Académie de Montpellier
Université Montpellier II
Sciences et Techniques du Languedoc

MÉMOIRE DE STAGE DE MASTER 2

effectué au Laboratoire d'Informatique de Robotique
et de Microélectronique de Montpellier

Spécialité : **DECOL**

Découverte de nouvelles relations spatiales par des méthodes de fouille de textes

Date de soutenance: 27/06/2014

par **Wissame LADDADA**

Sous la direction de

Sandra Bringay, Nicolas Béchet,
Mathieu Roche, Hugo Alatrista Salas

Remerciements

Je tiens à remercier en premier lieu mes encadrants Nicolas Béchet et Hugo Alatrista Salas qui m'ont bien encadré, écouté, aidé et encouragé durant les 5 mois de mon stage en étant toujours dans la bonne humeur. Je les remercie également pour leurs exigences qui m'ont permis de maintenir et de surpasser les limites de ma persévérance pour élaborer ce travail que j'espère sera à la hauteur de leurs attentes. Mes remerciements vont également à Mathieu Roche et Sandra Brigay dont les conseils et orientations m'ont beaucoup apportée.

Je souhaiterais remercier chaque membre du projet ANIMITEX, à commencer par ma reconnaissance envers la patience de Mauro Gaio sans qui une partie cruciale du travail n'aurait pas pu être aussi bien accomplie. Sans oublier bien évidement Bruno Crémilleux et Thierry Charnois dont les remarques et questions m'ont été d'une grande utilité.

Je remercie aussi Isabelle Mougenot et Maguelonne Teisseire pour le temps accordé à la lecture du rapport et dont les remarques me seront bien utiles tôt ou tard.

Enfin, je remercie toute l'équipe TATOO dont l'ambiance et l'esprit d'équipe a rendu mon stage agréable.

Mes derniers remerciements et respects reviennent à ma famille. A mon oncle qui m'a fait profiter de son expérience en recherche et de son compte de bibliothèque numérique. A mes soeurs et à mon frère qui m'ont toujours soutenue et enfin à mes *parents*...

Résumé

Le projet ANIMITEX (ANalyse d'IMages fondée sur des Informations TEXtuelles) a été proposé pour venir compléter l'analyse des images satellitaires. En effet, les moyens de télédétection ne permettent pas d'identifier tout type d'informations comme par exemple le type de culture d'un champs donné. Pour ce faire, l'idée du projet vise à mettre en relation les moyens de la télédétection et la fouille de textes. Dans ce contexte, les travaux de mon stage s'inscrivent dans la discipline d'Extraction d'Informations et plus précisément l'Extraction de Relations Spatiales à partir de données textuelles. Afin d'atteindre ce but, l'idée proposée est d'utiliser des approches de fouilles de données pour des finalités TALN (Traitement Automatique du Langage Naturel). En effet, en s'appuyant sur l'extraction de motifs séquentiels sous contraintes multiples, les résultats obtenus sont des patrons linguistiques relatifs aux relations spatiales.

Mots clés : Extraction d'Informations, Extraction de Relations Spatiales, la fouille de motifs séquentiels.

Abstract

The ANIMITEX project was proposed to complete the satellite image analysis. In fact, tasks using in remote sensing process do not allow the identification of a particular information as the type of specific crop field. To do this, the project's idea aims to put in place a relationship between remote sensing process and text mining process. In this research intership, we focus on Informations Extraction task, more precisely on Extraction of Spatial Relationships from textual data. To achieve this goal, we propose to use Data Mining approaches for NLP (Natural Language Processing) resolutions. Indeed, based on the Sequential Patterns Mining under multiple constraints, the obtained results are linguistic patterns related to Spatial Relationships.

Keyswords: Information Extraction, Extraction of Spatial Relationships, Sequential Patterns Mining.

Table des matières

1 Introduction **1**

2 Extraction d'Informations dans les textes **2**

3 Extraction de Relations **3**
3.1 Méthodes basées sur le TALN et leurs limites 4
3.2 Méthodes basées sur la fouille de données 5
3.2.1 Approches supervisées 5
3.2.2 Approches semi-supervisées 8
3.2.3 Approches non-supervisées 10
3.3 Discussion 12

4 Extraction de relations spatiales **15**
4.1 Problématique 15
4.1.1 Définitions 15
4.2 Méthodologie 16
4.2.1 Extraction des Entités Spatiales 16
4.2.2 Prétraitements 19
4.2.3 Extraction des motifs séquentiels 19

5 Expérimentations **26**
5.1 Exploration des motifs 26
5.2 Corpus : Midi-Libre 26
5.2.1 Tests et résultats 26
5.2.2 Discussion 28
5.2.3 Évaluation 29
5.3 Corpus : Randonnées 31
5.3.1 Tests et résultats 31
5.3.2 Évaluation 31
5.4 Corpus Wikipédia 33
5.4.1 Tests et Résultats 34
5.4.2 Évaluation 34
5.5 Impact des contraintes 35
5.5.1 Le Gap 35
5.5.2 La contrainte d'appartenance 36
5.5.3 Proposition d'une nouvelle contrainte : la pondération 37
5.6 Le lien avec ANIMITEX 37

6 Conclusion et perspectives **38**

Table des figures

1 Le lien entre EI et ER 2
2 Schéma global des méthodes supervisées existantes 5
3 Schéma global du principe des méthodes semi-supervisées 8
4 Schéma global des méthodes non-supervisées existantes 10
5 Schéma global de la démarche 16
6 Schéma global résumant la Reconnaissance des Entités Spatiales 17
7 Exemple de résultats de la chaîne de traitement pour la RES 18
8 Résultats des premiers travaux 22
9 Intérêt de la généralisation 23
10 L'outils Camelis 27
11 Extrait du corpus Randonnées 30
12 Extrait du corpus Wikipedia 33

Liste des tableaux

1 Récapitulatif des méthodes d'Extraction de Relations 13
2 Exemples de SDB 20
3 Lexique pouvant introduire des relations 23
4 Présence de syntagme exprimant une relation avant l'ES 24
5 Présence de syntagmes exprimant une relation après l'ES 25
6 Exemples de SDB dans les relations spatiales 25
7 Premiers résultats sur un corpus bruité 27
8 Exemples de patrons extraits du Midi Libre 28
9 Évaluation des premiers patrons 29
10 Exemples de patrons extraits du corpus Randonnées 31
11 Évaluation des patrons extraits du corpus Randonnées sur le corpus Midi Libre 32
12 Exemples de patrons extraits du corpus Wikipédia 34
13 Évaluation des patrons extraits de Wipedia sur le corpus Midi Libre 35
14 Impact de la contrainte de Gap 36
15 Impact de la contrainte d'appartenance facultative 37

1 Introduction

L'utilisation des technologies informatiques dans différents domaines a permis de produire une importante masse de données numériques, notamment les informations textuelles qui ne cessent de croître (articles de presse, blogs, etc.). Par conséquent, les données présentes de nos jours sont hétérogènes et non structurées. Leur exploitation devient de ce fait un défi pour les experts du traitement du langage. Au fil des années, cette exploitation a attiré plus d'un domaine : question answering [11][26], biomédical [6][8], construction ou alimentation d'ontologies [7][3] ou encore le business intelligence [28], ceci a engendré la discipline d'Extraction d'Informations (EI) qui a pour objectif d'extraire des détails essentiels pour un domaine particulier à partir de données textuelles.

L'Extraction d'Informations, à ne pas confondre avec la Recherche d'Informations, est un processus du Traitement Automatique du Langage Naturel (TALN) qui a induit d'autres domaines dans le milieu de la recherche dont les plus connus sont la Reconnaissance d'Entités Nommées (REN) et l'Extraction de Relations (ER) à partir de données textuelles. La Reconnaissance d'Entités Nommées vise à identifier des termes précis pour faciliter l'Extraction de Relations. Cette dernière se définit comme le fait de découvrir un lien sémantique entre entités, ce qui permettra d'extraire une information pertinente dans un domaine spécifique, par exemple les relations entre protéines, personnes, entités spatiales ou entre deux types d'entités différentes comme entre gènes et maladies.

Par ailleurs, une fois les informations extraites, nous pouvons nous poser la question : à quelle finalité pouvons nous les exploiter ? Dans ce contexte et comme réponse à cette question, les données textuelles peuvent venir enrichir et compléter d'autres applications. Le projet ANIMITEX (ANalyse d'IMages fondée sur des Informations TEXtuelles), sur lequel portent les travaux de ce stage, s'inscrit justement dans le cadre de cette perspective. Il a pour objectif de compléter l'analyse des images satellites qui devient de plus en plus ardue en raison de la répétitivité temporelle croissante qui amplifie la masse de ces images. Cet enrichissement s'appuie sur différentes données textuelles existantes et a pour contribution d'atténuer l'effort humain considérable consacré à l'amélioration de cette analyse. Pour ce faire, le projet fait appel aux méthodes d'EI à partir de textes qui représentent des données liées aux domaines pouvant apporter une analyse complémentaire aux images satellites. Les éléments textuels à découvrir sont de ce fait de nature spatiale. En d'autres termes, il est question d'extraire des relations pertinentes entre Entités Nommées qui font référence à un lieu combiné à des indicateurs spatiaux tels que l'orientation ou la distance, par exemple la phrase *"Montpellier est à 37 Km de Sète"* peut être parmi les relations à extraire, où *Montpellier* et *Sète* sont deux entités spatiales reliées par *est à 37 Km*.

Le but de ce stage est donc de découvrir les relations spatiales présentes dans des données textuelles avec une intervention humaine minimale. Dans la démarche faite en vue d'atteindre cet objectif, le travail est scindé en plusieurs parties. La première est consacrée à une veille exhaustive pour les diverses approches adoptées dans l'Extraction de Relations à partir de documents textuels. Nous aborderons dans cette partie le concept général de l'Extraction d'Informations (Section 2) ainsi que celle des relations (Section 3). Nous consacrerons également une sous partie pour les différents types de méthodes ayant traité cette problématique d'ER entre Entités Nommées. Ensuite nous passerons à la seconde partie (Section 4) pour développer l'approche adoptée au sein de ces travaux tout en évoquant les notions de bases relatives à la méthode appliquée. Cette dernière se

base comme nous allons le détailler sur la notion de l'extraction des motifs séquentiels. Dans la présente partie, nous analyserons également les Relations Spatiales d'un point de vue morphosyntaxique. De là, nous rapporterons les résultats obtenus, accompagnés d'une discussion suite à une évaluation (Section 5). Enfin, la dernière partie (Section 6) fera objet de conclusion et de possibles perspectives qui peuvent améliorer notre approche.

2 Extraction d'Informations dans les textes

Ces dernières années, l'Extraction d'Informations à partir de documents écrits en langage naturel a suscité de forts intérêts. Elle offre l'avantage de diminuer l'intervention humaine que peut exiger l'extraction d'un détail utile pour un utilisateur afin de résoudre un problème donné, notamment dans le cas où les données sont d'une taille importante, autrement, la tâche coûterait en temps et en effort. L'Extraction d'Informations vise à prélever automatiquement des propriétés sémantiques à partir de données textuelles pas nécessairement structurées. Généralement, il s'agit dans un premier temps de repérer les entités potentiellement intéressantes pour les besoins demandés. Par exemple dans la phrase : *Roger Federer est nommé ambassadeur de l'UNICEF*, le système identifie *Roger Federer* comme étant une entité de type PERSON et *UNICEF* de type ORGANIZATION. Cependant, d'autres types d'extractions existent comme dans le web sémantique où il est question de repérer par exemple des relations de type hyponyme [34]. Dans nos travaux, nous nous intéressons particulièrement aux extractions entre Entités Nommées.

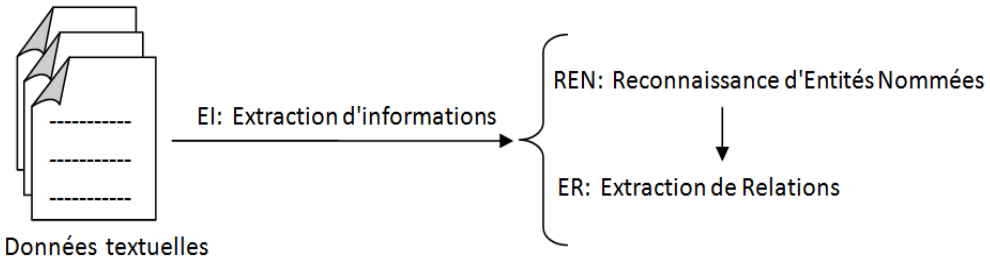

FIGURE 1 – Le lien entre EI et ER

Par conséquent, l'identification des Entités Nommées fait partie du processus d'Extraction d'Informations (cf. Figure 1). L'objectif ciblé est de reconnaître des termes qui font référence à des catégories comme *localisation*, *personne*, *organisation*, *nom de protéine*, ou *entité spatiale* comme dans le cas qui nous intéresse. Ces entités peuvent être identifiées de différentes manières. Dans Michelakis et al. [27], les auteurs utilisent une méthode basée sur des règles pour reconnaître les entités *personne* et *le numéro de téléphone*. Dans d'autres travaux, nous trouvons également des méthodes statistiques basées sur le Conditional Random Field (CRF) comme dans les travaux de Zhang et al. [38]. D'autres exploitent des ressources externes comme YAGO2, GeoNames ou FreeBase dont les descriptions sont comme suit :

FreeBase[1] [10] : Est un projet lancé en 2007 qui consiste à créer une base de connaissances où les données sont structurées via le Crowdsourcing. FreeBase est l'une des premières applications du web sémantique. Les informations qui la composent,

1. FreeBase : `http://www.freebase.com/`

portent sur des célébrités, des places et d'autres types de domaines. Elle contient à ce jour, plus de 40 mille thèmes, plus de 2 millions faits et environ 43 millions d'entités.

YAGO2[2] [22] : Est une extension de la base de connaissances YAGO (Yet Another Great Ontology) avec une dimension spatio-temporelle provenant de Wikipédia et de Wordnet. Elle tire aussi profit des informations de GeoNames. Actuellement, la base contient plus de 10 millions d'entités et près de 120 millions de faits les reliant.

GeoNames[3] : Est une base de données géographiques qui contient pas moins de 10 millions de noms de places dans le monde avec diverses informations les concernant. Des utilisateurs peuvent contribuer à corriger les données ou encore à en rajouter.

Les différentes recherches menées pour ce sous problème de la Reconnaissance d'Entités Nommées, nous conduit à conclure que plusieurs Entités Nommées ont été définies selon les besoins du problème à résoudre : PERSON, LOCATION, DATE, GEOGRAPHICAL ENTITY, ORGANIZATION, PROTEIN NAME, DISEASE NAME, etc. Une fois les entités identifiées, il ne restera plus qu'à déterminer le lien les unissant (cf. Figure 1), autrement dit, la relation existante entre deux entités.

3 Extraction de Relations

L'Extraction de Relations (ER) est un sous domaine du processus d'Extraction d'Informations à partir des données textuelles. Elle vise à extraire des relations entre deux arguments souvent de type Entités Nommées. Son adaptation a été étendue dans plusieurs domaines traitant du langage naturel. Compte tenu des divers secteurs auxquels l'ER peut être appliquée, différents types de relations sont apparus au fil du temps selon le besoin et le domaine traité. Les plus rependues sont les relations définies dans ACE, FreeBase ou bien les relations géographiques ou entre protéines. Leurs descriptions est comme suit :

Les relations d'ACE : Acronyme de Automatic Content Extraction, c'est un programme ayant pour objectif d'identifier certaines entités ainsi que la relation les liant. De ce fait, plusieurs relations sont identifiées : *Role*, *Part*, *Located*, *Near*, et *Social*, un sous type est attribué à chaque type, par exemple, la relation Role relie l'entité PERSON avec ORGANIZATION et a un sous type qui peut être *Founder*, *Client*, *Affiliation*, etc. Il en est de même pour les autres types. Dans l'ensemble, 24 types et sous-types ont été définis.

Les relations de FreeBase : Dans la base FreeBase, certaines relations ont une similarité avec les relations existantes dans ACE. Par exemple, la relation $PERSON/place-of-birth$ peut relier les entités PERSON et LOCATION. Cependant, d'autres relations sont définies comme les relations géographiques tel que *geography/river/mouth*. Par exemple la phrase *l'embouchure (mouth) de La Loire est l'océan Atlantique* (qui peut se traduire par *La Loire se jette dans l'océan Atlantique*) reflète ce type de relations. Cette terminologie pourrait être intéressante pour découvrir des relations spatiales plus spécifiques telle qu'une relation entre rivière et océan.

2. YAGO2 : `http://www.mpi-inf.mpg.de/yago-naga/yago/`
3. GeoNames : `http://geonames.org`

D'autres relations existent par exemple dans le domaine médical. Un grand nombre de ces relations sont référencées dans l'ontologie UMLS (Unified Medical Language System)[9]. Elle englobe les relations entre maladies et protéines et d'autres relations spécifiques au domaine.

Pour effectuer l'Extraction de Relations, différentes méthodes ont été suggérées. Cependant, nous pouvons distinguer deux types d'approches, la première est basée sur des patrons qui relèvent des méthodes du TALN, quand à la deuxième, elle s'appuie sur les méthodes dérivées de la fouille de données venues pallier les limites du premier type d'approches. L'état de l'art et la discussion à propos de ces méthodes feront l'objet de ce qui va suivre.

