Utilisation de la relation ” Verbe - Préposition - Toponyme” pour un inventaire lexical automatique

Tien Nguyen Van, Mauro Gaio

UTILISATION DE LA RELATION « VERBE – PREPOSITION – TOPONYME » POUR UN INVENTAIRE LEXICAL AUTOMATIQUE

Van Tien NGUYEN, Mauro GAIO
Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex
vantien.nguyen@univ-pau.fr, mauro.gaio@univ-pau.fr

Résumé : Nous proposons une approche, permettant à partir d'un modèle, d'extraire et d'interpréter des informations à connotation géographique à partir d'une analyse automatique d'un corpus de textes littéraires (récits de voyages dans les Pyrénées au XIX^{e} siècle). Il s'agit de la combinaison d'une approche lexico-syntaxique permettant le marquage et l'interprétation d'expressions contenant au moins une entité nommée géographique avec une analyse grammaticale ciblée impliquant des verbes de déplacement (ou de perception) permettant le marquage d'expressions de mouvement et d'expressions spatiales. L'inventaire lexical obtenu à l'aide de cette démarche est ensuite exploité à des fins d'enrichissement d'une ontologie géographique construite par l'IGN.

Mots-clés : extraction de concepts, modélisation spatiale, lexique géographique, grammaire hors contexte, enrichissement d'ontologie.

1 Introduction

L'un des buts du projet GEONTO[1] est de créer une ontologie initiale spécifique au domaine géographique puis de l'enrichir de manière automatique. L'ontologie initiale a donc été créée en collaboration avec l'équipe de recherche du COGIT de l'IGN impliquée dans le projet. Dans le cadre de cet article, nous présentons une méthode afin de réaliser automatiquement, à partir d'un ensemble de textes, un inventaire lexical potentiellement à connotation géographique. Cet inventaire devant par la suite servir à enrichir les concepts de l'ontologie ci-dessus énoncée. Nous nous sommes exclusivement intéressés à des situations dans lesquelles le mot ou le groupe de mots se retrouvent à proximité d'une entité nommée géographique et sont impliqués dans une relation de dépendance grammaticale avec un verbe de déplacement (ou verbe de perception) et éventuellement avec une préposition.

La problématique est détaillée en -2-. En -3- nous discutons des travaux existants relatifs au traitement automatique de langue (TAL) et les ressources lexicales. Notre méthode et le résultat d'expérimentation sont présentés en -4- et -5-.

2 La problématique

Le lexique à constituer doit être obtenu à partir de l'extraction des syntagmes nominaux employés pour leur connotation géographique (territoire aride, au sud de l'étroite vallée,…) dans le fonds documentaires constitués de plusieurs centaines de récits de voyage. Afin d'opérer automatiquement cette extraction de manière ciblée, il faudrait disposer des modèles permettant de différencier les syntagmes à connotation géographique parmi tous ceux contenus dans des textes. Ci-après quelques extraits de notre corpus, afin d'illustrer nos propos :

« […] Depuis quelques temps une vive curiosité avait porté mes regards vers la Maladetta[. . .] Je parlai de mes intentions à plusieurs guides de Luchon […]

*[…]Après avoir contemplé, avec une admiration mêlée d'effroi, la **charpente altière** des MontsMaudits, nous songeâmes bientôt à descendre sur le **territoire aride** au sud de la **région** d'Aragon. Le temps était menaçant : de légers brouillards parcouraient les **hauteurs**, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant de l'ouest des Pyrénées, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabineros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement, vers le **pied** de la Maladetta, laissant à notre droite les **roches calcaires** de la Pèna-Blanca. Arrivés au fond de la **vallée** du Plan-des-Etangs, qui est plus élevée que sa voisine, la **vallée latérale** de l'**hospice** de Bagnères, de 446 mètres, nous laissâmes derrière nous une **cabane** habitée pendant l'été par des bergers espagnols, pour remonter, par un **plan***

[1] GEONTO http ://geonto.lri.fr/, est un projet ANR (ANR-07-MDCO-005-04).

