Extraction de Localisations dans les MicroBlogs

Thi Bich Ngoc Hoang, Josiane Mothe

Extraction de Localisations dans les MicroBlogs

Thi-Bich-Ngoc Hoang*,**, Josiane Mothe*

*Université de Toulouse et IRIT, UMR5505 CNRS, France
Prénom.Nom@irit.fr
**University of Economics, the University of Danang, Vietnam

Résumé. La circulation de l'information est de plus en plus rapide. Les applications comme WhatsApp ou Twitter permettent d'échanger des informations sur des évènements de façon quasi instantanée. Il s'agit de ressources précieuses desquelles peuvent être extraites des informations sur des événements (temps, localisation ou entité concernée). Nous nous centrons ici sur l'aspect localisation qui a de nombreuses applications aussi bien dans le cadre d'outils géo-spatialisés que pour des recommandations personnalisées. Dans le contexte de microblogs, les outils développés en traitement du langage naturel ne sont pas suffisants compte tenu de la forme des messages; par exemple les tweets ne sont pas linguistiquement corrects. Par ailleurs, le nombre important de messages à traiter est également un challenge. Dans ce article, nous présentons un modèle pour prédire si un microblog (tweet) contient une localisation ou non et nous montrons que cette prédiction améliore l'efficacité de l'extraction de localisations des tweets.

1 Introduction

De nombreux travaux actuels s'intéressent aux microblogs et à leur exploitation. Par exemple, SanJuan et al. (2012) ont introduit une tâche d'évaluation à CLEF [1] concernant la contextualisation de tweets pour aider à leur compréhension. Dans TREC [2], la tâche vise à proposer des recommandations contextuelles aux utilisateurs (Ounis et al., 2011). La tâche CLEF a évolué récemment pour prendre en compte différents besoins qui peuvent être utiles aux utilisateurs dans le cadre d'événements comme les festivals (Goeuriot et al., 2016; Ermakova et al., 2017; Goeuriot et al., 2018).

Un événement possède trois composants essentiels ((Sundheim, 1996)) : (a) une localisation qui indique *où* l'événement se passe ; (b) une temporalité qui indique *quand* l'événement se passe ; (c) une information sur l'entité concernée qui indique *sur quoi* ou *sur qui* porte l'événement.

Cet article, dont une version étendue a été publiée dans le journal IPM (Hoang et Mothe, 2018), est centré sur la dimension de localisation qui est vitale pour les applications géo-spatiales (Munro, 2011). Par exemple, l'une des premières informations transmises aux sys-

1. CLEF est un programme de recherche européen centré sur l'évaluation de tâches de recherche d'information `http://www.clef-initiative.eu/`
2. TREC est un autre programme d'évaluation en RI patronné par le NIST USA `trec.nist.gov`

tèmes de secours en cas de catastrophe est l'endroit où la catastrophe s'est produite (Lingad et al., 2013). Les informations de localisation sont parfois présentes dans les microblogs quasi simultanément aux événements eux-mêmes. Par exemple, les utilisateurs de messagerie instantanée comme Twitter sont susceptibles de transmettre des mises à jours très régulières et les utilisateurs eux-mêmes trouvent la localisation de l'information très importante (Vieweg et al., 2010).

Au cours des dernières années, plusieurs systèmes de reconnaissance d'entités nommées (EN) traitent du problème de l'extraction de localisations spécifiées dans les documents (Bontcheva et al., 2013; Etzioni et al., 2005) ; mais ces systèmes ne fonctionnent pas bien sur des textes informels. En effet, les analyseurs de texte utilisent des fonctionnalités telles que le type de mot, les lettres en majuscules et le contexte agrégé, qui ne sont souvent pas exacts dans des microblogs bruités, non structurés et courts (Huang et al., 2015).

L'identification ou extraction de localisations repose principalement sur : 1) la recherche et comparaison du texte avec les noms d'entités dans des répertoires, 2) l'utilisation de la structure et du contexte du texte. Le premier type de méthodes est simple mais limite l'extraction à une liste prédéfinie de noms, alors que le second est capable de reconnaître les noms même s'ils ne figurent pas sur la liste (Huang et al., 2015).