3.1 Méthodes basées sur le TALN et leurs limites

Les méthodes basées uniquement sur du TALN font partie des approches à avoir été proposées pour l'Extraction de Relations entre entités. Parmi ces relations, nous pouvons trouver les interactions entre protéines ainsi que les relations spatiales. Les méthodes basées sur le TALN se basent sur l'élaboration manuelle de règles (patrons morpho-syntaxiques) afin d'extraire l'information désirée. S'inspirant de cette méthodologie, Blaschke et al. [8] utilisent des patrons générés manuellement dont le but est d'extraire des relations entre protéines. Les auteurs se reposent sur un modèle qui fait correspondre les patrons construits sur un corpus spécifique au domaine tout en utilisant un dictionnaire d'Entités Nommées caractérisant le nom des protéines. Ces dernières sont fixées avant d'entamer la procédure d'extraction. Les interactions ne sont pas validées sur le simple fait d'une correspondance avec le patron. La négation et la fréquence de la relation détectée sont également prises en considération. Les patrons définis sont une suite de séquences basés sur la morphologie et la syntaxe d'une expression décrivant une interaction donnée, ceci en s'appuyant sur l'analyse syntaxique et les étiquetages Part-Of-Speech (POS-tag). Plusieurs patrons sont définis dans l'approche. Par exemple *[proteins] (0-5) [verbs] (6-10) [proteins]* où la parenthèse (0-5) signifie qu'il peut y avoir de 0 à 5 termes entre *[proteins]* et *[verbs]*. Citons également *[proteins] (*) not (0-3) [verbs] (*) by (*) [proteins]*, où nous remarquons que la négation a été prise en considération. Les résultats obtenus au terme des expérimentations montrent une bonne performance du système, néanmoins, les inconvénients comme les erreurs rencontrées lors de la détection des Entités Nommées et les définitions de tous les patrons possibles rendent l'approche moins attractive.

Ce type d'approche a également été utilisé par Zhang et al. [37] afin d'extraire des relations entre entités spatiales représentant des entités géographiques (exemple : Hanshan ou Yangtze River [4]). Leur idée est de se référer à un corpus annoté afin d'extraire manuellement des patrons syntaxiques exprimant une relation spatiale. Ces patrons sont par la suite transformés en règles étiquetées par un système dans le but de les faire correspondre à des données et d'extraire ainsi l'information voulue. Par exemple, le patron syntaxique *geoNE1 + Verb + geoNE2* sera soumis au système qui génère les règles syntaxiques et donnera le résultat de la règle suivante :
((geoNE) : geoNE1
(Segment. pos == v) : connect
(geoNE) : geoNE2)
Ce qui permettra de capturer toutes les phrases où une entité représentant un lieu

4. Le corpus est exprimé dans la langue Chinoise

géographique est suivie par un verbe et par une autre entité géographique étiquetée *geoNE2*. Nous pouvons constater que ce type de démarche peut engendrer des erreurs lors de la construction des règles syntaxiques, particulièrement si la phrase est longue.

Ces approches ne sont pas beaucoup utilisées. En effet, l'Extraction d'Informations basée que sur les méthodes du TALN a des limites dues au fait que les données sont sous différents formats qui changent d'un domaine à un autre ou d'une langue à une autre. Ce qui nous permet de dire que les patrons ne suivent pas forcément les mêmes structures morphologiques ou syntaxiques. Ainsi, définir ou présager toutes les combinaisons possibles serait difficiles à accomplir, et bien que le nombre des règles soit optimal, leur coût de construction en temps et en effort humain est non négligeable. Par conséquent, les recherches ont exploité d'autres méthodes basées sur la fouille de données. Ce contexte fera l'objet de la sous-section suivante.

3.2 Méthodes basées sur la fouille de données

Les données textuelles sont variées du point de vue de leurs structures ou du domaine auquel elles appartiennent. Les méthodes TALN développées pour extraire des informations s'avèrent avoir des limites pour faire face au problème d'hétérogénéité. C'est dans ce but que la combinaison entre les méthodes de fouille de données et les méthodes de TALN a été proposée. D'après les recherches menées sur l'Extraction de Relations, celle-ci peut être abordée selon trois types d'approches qui reflètent les méthodes d'apprentissage adoptées. Suivant cet aspect, les approches sont soit supervisées, soit semi-supervisées ou bien non supervisées. La suite de cette section est consacrée aux méthodes existantes qui ont contribué à l'ER.

3.2.1 Approches supervisées

Ce type d'approches s'appuie sur des données annotées pour extraire des relations souvent ciblées avant le lancement du processus d'extraction dont les méthodes supervisées existantes peuvent se résumer avec le schéma de la Figure 2.

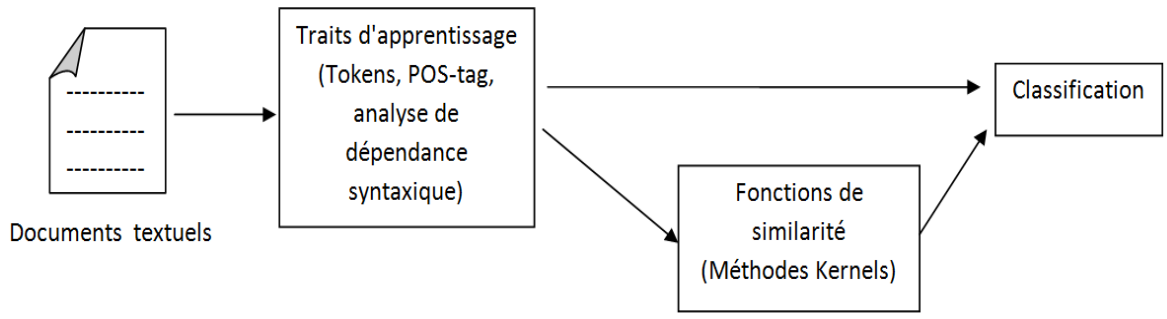

FIGURE 2 – Schéma global des méthodes supervisées existantes

Comme nous pouvons le constater sur le schéma (cf. Figure 2), l'Extraction de Relations peut être considérée en tant que processus de classification en s'appuyant sur certaines caractéristiques. Dans les travaux de Kambhatla [24], l'auteur identifie les 24 types et sous-types de relations présentes dans le corpus ACE [5] (relative-location, other,

5. Corpus ACE : `http://www.itl.nist.gov/iad/894.01/tests/ace/`

part-of, subsidiary, etc.). Pour ce faire, le modèle d'entropie maximale est appliqué pour une classification discriminante des types de relations. Dans l'approche, les données sont représentées en se basant sur différents traits d'apprentissage (syntaxiques, sémantiques, lexicaux). Ils sont déduits à partir des instances définissant les Entités Nommées, des mots se trouvant entre elles, ainsi qu'à partir des arbres de dépendance syntaxiques. (dans l'exemple *"Microsoft headquarters is located at Redmond", Microsoft* est une instance d'une entité de type ORGANIZATION). Les instances identifiées peuvent être des noms propres, des noms communs ou encore des pronoms personnels, quant aux mots reliant les deux entités, ils doivent être extraits (ex : *"is located"* en considérant l'exemple précédent) en indiquant leur nombre et le type de phrase (verbale, nominale ou prépositionnelle) dans laquelle les deux instances apparaissent. Tous les mots constituant la phrase sont étiquetés de façons à renseigner la dépendance entre les termes.

Afin d'analyser minutieusement les phrases, d'autres travaux ont émergé utilisant des méthodes à noyaux (MK) et des fonctions à noyaux complémentaires introduites dans Lodhi et al. [23]. Ces travaux permettent d'incorporer différents traits d'apprentissage plus enrichissants qui serviront à entraîner le SVM (Support Vector Machine) [15]. Ce dernier est un algorithme d'apprentissage qui aide à classer et séparer les données convenablement en définissant un classifieur ou des fonctions. Dans la phrase *The headquarters of Microsoft are located at Redmond*, il est important d'extraire le terme *headquarters* qui se trouve avant la première entité et non pas entre les deux, ceci permettra d'identifier la relation LOCATION-ORGANIZATION. Zhao et al. [39] ont appliqué cette approche dans le but d'améliorer les résultats de l'extraction en exploitant explicitement les traits combinés dans les travaux de Kambhatla [24]. Pour ce faire, les auteurs se focalisent sur les informations caractérisant la phrase d'où se fait l'extraction, à savoir, les termes la constituant et l'analyse de la dépendance syntaxique. Ces trois informations d'où sont dérivés les traits d'apprentissage, sont représentées par des vecteurs qui définissent les attributs de la relation à extraire ($1^{\text{ière}}$ entité, $2^{\text{ième}}$ entité, mots constituant la phrase, mots pertinents, informations sur la dépendance entre les mots). Ces attributs sont à leur tour détaillés : un type d'entité (ex : ORGANIZATION, LOCATION, etc.), un sous-type d'entité (ex : *city, country*, etc.), et un type pour chaque instance (ex : *Google : nom propre, he : pronom personnel, senator : nom commun*). Le mot est finalement enrichi par un étiquetage POS-tag pour déterminer sa nature ainsi que son lemme. Par exemple dans la phrase *"...are the Soldiers honored by the President"* l'entité *Soldiers* sera détaillée comme suit : ((“soldiers”, NNS, soldier, dépendance), “ORG”, “GOV”, “NAME”) où la première parenthèse décrit le terme *Soldiers* avec sa nature et son lemme, ORG pour ORGANIZATION qui est le type de l'entité, GOV pour GOVERNMENT qui définit le sous type de l'entité et NAME pour définir le type de l'instance *Soldiers*. La notion de dépendance quant à elle, est représentée par un ensemble de vecteurs spécifiant les caractéristiques du mot courant et son lien avec d'autres mots (l'arc de dépendance). Comparés aux travaux de Kambhatla [24], les résultats montrent une meilleure performance. Dans ces derniers travaux, c'est à l'utilisateur de choisir les traits d'apprentissage à prendre en considération. A contrario, dans Zhao et al. [39], tous les traits sont incorporés grâce aux fonctions à noyaux supplémentaires et sont soumis à SVM.

Dans l'approche précédente, les MK sont introduites pour mieux analyser la similarité entre les différents attributs des relations, améliorant ainsi l'extraction qui est basée sur le modèle SVM. Ce dernier a éveillé de l'intérêt dans d'autres travaux comme Zhou et al. [40] qui émettent l'hypothèse que l'utilisation de WordNet améliore l'extraction des relations ACE. Dans leur démarche, ils utilisent les même traits d'apprentissage des

travaux précédents, mais sans la définition de fonctions supplémentaires. L'inconvénient réside dans ce cas dans le choix des traits à prendre en considération. Nous trouvons également l'approche de Zang et al. [38] qui supposent que le modèle SVM peut être adapté dans le but d'extraire les relations entre deux entités spatiales. Ces dernières sont identifiées avec l'étiquetage CRF (Conditional Random Fields). Dans ces travaux, les auteurs définissent une méthode pour une classification multi-étiquetage. Ils se basent également sur les traits d'apprentissage dérivés des mots qui constituent la relation, comme l'aspect lexicale de l'entité spatiale (elle est constituée d'un ou de deux mots), son type (nom propre, pronom personnel, nom commun), sa catégorie géographique (rivière, lac, océan, etc.) ou encore la catégorie de la relation spatiale (20 catégories ont été définies : extend connection(EC), discrete connection (DC), northwest (NW), etc.). Représentés sous forme de vecteurs, ces traits d'apprentissage sont par la suite soumis à SVM pour prédire les relations spatiales. Même si le modèle peut extraire ce type de relations, elles peuvent toutefois être complexes, il faut donc avoir à disposition un large corpus annoté. Cette dépendance rejoint l'inconvénient des méthodes de TALN utilisées dans les travaux de Zhang et al. [37], qui visent également à extraire les relations spatiales, car la construction du corpus demande aussi du temps.

Nous avons pu constater que plusieurs contributions dans la catégorie des méthodes supervisées se sont focalisées sur les MK et les fonctions à noyaux. Ces dernières sont décrites dans Lodhi et al. [23] dans le cadre de la classification des documents, comme étant la similarité entre deux objets, mesurée par une fonction fondée sur le produit scalaire. À la base, ces méthodes ont été définies pour analyser les séquences de lettres dans un seul mot. Cette notion a été déployée pour analyser les séquences de mots dans les phrases pour l'extraction des relations. Dans les travaux de Bunescu et al. [13], les auteurs identifient les interactions entre les protéines à partir de textes médicaux. Pour ce faire, ils se focalisent sur l'analyse de la similarité du contexte qui entoure les entités (mots avant entités, mots entre entités et mots après entité). Par conséquent, trois fonctions à noyaux ont été définies, une pour chaque contexte. Ces relations sont définies avec les contextes, autrement dit elles sont constituées des termes de la phrase contenant deux entités. Après quelques prétraitements, ces relations sont représentées par des patrons (ex : "P1 interacts with P2") résultant des phrases enrichies avec des traits dont l'étiquetage POS-tag et l'étiquetage plus général (Active Verb, Passive Verb, etc.) ainsi qu'avec des informations de WordNet. Les résultats obtenus montrent une meilleure performance que les travaux de Blaschke et al. [8] mais l'effort manuel évité dans la construction des règles syntaxiques est transféré à l'annotation du corpus. Suivant ce même principe des MK, Zelenko et al. [36] proposent une approche pour l'extraction des relations ACE fondée sur l'analyse de la similarité entre non pas les sous-séquences mais entre la représentation syntaxique peu détaillée sous forme d'arbre qui décrit la relation, limitant ainsi les erreurs qui peuvent dérivées des étiquetages. Culotta et al. [16] utilisent quant à eux l'arbre de dépendance syntaxique en incorporant certains traits dans les noeuds (les mots, POS-tag, le type d'entité, et des informations provenant de WordNet). Ces noeuds sont par la suite formalisés sous la forme de vecteurs qui permettrons de calculer la similarité suivant la dépendance et les traits incorporés. Plus récemment Mu et al. [29] ont développé une approche fondée elle aussi sur les MK qui consiste à filtrer la phrase selon les entités qu'elle contient. Si une phrase contient plus que deux entités les liens peuvent varier, ce qui nécessite un filtrage, autrement dit, certaines entités peuvent appartenir à une relation mais jamais à une autre. Par exemple les entités PERSON et ORGANIZATION ont une relation AFFILIATION mais pas une relation KINSHIP (lien de parenté). Dans ce cas, la fonction de similarité définie, compare également les

relations selon leurs types, ce qui améliore l'extraction avec les MK. En général, les MK aboutissent à de meilleures performances que les méthodes initiales mais l'inconvénient se présente dans la complexité des calculs de similarité définis par les fonctions à noyaux.

Comme nous pouvons le constater, plusieurs contributions avec un apprentissage supervisé ont été proposées pour extraire différentes relations en évitant les travaux manuels qu'exigent les méthodes du TALN. Cependant, ces méthodes dépendent d'un large corpus annoté qui influence les résultats. L'élaboration de ce corpus demande aussi un travail manuel tout autant que la construction des règles syntaxiques des méthodes du TALN. Ce désavantage à conduit à l'émergence d'autres approches qui ne demandent pas beaucoup d'investissements manuels. Leur principe est développées dans les deux prochaines sous-sections.

3.2.2 Approches semi-supervisées

Ces approches prennent en entrée un ensemble d'exemples de taille réduite pour extraire des relations à partir d'un large corpus. Cette idée vient pallier le problème de l'effort manuel considérable qu'exigent les approches supervisées pour leur utilisation des données annotées. La procédure décrite peut se résumer dans la Figure 3.

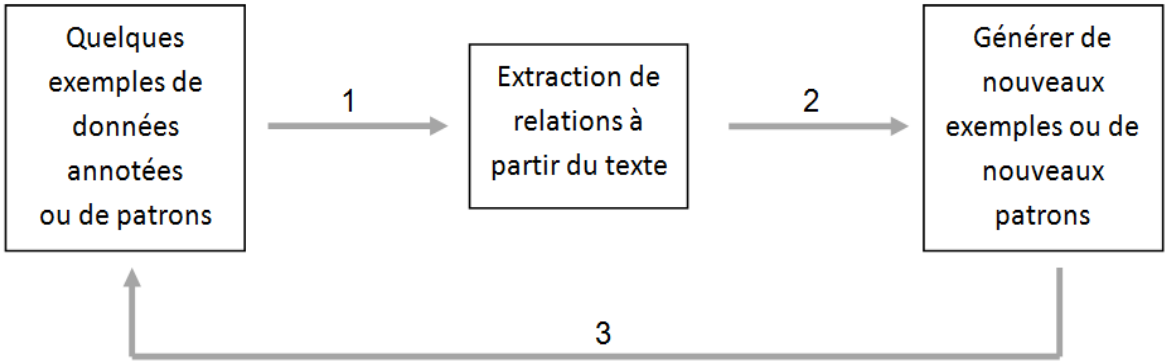

FIGURE 3 – Schéma global du principe des méthodes semi-supervisées

Ce principe fut introduit par Brin [12] qui développa le système DIPRE pour Dual Iterative Pattern Expansion. Il utilise en entrée un ensemble d'exemples de relations entre auteur et livre pour extraire certains patrons avec lesquels d'autres exemples de relations sont extraits. Cet outil commence par chercher les phrases où la paire (auteur, livre) apparaît et extrait pour chaque phrase un ensemble d'attributs (*author, title, order, url, prefix, middle, suffix*) où les éléments représentent :

- l'auteur pour *author*,
- le titre du livre pour *title*,
- *order* correspond à l'ordre dans lequel apparaît l'auteur par rapport au titre du livre dans la phrase,
- *prefix* (*suffix*) correspond aux mots se trouvant avant (après) le premier (le deuxième) élément (*author* ou *title* selon l'ordre) apparaissant dans la phrase,
- *middle* représente les mots se trouvant entre auteur et titre.

