Découverte de nouvelles entités et relations spatiales à partir d'un corpus de SMS

Sarah Zenasni[1, 3] Eric Kergosien[2]
Mathieu Roche[1, 3] Maguelonne Teisseire[1, 3]
(1) UMR TETIS (IRSTEA, CIRAD, AgroParisTech), Montpellier, France
(2) GERiiCO, Univ. Lille 3, France
(3) LIRMM, CNRS, Univ. Montpellier, France
{sarah.zenasni, mathieu.roche, maguelonne.teisseire}@teledetection.fr,
eric.kergosien@univ-lille3.fr

RÉSUMÉ

Dans le contexte des masses de données aujourd'hui disponibles, de nombreux travaux liés à l'analyse de l'information spatiale s'appuient sur l'exploitation des données textuelles. La communication médiée (SMS, tweets, etc.) véhiculant des informations spatiales prend une place prépondérante. L'objectif du travail présenté dans cet article consiste à extraire ces informations spatiales à partir d'un corpus authentique de SMS en français. Nous proposons un processus dans lequel, dans un premier temps, nous extrayons de nouvelles entités spatiales (par exemple, *motpellier*, *montpeul* à associer au toponyme *Montpellier*). Dans un second temps, nous identifions de nouvelles relations spatiales qui précèdent les entités spatiales (par exemple, *sur, par, pres,* etc.). La tâche est difficile et complexe en raison de la spécificité du langage SMS qui repose sur une écriture peu standardisée (apparition de nombreux lexiques, utilisation massive d'abréviations, variation par rapport à l'écrit classique, etc.). Les expérimentations qui ont été réalisées à partir du corpus 88milSMS mettent en relief la robustesse de notre système pour identifier de nouvelles entités et relations spatiales.

ABSTRACT

Discovering of new Spatial Entities and Relations from SMS

Within the context of the currently available data masses, many works related to the analysis of spatial information are based on the exploitation of textual data. Mediated communication (SMS, tweets, etc.) conveying spatial information takes a prominent place. The objective of the work presented in this paper is to extract the spatial information from an authentic corpus of SMS in French. We propose a process in which, firstly, we extract new spatial entities (e.g. *motpellier*, *montpeul* associate with the place names *Montpellier*). Secondly, we identify new spatial relations that precede spatial entities (e.g. *sur, par, pres,* etc.). The task is very challenging and complex due of the specificity of SMS language which is based on weakly standardized writing (lexical creation, massive use of abbreviations, textual variants, etc.). The experiments that were carried out from the corpus 88milSMS highlight the robustness of our system in identifying new spatial entities and relations.

MOTS-CLÉS : Entités spatiales, Relations spatiales, Mesure de Similarité, Corpus de SMS.

KEYWORDS: Spatial Entities, Spatial Relations, Similarity Measure, SMS Corpus.

