèmes de la prédiction du type de RS et l'identification de nouvelles formes de RS dans les messages courts.

Première contribution

Dans la première partie de ce chapitre, nous posons l'hypothèse que les corpus de messages courts (SMS, tweets, etc.) contiennent de nouvelles formes/variantes de RS. L'extraction de ces relations à partir de ces corpus pourrait permettre d'être plus exhaustif dans l'identification de l'information spatiale exprimée, et ainsi plus complet à terme pour proposer une aide à la construction de la représentation spa-

tiale décrite dans ce type de corpus. Dans ce but, nous proposons une chaîne de traitements qui combine une analyse grammaticale (étiquetage grammatical) et une approche de fouille de textes (n-grammes de mots) afin d'identifier les différentes variantes de RS précédant les ESA à partir de corpus non-standards. Nous proposons dans un second temps des patrons généraux pour identifier ces relations. Cette phase nous permet d'enrichir un dictionnaire préexistant avec de nouvelles variantes de RS provenant de corpus de messages courts.

Deuxième contribution

Le travail présenté dans la deuxième partie de ce chapitre se situe dans le contexte global de la problématique d'identifier le type de RS parmi les trois types des RS que sont la région, la direction et la distance. L'objectif est de proposer une typologie de relations permettant d'identifier, finement et de manière automatique, les types des RS exprimés dans les textes. À cet effet, nous proposons une méthode hybride, combinant des informations lexicales et contextuelles à une approche de fouille de textes pour prédire le type de RS. Étant donné qu'il n'existe pas encore, à notre connaissance, de corpus français annoté par des types des RS, nous avons choisi d'utiliser un corpus anglais reconnu dans le domaine (Kordjamshidi et al., 2011).

Ce chapitre est organisé comme suit : la section 4.2.1 présente l'approche que nous mettons en œuvre pour la phase d'identification de nouvelles variantes de RS. La section 4.2.2 présente une évaluation des performances de notre approche à travers différentes expérimentations pour la phase d'identification de nouvelles variantes de RS. Nous détaillons, en section 4.3.1, les deux méthodes et leur combinaison pour la prédiction du type de RS identifiées. La section 4.3.2 décrit le protocole expérimental et les résultats obtenus pour la phase de prédiction du type de RS. La section 4.4 conclut le chapitre en discutant les résultats obtenus sur nos travaux.

4.2 Extraction de nouvelles variantes/formes de relations spatiales à partir de données textuelles non-standards

Dans la première partie de ce chapitre, nous proposons d'identifier de nouvelles formes de RS contenues dans les messages courts (cf. sous-sections 4.2.1.1, 4.2.1.2 et 4.2.2.3). Cette tâche bénéficie de la grande quantité de connaissances géographiques exprimées dans divers textes en langage naturel afin d'obtenir une vision plus complète de la connaissance spatiale.
Le Tableau 4.1 présente un récapitulatif des paramètres utilisés dans les différentes

approches définies dans cette section pour l'identification de nouvelles formes/variantes de RS dans les messages courts. .

Paramètres	Définition
m	Le nombre des étiquettes les plus fréquents qui précédent les ESA
$S2$	Seuil pour la sélection des mots les plus fréquents des « m » étiquettes
N	paramètre de fenêtrage pour sélectionner les n mots précédant les ESA

TABLE 4.1 – Tableau récapitulatif des paramètres associés à la deuxième contribution.

4.2.1 Méthodologie

Afin d'extraire les RS, nous exploitons un dictionnaire initial regroupant les cinq formes de RS que sont l'inclusion (« dans », etc.), la direction (« au nord », etc.), l'adjacence (« à côté », etc.), la distance (« à X km de », etc.) et la relation géométrique (« entre X et Y », etc.) (Lesbegueries et al., 2006).

Dans l'objectif d'identifier de nouvelles formes de RS, nous proposons une approche composée de trois étapes décrite Figure 4.1. La première étape consiste à identifier de nouvelles RS précédant les ESA (par exemple, les RS « sur », « avant », etc.). Puis, la deuxième étape vise à identifier de nouvelles formes d'expressions de RS présentes dans les corpus de messages courts (par exemple, les RS « pres », « àu », etc.). Finalement, la troisième étape correspond à déterminer pour chaque RS la syntaxe la plus pertinente (par exemple, « a coté PRP », « pres PRP », etc.).

Dans la suite de cette section, nous illustrons nos propos à partir des exemples suivants, provenant de la liste d'exemples présentés en Chapitre 1, Section 1.1.3 (les nouvelles formes de RS sont en gras).

(SMS 19) *Pu... **Dvant belevile** ya une riviere avc les ècler sa fai flipè*

(SMS 20) *Nous avons laissé à son stage de musique. et moi poursuivons **sur Montpellier**. Nous avons passé le 1er bouchon à **Valence** et roulons vers les prochains : **Montelimar** et **avant Montpellier !!!***

(SMS 21) *Nous sommes arretés sur la voie, juste **avant la gare***

(SMS 22) *Oui je dors bd voltaire et oui on passe **par st lazar**, pk ? Tu y va vers quelle heure toi ?*

(SMS 23) *Ça marche pour 19 heures ou plus tard. Rdv devant la grosse tête **près de l'église Ste Eustache** ?*

(SMS 24) *Fin c'est pas sur car **collioure** c'est plus **pres de montpellier** :) t'es deja allée la bas ?*

(SMS 25) *C sur ! Je lui ai demandé ou il habitait exactement je le pensais pas **loin de gueret**. Il est **a coté de masgot a 20min de gueret** ms de l autre coté ! Je lui ai dit que ca faisait loin du coup si jms il devait venir me chercher. Il*

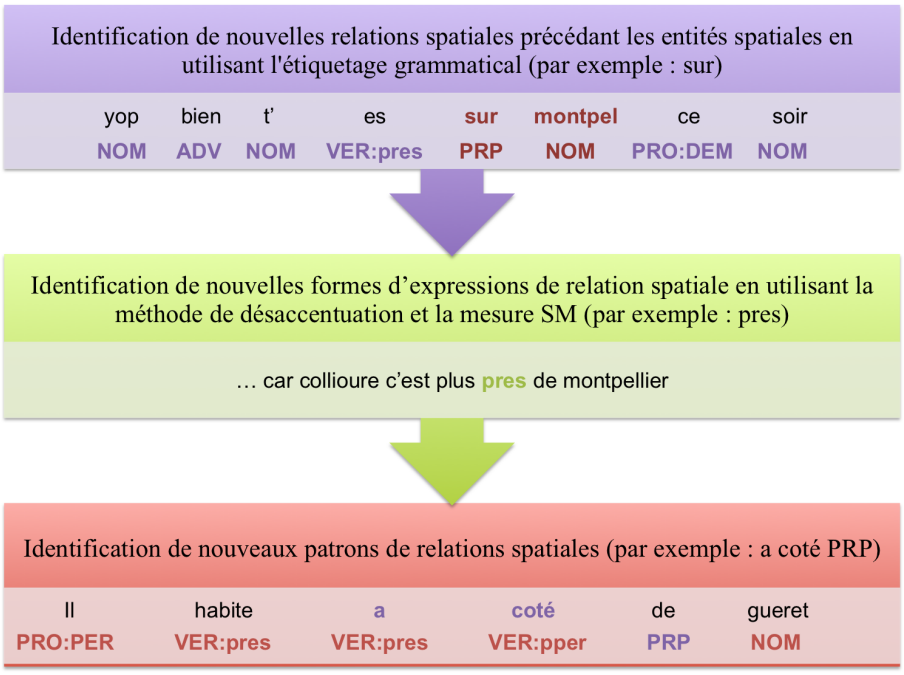

FIGURE 4.1 – Processus d'identification de nouvelles RS.

a dit qu un de ses amis etait encore plus loin et n hesitait pas a aller voir sa "chérie" s il fallait

(SMS 26) *Normalement je vais l'avoir a 41. J'espère ! Je suis juste* ***a cote de la gare****, mais j'attend qu'on me fasse ma prise de sangg*

(SMS 27) *Tu crois tu pourrai me rejoindre au creps* ***à coté du stade philippides***

(SMS 28) *Ce soir y à une white party que des teissiers* ***à coté de l'église saint roch***

(SMS 29) *j'suis deg c l'anniv de mariage de <PRE_5/> en* ***suisse****, c prevu depuis 6mois ! ! !*

(SMS 30) *je* ***suis ds le parc***

(SMS 31) *Cern* ***à 731 km de distance en Suisse***

(SMS 32) *j'ai vu que <PRE_6/>* ***est originaire de Beaumont****, qui se trouve a 20 min de chez moi*

(SMS 33) *le <PRE_4/> soient rentabilisés en le faisant payer des impots s'il* ***bosse en France***

(SMS 34) *je dois recup mon frère sur la route il* ***bosse a portel***

(SMS 35) *Je suis* ***a 5 min de la gare***

(Tweet 1) Jv aller fr un tour **dvant gambetta** lundi

(Tweet 3) pecnologic puis mes amis qui ont ét'e dans d'autres villes **avant tokyo** ont vraiment détesté tokyo je suis pas des pires tbh ça va

(Tweet 9) t'habites dans une ville inconnue **a coté de Lille** kestu parles

(Tweet 12) Demain je vais aller faire un ptit tour au nouveau magasin Nike **a cote du grand stade**

(Tweet 13) Qui serais me depaner paypal qui habite **pres de valencienne** que je lui rendre l'argent en IRL SVP c'est urgent

(Tweet 15) Malik il avait etait monstrueux **àu Sébastopol**

4.2.1.1 Identification de nouvelles variantes de relations spatiales

Notre première étape consiste à enrichir le dictionnaire initial des RS avec de nouvelles formes de RS présentes dans les corpus de messages courts en combinant une analyse morphosyntaxique (étiquetage grammatical) et une approche fréquentiste (nombre d'occurrences).

L'étiquetage grammatical est une des tâches importantes dans le traitement du langage naturel. L'identification de l'information grammaticale de chaque mot est une ressource précieuse lors de l'extraction d'informations à partir de corpus textuels (Bruce, 2012).

L'étiquetage grammatical est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes (par exemple, le nom commun, l'adjectif, etc.). Plusieurs méthodes ont été proposées pour étiqueter automatiquement les textes (Schmid, 1994 ; Toutanova and Manning, 2000). Cependant peu d'approches s'intéressent à l'étiquetage de textes non-standards. Nous pouvons citer cependant les travaux de Gimpel et al. (2011) et Derczynski et al. (2013) qui ont conçu un POS tagger (part-of-speech tagger) pour des données de messages courts en anglais provenant de Twitter. À notre connaissance, il n'existe pas de POS tagger développé pour traiter des messages courts en français. Fort de ce constat, nous avons adopté TreeTagger (Schmid, 1994) qui donne des résultats pertinents sur d'autres types de corpus standards en langue française (Corpus de tests de laboratoire (Lossio-Ventura et al., 2016), corpus MULTITAG (Allauzen and Bonneau-Maynard, 2008), etc.).

Une fois le corpus étiqueté, nous sélectionnons l'étiquette du mot précédant l'ESA. Nous calculons ainsi, pour l'ensemble des ESA, la fréquence de chaque étiquette. Ensuite, nous faisons l'hypothèse que les mots associés aux m [1] étiquettes les plus fréquentes sont retenus comme des RS candidates. Le m le plus approprié est sélectionné en fonction des résultats (voir les expérimentations en Section 4.2.2.1). Puis, nous sélectionnons les mots les plus fréquents associés aux m étiquettes sélectionnées sur la base d'un seuil $S2$ (les résultats liés à ce seuil sont discutés en Section 4.2.2.1). Les mots sélectionnés sont alors ajoutés comme nouvelles variantes de RS.

1. m représente le nombre des étiquettes les plus fréquentes, il varie entre 1 et 3 étiquettes.

Figure 4.2 présente un récapitulatif de processus d'identification de nouvelles RS.

À partir de l'exemple de la Figure 4.2, nous avons obtenu pour les étiquettes *PRP*[2] et *Nom* qui précèdent les ESA, un nombre d'occurrences égal à 34 et 31 respectivement.
Puis, les étiquettes *PRP* et *Nom* sont sélectionnées comme les deux plus fréquentes. Ensuite, nous avons obtenu les relations « à », « de », « sur », etc. et les relations « dvant », « avant », etc. comme des relations candidates liées aux étiquettes *PRP* et *Nom* respectivement. Finalement, nous avons sélectionné les relations les plus pertinentes en fonction du seuil $S2$ correspondant au nombre d'occurrences à partir duquel une variante est estimée valide.

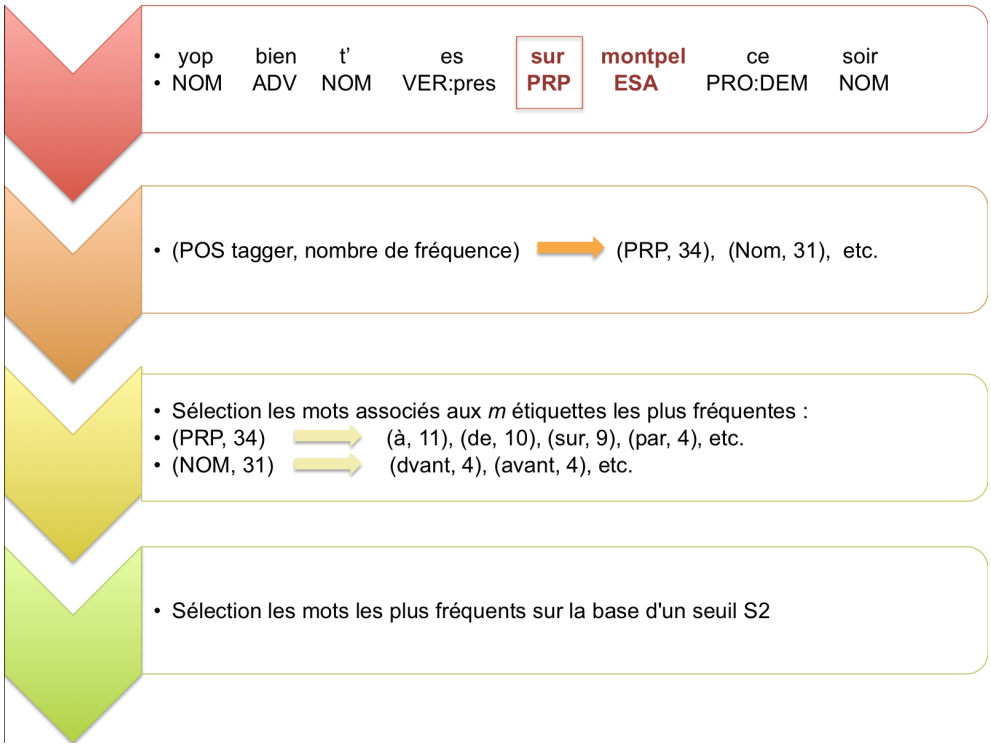

FIGURE 4.2 – Processus d'identification de nouvelles RS

4.2.1.2 Identification de nouvelles formes d'expression de relation spatiale

Dans l'objectif d'enrichir la liste des RS obtenues, nous cherchons maintenant à identifier de nouvelles formes d'expression de ces relations en écriture propre aux messages courts. Pour cela, nous calculons tout d'abord la similarité entre chaque

2. Préposition

RS et les N mots qui précèdent les ESA. Nous utilisons, comme liste d'entrée, une liste de RS enrichie par les nouvelles RS obtenues à l'aide de la méthode décrite dans la section précédente. Pour cette étape, nous réalisons, de la même manière que l'approche proposée dans le chapitre précédent, une première action de préparation du corpus en appliquant le module de désaccentuation. Puis nous utilisons à nouveau la mesure de similarité SM. Nous obtenons en résultat une liste de RS enrichies des candidats proches selon la mesure SM. Par exemple, en appliquant notre approche (désaccentuation + SM), nous pouvons dire que la relation candidate « pres » présente dans l'exemple (SMS 24) est proche de la relation initiale « près » (cf. Exemple (SMS 23)).

4.2.1.3 Identification de nouveaux patrons de relations spatiales

Dans le cas des messages courts, nous ne pouvons pas être assurés que la syntaxe et l'orthographe des RS seront utilisées en respectant les règles de la langue française classique. En effet, les règles d'orthographe sont régulièrement contournées (par exemple, « coté », « coté de » et/ou « a coté de »). Pour identifier l'interprétation morpho-syntaxique la plus pertinente pour chaque RS, nous proposons d'exploiter une version adaptée de la méthode C-value (Frantzi et al., 2000). Cette méthode vise à améliorer l'extraction des termes imbriqués, elle a été spécialement conçue pour l'extraction de termes constitués de plusieurs mots. En général, l'utilisation seule de la fréquence des occurrences pour extraire la relation « coté », qui est une sous-chaîne de « coté de » et « a coté de », n'est pas efficace. Dans ce cas, la fréquence de la relation « coté » est au moins aussi élevée que la somme des fréquences de toutes les relations les plus longues « coté de » et « a coté de ».

C-value offre une solution à ce problème : si une sous-chaîne comme « coté » a une fréquence du même ordre que « coté de », cette dernière chaîne est favorisée.

C-value (Frantzi et al., 2000)

C-value vise à combiner l'information linguistique et l'information statistique. L'information linguistique consiste à extraire les termes candidats du corpus en utilisant des filtres linguistiques basés sur l'étiquetage grammatical et une liste de mots vides. L'information statistique consiste à utiliser la mesure C-value, qui calcule le *termhood*[3] d'un terme en utilisant des caractéristiques statistiques (fréquence).

L'équation 4.1 présente la mesure de C-value (Frantzi et al., 2000). La mesure est construite en utilisant les caractéristiques statistiques de la relation candidate a. Ceux-ci sont : $|a|$ est la longueur de la relation candidate (nombre de mots), $f(a)$ est la fréquence des occurrences de a dans le corpus, T_a est l'ensemble des candidats extraits qui contiennent a, $P(T_a)$ est le nombre de relations candidates dans T_a, b

3. *termhood* sert à mesurer le degré qu'un terme est lié à des concepts à un domaine spécifique (Kageura and Umino, 1996)

est la relation candidate la plus longue, $f(b)$ est la fréquence des occurrences de b dans le corpus.

$$\text{C} - \text{value(a)} = \begin{cases} \log_2|a| \cdot f(a) & a \notin \textit{imbriquée} \\ \log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & \textit{sinon} \end{cases} \quad (4.1)$$

EC-value

Dans cette thèse, nous proposons d'étendre et adapter C-value à notre problématique. La *partie linguistique* (identification des candidats) de l'algorithme est modifiée. La liste des termes candidats est générée en combinant des n-grammes [4] et l'étiquetage grammatical (cf. Section 4.2.1.1). Concrètement, nous identifions les bi-grammes et tri-grammes contenant les RS précédemment détectées dans les messages courts.
À partir des exemples (SMS 27), (SMS 28) et (Tweet 9), les bi-grammes basés sur la RS initiale « coté » sont « coté de, coté du ». Et les tri-grammes basés sur la même RS sont « à coté du, à coté de, a coté de ».

Par la suite, nous considérons une paire *relation initiale et entité spatiale absolue* (RS, ESA). Puis, nous sélectionnons l'étiquetage grammatical des mots entre les deux éléments (RS et ESA). Par conséquent, un 3-tuple (RS, mot d'étiquetage grammatical, ESA) et 4-tuple (mot, RS, mot d'étiquetage grammatical, ESA) sont obtenus. Par exemple, dans la phrase (SMS 25) (cf. Section 4.2.1) « Il est **a coté de masgot** », nous obtenons un 3-tuple (coté, PRP, masgot) et un 4-tuple (a, coté, PRP, masgot)).
Ensuite, nous appliquons la mesure C-value afin de sélectionner le patron le plus pertinent.

En effet, dans l'approche statistique de la mesure C-value, une condition est ajoutée pour les relations qui n'apparaissent pas dans d'autres relations candidates plus longues, c'est-à-dire qui se trouvent directement avant l'ESA. Si la relation n'apparaît pas dans d'autres relations candidates, la relation sera ajoutée à la liste des relations finale.
Parmi les exemples présentés en Section 4.2.1, notre approche permet ainsi d'identifier les nouvelles formes de relation « **dvant** gambetta », « **sur** Montpellier », « **àu** Sébastopol », présentes respectivement dans les phrases (SMS 20), (Tweet 1) et (Tweet 15).

L'algorithme 1 présente la méthode de calcul C-value étendue que nous nommons **EC-value** (Extended C-value). Nous utilisons un seuil ($Threshold\_EC$) de C-value, et seules les chaînes avec EC-value au-dessus de ce seuil sont ajoutées à la liste des relations finale.

À noter que EC-value ne sélectionne pas toujours les formes les plus pertinentes de chaque RS. Par exemple, EC-value a sélectionné les deux patrons « coté PRP »

4. Un n-grammes est une séquence contiguë de n mots à partir d'une séquence de texte donnée, par laquelle les n-grammes sont formés (bi-grammes, tri-grammes, etc.).

Algorithme 1 : EC-value

Entrées : liste des termes candidats (terme candidat a)
Sorties : liste des termes les plus pertinents
pour chaque chaîne a qui n'apparaît pas dans d'autres termes **faire**
ajoutez a à la liste finale ;
fin pour
pour toutes les chaînes a de longueur maximale **faire**
calculer C-value(a) $= log_2|a| \cdot f(a)$
si C-value(a) $\geq Threshold\_EC$ **alors**
ajoutez a à la liste finale ;
fin si
fin pour
pour toutes les chaînes plus petites a en ordre décroissant **faire**
calculer C-value(a) $= log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b))$
si C-value(a) $\geq Threshold\_EC$ **alors**
ajoutez a à la liste finale ;
fin si
fin pour

et « a coté PRP » comme des patrons pertinents, voir la Section 4.2.2.3. Pour régler ce problème, nous proposons dans l'étape suivante d'améliorer la sélection des RS en utilisant l'algorithme SEC-value (Selection of Relevant Relations).

SEC-value

EC-value produit de bons résultats, néanmoins nous avons identifié du bruit lié à la présence de relations imbriquées. Ces relations sont présentes dans la liste des top k obtenues par EC-value. Par exemple, après avoir appliqué EC-value, nous avons obtenu les RS candidates « a coté PRP » et « coté PRP ». Cette dernière est une sous-chaîne de la RS candidates « a coté PRP ». Le but d'appliquer la EC-value est d'identifier, pour chaque RS, un seul patron qui est le plus pertinent (par exemple, « coté PRP » ou « a coté PRP »).
Une des solutions qui peut résoudre ce problème est d'affiner les résultats obtenus avec l'algorithme 1 (EC-value) en filtrant la liste de top k obtenu. Dans ce contexte, nous présentons l'algorithme 2 appelé SEC-value (Selection of Relevant Relations), qui utilise comme entrée la liste de top k ordonnée obtenue avec l'algorithme 1 (EC-value) et permet de sélectionner le patron le plus pertinent de chaque RS.