Supposons que la paire donnée en entrée soit (Paulo Coelho, Alchemist) et que les phrases où cette paire apparaît soient :
"the author Paulo Coelho wrote Alchemist in 1988",
"is the author Paulo Coelho who wrote Alchemist in Portuguese",
"buy Alchemist written by Paulo Coelho on Amazon".
Dans ce cas le système extrait les attributs :
(Paulo Coelho, Alchemist, true, url1, the author, wrote, in 1988),
(Paulo Coelho, Alchemist, true, url2, the author, who wrote, in Portuguese),
(Paulo Coelho, Alchemist, false, url3, buy, by, on Amazon).
Une fois l'étape de l'extraction des attributs achevée, les patrons sont générés en regroupant l'ensemble des attributs vérifiant que *order* et *middle* correspondent. Le *prefix final* (suffix final) correspondra à la chaîne commune entre le *prefix* (suffix) des attributs. Le patron généré dans l'exemple précédent sera donc [the author, (expression régulière), wrote, (expression régulière), in] où une expression régulière peut définir différents auteurs ou différents livres. Avec ce patron, d'autres paires peuvent être détectées, générant ainsi d'autres relations et d'autres patrons suivant un processus itératif.

Un mécanisme similaire a été adopté par Agichtein et al. [1] qui essayent d'extraire des patrons définissant une relation entre les entités LOCATION et ORGANIZATION dans le but de structurer les données qui peuvent être utiles pour répondre à des requêtes émises par des utilisateurs. Pour ce faire, la démarche suivie commence par donner un ensemble de patrons, par exemple : $< ORGANIZATION >$*'s headquarters in* $< LOCATION >$ et $< LOCATION >$*-based* $< ORGANIZATION >$. Afin de réduire l'espace de recherche, les auteurs procèdent à la détection des Entités Nommées, et comme dans le système DIPRE, ils extraient également des phrases dans lesquelles ces entités sont présentes. La différence entre les deux systèmes se présente dans l'attribution d'un poids à chaque terme, déterminé avec la fréquence normalisée, pour éviter que les patrons ne soient trop spécifiques. Un regroupement des phrases est élaboré en se basant sur une fonction de similarité. Chaque groupe génère un patron centroide, dont les caractéristiques ont une similarité avec les autres patrons du cluster, évalué grâce à une mesure de confiance qui est définie avec la fréquence de la paire $(ORGANIZATION, LOCATION)$ apparaissant dans l'itération, et la LOCATION n'apparaissant pas avec la même ORGANIZATION.

D'autres systèmes semi-supervisés extraient des relations entre deux phrases nominales, Banko et al. [5] utilisent l'analyse de dépendance syntaxique et des étiquetages pour un apprentissage semi-automatique afin de découvrir ce type de relations. S'inspirant de ces travaux, Fader et al. [18] développent un système qui réduit les erreurs rencontrées dans la génération des patrons, ceci en posant des contraintes syntaxiques sur les relations à extraire.

La majorité des démarches semi-supervisées s'intéressent à des relations assez simples du point de vue de la complexité syntaxique, lexicale ou sémantique. Appliquer ces méthodes à des relations plus complexes, ne nous donnerait pas forcément de bonnes performances car les exemples des paramètres d'entrée ne seront pas à la hauteur pour représenter ce type de relations, à moins d'avoir plusieurs exemples, ce qui rendrait la méthode supervisée. L'idéal serait de ne pas dépendre d'exemples prédéfinis avant le processus d'extraction. La sous-section suivante fait l'objet de ce principe qui ne requiert aucun exemple.

3.2.3 Approches non-supervisées

Les approches non-supervisées ne dépendent d'aucune utilisation d'exemples annotés dont l'élaboration exige un effort manuel considérable. De ce fait, plusieurs travaux se sont intéressés à la tâche d'Extraction d'Informations entre entités en introduisant des méthodes non-supervisées. Le schéma global des approches existantes est illustré dans la Figure 4.

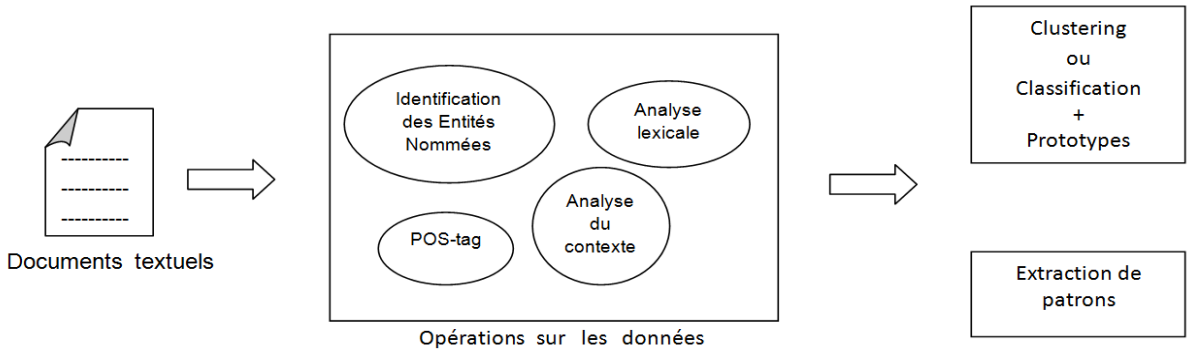

FIGURE 4 – Schéma global des méthodes non-supervisées existantes

Parmi les travaux basés sur du clustering, nous trouvons ceux de Hasegawa et al. [21] dont l'idée est de regrouper les paires d'entités en s'appuyant sur le clustering hiérarchique. Sa démarche consiste dans un premier temps à identifier les Entités Nommées avec l'étique-tage du système OAK [6]. Ensuite, les mots délimités par les entités, qui sont considérés comme le contexte de la relation, sont collectés et pondérés avec le produit TF*IDF(Term Frequency et Inverse Document Frequency) ce qui contribuera à regrouper les contextes selon leur similarité, en d'autres termes, les paires sont regroupées selon les mots qui les relient et en ne comparant que les paires ayant les mêmes types, formant ainsi des groupes dont le nombre est inconnu à l'avance. Une fois le regroupement accompli, chaque cluster obtenu est étiqueté avec son terme le plus fréquent.
Dans l'exemple suivant :
$< Person >$to govern$< GPE >$
$< Person >$to quit the government$< GPE >$
$< Person >$to be a member of the government$< GPE >$
où GPE siginifie *Geo-Political Entity*, toutes les phrases sont regroupées dans le cluster *government*. Le clustering a aussi été adopté par Eichler et al. [17] mais avec des mesures de similarité différentes. La comparaison est établie en s'appuyant sur les caractéristiques des phrases déduites de l'étape du prétraitement, à savoir, les informations de l'analyse de dépendance syntaxique et celles des Entités Nommées, mais aussi les informations lexicales de WordNet. Cependant l'extraction ne se base que sur les relations avec un verbe, ce qui ne généralise pas toutes les relations existantes.

D'autres méthodes non-supervisées plus originales sont apparues au fil des années dans le but d'extraire différents types de relations. Récemment, Nakashole et al. [30] ont mis au point un système permettant de créer un réseau sémantique non pas entre

6. OAK est un ensemble d'outils permettant d'analyse les documents écrit en Anglais. S. Sekine 2001 (New York University). `http://nlp.cs.nyu.edu/oak/manual.html`

les mots mais entre des patrons dont les caractéristiques représentent les relations entre Entités Nommées qu'ils extraient d'un texte. Prenons par exemple deux patrons : $< PERSON >$ *president of* $< ORGANIZATION >$ et $< PERSON >$ *founder of* $< ORGANIZATION >$, ces deux relations sont synonymes et un lien sémantique est créé entre les deux phrases. Pour ce faire, l'approche proposée est composée de plusieurs étapes. Dans un premier temps, les phrases sont soumises individuellement à un analyseur syntaxique afin d'obtenir l'arbre de dépendance permettant ainsi l'amélioration de la précision d'extraction des patrons dans les longues phrases. Ces dernières doivent contenir au minimum deux entités qui sont détectées grâce à l'utilisation de la base de connaissance YAGO2 ou Freebase. En se référant à l'arbre de dépendance syntaxique, les patrons sont générés et enrichis avec des adverbes ou des adjectifs ce qui leur donne un sens plus complet. Par la suite, ils sont transformés en séquences de mots accompagnés d'un étiquetage grammatical, de termes génériques et d'une signature ontologique qui permet d'attribuer un sous-type aux entités comme par exemple $<$ *singer* $>$. Avant de procéder à la construction de la taxonomie qui donnera le réseau sémantique, des généralisations syntaxiques et sémantiques sont appliquées en suivant certaines contraintes afin d'éviter la génération de patrons synonymes avec un sens contradictoire. Citons par exemple : $< PERSON >$ loves $< PERSON >$ et $< PERSON >$ hates $< PERSON >$ qui sont deux patrons ayant pour généralisation syntaxique $< PERSON >[vb]< PERSON >$. Si aucune contrainte n'est posée, ces deux patrons contradictoires seront synonymes. Vient à la fin l'étape de la construction de la taxonomie qui est fondée sur la construction d'un arbre préfixé (prefix-tree) où les éléments de l'arbre sont constitués des instances des entités et des informations de dépendance. Si deux patrons ont une paire d'entités en commun alors ils partagent le même préfixe représenté par un arc. Le graphe final obtenu est le système PATTY qui représente les synonymes entre patrons.

Certaines approches non-supervisées ne se limitent pas qu'aux relations entre Entités Nommées. Celle proposée par Corro et al. [14] a pour but de détecter des relations variées en se basant sur des propriétés grammaticales (sujet, verbe, complément d'objet) avec un large corpus. Néanmoins, les informations extraites peuvent refléter une relation entre Entités Nommées. En effet, dans la phrase *Nelson Mandela was born at Mvezo*, la relation extraite est sous la forme *("Nelson Mandela", "was born", "at Mvezo")* ce qui représente une relation entre deux Entités Nommées *(Nelson Mandela et Mvezo)*. L'application de la Reconnaissance des Entités Nommées donnera la possibilité de limiter le type de relations à celles qui nous intéressent mais l'utilisation de l'analyseur de dépendance syntaxique pourrait être lourde dans un large corpus. L'étude effectuée par Loglisci et al. [25] nécessite elle aussi une analyse de dépendance syntaxique, mais elle vise à extraire un autre type de relations : des relations entre entités spatiales. Leur méthode passe en outre par un processus de TALN dont la segmentation qui comporte les deux étapes de la tokenisation et la séparation en phrases, le POS-tagging et la lemmatisation. La relation à extraire est entre entités géographiques qui sont identifiées en exploitant l'ontologie GeoNames. Par la suite, un graphe est établi où les noeuds et les arcs correspondent respectivement aux mots et aux relations grammaticales entre deux termes. Le chemin existant entre deux entités, représentera la relation de dépendance qui est soumise à un classifieur basé sur des prototypes. Ces derniers sont obtenus grâce à l'ontologie fondamentale SUMO [7].

D'autres types de relations sont extraits avec les méthodes non-supervisées. Nous pouvons observer les relations entre gènes sur lesquelles les travaux de Béchet et al. [6]

7. SUMO (Suggested Upper Merged Ontology) : `http://www.ontologyportal.org/`

ont porté. Leur approche adopte la notion d'extraction de motifs séquentiels (Agrawal et Srikant [2]) pour identifier le lien entre les gènes et les maladies rares en suivant certaines contraintes, afin de minimiser l'espace des motifs découverts, comme une contrainte pour la fréquence, une contrainte pour minimiser les redondances et une autre pour vérifier la contiguïté des séquences. De ce fait, les auteurs appliquent une extraction de motifs sous plusieurs contraintes. Exploitant le même principe que cette approche, Alatrista et Béchet [4] abordent le problème qui fait l'objet du présent stage, à savoir, l'extraction des relations entre entités spatiales. Leurs travaux montrent certains désavantages dont le plus contraignant est la taille importante de l'espace des motifs découverts qui demande un effort humain considérable pour la validation des motifs.

Pour les méthodes non-supervisées, l'extraction s'appuie souvent sur l'analyse syntaxique, ce qui peut être un inconvénient dans un large corpus. Nous pouvons remarquer dans les travaux de Eichler et al. [17] que le taux d'erreur observé est assez élevé. Ceci est peut être dû au choix de l'outil de détection. Il serait dans ce cas plus intéressant d'utiliser ou de compléter la détection avec des ressources extérieures richement alimentées comme des ontologies ou des lexiques. Aussi, dans certaines approches, l'extraction avec les patrons est basée principalement sur des POS-tag, ce qui peut induire des erreurs, notamment si les phrases sont complexes comme dans le cas des relations spatiales. Ces dernières sont par conséquent non comparables avec des relations plus simples comme entre PERSON et ORGANIZATION où le lexique et la syntaxe peuvent plus ou moins être prédits.

3.3 Discussion

Le tableau présente une vue sur l'ensemble des méthodes citées précédemment qui ont contribué à l'Extraction de Relations. Ce récapitulatif souligne les principales caractéristiques qui peuvent différencier une approche d'une autre.

L'Extraction de Relations a fait l'objet de beaucoup de recherches et plusieurs méthodes ont été proposées dans le but d'améliorer les résultats obtenus. Lors de la partie précédente, nous avons fait un tour d'horizon sur les approches existantes. Les méthodes basées que sur du TALN sont les premières à avoir été abordées. Elles reposent sur des patrons élaborés manuellement en étudiant l'aspect syntaxique et morphologique des phrases et des relations visées pour l'extraction. Bien que ces méthodes aient tendance à avoir une bonne précision, les utiliser pour extraire les informations dans le cadre des travaux à réaliser dans ce stage, à savoir, l'extraction des relations spatiales, revient à prédire toute la terminologie qui leurs est liée, ainsi que toutes les combinaisons syntaxiques et morphologiques correspondantes aux phrases qui sont susceptibles de porter une information spatiale, ce qui consomme beaucoup en effort humain et en temps. Cet inconvénient rend ces méthodes moins attractives.

En effet, le désavantage des méthodes TALN a conduit à l'élaboration d'approches plus souples, il s'agit de celles basées sur les méthodes de fouille de données. En considérant l'aspect d'apprentissage des méthodes, comme expliqué précédemment (Section 3.2), trois types sont distingués : approches supervisées, semi-supervisées et non-supervisées. La première catégorie considère l'extraction de relations comme étant un processus

1. MK : Méthode à noyaux
2. POS : Part-Of-Speech

<table>
<tr><td colspan="3">Approches</td><td colspan="4">Caractéristiques</td></tr>
<tr><td colspan="3">Méthodes TALN</td><td colspan="4">[8] : Patrons syntaxiques construits manuellement + étiquetage et analyse syntaxique</td></tr>
<tr><td colspan="3"></td><td colspan="4">[37] : Patrons syntaxiques construits manuellement + étiquetage des entités géographiques</td></tr>
<tr><td rowspan="5">Méthodes orientées fouilles de données</td><td rowspan="3">Supervisées</td><td></td><td>Analyse syntaxique</td><td>POS-tag[2]</td><td>Ressources</td><td>Autres</td></tr>
<tr><td>Sans MK[1]</td><td>[24]</td><td>[24]</td><td>Absence de ressources</td><td>Classification basique</td></tr>
<tr><td>Avec MK</td><td>[39][16][40] [38]</td><td>[13][39] [36][16][40]</td><td>WordNet : [16] [13][40]</td><td>Toutes s’appuient sur SVM</td></tr>
<tr><td colspan="2">Non-supervisées</td><td>[17][30][14] [25]</td><td>[21][30][6] [14][25]</td><td>WordNet : [17]
YAGO2 : [30]
FreeBase : [30]
Geonames : [25]
SUMO : [25]</td><td>Clustering : [21][17]
Classifieur : [25]
Motifs séquent. : [6][4]
Patrons : [30]</td></tr>
<tr><td colspan="2" rowspan="4">Semi-supervisées</td><td colspan="4">[12] : Quelques exemples de relations pour extraire des patrons</td></tr>
<tr><td colspan="4">[1] : Quelques exemples de patrons pour en extraire d’autres</td></tr>
<tr><td colspan="4">[5] : Apprentissage semi-automatique avec des exemples, POS-tag et l’analyse syntaxique</td></tr>
<tr><td colspan="4">[18] : Contraintes pour corriger les patrons de [5]</td></tr>
</table>

Table 1 – Récapitulatif des méthodes d’Extraction de Relations

de classification, qui requiert la disponibilité des données annotées pour entraînement ainsi que le choix de certains traits d’apprentissage. Ces derniers concernent l’aspect syntaxique comme les POS-tag et l’arbre de dépendance syntaxique, l’aspect lexical qui consiste à prendre en considération les mots qui entourent les Entités Nommées, ainsi que l’aspect sémantique qui s’appuie sur des ressources extérieures comme WordNet. Ces informations d’apprentissage sont utilisées dans les tests individuellement ou en les combinant. Par conséquent, la classification dépend de ces données qui influencent les résultats. Cependant, choisir les traits d’apprentissage adéquats à l’extraction des relations spatiales demande de faire plusieurs essais car nous ignorons quels sont les meilleurs aspects (syntaxiques, lexicaux et sémantiques) à prendre en considération pour ce type de relations. Dans cette même catégorie des approches supervisées, les méthodes à noyaux (MK) ont également été abordées dans le but de représenter explicitement les traits d’apprentissage. Ces méthodes intègrent des fonctions de similarité pour une meilleure classification en considérant la correspondance entre les mots ou les arbres de dépendance syntaxique. Bien qu’il y ait une meilleure performance pour cette classification, le plus grand désavantage réside dans la disponibilité des données annotées et ceci dans toutes les méthodes supervisées. Par conséquent, employer ce premier type d’approches dans la découverte des relations spatiales, nous contraindra à avoir des exemples d’entraînement annotés assez diversifiés pour représenter les relations à extraire, or ces exemples ne sont pas toujours disponibles, les collecter et les annoter consommera du temps.

L’inconvénient lié à la disponibilité des données annotées a suscité un attrait pour des méthodes qui exigent un minimum d’entraînement. Ces approches suivent un apprentissage semi-supervisé. Elles ne se contentent que d’un petit ensemble d’exemples de patrons ou de données spécifiques sur lesquels elles se basent pour en extraire d’autres.