1 Introduction

Au cours de ces dernières années, de nombreux travaux liés à l'analyse de l'information spatiale ont été réalisés sur différents types de données volumineuses, que ce soit les images satellites, les données GPS, ou encore les données textuelles. Concernant les données textuelles, l'extraction d'entités spatiales (ES) a été appliquée sur divers types de corpus tels que les articles de presse, les tweets, etc. La plupart des travaux s'appliquent à identifier et extraire les ES simples correspondant à un nom toponymique ou nom de lieu (par exemple *Paris*, *gare Montparnasse*). Des recherches plus récentes se concentrent sur la reconnaissance des ES plus complexes intégrant des relations spatiales (par exemple *au nord de Paris*, *près de l'église Saint-Paul*). Ces derniers travaux ont pour objectif de préciser l'ES extraite (notamment sa localisation), ou encore de valider des ES lorsqu'une ambiguïté existe sur le type de l'entité nommée extraite (lieu, organisation, etc.). Le développement des technologies de communication a contribué à l'émergence de nouvelles formes de textes écrits que les scientifiques doivent étudier en fonction de leurs particularités. La communication par SMS (Short Message Service), notamment, est devenue un phénomène social. Des millions de messages sont échangés chaque jour pour communiquer, participer à des concours, obtenir des informations (nom de lieu, localisation, etc.), etc. En raison des spécificités du langage SMS (création lexicale, utilisation massive d'abréviations, présence de fautes, etc.), il est difficile d'utiliser ces données qui sont, par plusieurs aspects, différentes des autres types de corpus (Cooper *et al.*, 2005; Cooper & Manson, 2007). En effet, outre le volume de données à traiter, ce type de corpus se caractérise par une multitude d'expressions différentes ou variantes pour exprimer une même entité nommée et plus particulièrement une même ES. Nos travaux entrent dans ce cadre et nous proposons une méthode originale combinant une analyse statistique (mesure de similarité), une analyse lexicale (désaccentuation (Kobus *et al.*, 2008; Figuerola *et al.*, 2001) et identification de préfixes similaires) et une analyse contextuelle pour l'identification et l'extraction d'ES simples et complexes à partir de corpus textuels volumineux de SMS, type de données textuelles encore peu traité à ce jour. La suite de cet article est organisée de la façon suivante. La section 2 présente une brève introduction des travaux liés à l'extraction d'ES. Puis, nous décrivons en section 3 l'approche proposée. Nous détaillons en section 4 le protocole expérimental et les résultats obtenus. La section 5 présente la conclusion et les perspectives de nos travaux.

2 État de l'art

De nombreux travaux se préoccupent d'extraire les entités spatiales. Une première famille se concentre sur l'identification des ES simples, aussi nommées entités spatiales absolues (ESA) à la différence des entités spatiales plus complexes intégrant des relations spatiales (RS) appelées entités spatiales relatives (ESR) (Lesbegueries *et al.*, 2006). Parmi ces travaux, (Florian *et al.*, 2003) proposent un cadre expérimental classifieur-combinaison pour la reconnaissance d'entités nommées dans lequel quatre classifieurs d'entités nommées statistiques sont combinés dans des conditions différentes. Dans (Malouf, 2002), les modèles de Markov cachés sont utilisés pour résoudre la tâche de reconnaissance d'entités nommées pour la langue anglaise. Une deuxième famille plus réduite de travaux cherche à prendre en compte les relations spatiales. Nous pouvons notamment citer (Nguyen *et al.*, 2010) qui proposent une méthode s'appuyant sur les relations n-aires pour l'enrichissement d'une ontologie géographique à partir de l'analyse automatique d'un corpus textuel. À partir de ces relations et d'un lexique de termes géographiques, des relations spatio-temporelles sont identifiées dans un

contexte de descriptions d'itinéraires. Dans (Roberts *et al.*, 2013), une approche pour reconnaître les relations spatiales entre des événements mentionnés est définie. Les auteurs proposent une méthode par apprentissage supervisé pour (1) la reconnaissance d'une relation spatiale entre deux événements mentionnés, et (2) la classification des paires d'événements spatialement connexes selon l'une des cinq relations de confinement spatiales traitées. Dans (Weissenbacher & Nazarenko, 2007), les auteurs proposent de prédire les relations à l'aide de réseaux bayésiens. D'autres travaux s'appuient sur des ressources structurées, parmi lesquels nous pouvons citer (Blessing & Schütze, 2010) décrivant une méthode pour annoter automatiquement les relations entre les entités spatiales en s'appuyant sur des sources de données infoboxes de Wikipédia - Allemagne. (Pustejovsky *et al.*, 2012) proposent un système d'annotation sémantique nommé ISO-Space pour l'annotation d'informations spatiales et spatio-temporelles en langage naturel. Bien que ces travaux soient intéressants dans notre contexte d'étude, ils ne permettent pas de prendre en compte la multitude des formes utilisées par les usagers des téléphones mobiles pour exprimer des ES, qu'elles soient absolues ou relatives.