En effet, nous faisons l'hypothèse que si EC-value de la relation la plus longue (a) (e.g. a coté PRP) est supérieure à la relation imbriquée (b) (e.g. coté PRP), la relation imbriquée (b) est supprimée. Sinon, si la soustraction de la valeur de EC-value du terme imbriqué (b) et la valeur EC-value du terme le plus long (a) est

inférieure à 25% [5] de la valeur EC-value du terme imbriqué (b), le terme imbriqué (b) est supprimé. Sinon, le terme le plus long (a) est supprimé.

Par exemple, si EC-value (a coté PRP) = 104.60 et EC-value (coté PRP) = 94.66, alors la relation « coté PRP » est supprimée (cf. Tableau 4.5(b) Section 4.2.2.3).

Par exemple, si EC-value (a coté PRP) = 34.86 et EC-value (coté PRP) = 67, alors nous vérifions si [EC-value (a coté PRP)] - [EC-value (coté PRP)] <25% [EC-value (coté PRP)].
Si [EC-value (a coté PRP)] - [EC-value (coté PRP)] <25% [EC-value (coté PRP)] la relation « coté PRP » est supprimée sinon la relation« a coté PRP » est supprimée (cf. Tableau 4.5(a) Section 4.2.2.3). En considérant l'exemple précédant, nous avons obtenu pour $[67] - [34.86] < 25\%\ [67]$, $38.14 < 16.75$ ce qui signifie que « a coté PRP » sera supprimée.

Le processus est formalisé dans l'algorithme 2.

Algorithme 2 : SEC-value

Entrées : top k de la liste ordonnée des relations candidates ;
Sorties : $Patrons_{Gr}$: liste des patrons les plus pertinents ;

```
pour chaque chaîne imbriquée a apparaît dans d'autres relations b dans la liste
de top k faire
  si C-value(b) ≥ C-value(a) alors
    ajouter(b) dans Patrons_Gr ;
  sinon
    si [ C-value(a) ] - [ C-value(b) ] > 25% [ C-value(a) ] alors
      ajouter(b) dans Patrons_Gr ;
    sinon
      ajouter(a) dans Patrons_Gr ;
    fin si
  fin si
fin pour
```

Une caractérisation par des patrons représentatifs des RS dans les messages courts est importante afin de faciliter la tâche d'extraction de ces RS.
EC-value et SEC-value visent à améliorer l'identification des patrons les plus pertinents pour valoriser au mieux l'ensemble des informations dans cette nouvelle source. Elles sont conçues pour extraire les termes les plus longs. EC-value permet de sélectionner les termes candidats et attribuer un score de C-value à chaque terme.

5. Une étude est menée pour déterminer le seuil le plus approprié, c'est-à-dire 10%, 15%, 20%, 25% et 30%. Les mêmes résultats sont observés en utilisant 25% et 30%, pour cette raison, 25% est retenu pour nos expérimentations.

SEC-value permet de privilégier l'extraction des termes les plus longs, en pénalisant les termes imbriqués à l'aide d'un seuil.

Dans la section 4.2.2.3, nous décrivons le seuil utilisé, nos résultats et nous analysons enfin la pertinence de notre approche.

4.2.2 Expérimentations

Dans cette section, nous évaluons la capacité du système à identifier les RS et les nouvelles formes d'expression de RS dans les messages courts. Afin d'évaluer l'efficacité de notre système, nous évaluons notre approche sur deux types de corpus de messages courts non-standards, c'est-à-dire, SMS et tweets (cf. Chapitre 3, Section 3.4.1). L'évaluation des RS extraites reposent sur deux séries d'expérimentations.

4.2.2.1 Identification de nouvelles relations spatiales

La première série concerne l'identification de nouvelles RS en utilisant le dictionnaire d'ESA et l'étiquetage morpho-syntaxique. Nous avons fait varier le paramètre susceptible d'influencer les résultats (cf. Tableau 4.2), c'est-à-dire le seuil $S2$ relatif au nombre d'occurrences (cf. Section 4.2.1.1). Nous évaluons l'approche avec $S2$ allant de 0 à 10 en terme de Micro-Précision.

Le Tableau 4.2 montre les performances de notre système pour identifier de nouvelles RS dans les deux corpus pour les trois étiquettes les plus fréquentes (préposition, nom et verbe). Nous avons calculé le nombre d'occurrences des étiquettes des mots précédant l'ESA (cf. Section 4.2.1.1). Le Tableau 4.2 montre que si nous sélectionnons les mots avec la fréquence égale à 2, nous obtenons un score de Micro-Précision de 0.90 et 0.75 pour le corpus de SMS et de tweets respectivement. En effet, nous obtenons pour le corpus de SMS 9 relations pertinentes dont 5 sont de nouvelles relations (« sur, par, dvant, avant » et « a »). Et pour le corpus de tweets, nous obtenons 6 relations pertinentes dont 3 sont de nouvelles relations (« sur, dvant » et « avant »).

Les Tableaux 4.3 et 4.4 présentent un résumé des résultats obtenus dans cette section. Pour conclure, le meilleur résultat de F-mesure est obtenu avec $S2 = 2$ pour les corpus de SMS et de tweets. Les valeurs de la Micro-F-mesure sont respectivement 0.68 et 0.44 (cf. Tableaux 4.3 et 4.4). Concrètement, nous obtenons les nouvelles RS « sur », « par », « a », « dvant » et « avant » pour le corpus de SMS, et « sur », « par » et « dvant » pour le corpus de tweets. Nous obtenons des résultats similaires en terme de nouvelles relations spatiales identifiées en appliquant notre méthode sur deux types de corpus.

4.2.2.2 Identification de nouvelles formes d'expression de relation spatiale

Pour identifier les relations « dans, entre, direction, côté », etc. et de nouvelles formes d'expression de RS comme « coté, pres, àu » et « cote », une mesure de

(a) Dans le corpus de SMS.

$S2$	Micro-Précision	Relations pertinentes	Nouvelles relations	Relations non-pertinentes
10	0.66	2	1	1
8	0.75	3	2	1
5	0.75	3	2	1
2	**0.90**	**9**	**5**	**1**
0	0.42	12	6	18

(b) Dans le corpus de tweets.

$S2$	Micro-Précision	Relations pertinentes	Nouvelles relations	Relations non-pertinentes
10	0.66	2	1	1
8	0.66	2	1	1
5	0.75	3	1	1
2	**0.75**	**6**	**3**	**2**
0	0.23	10	5	33

TABLE 4.2 – Résultats d'extraction de nouvelles RS en termes de micro-précision (variation de S)

S2	Micro-Précision	Micro-Rappel	Micro-F-mesure
10	0.66	0.13	0.21
8	0.75	0.19	0.30
5	0.75	0.19	0.30
2	**0.90**	**0.56**	**0.68**
0	0.42	0.68	0.51

TABLE 4.3 – Résultats liés à l'extraction de nouvelles RS à partir du corpus du SMS

S2	Micro-Précision	Micro-Rappel	Micro-F-mesure
10	0.66	0.10	0.17
8	0.66	0.10	0.17
5	0.75	0.16	0.26
2	**0.75**	**0.32**	**0.44**
0	0.23	0.52	0.31

TABLE 4.4 – Résultats liés à l'extraction de nouvelles RS à partir du corpus du tweets

similarité basée sur la méthode de désaccentuation est appliquée. Les paramètres utilisés dans cette section sont la mesure SM et le seuil $S1 > 0.80$ (cf. Chapitre 3, Section 3.4.3).

La liste enrichie des RS obtenues dans la section 4.2.2.1 avec le seuil $S2 > 2$ est utilisée comme liste d'entrée. Le Tableau 4.5 montre les résultats d'évaluation pour l'identification de nouvelles formes d'expression de RS. Les résultats sont calculés en termes de précision, rappel et F-mesure. Comme illustré dans le Tableau 4.5, l'utilisation de la mesure de similarité améliore l'identification de RS, en comparant au Tableau 4.3, avec un score de F-mesure de 0.97 et 0.68 respectivement pour le corpus de SMS.

Corpus	Micro-Précision	Micro-Rappel	Micro-F-mesure
Corpus de SMS	0.93	1	0.97
Corpus de tweets	0.89	0.94	0.91

TABLE 4.5 – Résultats liés à l'extraction de RS standards et variantes

Comme résultat des deux sections précédentes (Sections 4.2.2.1 et 4.2.2.2), deux nouveaux types de RS sont identifiés : par la reconnaissance de nouvelles formes scripturales (par exemple, « pres », « coté », « cote » et « àu », voir les phrases (SMS 24), (SMS 25) et (SMS 26)), et par extraction de vocabulaires spécifiques aux messages courts (par exemple, « sur », « par », « dvant », « a » et « avant », voir les phrases (SMS 20), (SMS 22), (SMS 19) et (SMS 21)).

4.2.2.3 Identification de nouveaux patrons de relations spatiales

Dans la deuxième partie des expérimentations, le système est testé en utilisant les dictionnaires enrichis de RS et ESA afin d'identifier les patrons des RS les plus appropriés.
Les Tableaux 4.6, 4.7 et 4.8 présentent un exemple de listes ordonnées pour les relations « coté », « pres » et « bord ». Ils donnent les valeurs de C-value pour tous les syntagmes (relations candidates) associés à ces relations initiales. En outre nous avons également noté les fréquences des sous-chaînes dans des termes candidats plus longs (deuxième colonne), le nombre de ces termes candidats plus longs (troisième colonne) et les fréquences totales (quatrième colonne).
À partir des exemples des Tableaux 4.6 et 4.8, 4 différents syntagmes (relations candidates) contiennent la RS initial « coté » (cf. Tableaux 4.6) et 2 différents syntagmes contiennent la RS initial « bord » (cf. Tableau 4.8) ont été extraites à partir de chaque corpus (SMS et tweets). Ces syntagmes sont considérés comme pertinents selon les scores élevés obtenus avec C-value. Tandis que, à partir de l'exemple du Tableau 4.7, nous remarquons que les syntagmes identifiés pour la RS initiale « pres »

ne sont pas tous pertinents. Par exemple, le syntagme « wasquehal pres PRP » qui a obtenu un score égale à 1.58 qui est un score très faible.

(a) Dans le corpus de SMS

C-value	$\sum f(b)$	P(Ta)	Freq	Relations candidates
69.73	0	0	44	à coté PRP
67.0	78	3	93	coté PRP
34.86	0	0	22	a coté PRP
19.01	0	0	12	du coté PRP

(b) Dans le corpus de tweets

C-value	$\sum f(b)$	P(Ta)	Freq	Relations candidates
104.60	0	0	66	a coté PRP
94.66	121	3	135	coté PRP
66.56	0	0	42	à coté PRP
20.60	0	0	13	du coté PRP

TABLE 4.6 – Résultats de C-value pour les relations candidates qui contiennent « coté »

(a) Dans le corpus de SMS

C-value	$\sum f(b)$	P(Ta)	Freq	Relations candidates
6.33	0	0	4	plus pres PRP
8	4	1	12	pres PRP

(b) Dans le corpus de tweets

C-value	P(Ta)	$\sum f(b)$	Freq	Relations candidates
34.66	4	3	36	pres PRP
3.16	0	0	2	habite pres PRP
1.58	0	0	1	wasquehal pres PRP
1.58	0	0	1	marins pres PRP

TABLE 4.7 – Résultats de C-value pour les relations candidates qui contiennent « pres »

Pour valider automatiquement nos relations candidates, nous avons réalisé une série d'expérimentations afin d'identifier le seuil le plus pertinent pour C-value. 109 relations sont identifiées comme candidates à enrichir la liste de relations avec le score de C-value supérieur à 0 dans le corpus de tweets et 90 sont identifiées dans le corpus de SMS. Si nous incluons également ceux dont la valeur de C-value est égal à 0, le nombre monte à 132 relations candidates pour le corpus de tweets et 113 pour le corpus de SMS. Et si le seuil est supérieur à 2, le nombre de relations

(a) Dans le corpus de SMS

C-value	P(Ta)	$\sum f(b)$	Freq	Relations candidates
14.26	0	0	9	au bord PRP
10	9	1	19	bord PRP

(b) Dans le corpus de tweets

C-value	P(Ta)	$\sum f(b)$	Freq	Relations candidates
166.42	0	0	105	au bord PRP
38.0	105	1	143	bord PRP

TABLE 4.8 – Résultats de C-value pour les relations candidates qui contiennent « bord »

candidates diminue à 106 pour le corpus de tweets et 85 pour le corpus de SMS. Nous choisissons 2 comme seuil de C-value (cf. Tableau 4.2), c'est-à-dire que les relations avec C-value supérieure à 2 seront incluses dans la liste finale. Les Tableaux 4.9 et 4.10 présentent les 10 meilleurs candidats de la liste ordonnée avec SEC-value (cf. Algorithme 2) et la fréquence pour les deux corpus SMS et tweets. SEC-value fournit une syntaxe plus complexe pour les relations, qui sont *RS + PRP* et *mot + RS + PRP*. À partir des Tableaux 4.9 et 4.10, nous obtenons respectivement les patrons suivants « à coté PRP », « autour PRP », « au fond PRP », etc. au lieu de « coté », « autour », « fond », et « a côté PRP », « autour PRP », « au bord PRP », etc. au lieu de « côté, autour », « bord ».

(a) SEC-value		(b) Fréquence	
C-value	Relations candidates	Fréquence	Relations candidates
205	avant PRP	18698	de
100	loin PRP	15065	est
69.73	à coté PRP	13430	a
46	près PRP	9869	à
43	coeur PRP	5224	au
36.45	à côté PRP	4055	sur
36	autour PRP	3441	dans
33	cote PRP	1499	par
28.52	au fond PRP	1056	avant
22.18	au milieu PRP	812	vers

TABLE 4.9 – Extrait de top 10 de liste ordonnée pour le corpus de SMS : SEC-value vs Fréquence

Les résultats sont évalués en termes de précision pour les k premiers candidats

(a) SEC-value		(b) Fréquence	
C-value	Relations candidates	Fréquence	Relations candidates
1188	avant PRP	181367	de
768.70	à côté PRP	128186	est
660	fond PRP	89302	a
615	loin PRP	76503	à
345	autour PRP	36332	dans
328	près PRP	33354	au
221.89	la sortie PRP	32402	sur
201.29	a côté PRP	9704	par
166.42	au bord PRP	5479	avant
152.15	au milieu PRP	2980	devant

TABLE 4.10 – Extrait de top 10 de liste ordonnée pour le corpus de tweets : SEC-value vs Fréquence.

ordonnés avec EC-value (P@k, où $k = 5$, 10 et 20) (cf.Tableau 4.11). Ils sont comparés aux résultats correspondants de SEC-value et de la fréquence. Le Tableau 4.11 montre la précision de la liste finale ordonnée extraite avec EC-value (cf. Algorithme 1) et SEC-value (cf. Algorithme 2).

Les évaluations montrent que l'utilisation de SEC-value donne les meilleurs résultats que les mesures EC-value et fréquence (cf. Tableau 4.11).

		P@5	P@10	P@20
SMS	$EC - value$	0.8	0.9	0.8
	$SEC - value$	**1**	**1**	**1**
	Fréquence	0.6	0.8	0.8
Tweet	$EC - value$	0.8	0.8	0.75
	$SEC - value$	**1**	**1**	**0.95**
	Fréquence	0.6	0.8	0.8

TABLE 4.11 – Résultats en termes de précision dans les corpus de SMS et de tweets.

4.2.2.4 Identification de relations spatiales à partir du corpus Midi Libre

Dans cette section, nous proposons de comparer et discuter les résultats obtenus sur les différents corpus non-standards (SMS et tweets) à ceux obtenus sur un corpus standard (articles de presse). Cela permet d'étudier la généricité de notre approche. Par conséquent, nous menons nos expérimentions avec un extrait des articles du

journal Midi Libre, sélectionnés aléatoirement, contenant 173 ESA (cf. Chapitre 3). En effet, un tel corpus est adéquat à notre cas d'étude, il permettra de connaître le vocabulaire spécifique ainsi que les particularités de la syntaxe des informations spatiales d'usages littéraires.
Pour identifier de nouveaux patrons de RS (cf. Section 4.2.2.3), la précision est calculée pour chacun des trois top k (P@k, où $k = 5$, 10 et 20) de EC-value. Ensuite, les scores de précision sont comparés aux résultats correspondants de SEC-value et de la fréquence. Le Tableau 4.12 présente le top 10 des relations candidates ordonnées par valeur de SEC-value et de fréquence d'apparition dans le corpus Midi Libre. Les résultats du top k relations sont évalués en termes de précision (cf. Tableau 4.13). Ces résultats confirment le bon comportement de notre mesure SEC-value.

Pour l'identification de nouvelles RS (cf. Section 4.2.1.1), le système identifie de façon erronée les termes « de, a, tram, du », etc. comme nouvelles relations. Les relations « sur, par, dvant », et « avant » (cf. Section 4.2.2.1) n'existent pas dans le corpus standard comme RS, montrant que les nouvelles RS identifiées en Section 4.2.2.1 sont liées aux corpus de messages courts.

(a) SEC-value.		(b) Fréquence.	
C-value	Relations candidates	Fréquence	Relations candidates
90	autour PRP	824	à
71.47	route PRP	557	est
40.33	près PRP	541	dans
34.66	centre PRP	169	entre
31.69	du côté PRP	104	centre
28.52	au coeur PRP	93	devant
26	bord PRP	91	autour
20.60	la sortie PRP	90	route
17.92	la direction PRP	81	vers
15.84	au milieu PRP	68	côté

TABLE 4.12 – Extrait de top 10 de liste ordonnée pour le corpus de Midi Libre : SEC-value vs Fréquence.

		P@5	P@10	P@20
Midi Libre	$EC - value$	1	0.8	0.6
	$SEC - value$	**1**	**1**	**1**
	Fréquence	0.8	0.9	0.85

TABLE 4.13 – Résultats en termes de précision du corpus Midi Libre.

Les résultats obtenus de SEC-value, sur le corpus de Midi Libre, montrent que SEC-value fournit des syntaxes plus complexe pour les RS qui sont *RS + PRP* et *mot + RS + PRP*.

4.2.2.5 Analyse des résultats

Dans cette section, nous avons proposé une méthode pour identifier de nouvelles RS et de nouvelles variantes de RS contenues dans les messages courts. Dans ce sens, la particularité de chaque corpus (standard et messages courts) est mise en évidence. Nos expérimentations nous ont conduit à plusieurs conclusions importantes.

Premièrement, la plupart de nouvelles variantes d'ESA dans le corpus standard apparaissent en raison d'erreurs d'orthographe, la non prise en compte des accents et l'utilisation de l'abréviation de « saint ». Par exemple, notre système à détecté les ESA « villen **ue**ve-lés-béziers », « saint-etienne » et « st-louis, st-andré » comme nouvelles variantes d'ESA. Cependant, de nouvelles variantes d'ESA sont automatiquement identifiées dans les corpus de messages courts comme « motpellier », « bezier », « st-eloi », etc.

Deuxièmement, notre méthodologie permet d'identifier de nouvelles relations dans les messages courts telles que « sur », « par », « dvant » et « avant » (cf. Exemples (SMS 20), (SMS 22), (Tweet 1) et (Tweet 3)).
Et troisièmement, trois catégories spécifiques existent pour les RS (cf. Tableau 4.14) : 1) RS liées au corpus de SMS ; 2) RS liées au corpus de tweets ; 3) RS liées au corpus de Midi Libre.

Relations liées au corpus de SMS	Relations liées au corpus de tweets	Relations liées au corpus de Midi Libre
à coté PRP	fond PRP	route PRP
coeur PRP	a côté PRP	centre PRP
a coté PRP	au bord PRP	du côté PRP
cote PRP		au coeur PRP
au fond PRP		bord PRP
		la direction PRP

TABLE 4.14 – Extrait des relations spécifiques associées à chaque corpus.

SMS ∩ tweets ∩ Midi Libre	SMS ∩ tweets	SMS ∩ Midi Libre	tweets ∩ Midi Libre
autour PRP	autour PRP	autour PRP	autour PRP
près PRP	près PRP	près PRP	près PRP
	avant PRP		la sortie PRP
	loin PRP		au milieu PRP
	à côté PRP		
	sur		
	dvant		
	avant		

TABLE 4.15 – Extrait des relations en commun dans les corpus.

Les résultats obtenus dans le Tableau 4.15 s'expliquent par la particularité de chaque type de corpus (corpus de messages courts et corpus standard). Dans les

messages courts, les utilisateurs ont tendance à utiliser une forme non-standard d'écriture qui s'affranchit souvent de certaines règles de grammaire, d'orthographe et des ponctuation (Choudhury et al., 2007). Les messages les plus courants concernent l'obtention d'informations pour rencontrer quelqu'un, trouver des endroits tels que des boutiques, des bibliothèques, etc. ou pour suggérer des endroits à certains comme des restaurants, des clubs de fitness, etc. Ainsi, les utilisateurs manipulent surtout des relations de type adjacence (par exemple, a coté PRP, à coté, au bord PRP, loin PRP, etc.) pour indiquer les localisations en utilisant des repères. Par exemple, « Il est **a coté de masgot a 20min de gueret** » et « tu pourrai me rejoindre au creps **à coté du stade philippides** » (cf. (SMS 25) et (SMS 26)).

En revanche, les articles de presse sont des publications rédigées, contenant des informations structurées rédigées en français élaboré destinées au grand public. C'est la raison pour laquelle il existe peu de relations communes extraites des corpus SMS et Midi Libre autres que les relations spatiales classiques. Les tweets sont diffusés publiquement sur la plate-forme Twitter, à des fins politiques, commerciales, et plus généralement pour transmettre de l'information au plus grand nombre. Cela explique le fait qu'une partie des usagers s'expriment avec un langage plus proche de celui de la presse et que notre approche identifie des formes de RS similaires dans les tweets et dans la presse Midi Libre.

Dans la section suivante, nous présentons une approche qui permet de prédire automatiquement le type de RS. La prédiction automatique du type de RS est une tâche importante car elle permet de comprendre l'information spatiale présente dans le texte. En effet, l'analyse des RS et l'identification de leur type permettent d'identifier une représentation plus précise de l'entité spatiale exprimée.

4.3 Prédiction du type de relations spatiales standards

Dans la section 4.3.1, nous proposons de comparer et adapter à notre problématique deux approches éprouvées de la littérature (*prédiction du type de relations spatiales par comparaison de chaînes de caractères* et *Prédiction du type de relations spatiales par proximité contextuelle*). Ces deux approches sont ensuite combinées (cf. Section 4.3.1.3). Le Tableau 4.16 présente un récapitulatif des paramètres utilisés dans cette section.