***rocailleux**, jusqu'au **gouffre** de Tourmon, qui absorbe les **eaux d'un torrent rapide**, descendant de la partie orientale du **glacier** de la Maladetta[…] »*

Comme nous pouvons le constater dans ces exemples, les termes à extraire sont très souvent associés à des entités nommées géographiques. Cette observation est corroborée par les travaux de Vandeloise (1986) sur le couple (cible, site) et de Borillo (1998) sur le couple (entité concrète, repère spatial). Toutefois, si cette observation est intéressante, elle reste incomplète car des expressions considérées comme non géographiques peuvent également être associées à une entité nommée de lieu (ex : guides de Luchon, mes regards vers la Maladetta), des expressions géographiques peuvent exister sans être associées à un nom de lieu comme « cabane » et d'autre part. L'étude de notre corpus a permis d'observer que très fréquemment ce couple se trouve en relation, au sein d'une même phrase, avec des verbes de déplacement (« remonter jusqu'au gouffre de Tourmon ») ou des verbes de perception (« contempler la charpente altière des Monts-Maudits »). Enfin, dans plusieurs cas la construction de ce couple fait appel à des relations spatiales afin de faire référence à un lieu complexe (« descendre sur le territoire aride au sud de la région d'Aragon »).

Nous proposons donc un modèle et son opérationnalisation afin de permettre, par la prise en compte de ces observations, un traitement automatique. Cela nous mène à proposer le modèle VPT, des détails de ce modèle et son opérationnalisation sont ensuite proposés.

3 Etat de l'art et travaux connexes

Le gazetteer et le problème de détection des entités nommées

La détection des entités nommées géographiques, et de manière plus générale des entités nommées (personnes, entreprises,…) est une problématique reconnue comme jouant un rôle important dans nombreux traitements automatiques de la langue Sagot et al.(2008) et notamment dans le cas de l'extraction automatique d'information Poibeau(2003). Nous nous intéressons exclusivement aux entités nommées géographiques que nous considérons comme étant un groupe nominal dont le noyau est un nom de lieu ou nom toponymique. Dans de nombreux travaux comme par exemple Rocío et al.(2010), ou au sein de notre équipe Loustau et al.(2008), Palacio(2010), la détection entités nommées géographiques est élaborée à l'aide des gazetteers[2]. Il existe plusieurs gazetteers accessibles par Internet tels que : Geonames, BDNyme, Word Gazetteer, GEOnet Names Serve (GNS) [3],…

Dans notre travail, les noms toponymiques une fois repérés permettent de déclencher le processus de marquage/interprétation des expressions évoquant la ou les relations spatiales ainsi que le syntagme verbal de déplacement ou de perception afin de construire une structure en traits sémantiques permettant d'isoler les informations à extraire.

Expression spatiale dans le texte et la modélisation spatiale

Selon Borillo(1998), un lieu est une portion de l'espace matériel dans lequel nous nous situons et nous évoluons. Nous considérons donc que dans l'expression « la partie orientale du glacier de la Maladetta », le repère spatial peut être déduit de la même manière que dans l'exemple précédent via le nom toponymique « Maladetta » et l'entité concrète est incarnée ici par l'expression « partie oriental du glacier ». Cette expression contient une précision de localisation (que appelons indirection) au sein de l'entité concrète. Cette précision est exprimée via une relation spatiale nécessitant une interprétation.

Il existe deux types d'approches permettant de raisonner et donc d'interpréter au travers de relations spatiales : des approches quantitatives comme celle proposée par Balbiani et al.(2000) ou par Vieu(1997), et des approches qualitatives telles que celle proposée par Allen(1991), ou par Freksa(1992) ou encore par Frank(1996). Les approches quantitatives prennent en compte les aspects mesurables relatifs aux lieux tels que la longitude et la latitude, tandis que les approches qualitatives

---

[2] Un gazetteer est un dictionnaire ou répertoire géographique dont les entrées sont des noms toponymiques. A chaque entrée du dictionnaire peuvent être associées des informations comme l'appartenance à une ou plusieurs structures administratives (commune, région, pays,…), la caractéristique physique (montagne, rivière, route,…), des données statistiques, une géométrie exprimée dans un référentiel géographique.