Stanford NER est un système d'extraction d'EN qui s'appuie sur une méthode d'apprentissage automatique (Toutanova et al., 2003) ; il fonctionne bien sur les nouvelles mais mal sur les microblogs. Récemment, Bontcheva et al. (2013) ont adapté leur système GATE d'extraction d'EN pour les tweets. Ils ont également adapté l'analyseur de Stanford pour les collections de tweets. Leurs propositions ont permis d'augmenter la performance en termes de mesure F de 60 % à 80 %, principalement pour l'extraction de personnes, d'organisations et du temps, mais pas en ce qui concerne les lieux. Ritter et al. (2011) a abordé le problème de l'extraction d'EN pour les microblogs en utilisant un modèle probabiliste et une base de données ouverte (Freebase) comme source d'apprentissage. Leurs expériences montrent que leur approche surpasse les outils existants sur les tweets pour les entités de localisation avec une mesure F de 77%. Alors que Gate NLP est plus efficace en termes de rappel, Stanford NER et Ritter sont plus efficaces en termes de précision (Bontcheva et al., 2013; Hoang et al., 2017). Dans cet article, nous introduisons une méthode qui combine ces outils pour cibler des applications orientées vers le rappel ou orientées vers la précision. Nous proposons également une méthode prédisant si les microblogs contiennent une localisation. Nous proposons également de filtrer les localisations extraites à l'aide de DBpedia pour augmenter la précision des outils.

L'article est organisé comme suit : dans la section 2, nous présentons les résultats obtenus par une méthode de fusion de différents outils d'extraction. Dans la section 3, nous présentons nos propositions pour la prédiction de la présence d'une localisation dans un tweet et montrons que cette prédiction améliore significativement l'efficacité de l'extraction. Finalement, nous concluons cet article en Section 4. Ces travaux ont également donné lieu à publication dans la revue Information Processing Management (Hoang et Mothe, 2018).

2 Combinaison des méthodes d'extraction de la littérature

Plusieurs méthodes se sont intéressées à l'extraction de localisation dans des textes comme Ritter tool (Ritter et al., 2011), Gate NLP framework (Bontcheva et al., 2013) et Stanford NER (Finkel et al., 2005). Nous avons étudié la combinaison de ces trois méthodes : nous avons

extrait les localisations identifiées par chacun des trois outils et les avons fusionnés. Nous avons également considéré leur filtrage après extraction en nous appuyant sur la base DBpedia (`http://dbpedia.org/snorql/`). Pour les évaluations, nous avons utilisé deux collec-

TAB. 1 – *Description of data used for training and testing.*

	Ritter's dataset	MSM2013 dataset
Training	142 TCL, 1420 TNL	331 TCL, 1655 TLN
Testing	71 TCL, 761 TNL	165 TCL, 664 TNL

tions standards : la collection Ritter (Ritter et al., 2011) et la collection MSM2013 (Cano Basave et al., 2013). La collection Ritter contient 2 394 tweets dont 213 (soit $8,8\%$) avec localisation et 2 181 sans. MSM2013 contient 2 815 tweets dont 496 (soit $17,6\%$) avec localisation et 2 319 sans. Les localisations dans ces collections sont annotées ; elles contiennent un ensemble d'apprentissage et un ensemble de test. La Table 1 présente les caractéristiques de ces collections.

Les résultats pour le rappel, la précision et la mesure F sont présentés dans la table 2 (nous n'avons pas testé l'ensemble des combinaisons possibles et laissons cela pour des travaux futurs). Nous avons utilisé le T-test avec comme fonction témoin l'extraction par l'outil Ritter (première ligne de la table 2).

TAB. 2 – *Résultats de la combinaison des modèles Ritter, Gate et Stanford et du filtrage avec DBPedia. Rappel - R(%), Précision - P(%), Mesure F - F(%). (*) indique un résultat statistiquement significatif par rapport à la fonction témoin.*

	Données Ritter			**Données MSM2013**		
	R(%)	**P(%)**	**F(%)**	**R(%)**	**P(%)**	**F(%)**
Ritter (témoin)	71	82	77	61	80	69
Ritter +Stanford+DBp	77*	79	**78**	72*	79	**75***
Ritter+Gate+DBp	78*	71	74	74*	77	**75***
Ritter+Stanford	80*	64	72	**78***	72	**75***
Ritter+Gate	**82***	56	66	**78***	64	71
Ritter+DBp	45	**97***	62	48	**88***	62

Comme le montre la table 2, la combinaison de l'outil Ritter et de Stanford-NER filtré par DBpedia donne la meilleure mesure F. Pour MSM2013, la mesure F augmente de 69 % à 75 %. Lorsque l'on s'intéresse à une forte précision, c'est la combinaison de Ritter avec le filtrage BDPedia qui est la plus efficace (dernière ligne) alors que pour le rappel, il s'agit de la combinaison de Ritter avec Gate (avant dernière ligne).