4.3.1 Méthodologie

Afin de prédire les types des RS exprimés dans les textes, nous proposons l'approche décrite en figure 4.3.

Paramètres	Définition
K	K classes majoritaires sélectionnées par la première méthode *Prédiction de type par comparaison de chaînes de caractères*
K'	K' classes majoritaires sélectionnées par la deuxième méthode *Prédiction de type par proximité contextuelle*
n	paramètre de fenêtrage pour sélectionner les n mots autour de la RS
m	le nombre de mots de la RS

TABLE 4.16 – Tableau récapitulatif des paramètres associés à la première contribution.

Premièrement, pour prédire le type de RS identifiées, nous utilisons la comparaison de chaînes de caractères entre la relation candidate et l'ensemble d'apprentissage à travers les méthodes String Matching (SM) (Maedche and Staab, 2002) et Lin (Lin, 1998). Rappelons que ces méthodes permettent de calculer la proximité lexicale entre deux chaînes de caractères. Ensuite, nous adoptons l'algorithme de fouille de données des K plus proches voisins ($KPPV$) (Bhatia and Vandana, 2010) pour retourner la classe majoritaire attribuée à chaque RS candidate. Le $KPPV$ attribue, pour chaque élément à classer, la classe majoritaire de ces K plus proches voisins. En effet, l'algorithme recherche les K voisins les plus proches parmi les éléments d'apprentissage pré-classés en fonction d'une mesure de similarité, à savoir, les K éléments les plus similaires. Puis, l'algorithme sélectionne la classe majoritaire.
Nous avons choisi $KPPV$ qui est un algorithme performant et simple à mettre en œuvre. Il est l'un des 10 meilleurs algorithmes de fouille de données (Wu et al., 2008). $KPPV$ est approprié à nos expérimentations car il permet une interprétation aisée des résultats d'apprentissage automatique (Al-Shalabi et al., 2006 ; Yang and Liu, 1999).

Deuxièmement, dans cette phase de prédiction du type de RS, nous proposons d'utiliser le contexte de la relation comme indicateur du type. Dans ce sens, les mots autour de la relation sont extraits et pondérés avec des mesures statistiques classiques que sont le nombre d'occurrences, TF-IDF (Salton and Buckley, 1988) et la confiance (Agrawal et al., 1993). Puis, nous appliquons l'algorithme de fouille de données $KPPV$ pour sélectionner la classe, c'est-à-dire le type de RS à attribuer.
Dans un troisième temps, afin d'améliorer le marquage des relations, nous proposons de combiner les deux premières approches, i.e. lexicale et contextuelle, et d'y attacher ensuite l'algorithme de $KPPV$.

Dans la suite de cette section, nous illustrons nos propos à partir des exemples suivantes extraites d'un corpus en langue anglaise SpRL [6] (Spatial Role Labeling) (Kordjamshidi et al., 2011). Le type de RS est mis en forme en gras.

6. http ://liir.cs.kuleuven.be/sprl/sprl.php

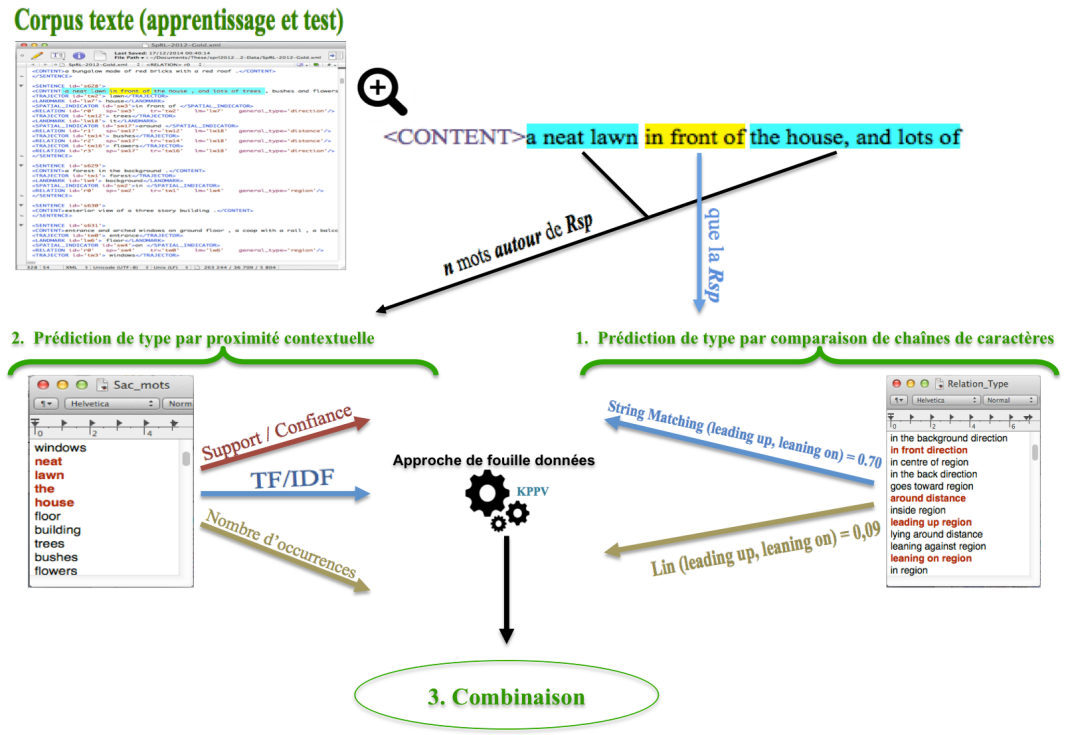

FIGURE 4.3 – Processus proposé pour la prédiction du type de RS.

(SpRL 1) *Stairs are **leading up** (type région) to the entrance.*

(SpRL 2) *Four locals are **sitting on** (type région) a bench in a canteen kitchen, **leaning on** (type région) a red brick wall.*

(SpRL 3) *movable wooden walls are **leaning on** (type région) the wall in the background .*

(SpRL 4) *About 20 kids in traditional clothing and hats waiting **on** (type région) stairs.*

(SpRL 5) *a dark-skinned , dark-haired boy in a grey and blue jumper is standing **next to** (type région) a desk in a classroom.*

(SpRL 6) *one kid is standing **next to** (type région) the table.*

(SpRL 7) *one girl is standing **at** (type région) her desk , another one is walking **to** (type région) her desk , three other boys are standing and looking at a picture of Minnie Mouse **on** (type région) the wall.*

(SpRL 8) *the courtyard of an orange , two-storey building with a footpath **to** (type région) a swimming pool in the shape of an eight and small palm trees **to the left and right** (type direction).*

(SpRL 9) *there is a wooden table and a wooden commode **to the left** (type direction).*

(SpRL 10) *a white bungalow with big windows , stairs* ***to the left and the right*** *(type direction) , a neat lawn and flowers* ***in front of*** *(type direction) the house and trees* ***at the back*** *(type direction).*

(SpRL 11) *Group of tourists is standing* ***in front of*** *(type direction) a buildilng.*

(SpRL 12) *one kid is* ***about to*** *(type région) slide down the slide , with one grown up waiting to catch it.*

(SpRL 13) *cars are parked* ***along the left side of*** *(type région) the street as well.*

(SpRL 14) *small palm trees* ***along*** *(type région) a footpath* ***next to*** *(type région) a gravel area .*

4.3.1.1 Prédiction du type de relations spatiales par comparaison de chaînes de caractères

Parmi les nombreuses mesures de similarité existantes, nous avons choisi deux méthodes *String Matching (SM)* et *Lin* qui sont classiquement utilisées dans la littérature car elles produisent des résultats pertinents (Duchateau et al., 2008) (cf. Chapitre 3, Section 3.2).

String Matching

Rappelons que la mesure String Matching est une mesure fondée sur la distance de Levenshtein (voir plus de détails en Chapitre 3, Section 3.2). À partir de l'exemple présenté dans la Figure 4.4, nous obtenons E(leading up, leaning on)=3. En effet, il y a trois opérations permettant de passer de la chaîne « leading up » à « leaning on ».

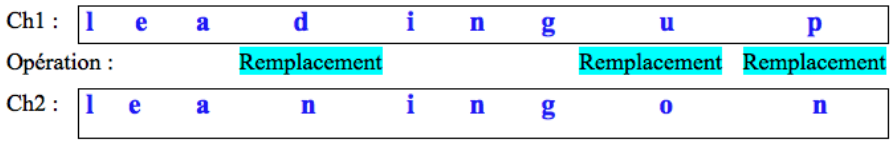

FIGURE 4.4 – Distance d'édition Levenshtein (E) pour les relations « leading up » et « leaning on »

Après avoir calculé la distance E, nous appliquons la formule (4.2) pour calculer la valeur de SM, normalisée entre 0 et 1.

$$\mathrm{SM}(Ch_1, Ch_2) = \max[0, \frac{(\min(|Ch_1|, |Ch_2|) - \mathrm{E}(Ch_1, Ch_2))}{\min(|Ch_1|, |Ch_2|)}] \qquad (4.2)$$

À partir de l'exemple des phrases (SpRL 1) et (SpRL 2), nous obtenons la similarité (cf. Exemple 4.3.1) :

Exemple 4.3.1 $SM(leading\ up, leaning\ on) = \max[0,\ (10-3)/10] = 0.70$

Sur la base de ces mesures, nous avons retourné, pour chaque relation candidate pour laquelle nous souhaitons prédire la classe, les similarités obtenues avec l'ensemble des relations de l'ensemble d'apprentissage (voir l'explication détaillée en Section 4.3.2). Nous déterminons ainsi les K relations les plus proches afin de prédire la classe à associer à la relation candidate (selon l'algorithme des K plus proches voisins, $KPPV$). Pour les cas particuliers tels que les relations composées de deux mots, dont le deuxième mot est une RS « next to », « sitting on », etc., nous faisons l'hypothèse que ces relations sont du même type que celui des relations « to », « on », etc. Si nous prenons les relations « next to » et « sitting on » dans les exemples (SpRL 5) et (SpRL 2) de type *région*, nous pouvons remarquer qu'ils ont le même type que les relations « to » et « on » dans les exemples (SpRL 8) et (SpRL 4).

Lin

La mesure Lin est une mesure de similarité fondée sur l'identification des n-grammes de caractères. La formule (4.3) présente la mesure Lin normalisée entre 0 et 1 :

$$\mathrm{Lin}(Ch_1, Ch_2) = \frac{1}{[1 + |\mathrm{T}(Ch_1)| + |\mathrm{T}(Ch_2)| - 2 \times |\mathrm{T}(Ch_1) \cap \mathrm{T}(Ch_2)|]} \tag{4.3}$$

Généralement, la valeur de n varie entre 2 et 5. En posant $n = 3$ (tri-grammes notée T), nous obtenons le résultat ci-dessous en reprenant l'exemple de la section 4.3.1 :
|**T** (leading up)| = {**lea**,ead,adi,din,**ing,ng** ,g u, up} = 8
|**T** (leaning on)| = {**lea**,ean,ani,nin,**ing,ng** ,g o, on} = 8
|**T** (leading up) ∩ **T** (leaning on)| = 3 (trois tri-grammes en commun)

À partir de l'exemple des phrases (SpRL 1) et (SpRL 2), nous obtenons les résultats suivants :

Exemple 4.3.2 $\mathrm{Lin}(leading\ up, leaning\ on) = \frac{1}{[(1+8+8)-(2\times 3)]} = 0.09$

Sur la base de ces mesures de similarité, nous avons appliqué l'algorithme $KPPV$ qui retourne la classe majoritaire pour la relation candidate. Notons que les informations lexicales ne sont pas toujours suffisantes. En effet, deux expressions peuvent être lexicalement éloignées mais sémantiquement très proches. Par exemple, les relations « along the left side of » (cf. Exemple (SpRL 13)) et « along » (cf. Exemple (SpRL 14)) sont de même type *région*, alors qu'en se basant sur la proximité lexicale, la relation « along the left side of » est assignée au type *direction*. Pour résoudre un tel problème nous proposons, dans la section suivante, de prendre en compte le contexte des relations pour prédire leur classe.

4.3.1.2 Prédiction du type de relations spatiales par proximité contextuelle

À cette étape, nous faisons l'hypothèse que les mots présents autour des relations (*toute la phrase* ou les *n mots autour de la relation*), que nous nommons *monde lexical*, vont nous permettre d'améliorer l'identification du type de RS. Le contexte est défini de deux façons : (i) *toute la phrase*, (ii) *n mots autour de la relation*.

Nous nous appuyons ensuite sur une approche dite sac de mots *SDM*, qui est l'une des premières méthodes de représentation des textes utilisée dans les systèmes de recherche d'information (Smail, 2009). Cette approche suppose que la sémantique d'un mot est reliée à l'ensemble des contextes dans lesquels il apparaît (Smail, 2009). Elle utilise des méthodes classiques afin de donner un poids aux termes. Nous comparons différents facteurs de pondération : nombre d'occurrences, TF-IDF (Term Frequency, Inverse Document Frequency) (Salton and Buckley, 1988) et la confiance (Agrawal et al., 1993) afin de sélectionner celui qui nous permet de construire le monde lexical le plus à même d'identifier le type de RS pertinent.

Nombre d'occurrences (nbre_occ)

Ici, nous calculons le poids d'un terme i en comptant les occurrences de ce mot dans le document j. Plus précisément, nous cherchons à connaître la fréquence d'apparition du terme i avec la même relation dans toutes les phrases. Par exemple, le terme « wall » apparaît trois fois avec la relation « leaning on ». Une fois le nombre d'occurrences de tous les couples terme/relation calculés, nous obtenons pour chaque relation un vecteur dont un exemple est présenté dans le Tableau 4.17.

(RS, nbre_occ)	(mot 1, nbre_occ)	(mot 2, nbre_occ)	...	(mot n, nbre_occ)
(leaning on, 3)	(wall, 3)	(locals, 1)	...	(mot n, nbre_occ)

TABLE 4.17 – Extrait d'un vecteur de nombre d'occurrences obtenu à partir des deux phrases (SpRL 2) et (SpRL 3)

Term Frequency, Inverse Document Frequency (TF-IDF)

TF-IDF est une mesure statistique permettant d'évaluer l'importance d'un terme contenu dans un document, en fonction de la fréquence à laquelle il apparaît dans plusieurs documents. Le TF-IDF est calculé comme suit :

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \tag{4.4}$$

Où
$w_{i,j} = tf_{i,j} \times idf_i$,

$tf_{i,j}$ = nombre d'occurrences de i dans j,
df_i = nombre de documents contenant i,
N = nombre total de documents.

Nous obtenons X vecteurs représentant chaque relation, dont la longueur est égale au nombre de termes du corpus et le poids de chaque terme égal à $w_{i,j}$. Le Tableau 4.18 est un extrait du vecteur TF-IDF de la relation « leaning on ».

(SR, nbre_occ)	(mot 1, $w_{i,j}$)	(mot 2, $w_{i,j}$)	...	(mot n, $w_{i,j}$)
(leaning on, 3)	(wall, 5.05)	(locals, 5.74)	...	(mot n, $w_{i,j}$)

TABLE 4.18 – Extrait du vecteur de TF-IDF pour les phrases (SpRL 2) et (SpRL 3)

La confiance

Nous calculons la confiance (*conf*) telle que définie par (Agrawal et al., 1993), définissant la probabilité qu'une transaction contienne le mot Y étant donné qu'elle contient la relation X. La confiance est définie par :

$$conf = \frac{Nombre\ de\ phrases\ contenant\ la\ relation\ X\ et\ le\ mot\ Y}{Nombre\ de\ phrases\ contenant\ la\ relation\ X} \tag{4.5}$$

Par exemple, si trois phrases contiennent la relation « leaning on » et une phrase contient la relation « leaning on » et le terme « locals », nous calculons la confiance associée au mot « locals » du monde lexical comme suit : $conf = 1/3 = 0.33$. Comme pour les étapes précédentes, nous mettons en avant dans le Tableau 4.19, un extrait du vecteur de la relation « leaning on » et les termes des phrases (SpRL 2) et (SpRL 3), le poids de chaque terme ici étant égal à *conf*.

(SR, nbre_occ)	(mot 1, conf)	(mot 2, conf)	...	(mot n, conf)
(leaning on, 3)	(wall, 1.0)	(locals, 0.33)	...	(mot n, conf)

TABLE 4.19 – Extrait du vecteur de confiance pour les deux phrases (SpRL 2) et (SpRL 3)

Sur la base des trois pondérations proposées (*nbre_ occ, TF-IDF, conf*), nous mesurons la proximité fondée sur le *cosinus* (Baeza-Yates et al., 1999) entre les mondes lexicaux propres aux relations candidates et ceux propres aux relations de l'ensemble d'apprentissage. La mesure *cosinus* est largement utilisée dans le domaine

de la recherche d'information. Elle a été étendue pour prendre en compte les poids et est devenue la mesure standard des vecteurs pondérés (Bannour et al., 2011). La mesure *cosinus* est définie par la formule (4.6).

$$Cos(\overrightarrow{d_j}, \overrightarrow{d_k}) = \frac{\overrightarrow{d_j} \ . \ \overrightarrow{d_k}}{\mid \overrightarrow{d_j} \mid . \mid \overrightarrow{d_k} \mid} = \frac{\sum_{i=1}^{n} w_{i,j} \ . \ w_{i,k}}{\sum_{i=1}^{m} w_{i,j}^2 \ \sum_{i=1}^{n} w_{i,k}^2} \tag{4.6}$$

Une fois l'ensemble des mesures de proximité calculées, nous appliquons l'algorithme $KPPV$ pour déterminer les K' relations les plus proches et nous affectons chaque relation candidate à la classe identifiée comme la plus proche.

En outre, dans cette étape, les termes peuvent appartenir au monde lexical de plusieurs types de relation. Dans ce cas, nous choisissons le type de relation pour lequel le terme est le plus fréquent (par exemple, (i) si le mot « chairs » se trouve une seule fois avec la relation « around » qui est de type *distance*, (ii) et trois fois avec la relation « in front of » qui est de type *direction*, nous supprimons ce mot de contexte de (i)).
Pour réduire l'ambiguïté et améliorer l'identification automatique du type de RS, nous proposons de combiner les deux approches précédentes *par comparaison de chaîne de caractères* et *par proximité contextuelle*.

4.3.1.3 Combinaison des approches par comparaison de chaînes de caractères et par proximité contextuelle

Sur la base d'une analyse qualitative, qui vise à donner sens et à comprendre les résultats obtenus par notre système, nous distinguons deux principales conclusions pour la prédiction automatique du type de RS.
Premièrement, nous remarquons que l'approche *par comparaison de chaînes de caractères* donne généralement de meilleurs résultats. Deuxièmement, nous notons que l'approche *par proximité contextuelle* donne des résultats sensiblement meilleurs lorsque les RS se composent de plus de 4 mots. Par exemple, si nous prenons la relation « along the left side of », présente dans l'exemple (SpRL 13), elle est associée au type direction dans la première étape (cf. Section 4.3.1.1), alors que dans la deuxième étape 4.3.1.2 est considérée comme type région qui est la classe pertinente.

Au regard de cette première analyse, nous faisons l'hypothèse que si les relations se composent de plus de m [7] mots, nous privilégions l'approche *par proximité contextuelle Cos*, sinon nous choisissons l'approche *par comparaison de chaînes de caractères SM*. La section 4.3.2 détaille les résultats de nos expérimentations menées sur un corpus en langue anglaise.

4.3.2 Expérimentations

Étant donné qu'il n'existe pas encore, à notre connaissance, de corpus français annotés avec des RS et les types associés, nous avons choisi un corpus en langue

7. Nos expérimentations ont montré que $m = 4$ donne les résultats les plus pertinents

anglaise SpRL (Spatial Role Labeling) (Kordjamshidi et al., 2011) qui représente un benchmark reconnu dans le domaine pour mener à bien ces expérimentations. Le corpus est composé de 1 213 phrases annotées contenant les trois types des RS région, direction et distance. Nous avons procédé à une série d'expérimentations dans lesquelles nous avons fait varier les paramètres susceptibles d'influencer les résultats des mesures de performance : K et K' pour l'algorithme de $KPPV$ et n paramètre de fenêtrage pour sélectionner les n *mots autour de la relation*.

4.3.2.1 Prédiction du type de relations spatiales par comparaison de chaînes de caractères

Afin d'estimer l'efficacité des différentes méthodes, nous appliquons un processus de validation croisée. Dans notre cas, le corpus est divisé en 3 partitions et chaque partition contient 31 relations (18 régions, 10 directions, 3 distances). Le jeu d'apprentissage est constitué successivement de 2 des 3 partitions et le jeu de test permettant d'obtenir les résultats présentés est constitué de la partition restante. Le Tableau 4.20 montre les résultats de *String Matching 1* obtenus à partir de la première série d'expérimentations, en appliquant l'algorithme SM uniquement. Le Tableau 4.21 présente les résultats de *String Matching 2* obtenus en appliquant les propositions présentées en section 4.3.1.1 de relations composées de deux mots. Une différence notable est observée entre les deux méthodes *String Matching 1*, *String Matching 2* en obtenant une amélioration de 5% en terme d'exactitude avec $K = 1$. Cette amélioration indique que nous identifions le même type pour les : (i) relations composées de deux mots, dont le deuxième mot est une relation spatiale (par exemple, « next to », « sitting on », etc.), et (ii) relations non composées qui correspondent aux deuxièmes mots des relations composées (par exemple, « to », « in », etc.).

Le Tableau 4.22 présente les résultats obtenus avec la mesure Lin. Nous pouvons remarquer que la mesure de similarité *String Matching 2* donne des résultats satisfaisants comparativement à la mesure *Lin* quelque soit la valeur de K avec un score maximal de 0.82 d'exactitude (accuracy) lorsque $K = 1$.

K	Précision	Rappel	F-mesure	Exactitude
1	0.70	0.66	0.68	0.77
3	0.50	0.55	0.52	0.74
5	0.48	0.52	0.50	0.73

TABLE 4.20 – Résultats de la mesure *String Matching 1*

Le Tableau 4.23 représente des extraits des résultats obtenus avec la mesure *String Matching 2*. Les relations en gras représentent des exemples de relations bien classées par cette approche. Malgré les résultats significatifs obtenus, nous pouvons remarquer que le problème reste néanmoins loin d'être résolu. Certaines limites sont

K	Précision	Rappel	F-mesure	Exactitude
1	**0.80**	**0.69**	**0.74**	**0.82**
3	0.53	0.57	0.55	0.79
5	0.51	0.54	0.53	0.76

TABLE 4.21 – Résultats de la mesure *String Matching 2*

K	Précision	Rappel	F-mesure	Exactitude
1	**0.70**	**0.58**	**0.63**	**0.75**
3	0.59	0.53	0.56	0.73
5	0.55	0.44	0.49	0.69

TABLE 4.22 – Résultats de la mesure *Lin*

apparues, notamment pour la prédiction du type pour les relations longues (par exemple, les relations « Along the left side of » et « To the left and the right »). Ainsi, les résultats peuvent être améliorer si nous prenons en considération les mots qui se trouvent autour de la relation.