[3] http://geonames.org, http://www.ign.fr, http://www.world-gazetteer.com/, http://earth-info.nga.mil/gns/html/

opèrent sur des représentations symboliques. Selon ces approches, les relations spatiales peuvent être catégorisées en trois classes principales : topologiques comme décrite par Egenhofer et al.(1991) (ex : dans, à l'intersection, etc.), directionnelles formalisées par Ligozat(1998)) (ex : au sud de, etc.), et métriques (ex : à 10km de, etc.). Afin d'obtenir une représentation automatique proche du lieu nous prenons en compte l'évocation des relations spatiales grâce à une approche hybride Gaio et al.(2008).

Expression de déplacement

Selon Talmy(2000), dans les langues latines comme le français, le mouvement est caractérisé par le verbe. Dans notre corpus, d'après une étude réalisée dans notre équipe par Loustau et al.(2008), l'expression du déplacement est essentielle dans un récit de voyage. Plusieurs travaux linguistiques comme ceux de Boons(1987), de Laur(1991) et de Sarda(2000) ont été réalisés afin d'étudier le rôle des verbes de déplacement dans la langue. Ces auteurs ont proposé une catégorisation des verbes de déplacement via leur polarité. En synthèse nous dirons que les polarités sont : initiale (ex : quitter), médiane (ex : visiter), ou finale (ex : arriver). D'autre part, dans un écrit, en particulier dans un récit de voyage, lorsque le narrateur souhaite rendre compte de certaines actions ou sensations, les verbes de perception (ex : voir) acquièrent une importance particulière.

TAL et la grammaire hors contexte

A des fins d'extraction d'information, il est indispensable d'utiliser les outils de TAL. Ces outils permettent de traiter les textes sur différents niveaux. Pour le prétraitement du corpus, nous avons besoin d'une analyse morphosyntaxique de texte. Pour cette étape des outils tels que TreeTagger Schmidt(1994) et Melt Denis et al.(2009)) peuvent être utilisés[4].

Les grammaires hors contexte[5] sont souvent utilisées en TAL. Ces grammaires se composent d'un ensemble de règles qui permettent de remplacer une séquence d'expression (nom, adjectif, verbe, etc.) par un nouvel identifiant unique d'un niveau d'abstraction plus élevé (syntagme nominal, syntagme verbal, etc.). Dans le cas de ce travail, la grammaire hors contexte est utilisée pour marquer non seulement des informations à un niveau d'abstraction syntaxique plus élevé (groupes de noms propres, groupes de nom communs) mais également à un niveau sémantique (ex : verbe de déplacement, nom toponymique, etc.) grâce à l'utilisation combinée de ressources lexicales hétérogènes.

4 Opérationnalisation

4.1 Le modèle Verbe-Préposition-Toponyme (VPT)

Le modèle tel que schématisé dans la fig.1 combine de manière parcimonieuse les travaux précédemment évoqués relatifs à l'expression spatiale dans la langue, aux relations spatiales et au lexique verbal. Ce modèle décrit un triplet (VPT) qui se compose d'un verbe en général de déplacement mais également de perception (V), d'une préposition (P), et d'un Toponyme (T). Le Toponyme est défini de façon récursive à partir des noms toponymiques, des relations spatiales (ou indirections), et des termes associés.

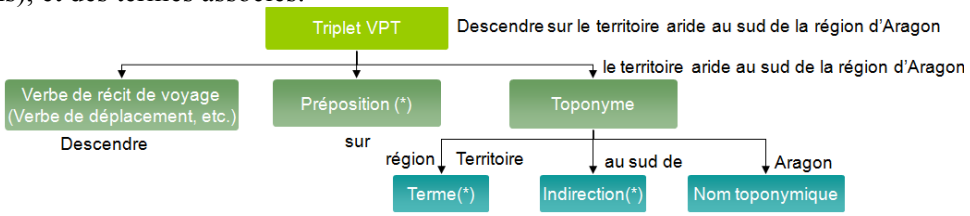

FIG. 1 – Schématisation du modèle proposé

[4] Pour la version actuelle de notre chaîne de traitement (section 4), nous utilisons TreeTagger. Toutefois, cet analyseur produit des erreurs dans certains cas (section 5). Nous envisageons donc de tester Melt dans une version ultérieure en espérant qu'il soit plus robuste et permette de réduire certaines erreurs.

