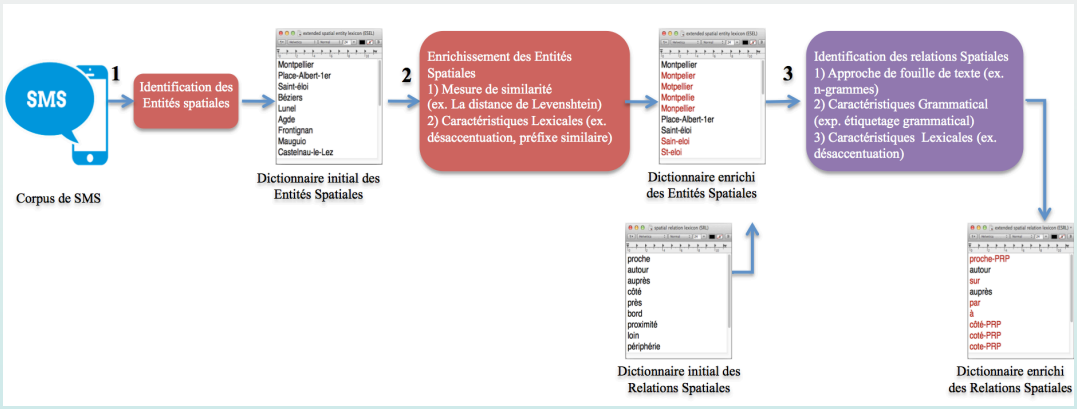
Découverte de nouvelles entités et relations spatiales à partir d'un corpus de SMS

Sarah Zenasni[1, 2], Eric Kergosien[3], Mathieu Roche[1, 2], Maguelonne Teisseire [1,2]
sarah.zenasni@teledetection.fr
[1] UMR Tetis CIRAD, Irstea, AgroParisTech, France
[2] LIRMM, Université de Montpellier, France
[3] Univ. Lille 3, GERiICO, France

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

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.



Méthodologie :

1. Identification et Enrichissement des Entités Spatiales Absolues (ESA)

a. Nous calculons la similarité entre ESA et tout le corpus de SMS en utilisant les mesures de similarité String Matching (SM) et Lin.

b. Nous appliquons des caractéristiques lexicales :

1 la désaccentuation afin d'améliorer l'identification de nouvelles ESA (par exemple **Sète** et **Sete** qui sont alors considérées comme identiques alors que la valeur de comparaison initiale de la mesure SM est de **0,75**).

2 Nous vérifions si deux termes ont le même préfixe car il est possible malgré tout qu'ils soient proches selon la mesure SM mais avec une signification très différente (par exemple **Lattes** et **pattes**).

2. Identification des Relations Spatiales (RS)

a. Etape 1 :

1. Nous étiquetons morpho-syntaxiquement le corpus de SMS en utilisant le TreeTagger.
2. Puis, nous appliquons une approche fréquentiste (nombre d'occurrences) afin de pouvoir d'identifier les étiquètes les plus fréquentes.
3. Nous sélectionnons les mots associés aux deux étiquettes les plus fréquents sur la base d'un seuil S comme des RS candidates.

b. Etape 2 :

1. Nous calculons la similarité entre chaque RS et les N mots qui précèdent les ESA en utilisant SM et la désaccentuation (par exemple : pres, coté, cote, etc).
2. Nous utilisons la méthode des n-grammes de mots pour sélectionner les n mots qui se trouvent entre RS et ESA.
3. Nous généralisons ces n mots sur la base des informations grammaticales 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 + Préposition**).

Expérimentations :

Nous évaluons notre approche sur le corpus de SMS **88milSMS**.

1. Notre système a été capable d'identifier 37 ESA standards (par exemple : Montpellier, béziers, saint-éloi, etc.) et 17 nouvelles variantes d'ESA (motpellier, bezier, st-eloi, etc.).

	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

2. a. Nous avons obtenu le meilleur résultat avec **S = 5** relatif au nombre d'occurrences.

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

2. b. Nous avons fait varier n propre aux n-grammes de mots pour identifier la valeur la plus adaptée. La syntaxe qui semble la plus pertinente correspond à la structure RS + PRP (par exemple, **près de** , **proche du**).

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

Perspectives :

1. Nous souhaitons 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.

2. Nous envisageons ensuite d'approfondir l'étude de la généricité de notre méthode sur un corpus standard d'articles de presse Midi Libre.