Relation spatiale	type de la relation	type prédit
In front of	direction	**direction**
Sitting in	région	**région**
Along the left side of	région	direction
To the left and the right	direction	région
Near	distance	région

TABLE 4.23 – Extraits des classes obtenues avec la mesure *String Matching 2*

4.3.2.2 Prédiction du type de relations spatiales par proximité contextuelle

Dans cette série d'expérimentations, la prédiction des RS est effectuée sur la base de l'approche *SDM* classique avec suppression des *mots vides*[8]. Nous appliquons les deux contextes (*toute la phrase*, *n mots autour de la relation*) et nous évaluons l'approche pour chaque monde lexical avec K' variant de 1 à 5. Le Tableau 4.24 présente les résultats obtenus en terme d'exactitude pour chaque facteur de pondération (nombre d'occurrences, TF-IDF et confiance) en fonction de chaque contexte. Notons que pour le contexte *n mots autour de la relation*, le n varie de 1 à 3 car le

8. Les mots vides (*Stop word*) sont des mots fonctionnels qui ne sont pas porteurs de sens (déterminants, pronoms, conjonctions, adverbes) (Vergne, 2004). La liste utilisée est disponible ici : http ://xpo6.com/list-of-english-stop-words/

corpus est un ensemble de phrases. Dans le Tableau 4.24, nous pouvons constater que le contexte *n mots autour de la relation* donne des résultats supérieurs à ceux du contexte *toute la phrase*. Le monde lexical fondé sur le TF-IDF avec $K' \geq 3$ donne des résultats satisfaisants. Comme conclusion de cette série d'évaluations, le meilleur score (exactitude de 0.67) est obtenu avec $K' = 5$ et $n = 2$.

K'		*toute la phrase*	*n termes autour de RS*		
			$n = 1$	$n = 2$	$n = 3$
1	nombre d'occurrences	0.61	0.62	0.62	0.60
	TF-IDF	0.56	0.60	0.51	0.53
	Confiance	0.56	0.62	0.62	0.60
3	nombre d'occurrences	0.62	0.60	0.63	0.58
	TF-IDF	0.45	0.63	**0.66**	**0.63**
	Confiance	0.40	0.60	0.63	0.56
5	nombre d'occurrences	0.58	0.65	0.67	0.57
	TF-IDF	0.40	0.64	**0.67**	**0.66**
	Confiance	0.41	0.64	0.67	0.56

TABLE 4.24 – Résultats de la méthode par proximité contextuelle (notée *Cos*) en terme d'exactitude

Le Tableau 4.25 présente des extraits des résultats obtenus en utilisant la mesure Cosinus avec $n = 2$ et $K' = 5$. Les relations en gras représentent des exemples de relations bien classées par cette approche. En prenant en compte le contexte de la relation, nous remarquons que les résultats obtenus sont nettement moins bons par rapport à l'étape précédente (cf. Section 4.3.2.1). Par contre, nous pouvons observer à partir du Tableau 4.25, que le système prédit correctement le type pour les relations longues. Dans ce contexte, il serait intéressant de combiner la présente méthode avec l'approche précédente afin de tirer avantage des deux méthodes.

Relation spatiale	type de la relation	type prédit
In front of	direction	région
Sitting in	région	direction
Along the left side of	région	**région**
To the left and the right	direction	**direction**
Near	distance	région

TABLE 4.25 – Extraits des résultats obtenus avec la mesure Cosinus.

4.3.2.3 Combinaison des approches par comparaison de chaînes de caractères et par proximité contextuelle

Dans cette section, nous présentons les résultats de la combinaison des deux méthodes (par comparaison de chaînes de caractères et prise en compte du monde lexical). Nous avons réalisé une série d'expérimentations pour identifier la combinaison la plus adaptée des paramètres, i.e. $K = 1$ pour SM et $K' = 5$, $n = 2$ pour Cos en utilisant le monde lexical fondé sur TF-IDF. Ceci nous permet d'obtenir un score d'exactitude de **0.84**. Ainsi, la combinaison se comporte mieux que chaque méthode individuellement.

Le Tableau 4.26 présente des extraits des résultats obtenus en combinant les deux approches, comparaison de chaînes de caractères et prise en compte du monde lexical. Les relations en gras représentent des exemples de relations bien classées en utilisant la combinaison. Nous notons que la combinaison améliore la prédiction du type de relations en tirant parti des deux méthodes. La combinaison permet de compléter l'information en identifiant les quatre relations « In front of », « Sitting in », « Along the left side of » et « To the left and the right » au lieu d'identifier seulement les relations « In front of », « Sitting in » dans la première étape ou les relations « Along the left side of » et « To the left and the right » dans la deuxième étape. Néanmoins, notre approche n'a pas prédit le bon type pour la relation « Near ». Cela peut être expliqué par le manque de données d'apprentissage liées aux relations de type distance.

Relation spatiale	type de la relation	type prédit
In front of	direction	**direction**
Sitting in	région	**région**
Along the left side of	région	**région**
To the left and the right	direction	**direction**
Near	distance	région

Table 4.26 – Extraits des résultats obtenus en utilisant la combinaison des deux approches précédentes.

Dans cette partie du chapitre, nous avons présenté une méthode permettant de combiner différentes approches statistiques et sémantiques afin de prédire automatiquement le type de RS à partir d'un corpus de textes courts. Cette proposition est expérimentée sur un corpus standard anglais étant donné qu'il n'existe pas encore, à notre connaissance, de corpus français annoté avec des RS et les types associés. Nous envisageons, comme perspective, de prédire le type des relations spatiales présentent dans les corpus de messages courts en langue française. Ce travail nécessitera une phase d'annotation préalable pour constituer le corpus d'apprentissage pour le français. En effet, la prédiction des types associés aux nouvelles formes/variantes de

RS donne des indications plus précises quant à la représentation spatiale des informations contenues dans les données textuelles dites non standards.
Comme nous avons pu le voir dans les exemples de messages courts présentés en Chapitre 1, Section 1.1.3, les relations spatiales peuvent être exprimées différemment comparativement aux corpus standards (par exemple, « ds » et « sur » de type inclusion et « dvant » et « avant » de type adjacence).

4.4 Conclusion

Nous avons présenté dans ce chapitre de nouvelles approches pour l'extraction et la classification des relations spatiales. Dans une première étape, nous proposons une analyse comparative de deux approches et de leur combinaison pour l'identification automatique du type de relations spatiales.

Pour chaque relation spatiale identifiée, nous proposons de définir son monde lexical afin d'améliorer la prédiction du type de relation. Puis nous avons proposé une méthode combinant plusieurs mesures de similarité et pondération ainsi que des méthodes d'apprentissage supervisé (SM, Lin, TF-IDF, confiance, nombre d'occurrences et $KPPV$). Nos résultats montrent que la combinaison améliore la qualité de la prédiction. Cela nous permet d'explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des différentes approches (lexicale et contextuelle).

Dans une deuxièmement étape, nous proposons une méthode pour l'identification automatique de nouvelles relations spatiales à partir de deux types de corpus de messages courts. En utilisant le dictionnaire enrichi généré par notre méthode décrite dans le chapitre précédent (cf. Chapitre 3), nous identifions des nouvelles relations spatiales en combinant plusieurs approches proposées par la communauté du TALN (étiquetage grammatical, nombre d'occurrences et n-grammes de mots). Nous avons constaté que la combinaison de ces approches nous a permis d'obtenir des résultats très satisfaisants pour l'identification de nouvelles formes/variantes de relations spatiales avec un score de F-mesure de 0.97 et 0.91 pour les deux corpus SMS et tweets respectivement.

En outre, nous avons observé quelques différences dans les résultats de reconnaissance de relations spatiales entre les corpus de SMS et tweets. Pour l'identification de nouvelles relations spatiales qui précèdent les entités spatiales, les résultats du corpus SMS étaient constamment meilleurs que les résultats du corpus de tweets.

En dépit des résultats encourageants, certaines relations ne sont pas encore identifiées (cf. par exemple, « ds » et « à 731 km de distance en » voir (SMS 30) et (SMS 31)). Les erreurs d'identification sont liées à la fréquence d'apparition de ces relations dans les corpus. Notons que la fréquence d'apparition de certaines relations est limitée dans le corpus. À partir de l'exemple (SMS 30), la relation « ds » dans l'exemple (SMS 30) apparaît une seule fois dans l'échantillon de 1000 SMS. Notre approche (cf. Section 4.2.1.1) ne permet pas d'extraire cette relation du fait du seuil

$S2$ [9]. La suppression de ce seuil ($S2$) ajoute un bruit important aux résultats obtenus par notre approche.

Dans ce cas, en tenant compte uniquement des informations présentes dans le corpus, cela ne permet pas de valider si la relation est pertinente ou non.

Afin de résoudre ce problème, dans le chapitre suivant (cf. Chapitre 5), nous détaillons la troisième de nos propositions ayant pour objectif d'identifier les relations sémantiques présentes entre les entités nommées et une entité spatiale. Par exemple, les relations sémantiques « est originaire de » et « bosse en » dans les expression « <PRE_6/> **est originaire de Beaumont** » et « le <PRE_4/> soient rentabilisés en le faisant payer des impots s'il **bosse en France** » (cf. Exemples (SMS 32) et (SMS 33)). L'objectif ici est d'identifier de nouveaux types de relations sémantiques faisant intervenir une entité spatiale, et ainsi d'enrichir la typologie des relations spatiales définies dans la communauté scientifique. L'identification des différents types de relations peut toutefois servir pour compléter et valider les informations contenues autour les entités spatiales.

9. Pour rappel, le seuil $S2$ permet de sélectionner les mots les plus fréquents des « m » étiquettes.

Chapitre

5 Extraction de relations sémantiques entre entités nommées

Contents

5.1 Introduction

Dans ce chapitre, nous détaillons la troisième de nos propositions ayant pour objectif d'identifier les relations sémantiques RS_m entre les entités nommées (EN) avec la présence d'au moins une entité spatiale (ES). Nous entendons par les EN, les ES, $Organisation$ et $Personne$.
Nous cherchons ici à extraire de nouvelles RS_m impliquant des entités spatiales à partir de corpus de messages peu standardisés, en particulier les SMS afin d'améliorer encore la compréhension de ce type de messages lorsque des informations spatiales sont mentionnées explicitement. Pour atteindre cet objectif, nous proposons, comme première étape, d'extraire l'ensemble des relations candidates associant les entités nommées et les entités spatiales. Puis, nous proposons, dans une deuxième étape, de les valider.

Dans la suite de ce chapitre, nous nous appuierons sur une représentation pré-traitée des SMS indiquée ci-dessous pour lesquels nous cherchons à extraire les RS_m qui se trouvent entre une EN et une ES afin d'identifier les différents types de relations qui peuvent servir à compléter et valider les informations contenues autour les ES. Cette représentation est obtenue en appliquant le processus d'annotation présenté dans la Section 5.3.1. À partir des exemples ci-dessous, les balises <PRE_6/> et <PRO> représentent l'EN personne, la balise <ORG> représente l'EN organisation (voir plus de détail sur le processus d'annotation en Section 5.3.1).

(1) ... <PRE_6/> **est originaire de** <ES>Beaumont</ES> ...
(2) ... <PRO>il</PRO> **bosse en** <ES>France</ES>...
(3) <PRO>Je</PRO> **suis a 5 min de** <ES>la gare</ES>
(4) <PRO>On</PRO> **est ver** <ES>beziers</ES>
(5) ... <PRO>je</PRO> **suis ds** <ES>le parc</ES>
(6) ... <ORG>Cern</ORG> **à 731 km de distance en** <ES>Suisse</ES> ...
(7) ... <ES>masgot</ES> **a 20min de** <ES>gueret</ES>
(8) ... <ORG>creps</ORG> **à coté du** <ES>stade philippides</ES>
(9) ... <PRO>je</PRO> **bosse en** <ORG>presta</ORG> **à** <ES>Chatou</ES>
(10) ... <PRE_5/> **en** <ES>suisse</ES> ...
(11) ... <ORG>La clinique</ORG> **est dans la partie entre** <ES>gambetta</ES> **et** <ES>l'avenue paul bringuier</ES>
(12) ... <PRO>Je</PRO> **veux voyagé ds tte** <ES>l'australie</ES> ...
(13) ... <PRO>il</PRO> **est remonte hier sur** <ES>paris</ES> ...

La fréquence d'apparition des relations dans le corpus peut se révéler trop limitée pour permettre d'utiliser des critères de sélection statistiques. Par exemple, nous pouvons relever une relation pertinente qui apparaît une seule fois dans le corpus (e.g. la relation « a 20min de » dans l'exemple (7) apparaît une seule fois dans le corpus de SMS). Dans ce cas, la prise en compte unique des informations présentes dans le corpus ne permet pas de valider la pertinence ou non de la relation de façon automatique. À cet effet, nous avons opté pour l'utilisation des données externes pour permettre de résoudre ce problème.

Le Web étant une grande ressource de données, nous proposons de l'exploiter afin de valider les RS_m extraites à partir du corpus de SMS. L'utilisation du Web est bénéfique, elle permet d'accéder à toutes sortes de données quel que soit le domaine (par exemple, des articles de presse, des pages de sites web, des tutoriels, des réseaux sociaux, etc.). Nous faisons ainsi l'hypothèse, que la prise en compte des informations externes provenant du Web, et plus précisément des moteurs de recherche, nous permet de valider comme information pertinente une relation sémantique identifiée entre deux EN dans les corpus de messages courts que nous traitons dans nos travaux.

5.2 Le Web comme ressource complémentaire

De nos jours, le Web est devenu une source de données incontournable. En effet, le manque de données d'entraînement disponibles a amené les chercheurs à recourir à des sources externes et ouvertes issues du Web pour évaluer différents systèmes (Bollegala et al., 2009 ; Chen et al., 2006). Cependant, le coût du traitement de tous les documents du Web peut se révéler très élevé. Les travaux de Berger and Mittal (2000) ; Turpin et al. (2007) ; Varadarajan and Hristidis (2006) ; Wu et al. (2004) ont alors exploré des algorithmes et des structures de données associés à un moteur de recherche pour améliorer les performances en termes de temps d'exécution et de pertinence des résultats. De manière plus précise, ces travaux ont généré des extraits (snippets) représentatifs des documents retournés afin de permettre à l'utilisateur de localiser rapidement les informations souhaitées. La snippet constitue une *mini fiche* de présentation de la page comprenant un titre, une description et parfois des informations supplémentaires (images, votes, prix, etc.). Les snippets sont générées à partir d'une requête. Elles présentent des extraits des documents dans lesquels des mots de cette requête sont présents.

En effet, si ces extraits contiennent l'information désirée, le coût de recherche peut être largement réduit. De ce fait, la qualité de ces extraits a un impact significatif sur les coûts de recherche.

D'autres chercheurs se sont focalisés à exploiter les données statistiques retournées par un moteur de recherche tel que le nombre de pages Web. Ces résultats de recherche (snippets, nombre de recherches) sont retournés sous forme de pages web

qui sont générées automatiquement par un moteur de recherche, tel que Google, en fonction des requêtes saisies (cf. Figure 5.1). Google (Brin and Page, 2012) est le moteur de recherche le plus utilisé dans le monde avec 5.5 milliards de requêtes effectuées chaque jour [1] et 130 Mille milliards de pages indexées en Novembre 2016 [2].

Les résultats sont présentés sous forme d'une liste triée en fonction d'un critère de pertinence. Le classement de ces pages web est donné à l'aide de l'algorithme PageRank (Brin and Page, 1998) qui analyse, entre autre, les liens concourants afin d'attribuer à chaque page un score (poids). Les pages avec le meilleur score PageRank auront plus de poids pour Google et seront donc classées en début de la liste des résultats.
Une page de résultat est, en général, affichée de la manière suivante (cf. Figure 5.1) :

1. Le nombre de résultats d'une requête donnée sur un moteur de recherche représente un ensemble de liens pointant vers les ressources considérées comme pertinentes pour une requête donnée (cf. Figure 5.1 (a)) ;
2. Les snippets représentent les informations associées aux résultats de la recherche d'une requête donnée. Ces informations représentent des extraits des documents pertinents, selon une requête donnée, dans lesquels des mots de la requête sont présents. Par exemple, avec le moteur de recherche Google qui est largement utilisé aujourd'hui, les snippets sont composées des éléments suivants :
   — Un titre de couleur bleue qui correspond au titre de la page web,
   — Une URL de couleur verte qui correspond à l'adresse Web du site en question,
   — Un aperçu du contenu de la page web appelé également *méta description* qui s'affiche sous l'URL du site et inclut les mots de la requête en gras afin d'aider les utilisateurs à évaluer la pertinence de la page qu'ils recherchent.

Nous présentons dans les deux sous-sections 5.2.1 et 5.2.2, les travaux connexes de la communauté scientifique Fouille de Textes qui utilisent le Web comme ressource.

5.2.1 Utilisation des informations statistiques issues des moteurs de recherche

Le nombre de pages fournies par le moteur de recherche est devenu un facteur essentiel pour les approches de fouille de textes. Un tel critère dit *quantitatif* s'appuie en général sur le nombre de pages web retournées par les moteurs de recherche (cf. Figure 5.1 (a)). Nous pouvons, par exemple, citer les travaux mesurant les associations de termes. Nous présentons dans la suite de cette section une synthèse des travaux exploitant ces critères quantitatifs issus des moteurs de recherche.

1. http ://www.webrankinfo.com/dossiers/google/chiffres-cles
2. http ://www.arobasenet.com/2016/11/chiffres-index-google-3490.html

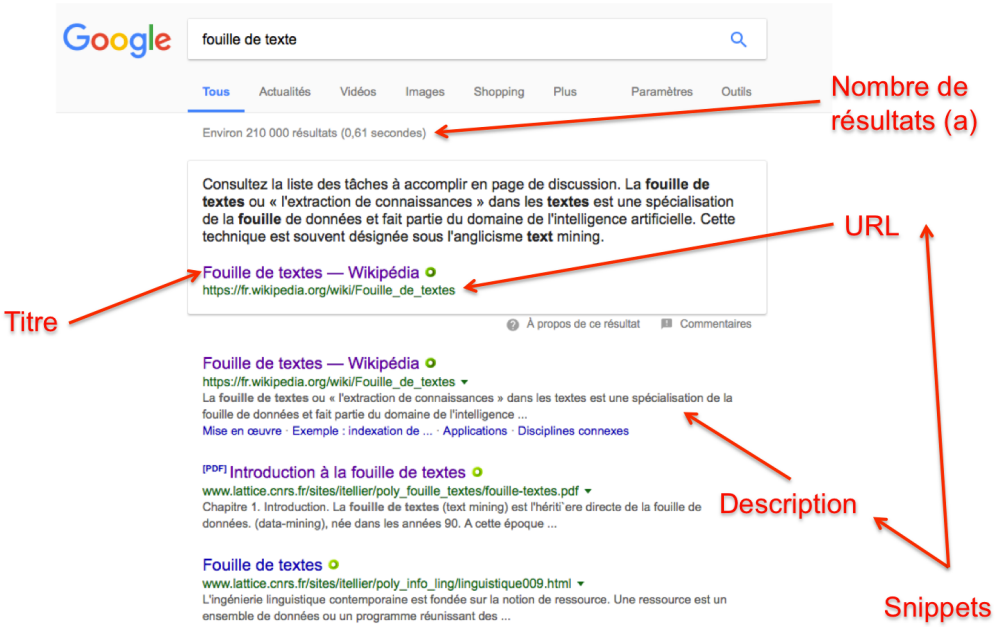

FIGURE 5.1 – Page Google issue de la requête *fouille de texte*.

Turney (2001) a présenté une méthode pour déterminer des synonymes appropriés en se basant sur le requêtage des données acquises via le moteur de recherche AltaVista[3]. Turney (2001) a proposé un algorithme d'apprentissage non supervisé, PMI-IR (Pointwise Mutual Information and Information Retrieval), basé sur la co-occurrence de mots sur le Web pour mesurer la similarité des paires de mots (*mot*, $choix_i$). L'objectif de PMI-IR est de choisir un synonyme ($choix_i$), pour un *mot* donné, parmi une liste ($choix_i$) existante (issue du TOEFL[4]) en attribuant un score à chaque $choix_i$ comme défini équation 5.1. Le $choix_i$ qui donne le meilleur score est assigné comme *synonyme* au *mot*. PMI-IR a été évalué à l'aide des exercices du TOEFL, obtenant un score de 73.75% de réponses correctes.

$$score(choix_i) = \frac{nb(mot\ AND\ choix_i)}{nb(choix_i)} \tag{5.1}$$

avec $nb(mot\ AND\ choix_i)$ le nombre de documents contenant le *mot* et $choix_i$, $nb(choix_i)$ le nombre de documents contenant $choix_i$.

Roche and Prince (2007) ont proposé une mesure appelée *AcroDef* qui exploite des ressources du Web dans l'objectif d'identifier la définition pertinente d'un acronyme présent dans un document. Pour déterminer l'expansion d'un acronyme, les auteurs ont proposé d'utiliser le contexte pour identifier l'expansion la plus

3. http ://www.altavista.com/
4. Test of English as a Foreign Language

pertinente pour un acronyme donné. Ils ont défini le contexte C par un ensemble de mots significatifs présents dans la page où l'acronyme est présent. Les auteurs ont classé cette liste de co-occurrences à l'aide de mesures statistiques. Pour ce faire, la mesure $AcroDef$ utilise différentes mesures, nous donnons ci-dessous la mesure $AcroDef_{IM3}$ fondée sur l'Information Mutuelle au Cube $IM3$ (Daille, 1994).

$$AcroDef_{IM3}(a^j) = \frac{nb(\cap_{i=1}^{n} a_i^j + C)^3}{\sqcap_{i=1}^{n} nb(a_i^j + C \mid a_i^j \notin M_{stop-words})} \quad où\ n \geq 2 \qquad (5.2)$$

avec $a_i^j + C$ les pages contenant les mots a_i^j formant l'expansion avec tous les mots du contexte C et nb le nombre de pages retournées par le moteur de recherche Exalead [5].