Pour certains domaines où les relations se ressemblent comme par exemple les relations entre ORGANISATION et LOCALITE, adapter ce type d'approches peut donner des résultats satisfaisants étant donné que les exemples peuvent être assez représentatifs pour en extraire d'autres. Les paramètres d'entrée sont dans ce cas une bonne base. En revanche, les exploiter pour les relations spatiales ne nous donnerait pas la certitude d'avoir de bonnes extractions car ce sont des relations plus complexes et variées du point de vue syntaxique, sémantique et lexical.

Le dernier type d'approches, non-supervisées, a le grand avantage de ne pas dépendre des données annotées, les méthodes adaptées sont diverses, certaines utilisent le clustering en se basant sur différentes similarités, d'autres s'appuient sur l'extraction automatique de patrons ou sur les données grammaticales pour extraire les relations. Cependant, la majeur partie de ces méthodes dépend des analyseurs de dépendance syntaxique ce qui peut induire des erreurs. De plus, certaines de ces méthodes extraient des relations générales (sujet, verbe, complément), les appliquer à des relations bien spécifiques telles que les relations spatiales serait difficile à adapter. Toujours dans ce même type d'approches non-supervisées, une méthode fondée sur les motifs séquentiels a été proposée, son principe ne requiert ni données annotées ni analyse de dépendance syntaxique. La méthode peut être adaptée pour extraire différentes relations y compris les relations spatiales. Comme mentionné auparavant (Section 3.2.2), des premiers travaux ont été réalisés. Ils exploitent ces motifs séquentiels afin d'extraire les relations entre entités spatiales dans le cadre du projet ANIMITEX, et les avantages qu'offrent ces méthodes non-supervisées expliquent le choix de leur adaptation. Dans les dernières expérimentations [4], les résultats aboutissent à une taille importante de l'espace des motifs extraits. Cet inconvénient est dû à l'existence de motifs généraux qui ne représentent pas forcément une relation spatiale. Des contraintes, dont les détails seront expliqués dans la prochaine section, ont été proposées afin de limiter le nombre de motifs et en extraire des plus pertinents. Cependant, ces contraintes restent insuffisantes. Le but des travaux à venir est d'améliorer ces premières contributions en étudiant plus explicitement les aspects sémantiques ou morphosyntaxiques relatifs aux relations spatiales présentes dans les données textuelles. Dans les sections qui suivent, nous détaillerons le processus que nous avons appliqué afin d'améliorer les travaux avec la généralisation ou l'apprentissage via d'autres corpus non hétérogènes comme nous le verrons dans les expérimentations.

4 Extraction de relations spatiales

Après avoir vu les différentes approches qui ont contribué à la résolution de la problématique d'Extraction de Relations avec leurs avantages et leurs inconvénients, nous allons maintenant consacrer les parties à suivre à l'extraction des relations spatiales qui représente l'objectif principal des travaux de ce stage. Nous allons ainsi voir la méthode appliquée pour aboutir à l'extraction de ces relations. Cette méthode se base sur l'extraction des motifs séquentiels qui permettront la génération de patrons relatifs aux relations spatiales. De ce fait, la méthode est une combinaison entre les méthodes TALN et la fouille de données.

4.1 Problématique

La finalité de ce stage s'inscrit dans le cadre du projet ANIMITEX (ANalyse d'Images fondée sur les Informations TEXtuelles). Comme l'indique l'intitulé, le but de ce projet est d'apporter des informations complémentaires à l'analyse des images satellitaires en exploitant des données textuelles. Cette perspective s'appuie sur L'extraction des relations spatiales à partir de textes.
Nous présentons ci-dessous, des définition clés pour la suite du rapport, ainsi que la description de la problématique via un exemple.

4.1.1 Définitions

Entité Spatiale (ES) : dans ces travaux, une Entité Spatiale correspond à une localité comprenant une Entité Nommée Géographique qu'on appelle toponyme. Par exemple *Montpellier* ou *l'étang de Thau* sont des Entités Spatiales.

Relation spatiale : correspond au lien sémantique existant entre Entités Spatiales. En d'autres termes, une relation spatiale définit la position géographique d'une Entité Spatiale par rapport à une ou d'autres Entités Spatiales ou simplement la position d'une Entité Spatiale. Par exemple "*à 30 km de la mer*" décrit une relation spatiale.

Type de relations spatiales : le lien existant entre les ES peut être exprimé avec plusieurs expressions décrivant différents types de relations. Ces dernières peuvent être de type : distance (*à 30 km de*), ou bien, orientation (*à l'est de*) ou adjacence (*à coté de*) ou encore inclusion (*dans la zone de*). D'autres relations spatiales peuvent êtres observées, notamment celles qui sont relatives aux routes ou aux rivières (*de A jusqu'à B*, *A traverse B*, *A débouche dans B*, *relie A à B*).

Pour expliciter la problématique, considérons l'exemple des phrases qui suit :

Exemple : soient les phrases suivantes :
(1) : *Montpellier se situe à 10 km de la mer Méditerranée.*
(2) : *Caen est à deux heures au nord-ouest de Paris.*

Chacune des deux phrases contient une relation spatiale. La première représente le lien entre *Montpellier* et *la mer Méditerranée*, quant à la deuxième, c'est une relation entre *Caen* et *Paris*. Notre but est d'extraire automatiquement des patrons linguistiques qui vont représenter ce type de relations. Par exemple, le patron $\{PREP\ au\}\{N$

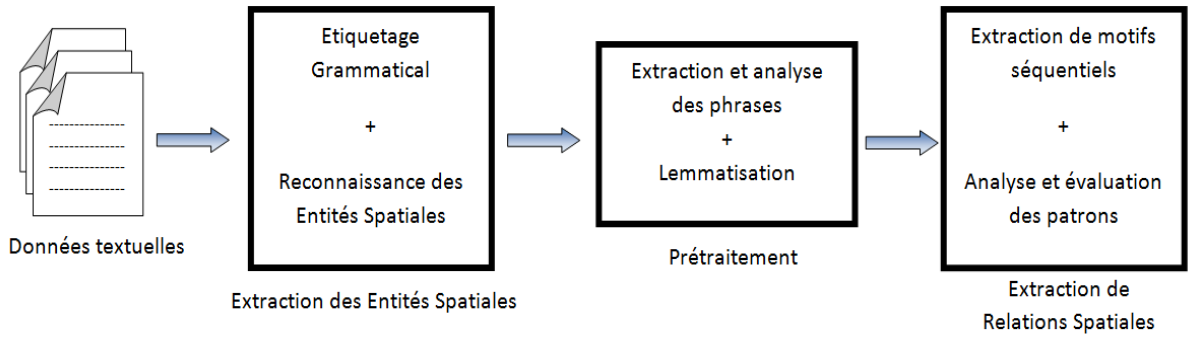

FIGURE 5 – Schéma global de la démarche

nord-ouest}{*DET de*}{*ES*} permettra l'extraction de la deuxième phrase. Pour ce faire, il est primordial dans un premier temps de détecter les phrases qui peuvent contenir une relation spatiale car c'est sur ces phrases que l'approche va s'appuyer pour extraire les patrons. Noyées dans un texte, la phrase *(1)* et la phrase *(2)* peuvent être détectées par la reconnaissance de *Montpellier*, *la mer Méditerranée*, *Caen* et *Paris*. Cette étape d'identification des ES constitue la partie de *la Reconnaissance des Entités Spatiales* (RES). Une fois détectées, ces dernières contribueront à identifier les phrases susceptibles d'être porteuses de relations spatiales. A partir de là, nous pouvons appliquer la méthode d'extraction de motifs séquentiels pour extraire les patrons. De ce fait, notre démarche est scindée en deux parties principales comme illustré dans la figure précédente (cf. Figure 5), à savoir, l'extraction des ES et l'extraction des relations spatiales.

L'identification de ce type de relations relève un défit dans le Traitement Automatique du Langage Naturel, car leur structure est complexe. Nous sommes en effet confrontés à une complexité syntaxique. Pour mieux comprendre ce verrou, prenons par exemple les phrases suivantes :
(1) : *Proche de la mer Méditerranée, Montpellier attire les vacanciers.*
(2) : *La Loire est orienté sud-nord jusqu'aux environs de Briare.*

En considérant la position de la relation spatiale par rapport à l'ES, nous pouvons constater que d'un point de vue syntaxique, la relation peut se trouver très proche avant l'ES (*Proche de la mer Méditerranée*) ou très proche après l'ES (*La Loire est orienté sud-nord*). Ce qui nous ramène à nous interroger sur ce point : est-il plus fréquent que la relation soit avant ou après l'ES ? Quels sont les types de relations dans les deux cas ?

Dans la suite de cette section, nous détaillerons la méthodologie adoptée dans nos travaux pour répondre à ces questions et extraire par la suite les relations qui nous intéressent.

4.2 Méthodologie

4.2.1 Extraction des Entités Spatiales

Comme expliqué dans la problématique, nous devons passer par l'étape de la Reconnaissance des Entités Spatiales afin d'identifier et constituer l'ensemble des phrases qui seront la base de l'extraction des relations. Dans cette première étape, nous nous sommes basés sur les travaux effectués par Gaio et al.[20]. Dans leur approche, les auteurs visent à annoter automatiquement certaines expressions spatiales entre autres les ES. La méthodologie suivie dans le cas de la Reconnaissance des Entités Spatiales peut

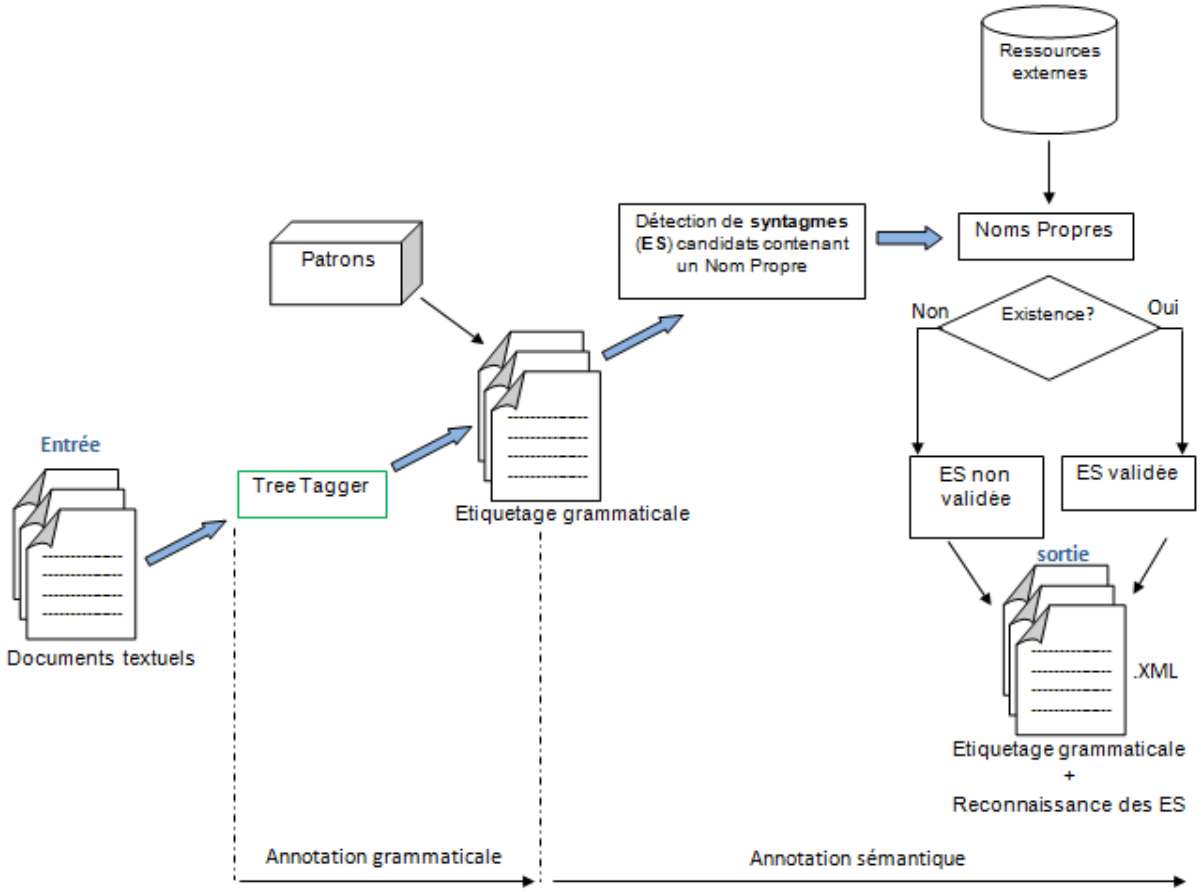

FIGURE 6 – Schéma global résumant la Reconnaissance des Entités Spatiales

être décrite par la figure ci-dessus (cf. Figure 6). Pour réaliser ce procédé, le processus établi passe par une chaîne de traitement englobant deux principales annotations : grammaticales et sémantiques.

-Annotation grammaticale : la présente annotation consiste à associer une catégorie grammaticale aux mots constituant les données textuelles (exemple : V pour verbe ou encore ADV pour adverbe). Cet étiquetage grammatical est réalisé avec l'outil Tree-Tagger.

-Annotation sémantique : pour ce type d'annotation, des patrons ont été définis afin d'identifier certains types de syntagmes, comme pour le cas des ES par exemple (*l'étang de Thau*) ou encore certains syntagmes pouvant introduire les relations spatiales (proche de, centre de, etc.).

La Reconnaissance des Entités Spatiales se base sur les noms propres et des ressources externes, particulièrement Geonames et BD Nyme[8]. Un nom propre (exemple : *Montpellier*) ou un syntagme contenant un nom propre (exemple : *l'étang de Thau*) sont potentiellement considérés comme étant des ES. Cependant, leur validation comme telle repose sur la définition suivante : une ES est validée si les noms propres qui la composent sont présents dans les ressources utilisées.

8. BD Nyme : base de données toponymiques http ://professionnels.ign.fr/bdnyme

La figure qui suit (cf. Figure 7) décrit quelques exemples des résultats que nous pouvons obtenir suite à cette chaîne de traitements. Ces résultats sont exploités dans le cadre de nos travaux. Les deux premiers exemples représentent une annotation sémantique qui définit une ES. Dans le premier exemple, *Sète* est reconnue comme étant une ES car le nom propre est présent dans les ressources de Geonames et BD Nyme. Quant au deuxième exemple, il décrit un syntagme identifié par un patron et reconnu comme étant une ES suite à la présence de *Frontignan* dans les ressources. Le troisième exemple représente un syntagme pouvant introduire une catégorie de relations spatiales qui est ici de type *adjacence*. Nous pouvons également remarquer qu'il y a un étiquetage grammatical associé aux mots présents dans le corpus.

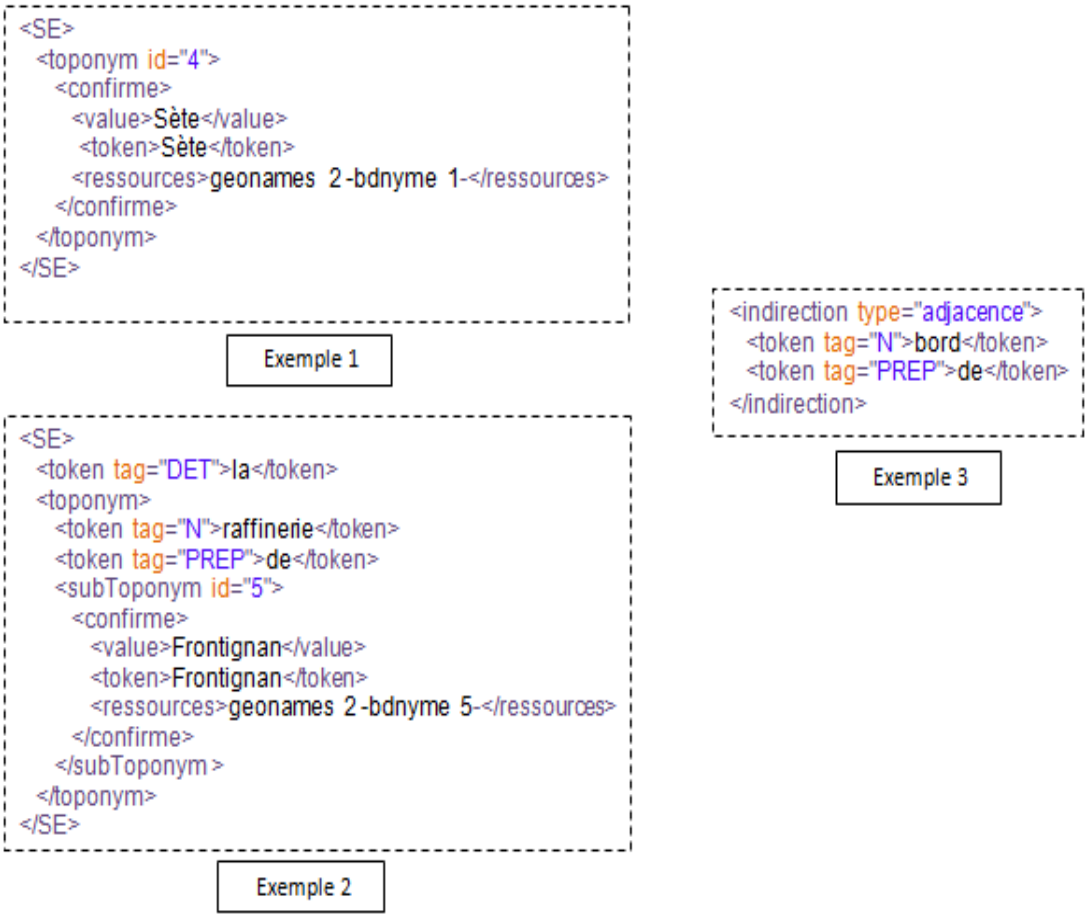

FIGURE 7 – Exemple de résultats de la chaîne de traitement pour la RES

L'extraction des ES nous permettra dans un premier temps d'identifier les phrases susceptibles de contenir une relation spatiale. Cependant, cette étape ne suffit pas à elle seule pour éviter du bruit, que ce soit dans les phrases ou dans les patrons. Par conséquent, une étape intermédiaire de prétraitement que nous détaillons ci-dessous semble nécessaire avant de procéder à l'extraction.

