Identification de lieux dans les messageries mobiles

Clément Doumouro, Adrien Ball, Joseph Dureau, Ramzi Ben Yahya, Sylvain Raybaud
Snips Research, 18 rue Saint Marc, 75002 Paris, France
prenom.nom@snips.ai

RÉSUMÉ

Nous présentons un système d'identification de lieux dans les messageries typiquement utilisées sur smartphone. L'implémentation sur mobile et son cortège de contraintes, ainsi que la faible quantité de ressources disponibles pour le type de langage utilisé rendent la tâche particulièrement délicate. Ce système, implémenté sur Android, atteint une précision de 30% et un rappel de 72%.

ABSTRACT

Place extraction from smartphone messaging applications.

We propose a place extraction system for smartphone messaging applications. On device implementation comes with specific constraints regarding computation costs and model size ; the messaging language style is also very specific and very little data is available to train models on. Our system achieves a precision of 30% and a recall of 72% on the data we collected and labelled.

MOTS-CLÉS : extraction de lieu, reconnaissance d'entités nommées, sms, smartphone.

KEYWORDS: place extraction, named entity recognition, text message, smartphone.

1 Introduction

L'identification de lieux dans les services de messagerie sur smartphone (SMS, Facebook messenger, etc.) est capitale pour proposer des services contextualisés à ses utilisateurs. Nous estimons cependant que cela ne peut se faire au détriment du respect de leur vie privée. La meilleure garantie est d'effectuer tous les traitements directement sur l'appareil. Cela amène des contraintes sur les capacités de calcul et de stockage. En outre, le style de langue utilisé dans ces applications (capitalisation hasardeuse, « langage texto », etc.) met en échec les méthodes de l'état de l'art, souvent dépendantes de grands corpus annotés. Des corpus similaires pour les messageries mobiles sont rares et aucun, à notre connaissance, n'est annoté avec des informations de lieu. Nous avons donc collecté et annoté un corpus d'environ 2200 « messages Facebook » pour entraîner le système. Les messages sont en anglais mais le système est générique et a été testé avec succès, bien que qualitativement seulement, dans d'autres langues (voir capture d'écran, Figure 1).

2 Description du système

NOTRE ARCHITECTURE à double classification est inspirée de celle décrite dans (Sitter & Daelemans, 1997), en remplaçant le classificateur bayésien naïf par des arbres de décisions, moins coûteux en ressources et offrant des performances comparables. Notre algorithme utilise également des listes

de noms de lieux automatiquement extraites de services en ligne. Mais ces noms sont souvents ambigüs et leur présence dans un texte ne désigne pas toujours un lieu.

UNE PRÉSÉLECTION DE CANDIDATS repère les phrases susceptibles de contenir des informations de lieux. Cela compense partiellement le fort déséquilibre entre les proportions de phrases contenant ou non des lieux, qui nuit à la qualité des modèles entraînés (Sitter & Daelemans, 1997). En effet, un peu moins de 3% des messages du corpus de test mentionnent effectivement un lieu. Des caractéristiques de surface sont utilisées pour la classification (capitalisation, distributions de fréquences des mots, présence des mots dans des listes de lieux, etc.). La précision de cette sélection est de 8% pour un rappel de 91%, mais c'est ce dernier qui est primordial à cette étape.

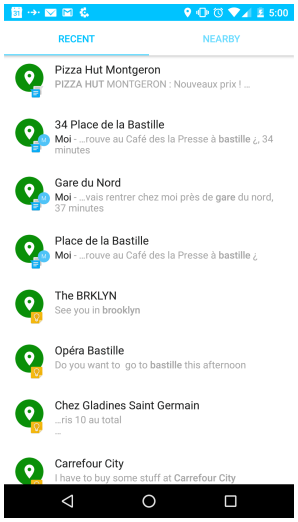

FIGURE 1 – Capture d'écran.

LA DÉTECTION DE LIEU se fait en étiquetant chaque mot comme faisant partie ou non d'une désignation de lieu. Un premier arbre parcourt la phrase et étiquette chaque mot en utilisant des caractéristiques similaires à celles du sélecteur de candidat, ainsi que l'étiquette du mot précédent ; un second fait de même en parcourant la phrase à l'envers ; enfin une régression logistique prédit l'étiquette finale en combinant les deux prédictions et un jeu de caractéristiques réduit. Une séquence continue de mots étiquetés positivement identifie un lieu.

UN FILTRAGE A POSTERIORI, basé sur Word2Vec (Mikolov *et al.*, 2013) et entraîné sur des données issues de forums, élimine un certain nombre de faux positifs en mesurant leur proximité sémantique avec des mots « déchets » connus (onomatopées, insultes,..). Nous obtenons finalement une précision de 30% et un rappel de 72%. La spécificité du classifieur, indépendante de la proportion de messages contenant des lieux, est supérieure à 95%.

3 Conclusion

Nous obtenons des résultats nettement plus précis qu'avec Stanford NER tagger ou un service d'extraction de lieux comme Indico. Ce dernier se révèle très sensible à la capitalisation et aux fautes d'orthographes, probablement parce qu'il est entraîné sur du texte essentiellement bien construit (cf. comparaison figure 2). Après ce prototype nos efforts vont porter sur la constitution de corpus plus importants et l'implémentation d'étiqueteurs POS légers comme (Brill, 1992).

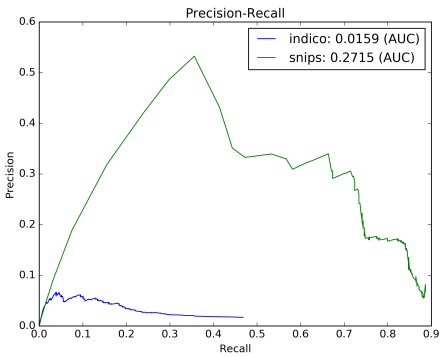

FIGURE 2 – Précision et rappel.

Références

BRILL E. (1992). A simple rule-based part of speech tagger.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality.

SITTER A. D. & DAELEMANS W. (1997). Information extraction via double classification.