Bollegala et al. (2007) ont proposé une méthode automatique pour mesurer la similarité sémantique entre deux mots en utilisant le moteur de recherche Google. Tout d'abord, les auteurs ont récupéré le nombre de pages associées à des requêtes avec des mots X et Y considérés individuellement puis conjointement ($X\ AND\ Y$). Puis, quatre mesures de co-occurrence ont été modifiées : Jaccard (Jaccard, 1901), Overlap (Simpson, 1943), Dice (Smadja et al., 1996) et Pointwise Mutual Information (PMI) (Church and Hanks, 1990). Cette méthode permet de calculer une *similarité sémantique* en s'appuyant sur le nombre de pages retournées. Ensuite, des patrons lexico-syntaxiques sont extraits à partir des snippets en utilisant les n-grammes de mots. Puis, la fréquence de chaque patron extrait est calculée. Finalement, les auteurs ont utilisé les machines à vecteurs de support (support vector machine, SVM) pour déterminer la combinaison optimale des scores de similarité basés sur le nombre de pages retournées et le meilleur classement des patrons.

Cilibrasi and Vitanyi (2007) ont proposé une métrique de distance entre deux termes, nommée NGD (Normalized Google Distance), basée sur leur corrélation dans les résultats de recherche retournés par Google. NGD est fondée sur le nombre de pages Web, dans lesquels les termes se trouvent individuellement et conjointement, pour calculer la distance sémantique entre les termes. L'hypothèse est la suivante : si deux termes X et Y se trouvent souvent ensemble sur la même page Web, ils sont sémantiquement liés. NGD retourne un score de similarité entre 0 et 1, où 0 représente une correspondance parfaite et 1 l'absence de similarité entre les chaînes comparées. La mesure NGD est définie par l'équation 5.3.

$$NGD(X,Y) = \frac{\max\{\log f(X), \log f(Y)\} - \log f(X,Y)}{\log N - \min\{\log f(X), \log f(Y)\}} \qquad (5.3)$$

avec $f(X),\ f(Y)$ le nombre de pages contenant $X,\ Y$, $f(X,Y)$ le nombre de pages contenant à la fois X et Y, et N le nombre total de pages Web indexées par Google.

5. http ://www.exalead.fr/

5.2.2 Utilisation du contenu textuel du Web

L'utilisation des snippets est également devenue une alternative pour les approches de recherche et d'extraction d'information. En effet, les snippets permettent d'améliorer le taux de clic sur les résultats retournés par un moteur de recherche et donc minimiser le temps nécessaire pour une recherche donnée, et cela car les snippets se composent de courts résumés qui représentent des fragments de texte contenant des termes d'une requête. Les traitements qui peuvent être effectués sur ces snippets sont plus simples étant donnée la taille réduite du contenu à traiter. Nous présentons par la suite quelques travaux qui exploitent les snippets pour des tâches de fouille de textes.

Chen et al. (2006) ont proposé un modèle de double checking (Web Search with Double Checking WSDC) utilisant des snippets retournées par un moteur de recherche (Google) pour calculer la similarité sémantique entre les termes. Pour calculer la similarité entre chaque paire de termes X et Y, les auteurs ont extrait des snippets correspondant à chaque terme. Ensuite, les occurrences de X dans les N meilleures snippets de la requête Y et les occurrences de Y dans les N meilleures snippets de la requête X ont été calculées. Ces valeurs sont combinées selon différentes mesures, à savoir Dice, Cosinus, Jaccard et Overlap Ratio. Les auteurs ont également proposé une nouvelle mesure appelée Co-occurrence Double-Checking (CODC) pour calculer la similarité entre X et Y (cf. Formule 5.4).

$$CODC(X,Y) = \begin{cases} 0 & if\ f(Y@X)=0\ or\ f(X@Y)=0 \\ e^{\log\left[\frac{f(Y@X)}{f(X)} \times \frac{f(X@Y)}{f(Y)}\right]^{\alpha}} & otherwise. \end{cases} \tag{5.4}$$

Avec $f(Y@X)$ le nombre d'occurrences de Y dans le top N snippets de la requête X avec le moteur de recherche Google, $f(X)$ le nombre d'occurrences de X dans le top N snippets de la requête X, et α une constante déterminée expérimentalement à 0.15.

Geleijnse and Korst (2007) ont proposé une méthode pour enrichir une ontologie avec les personnages historiques et leurs principales informations biographiques en utilisant le résultat d'un moteur de recherche (Google). Ils ont également proposé d'identifier le réseau social entre les personnages trouvés en partant de l'hypothèse que deux personnes sont liées lorsqu'ils sont souvent mentionnés dans le même contexte. Pour atteindre cet objectif, les auteurs ont extrait les informations provenant des snippets retournées par le moteur de recherche Google pour des requêtes données. Pour obtenir des snippets pertinentes, les auteurs ont utilisé l'instance (année de naissance - année de décès), des personnages historiques, pour formuler les requêtes (e.g. 1685-1750) à partir une ontologie du domaine. Les auteurs s'appuient en effet sur le fait que les noms des personnes historiques sont souvent suivis de textes avec une fenêtre temporelle (e.g Vincent van Gogh (1853 - 1890)). Puis, ils ont utilisé une approche à base de règles (rule-based approach) pour identifier les noms des personnes dans les snippets. Les noms des personnes

historiques sont reconnus en utilisant un ensemble de règles simples. Par exemple, un nom est reconnu lorsqu'il est placé avant la fenêtre temporelle et contient deux ou trois mots en majuscule (e.g. en utilisant la requête « 1685-1750 », les auteurs ont obtenu « Johann Sebastian Bach » comme un nom de personnage historique). Finalement, les auteurs ont requêté les noms extraits avec le motif « was » (par exemple, « Johann Sebastian Bach was ») pour identifier des informations biographiques supplémentaires (nationalité, genre, profession, etc).

Xu et al. (2016b) ont proposé une méthode qui permet de générer automatiquement un graphe de relations spatio-temporelles (STRG : spatial temporal relation graph) pour chaque paire de concepts donnée (par exemple, recherche la paire de concepts « apple » et « computer » dans Google.). Afin de générer le STRG, les auteurs ont exploité les données issues des pages web retournées par un moteur de recherche (Google). Les nœuds de ce graphe sont les mots des relations (RW) qui co-occurrent sur le Web avec les deux concepts. Afin d'identifier ces mots de relation, les auteurs ont d'abord requêté la paire de concepts dans Google. Puis ils ont sélectionné les noms les plus fréquents dans les snippets retournées (par exemple, les cinq premiers mots de relation entre les deux concepts « apple » et « computer » sont « Macbook, Macintosh, iphone, ipad, watch »). Quant aux arêtes, les auteurs leur ont attribué un poids en calculant le PMI entre les mots de la relation (les nœuds). Finalement, les auteurs ont supprimé les arêtes dont le poids est faible afin d'optimiser le graphe de relations en maintenant la condition qui était de préserver sa connectivité.

Inspiré par les études de Bollegala et al. (2007) ; Geleijnse and Korst (2007), nos travaux exploitent : (1) le nombre de pages et (2) les snippets obtenues à partir de requêtes données. Ceci représente deux sources d'informations utiles et complémentaires fournies par la plupart des moteurs de recherche. De manière concrète, dans ce chapitre, nous présentons une approche d'identification et de validation de relations sémantiques (spatiales et non spatiales) entre les entités nommées (Personne, Organisation et Localisation) extraites à partir d'un corpus de SMS.
Notre méthode repose sur l'exploitation d'informations statistiques fournies par les moteurs de recherche comme première étape puis sur l'utilisation des informations extraites des snippets comme deuxième étape. Finalement, nous proposons de combiner les deux étapes afin d'améliorer les résultats.

5.3 Notre approche

Dans cette section, nous présentons une méthode originale permettant d'exploiter les données issues du Web comme support de connaissances pour extraire de nouvelles relations sémantiques (RS_m), spatiales et non spatiales, entre entités nommées (EN) à partir d'un corpus de messages courts. En nous basant sur les travaux

présentés en section 5.2, nous proposons une méthode qui repose sur trois phases. En effet, nous proposons une approche décrite en Figure 5.2 comprenant les quatre étapes suivantes : Phase 1 : (1) Identification des relations candidates, Phase 2 : (2) Validation de ces relations par généralisation/spécialisation, (3) Validation de ces relations par contextualisation et Phase 3 : (4) Combinaison des deux étapes de validation (2) et (3). Le moteur de recherche utilisé est Google. Nous détaillons par la suite chacune des étapes.

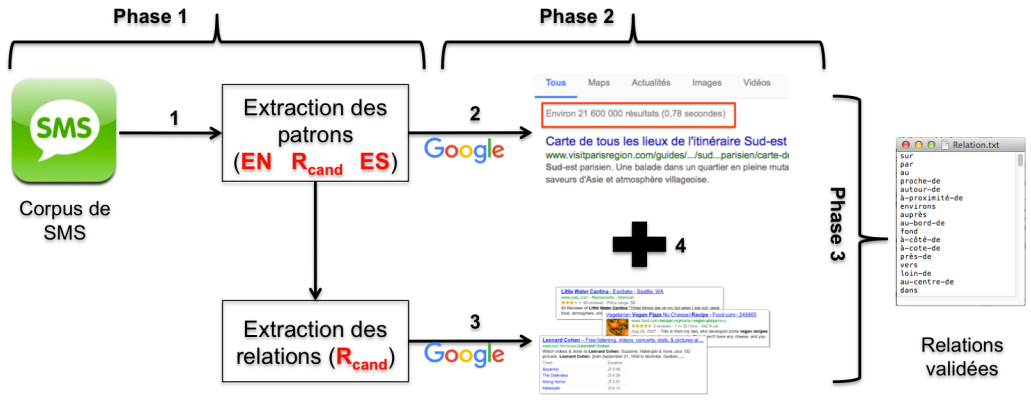

FIGURE 5.2 – Processus global d'extraction et de validation de relations sémantiques associées aux informations spatiales.

5.3.1 Extraction des relations candidates (Phase 1)

La première étape de notre processus consiste à extraire les R_{cand} entre EN présentes dans le corpus traité. Dans cette section, nous concentrons nos travaux sur le corpus SMS contenant les annotations d'entités nommées <PRE>, <NOM> et <SUR>. Dans ce contexte, nous proposons d'identifier les R_{cand} en construisant des patrons que nous nommons les patrons généralisés ($patron_{GR}$). Prenons l'exemple (1) de la liste présentée page 96 (Section 5.1), l'expression « *Personne* est originaire de ES » est un patron généralisé car nous avons substitué les instances des EN par leur type (par exemple, l'instance « Beaumont » est substituée par ES). Chaque $patron_{GR}$ est représenté par le triplet (entité1, R_{cand}, entité2). Ainsi les $patron_{GR}$ extraits doivent respecter des contraintes syntaxiques que nous définissons ci-dessous :

— Les $patron_{GR}$ sont extraits dans une fenêtre de taille fixe (taille de la fenêtre égale à 10 mots [6]). Étant donné que nous travaillons avec un corpus d'écrits non-standards, qui ne respecte pas les règles de la ponctuation, l'utilisation d'une phrase comme fenêtre est inappropriée ;

6. Ce paramètre a été déterminé de manière empirique et au regard des travaux de la littérature (Lison and Kutuzov, 2017)

— Les patrons doivent impliquer au moins une entité spatiale (ES). De manière précise, les trois patrons que nous retenons sont : $Organisation\ R_{cand}\ ES$, $Personne\ R_{cand}\ ES$ et $ES\ R_{cand}\ ES$.

Afin de sélectionner les $patron_{GR}$ contenus dans le corpus de SMS, nous avons extrait trois types d'EN (ES, $Organisation$ et $Personne$).

Extraction des ES. Dans ce but, nous avons dans un premier temps utilisé notre approche décrite dans le chapitre 3 afin d'identifier les ES dans le corpus de SMS.

Extraction des Organisations. Puis, dans un second temps, nous avons utilisé CasEN version Quaero (pour le français) pour identifier les organisations présentes dans le corpus. En effet, le choix de cet outil est motivé par l'intégration aisée de dictionnaires spécifiques dont nous disposons. CasEN est implanté sous le logiciel CasSys de la plate-forme Unitex et est librement mis à disposition des utilisateurs sous licence LGPL-LR. La plate-forme Unitex permet une écriture et une maintenance faciles des patrons (transducteurs) en les présentant à l'utilisateur sous forme de graphes.

Extraction des Personnes. Finalement, dans le cadre du projet sud4science[7] et dans le but de garantir la mise à disposition de corpus de SMS, celui-ci a été entièrement anonymisé afin de masquer l'identité des individus et de supprimer les indications personnelles qui permettraient de les reconnaître (Patel et al., 2013). Cette phase d'anonymisation a été réalisée avec un logiciel appelé Seek&Hide, implémenté dans le cadre du projet, qui est un système semi-automatique permettant de désambiguïser les SMS et de décider si l'anonymisation doit ou non être effectuée. La détection automatique repose sur l'utilisation de deux types de dictionnaires : un dictionnaire qui contient des mots (liste de prénoms enrichie) qui doivent être rendus anonymes et un anti-dictionnaire qui contient des mots qui ne nécessitent pas d'anonymisation (mots du "langage courant"). L'identification repose donc sur des informations lexicales. Enfin, des étudiants en Science du Langage ont validé les identifications automatiques proposées par le système Seek&Hide. Le processus complet est décrit dans (Panckhurst et al., 2013).
Les noms des personnes sont remplacés par les balises <PRE> qui représentent les prénoms, <NOM> qui représentent les noms et <SUR> qui représentent les surnoms. Dans notre cas, nous avons considéré les trois balises (<PRE>, <NOM>, <SUR>) comme EN $Personne$. Aussi, dans le but d'identifier toutes les RS_m présentes dans le corpus, nous avons considéré les pronoms personnels <PRO>

7. http ://sud4science.org

comme *EN Personne*.

Une fois toutes les entités identifiées dans le corpus, nous sélectionnons une liste de $patron_{GR}$. Nous pouvons notamment citer les patrons « *Personne* est originaire de *ES* » et « *ES* a 20min de *ES* » provenant des exemples (1) et (7). En se fondant sur la liste des $patron_{GR}$ (candidats), nous proposons de les valider et d'identifier les différents types de relations à partir des propositions décrites dans les sous-sections suivantes.

5.3.2 Web_{GS} : Validation des relations par généralisation/spécialisation (Phase 2)

Une fois les $patron_{GR}$ identifiés (cf. Section 5.3.1), nous proposons d'extraire, pour chaque $patron_{GR}$, tous les patrons spécialisés $patron_{SP}$ présents dans le corpus de SMS. Un $patron_{SP}$ représente les instances des $patron_{GR}$ mentionnées dans le corpus. Nous associons ainsi à chacune des EN les noms propres, noms communs, pronoms associés présents dans le corpus. Par exemple, pour le $patron_{GR}$ « *Organisation* à coté du *ES* », nous identifions l'instance $patron_{SP}$ « creps à coté du stade philippides ». En effet, à cette étape, nous nous concentrons sur l'instance de l'EN elle-même (par exemple, « creps » et « stade philippides », voir exemple (8) de la Section 5.1) et non pas sur le type de l'EN (*Personne*, *Organisation* et *ES*).

Notons que le corpus de SMS est passé par une phase d'anonymisation, c'est-à-dire que l'EN de type *Personne* n'est pas instanciée dans le corpus. Seules les balises d'anonymisation *Personne* sont présentes. Dans ce contexte, nous avons proposé d'associer à l'entité *Personne* les prénoms les plus fréquents retournés par Google (e.g. $patron_{GR}$: *Personne* est originaire de *ES* $\Rightarrow$ $patron_{SP}$: « Louis est originaire de Beaumont », etc.). À noter que tous les $patron_{SP}$ générés ne sont pas nécessairement pertinents dans notre corpus de SMS.

Notre proposition consiste alors à prendre en considération les $patron_{SP}$ et leurs composants (*EN*, R_{cand} et *ES*) et la page de résultats d'un moteur de recherche afin de filtrer et valider les R_{cand} les plus pertinentes.

Concrètement, dans un premier temps, nous soumettons les $patron_{SP}$ ainsi que chaque élément de ces $patron_{SP}$ comme requête à un moteur de recherche tel que Google. À partir de chacune de ces requêtes, nous récupérons le nombre de résultats, ce qui nous permet de connaître approximativement le nombre de pages indexées contenant ce patron.

Afin de valider et filtrer les relations R_{cand}, nous proposons de mesurer la co-occurrence entre les trois éléments constituant un $patron_{SP}$. Dans ce sens, nous proposons de calculer ce que nous nommons *Web généralisation/spécialisation* (Web_{GS}). La phase de *généralisation* de Web_{GS} considère tous les $patron_{GR}$ présents dans le corpus. Et la phase de *spécialisation* exploite les instances ($patron_{SP}$) présentes dans

le corpus. La mesure Web_{GS} que nous proposons est une mesure fondée sur la fréquence d'occurrences des requêtes sur le Web.
Par exemple, en considérant la requête *"Cern près de Genève"* qui représente un $patron_{SP}$, nous obtenons 2520 résultats via le moteur de recherche Google. Puis, nous requêtons indépendamment toujours via Google les éléments de ce $patron_{SP}$ que sont *"Cern"* ($Organisation$), *"près de"* (R_{cand}) et *"Genève"* (ES) afin de relever le nombre de pages web associé. Nous obtenons respectivement, pour chaque élément, $2,130,000$, $185,000,000$ et $12,600,000$ pages. Ce processus a pour but de mesurer l'association entre les trois éléments composant le $patron_{SP}$.

Web_{GS_Nbre} permet de filtrer les R_{cand} extraites en attribuant le nombre de résultats de la requête $patron_{SP}$ retourné par un moteur de recherche (cf. Équation (5.5)). Nous faisons l'hypothèse que si le $patron_{SP}$ existe au moins une seule fois sur le Web, il représente une information pertinente.

$$Web_{GS\_Nbre}(x, y, z) = nb(x, y, z) \tag{5.5}$$

La deuxième de nos propositions s'appuie sur le coefficient de Dice (Smadja et al., 1996).

Coefficient de Dice. Le coefficient de Dice appliqué aux données textuelles est une mesure de qualité qui permet de calculer la similarité de deux éléments en fonction de leurs fréquences d'apparition (i.e. f) dans un corpus. Le coefficient de Dice est donné par l'équation (5.6).

$$Dice(x, y) = \frac{2 \times f(x, y)}{f(x) + f(y)} \tag{5.6}$$

Par ailleurs, une telle mesure peut utiliser les résultats de requêtes Web comme fréquences d'apparition (Roche and Prince, 2007). Dans ce contexte, l'exemple 5.3.1 présente le résultat obtenu avec la mesure Dice pour le couple (Paris, France). Dans ce cadre *f(Paris, France)* correspondant au nombre de pages retournées contenant les deux entités conjointement *"Paris, France"*, *f(Paris)* correspondant au nombre de pages contenant l'entité *Paris*, *f(France)* représente le nombre de pages Web contenant l'entité *France*. Le score 0.37 montre qu'il existe un lien sémantique entre les deux entités « Paris » et « France ». Plus la valeur est proche de 1, plus les chaînes sont similaires.

Exemple 5.3.1
$Dice(Paris, France) = \frac{2 \times f(Paris, France)}{f(Paris) + f(France)} = \frac{2 \times 1{,}030{,}000{,}000}{1{,}870{,}000{,}000 + 3{,}730{,}000{,}000} = 0.37$

À partir de l'exemple « creps à coté du stade philippides » (cf. Exemple (8), Section 5.1), Dice ne permet pas de calculer la similarité entre les trois éléments « creps », « à coté du » et « stade philippides ». En effet, Dice est une mesure définie pour l'association entre deux éléments (termes) ce qui n'est pas adapté pour

l'association entre trois termes. Ainsi, Dalbelo Basic et al. (2006) ont proposé une extension de la formule de Dice originale à trois termes (cf. Équation (5.7)). Le choix de cette mesure est motivé par les argumentaires liés à son intérêt dans Roche and Kodratoff (2009), elle donne généralement de bons résultats.

$$Dice(x, y, z) = \frac{3 \times f(x, y, z)}{f(x) + f(y) + f(z)} \tag{5.7}$$

Dans notre cas, x, y, z représentent les éléments de $patron_{SP}$ (cf. Équation (5.8)) ; $nb(x, y, z)$ représente le nombre de pages retournées avec de la requête $patron_{SP}$.

$$Web_{GS}(x, y, z) = \frac{3 \times nb(x, y, z)}{nb(x) + nb(y) + nb(z)} \tag{5.8}$$

L'exemple 5.3.2 représente le résultat d'une requête (*"creps à coté du stade philippides"*) :

Exemple 5.3.2
$Web_{GS}(creps, à\ coté\ du, stade\ philippides) = \frac{3 \times nb(creps, à\ coté\ du, stade\ philippides)}{nb(creps) + nb(à\ coté\ du) + nb(stade\ philippides)}$
$= \frac{3 \times 27}{2{,}350{,}000 + 17.700{,}000 + 9{,}760} = 0.00000404$

$nb(creps, à\ coté\ du, stade\ philippides)$ est le nombre de résultats de la requête "$patron_{SP}$" ; $nb(creps)$ est le nombre de résultat du premier élément de $patron_{SP}$ ($Organisation$) ; $nb(à\ cotédu)$ représente le nombre de résultats du deuxième élément de $patron_{SP}$ (à savoir R_{cand}) ; $nb(stadephilippides)$ correspond au nombre de résultats du troisième élément de $patron_{SP}$ (ES).

À partir de l'exemple ci-dessus, nous avons obtenu le score 0.00000404 qui indique qu'il existe une association entre les trois termes « creps », « à coté du » et « stade philippides ». Ce score est assez faible, cependant, étant donné que notre objectif est de mesurer l'association entre les trois éléments, c'est-à-dire vérifier s'il existe un lien entre eux, ce score nous semble suffisant.

Dans notre démarche, nous proposons de changer d'échelle[8] pour que les résultats retournés par Web_{GS} soient exploitables. Le principe que nous proposons est de multiplier Web_{GS} par un coefficient $Coef$ (cf. Équation (5.9)). Une telle opération a l'avantage de maintenir le classement des relations associées à la fonction de rang Web_{GS} tout en facilitant l'analyse.
Ainsi, le changement d'échelle proposé entraîne une modification des valeurs des Web_{GS} mais ne produit aucun effet de bord sur l'interprétation des résultats.

$$WebCoef_{GS}(x, y, z) = \frac{3 \times Coef \times nb(x, y, z)}{nb(x) + nb(y) + nb(z)} \tag{5.9}$$

8. Un changement d'échelle permet de passer d'un repère à un autre en multipliant la taille de tous les objets par un même facteur. Ce changement n'a pas d'influence sur les propriétés des objets.