[5] Formellement, un langage est hors-contexte si et seulement si il existe un automate à pile qui le reconnaît.

Le caractère étoile (*) dans la fig.1 signifie que le composant correspondant pourra être présent zéro ou plusieurs fois. Les autres composants doivent y apparaître au moins une fois. Voici quelques exemples extraits du corpus :
– remonter à Gavarnie – contempler la charpente altière des Monts-Maudits – remonter jusqu'au gouffre de Tourmon – arriver au fond de la vallée du Plan-des-Etangs – franchir l'arête occidentale de la Frondella au petit col Wallon – passer sur le versant de Cauterets par la brèche de Courouaou de Bouc – etc.

Comment construire automatiquement ce modèle VPT ? Et comment sert-t-il à extraire un lexique à connotation géographique ? Cela est réalisé par une chaîne de traitement complète au sein de laquelle nous avons défini une grammaire et utilisé ou construit diverses ressources lexicales : liste de verbes de déplacement et de perception avec leur polarité, des gazetteers, liste d'expressions évoquant des relations spatiales et leur correspondance dans le modèle hybride proposé par Gaio et al.(2008), ontologie de concepts topographiques, thésaurus de termes pour l'indexation documentaire.

4.2 La chaîne de marquage des triplets VPT

L'objectif (fig.2) est de marquer les triplets VPT, puis d'en extraire des expressions selon un certain filtre.

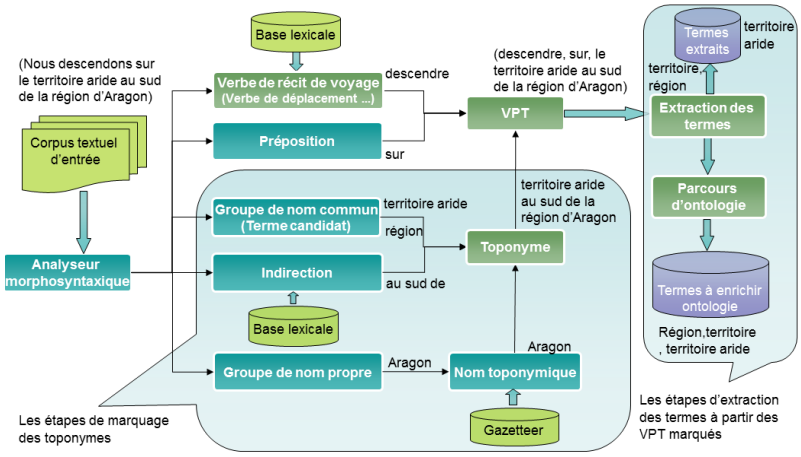

FIG. 2 – Illustration de la chaîne de marquage des triplets VPT

L'entrée de notre chaîne correspond à un texte dont les mots ont été soumis au préalable à un étiquetage morphosyntaxique. Seuls les groupes de mots reconnus selon un ensemble de patrons obtiennent un sur-étiquetage soit par extension du nom (propre ou commun), soit par détection de la sémantique (relation spatiale, verbe de déplacement ou de perception, nom de lieu), puis selon des règles de composition le toponymique est isolé et par le suite le triplet VPT. Les patrons sont construits grâce à des règles de la grammaire hors-contexte et aux ressources lexicales que nous allons successivement étudier en détail.