3 Méthodologie pour l'extraction d'entités spatiales

Dans l'objectif d'extraire des ESA et ESR présentes dans des corpus textuels constitués de SMS, nous proposons une approche, décrite en figure 1, comprenant les trois étapes suivantes : (1) Identification des entités spatiales absolues, (2) Enrichissement des entités spatiales absolues et (3) Identification des relations spatiales. Nous détaillons par la suite chacune des étapes.

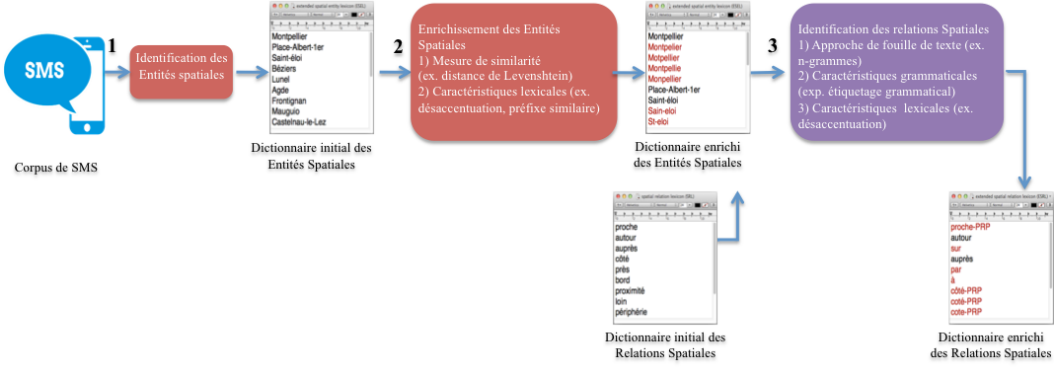

FIGURE 1 – Processus global.

3.1 Identification des entités spatiales absolues

Une première étape consiste à identifier les ESA correctement écrites (variantes standards) et présentes dans les SMS. Pour cela, une approche lexicale classique s'appuyant sur différentes sources de données est appliquée. Nous nous appuyons notamment sur les bases BDTopo et BDCarto de l'IGN ainsi que sur un ensemble de listes de noms de lieux fournies par la Métropole de Montpellier (rue, quartiers, etc.), le corpus de SMS étant relatif à la zone géographique de Montpellier. A noter qu'une approche

classique par patrons intégrant des règles lexicales comme l'identification de mots capitalisés est inefficace étant donnée l'écriture peu standardisée du langage SMS.

3.2 Enrichissement des entités spatiales absolues

Nous proposons, dans un deuxième temps, d'identifier et extraire de nouvelles variantes d'ESA, correspondant à des formulations différentes des ESA existantes, en calculant la similarité entre le dictionnaire initial des ESA et les mots issus du corpus de SMS. Parmi les nombreuses mesures de similarité existantes, nous avons testé les mesures de Lin (Lin, 1998) et String matching [1] (SM) (Maedche & Staab, 2002), classiquement utilisées dans la littérature et qui donnent des résultats pertinents (Duchateau *et al.*, 2008; Zenasni *et al.*, 2015). Au regard des résultats obtenus, nous avons retenu la mesure SM et nous présentons en section 4.1 les résultats associés. Dans un premier temps, nous calculons la similarité entre les ESA (dictionnaire initial) et tous les mots issus du corpus de SMS sans prendre en considération les caractères accentués. Ce type de normalisation permet d'améliorer de manière significative la reconnaissance des ES dans les SMS. Nous pouvons notamment citer les ESA *Sète* et *Sete* qui sont alors considérées comme identiques alors que la valeur de comparaison initiale de la mesure SM est de $0,75$ (valeur allant de 0 pour deux termes distincts à 1 pour deux termes similaires). Dans un deuxième temps, nous appliquons une analyse lexicale pour vérifier si deux termes ont le même préfixe. Dans le cadre du traitement des variantes d'une ESA, nous avons remarqué que, en général, les utilisateurs utilisent les mêmes préfixes que les ESA standards afin qu'elles restent compréhensibles. Dans ce sens, nous faisons l'hypothèse que deux mots qui n'ont pas le même préfixe ne représentent pas le même lieu. Il est possible malgré tout qu'ils soient proches selon la mesure SM mais avec une signification très différente (par exemple *Lattes* et *pattes*). Nous verrons dans les expérimentations menées en section 4 que ces traitements améliorent significativement la qualité de l'extraction des ES.