L'exemple 5.3.3 présente le résultat de la même requête utilisée pour l'exemple précédent mais cette fois-ci en multipliant par un coefficient (*"creps à coté du stade philippides"*) :

Exemple 5.3.3

$WebCoef_{GS}(creps, à \quad coté \quad du, stade \quad philippides) = \frac{3\times1.000.000\times nb(creps,à\ coté\ du,stade\ philippides)}{nb(creps)+nb(à\ coté\ du)+nb(stade\ philippides)} = \frac{3\times1,000,000\times27}{2,350,000+17,700,000+9760} = 4.04$

$WebCoef_{GS}$ de deux termes

En respectant le principe de la mesure de Dice originale (cf. Équation (5.6)), nous proposons de décomposer le $patron_{SP}$ en deux afin de mesurer l'association entre chaque paire. Ce principe repose sur le fait que plus le nombre de termes, qui doit mesurer l'association entre eux, est réduit, plus la probabilité que les deux paires traitées soient présentes dans les mêmes textes ou les mêmes documents est élevée.

Dans ce cas, nous mesurons l'association entre les deux premiers éléments x et y puis les deux derniers éléments y et z. Ensuite, nous calculons indépendamment la somme et le produit des deux scores (cf. Équations (5.11) et (5.10)). Le produit permet de pondérer les résultats d'une requête par rapport à l'autre (cf. Équation 5.10). La somme consiste, quant à elle, à cumuler les résultats des deux requêtes (cf. Équation 5.11).

$$WebCoef_{GS\_Mult}(x, y, z) = \frac{2 \times Coef \times nb(x, y)}{nb(x) + nb(y)} \times \frac{2 \times Coef \times nb(y, z)}{nb(y) + nb(z)} \quad (5.10)$$

$$WebCoef_{GS\_Add}(x, y, z) = \frac{2 \times Coef \times nb(x, y)}{nb(x) + nb(y)} + \frac{2 \times Coef \times nb(y, z)}{nb(y) + nb(z)} \quad (5.11)$$

Les exemples 5.3.4 et 5.3.5 présentent les résultats de la requête *"creps à coté du stade philippides"* en calculant $WebCoef_{GS\_Mult}$ et $WebCoef_{GS\_Add}$:

Exemple 5.3.4

$Web_{GS}(creps, à \quad coté \quad du, stade \quad philippides) = \frac{2\times Ceof\times nb(creps,à\ coté\ du)}{nb(creps)+nb(à\ coté\ du)} \times \frac{2\times Ceof\times nb(à\ coté\ du,stade\ philippides)}{nb(à\ coté\ du)+nb(stade\ philippides)} = \frac{2\times1,000,000\times8}{2,350,000+17,700,000} \times \frac{2\times1,000,000\times348}{17,700,000+9760} = 31.44$

Exemple 5.3.5

$Web_{GS}(creps, à \quad coté \quad du, stade \quad philippides) = \frac{2\times Ceof\times nb(creps,à\ coté\ du)}{nb(creps)+nb(à\ coté\ du)} + \frac{2\times Ceof\times nb(à\ coté\ du,stade\ philippides)}{nb(à\ coté\ du)+nb(stade\ philippides)} = \frac{2\times1,000,000\times8}{2,350,000+17,700,000} + \frac{2\times1,000,000\times348}{17,700,000+9760} = 40.10$

Finalement, nous comparons les mesures proposées ci-dessus (Web_{GS_Nbre}, $WebCoef_{GS}$, $WebCoef_{GS\_Mult}$ et $WebCoef_{GS\_Add}$) afin d'en déterminer la plus pertinente. Les résultats de comparaison sont présentés en Section 5.4.1. Une fois la mesure la plus pertinente sélectionnée, nous attribuons à chaque $patron_{SP}$ un score Web_{GS}.

Notons que chaque $patron_{GR}$ représente un ensemble d'instances $patron_{SP}$. Par exemple, les instances {« Louis en suisse », « Lucas en suisse », « Chloé en suisse »} sont associées au $patron_{GR}$ « *Personne en ES* ». Nous obtenons, comme résultat de cette étape, un ensemble de valeurs de Web_{GS} pour chaque $patron_{GR}$. Puis, afin d'associer la valeur globale au $patron_{GR}$, nous calculons Web_{GS_max}, Web_{GS_sum} ou Web_{GS_moy}, comme suit :

$$\begin{cases} Web_{GS\_max} = \max\limits_{1..n}(Web_{GSi}) & \text{(5.13a)} \\ Web_{GS\_sum} = \sum\limits_{i=1}^{n} Web_{GSi} & \text{(5.13b)} \\ Web_{GS\_moy} = \dfrac{\sum_{i=1}^{n} Web_{GS1}}{n} & \text{(5.13c)} \end{cases} \tag{5.13}$$

Web_{GS} représente les mesures proposées dans cette section, à savoir Web_{GS_Nbre}, $WebCoef_{GS}$, $WebCoef_{GS\_Add}$ et $WebCoef_{GS\_Mult}$; n représente le nombre d'instances $patron_{SP}$ de chaque $patron_{GR}$. Web_{GSi} est la valeur obtenue pour la ième instance ; Web_{GS_max} est la valeur maximale des Web_{GS} pour les n instances ; Web_{GS_sum} est la somme des Web_{GS} pour les n instances ; Web_{GS_moy} est la moyenne des Web_{GS} pour les n instances. Ces différentes propositions sont évaluées en Section 5.4 de ce chapitre.

À partir de l'exemple « <PRE_5/> en suisse » (cf. (10), Section 5.1), nous obtenons $Web_{GS\_Nbre\_max}$:

Exemple 5.3.6
$Web_{GS\_Nbre\_max} = \max[Web_{GS\_Nbre}(Lucas, en, suisse),\ Web_{GS\_Nbre}(Louis, en, suisse),\ Web_{GS\_Nbre}(Emma, en, suisse),\ Web_{GS\_Nbre}(Chloè, en, suisse)] = \max[10,\ 82300,\ 20100,\ 3] = 82,300$

L'approche présentée dans cette section s'appuie sur des mesures d'association des triplets constituant les $patron_{SP}$. Ces mesures sont fondées sur la fréquence de leur apparition sur le Web, en l'occurrence les résultats du moteur de recherche Google. La limite de ce type d'approches est que la probabilité d'identifier l'association entre les éléments de $patron_{SP}$ peut être faible. En effet, la probabilité de trouver les trois éléments consécutifs de $patron_{SP}$ dans un même texte indexé par le moteur de recherche est faible. Par exemple, la requête *"Louis est originaire*

de Beaumont" ne retourne aucun résultat avec Google alors que la relation « est originaire de » est tout à fait pertinente.

Rappelons que notre objectif est de valider les relations R_{cand}. Dans la suite de ces travaux, nous considérons que les requêtes à partir des seules relations candidates R_{cand} (par exemple, « est originaire de ») consolident le processus de validation en prenant en compte des informations contextuelles.

5.3.3 Web_{Cont} : Validation des relations par contextualisation

Dans la section précédente, nous avons proposé des fonctions de rang pour valider les relations les plus pertinentes. Dans cette section, nous décrivons comment les valider en nous appuyant, contrairement à l'approche Web_{GS} (cf. Section 5.3.2), sur le contenu des pages Web retournées par les moteurs de recherche.

Suite à la soumission d'une requête (cf. Section 5.3.2), le moteur de recherche présente les résultats sous forme d'une liste de sites Web. Pour chacun de ces sites, un aperçu du contenu de la page Web est présenté (snippets), justifiant le résultat retourné en mettant en gras les termes initialement présents dans la requête (cf. Figure 5.1). Par défaut, les moteurs de recherche organisent les résultats retournés sous forme de pages. Ces pages contiennent un nombre de résultats spécifiques. Par exemple Google affiche 10 résultats par page lors d'une recherche. Il est néanmoins possible d'en afficher plus.

Au cours de cette étape, nous nous concentrons sur les R_{cand} et les informations contextuelles issues des snippets. En effet, nous supposons que si nous requêtons les R_{cand} entre deux termes inconnus (" $*$ R_{cand} $*$ "), et nous extrayons les snippets associées à cette requête, nous pouvons valider la R_{cand} si cette dernière se trouve entre une EN ($Personne$, $Organisation$ ou ES) et une ES.

Le moteur de recherche offre de nombreuses fonctionnalités pour affiner une recherche. L'une de ces fonctionnalités est de chercher un terme connu entre deux termes inconnus en utilisant l'astérisque[9] (*). Par exemple, la requête ** est originaire de ** permet de retrouver l'expression « Caroline Louat **est originaire de** Chemilly sur Yonne ».

De ce fait, nous traitons chaque requête " $*$ R_{cand} $*$ " avec un moteur de recherche pour obtenir les k meilleurs documents récupérés et leur snippets. Puis, en mobilisant des outils d'extraction d'entités nommées, nous identifions les EN

9. * : L'astérisque utilisée dans une requête signale à Google qu'il y a un mot-clé inconnu dans la requête de recherche d'un utilisateur mais que ce dernier autorise à le remplacer librement. Dans ce cas, le moteur de recherche se chargera de proposer le mot clé en question qui correspond le mieux à la requête (http ://cognitos.ca/fr/blogue/13-comment-effectuer-une-recherche-sur-google-comme-un-pro).

(*Personne*, *Organisation*, et *ES*) présentes dans les snippets. Cette tâche est essentielle afin de vérifier la présence des R_{cand} entre une *EN* et une *ES* pour le processus de validation.

Pour identifier les *ES* présentes dans les snippets, nous utilisons notre approche décrite dans le Chapitre 3 (Section 3.3). Ensuite, nous utilisons Polyglot (Al-Rfou et al., 2015) pour identifier les autres types d'*EN*, c'est-à-dire *Personne* et *Organisation*.

Une fois les snippets annotées, notre approche permet d'identifier et extraire l'ensemble des $patron_{GR}$ présents, c'est-à-dire *Organisation* R_{cand} *ES*, *Personne* R_{cand} *ES*, ou *ES* R_{cand} *ES*. Puis, nous calculons ce que nous nommons Web_{cont} (*Web contextualisation*), qui représente le nombre d'occurrences de chaque patron dans les k snippets.

Par exemple, en lançant la requête *"* est originaire de *"* (cf. Exemple (1) Section 5.1) sur Google, nous obtenons parmi les 20 premières snippets, la snippet suivante :
*"**Titre**" : "Caroline Louat est originaire de Chemilly sur Yonne et réside à ...",*
*"**description**" : "Caroline Louat est originaire de Chemilly sur Yonne et réside à Saint Lambert au Québec. Par Catherine Marchesinle mercredi 15 février 2017. Podcasts ..."*

Puis, en annotant la snippet ci-dessus avec les *EN Personne*, *Organisation* et *ES*, nous obtenons la présentation suivante : *"**Titre**" : "<I-PER>Caroline Louat</I-PER> est originaire de <ES>Chemilly sur Yonne</ES> et réside à ...",*
*"**description**" : "<I-PER>Caroline Louat</I-PER> est originaire de <ES>Chemilly sur Yonne</ES> et réside à <ES>Saint Lambert</ES> au <ES>Québec</ES>. Par <I-PER>Catherine Marchesinle</I-PER> mercredi 15 février 2017. Podcasts ..."*

Finalement, une fois la snippet annotée, nous extrayons tous les $patron_{GR}$ présents dans cette snippet. En effet, à partir de snippet ci-dessus, nous avons obtenu les deux $patron_{GR}$ *Personne* est originaire de *ES* et *ES* au *ES*.

Comme le montrent les expérimentations détaillées en Section 5.4 (Sections 5.4.1 et 5.4.2), les deux approches présentées ci-dessus (cf. Sections 5.3.2 et 5.3.3) sont perfectibles.

En effet, en considérant la R_{cand} « est originaire de », cette dernière n'a pas été validée par notre première approche (Web_{GS}), cependant elle a été validée par notre deuxième approche (Web_{Cont}). Dans ce sens, nous proposons de les combiner afin d'améliorer le marquage des relations sémantiques.

5.3.4 Combinaison (Phase 3)

En analysant qualitativement les résultats obtenus (cf. Section 5.4), nous remarquons que Web_{GS} (Section 5.3.2) a un meilleur comportement pour les $patron_{SP}$ qui ont obtenu un nombre de résultats (i.e. le nombre de pages retournées) très élevé. Cependant, Web_{Cont} (Section 5.3.3) se comporte mieux pour les $patron_{SP}$ qui ont obtenu un faible score de Web_{GS} (*Web généralisation/spécialisation* voir Section 5.3.2).

Par exemple, les $patron_{SP}$ qui contiennent les relations R_{cand} comme « sur », « à », « a », « en », etc. ont obtenu un nombre de résultats élevé avec la première approche (e.g. le nombre de résultats retourné par Google pour la requête *"Louis en suisse"* est égal à $82,300$). Tandis que, pour la deuxième approche, la probabilité de trouver ces R_{cand} (e.g. « sur », « à », « en », etc.) entre une EN et une ES dans les k meilleurs snippets est extrêmement faible car ce type de relations représente des prépositions qui sont très utilisées dans différents contextes. Par exemple, pour la R_{cand} « sur », nous n'avons trouvé aucun $patron_{GR}$ dans les 20 premières snippets retournées par Google.

Au regard de cette première analyse, nous faisons l'hypothèse que R_{cand} sera considérée comme pertinente si Web_{GS} retourne une valeur supérieure à 0. Sinon, nous calculons la valeur de Web_{Cont}. Si cette valeur est supérieure à 0, nous validons les R_{cand}, sinon nous les considérons comme non pertinentes.

En d'autres termes, si le nombre de pages Web issues de requêtes est élevé, nous privilégions la première approche par généralisation/spécialisation, sinon nous favorisons la deuxième approche par contextualisation. Par exemple, le $patron_{SP}$ « Louis est originaire de Beaumont » obtient un score de $WebCoef_{GS}$ égal à 0 et un score de Web_{Cont} égal à 4 en considérant les 20 meilleurs snippets. Cela permet de valider la R_{cand} « est originaire de » qui est tout à fait pertinente.

Nous détaillons dans la section suivante les résultats de nos expérimentations sur les différentes propositions présentées dans cette section.

5.4 Expérimentations

Cette section présente les expérimentations menées pour valider nos propositions et les résultats obtenus. Nous avons réalisé nos expérimentations avec le corpus de SMS 88milSMS. Nous avons mesuré la qualité de nos résultats sur le même échantillon de 1000 SMS utilisé pour les contributions précédentes.
Dans nos expérimentations, les $patron_{SP}$ sont soumis sous forme de requêtes au moteur de recherche Google [10]. L'objectif de ces expérimentations est d'évaluer les R_{cand} sélectionnées par nos approches automatiques. Nous avons alors évalué les

10. https ://www.google.fr

trois processus de sélection décrits dans ce chapitre. La première étape consiste à évaluer les R_{cand} par généralisation/spécialisation (cf. Section 5.4.1). La deuxième étape concerne la validation des R_{cand} par contextualisation (cf. Section 5.4.2). La troisième présente les résultats obtenus avec la combinaison des deux processus précédents (cf. Section 5.4.3).

5.4.1 Validation des relations par généralisation/spécialisation

L'idée que nous avons mise en œuvre dans cette étude est de soumettre les requêtes $patron_{SP}$ au moteur de recherche Google afin de valider les R_{cand} extraites à partir d'un corpus de SMS. Afin de soumettre les $patron_{SP}$ au moteur de recherche Google, nous avons associé à l'entité *Personne* des prénoms provenant de la liste des prénoms les plus donnés en France en 2015 [11]. En suivant ce principe, le Tableau 5.1 présente le $patron_{SP}$ « <PER/> en <ES>suisse</ES> » et le nombre de résultats retourné par Google pour chaque $patron_{SP}$.

<PER/> en <ES>suisse</ES>	Web_{GS_Nbre}
*"**Lucas** en suisse"*	10
*"**Louis** en suisse"*	82,300
*"**Emma** en suisse"*	20,100
*"**Chloé** en suisse"*	3

TABLE 5.1 – Extrait de résultats de recherche de $patron_{GR}$ « *Personne* R_{cand} *ES* ».

Pour valider semi-automatiquement les R_{cand}, nous menons une série d'expérimentations afin d'identifier la mesure Web_{GS} (*Web généralisation/spécialisation*) la plus pertinente. Pour évaluer la qualité des résultats obtenus, nous avons calculé la précision, le rappel et la F-mesure (cf. Chapitre 3, Section 3.4.3).

En effet, des annotations manuelles ont été effectuées sur l'échantillon sélectionné. En utilisant le corpus annoté manuellement et le corpus annoté par notre système, nous évaluons la capacité de notre système à extraire toutes les relations pertinentes présentes dans cet échantillon (c'est-à-dire, calculer le rappel).

Sur un échantillon qui contient 100 $patron_{SP}$, et en utilisant Web_{GS_Nbre} qui représente la présence du $patron_{SP}$ tel qu'il est sur le Web (cf. Section 5.5), nous avons pu identifier 28 relations pertinentes, dont quelques exemples sont présentés dans le Tableau 5.2. Dans le cadre de nos expérimentations sur ce jeu de données, nous avons obtenu une précision égale à 0.88 et un rappel de 0.50 (cf. Tableau 5.3). Le Tableau 5.3 présente les résultats obtenus avec les différentes mesures de Web_{GS}. Nous remarquons que les mesures $WebCoef_{GS}$, Web_{GS_Nbre} et $WebCoef_{GS\_Mult}$ ont obtenu un faible score du rappel, en particulier pour le $WebCoef_{GS}$ qui a obtenu un

11. http ://www.journaldesfemmes.com/prenoms/classement/prenoms/les-plus-donnes

score de 0.25. Ces mesures reflètent la spécificité des patrons extraits (forte précision et faible rappel). Cependant, en utilisant la mesure $WebCoef_{GS\_Add}$, le score du rappel est augmenté à 0.80. En effet, nous avons obtenu une F-mesure égale à 0.83 en identifiant 45 relations pertinentes en utilisant un coefficient égal à $1,000,000$. En considérant, par exemple, le $patron_{SP}$ « <PER>on<PER/> est ver <ES>beziers</ES> », la relation candidate « est ver » est validée en utilisant $WebCoef_{GS\_Add}$. En effet, nous avons tout d'abord mesuré l'association entre les deux éléments *"on est ver"*, qui se trouvent 5700 fois dans le Web. Puis, nous avons mesuré l'association entre les deux éléments *"est ver beziers"* qui est présente 2 fois sur le Web. Finalement, nous avons calculé $WebCoef_{GS\_Add}$ qui donne un score égal à 17.

En revanche, en considérant, par exemple, les $patron_{SP}$ « <PER/> est originaire de <ES>Beaumont</ES> » et $patron_{SP}$ « <ES>masgot</ES> a 20min de <ES>gueret</ES> », aucun résultat n'est retourné par Google (cf. Tableau 5.2). Cela s'explique par l'absence d'association entre R_{cand} et les $EN\ Personne$ et ES. Ce manque d'information a été une motivation pour la proposition décrite en Section 5.3.3. La Section 5.4.2 décrit les résultats de nos expérimentations associées à cette contribution.

$patron_{GR}$	Web_{GS_Nbre}
PER a ES	6250
PER bosse en ES	4080
PER ds ES	5
ES a 20min de ES	0
PER est originaire de ES	0

TABLE 5.2 – Extrait de $patron_{SP}$ identifiés par la mesure Web_{GS_Nbre}.

	Précision	**Rappel**	**F − mesure**
$WebCoef_{GS}$	0.87	0.25	0.39
$WebCoef_{GS\_Add}$	0.88	**0.80**	**0.83**
$WebCoef_{GS\_Mult}$	**0.94**	0.59	0.71
Web_{GS_Nbre}	0.88	0.50	0.64

TABLE 5.3 – Résultats de Web_{GS}.

5.4.2 Validation des relations par contextualisation

Dans la deuxième série d'expérimentations, nous nous penchons sur les R_{cand} extraites en prenant en compte les informations contextuelles issues des snippets. Pour évaluer nos résultats, nous nous appuyons sur les k meilleurs snippets résultantes du moteur Google. Dans notre cas, nous nous limitons à $k = 20$, en ne prenant pas en

compte les pages web contenant la conjugaison, la traduction, etc[12]. Le Tableau 5.4 présente un extrait des snippets des R_{cand} « a 20min de » et « est originaire de ». En lançant les deux requêtes *"* a 20min de *"* et *"* est originaire de *"* sur le moteur de recherche Google, nous avons obtenu pour chaque R_{cand}, les deux snippets présentes dans le Tableau 5.4. Ces snippets ont été sélectionnées parmi les 20 premières retournées par Google. Chaque snippet est présentée par son titre et sa description.

R_{cand}	snippets
a 20min de	"**Titre**" : "Cosy appartement ***a 20min de*** l'aéroport - Appartements à louer à ...", "**Description**" : "Chambre privée pour 31 31 €. Mon logement est proche de l'aéroport (20min en voiture) nous assurons également le transfert de et vers l'aéroport), nous avons ..."
	"**Titre**" : "Chambre avec Vue sur Parc ***a 20min de*** Fontainebleau - Maisons à ...", "**Description**" : "Chambre privée pour 70 €. Chambre au calme , dans une maison de style "île de France" Accès direct de l'autoroute A77 (5 min) qui rejoint la A6 (10min) ..."
est originaire de	"**Titre**" : "Vrai ou faux ? Usain Bolt **est originaire de** la Barbade. - Quipo Quiz", "**description**" : "Usain Bolt, l'athlète qualifié d'homme le plus rapide du monde, est Jamaïcain."
	"**Titre**" : "Caroline Louat ***est originaire de*** Chemilly sur Yonne et réside à ...", "**Description**" : "Caroline Louat ***est originaire de*** Chemilly sur Yonne et réside à Saint Lambert au Québec. Par Catherine Marchesinle mercredi 15 février 2017. Podcasts ..."

TABLE 5.4 – Extraits de snippets des R_{cand} « a 20min de » et « est originaire de »

Le Tableau 5.5 présente les résultats obtenus dans cette section en terme de précision, rappel et F-mesure sur les 20 premières snippets retournées par le moteur de recherche Google pour chaque R_{cand}. Ce tableau met en exergue une précision élevée (0.84) et un rappel assez faible (0.29), le score F-mesure étant par conséquent assez faible, i.e. 0.43. Notons que les $patron_{GR}$ avec les R_{cand} les plus représentatifs (par exemple, « au centre de », « en », « a », « dans », etc.) n'apparaissent pas dans les 20 premières snippets retournées. Ceci explique le résultat du rappel assez faible.