4.2.2 Prétraitements

Sélection des phrases

Suite à une analyse de plusieurs phrases, nous avons constaté que la majorité des phrases contenant au minimum deux ES sont porteuses de relations spatiales. Cependant, cette hypothèse peut générer quelques bruits. En effet, dans le corpus du Midi Libre (Section 5) nous pouvons repérer de courtes phrases contenant deux ES mais elles ne sont pas pour autant porteuses de relations spatiales. Par exemple, l'expression "*Entrée gratuite à 20 h. Montpellier, Sussargues et Nîmes.*" va être divisée en deux phrases : la première "*Entrée gratuite à 20 h .*" car elle se termine par un point et la deuxième sera "*Montpellier, Sussargues et Nîmes*". En se basant sur l'hypothèse citée précédemment, cette dernière phrase va être sélectionnée alors qu'elle ne contient aucune relation spatiale. Pour éviter ce type de bruit, les phrases doivent contenir au minimum cinq mots dont deux ES. Ceci dans le but d'éviter de perdre de l'information. Par exemple, "*Montpellier est proche de Sète*" est une phrase qui est pertinente et contient cinq mots. Et c'est pour ce motif que le nombre minimal a été fixé à cinq. Nous pouvons également trouver de longues phrases contenant deux ES. Dans ce dernier cas, les phrases peuvent contenir des syntagmes et un lexique non utiles à notre objectif. Par conséquent, il nous semble pertinent de prendre les "n" premiers mots entourant les ES. Ce dernier raisonnement sera mieux détaillé dans la prochaine sous section consacrée à la fouille de données séquentielles (sous section C.2). En plus de ce prétraitement, nous nous sommes également intéressés aux ponctuations. Non omises, ces dernières peuvent faire converger les résultats vers des patrons trop spécifiques où les ponctuations apparaîtront. Ce point a été donc pris en compte en supprimant les ponctuations.

La lemmatisation

Le principe de la méthodologie que nous suivrons dans ces travaux repose sur la notion de fréquence : plus une liste ordonnée de mots apparaît dans les données textuelles, plus sa chance d'apparaître dans les résultats augmentera. Il serait dans notre cas plus avantageux de passer par la lemmatisation. Cette dernière consiste à mettre les mots sous leurs formes canoniques (leur lemme, verbe à l'infinitif, les mots au singulier). La lemmatisation permettra au lexique d'être plus général et moins spécifique. Ainsi, plusieurs mots peuvent être représentés par une seule forme et donc par un seul mot. Dans ce cas, les lemmes vont apparaître plus souvent dans le corpus, ce qui induit une augmentation dans la probabilité de leur apparition dans les résultats. Par exemple la phrase "*les villes sont proches de la mer*" deviendra " *la ville être proche de la mer*".

4.2.3 Extraction des motifs séquentiels

Une fois les ES identifiées et les prétraitements achevés, le processus passe à la seconde étape, et comme expliqué dans la section précédente, la méthode adoptée s'appuie sur l'extraction des motifs séquentiels sous contraintes multiples [6]. Le résultat attendu par cette méthode est d'avoir à la fin des patrons relatifs aux relations spatiales qui contribueront à leur extraction dans les textes.

ID	*Séquences*
1	<(d e) (a b c) (d f)>
2	<(a b) (d e) (c) (f)>
3	<(e) (a b) (d f)>

TABLE 2 – Exemples de SDB

A. Concepts et définitions

Le processus d'extraction de motifs séquentiels utilise en entrée un ensemble de séquences appelé base de séquences qu'on note SDB. Un tuple de la base *(Sid,S)* représente donc une séquence S identifiée par Sid. Une séquence S correspond à une liste ordonnée d'itemsets, ce qui se résume par $S =< I_1, I_2...I_m >$ où I représente un itemset. Ce dernier est constitué d'un ensemble de littéraux appelés items $I = (i_1, i_2...i_n)$. Par exemple $S =<(d\ e)\ (a\ b\ c)\ (d\ f)>$ est une séquence contenant trois itemsets donc sa taille est égale à 3, *(d e)*, *(a b c)* et *(d f)* où l'itemset *(a b c)* comprends trois items *a*,*b* et *c*.
Une séquence $S_1 =< I_1, I_2, ..., I_n >$ est une sous séquence de $S_2 =< I'_1, I'_2, ..., I'_m >$ notée $S_1 \preceq S_2$ si S_1 est incluse dans S_2, c'est à dire si $\exists\ j_i \in [1, m]$ tq $I_1 \subseteq I'_{j_1}$ ..... $I_n \subseteq I'_{j_n}$. Par exemple si $S_1 =<(d)(b\ a)>$ et $S_2 =<(d)(e\ c)(b\ a)>$ alors S_1 est dite incluse dans S_2 et on note $S_1 \preceq S_2$. Nous pouvons qualifier S_2 d'une séquence supérieure de S_1.
On appelle *support* d'une séquence noté *sup(S)* le nombre de séquences dans lesquelles apparaît la séquence S dans la base de séquences.
Une séquence S est dite fréquente si son support *sup(S)* est supérieur ou égal à un seuil minimal appelé support minimal et noté sup_{min}.

En prenant comme exemple la SDB décrite dans la table ci-dessus (cf. Table 2) et les séquences $S =<(d\ e)\ (f)>$ et $S' =<(a\ b)\ (c)>$, *sup(S)* sera égale à 2 car S apparaît dans *S1* et *S2* de la SDB. Quant à *sup(S')*, il sera égal à 1 car *S'* n'apparaît que dans la première séquence de la SDB. Prenons maintenant un sup_{min} égale à 2. Dans ce cas, la séquence S est fréquente car son support $sup(S) \geq sup_{min}$. La séquence *S'* quant à elle n'est pas fréquente car son support $sup(S) < sup_{min}$.

B. Algorithme sous contraintes

L'algorithme utilisé dans ces travaux repose sur celui proposé par Béchet et al. [6]. Son principe se base sur une approche hybride regroupant deux autres algorithmes : CloSpan [35] et BIDE [32]. L'algorithme permet de réduire l'espace de recherche des motifs grâce à différentes contraintes. En effet, les auteurs ont proposé un algorithme permettant l'extraction de motifs séquentiels sous contraintes multiples. Ces dernières sont détaillées ci-dessous :

-***La contrainte de fréquence*** : pose la condition que chaque motif extrait doit être fréquent. Prenons l'exemple suivant : soit le motif $S =<(a)\ (e)>$, la SDB présentée dans la Table 2 et un sup_{min} égale à 2. Pour cet exemple, *sup(S)=1*, de ce fait S n'est pas un motif fréquent et ne sera dans ce cas pas extrait et cela en est de même pour ses motifs supérieurs. Ceci en raison de la propriété suivante : Si $S_1 \preceq S_2$ alors $sup(S_1) \geq sup(S_2)$

-***La contrainte de fermeture*** : stipule qu'un motif candidat S est fermé s'il n'existe pas *S'* tel que $S \preceq S'$ et *sup(S)=sup(S')*. Supposons un $sup_{min} = 2$ et un motif candidat S =<*(e) (d)*>. En considérant la SDB précédente (cf. Table 2) nous pouvons constater que *sup(S)=2* car nous trouvons S dans S_1 et S_3. S est fréquent mais il existe un motif S' =<*(e) (b) (d)*> dont le support est égal à 2 et $S \preceq S'$ donc S n'est pas un motif fermé.

-***La contrainte de Gap*** : elle est posée pour vérifier la contiguïté entre les itemsets. En d'autres termes, elle met une condition quant au nombre d'itemsets pouvant exister entre deux itemsets voisins d'un motif candidat en considérant l'ensemble des séquences. Le nombre doit être compris dans un intervalle *[M,N]* d'où l'annotation $S_{[M,N]}$. Par exemple $S_{[0,1]}$ =<*(e) (f)*> apparaît trois fois dans la SDB, dans S_1, S_2 et S_3 car le nombre d'itemsets autorisé entre *(e)* et *(f)* est soit 0 ou bien 1. Par contre $S_{[1,2]}$ =<(a) (f)> n'apparaît qu'une seule fois, et ceci dans S_1 car dans ce cas, le Gap est compris entre 1 et 2.

-***La contrainte d'appartenance*** : elle permet de spécifier un ensemble d'items obligatoires ou facultatifs qu'un motif extrait peut contenir. Par exemple si on pose la condition qu'un motif doit obligatoirement avoir l'item *(a)* et un item parmi {*d, c*} alors le motif S' =<*(e) (b) (d)*> ne sera pas extrait car il ne contient pas l'item obligatoire *(a)*. Le motif S' =<*(a) (e) (f)*> ne sera pas non plus extraits car il ne contient aucun item de la liste facultative, à savoir {*d, c*}. En revanche, le motif S' =<*(a) (c) (f)*> sera extrait car il satisfait les deux appartenances. Formellement nous pouvons définir cette contrainte comme suit : $L1 \wedge L2$ où $L1 = (i_1 \wedge i_2 \wedge ... \wedge i_n)$ et $L2 = (i'_1 \bigvee i_2 \bigvee ... \bigvee i_n)$ tel que L1 correspond aux contraintes d'appartenance obligatoires et L2 aux contraintes d'appartenance facultatives.

-***La contrainte de la longueur*** : avec cette contrainte, l'utilisateur a la possibilité de limiter le nombre d'itemsets constituant le motif extrait. Par exemple, si on pose la condition que le motif doit avoir une taille comprise entre 3 et 6 alors S' =<*(a) (e c b)*> ne sera pas extrait car il ne contient que 2 itemsets (il est de taille 2).

La difficulté qui se pose dans cette approche est de prendre en considération toutes les contraintes. En effet, l'introduction de Gap peut invalider la propriété d'anti-monotonie du support. Par conséquent, les auteurs ont proposé un algorithme efficace permettant de respecter toutes les contraintes.

L'algorithme calcule dans un premier temps les motifs fréquents à un seul item qui vont contribuer à la construction d'une base projetée. Pour ce faire, l'algorithme extrait à partir de la SDB les suffixes de ces motifs considérés comme préfixes dans les séquences. Cette base projetée est construite en respectant la contrainte de fréquence et va faire état d'une nouvelle base de séquences à partir de laquelle l'algorithme calculera cette fois-ci les motifs fréquents contenant deux items et construira de ce fait, une autre base projetée à partir de ces nouveaux motifs fréquents. Ainsi de suite, jusqu'à la réduction totale de la taille de la base projetée. Les motifs sont générés par une extension à droite. Ce procédé s'exécute récursivement pour vérifier si chaque motif généré respecte les contraintes. Par exemple, pour maintenir la contrainte de la fermeture, les auteurs se basent sur les travaux de Wang et al. [33] dont le principe stipule qu'un motif peut être étendu par un item ou un itemset et ceci à gauche ou à droite du motif. Avec ce procédé, l'algorithme peut vérifier si un motif respecte la contrainte ou non.

C. Les relations spatiales et les motifs séquentiels

L'extraction de motifs séquentiels permet d'extraire automatiquement des sous séquences à partir d'un ensemble de séquences en combinant les différents items les constituant tout en respectant un certain nombre de contraintes qui peuvent être paramétrées selon les résultats voulus. Nous voulons dans notre cas extraire automatiquement des patrons linguistiques qui nous permettront d'extraire des relations spatiales. Les patrons sont une combinaison catégories grammaticales et de mots. Par exemple le patron $\{PRO\ se\}\{V\ situe\}\{PREP\ à\}\{ES\}$, peut représenter un patron pertinent pour extraire des relations spatiales. Nous pouvons imaginer que les patrons que nous voulons extraire sont des motifs séquentiels dont les itemsets sont constitués d'un mot et de sa catégorie grammaticale. Nous nous sommes donc basés sur ce contexte pour extraire les relations spatiales représentées par des patrons. En d'autres termes, les séquences vont être décrites par des itemsets allant au maximum jusqu'à la taille 2 (deux items), pour un mot et sa catégorie grammaticale, sauf pour l'ES qui est représentée par l'item *TOPO* pour toponyme.

Des premiers travaux ont été accomplis [4] avec la méthode des motifs séquentiels sans passer par l'étape de prétraitement proposée dans ce rapport (Section 4.2.2). Avec une contrainte de Gap $\in [0,0]$, les résultats ont montré que les motifs extraits étaient généraux (cf. Figure 8). Par conséquent, choisir les meilleurs patrons décrits par les motifs consomme du temps car la majorité des patrons sont représentés par les catégories grammaticales. Extraire des relations avec ce type de patrons à partir de textes hétérogènes peut générer du bruit et converger ainsi vers un résultat non souhaitable. Cette absence de termes dans les patrons résultants est dû au fait que les relations spatiales sont riches du point de vue lexical donc les mots ont une faible probabilité quant à leur apparition dans les motifs. De plus, la lemmatisation n'a pas été prise en considération. Le résultat souhaitable est qu'après extraction de motifs, ces derniers représenteront des patrons plus spécifiques comme $\{PRO\ se\}\{V\ situer\}\{PREP\ à\}\{TOPO\}$ et non pas $\{PRO\}\{V\ \}\{PREP\}\{TOPO\}$ qui est un patron général, pouvant extraire une phrase qui ne comporte pas de relations spatiales. Par exemple on peut extraire à partir de ce patron la phrase *"Il se rend à Montpellier"*, comme on peut extraire *"Versailles se situe à Paris"*.

Motif séquentiel	Support
<(TOPO)(DET)(NOM)(PREP)(DET)(TOPO)(NOM)>	7
<(TOPO)(DET)(PREP)(DET)(TOPO)(TOPO)>	7
<(TOPO)(DET)(PREP de)(N)(PREP)(TOPO)>	7
<(V)(NOM)(PREP)(NOM)(CONJC)(PREP)(DET)(TOPO)>	7
<(TOPO)(DET)(NOM)(DET la)(TOPO)>	8
<(DET)(NOM)(PREP)(NOM)(PREP)(TOPO)(CONJC)(DET)(TOPO)>	8
...	...

FIGURE 8 – Résultats des premiers travaux

C.1 Amélioration : généralisation des relations spatiales

Étant donné que l'extraction des motifs se base sur la fréquence et que les relations ont un lexique riche, l'idéal serait d'augmenter la probabilité de l'apparition des mots relatifs aux relations spatiales dans les patrons. L'idée proposée est de faire une généralisation

selon la catégorie des relations. En d'autres termes, nous allons regrouper un ensemble de différents mots dans une seule classe. Chaque classe représentera un seul type de relations. Ce principe permettra d'augmenter le support d'une séquence donnée si cette dernière contient un syntagme pouvant introduire une relation.

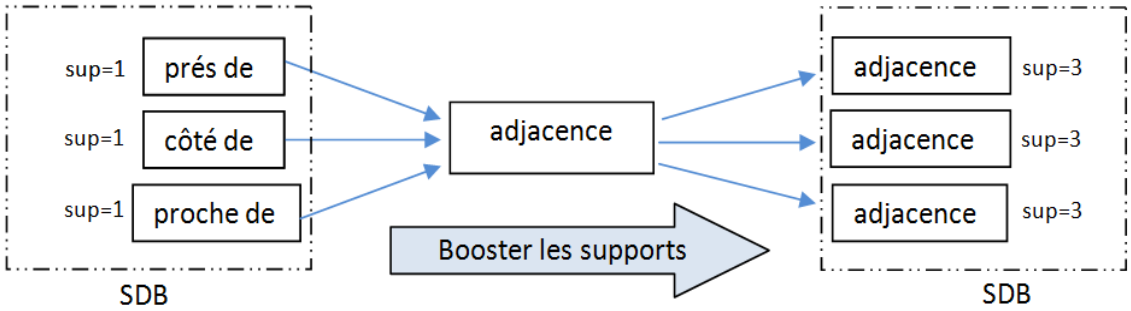

FIGURE 9 – Intérêt de la généralisation

Nous avons vu précédemment (Section 4.1) qu'il peut exister quatre types de relations : l'orientation, la distance, l'adjacence et l'inclusion. Pour faciliter cette tâche de généralisation, nous pouvons exploiter les travaux de Gaio et al. [20]. En effet, dans l'exemple 3 de la Figure 7, l'expression *"bord de"* est annotée dans la balise *indirection* comme étant de type adjacence. Grâce à la généralisation, au lieu de représenter cette expression telle qu'elle est dans les séquences ($\{N\ bord\}\{PREP\ de\}$) nous la considérons comme étant un seul item décrit par ($\{adjacence\}$). La figure ci-dessus illustre ce principe (cf. Figure 9).
Le lexique le plus répondu avec lequel une relation peut être introduite est décrit dans la table ci-dessous (cf. Table 3). Chaque colonne représentera donc une catégorie de généralisation.

inclusion	*distance*	*adjacence*	*orientation*
autour	kilometre	entre	sud
proche	kilomètre	milieu	nord
auprès	km	intérieur	sud-ouest
bord	hectometre	centre	nord-ouest
fond	lieue	partie	sud-est
côté	pieds		nord-ouest
périphérie	decametre		
lisière			
près			
tout près			
environs			

TABLE 3 – Lexique pouvant introduire des relations

C.2 Le choix des séquences

Suite aux différentes questions posées concernant la structure des relations spatiales, nous avons établi un protocole pour mieux observer leur syntaxe et choisir de ce fait une meilleure stratégie quant aux choix des séquences. Ceci en s'appuyant sur l'approche décrite dans ce rapport et un corpus provenant du Midi Libre (Section 5). Le protocole établi suit deux expérimentations : la première pour analyser la syntaxe ou le lexique se trouvant avant l'ES. Quant à la deuxième, elle est consacrée à ce qui se trouve après l'ES. Ces deux expérimentations ainsi que les résultats obtenus sont détaillées ci-dessous (C.2.1 et C.2.2).