3.3 Identification des relations spatiales

Sur la base des résultats obtenus dans les étapes précédentes, nous nous appuyons sur un lexique regroupant les différentes formes des cinq relations spatiales d'ordre topologique que sont l'inclusion (dans, etc.), l'orientation (au nord, etc.), l'adjacence (à côté de, etc.), la distance (à X km de, etc.) et la relation géométrique (entre X et Y, etc.) (Lesbegueries *et al.*, 2006). Notre première heuristique consiste à enrichir le dictionnaire initial des relations spatiales (RS) en combinant une analyse morpho-syntaxique qui s'appuie sur le TreeTagger (Helmut, 1994) et une approche fréquentiste (nombre d'occurrences). Une fois le corpus étiqueté, pour chaque ES, nous calculons le nombre d'occurrences des étiquettes des mots précédant l'ES. Nous obtenons ainsi, pour l'ensemble des ES, un vecteur constitué de couples *(étiquettes morho-syntaxique, nombre d'occurrences)* (par exemple, *(PRP* [2]*, 981)*, *(NOM, 460))*. Ensuite, nous faisons l'hypothèse que les mots associés aux deux étiquettes les plus fréquentes sont retenus comme des RS candidates. Puis, nous sélectionnons les mots les plus fréquents sur la base d'un seuil S (les résultats liés à ce seuil sont discutés en section 4.2). Les mots sélectionnés sont alors ajoutés comme nouvelles RS. Dans l'objectif d'enrichir la liste des RS obtenue, nous cherchons maintenant à identifier de nouvelles formes d'expressions de ces relations en langage SMS. Pour cela, nous calculons tout d'abord la similarité entre chaque RS et les N mots

1. Mesure fondée sur la distance de Levenshtein.
2. Préposition

qui précèdent les ES. Pour cette étape, nous réalisons une première action de préparation du corpus en appliquant le module de désaccentuation puis nous utilisons à nouveau la mesure de similarité *SM*. Nous obtenons en résultat une liste de relations spatiales enrichies de candidats proches selon la mesure SM (par exemple, *pres*). Nous proposons ensuite de sélectionner les mots précédant l'ESA. De manière concrète, nous utilisons la méthode des n-grammes de mots pour sélectionner les n mots précédant une entité spatiale (frontière droite) et incluant une RS (frontière gauche). Puis nous généralisons ces n mots sur la base des informations grammaticales (étiquettes morpho-syntaxiques) qui leur sont associées. Par exemple, les *n-grammes de mots près de, près du, près des, etc.* qui s'appuient sur la RS initiale *près* sont associés au patron *près+PRP*. Enfin, un patron plus général peut être proposé en utilisant les informations spatiales issues de nos dictionnaires de RS. Ainsi, le patron général *RS+PRP* permet d'instancier les n-grammes de mots *près de, près du, proche de, etc.*

4 Expérimentations

Dans cette section, nous présentons une série d'expérimentations permettant d'évaluer les méthodes d'identification automatique des variantes d'ES et de RS. Afin d'évaluer la capacité de notre système, nous évaluons notre approche sur le corpus de SMS *88milSMS*[3]. Le corpus est composé de plus de 88000 SMS authentiques en français qui ont été recueillis par une équipe pluridisciplinaire de linguistes et d'informaticiens issus de Montpellier en 2011 dans le cadre du projet Sud4Science Languedoc Roussillon. Deux dictionnaires initiaux des ES et RS ont été utilisés pour la tâche d'extraction. Le dictionnaire des ES contient une description complète des 7008 éléments géographiques en France, des rivières, noms de villes, communes, quartiers, lieux patrimoine, stations de tram ... etc. Pour les RS, nous nous appuyons sur les dictionnaires issus des travaux de (Lesbegueries *et al.*, 2006) qui contiennent 163 relations spatiales.