4.3 La grammaire

Dans notre chaîne de traitement, les étiquettes sont assignées au fur et à mesure. En effet, un groupe de noms communs, ou terme candidat : « territoire aride », est marqué à partir des noms communs « territoire » et des adjectifs « aride » préalablement étiquetés par l'analyseur morphosyntaxique. La fig.3 présente notre grammaire de marquage de 4 cas distincts de la catégorie « groupe de nom commun » :

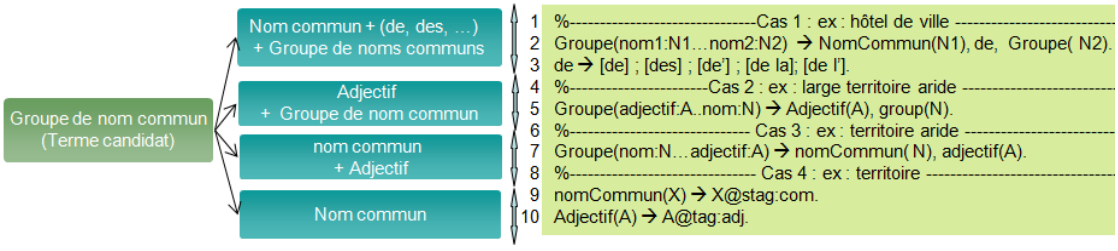

FIG. 3 – La grammaire de marquage des groupes de noms communs

De même, « un groupe de nom propres » est repéré par des noms propres précédemment étiquetés par l'analyseur morphosyntaxique, par exemple Mont de Marsan :
groupePropre(GP) > nomPropre(NP1), de, nomPropre(NP2).

Ensuite, les groupes de noms propres sont cherchés dans des gazetteers afin d'être validés comme nom toponymiques. Dans l'étape suivante, les Toponymes sont étiquetés, voici une des règles :
Toponyme(T) > groupeNomCommun(G1), indirection(I), groupeNomCommun(G2), de, nomToponymique(NT)

À droite de la règle, les groupes de nom commun G1 « territoire », G2 « région », le nom toponymique NT « Aragon », les indirections « au sud de » sont précédemment étiquetés. Dans ce cas, il s'agit d'un Toponyme dit complet « territoire aride au sud de la région d'Aragon ». Nous définissons ainsi des règles pour des Toponymes partiels :

> Nom toponymique — *le Gave de Pau* ; syntagme nominal associé au nom toponymique — *dans les plaines d'Espagne*; groupe de nom commun + indirection + nom toponymique —*nos logements respectifs à Bagnères-de-Luchon;* Indirection + syntagme nominal + nom toponymique — *au sud de la vallée du Plan-des-Etang.*

Enfin les triplets VPT (*descendre sur le territoire aride au sud de la région d'Aragon*) sont étiquetés à partir des verbes « descendre », des prépositions « sur) »et des Toponymes « territoire aride au sud de la région d'Aragon », voici un des règles :
VPT(verbe :V...pre :P...toponyme :T) >Verbe(V),Preposition(P),Toponyme(T).

Dans celle-ci, les verbes sont préalablement marqués à l'aide d'une base lexicale. Le marquage des triplets VPT dépend de la nature du verbe (transitif direct, transitif indirect ou intransitif). Dans ce cas, il s'agit d'un triplet VPT dont le verbe est associé aux prépositions.

A l'heure actuelle les règles de notre grammaire couvrent 4 cas de groupes de noms communs, 14 cas de groupes de noms propres, 10 cas de toponymes, et 15 cas de triplets VPT.

4.4 Les ressources utilisées

Nous avons construit une base lexicale des verbes de 75 verbes de déplacement, construite en nous appuyant sur les travaux précédemment cités. Afin d'étendre la capacité de repérage d'un lexique à connotation géographique nous y avons ajoutons 29 verbes de perception et nous étudions actuellement l'intérêt de traiter également 59 verbes qui, compte tenu de leur contexte d'utilisation nous nommons par convenance : *topographiques* « nous nous **abattions** péniblement sur les versants qui dominent l'hospice de Bénasque et la vallée de l'Essera ». Le tab.1 présente la distribution des verbes plus fréquents dans les corpus en fonction de leur catégorie. La deuxième colonne des tableaux indique le nombre de fois où le verbe est associé au triplet VPT. La dernière colonne est le pourcentage de cette association par rapport au nombre d'occurrences du verbe dans tout le corpus étudié.

