Représentations vectorielles de corpus collaboratifs sur la ville de Paris

Carmen Brando*, Catherine Dominguès**

*EHESS, 190-198 Avenue de France, Paris
carmen.brando@ehess.fr
**Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint Mandé
catherine.domingues@ign.fr

Ce travail vise à analyser un corpus collaboratif produit dans le contexte d'appels à contributions menés par la Mairie de Paris. Ces contributions sont reçues sur la plateforme en ligne "Madame la maire, j'ai une idée" ; ces contributions sont sollicitées par la ville qui souhaite recueillir les opinions des Parisiens sur certaines thématiques. Ces derniers déposent des témoignages, des propositions sur leur environnement urbain ainsi que des demandes sur des projets d'aménagement qu'ils souhaitent voir entrepris par la mairie. Ce corpus que nous appelons dorénavant le *corpus mairie de Paris* est composé de quatre sous-corpus thématiques : 1) reconquête des berges de Seine, sous-corpus *berges*, 150 contributions, 18 244 mots, 2) construisons notre métropole, sous-corpus *métropole*, 342 contributions, 63 357 mots, 3) budget participatif, sous-corpus *budget*, 5116 contributions, 1 002 085 mots, et 4) réinventons nos places, sous-corpus places, 338 contributions, 69 953 mots. Ce dernier est lui-même divisé en sept sous-corpus dédiés à sept places parisiennes choisies par la mairie, dont la place de la Bastille, la place Gambetta, la place d'Italie, la place de la Nation.

Pour explorer ces corpus, nous employons des représentations vectorielles de mots, en particulier le modèle de réseau neuronal *word2vec* (Mikolov et al., 2013). ll s'agit de construire la représentation des mots d'un texte dans un espace vectoriel de grande dimension ; un mot est modélisé en tant que vecteur de nombres réels et des mots apparaissant dans des contextes similaires seront représentés par des vecteurs plus proches que d'autres mots apparaissant dans des contextes différents. Cette représentation permet de calculer la ressemblance entre mots par la similarité cosinus et de retrouver, par conséquent, des relations sémantiques entre les mots. Il est également possible d'effectuer des opérations arithmétiques sur les vecteurs (par exemple, la soustraction) pour retrouver d'autres types de relations (Levy et Goldberg, 2014).

Les représentations vectorielles de mots remportent des succès dans de nombreuses tâches de traitement du langage naturel et plus récemment, en recherche d'information. En effet, Despres et al. (2016) proposent de représenter le contexte d'un document par un vecteur de mots qui sera utilisé en tant que modèle de langues pour améliorer la recherche dans des documents quand on s'intéresse à des documents multilingues. En outre, Muchemi et Grefenstette (2016) utilisent ces modèles pour produire un vocabulaire spécifique à la thématique d'un corpus (ex : l'astronomie) sous forme de taxonomie, sans avoir à contraster le corpus avec un corpus généraliste déjà constitué, par exemple un texte encyclopédique. Dans ce travail, nous utilisons *word2vec* afin de retrouver, par similarité sémantique, des mots qui désignent des besoins ou des préoccupations inattendus des citoyens (autrement dit, des sujets qui n'ont pas été

identifiés par la mairie comme une thématique pertinente mais qui préoccupent ou intéressent les Parisiens) ou bien des thématiques transversales à tous les corpus. Nous définissons donc deux types d'expériences. Premièrement, nous faisons contraster le vocabulaire utilisé dans les différents groupes de contributions concernant chacun une place parisienne. Deuxièmement, nous comparons le *corpus mairie de Paris* et un corpus encyclopédique français mis à disposition par l'équipe Alpage de l'Inria [1] afin de faire ressortir les spécificités du sens des mots dans notre corpus, par exemple, l'adjectif *vert* renvoie clairement au champ lexical de l'écologie dans notre corpus, et non à la couleur. Nous explorons également les mots similaires à certaines expressions importantes dans notre contexte comme par exemple, Grand Paris (y compris plusieurs variantes orthographiques). Du point de vue méthodologique, *word2vec* construit les vecteurs de mots en utilisant un modèle skip-gram dont la paramétrisation est choisie empiriquement ; nos corpus sont lemmatisés par TreeTagger. Il est également nécessaire de choisir les mots de départ à partir desquels nous recherchons les mots les plus similaires selon le modèle. Pour cela, nous sélectionnons ces mots d'emploi significativement fréquents par un test de chi-2 (voir par exemple Ascone et al. (2016)) à l'aide des logiciels R et Iramuteq.

Dans notre présentation, nous exposerons en détail la méthodologie ainsi que les résultats obtenus en évoquant les limitations liées à la taille des corpus d'entraînement des modèles pour *word2vec*, taille qui a probablement une incidence sur la pertinence des résultats. Quelques pistes de travail futur consistent à classifier automatiquement les relations trouvées en leur attribuant des types, inconnus a priori, grâce à des méthodes en open information extraction (Gábor et al., 2016), ce qui nous permettrait d'accéder à la sémantique des relations.

Références

Ascone, L., C. Dominguès, et J. Longhi (2016). Perception de l'ambiance sonore d'un lieu selon sa représentation visuelle : une analyse de corpus. *Corela 14-1*, 679–696.

Despres, N., S. Lamprier, et B. Piwowarski (2016). Apprentissage de Modèles de Langue Neuronaux pour la Recherche d'Information. In *Conférence en Recherche d'Infomations et Applications*, Toulouse, France, pp. 717–732.

Gábor, K., I. Tellier, T. Charnois, H. Zargayouna, et D. Buscaldi (2016). Détection et classification non supervisées de relations sémantiques dans des articles scientifiques. In *JEP-TALN-RECITAL 2016*, Volume 2 of *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, Paris, France.

Levy, O. et Y. Goldberg (2014). Linguistic regularities in sparse and explicit word representations. In R. Morante et W. tau Yih (Eds.), *CoNLL*, pp. 171–180. ACL.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, et K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

Muchemi, L. et G. Grefenstette (2016). Word Embedding and Statistical Based Methods for Rapid Induction of Multiple Taxonomies. working paper or preprint.

1. http ://alpage.inria.fr/depglove/process.pl