4.1 Identification des entités spatiales absolues

Pour mener à bien ces expérimentations, nous évaluons la capacité du système à identifier de nouvelles variantes d'ESA. Le tableau 1 indique, dans la colonne *Similarité de base* (baseline), les résultats obtenus avec l'application de l'algorithme *SM* uniquement. Les expérimentations intitulées *Similarité + caractéristiques lexicales* présentent les résultats issus de notre heuristique décrite en section 3.2, à savoir l'application d'un processus de désaccentuation et l'identification de préfixes similaires. Les résultats sont évalués en termes de *macro-, micro- précision* et *rappel*. La macro-moyenne consiste à calculer la moyenne des précisions et rappels des entités. La micro-moyenne calcule la précision et rappel sur la base de l'ensemble des instances des entités. Nous avons calculé la précision et le rappel sur un échantillon de 1000 SMS. Un groupe de trois annotateurs a extrait toutes les ESA de 1000 SMS manuellement. Ceci nous a permis d'évaluer la qualité de notre méthode d'extraction des ESA puis d'intégrer les nouveaux éléments à nos dictionnaires pour la phase d'extraction des RS. Sur l'échantillon de 1000 SMS, notre système a été capable d'identifier 37 ESA standards (par exemple : *Montpellier, béziers, saint-éloi* ...) et 17 nouvelles variantes d'ESA (*motpellier, bezier, st-eloi* ...). Les résultats présentés dans le tableau 1 montrent que l'utilisation de notre heuristique permet d'améliorer, de manière significative l'identification des ESA. Par exemple, l'algorithme de base associe le toponyme *lattes* aux mots *pattes*, *mattes*, *battes*, *nattes*, *flattes*, *luttes*, *latte* et *béziers*

3. http ://88milsms.huma-num.fr

est associé à *beziers*. Après l'application de nos propositions, *lattes* est seulement associé à *latte*, et *béziers* est associé à *beziers* mais aussi à *bezier* et *bezies* qui sont des éléments tout à fait pertinents.

	Similarité de base			Similarité + caractéristiques lexicales		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
Micro	0,32	0,77	0,44	0,83	0,86	0,84
Macro	0,32	0,86	0,46	0,86	0,90	0,87

TABLE 1 – Résultats liés à l'extraction des ESA sur un échantillon de 1000 SMS.

4.2 Identification des relations spatiales

Ces expérimentations liées à l'évaluation des relations spatiales extraites reposent sur deux séries d'expérimentations. La première série concerne l'identification de nouvelles RS en utilisant le dictionnaire d'ESA et l'étiquetage morpho-syntaxique. Nous avons fait varier les paramètres susceptibles d'influencer les résultats (cf. Table 2), c'est-à-dire le seuil S relatif au nombre d'occurrences (cf. Section 3.3) et nous avons fixé N à un mot précédant les ESA. Ces résultats mettent en relief qu'il est préférable d'appliquer un seuil S au delà de 5. Dans la deuxième série d'expérimentations, le système est testé à l'aide des dictionnaires enrichis des RS et ESA. Nous avons fait varier n propre aux n-grammes de mots pour identifier la valeur la plus adaptée, à savoir $n = 2$ (cf. Table 3). Par ailleurs, la syntaxe qui semble la plus pertinente correspond à la structure ***RS + PRP*** (par exemple, *près de*, *proche du*). Cette évaluation qualitative devra être approfondie dans nos futurs travaux.