C.2.1 Le préfixe de l'Entité Spatiale

Notre but ici est d'analyser la syntaxe et le lexique se trouvant avant l'ES. De ce fait, nous avons testé l'approche en choisissant comme séquence les mots se trouvant avant chaque ES, présente dans les phrases résultantes de l'étape du prétraitement (sélection des phrases et lemmatisation). Ce protocole a été testé pour des séquences dont le nombre d'items dans un itemset est égale à 1, c'est à dire sans catégorie grammaticale. Par exemple *S=<(soir) (sur) (le) (D114) (adjacence) (TOPO)>* où "*adjacence*" remplace "*près de*" et "*TOPO*" remplace l'ES. Il a également été testé sur des séquences avec les catégories grammaticales (au maximum deux items dans un itemset). Par exemple : *S=<(N soir) (PREP sur) (DET le) (N D114) (adjacence) (TOPO)>*.
En faisant varier la contrainte de Gap de 0 à 2, les résultats obtenus quant à la présence (+) ou l'absence (-) des syntagmes sont résumés dans la table suivante :

Gap \ Syntagme	*inclusion*	*adjacence*	*orientation*	*distance*
Séquences sans catégories grammaticales ($sup_{min} = 7$)				
$[0-0]$	+	−	+	+
$[0-1]$	+	+	+	+
$[0-2]$	+	+	+	+
Séquences avec catégories grammaticales ($sup_{min} = 50$)				
$[0-0]$	−	−	−	−
$[0-1]$	+	−	−	−
$[0-2]$	+	+	+	+

TABLE 4 – Présence de syntagme exprimant une relation avant l'ES

Discussion

Nous remarquons que les syntagmes exprimant une adjacence, orientation, inclusion ou distance apparaissent dès un Gap $\in$ [0-0] ce qui signifie que ces expressions sont très proches avant l'ES car suivant le principe de la contrainte de Gap, celle-ci permet d'ignorer des itemsets. Or, quand le Gap $\in$ [0-0] les motifs extraits sont des itemsets consécutifs dans les séquences et aucun itemset n'est ignoré. Cependant, nous remarquons que pour le cas des séquences à catégories grammaticales, la relation apparaît à partir du Gap [0-1]. Ceci s'explique par le fait que pour ces séquences, le support minimal a été augmenté dans le but de rendre l'extraction des motifs moins complexe.

C.2.2 Le suffixe de l'Entité Spatiale

Le but ici est d'analyser la syntaxe se trouvant après l'ES. Pour ce faire nous avons pris comme séquence les mots se trouvant après chaque ES présente dans les phrases résultantes de l'étape du prétraitement. Pour les expérimentations nous avons suivi le même principe que pour le préfixe des ES. Par exemple *S=< (TOPO) (PREP dans) (DET le) (N refuge) (ADV spécialement) (V construire)>* pour les séquences avec catégories grammaticales et *S=< (TOPO) (dans) (le) (refuge)(spécialement) (construire)>* pour les séquences sans catégories grammaticales. Les résultats obtenus sont résumés dans la table suivante :

Gap \ Syntagmes	*inclusion*	*adjacence*	*orientation*	*distance*
Séquences sans catégories grammaticales ($sup_{min} = 7$)				
$[0-0]$	$-$	$-$	$-$	$-$
$[0-1]$	$-$	$-$	$-$	$-$
$[0-2]$	$+$	$-$	$-$	$-$
Séquences sans catégories grammaticales ($sup_{min} = 50$)				
$[0-0]$	$-$	$-$	$-$	$-$
$[0-1]$	$-$	$-$	$-$	$-$
$[0-2]$	$-$	$-$	$-$	$-$

TABLE 5 – Présence de syntagmes exprimant une relation après l'ES

-Discussion Les syntagmes exprimant une adjacence, orientation, inclusion ou distance n'apparaissent que lorsque le Gap $\in$ [0-2] ce qui signifie que ce type de syntagmes se trouve fréquemment bien après l'ES. Cependant, et au travers de ces résultats, nous avons constaté qu'une proposition subordonnée revient souvent après l'ES. Par exemple $\{TOPO\}\{PRO\ où\}\{PRO\}\{V\}$ ou encore $\{TOPO\}\{PRO\ qui\}\{V\}\{DET\}$ et ceci pour introduire une description de l'ES. Nous avons également remarqué qu'il y avait beaucoup de verbes pertinents pour exprimer une relation spatiale comme $\{TOPO\}\{en\}\{passer\}\{par\}$.

Ces deux expérimentations nous montrent que même si les syntagmes pouvant exprimer les quatre types de relations se trouvent fréquemment avant l'ES, il ne faut néanmoins pas négliger ce qu'il y a après l'ES car une relation pourrait être exprimée par un verbe qui se trouve justement après l'ES. Suite à ce constat, nous avons opté pour des séquences englobant ce qu'il y a après et avant l'ES. La table ci-dessous (cf. Table 6) décrit des exemples de la représentation des séquences.

ID	*Séquences*
1	Montpellier se situe à 10 km de
	(TOPO)(PRO se)(V situer)(PREP à)(distance)(DET de)
2	Caen être à deux heures au nord-ouest
	(TOPO)(V être)(PREP à)(N deux)(N heure)(PREPDET au)(adjacence)
3	est reliée à l'Espagne par le tunnel
	(V être)(V relier)(PREP à)(TOPO)(PREP par)(DET le)(N tunnel)

TABLE 6 – Exemples de SDB dans les relations spatiales

Autrement dit, pour chaque ES présente dans les phrases sélectionnées dans le prétraitement, nous générons les séquences en prenant les mots entourant l'ES. Par conséquent, chaque séquence comptera une seule ES remplacée par le nom générique "*TOPO*". Une fois toute la démarche développée, nous l'avons testée sur différents corpus. Les détails de ces expérimentations sont rapportés et interprétés dans la section suivante (Section 5).

5 Expérimentations

La présente section détaillera les expérimentations faites lors de ces travaux ainsi que l'interprétation des résultats obtenus. En effet, différents tests ont été effectués afin de comprendre l'intérêt des paramètres relatifs à la fouille de données séquentielles, notamment les contraintes à adapter, ou encore l'intérêt de passer par un prétraitement judicieux. Les expérimentations se sont basées sur trois corpus : le contenu du premier est hétérogène d'un point de vue contexte, le deuxième décrit le récit de randonneurs, quant au dernier, il décrit les communes de la France et ses rivières.

5.1 Exploration des motifs

Avant d'entamer les protocoles expérimentaux, nous expliqueront comment nous pouvons procéder afin de faciliter la navigation dans les motifs résultants. Ces derniers s'avèrent être en effet nombreux même avec l'utilisation de certaines contraintes qui permettent de réduire l'espace de recherche. Par exemple, la contrainte du support minimal est très pertinente pour élaguer les motifs. En suivant la propriété "*tout motif supérieur à un motif non fréquent est non fréquent*" l'algorithme interrompra la recherche des motifs supérieurs si le motif courant n'est pas fréquent, ce qui réduira le nombre des motifs extraits mais également la complexité du calcul. De façon similaire à cette contrainte relative à la fréquence, celle de fermeture permet également de réduire cet espace de recherche tout en réduisant aussi la complexité du calcul.

Pour ne pas se heurter à la difficulté de naviguer dans les motifs, nous pourrons exploiter l'outil *Camelis* (cf. Figure 10) développé par S.Ferré[19]. Le fonctionnement de cet outil repose sur l'avantage que les motifs extraits sont partiellement ordonnés. En effet, dans les résultats obtenus, certains patrons sont inclus dans d'autres patrons, avec un support évidemment différent pour respecter la contrainte de fermeture. Par exemple, si le patron1=$\{PRO\ se\}\{V\ situer\}\{PREP\}\{TOPO\}$ est validé alors il en sera de même pour le patron2 $\{PRO\ se\}\{V\ \}\{PREP\}\{TOPO\}$.

5.2 Corpus : Midi-Libre

5.2.1 Tests et résultats

Pour mener nos expérimentations, nous avons collecté dans un premier temps des articles provenant de la presse locale Midi-Libre. Ce choix se justifie par le fait que les informations qui le constituent, portent beaucoup sur des faits réels qui peuvent rapporter de nouvelles informations pertinentes quant à l'analyse des images satellites. Nous pouvons trouver par exemple un article sur la construction d'un nouvel immeuble. Cette information aidera les experts de la télédétection à préciser à quoi correspond un chantier donné. Une fois collectés, 104 articles ont été traités via la chaîne de traitement de la Reconnaissance des Entités Spatiales. Nous avons donc fait des premiers tests en nous

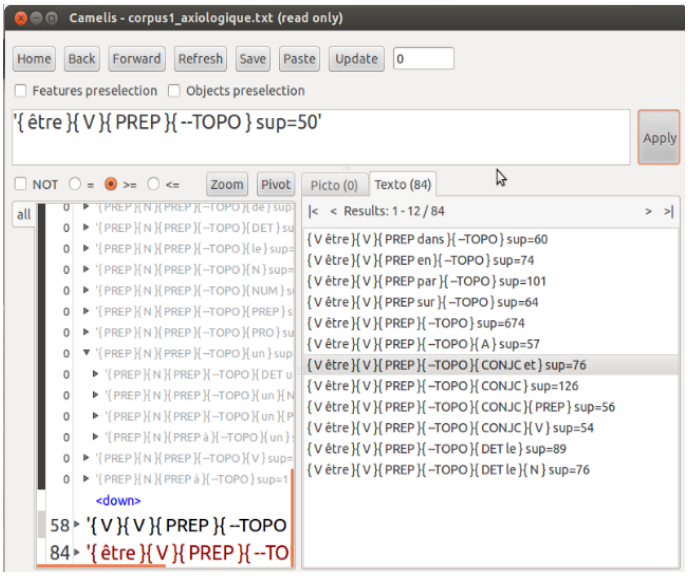

FIGURE 10 – L'outils Camelis

appuyons sur ces données et ce sans généralisation et sans lemmatisation. Les résultats constatés ont montré que les données étaient bruitées notamment avec des informations portant sur des concerts apportant des informations où aura lieu l'événement. Étant des noms de rue, ces lieux ne sont pas forcément présents dans les ressources externes qui sont utilisées dans la Reconnaissance des Entités Spatiales, ce qui a engendré du bruit dans les motifs. La table qui suit (cf. Table 7) donne les détails des paramètres utilisés et le motif (ID = 1) qui nous a permis de constater que le corpus était bruité.

ID	*motifs*	*sup*	sup_{min}	*Gap*
1	{h}{NPr Secret}{NPr Place}{TOPO}	24	10	[0-0]
2	{TOPO}{PREP}{DET la}{N}{PUN ,}	12	10	[0-0]

TABLE 7 – Premiers résultats sur un corpus bruité

Nous remarquons que le premier motif a un support de 24 avec un Gap entre [0-0]. Autrement dit, les éléments du patron se répètent dans 24 séquences et ce consécutivement. Ce patron correspond à une adresse où aura lieu un événement. D'ailleurs on remarque le "*h*" en tout début du motif qui signifie " *heure*". Le deuxième patron décrit sur la table, démontre pourquoi nous avons omis les ponctuations. Nous remarquons que même si le patron peut être pertinent, la présence de la ponctuation pourrait compromettre son application sur un texte. Plus explicitement, si un fragment du texte peut correspondre à la relation présente dans le patron et que ce fragment ne contient pas de ponctuations, il ne sera donc pas extrait. Ce type de patrons est qualifié de trop spécifique. De plus, le nombre de séquences résultantes de ce premier corpus n'est pas en adéquation avec notre objectif : seulement 368 séquences ont été générées sur un ensemble de 1964 phrases extraites. Ce qui n'est pas suffisant pour la fouille de données.

Suite à ce constat, la collecte de 6500 nouveaux articles, toujours à partir du Midi libre, a été effectuée en se basant sur le filtrage qui suit : les articles portant sur les événements comme les concerts, le sport ou encore le cinéma ont été écartés, réduisant de ce fait le bruit et augmentant la taille des données. Ce nouveau corpus a été annoté

pour l'identification des ES mais aussi pour procéder à la généralisation.

Pour les nouvelles expérimentations, le corpus a été divisé en deux parties : deux tiers pour l'apprentissage des patrons et l'autre tiers pour l'évaluation des patrons extraits. Nous avons suivi deux protocoles expérimentaux. Pour ce qui est du premier, les séquences sont constituées d'itemsets ne comportant pas de catégories grammaticales et ne comportent donc que les lemmes. Quant au deuxième protocole, il stipule que les itemsets des séquences comportent au maximum deux items : le lemme et sa catégorie grammaticale. La table qui suit (cf. Table 8) rapporte quelques exemples de patrons qui sont soit nouveaux ou bien enrichis en raison de la présence des syntagmes définis dans la généralisation. Les paramètres des contraintes adoptées sont comme suit :
-Premier protocole (Séquences sans catégories grammaticales) :
support minimal= 10, Gap compris entre [0-2], taille du motif à extraire comprise entre 4 et 10 et contrainte d'appartenance obligatoire pour l'item {TOPO}.
-Deuxième protocole (Séquences avec catégories grammaticales) :
support minimal= 50, Gap compris entre [0-2], taille du motif à extraire comprise entre 4 et 10 et contrainte d'appartenance obligatoire pour l'item {TOPO}.

Protocole	*Patrons*	*Enrichi*	*Nouveau*
1	{*de*}{*TOPO*}{*jusqu'*}{*au*}		✓
	{*de*}{*TOPO*}{*inclusion*}{*et*}	✓	
	{*le*}{*relier*}{*TOPO*}{*à*}		✓
	{*a*}{*TOPO*}{*sur*}{*le*}		✓
	{*situer*}{*inclusion*}{*TOPO*}{*et*}{*le*}	✓	
	{*TOPO*}{*en*}{*passer*}{*par*}		✓
2	{*PREP*}{*N*}{*inclusion*}{*TOPO*}	✓	
	{*DET le*}{*N*}{*inclusion*}{*TOPO*}	✓	
	{*PREP*}{*N*}{*adjacence*}{*TOPO*}	✓	
	{*PRO se*}{*V*}{*PREP*}{*TOPO*}		✓
	{*PREP*}{*TOPO*}{*PREP sur*}{*DET*}{*N*}		✓
	{*N*}{*inclusion*}{*TOPO*}{*CONJC et*}	✓	

Table 8 – Exemples de patrons extraits du Midi Libre

5.2.2 Discussion

Les patrons obtenus montrent que la généralisation et la lemmatisation ont apporté une amélioration aux résultats et que l'annotation des syntagmes effectuée lors de la Reconnaissance des Entités Spatiales est bien robuste. En effet, comparés aux résultats des premiers travaux (cf. Figure 7), nous remarquons que ces patrons sont plus riches et plus spécifiques car les lemmes sont présents. Ce qui permettra de générer moins de bruit pour l'extraction des fragments correspondant dans les textes. Nous constatons également que les termes génériques (*inclusion, adjacence, distance, orientation*) qui remplacent les types de syntagmes introduisant des relations sont très fréquents dans le corpus. Ce qui est peut être du au fait que les rédacteurs cherchent à véhiculer l'information en toute simplicité avec les termes les plus courants. Par ailleurs, les patrons contenant les syntagmes sont qualifiés d'enrichis, car il n'y a pas que le syntagme dans le patron, mais il est associé à

son contexte, ce qui ramènera à extraire plus efficacement des relations à partir du texte. Les patrons qualifiés de nouveaux sont majoritairement extraits via le premier protocole. Ceci est dû au fait qu'avec des séquences dont la taille des itemsets est fixée à 1 (premier protocole), nous pourrons baisser le support minimal et avoir de ce fait des résultats qui peuvent être pertinents. Nous pouvons aussi confirmer que les verbes sont bien présents dans les relations, on y trouve par exemple le patron $\{TOPO\}\{en\}\{passer\}\{par\}$. En revanche, parmi les patrons extraits, nous avons remarqué la présence du verbe *avoir* juste après l'ES, comme par exemple, $\{TOPO\}\{V\ avoir\}\{V\}\{DET\ un\}$ et ce dès un Gap entre [0-0] c'est à dire que l'auxiliaire suit directement "*TOPO*". Or, dans le contexte des relations spatiales, une ES ne peut pas être suivie par cet auxiliaire car ce verbe va probablement être employé avec un verbe transitif, ce qui est sémantiquement pas cohérent dans les relations spatiales. Ce résultat s'explique par les ambiguïtés rencontrées lors de la Reconnaissance des Entités Spatiales. Par exemple, dans la phrase "*Hollande a fait une adhésion...*", l'Entité Nommée "*Hollande*" est reconnue comme étant une ES, ce qui peut engendrer du bruit dans les patrons.

5.2.3 Évaluation

Cette étape d'évaluation consiste à trouver le fragment correspondant aux patrons validés dans un texte. Dans notre cas, comme mentionné plus haut, un tiers du corpus est consacré à l'évaluation. En appliquant les patrons extraits sur ce tiers de données, les relations extraites à partir du texte sont prometteuses et encourageantes pour la suite du projet ANIMITEX. En effet, comme nous allons le constater, les informations extraites sont pertinentes et représentent des relations spatiales.