3 Prédiction de la présence de localisation

Prévoir qu'un tweet contient un nom de lieu n'est pas simple car les tweets sont généralement écrits dans un langage pseudo-naturel et peuvent ne pas correspondre à des phrases grammaticalement correctes. Les outils usuels de traitement automatique de la langue rencontrent alors des difficultés. Nous proposons un ensemble de caractéristiques pour représenter

les tweets et nous étudions la pertinence de cette représentation dans un modèle prédictif basé sur un apprentissage automatique. La Table 3 présente ces caractéristiques ; (Hoang et Mothe, 2018) présente plus de détails.

TAB. 3 – *Caractéristiques pour prédire la présence de localisation dans un tweet.*

Nom	**Description**	**Exemple**
1. Geography gazetteer	Contient un terme qui apparaît dans Gate geography gazetteer	Today I got a new job ; tomorow jI will be staying in **Dublin**
2. Prep+PP	Contient une préposition juste avant un nom propre	- RT @RMBWilliams : Here **in Gainesville** ! - Greek Festival **at St Johns** before ASPEN !
3. PP	Nombre de nom propres	going **to** alderwood :). # PP : 1
4.Prep	Contient une des 7 prépositions anglaises de lieu ou de mouvement : *at, in, on, from, to, toward, towards*	- Feeling really good after great week **in** our London offices - @Strigy got mine **in** bbt aintree today
5. Place+PP	Contient un terme spécifiant un lieu (*town, city, state, region, country*) juste avant ou après un nom propre	- The football fever : Ohio head coach Frank Solich says Ohio **state** knows they have a special team and season underway
6. Time	Contient une expression de temps (*today, tomorrow, weekend, tonight...*)	- Headed to da gump **today** alabama here I come - Come check out Costa Lounge **tonight** !
7. DefArt+PP	Contient un article défini juste avant un nom propre	- Beautiful day ! Nice to get away from **the Florida** heat
8. Htag	Contient un hashtag	**#Brazil**
9. Adj	Nombre d'adjectifs	- **Bad** time for leicester fans. # Adj :1
10. Verb	Nombre de verbes	- Willingham **took** a turn. # Verb : 2

Les caractéristiques "PP","Adj","Verb" sont des entiers alors que les autres sont des valeurs booléennes.

Nous avons utilisé les mêmes collections que dans la section 2 pour l'apprentissage de notre modèle pour la détection de la présence d'une localisation. Nous avons utilisé différents algorithmes d'apprentissage : Naive Bayes (NB), Support Vector Machine (SMO) et Random Forest (RF) avec une validation croisée. Lors de l'apprentissage, il est possible d'optimiser différents critères ; nous avons choisi la précision et les vrais positifs comme critères d'optimisation. Nous ne présentons pas ici le détail des résultats mais nous obtenons une mesure F d'environ $0,65$ et une précision (accuracy) de $0,80$ à $0,92$ en fonction des cas. RF permet d'obtenir les meilleurs résultats.

TAB. 4 – *Efficacité de l'algorithme Ritter pour les collections Ritter et MSM2013 en termes de Rappel, Précision, mesure F, sur l'ensemble de tests tel que décrit dans le Tableau 1 et les tweets que nous prédisons comme contenant une localisation (avec RF). Les valeurs statistiquement différentes de la fonction témoin sont indiquées par une étoile (*). Le nombre entre parenthèses est le meilleur résultat des trois tirages.*

	Ritter dataset			**MSM2013 dataset**		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
Baseline	69	85	75	60	80	69
Accuracy	45(51)	**96*(98)**	61(66)	37(40)	**89*(92)**	52(55)
TP	45(51)	**96*(98)**	61(66)	37(40)	**89*(92)**	52(55)

Le modèle appris permet donc de prédire si un nouveau tweet contient une localisation ou non. C'est sur ces seuls tweets que nous extrayons ensuite les localisations. La table 4 présente les résultats que nous avons obtenus lors de l'extraction des emplacements des tweets prédits comme contenant une localisation. Un seul tirage apprentissage/test ne permet pas de conclure sur les résultats, nous avons donc utilisé trois tirages et rapporté les valeurs moyennes. Le nombre entre parenthèses est le meilleur résultat des trois tirages.

4 Conclusion

Nous avons proposé une approche pour l'extraction de localisations et un modèle pour prédire la présence de localisations dans les tweets. Notre approche d'extraction de localisations repose sur la fusion de méthodes d'extraction existantes et améliore de manière significative les performances lorsque nous visons soit des applications axées sur le rappel, soit au contraire sur la précision. Nous avons montré que : (1) la fusion des localisations extraites par les outils Ritter et Stanford puis filtrées par DBpedia augmente la mesure F. (2) la fusion des localisations extraites par Ritter et Gate améliore considérablement le rappel alors que l'utilisation de DBpedia pour filtrer les entités de localisation reconnues par Ritter augmente considérablement la précision.