Verbe	VPT	Taux
arriver	127	19.4%
passer	114	14.4%
aller	109	10.0%
venir	80	9.9%
descendre	79	14.7%
partir	79	21.5%
monter	78	12.8%
conduire	77	25.9%
quitter	51	23.1%
suivre	46	8.1%

(a) Verbe de déplacement

Verbe	VPT	Taux
voir	121	8.5%
apercevoir	26	12.6%
reconnaître	20	8.1%
paraître	17	3.5%
jeter	16	11.1%
découvrir	16	8.7%
aimer	13	8.5%
explorer	11	15.5%
coucher	11	8.3%
entendre	10	5.7%

(b) Verbe de perception

Verbe	VPT	Taux
plonger	13	16.4%
élever	45	8.8%
dominer	35	13.9%
séparer	31	15.0%
situer	25	20.3%
rencontrer	24	11.1%
prolonger	22	23.9%
étendre	20	10.1%
naître	18	20.0%
placer	18	11.3%

(c) Verbe dit topographique

TAB.1 – Classification et distribution des verbes plus fréquents dans les corpus

Chaque verbe dans la base est décrit par deux informations importantes : la catégorie du verbe (verbe de déplacement, verbe de perception, etc), et une forme de « polarité ». Concernant les verbes de déplacement celle-ci est de trois type « initiale », « médiane » et « finale ». Pour les verbes dits de perception nous avons considérés qu'ils se comportaient de manière équivalente aux verbes de déplacements médians. Enfin, pour la dernière famille de verbes une étude empirique est en train d'être menée. Dans la construction des triplets VPT, nous distinguons deux types de verbes : verbes

associés aux prépositions (e.g. verbes transitifs indirects), et verbes non associés à une préposition (e.g. verbes transitifs directs). Concernant les verbes de déplacement, ils peuvent être transitifs directs (visiter, traverser,…), ou transitifs indirects (aller, arriver, venir,…), alors que tous les verbes de perception ont été considérés comme étant transitifs directs (voir, contempler,…).

Enfin, les indirections du modèle VPT sont repérées à l'aide d'une base lexicale construite à partir des relations spatiales modélisées selon l'approche hybride Gaio et al. (2008) on distinguera : les relations topologiques, les relations directionnelles et les relations métriques discrétisées pour être traitées comme une combinaison des deux précédentes. Afin de valider les groupes de noms propres comme les noms toponymiques, nous utilisons deux
gazetteers : BDNyme de l'IGN qui comporte 44315 noms de lieu français, et Geonames qui en contient 118301.

4.5 Quelques éléments sur l'enrichissement

Comme déjà mentionné l'objectif premier de ce travail consiste à utiliser des textes grand public pour l'enrichissement d'une ontologie de domaine spécifique. Après avoir extrait à partir des triplets VPT marqués, les termes trouvés sont comparés avec les termes utilisés comme label de concept dans l'ontologie de référence. Comme montré dans la fig.2, cette tâche est réalisée par le module *parcours d'ontologie* qui vérifie si un terme existe dans une ontologie. Sinon, il est retenu comme candidat à l'enrichissement. À cette étape, soit l'enrichissement est semi-automatique le terme est alors proposé à l'expert afin qu'il choisisse le meilleur emplacement dans l'ontologie pour son insertion. Soit l'enrichissement est automatique dans ce cas nous nous appuyons sur une ressource tierce générique (tel que wordnet ou un thésaurus générique tel que RAMEAU de la BnF) afin de déduire via les relations de subsomption l'emplacement le plus adéquat pour son insertion dans l'ontologie. Dans notre exemple les trois concepts « région », « territoire », et « territoire aride » sont candidats à enrichir l'ontologie initiale de l'IGN proposée par Abadie et al.(2010).