12. Par exemple, les résultats liés aux pages Web telles que *http ://www.linguee.fr*, *http ://www.reverso.net*, etc.

Précision	**Rappel**	**F-mesure**
0.84	0.29	0.43

TABLE 5.5 – Résultats de Web_{Cont} sur les 20 premiers snippets retournés par Google

5.4.3 Combinaison

Dans cette section, nous présentons les résultats obtenus en combinant les deux approches Web_{GS} et Web_{Cont}. Comme nous pouvions le prévoir, le meilleur score de F-mesure est obtenu par la combinaison de nos deux propositions à savoir, Web_{GS} et Web_{Cont} (cf. Tableau 5.6).
Nous pouvons remarquer une amélioration de la F-mesure, ce qui tend à prouver que la combinaison améliore globalement la qualité d'identification et de validation des R_{cand} extraites. Ce résultat est encourageant et montre la faisabilité de l'approche et sa capacité à identifier et à valider automatiquement des relations sémantiques entre les EN faisant intervenir des ES (par exemple : « ds », « a 5 min de », « a 20min de », « est originaire de », etc.).

	Précision	**Rappel**	**F-mesure**
généralisation/spécialisation-$WebCoef_{GS\_Add}$	0.92	0.83	0.86
contextualisation-Web_{Cont}	0.84	0.29	0.43
Combinaison	**0.92**	**0.86**	**0.89**

TABLE 5.6 – Résultats de la combinaison de Web_{GS} et Web_{Cont}

5.5 Discussion

Dans ce chapitre, nous avons présenté une nouvelle approche pour l'identification des relations sémantiques dans des SMS. En proposant une approche qui exploite des données externes (le nombre de pages et les snippets) provenant des réponses à des requêtes du moteur de recherche Google, nous avons identifié de nouvelles variantes de relations sémantiques issues de l'écriture SMS. Nous pouvons citer les exemples suivants : la relation « est ver » qui se trouve dans le $patron_{GR}$ « $Personne$ **est ver** ES », la relation « suis ds » qui se trouve dans le $patron_{GR}$ « $Personne$ **suis ds** ES » (cf. Exemples (4) et (5), Section 5.1). Nos résultats montrent que nous avons pu obtenir une méthode robuste qui permet de prendre en compte les spécificités de l'écriture de SMS, avec un score de F-mesure de 0.86 obtenu par la mesure $WebCoef_{GS\_Add}$ (Web généralisation/spécialisation) et 0.43 obtenu par Web_{Cont} (Web contextualisation) (cf. Tableaux 5.3 et 5.5). La combinaison des deux

approches donne le meilleur résultat de 0.89 en terme de F-mesure (cf. Tableau 5.6).

Ces travaux tendent à montrer qu'il est pertinent d'exploiter une certaine forme de *popularité* des expressions indexées par les moteurs de recherche pour valider des candidats extraits à partir de corpus très spécifiques.

Afin de réaliser ce travail, nous nous sommes appuyés sur Google search API [13] qui permet d'interroger un navigateur web à partir d'un programme Python. Cette API nous permet de retourner le nombre de résultats de recherche, les liens url, les titres ainsi que les descriptions des documents retournés pour une requête donnée.

Cette API permet d'exécuter 100 requêtes par jour et ce nombre peut-être augmenté à une valeur maximale de 1000 requêtes [14]. Dans notre étude, pour calculer les scores des Web_{GS} (cf. Section 5.3.2), nous avons évalué le nombre de requêtes réalisées. Pour la première mesure Web_{GS_Nbre} (cf. Équation (5.5)), nous avons effectué 100 requêtes correspondant aux 100 $patron_{SP}$. Pour Web_{GS} et $WebCoef_{GS}$ (cf. Équations (5.8) et (5.9)), nous avons requêté 100 $patron_{SP}$ ainsi que les trois éléments composant chaque instance, ce qui produit 400 requêtes. Finalement, pour les deux mesures $WebCoef_{GS\_Add}$ et $WebCoef_{GS\_Mult}$, nous avons exécuté 200 requêtes ce qui résulte à un total de 1300 requêtes.

Notons que bien que le nombre de pages retourné par Google search API soit parfois légèrement différent des résultats de Google, la tendance reste du même ordre. De ce fait, il apparaît pertinent d'utiliser l'API.

L'approche peut avoir plusieurs voies d'amélioration. La plupart des erreurs manuellement identifiées sont liées à la spécificité de l'écriture SMS et à la complexité des relations. Nous pouvons notamment mentionner ici les exemples suivants : « $Organisation$ est dans la partie entre ES et ES », « $Personne$ veux voyagé ds tte ES », « $Personne$ est remonte hier sur ES », etc. (cf. Exemples (11), (12) et (13)).

L'une de nos proposition pour résoudre ce problème est d'utiliser les k-skip-n-grammes (Guthrie et al., 2006) pour identifier ces R_{cand}. En effet, les k-skip-n-grammes nous permettent d'identifier les différentes combinaisons de R_{cand} en relâchant des contraintes liées à la présence exacte d'expression. Par exemple, pour la relation « est remonte hier sur » les k-skip-n-grammes nous permettant d'obtenir les combinaisons suivantes : « est remonte hier », « est remonte sur », « est hier sur » et « remonte hier sur ». En utilisant ces trois derniers syntagmes dans notre approche détaillée en Section 5.3.3, nous pouvons enrichir les recherches Web permettant de valider certaines expressions très spécifiques.

Enfin, nous proposons de classifier ces relations selon leurs types. Pour atteindre cet objectif, nous proposons, dans un premier temps, d'identifier les différents types de relations sémantiques (spatiales et non spatiales) en créant un thésaurus conte-

13. http ://www.google.com/apis/
14. https ://developers.google.com/analytics/devguides/reporting/core/v3/limits-quotas

nant des relations et leurs types (par exemple, « dans » de type *inclusion*, « vers » de type *direction*, etc.). Ensuite, nous proposons de calculer la similarité lexicale entre les noms de nouvelles relations (par exemple, « ver », « ds », etc.) et les relations présentées dans le thésaurus (par exemple, « vers », « dans », etc.). En effet, l'identification de type de relations permet d'affiner la compréhension de l'information présente autour de la relation.

Dans ce contexte, nous envisageons comme perspectives d'améliorer notre système d'identification de relations sémantiques ainsi que d'étudier les différents types de ces relations.

Conclusion Générale et Perspectives

Chapitre

6 Conclusion et Perspectives

Contents

6.1 Conclusion et Perspectives

Ce travail de thèse a été initié dans le but d'identifier et d'étudier les variantes d'expression de l'information spatiale, à savoir les entités spatiales absolues et relatives, à partir de messages courts. Nous commençons par rappeler les principales contributions formalisées dans le cadre de cette thèse. Puis, nous présentons quelques perspectives qui s'inscrivent dans la continuité directe de ces travaux.

6.1.1 Synthèse des travaux

Dans la communauté scientifique, de nombreux travaux liés à l'analyse et l'extraction d'informations spatiales s'appuient sur l'exploitation de données textuelles standards provenant de différents types de sources (articles de presse, documents techniques, revues, etc.) afin d'améliorer l'efficacité des systèmes de recherche d'information.

Récemment, avec le développement du Web et de la communication médiée, les besoins liés à l'identification et l'analyse fine des informations spatiales est encore plus notable. Aussi, les diverses avancées technologiques ont rendu la création, l'acquisition, le stockage et le partage de documents numériques de plus en plus accessibles. Face aux coûts élevés que peut représenter le traitement manuel de ces masses de données textuelles, les analyses automatiques deviennent indispensables.

Les informations textuelles disponibles comportent souvent une multitude d'expressions différentes ou variantes pour exprimer une même entité ou relation spatiale. Ceci rend la reconnaissance et l'extraction automatique de ces informations difficiles. Ainsi, les méthodes classiques d'extraction d'informations ne s'avèrent pas adaptées au traitement et à l'analyse de corpus de messages courts. En effet, en raison des particularités lexicales et syntaxiques de ce type de corpus, les performances de ces approches diminuent nettement. De plus, l'absence de ressources linguistiques spécialisées (lexiques, corpus annotés) ne permet pas d'adapter les outils existants à ce type de texte.

Dans ce contexte, nous avons proposé des méthodes génériques adaptées aux différents types de corpus traités (SMS, tweets, articles de presse) permettant d'extraire de façon précise et exhaustive les informations spatiales.

6.1.2 Contributions

Afin de relever les défis mentionnés précédemment, nous avons donc proposé dans cette thèse des approches pour l'identification automatique de nouvelles variantes d'entités spatiales absolues ainsi que des relations sémantiques (spatiales et non-spatiales) à partir des corpus de messages courts français. Ces approches sont fondées sur trois principales contributions.

Notre première contribution se concentre sur le problème de reconnaissance et d'identification de nouvelles variantes d'entités spatiales absolues à partir de corpus

de messages courts. Dans ce sens, nous avons proposé une approche qui exploite des connaissances linguistiques combinées à des dictionnaires (liste de pays, villes, etc.) qui sont pré-définis manuellement par des experts. En effet, notre approche combine l'analyse statistique et lexicale. L'analyse statistique consiste à calculer la similarité entre deux chaînes de caractères en utilisant différentes mesures de similarité. Ces dernières permettent de calculer une certaine proximité lexicale. Quant à l'analyse lexicale, elle repose sur l'utilisation des méthodes de désaccentuation ainsi que la prise en compte des préfixes similaires pour identifier des proximités lexicales.

Deux corpus de messages courts français ont été utilisés afin d'évaluer nos propositions. Les expérimentions menées ont montré que la combinaison de l'analyse statistique et lexicale améliore de manière significative l'extraction de nouvelles variantes d'entités spatiales absolues. La comparaison de notre méthode avec celles existantes de l'état de l'art a mis en exergue les bonnes performances de nos propositions permettent d'enrichir les dictionnaires d'entités spatiales absolues.

Notre deuxième contribution est constituée de deux parties : (i) la prédiction du type de relations spatiales et (ii) l'identification de nouvelles variantes/formes d'expression de relations spatiales.

La première partie présente une typologie de relations entre entités spatiales absolues permettant d'identifier, finement et de manière automatique, les types de relations spatiales exprimées dans les textes. Ainsi, nous avons proposé une méthode hybride combinant des informations lexicales et contextuelles à une approche de fouille de textes pour prédire le type de relation spatiale. Plus précisément, nous avons proposé de prédire une classe relative aux relations candidates en utilisant des mesures de similarité lexicale et un algorithme de classification (K plus proches voisins). Par ailleurs, nous avons pris en considération le contexte des relations afin d'améliorer la prédiction. Ensuite, nous avons combiné plusieurs mesures de similarité vectorielle avec différentes méthodes de pondération (*TF-IDF, nombre d'occurrences* et *confiance*). Finalement, nous avons proposé une analyse comparative de ces deux approches (lexicales et contextuelles) et proposé une combinaison pour l'identification automatique du type de relations spatiales.

Étant donné qu'il n'existe pas encore, à notre connaissance, de corpus français annoté par des types de RS, les expérimentations ont été réalisées sur un corpus anglais annoté (SpRL (Kordjamshidi et al., 2011)). Le corpus représente un benchmark reconnu dans le domaine. Nos résultats montrent que la combinaison des deux méthodes se comporte mieux que chaque méthode individuellement. En effet, elle permet d'explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des différentes approches (lexicales et contextuelles).

La deuxième partie de notre contribution liée à l'étude des relations spatiales repose sur l'hypothèse que les corpus de messages courts contiennent de nouvelles formes/variantes de relations spatiales. La prise en compte de ces dernières est cruciale pour les études associées à la spatialité véhiculée dans les textes en communication médiée. Nos propositions méthodologiques consistent à identifier de nouvelles

formes de relations spatiales dans des corpus non-standards. Ainsi, une chaîne de traitements qui combine une analyse morphosyntaxique et une approche fréquentiste a été mise en œuvre.

Nous avons ensuite mesuré les performances de notre proposition à l'aide de deux corpus de messages courts, à savoir les SMS et les tweets. Les résultats de nos expérimentations montrent que notre approche nous a permis d'obtenir des résultats très satisfaisants pour l'identification de nouvelles formes/variantes de relations spatiales.

Finalement, la troisième contribution concerne l'identification des relations sémantiques impliquant des entités spatiales absolues. L'objectif de cette contribution est d'identifier de nouveaux types de relation sémantique faisant intervenir une entité spatiale absolue, et ainsi d'enrichir la typologie des relations spatiales définies dans la communauté scientifique. À cet effet, nous avons proposé, comme première étape, d'extraire l'ensemble des relations candidates associant les entités nommées (*Personnes* et *Organisation*) et les entités spatiales absolues. Nous avons ainsi annoté notre corpus à l'aide d'outils spécifiques en utilisant notamment celui de notre première contribution. Puis, nous avons extrait les relations candidates que nous validons en exploitant les données issues du Web. Plus précisément, nous étudions dans quelle mesure la combinaison « entités nommées + relation » est présente sur le Web. Cette étude s'appuie notamment sur des résultats de moteurs de recherche, la mise en place de fonctions de rang ainsi que sur l'étude contextuelle des snippets.

Ces travaux tendent à montrer qu'il est pertinent d'exploiter une certaine forme de *popularité* des expressions indexées par les moteurs de recherche pour valider et filtrer des candidats extraits à partir de corpus très spécifiques.

Ces trois contributions permettent d'identifier et extraite de nouvelles formes d'expression de l'information spatiale exprimée dans des corpus textuels de type non-standard. Les évaluations menées dans cette thèse, sur des corpus réels principalement en français (SMS, tweets et articles journalistiques), soulignent l'intérêt de ces contributions. Les ressources produites dans le cadre de cette thèse (corpus de messages courts annotés et lexiques) sont également mis à disposition en ligne pour permettre à la communauté scientifique d'étendre la démarche (Zenasni et al., 2017a;b).

6.1.3 Perspectives

Dans cette thèse, nous avons proposé différentes contributions d'extraction automatique de nouvelles formes d'expression des entités spatiales absolues à partir de messages courts. Nous nous sommes aussi intéressés à l'identification automatique des relations spatiales et sémantiques à partir de ces corpus. Bien que nous ayons montré la faisabilité de nos propositions sur des données réelles, il y a encore quelques limites et des améliorations à réaliser. Nous discutons, dans ce qui

suit, de nouvelles contributions possibles qui nous semblent pertinentes.

En ce qui concerne notre première proposition, à savoir l'identification de nouvelles variantes d'entités spatiales absolues, notre système a retourné des erreurs liées :

i *aux abréviations des ESA, ce qui complique l'identification.* Par exemple, notre approche ne permet pas actuellement d'associer automatiquement l'expression « mtp » à l'entité spatiale absolue standard « Montpellier », ou encore « brdx » à l'entité spatiale absolue standard « bordeaux ». Toujours dans un objectif d'identifier automatiquement les nouvelles formes d'expression des entités spatiales absolues, une solution serait de calculer la similarité entre les entités spatiales absolues et les mots du corpus sans considérer les voyelles (par exemple, calculer la similarité entre « brdx » avec l'entité spatiale absolue « bordeaux » sans voyelles) ;

ii *aux mots très similaires considérés comme de nouvelles variantes d'entités spatiales absolues.* Nous pouvons notamment citer l'exemple de « bassin » qui a été associé de manière incorrecte à l'entité spatiale absolue standard « Bassan ». Une première solution pourrait être d'intégrer dans le processus une validation manuelle menée par des experts du territoire d'étude. Nous estimons par ailleurs que la prise en compte du contexte de l'entité spatiale absolue devrait nous permettre de filtrer les résultats obtenus afin d'améliorer la précision ;

iii *aux entités ambiguës qui ont été considérées comme des entités spatiales absolues.* Par exemple, en section 3.1, les messages (SMS 15) et (SMS 16) contiennent l'expression « tahiti » qui ne référence pas toujours une entité spatiale absolue. En effet, nous pouvons remarquer que l'entité nommée « tahiti » dans l'expression « on part à tahiti » représente bien une ESA (cf. (SMS 15)). En revanche, la même EN dans l'expression « gel douche tahiti » représente un objet et non une ESA (cf. (SMS 16)). Un problème similaire est présent avec l'entité « marrakech » dans l'exemple (SMS 17) qui représente un évènement. Cet exemple contient l'information thématique « marrakech du rire » qui est un festival international annuel. L'une des solutions possibles est de désambiguïser les entités nommées en définissant de nouvelles règles de marquage s'appuyant notamment sur des lexiques pour identifier et extraire d'autres types d'entités nommées. La prise en compte du contexte dans les messages courts peut également nous permettre de mener à bien cette étape de désambiguïsation (Tahrat et al., 2013).

Aussi, nous proposons comme perspectives de :

iv *structurer ces ESA sous forme d'un thésaurus.*

v *d'améliorer les outils existants tels que Stanford NER et Polyglot à l'aide de règles apprises à partir du corpus annoté d'ESA (Zenasni et al., 2017a).*

Pour les deux contributions visant à identifier les relations spatiales et sémantiques, nous proposons :

i *d'améliorer l'identification des relations sémantiques.* La plupart des erreurs manuellement identifiées sont liées à la spécificité de l'écriture SMS et à la complexité des relations. L'une de nos propositions pour résoudre ce problème est d'utiliser les k-skip-n-grammes (Guthrie et al., 2006) pour identifier ces relations candidates. En effet, les k-skip-n-grammes nous permettent d'identifier les différentes combinaisons de relations candidates en relâchant des contraintes liées à la présence exacte d'expression.

ii *de structurer sous forme d'un thésaurus la représentation des différents types de relations spatiales et sémantiques présentes dans les corpus de messages courts.* L'idée ici serait de formaliser et de classer dans une ressource unique la liste des relations spatiales et sémantiques identifiées dans les corpus en précisant différentes caractéristiques (synonymie, relations hiérarchiques, langue, forme standard ou variante, etc.). Par exemple, la ressource permettrait d'expliciter le fait que l'expression « dans » a pour synonymes « à l'intérieur de », ou encore « ds », et que cette expression est une relation d'inclusion qui est une relation spatiale de la langue française. Le résultat serait donc un thésaurus, éventuellement multilingue, dédié à la spatialité, réutilisable pour mener de nouvelles actions d'identification et d'extraction d'entités spatiales absolues.

Bibliographie

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216.

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-NER : Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015*, pages 586–594.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot : Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

Al-Shalabi, R., Kanaan, G., and Gharaibeh, M. (2006). Arabic text categorization using knn algorithm. In *the Proc. of Int. multi conf. on computer science and information technology CSIT06*.

Alfred, R., Leong, L. C., On, C. K., and Anthony, P. (2014). Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, 4 :300–306.

Allauzen, A. and Bonneau-Maynard, H. (2008). Training and evaluation of POS Taggers on the French multitag corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3373–3377.

Anis, J. (2001). *Parlez-vous texto ? Guide des nouveaux langages du réseau*. Éditions du Cherche Midi.

Arulanandam, R., Savarimuthu, B. T. R., and Purvis, M. (2014). Extracting crime information from online newspaper articles. In *Second Australasian Web Conference, AWC 2014, Auckland, New Zealand, January 2014*, pages 31–38.

Asch, V. V. (2013). Macro- and micro-averaged evaluation measures. `http://www.clips.uantwerpen.be/~vincent/pdf/microaverage.pdf`.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*. ACM Press, New York.

Bannour, S., Audibert, L., and Nazarenko, A. (2011). Mesures de similarité distributionnelle entre termes. In *IC2011*, pages 523–538.

Béchet, F. and Charton, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5338–5341.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., and Frieder, O. (2007). Varying approaches to topical web query classification. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 783–784.

Berger, A. L. and Mittal, V. O. (2000). Ocelot : A system for summarizing web pages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 144–151.

Bhatia, N. and Vandana (2010). Survey of nearest neighbor techniques. *CoRR*, abs/1007.0085.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble : a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5) :16–23.

Bilhaut, F., Dumoncel, F., Enjalbert, P., and Hernandez, N. (2007). Indexation sémantique et recherche d'information interactive. *CORIA*, 7 :65–76.

Blessing, A. and Schütze, H. (2010). Self-annotation for fine-grained geospatial relation extraction. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 80–88.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase : a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Bollacker, K., Tufts, P., Pierce, T., and Cook, R. (2007). A platform for scalable, collaborative, structured information integration. In *Intl. Workshop on Information Integration on the Web (IIWeb'07)*, pages 22–27.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. pages 757–766.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2009). A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, pages 803–812.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Borhaninejad, S., Hakimpour, F., and Hamzei, E. (2015). Tags extarction from spatial documents in search engines. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40 :111–113.

Borra, E. and Rieder, B. (2014). Programmed method : developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66(3) :262–278.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998). NYU : Description of the MENE named entity system as used in muc-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7) :107–117.

Brin, S. and Page, L. (2012). Reprint of : The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18) :3825–3833.

Bruce, L.-E. (2012). Processing electronic medical records : Ontology-driven information extraction and structuring in the clinical domain. Master's thesis.

Brun, C. and Ehrmann, M. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ester 2. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*.

Btoush, M. H., Alarabeyyat, A., and Olab, I. (2016). Rule based approach for arabic part of speech tagging and name entity recognition. *International Journal of Advanced Computer Science and Applications*, 7(6) :331–335.

Chen, H.-H., Lin, M.-S., and Wei, Y.-C. (2006). Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1009–1016.

Chen, J., Cohn, A. G., Liu, D., Wang, S., Ouyang, J., and Yu, Q. (2015). A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(1) :106–136.

Chinchor, N. and Marsh, E. (1998). Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pages 359–367.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., and Basu, A. (2007). Investigation and modeling of the structure of texting language. *IJDAR*, 10(3-4) :157–174.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1) :22–29.

Cilibrasi, R. L. and Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3) :370–383.

Ciravegna, F. (2001). (LP2), an adaptive algorithm for information extraction from web-related texts. In *In Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*.

Collin, O., Guerraz, A., Hiou, Y., and Voisine, N. (2013). Participation de orange labs à deft 2013. *Actes du neuvième DÉfi Fouille de Textes*, page 65.

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Cooper, R., Ali, S., and Bi, C. (2005). Extracting information from short messages. In *Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Proceedings*, pages 388–391.

Cooper, R. and Manson, S. (2007). Extracting temporal information from short messages. In *Data Management. Data, Data Everywhere, 24th British National Conference on Databases, BNCOD 24, Proceedings*, pages 224–234.

Cougnon, L.-A. and Ledegen, G. (2008). c'est écrire comme je parle. une étude comparatiste de variétés de français dans l'écrit SMS. *Actes du Congrès annuel de l'AFLS*.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université Paris Diderot - Paris 7.