5 Conclusion

Dans cet article, nous avons proposé une méthode d'identification automatique de nouvelles variantes d'entités et relations spatiales à partir des SMS. Pour l'identification des variantes d'ES, nous avons combiné l'analyse statistique (mesure de similarité), l'analyse lexicale (désaccentuation et préfixes similaires) et des mesures d'édition. Puis nous avons utilisé ces résultats associés à une analyse contextuelle pour identifier de nouvelles RS. Nos résultats montrent que la combinaison de différentes approches améliore la qualité de l'extraction des ES et RS. En perspectives, nous envisageons d'approfondir l'étude de la généricité de notre méthode sur un corpus standard d'articles de presse *Midi Libre*. Nous souhaitons ensuite exploiter un corpus de tweets (données également « bruitées » et de nature informelle avec l'utilisation de nombreuses variantes) afin de mettre en relief les spécificités lexicales et syntaxiques des différents modes de communication médiée.

S	Précision	Relations pertinentes	Relations non pertinentes
10	0,62	5	3
5	**0,67**	8	4
3	0,55	8	8

TABLE 2 – Résultats en terme de Précision (variation de S).

n	Précision	Nombre de nouvelles relations
2	**0,92**	13
3	0,88	15
4	0,42	7

TABLE 3 – Résultats en terme de Précision (variation de n).

Remerciements

Ce travail est soutenu par le ministère algérien de l'enseignement supérieur et de la recherche scientifique et le Cirad. Notre étude fait suite au projet sud4science LR (http ://sud4science.org/) coordonné par Rachel Panckhurst et soutenu par le CENTAL (Université Catholique de Louvain, Belgique), la MSH-M (Maison des Sciences de l'Homme de Montpellier) et la DGLFLF (Délégation générale à la langue française et aux langues de France).

Références

BLESSING A. & SCHÜTZE H. (2010). Self-annotation for fine-grained geospatial relation extraction. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, p. 80–88.

COOPER R., ALI S. & BI C. (2005). Extracting information from short messages. In *Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Proceedings*, p. 388–391, Alicante.

COOPER R. & MANSON S. (2007). Extracting temporal information from short messages. In *Data Management. Data, Data Everywhere, 24th British National Conference on Databases, BNCOD 24, Proceedings*, p. 224–234, Glasgow.

DUCHATEAU F., BELLAHSENE Z. & ROCHE M. (2008). Improving quality and performance of schema matching in large scale. *Ingénierie des Systèmes d'Information*, **13**(5), 59–82.

FIGUEROLA C. G., RODRÍGUEZ Á. F. Z. & BERROCAL J. L. A. (2001). Automatic vs manual categorisation of documents in spanish. *Journal of Documentation*, **57**(6), 763–773.

FLORIAN R., ITTYCHERIAH A., JING H. & ZHANG T. (2003). Named entity recognition through classifier combination. p. 168–171, Edmonton : CoNLL 2003.

HELMUT S. (1994). Probabilistic part-of-speech tagging using decision trees. In *In Proceedings of the International Conference on New Methods in Language Processing*, Manchester.

KOBUS C., YVON F. & DAMNATI G. (2008). Normalizing SMS : are two metaphors better than one ? In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference*, p. 441–448, Lyon.

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

LIN D. (1998). An information-theoretic definition of similarity. In *Proc. of the Fifteenth Int. Conf. on Machine Learning (ICML)*, p. 296–304.

MAEDCHE A. & STAAB S. (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Int. Conf. EKAW*, p. 251–263.

MALOUF R. (2002). Markov models for language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*.

NGUYEN V. T., GAIO M. & SALLABERRY C. (2010). Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. *CoRR*.

PUSTEJOVSKY J., MOSZKOWICZ J. & VERHAGEN M. (2012). A linguistically grounded annotation language for spatial information. *TAL*, **53**(2), 87–113.

ROBERTS K., SKINNER M. A. & HARABAGIU S. M. (2013). Recognizing spatial containment relations between event mentions. *IWCS*, p. 216–227.

WEISSENBACHER D. & NAZARENKO A. (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Proceedings of TALN*, p. 145–155.

ZENASNI S., KERGOSIEN E., ROCHE M. & TEISSEIRE M. (2015). Discovering types of spatial relations with a text mining approach. In *Foundations of Intelligent Systems - 22nd International Symposium, ISMIS Proceedings*, p. 442–451, Lyon.