Nous avons également fait l'hypothèse que nous pourrions augmenter la précision si nous pouvions prédire la présence de localisations dans les tweets. Nous avons donc introduit une méthode pour prédire si un tweet contient une localisation ou non. Nous avons défini de nouvelles caractéristiques pour représenter les tweets et évalué les paramètres d'apprentissage automatique. Les résultats montrent que : (3) Random Forest et Naive Bayes sont les meilleures solutions d'apprentissage pour ce problème(4) la modification des critères d'optimisation (précision ou vrai positif) ne change pas significativement la mesure F alors que ce changement d'optimisation a un vrai impact sur le taux de vrais positifs et de faux positifs. (5) pour l'extraction de localisation, nous avons amélioré la précision en nous concentrant uniquement sur les tweets prédits comme contenant une localisation par notre méthode. (6) le compromis entre l'augmentation de la précision et la diminution du rappel restent à étudier.

Dans les travaux futurs, nous souhaitons utiliser les méthodes d'encapsulation de mots (word embedding) pour la représentation afin d'étudier l'impact à la fois pour la prédiction de la présence de localisations et pour leur extraction dans les microblogs.

Références

Bontcheva, K., L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, et N. Aswani (2013). TwitIE : An open-source information extraction pipeline for microblog text. In *RANLP*, pp. 83–90.

Cano Basave, A. E., A. Varga, M. Rowe, M. Stankovic, et A.-S. Dadzie (2013). Making sense of microposts (# msm2013) concept extraction challenge.

Ermakova, L., L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, et E. SanJuan (2017). CLEF 2017 Microblog Cultural Contextualization Lab Overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 304–314.

Etzioni, O., M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, et A. Yates (2005). Unsupervised named-entity extraction from the web : An experimental study. *Artificial intelligence 165*(1), 91–134.

Finkel, J. R., T. Grenager, et C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the 43rd annual meeting on association for computational linguistics*, pp. 363–370. ACL.

Goeuriot, L., G. Linares, J. Mothe, P. Mulhem, et E. SanJuan (2018). Building Evaluation datasets for Cultural Microblog Retrieval. In *Language Resources and Evaluation Conference, LREC'18*.

Goeuriot, L., J. Mothe, P. Mulhem, F. Murtagh, et E. SanJuan (2016). Overview of the CLEF 2016 Cultural micro-blog Contextualization Workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 371–378. Springer.

Hoang, T. B. N., V. Moriceau, et J. Mothe (2017). Predicting locations in tweets. In *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Hoang, T. B. N. et J. Mothe (2018). Location extraction from tweets. *Information Processing & Management 54*(2), 129–144.

Huang, Y., Z. Liu, et P. Nguyen (2015). Location-based event search in social texts. In *International Conference on Computing, Networking and Communications (ICNC)*, pp. 668–672. IEEE.

Lingad, J., S. Karimi, et J. Yin (2013). Location extraction from disaster-related microblogs. In *Proc. of the 22nd international conference on world wide web*, pp. 1017–1020. ACM.

Munro, R. (2011). Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Proc. of the conference on computational natural language learning*, pp. 68–77. ACL.

Ounis, I., C. Macdonald, J. Lin, et I. Soboroff (2011). Overview of the trec-2011 microblog track. In *Proc. of the 20th Text REtrieval Conference*, Volume 32.

Ritter, A., S. Clark, O. Etzioni, et al. (2011). Named entity recognition in tweets : an experimental study. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534. ACL.

SanJuan, E., V. Moriceau, X. Tannier, P. Bellot, et J. Mothe (2012). Overview of the INEX 2012 Tweet Contextualization Track. *Initiative for XML Retrieval INEX*, 148.

Sundheim, B. M. (1996). Overview of results of the MUC-6 evaluation. In *Proc. of a workshop on held at Vienna, Virginia*, pp. 423–442. A.

Toutanova, K., D. Klein, C. D. Manning, et Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the Conference of the ACL on Human Language Technology*, pp. 173–180. ACL.

Vieweg, S., A. L. Hughes, K. Starbird, et L. Palen (2010). Microblogging during two natural hazards events : what twitter may contribute to situational awareness. In *Proc. of the SIGCHI*, pp. 1079–1088. ACM.

Summary

Applications such as WhatsApp or Twitter allows anyone to exchange information about events almost instantly. There are therefore valuable resources from which information about events (time, location or entity) can be extracted. In this paper, we focus on localization, which has many applications in the context of geo-spatialized tools or for personalized recommendations. In the context of microblogs, tools developed in natural language processing are not sufficient; for example, tweets are generally not linguistically correct. Moreover, the large number of messages to be processed is also a challenge to solve. in this paper, we present a model for predicting whether a short text contains a localization or not and we show that this prediction improves localization extraction.