Dalbelo Basic, B., Kolar, M., Snajder, J., and Petrovic, S. (2006). Comparison of collocation extraction measures for document indexing. *CIT. Journal of computing and information technology*, 14(4) :321–327.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all : Overcoming sparse and noisy data. In Angelova, G., Bontcheva, K., and Mitkov, R., editors, *RANLP*, pages 198–206.

Dinarelli, M. and Rosset, S. (2011). Models cascade for tree-structured named entity detection. In *IJCNLP*, pages 1269–1278.

Dini, L., Bittar, A., and Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine deft, taln-recital 2013. *Actes du neuvième DÉfi Fouille de Textes*, page 51.

Dittrich, A., Richter, D., and Lucas, C. (2015). Analysing the usage of spatial prepositions in short messages. In *Progress in Location-Based Services 2014*, pages 153–169.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, pages 837–840.

D'Souza, J. and Ng, V. (2015). Utd : Ensemble-based spatial relation extraction. In *SemEval@ NAACL-HLT*, pages 862–869.

Duchateau, F., Bellahsene, Z., and Roche, M. (2008). Improving quality and performance of schema matching in large scale. *Ingénierie des Systèmes d'Information*, 13(5) :59–82.

Egenhofer, M. J. (1991). Reasoning about binary topological relations. In *Proceedings of the Second International Symposium on Advances in Spatial Databases*, SSD '91, pages 143–160.

Egenhofer, M. J. and Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2) :161–174.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web : An experimental study. *Artif. Intell.*, 165(1) :91–134.

Even, F. (2005). *Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale*. PhD thesis, Université de Nantes.

Ezzat, M. (2014). *Acquisition de relations entre entités nommées à partir de corpus*. PhD thesis, Paris, INALCO.

Fairon, C., Klein, J. R., and Paumier, S. (2006). *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête "Faites don de vos SMS à la science"*. `http://questionsdecommunication.revues.org/272`, UCL Presses Universitaires de Louvain.

Fallery, B. and Rodhain, F. (2007a). Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique. In *XVI ème Conférence de l'Association Internationale de Management Stratégique AIMS*, pages 1–16.

Fallery, B. and Rodhain, F. (2007b). Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique. In *XVI ème Conférence de l'Association Internationale de Management Stratégique AIMS*, pages pp 1–16.

Figuerola, C. G., Zazo Rodríguez, A., and Luis Alonso Berrocal, J. (2001). Automatic vs manual categorisation of documents in spanish. *Journal of Documentation*, 57(6) :763–773.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 363–370.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. pages 168–171, Edmonton. CoNLL 2003.

Frantzi, K. T., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *Int. J. on Digital Libraries*, 3(2) :115–130.

Friburger, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*. PhD thesis, Tours.

Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1) :93–104.

Gaio, M. (2001). Traitements de l'information géographique : Représentations et structures. *Habilitation à diriger des recherches, Université de Caen*.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Geleijnse, G. and Korst, J. (2007). Creating a dead poets society : Extracting a social network of historical persons from the web. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 156–168.

Geleijnse, G., Korst, J., de Boer, V., et al. (2006). Instance classification using co-occurrences on the web. In *Proceedings of the ISWC 2006 workshop on Web Content Mining (WebConMine)*.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6) :721–741.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and A.Smith, N. (2011). Part-of-speech tagging for twitter : Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, pages 42–47.

Gotti, F. and Lapalme, G. (2014). Zodiac : Insertion automatique des signes diacritiques du français. In *Traitement Automatique des Langues Naturelles, TALN 2014, Marseille, France, 1-4 Juillet 2014, Démonstrations*, pages 19–20.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6 : A brief history. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 466–471.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4.

Hahn, U., Buyko, E., Landefeld, R., Mühlhausen, M., Poprat, M., Tomanek, K., and Wermter, J. (2008). An overview of jcore, the julie lab uima component repository. In *Proceedings of the LREC*, volume 8, pages 1–7.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages : Makn sens a #twitter. In *The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 368–378.

Han, B., Cook, P., and Baldwin, T. (2013). Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1) :5 :1–5 :27.

Hatmi, M., Jacquin, C., Morin, E., and Meignier, S. (2013). Named entity recognition in speech transcripts following an extended taxonomy. In *SLAM@ INTERSPEECH*, pages 61–65.

He, Z., Hong, J., and Bell, D. (2008). Schema matching across query interfaces on the deep web. In *British National Conference on Databases*, pages 51–62.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations : Recent Achievements and Future Directions*, DEW '09, pages 94–99.

Hill, L. L. (2000). Core elements of digital gazetteers : placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pages 280–290.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes : the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume : Short Papers*, pages 57–60.

Isozaki, H. and Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 :241–272.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406) :414–420.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Joachims, T. (1999). Svmlight : Support vector machine. *SVM-Light Support Vector Machine http ://svmlight. joachims. org/, University of Dortmund*, 19(4).

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., and Weibel, R. (2002). Spatial information retrieval and geographical ontologies an overview of the spirit project. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–388.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition : A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2) :259–289.

Kent, A., Berry, M. M., Luehrs, F. U., and Perry, J. W. (1955). Machine literature searching viii. operational criteria for designing information retrieval systems. *Journal of the Association for Information Science and Technology*, 6(2) :93–101.

Kergosien, E., Sallabery, C., Bessagnet, M.-N., Le Parc-Lacayrelle, A., and Chaudiron, S. (2017). Using a GIR tool in a Business Intelligence Context : the case of EGC conferences. In *Proceedings of the 7th International Conference on Information Systems and Economic Intelligence* .

Kim, J.-H., Kang, I.-H., and Choi, K.-S. (2002). Unsupervised named entity classification models and their ensembles. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7.

Kobus, C., Yvon, F., and Damnati, G. (2008). Normalizing SMS : are two metaphors better than one ? In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 441–448.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial role labeling : Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3) :4.

Kramdi, S. E., Haemmerlé, O., and Hernandez, N. (2009). Approche générique pour l'extraction de relations à partir de textes. In *Journées Francophones d'Ingénierie des Connaissances*, pages 97–108.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

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

Lesbegueries, J. and Loustau, P. (2006). Structuration d'information spatiale qualitative pour la recherche d'information. *Représentation et raisonnement sur le temps et l'espace (RTE 2006)*, 1 :4.

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

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Li, C. and Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 43–52.

Liao, Z. and Wu, H. (2012). Biomedical named entity recognition based on skip-chain crfs. In *Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on*, pages 1495–1498.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the Fifteenth Int. Conf. on Machine Learning (ICML)*, pages 296–304.

Ling, X. and Weld, D. S. (2012). Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 94–100.

Lingad, J., Karimi, S., and Yin, J. (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web*, pages 1017–1020.

Lison, P. and Kutuzov, A. (2017). Redefining context windows for word embedding models : An experimental study. *CoRR*, abs/1704.05781.

Liu, F., Vasardani, M., and Baldwin, T. (2014). Automatic identification of locative expressions from social media text : A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web*, LocWeb '14, pages 9–16.

Liu, X., Wei, F., Zhang, S., and Zhou, M. (2013). Named entity recognition for tweets. *ACM TIST*, 4(1) :3.

Liu, Y., Guo, Q., and Kelly, M. (2008). A framework of region-based spatial relations for non-overlapping features and its application in object based image analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(4) :461–475.

Loglisci, C., Ienco, D., Roche, M., Teisseire, M., and Malerba, D. (2012). Toward geographic information harvesting : Extraction of spatial relational facts from web documents. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pages 789–796.

Lopez, C., Partalas, I., Balikas, G., Derbas, N., Martin, A., Reutenauer, C., Segond, F., and Amini, M.-R. (2017). Cap 2017 challenge : Twitter named entity recognition. *CoRR*, abs/1707.07568.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016). Biomedical term extraction : overview and a new methodology. *Information Retrieval Journal*, 19(1-2) :59–99.

Luo, G., Huang, X., Lin, C.-Y., and Nie, Z. (2015). Joint named entity recognition and disambiguation. In *Proc. EMNLP*, pages 879–880.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Int. Conf. EKAW*, pages 251–263.

Malouf, R. (2002). Markov models for language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml : Annotation scheme, corpora, and tools. In *LREC*.

Mansouri, A., Affendey, L. S., and Mamat, A. (2008). Named entity recognition approaches. pages 339–344.

Martineau, C., Tolone, E., and Voyatzi, S. (2007). Les Entités Nommées : usage et degrés de précision et de désambiguïsation. In *26ème Colloque international sur le Lexique et la Grammaire (LGC'07)*, pages 105–112, Bonifacio, France.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I., and Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1) :69–96.

Maurel, D., Friburger, N., and Eshkol, I. (2009). Who are you, you who speak ? Transducer cascades for information retrieval. In *4th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 220–223.

Maynard, D. (2003). Multi-source and multilingual information extraction. *Expert Update*, 6(3) :11–16.

Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Moncla, L., Gaio, M., and Mustiere, S. (2014). Automatic itinerary reconstruction from texts. In *International Conference on Geographic Information Science*, pages 253–267.

Monteiro, D. M. et al. (2015). *A proposal for an architecture to extract information from SMS messages during emergency situations*. PhD thesis.

Munro, R. and Manning, C. D. (2012). Accurate unsupervised joint named-entity extraction from unaligned parallel text. In *Proceedings of the 4th Named Entity Workshop*, NEWS '12, pages 21–29.

Nadeau, D., Turney, P. D., and Matwin, S. (2006). Unsupervised named-entity recognition : Generating gazetteers and resolving ambiguity. In *Proceedings of the 19th International Conference on Advances in Artificial Intelligence : Canadian Society for Computational Studies of Intelligence*, AI'06, pages 266–277.

Nagesh, A., Ramakrishnan, G., Chiticariu, L., Krishnamurthy, R., Dharkar, A., and Bhattacharyya, P. (2012). Towards efficient named-entity rule induction for customizability. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 128–138.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comp. Surv.*, pages 31–88.

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Nguyen, V. T. (2012). *Méthode d'extraction d'informations géographiques à des fins d'enrichissement d'une ontologie de domaine*. PhD thesis, Pau.

Nguyen, V. T., Gaio, M., and Sallaberry, C. (2010). Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. *CoRR*.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9.

Ogren, P. V., Wetzler, P. G., and Bethard, S. (2008). Cleartk : A uima toolkit for statistical natural language processing. *Towards Enhanced Interoperability for Large HLT Systems : UIMA for NLP*, 32.

Oliva, J., Serrano, J. I., Del Castillo, M. D., and Iglesias, Á. (2011). SMS normalization : combining phonetics, morphology and semantics. In *Conference of the Spanish Association for Artificial Intelligence*, pages 273–282.

Panckhurst, R. (2009). Short Message Service (SMS) : typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, pages 33–52. Université Paul-Valéry Montpellier 3.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. (2013). Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. *Epistémé*, 9 :107–138.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. (2014). 88milSMS. a corpus of authentic text messages in french. produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo.

Park, Y. and Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing*, pages 126–133.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C., and Roche, M. (2013). Approaches of anonymisation of an SMS corpus. In *CICLing : Conference on Intelligent Text Processing and Computational Linguistics*, number 7816, pages 77–88.

Paumier, S. (2003). *De la reconnaissance des formes linguistiques à l'analyse syntaxique*. PhD thesis, Université de Marne-la-Vallée.

Plu, J., Rizzo, G., and Troncy, R. (2015). A hybrid approach for entity recognition and linking. In *Semantic Web Evaluation Challenge*, pages 28–39.

Pooja Pandey, Dhiraj Amin, S. G. (2016). Rule based stemmer using marathi wordnet for marathi language. *International Journal of Advanced Research in Computer and Communication Engineering*, 5 :278–282.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). Semeval-2015 task 8 : Spaceeval. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 884–894.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda : A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1*, pages 248–256.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155.

Raymond, C. (2013). Robust tree-structured named entities recognition from speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8475–8479.

Raymond, C. and Fayolle, J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Conférence Traitement automatique des langues naturelles, TALN'10*.

Reul, C., Köberle, P., Üçeyler, N., and Puppe, F. (2016). Expectation-driven text extraction from medical ultrasound images. In *MIE*, pages 712–716.

Rikitianskii, A., Harvey, M., and Crestani, F. (2014). A personalised recommendation system for context-aware suggestions. In *ECIR*, pages 63–74.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets : An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1524–1534.

Rizzo, G. and Troncy, R. (2011). NERD : evaluating named entity recognition tools in the web of data. In *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*.

Roberts, K., Skinner, M. A., and Harabagiu, S. M. (2013). Recognizing spatial containment relations between event mentions. *IWCS*, pages 216–227.

Roche, M. (2011). *Fouille de Textes : de l'extraction des descripteurs linguistiques à leur induction*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc.

Roche, M. (2012). Fonctions de rang et fouille du web pour l'identification et la catégorisation d'entités nommées. In *JADT'2012 : 11ièmes Journées internationales d'analyse statistique des données textuelles*, pages 859–870.

Roche, M. and Kodratoff, Y. (2009). Text and web mining approaches in order to build specialized ontologies. *Journal of Digital Information*, 10(4) :1–6.

Roche, M. and Prince, V. (2007). Acrodef : A quality measure for discriminating expansions of ambiguous acronyms. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 411–424.

Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386.

Salas, H. A., Kergosien, E., Roche, M., and Teisseire, M. (2014). Animitex project : Image analysis based on textual information. In *SIMBig : Symposium on Information Management and Big Data*, pages 49–52.

Sallaberry, C. (2013). *Geographical Information Retrieval in Textual Corpora*. FOCUS - Geographical Information Systems Series. Wiley-ISTE.

Sallaberry, C., Baziz, M., Lesbegueries, J., and Gaio, M. (2007). Une approche d'extraction et de recherche d'information spatiale dans les documents textuels-évaluation. In *CORIA*, pages 53–64.

Sallaberry, C., Royer, A., Loustau, P., Gaio, M., and Joliveau, T. (2009). Geostream : Spatial information indexing within textual documents supported by a dynamically parameterized web service. In *OGRS 2009 : International Opensource Geospatial Research Symposium*, page 14.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5) :513–523.

Saneifar, H., Bonniol, S., Laurent, A., Poncelet, P., and Roche, M. (2009). Processus d'extraction et de validation de la terminologie issue de logs. In *JFO'09 : 3èmes Journées Francophones sur les Ontologies*, pages 1–10.

Savary, A. and Piskorski, J. (2010). Lexicons and grammars for named entity annotation in the national corpus of polish. *Intelligent Information Systems, Siedlce, Poland*, pages 141–154.

Savoy, J. (1999). A stemming procedure and stopword list for general french corpora. *JASIS*, 50(10) :944–952.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester.

Sehgal, V., Getoor, L., and Viechnicki, P. D. (2006). Entity resolution in geospatial data integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 83–90.

Serrano, L. (2014). *Vers une capitalisation des connaissances orientée utilisateur : extraction et structuration automatiques de l'information issue de sources ouvertes*. PhD thesis, Universté de Caen.

Severo, M., Giraud, T., and Pecout, H. (2015). Twitter data for urban policy making : an analysis on four european cities. In *Handbook of Twitter for Research*, pages 132–155.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27 :623–656.

Shen, D., Sun, J.-T., Yang, Q., and Chen, Z. (2006). Building bridges for web query classification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 131–138.

Sileo, D., Pradel, C., Muller, P., and Van de Cruys, T. (2017). Synapse at cap 2017 ner challenge : Fasttext crf. *CoRR*, abs/1709.04820.

Simard, M. and Deslauriers, A. (2001). Real-time automatic insertion of accents in french text. *Natural Language Engineering*, 7(2) :143–165.

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science*, 241(1) :1–31.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational linguistics*, 22(1) :1–38.

Smail, N. (2009). *Contribution à l'analyse et à la recherche d'information en texte intégral : application de la transformée en ondelettes pour la recherche et l'analyse de textes*. PhD thesis, Université Paris-Est.

Smith, D. A. and Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pages 45–49, Stroudsburg, PA, USA.

Song, Y., Huang, J., Councill, I. G., Li, J., and Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 342–351.

Stern, R. and Sagot, B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.

Stoilos, G., Stamou, G., and Kollias, S. (2005). A string metric for ontology alignment. *The Semantic Web–ISWC 2005*, pages 624–637.

Tahrat, S., Kergosien, E., Bringay, S., Roche, M., and Teisseire, M. (2013). Text2geo : des données textuelles aux informations géospatiales. In *EGC : Extraction et Gestion des Connaissances*, volume 13, pages 407–412.

Tarrade, L. and Lopez, C. (2017). Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français "non standard". In *Actes TALN'2017*, pages 27–34.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. CoNLL 2003.

Tkachenko, M. and Simanovsky, A. (2012). Named entity recognition : Exploring features. In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, pages 118–127.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora : held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70.

Turney, P. D. (2001). Mining the web for synonyms : Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502.

Turpin, A., Tsegay, Y., Hawking, D., and Williams, H. E. (2007). Fast generation of result snippets in web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 127–134.

Usery, E. L. (1996). A feature-based geographic information system model. *Photogrammetric Engineering and Remote Sensing*, 62(7) :833–838.

Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, 6 :111–122.

Varadarajan, R. and Hristidis, V. (2006). A system for query-specific document summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 622–631.

Vergne, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. 2 :1158–1164.

Visser, U., Vögele, T., and Schlieder, C. (2002). Spatio-terminological information retrieval using the buster system. pages 93–100.

Wakao, T., Gaizauskas, R., and Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 418–423.

Widlöcher, A. and Bilhaut, F. (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN'05)*, pages 517–522.

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*.

Wu, H., Radev, D. R., and Fan, W. (2004). Towards answer-focused summarization using search engines. *Chapter 18*.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1) :1–37.

Wu, Y.-C., Fan, T.-K., Lee, Y.-S., and Yen, S.-J. (2006). Extracting named entities using support vector machines. In *Proceedings of the 2006 International Conference on Knowledge Discovery in Life Science Literature*, KDLL'06, pages 91–103.

Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., and Jin, Z. (2016a). Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv :1601.03651*.

Xu, Z., Xuan, J., Liu, Y., Choo, K.-K. R., Mei, L., and Hu, C. (2016b). Building spatial temporal relation graph of concepts pair using web repository. *Information Systems Frontiers*, pages 1–10.

Yang, T.-I., Torget, A. J., and Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49.

Yu, H., Hripcsak, G., and Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *JAMIA*, 9(3) :262–272.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2015). Discovering types of spatial relations with a text mining approach. In *Foundations of Intelligent Systems - 22nd International Symposium, ISMIS Proceedings*, pages 442–451, Lyon.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2016a). Découverte de nouvelles entités et relations spatiales à partir d'un corpus de SMS. In *23ème Conférence sur le Traitement Automatique des Langues Naturelles TALN 2016*, pages 403–410.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2016b). Extracting new spatial entities and relations from short messages. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems MEDES'16*, pages 189–196.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2017a). A corpus of 1000 authentic SMS in French with spatial labels. [Dataset]. doi :10.18167/DVN1/0ZGJRC, CIRAD Dataverse.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2017b). Dic-ES : Liste d'entités spatiales en français. [Dataset]. doi :10.18167/DVN1/LPY080, CIRAD Dataverse.

Zenasni, S., Kergosien, E., Roche, M., and Teisseire, M. (2018). Spatial information extraction from short messages. *Expert Systems With Applications*, 95 :351– 367.

Zhang, T. and Johnson, D. (2003). A robust risk minimization based named entity recognition system. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 204–207.

Chapitre

7

Annexe

.1 Extrait du corpus Midi Libre

(1) ... À **la sortie du village**, avant le pont, prenez le chemin à droite **en direction de la D5**. Au prochain croisement, tournez à gauche, poursuivez tout droit sur 1 km, puis à gauche toute **vers Vignogoul** pour une arrivée grandiose sur l'abbaye entrevue depuis le vignoble... 04 67 47 72 88 ...

(2) ... Jean-Marius Chanet, 56 ans, ferailleur **à Poussan** (Hérault) a été mis en examen hier pour meurtre et laissé en liberté sous contrôle judiciaire avec interdiction de résider **à Poussan**, à la suite du drame survenu mardi dans ce village situé **au nord du bassin de Thau**, au cours duquel ...

(3) ... Archéologie. Lun, mer, jeu, ven 10 h-12 h, 13 h 30- 17 h 30 ; sam, dim 14 h-18 h. Musée Henri-Prades, **route de Pérols**, Lattes. 2,50 et 2. 04 67 99 77 20 ...

(4) ... Départ Cruciera de Manolo, **route de Candillagues**, arrivée aux Arènes, boulevard Jean-Macé ...

(5) ... La déchetterie de Mèze, située **route de Villeveyrac**, est ouverte du lundi au samedi ...

(6) ... **Du côté de Bouzigues**, les autos se sont agglutinées jusqu'au rond-point où on les faisait repartir en sens inverse. Et **du côté de Poussan** le trafic s'est accumulé sans qu'un nouvel itinéraire ne soit proposé. Sidérant.Philippe MALRIC ...

(7) ... peut confirmer sa belle prestation chez l'AOC en recevant le bon dernier de la classe, **Villenueve-lès-Béziers** et ainsi se replacer dans la course à la qualification ...

(8) ... candidat du Front de gauche aux cantonales sur Montpellier IX (**La Paillade**) ...

(9) ... la canalisation d'eau potable longeant les marais de **la Grande Palude** et rejoignant le chemin ...

(10) ... les joueurs visiteront notamment **Agde**, **Le Cap d'Agde** et **Le Grau d'Agde**. Fabriquée par la maison d'édition officielle du Monopoly ...

(11) ... Dimanche, à **Gigea**... RELIGION **Gigean**, **Montbazin** et **Poussan** Paroisse du Bon Pasteur ...

(12) .. 18 h-19 h 30. Fermé lundi 13 août. Rue **Teyron**. 04 67 70 64 10.

Abstract

The extraction of spatial information from textual data has become an important research topic in the field of Natural Language Processing (NLP). It meets a crucial need in the information society, in particular, to improve the efficiency of Information Retrieval (IR) systems for different applications (tourism, spatial planning, opinion analysis, etc.). Such systems require a detailed analysis of the spatial information contained in the available textual data (web pages, e-mails, tweets, SMS, etc.). However, the multitude and the variety of these data, as well as the regular emergence of new forms of writing, make difficult the automatic extraction of information from such corpora.

To meet these challenges, we propose, in this thesis, new text mining approaches allowing the automatic identification of variants of spatial entities and relations from textual data of the mediated communication. These approaches are based on three main contributions that provide intelligent navigation methods. Our first contribution focuses on the problem of recognition and identification of spatial entities from short messages corpora (SMS, tweets) characterized by weakly standardized modes of writing. The second contribution is dedicated to the identification of new forms/variants of spatial relations from these specific corpora. Finally, the third contribution concerns the identification of the semantic relations associated with the textual spatial information.