Méthode d'Apprentissage pour Extraire les 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

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 concernant la contextualisation de tweets pour aider à leur compréhension, en particulier dans le cadre d'évènements comme les festivals (Goeuriot et al., 2016; Ermakova et al., 2017).

Un évènement possède trois composants essentiels : une localisation, une temporalité, une information sur l'entité concernée. Cet article est centré sur la dimension de localisation qui est vitale pour les applications géo-spatiales (Munro, 2011). 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 ; mais ces systèmes ne fonctionnent pas bien sur des textes informels.

Plusieurs méthodes se sont intéressées à l'extraction de localisation dans des textes comme Ritter tool (Ritter et al., 2011), Gate NLP (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.

Pour les évaluations, nous avons utilisé deux collections 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 résultats sont présentés dans la table 1 (avec le test statistique).

	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

TAB. 1 – *Résultats de la combinaison des modèles Ritter, Gate et Stanford et du filtrage avec DBPedia. Rappel - R(%), Précision - P(%), Mesure F - F(%).*

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).

Prévoir qu'un tweet contient un nom de lieu n'est pas simple car les tweets sont généralement écrits dans un langage pseudo-naturel. Les outils usuels de traitement automatique de la langue rencontrent alors des difficultés. Nous avons proposé un ensemble de caractéristiques pour représenter les tweets et nous avons étudié la pertinence de cette représentation dans un modèle prédictif basé sur un apprentissage automatique. Ces caractéristiques sont précisées et détaillées dans (Hoang et Mothe, 2018). Nous avons utilisé différents algorithmes d'apprentissage : Naive Baiyes (NB), Support Vector Machine (SMO) et Random Forest (RF) avec une validation croisée. Nous avons obtenu une mesure F d'environ $0,65$ et une précision (accuracy) de $0,80$ à $0,92$ en fonction des cas. RF permet d'optenir les meilleurs résultats.

Le modèle appris permet donc de prédire si un nouveau tweet contient une localisation ou non. Plus de détails sur cette approche sont disponibles dans (Hoang et Mothe, 2018). Dans nos travaux futurs, nous souhaitons analyser comment les localisations pourraient aider à la prédiction de la diffusion des tweets. Nous pourrions ainsi étanedre les travaux présentés dans Hoang et Mothe (2017).

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 *CLEF*, pp. 304–314.

Finkel, J. R., T. Grenager, et C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pp. 363–370.

Goeuriot, L., J. Mothe, P. Mulhem, F. Murtagh, et E. SanJuan (2016). Overview of the clef 2016 cultural micro-blog contextualization workshop. In *CLEF*, pp. 371–378. Springer.

Hoang, T. B. N. et J. Mothe (2017). Predicting Information Diffusion on Twitter - Analysis of predictive features. *Journal of Computational Science 22*.

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

Munro, R. (2011). Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Computational Natural Language Learning*, pp. 68–77.

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

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.