Patrons	**Fragments correspondants dans le texte**
$\{de\}\{TOPO\}\{inclusion\}\{et\}$	*...de Bagnas située entre Cap d'Agde et Marseillan-plage...*
$\{a\}\{TOPO\}\{sur\}\{le\}$	*...à Sète sur les voies ...*
$\{a\}\{TOPO\}\{dans\}\{le\}$	*...à Banyuls-sur-Mer dans les Pyrénées-Orientales...*
$\{le\}\{relier\}\{TOPO\}\{à\}$	*...la voie reliant Béziers à Cahors...*
$\{a\}\{TOPO\}\{et\}\{adjacence\}$	*...à Chamrousse et Domène près de Grenoble...*
$\{PREP\}\{TOPO\}\{PREP\ dans\}\{DET\}\{N\}$	*...à Marseillan dans un campement ...*
$\{DET\}\{N\}\{PREP\ sur\}\{TOPO\}$	*...d' hinterland sur Poussan...*
$\{PRO\ se\}\{V\}\{PREP\}\{TOPO\}$	*...se jeter dans l'Étang de Thau...*
$\{TOPO\}\{PREP\ sur\}\{DET\}\{N\}$	*...Marseillan sur la route des Parcs...*
$\{TOPO\}\{PRO\ qui\}\{PRO\}\{V\}$	*...Vic qui se trouve à Conques...*
$\{N\}\{inclusion\}\{TOPO\}\{CONJC\ et\}$	*...la D613 entre Gigean et Fabrègues...*
$\{PREP\}\{N\}\{adjacence\}\{TOPO\}$	*...sur la D114 près de Fabrègues...*
$\{PREP\}\{TOPO\}\{PREP\ sur\}\{DET\}\{N\}$	*...À Poussan sur le boulevard Prosper-Gervais...*

Table 9 – Évaluation des premiers patrons

Nous rapportons dans la Table9 quelques exemples. Il est à noter que comme les

patrons ont été extraits en suivant un Gap entre [0-2], dans la validation nous avons pris en considération cette contrainte.

Si nous regardons bien les résultats auxquels nous avons abouti via l'extraction des fragments correspondants aux patrons dans le texte, nous pouvons déduire que les patrons extraits via l'approche sont pertinents et prometteurs. Mais d'un autre coté, nous pouvons confirmer que les relations qui se trouvent dans ce premier corpus ne sont pas assez complexes, ou bien si elles existent, elle ne sont pas fréquentes. Et c'est peut être pour cette raison que l'approche, qui se base sur la fréquence, n'a pas permis de les extraire. Via cette évaluation, nous confirmons également que plus le patron est spécifique, moins il y a du bruit dans l'extraction de la relation qui lui correspond dans un texte. Par exemple, avec le patron $\{PRO\ se\}\{V\}\{PREP\}\{TOPO\}$, où les lemmes ne sont pas présents, le fragment "*se déplaceront à Sète*" a également été extrait, ce qui ne correspond pas à une relation spatiale. Il ne faut pas aussi négliger que le corpus de la presse diffuse beaucoup d'informations sur des personnes qui sont quelques fois reconnues par erreurs comme étant des ES, car le nom propre est présent dans les ressources Geonames et BD-Nyme, ce qui engendre du bruit dans les motifs extraits et donc dans les patrons. Ainsi, les patrons extraits via notre approche ne sont pas tous pertinents.

La constatation du manque de relations dans ce corpus et du bruit rencontré à cause des multiples noms propres qui ne sont pas forcément des ES, nous nous sommes dirigés vers une autre alternative concernant le corpus d'apprentissage. L'hypothèse émise est d'appliquer le processus d'apprentissage, en l'occurrence ici notre approche, sur d'autres corpus. Ces derniers doivent être riches en relations spatiales et non hétérogènes dans leurs contenus. Une fois l'approche appliquée et les patrons extraits, nous évaluerons ces derniers sur le premier corpus, à savoir, celui du Midi Libre, car c'est sur ce corpus que se basent les ambitions du projet ANIMITEX. Pour ce motif, deux autres corpus on été collectés, l'un porte sur le récit de randonnées quant à l'autre il correspond aux descriptions des communes de la France ainsi que ses fleuves et rivières. Ce dernier corpus a été collecté, comme nous allons l'expliquer, à partir de Wikipedia.

Déboucher au village de Saint-Christophe-sur-Dolaison et faite le tour de l'église pour observer son remarquable clocher à peigne. Poursuivre en quittant le bourg, prendre à droite pour passer devant des tables de pique-nique. Puis ressortir de l'autre côté de la départementale en empruntant un petit tunnel. Face à la croix, prendre à gauche. Traverser successivement les hameaux de Tallode, Liac et Lic en suivant le balisage blanc - rouge du GR 65. Vous arrivez ensuite au village de Ramourouscle. Au centre, bifurquer à gauche. Suivre la route jusqu'à la chapelle Saint-Roch. Traverser Monbonnet pour rejoindre la D589. La suivre à gauche sur 150m, avant de la traverser pour s'engager à droite sur un chemin montant entre les maisons. Traverser un plateau qui à un moment se scinde en deux. Prendre le chemin de gauche qui grimpe et monte vers la crête sombre de sapins. Arriver à un embranchement, le chemin continue vers la droite. Pénétrer dans la forêt en suivant le balisage blanc - rouge du GR 65. Vous atteignez le lac de l'Oeuf que vous longez par la gauche. Déboucher sur une route. La suivre à gauche sur 150m, avant de prendre rapidement un chemin qui s'abaisse à droite (signalisation). Continuer tout droit en restant sur ce chemin dans les bois. Il se poursuit par une petite route à découvert bordée de part et d'autre de champs cultivés. Couper la D589 pour prendre en face la route menant au village de Chier. Traverser la place centrale en passant devant la mairie.

FIGURE 11 – Extrait du corpus Randonnées

5.3 Corpus : Randonnées

Les données qui constituent ce corpus sont beaucoup moins ambiguës que le corpus provenant de la presse. En effet, ce deuxième corpus est caractérisé par des données homogènes dont l'idée principale tourne autours de la description de randonnées. La Figure11 illustre un extrait d'un texte appartenant à ce dernier corpus.

5.3.1 Tests et résultats

La terminologie du présent corpus est très intéressante. En effet, le lexique utilisé est relatif aux ES et leurs positions géographiques. Dans l'ensemble, près de 1400 documents ont été exploités. Et en suivant le deuxième protocole testé sur le premier corpus (les séquences avec catégories grammaticales) ainsi que les même paramètres des contraintes pour l'extraction des motifs, nous avons obtenus de meilleurs résultats. Quelques uns des patrons les plus pertinents sont résumés dans la Table 10.

Patrons	*Enrichi*	*Nouveau*
{*DET le*}{*N sentier*}{*PRO qui*}{*V*}{*TOPO*}		✓
{*TOPO*}{*V*}{*PREP à*}{*A gauche*}{*DET*}		✓
{*DET le*}{*N route*}{*V*}{*–TOPO*}		✓
{*V*}{*PREP jusqu'*}{*PREP à*}{*TOPO*}		✓
{*TOPO*}{*distance*}{*PREP*}{*N*}	✓	
{*PRO qui*}{*V*}{*PREP vers*}{*TOPO*}		✓
{*TOPO*}{*V traverser*}{*DET le*}{*N*}		✓
{*TOPO*}{*CONJC*}{*V rejoindre*}{*DET le*}		✓
{*TOPO*}{*PREP après*}{*DET le*}{*N*}		✓
{*TOPO*}{*PREP sur*}{*DET*}{*N droit*}		✓
{*TOPO*}{*PREP par*}{*DET le*}{*N*}{*PREP*}		✓
{*TOPO*}{*orientation*}{*PREP*}{*DET le*}	✓	
{*orientation*}{*TOPO*}{*PRO*}{*V*}	✓	
{*PREP à*}{*N droit*}{*PREP*}{*TOPO*}		✓

Table 10 – Exemples de patrons extraits du corpus Randonnées

Les résultats obtenus sur ce nouveau corpus sont meilleurs que les premiers d'un point de vue nouveauté des patrons. En effet, ces derniers sont plus riches et observent plus de lemmes que dans les premiers résultats. Cette pertinence est associée à la terminologie homogène et fréquente qui constitue le corpus. Donc pour pallier le problème de la fréquence, qui ne nous permet pas d'extraire des résultats pertinents sur un corpus dont les données sont hétérogènes et bruitées, nous pouvons avoir recours à d'autres données dont le contenu est riche en connaissances ciblées par l'extraction.

5.3.2 Évaluation

Afin de souligner la véracité de nos dires, nous avons évalué ces nouveaux patrons sur le premier corpus, ceci dans le but de voir si ces patrons peuvent nous extraire des fragments relatifs à des relations spatiales à partir du Midi Libre. Les résultats sont rapportés sur la Table 11.

Patrons	**Fragments correspondants dans le texte**
{*TOPO*}{*PREP après*}{*DET le*}{*N*}	*...situé sur Marseillan juste après le passage à niveau ...*
	...dans Coste après le pont.....
{*TOPO*}{*PREP sur*}{*DET*}{*N droit*}	*...Saint-Césaire et sur sa droite l' ancienne Sommières...*
	...Pont des Abîmes qui enjambe Vidourle sur la droite ...
{*DET le*}{*N sentier*}{*PRO qui*}{*V*}{*TOPO*}	*...le sentier qui monte vers Saint-Roman...*
{*DET le*}{*N chemin*}{*PRO qui*}{*V*}{*TOPO*}	*...le chemin des artilleurs qui serpente Rimbault...*
{*TOPO*}{*V*}{*PREP à*}{*A gauche*}{*DET*}	*...Vidourle remonter à gauche le parking...*
	...Pont des Abîmes et s' engager à gauche sur la piste ...
{*DET le*}{*N route*}{*V*}{*TOPO*}	*...la route reliant Villeveyrac à Mèze...*
	...la route menant à Pomerol...
{*TOPO*}{*PRO*}{*V*}{*PREP sur*}{*DET le*}	...à Montbazin sur la route menant à Cournonterral...
{*PREP à*}{*N droit*}{*PREP*}{*TOPO*}	*...à droite le chemin de halage Carnon...*
{*V*}{*TOPO*}{*PREP par*}{*DET*}	*...rejoignant Sète par et le début de la piste...*
	...contourner Barcelone par l' ouest...
{*DET le*}{*N*}{*PRO qui*}{*V*}{*TOPO*}	*...la seule voie qui relie Montpellier vers Espagne...*
	...La ligne qui reliait Caunes à Moux...
{TOPO}{distance}{PREP}{N}	*...Étang du Fangassier à 4 km de la mer...*
	...en Espagne à 4 km de la sortie pour Agde...
{*TOPO*}{*orientation*}{*PREP*}{*DET le*}	*...Barcelone par l' ouest depuis le chantier...*
{*TOPO*}{*PREP par*}{*DET le*}{*N*}{*PREP*}	*...à Londres par la route de la soie...*
	...aux Quintillan en passant par les terrasses basses...
{PRO qui}{V}{PREP vers}{TOPO}	*...la piste cyclable qui longe l'Étang de Thau vers la Crique de l'Angle...*
	...un chemin empierré qui monte vers Saint-Antoine...

TABLE 11 – Évaluation des patrons extraits du corpus Randonnées sur le corpus Midi Libre

Nous affirmons via cette évaluation que les patrons extraits à l'issu de cet apprentissage sur un deuxième corpus, peuvent effectivement nous extraire des relations bien pertinentes qui alimenteraient les ambitions du projet ANIMITEX. Nous déduisons également que ces patrons existent dans le corpus du Midi Libre puisque nous avons pu extraire des fragments qui leurs correspondent. Cependant, ils n'ont pas été extraits via l'apprentissage sur le premier corpus en raison de leur faible fréquence. Néanmoins, dans cette évaluation, quelques bruits persistent avec les patrons moins spécifiques. Bien que ce deuxième corpus possède une terminologie en rapport avec des ES, sémantiquement parlant, il représente des itinéraires d'où nous pouvons déduire implicitement des relations spatiales plutôt que des relations spatiales explicites. Par exemple à partir du fragment "*...la piste cyclable qui longe l'Étang de Thau vers la Crique de l'Angle...*", nous pouvons déduire qu'il y a "*une piste cyclable*" pas loin de "*l'Étang de Thau*" et ce dernier est proche de "*la Crique de l'Angle*". De ce fait, nous avons voulu tester notre approche sur un troisième corpus où les relations sont représentées comme nous allons voir d'une façon plus explicite, mais les relations implicites sont aussi présentes.

5.4 Corpus Wikipédia

Comme son nom l'indique, ce corpus a été collecté via l'encyclopédie collaborative "*Wikipédia*". Les données représentent une description des communes de la France et ses rivières. La collecte s'est faite automatiquement grâce à l'API "*Mediawiki*" [9] et les noms de toutes les communes publiées par "*l'INSEE*" [10]. Les informations sont extraites sous la forme de page web PHP donc un prétraitement était requis pour les besoins de la Reconnaissance des Entité Spatiales dont le traitement exige des données sous un format bien spécifique. Ce choix du corpus a été influencé par la présence de la description géographique des communes, notamment la localisation comme le montre l'extrait sur la Figure 12.

Lille est située dans le nord de la France, au centre du département du Nord, à une dizaine de kilomètres de la frontière belge. De par sa proximité avec la Belgique, Lille se trouve également toute proche de la frontière linguistique qui, au nord de la ville, sépare la région flamande de la région wallonne. Ainsi, par exemple, le poste-frontière de Menin est flamand, tandis que ceux de Comines-Warneton et de Mouscron sont wallons.

Elle s'est établie dans la vallée de la Deûle dont plusieurs bras, aujourd'hui pour la plupart couverts, parcourent la ville. Naviguée depuis l'époque gallo-romaine, la rivière, aménagée récemment en canal à grand gabarit, traverse la ville du sud-ouest au nord pour rejoindre la Lys.

FIGURE 12 – Extrait du corpus Wikipedia

9. http ://www.mediawiki.org/wiki/API :Main_page/fr
10. www.insee.fr

5.4.1 Tests et Résultats

Plus de 36000 articles ont été collectés mais compte tenu du temps que requiert le processus de Reconnaissance des Entités Spatiales, nous n'avons testé l'approche que sur 1400 articles. De même que pour les tests précédents, l'apprentissage s'est déroulé en suivant le premier protocole. Quelques exemples des résultats obtenus au terme de ce test sont résumés dans la Table 12.

Dans ce nouveau et dernier apprentissage, nous remarquons que là encore, l'approche a pu extraire de nouveaux patrons, et à première vue, il semble que les relations sont plus explicites. De plus, la terminologie de ce corpus ressemble un peu a celle des relations spatiales présentes dans le premier car nous observons la présence des syntagmes annotés par la chaîne de Reconnaissances des Entités Spatiales plus que dans le deuxième corpus. Cependant, les résultats obtenus ici sont plus spécifiques que ceux du premier, en raison de la présence des lemmes dans les patrons dû à leurs fréquences élevées dans le corpus.

Patrons	*Enrichi*	*Nouveau*
{*TOPO*}{*V être*}{*V*}{*PREP à*}		✓
{*TOPO*}{*V être*}{*V situer*}{*PREP à*}{*distance*}	✓	
{*TOPO*}{*V être*}{*V rattacher*}{*PREP à*}		✓
{*V être*}{*V arroser*}{*PREP par*}{*TOPO*}		✓
{*PREP à*}{*orientation*}{*TOPO*}{*PREP dans*}	✓	
{*V situer*}{*PREP à*}{*orientation*}{*TOPO*}{*PREP*}	✓	
{*TOPO*}{*PRO se*}{*V trouver*}{*PREP*}		✓
{*DET le*}{*N*}{*V être*}{*V traversé*}{*PREP par*}{*TOPO*}		✓
{*TOPO*}{*V fait*}{*V partie*}{*PREP de*}		✓
{*TOPO*}{*V être*}{*V*}{*PREP à*}{*orientation*}	✓	
{*PREP*}{*distance*}{*PREP de*}{*TOPO*}	✓	
{*DET le*}{*ADV plus*}{*A proche*}{*V être*}{*TOPO*}	✓	
{*distance*}{*PREPDET au*}{*orientation*}{*TOPO*}	✓	
{*N*}{*V relier*}{*TOPO*}{*PREP*}		✓

Table 12 – Exemples de patrons extraits du corpus Wikipédia

5.4.2 Évaluation

Maintenant que nous avons les nouveaux patrons, nous allons les évaluer sur le corpus dont se sert le projet ANIMITEX. Nous illustrons donc dans la Table 13, quelques exemples des extractions obtenues.

Bien que ces derniers patrons soient pertinents, leur évaluation sur le corpus d'ANIMITEX n'a pas abouti aux résultats souhaités. Ce qui peut être justifié pas la différence entre les relations présentes dans Wikipédia et celles présentes dans le premier corpus. Celles de Wikipédia sont plus complexes, on peut même remarquer des combinaisons de syntagmes dans une même phrase, ce qui est très rare dans le corpus du Midi Libre. Dans

Patrons	**Fragments correspondants dans le texte**
{*PREP*}{*distance*}{*PREP de*}{*TOPO*}	*...lac de Cambelliès à 2 km de Loupian ...*
	...ce massif majestueux situé à seulement 30 km de Millau...
	...un sentier paysager de 12 km d' Aniane...
{*V*}{*inclusion*}{*TOPO*}{*CONJC et*}	*...Montgolfier situé entre Lafarge et Rhône ...*
	...va être construit entre Loupian et ouest de Bouzigues...
	...une zone de végétation située entre Villeveyrac et de Poussan...
{*TOPO*}{*PRO se*}{*V trouver*}{*PREP*}	*...Perpignan se trouve à 11 km de la mer ...*
	...le Port de Sète se trouve à Canet...
{*N*}{*V relier*}{*TOPO*}{*PREP*}	*...la D 60 reliant Frontignan à Frontignan-plage ...*
	..la RD5 E8 qui relie Villeveyrac à Mèze...
	...trois lignes reliaient Montpellier à Mazamet...
	...la voie ferrée reliant Alès à Bessèges...
{*distance*}{*PREPDET au*}{*orientation*}{*TOPO*}	*...de l' Italie à 50 km au sud de Bologne et à 30 km de Ravenne ...*

TABLE 13 – Évaluation des patrons extraits de Wipedia sur le corpus Midi Libre

ce dernier, les relations ne sont pas très fréquentes, et quand elles sont présentes, elles sont exprimées avec un lexique et une syntaxe très courants.