5 Expérimentation et évaluation

5.1 Evaluation quantitative

Nous avons expérimenté notre méthode sur 12 livres ce qui fait un total de 2400 pages environ, fournis par la médiathèque de Pau (MIDR). Le tab. 2 présente quelques exemples. Le tab.3 indique l'apport de l'utilisation des verbes de perception et des verbes dits topographiques tandis que la précision sur chaque type de verbe reste stable. Parmi 323 termes extraits et validés par des experts (soit 1137 occurrences dans le corpus), 260 termes n'existent pas dans l'ontologie de l'IGN, 119 étant des termes composés comme : « débouché des ports », « panorama des cimes », « embranchement des routes »,…

Terme	Nombre d'extractions	Validé
vallée	117	Oui
lac	83	Oui
port	41	Oui
col	35	Oui
pic	34	Oui
route	23	Oui
pont	22	Oui
village	19	Oui
sommet	19	Oui
gave	17	Oui

(a)

Terme	Nombre d'extractions	Validé
frontière	1	Oui
camarades	1	Non
débouché des ports	1	Oui
direction	1	Non
traversée	1	Oui
grote	1	Oui
monts	1	Oui
jour	1	Non
col des pierres	1	Oui
malades	1	Non

(b)

TAB.2 – Nombre total d'apparitions pour quelques termes extraits : (a) > 10 fois ; (b) une fois

Type de verbe	Nombre d'extractions de termes valides	Précision
Verbe de déplacement	701	0,81
Verbe de perception	178	0,78
Verbe dit topographique	258	0,83
Tous	**1137**	**0,81**

TAB.3 – La précision sur les corpus expérimentés

L'extraction manuelle de termes à partir de corpus exige un important travail. L'intérêt principal de notre méthode est donc l'automatisation de ce travail d'extraction et sa capacité à être utilisée sur des corpus de taille très importante.

5.2 Quelques exemples de bruit et de silence

Ci-après quelques cas, illustrés par un exemple, dans lesquels les termes extraits n'ont pas une connotation géographique.

1) *Depuis que j'ai quitté le confort de la vie de Bordeaux, je trouve […]*
   Cas de polysémie des verbes, ici le verbe « quitter ».
2) *Il serait devenu un peu fier vis-à-vis de ses camarades d'Arrens[…]*
   Cas d'erreurs générées par les pré-traitements. Ici un faux étiquetage de l'analyseur morphosyntaxique, l'étiquette « verbe voir » a été donnée à « vis » dans « vis-à-vis ».

En analysant les résultats de l'expérimentation, nous avons également détecté des cas dans lesquels les termes à connotation géographique n'ont pas été extraits, ci-après quelques exemples :

1) *Pour ce qui est des variations du niveau du gouffre, il y a, en effet, une crue et une baisse[…]*
   Cas où le contexte phrastique est tel qu'il n'existe aucune indication qui permette d'identifier le terme.
2) *[…]le pic de Néthou , n'a été encore gravi par personne[…]*
   Cas où un pré-traitement supplémentaire est nécessaire. Ici transformation de la structure passive.
3) *Avant d'arriver à la fin de la vallée, nous traversons le bras de la Garonne et nous grimpons sur le plateau d'Esquierry[…]*
   Cas d'incomplétude des ressources. Ici « Esquierry » est un nom toponymique valide toutefois il n'est pas répertorié dans les gazetteers utilisées.

6 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode de modélisation et son opérationnalisation pour permettre de réaliser automatiquement un inventaire lexical à connotation géographique à partir d'un fonds documentaire. Pour cela, nous nous appuyons, d'une part, sur des lexiques et une structure locale permettant de modéliser l'information géographique contenue dans des textes, et d'autre part sur un ensemble de règles construites grâce à une grammaire hors contexte, ces trois aspects sont opérationnalisés au sein d'une chaîne automatique permettant de traiter en entrée un corpus de taille quelconque.

Notre méthode offre deux avantages majeurs (1) chaque élément du modèle est marqué par un module. Cela permet de traiter les cas complexes pour chaque élément du modèle avant de les rassembler. (2) les règles peuvent traiter des configurations complexes pouvant par exemple comporter plusieurs verbes, plusieurs noms toponymiques ou plusieurs triplets VPT.

