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(%).*