5.5 Impact des contraintes

Rappelons ici que les résultats obtenus tout au long de ces expérimentations ont suivi soit le premier protocole soit le deuxième (Section 5.2.1). Donc l'extraction s'est faite en respectant certaines contraintes relatives à l'algorithme. La contrainte de Gap a eu comme nous allons l'expliquer un impact non négligeable sur les résultats obtenus, c'est à dire les patrons extraits.

5.5.1 Le Gap

Cette contrainte permet de laisser un écart fixé par l'utilisateur entre les itemsets d'un motif candidat. Dans notre approche, elle permet de faire abstraction à certains termes dans les phrases au moment d'extraire les patrons. Pour être plus explicite, soient les deux séquences suivantes :

S1 =(V être) (V situer) (PREP à) (distance) (PREP de) (DET la) (N mer).
S2 =(V être) (V situer) (PREP à) (ADV seulement) (distance) (PREP de) (DET la) (N mer).

Supposons maintenant que ces deux séquences constituent une base de données de séquences (SDB) et que les paramètres d'extraction fixés sont comme suit : sup_{min}=2, un

Gap compris entre [0-0].

Si le motif candidat est *M= (V être) (V) (PREP à) (distance)* alors il ne sera pas extrait car *sup(M)=1* (il n'apparaît que dans la première séquence) ce qui ne vérifie pas la contrainte du support minimal. Mais avec un *Gap* compris entre [0-1], le motif serait extrait car *sup(M)=2* (il apparaît dans les deux séquences *S1* et *S2*). Autrement dit, avec ce *Gap*, l'algorithme d'extraction fait abstraction de l'adverbe "*seulement*". Nous pouvons de ce fait voir clairement que les résultats sont très influencés par cette contrainte.

La Table 13 illustre nettement cet impact. Nous pouvons dire que si nous avions fixé le *Gap* entre [0-0] nous n'aurions pas eu comme résultats les patrons qui sont marqués dans la table avec un *Gap* entre [0-2].

Patrons	*Gap[0-0]*	*Gap[0-2]*
{–*TOPO*}{*V être*}{*V situer*}{*PREP à*}{*distance*}	✓	
{*V être*}{*V arroser*}{*PREP par*}{*TOPO*}	✓	
{*PREP*}{*distance*}{*PREP*}{*TOPO*}	✓	
{*V*}{*inclusion*}{–*TOPO*}{*CONJC et*}	✓	
{*TOPO*}{*PRO se*}{*V trouver*}{*PREP*}	✓	
{*N*}{*V relier*}{*TOPO*}{*PREP*}		✓
{*TOPO*}{*V fait*}{*V partie*}{*PREP de*}		✓
{*DET le*}{*N sentier*}{*PRO qui*}{*V*}{*TOPO*}		✓
{*TOPO*}{*V*}{*PREP à*}{*A gauche*}{*DET*}		✓
{*TOPO*}{*PREP sur*}{*DET*}{*N droit*}		✓

Table 14 – Impact de la contrainte de Gap

5.5.2 La contrainte d'appartenance

A l'issu de tous les résultats obtenus, nous pouvons remarquer que mis à part l'ES, une relation doit obligatoirement contenir soit *une préposition* ou *un verbe* ou bien *un syntagme*. Nous pouvons prendre avantage de cette constatation pour diminuer le nombre de motifs extraits. Pour ce faire, nous utiliseront la contrainte d'appartenance facultative qui nous permet de spécifier quels sont les items qui doivent apparaître dans les motifs. Autrement dit, et suivant notre problématique, les patrons extraits doivent contenir au minimum un item qui doit être *une préposition* ou *un verbe* ou bien *un syntagme*. Cette contrainte permettra d'éviter d'avoir certains patrons non pertinents, notamment avec un Gap compris entre [0-2], comme par exemple{*TOPO*}{*CONJC*}{*DET*}{*N*}{*DET le*}. La Table 15 rapporte le nombre de motifs extraits dans le cas de l'utilisation de la contrainte d'appartenance facultative et leur nombre dans le cas contraire, c'est à dire sans l'utilisation de cette contrainte. Dans ce dernier tableau, nous pouvons remarquer qu'avec cette contrainte, le nombre de motifs a baissé mais dans certains corpus plus que dans d'autres. Ceci peut être expliqué par le lexique présent dans chacun des corpus et leurs syntaxes. Dans le second corpus, la structure des phrases est plutôt courte, donc avec le Gap, certains termes vont disparaître dans les motifs laissant beaucoup de vide et beaucoup de noms sans liens. En revanche, dans le premier il n'y a pas beaucoup de descriptions. Ici les phrases sont formelles et la majorité contient plus d'un verbe. Dans le dernier corpus,

il y a beaucoup de noms dans les phrases. Avec un Gap compris dans l'intervalle [0-2], les liens les combinant sont omis. Ce qui fait converger les résultats vers des motifs non intéressants, comme par exemple {*TOPO*}{*DET le*}{*N*}{*N*}.

Corpus	*Nbr motifs sans la contrainte*	*Nbr motifs avec la contrainte*
Midi Libre	48019	42353
Randonnées	13599	3817
Wikipédia	33012	8286

TABLE 15 – Impact de la contrainte d'appartenance facultative

5.5.3 Proposition d'une nouvelle contrainte : la pondération

Nous avons pu voir via ces expérimentations que la contrainte de la fréquence ne nous permettait pas d'extraire certains patrons bien que ces derniers soient pertinents. Ceci en raison de leur rare présence dans le corpus qui nous intéresse (Midi Libre). Mais après avoir testé l'approche sur d'autres corpus, nous sommes parvenus à extraire des relations qui n'étaient pas fréquentes dans le premier corpus. De là, nous pouvons déduire que certaines relations sont présentes dans le corpus mais la contrainte de la fréquence ne nous permet pas de les extraire. Autrement dit, certains items pertinents sont présents dans les séquences mais pas fréquemment. Hormis la fréquence, nous pouvons imaginer une autre mesure pour permettre l'extraction de ces items considérés comme importants dans les relations spatiales. En effet, par exemple si on collecte un lexique spécifique aux relations spatiales, nous pouvons le pondérer en fonction de sa pertinence dans les patrons que nous voulons extraire. Pour ce faire, nous pourrons exploiter la fréquence de ce lexique dans un corpus tel que Wikipedia ou celui des Randonnées, autrement, un corpus dont le contenu est riche en relations spatiales, pourrait être exploité et utilisé pour augmenter la valeur du support dans les séquences quand celle-ci contient un item pertinent. Cependant, bien que cette idée semble cohérente, elle n'est pour l'instant qu'une perspective à long terme que nous voudrions développer et adapter convenablement à l'algorithme. Sans omettre bien évidement la fréquence qui est très importante pour la bonne structuration des patrons.

5.6 Le lien avec ANIMITEX

Nous allons conclure ce rapport par l'explication de l'utilité des patrons extraits dans le projet ANIMITEX. Pour ce faire, nous donnons ici un exemple concret de comment pourrait-on exploiter ces relations spatiales.

Parmi les relations spatiales extraites grâce aux patrons à partir du premier corpus, certaines sont incluses dans des phrases où l'information qui est véhiculée peut porter sur des chantiers ou des incendies, comme par exemple :
-" *un feu s'est déclaré dans une zone de végétation située entre Villeveyrac et de Poussan*".
- *Un autre giratoire sera ensuite aménagé d'ici l'été à environ 1 km avant Mèze*".
- *Enfin un muret séparateur sera implanté entre Loupian et le giratoire ouest de Bouzigues*".

-...la Peyrade s'interrogent sur le Pont de Méréville en cours de construction au dessus de la RD 600...En se situant à 150 m de la Peyrade l'emplacement de cet ouvrage d'art d'une largeur de 8 m paraît déjà bien insolite".

Si nous apercevons, grâce à des processus de télédétection, des altérations sur une série d'images satellites, nous pourrons faire le lien avec un événement apparu dans les phrases extraites. Cet événement pourrait être par exemple un incendie, une installation de giratoire ou encore la construction d'un pont. Cependant, ce lien s'établit en s'appuyant sur l'endroit où l'altération a été constatée (exemple *à 150 de la Payrade*). En d'autre termes nous allons exploiter les relations spatiales existantes dans les phrases pour bien déterminer l'endroit. Nous devons aussi prendre en considération la date de l'apparition de l'information qui pourrait être récupérée à partir de la date de l'article publié, car les points communs entre les images et ces informations sont le temps, l'espace et l'événement. Ce dernier correspond à l'altération.

6 Conclusion et perspectives

Si l'Extraction d'Informations a suscité de l'intêret ces dernières années, c'est bien pour le rôle important qu'elle pourrait avoir sur différentes applications et le défi d'extraire des informations bien ciblées dû à la difficulté de la tâche. Dans notre cas, les travaux effectués ont fait l'objet d'extraction de relations spatiales à partir de documents textuels issus de la presse écrite. Ceci dans le but de compléter l'analyse des images satellites. Le présent rapport est ancré justement dans ce contexte. Pour mener à bien ces travaux de recherche, nous avons dans une première partie, présenté une veille exhaustive faisant référence à différentes méthodes qui ont contribué à l'Extraction de Relations pas forcément spatiales. Ces différentes méthodes peuvent être classées en deux principales catégories, à savoir, les approches à dominance TALN (Traitement Automatique du Langage Naturel) et celles à dominance fouille de données. Ces dernières sont à leur tours classées en trois sous catégories qui sont les méthodes supervisées, semi-supervisées et non-supervisés. Cette première partie a été conclue par une synthèse rapportant les principales caractéristiques de chaque approche.

La seconde partie fait état de description de l'approche adoptée dans nos travaux. Cette dernière est classée comme étant une méthode à dominance fouille de données et plus précisément une méthode non-supervisée en raison de son principe qui se base sur la fouille de données séquentielles, plus précisément, l'extraction de motifs séquentiels. Nous avons émis l'hypothèse que ce dernier principe peut être exploité pour des finalités TALN. En effet, les motifs extraits vont représenter des patrons linguistiques qui nous permettront d'extraire des relations spatiales à partir de données textuelles. Mais pour arriver aux résultats souhaités, l'approche proposée passe par de multiples étapes. La première consiste à identifier les ES via un traitement spécifique à la Reconnaissance des Entité Spatiales. Ensuite, nous avons établi un protocole de prétraitement permettant d'extraire les phrases susceptibles d'être porteuses de relations spatiales. De là, nous sommes passés à la lemmatisation. La dernière étape repose sur l'extraction des motifs séquentiels où nous devions judicieusement choisir les séquences sur lesquelles appliquer l'algorithme d'extraction de motifs séquentiels.

Une fois toutes les étapes achevées, nous avons testé l'approche dans un premier temps sur un corpus hétérogène pour les besoins du projet ANIMITEX. Dans ces

premières expérimentations, nous avons démontrés l'importance de la généralisation dans de tels corpus mais également l'importance de passer par un prétraitement judicieux. D'autres tests ont également été effectués sur deux autres corpus dont le contenu est spécifique aux relations spatiales, ceci dans le but de souligner l'importance de faire un apprentissage sur des données avec le moins de bruit possible.

L'une des perspectives à ne pas négliger est la validation des patrons par les experts, ce qui nous permettra probablement de valider plus de patrons et d'extraire ainsi d'autres relations. Il serait probablement intéressant d'identifier certains termes comme étant des ES comme la route, le sentier, le chemin, etc., comme l'a proposé Christian Sallaberry, membre du projet ANIMITEX. Dans un second temps, nous souhaiterons exploiter les méthodes supervisées, par exemple la classification afin de comparer les résultats avec ceux obtenus dans les présents travaux. Et Comme nous l'avons mentionné dans les expérimentations, une perspective à long terme serait de prendre en considération la contrainte de pondération et de l'intégrer dans l'algorithme de façon à respecter toutes les contraintes qui y sont définies.

Références

[1] Eugene Agichtein and Luis Gravano. Snowball : Extracting relations from large plain-text collections. *Proceedings of the fifth ACM conference*, pages 85–94, 2000.

[2] Rakesh. Agrawal and Ramakrishnan Srikant. Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.

[3] Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, and Nigel R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, pages 14–21, 2003.

[4] Hugo Alatrista-Salas and Nicolas Béchet. Fouille de textes : une approche séquentielle pour découvrir des relations spatiales le corpus. *Atelier CerGEO, dans le cadre de la conférence EGC*, 2014.

[5] Michele Banko, MJ Cafarella, and Stephen Soderland. Open information extraction for the web. *IJCAI International Joint Conference on Artificial Intelligence*, pages 2670–2676, 2007.

[6] Nicolas Béchet and Peggy Cellier. Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. *Actes des 23e journées francophones d'ingénierie des connaissances (IC 2012)*, pages 149–164, 2012.

[7] Christian Blaschke and Alfonso Valencia. Automatic ontology construction from the literature. *International Conference on Genome Informatics*, pages 201–213, 2002.

[8] Christian Blaschke and Alfonso Valencia. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, pages 14–20, 2002.

[9] Olivier Bodenreider. The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic acids research*, pages 267–270, 2004.

[10] Kurt Bollacker, Colin Evans, and Praveen Paritosh. Freebase : a collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1249, 2008.

[11] Eric Brill, Susan Dumais, and Michele Banko. An analysis of the AskMSR question-answering system. *In Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 257–264, 2002.

[12] Sergey Brin. Extracting patterns and relations from the world wide web. *Selected Papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, 1999.

[13] Razvan C. Bunescu and Raymond J. Mooney. Subsequence kernels for relation extraction. *The Neural Information Processing Systems*, 2005.

[14] Del Corro, Luciano and Rainer Gemulla. Clausie : clause-based open information extraction. *Proceedings of the 22Nd International Conference on World Wide Web*, pages 355–366, 2013.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, pages 273–297, 1995.

[16] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. *Proceedings of Annual Meeting on Association for Computational Linguistics*, pages 423–429, 2004.

[17] Kathrin Eichler, Holmer Hemsen, and G Neumann. Unsupervised Relation Extraction From Web Documents. *IN : Proceedings of Edition of the Language Resources and Evaluation Conference LREC*, pages 1674–1679, 2008.

[18] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, 2011.

[19] S. Ferré. Camelis : a logical information system to organize and browse a collection of documents. *Int. J. General Systems*, 38(4) :379–403, 2009.

[20] Mauro Gaio, Christian Sallaberry, and Tien Nguyen Van. Typage de noms toponymiques à des fins d'indexation géographique. *Traitement Automatique des Langues*, 53 :143–176, 2013.

[21] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. *Proceedings of Annual Meeting on Association for Computational Linguistics*, page 415, 2004.

[22] Johannes Hoffart, FM Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2 : A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, pages 28–61, 2013.

[23] Lodhi Huma, Shawe-Taylor John, Cristianini Nello, Watkins Chris, and Scholkopf Bernhard. Text classification using string kernels. *The Journal of Machine Learning Research*, pages 419–444, 2002.

[24] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 2004.

[25] Corrado Loglisci and Dino Ienco. Toward Geographic Information Harvesting : Extraction of Spatial Relational Facts from Web Documents. *The IEEE International Conference on Data Mining series (ICDMW) Workshops*, pages 789 – 796, 2012.

[26] Hirschman Lynette and Gaizauskas Rob. Natural language question answering : the view from here. *Natural Language Engineering*, pages 275–300, 2002.

[27] Eirinaios Michelakis, Rajasekar Krishnamurthy, Peter J. Haas, and Shivakumar Vaithyanathan. Uncertainty management in rule-based information extraction systems. *Proceedings of the international conference on Management of data*, pages 101–114, 2009.

[28] Raymond J. Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter*, pages 3–10, 2005.

[29] Kong Mu and Guo Qin. Improved Method of Relation Extraction Using Subsequence Kernel. *2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks*, pages 14–17, 2012.

[30] Ndapandula Nakashole, Gerhard Weikum, and Fabian M Suchanek. PATTY : A Taxonomy of Relational Patterns with Semantic Types. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145, 2012.

[31] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan : Mining sequential patterns by prefix-projected growth. pages 215–224, 2001.

[32] Jianyong Wang and Jiawei Han. Bide : Efficient mining of frequent closed sequences. page 79, 2004.

[33] Jianyong Wang, Jiawei Han, and Chun Li. Frequent closed sequence mining without candidate maintenance. *IEEE Transactions on Knowledge and Data Engineering*, pages 1042–1056, 2007.

[34] Bifan Wei, Jun Liu, Jian Ma, Qinghua Zheng, Wei Zhang, and Boqin Feng. Motif-based Hyponym Relation Extraction from Wikipedia Hyperlinks. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–13, 2013.

[35] Xifeng Yan, Jiawei Han, and Ramin Afshar. CloSpan : Mining Closed Sequential Patterns in Large Datasets. pages 166–177, 2003.

[36] Dmitry Zelenko, Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning*, pages 1083–1106, 2003.

[37] C Zhang, X Zhang, and Wenming Jiang. Rule-based extraction of spatial relations in natural language Text. *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference*, pages 1–4, 2009.

[38] Xueying Zhang, Chunju Zhang, Chaoli Du, and Shaonan Zhu. SVM based extraction of spatial relations in text. *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, pages 529–533, 2011.

[39] Shubin Zhao and Ralph Grishman. Extracting relations with integrated information using kernel methods. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 419–426, 2005.

[40] D Zhou, S Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 427–434, 2005.