Le patron VPT que nous avons proposé peut également être utilisé dans un but de détection de noms toponymiques non encore répertoriés dans les ressources. Par exemple, pour la phrase « nous grimpons sur le **plateau d'Esquierry** », le triplet VPT (grimper , sur, le plateau d'Esquierry) sera marqué, et on supposera dans un premier temps que « Esquierry » est un nom toponymique. Dans un second temps, on vérifiera si le terme « plateau » existe dans l'ontologie géographique enrichie, si tel est le cas « Esquierry » sera considéré comme un nom toponymique validé.

Dans un futur proche, nous avons l'ambition d'étendre notre méthode pour pouvoir extraire également des termes non directement attachés aux triplets VPT, par exemple, le terme « ville » dans la phrase « Jusqu'à cette **ville**, nous avons longé la Garonne ». Pour le traitement de tel cas, nous proposons un modèle basé sur les relations n-aires dont VPT fait partie.

Références

ABADIE N. & MUSTIÈRE S. (2010). Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. RIG, 20(2), 145–174.

ALLEN J. F. (1991). Planning as temporal reasoning. KR, 3–14.

BALBIANI P. & MULLER P. (2000). Le raisonnement spatial. Le temps, l'espace et l'évolutif en sciences du traitement de l'information. Cepadues Editions.

BOONS J.-P. (1987). La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. LANGUE FRANÇAISE, 76(76), 5–40.
BORILLO A. (1998). L'espace et son expression en français, L'essentiel. Orphrys.
BRODEUR J. (2004). Interopérabilité des données géospatiales : Élaboration du concept de proximité géosémantique. PhD thesis, U. Laval, Québec, CA.
DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In Proceedings of PACLIC 2009, Hong Kong, China.
EGENHOFER M. & R.D. F. (1991). Point-set topological spatial relations. IJGIS, 5(2), 161–174.
FRANK A. U. (1996). Qualitative spatial reasoning : Cardinal directions as an example. IJGIS, 10(3), 269–290.
FREKSA C. (1992). Using orientation information for qualitative spatial reasoning.
GAIO M., SALLABERRY C., ETCHEVERRY P., MARQUESUZAÀ C. & LESBEGUERIES J. (2008). A global Process to Access Documents' Contents from a Geographical Point of View. JVLC, 19(1), 03–23.
LAUR D. (1991). Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple. PhD thesis, U. Toulouse II, FR.
LIGOZAT G. (1998). Reasoning about cardinal directions. Visual Languages and Computing, 9(1).
LOUSTAU P., NODENOT T. & GAIO M. (2008). Spatial decision support in the pedagogical area: Processing travel stories to discover itineraries hidden beneath the surface. In 11th AGILE. 340–359, Girona, ESP.
PALACIO D. (2010). Combinaison de critères par contraintes pour la Recherche d'Information Géographique. PhD thesis, U. de Pau et des Pays de l'Adour, FR.
POIBEAU T. (2003). Extraction automatique d'information. Hermès Lavoisier.
ROCÍO A.-M. & ERICK L.-O. (2010). Geo information extraction and processing from travel narratives. In Transforming the Nature of Communication, 14th ICE, 363–373, Helsinki, FIN.
SAGOT B. & BOULLIER P. (2008). Sxpipe2 : architecture pour le traitement présyntaxique de corpus bruts. TAL, 49(2), 155–188.
SARDA L. (2000). L'expression du déplacement dans la construction transitive directe. Syntaxe et Sémantique, 121–137.
SCHMIDT H. (1994). Probabilistic part-of-speech tagging using decision trees. In ICNMLP, Manchester, UK.
TALMY L. (2000). Toward a Cognitive Semantics, chapter How language structures space. The MIT Press.
UITERMARK H. (2001). Ontology-Based Geographic Data Set Integration. PhD thesis, U. Twente, NL
VANDELOISE C. (1986). L'espace en français. Paris, France, Seuil.
VIEU L. (1997). Spatial representation and reasoning in artificial intelligence. In STR, 3–41.

Van Tien NGUYEN, Mauro GAIO

Université de Pau et des Pays de l'Adour