

Retour d'expérience sur la détection automatique de métaphores dans des textes de Géographie

Max Beligné**, Aleksandra Campar*, Jean-Hugues Chauchat*, Mélanie Lefeuvre*, Isabelle Lefort**, Sabine Loudcher*, Julien Velcin*

*Laboratoire ERIC (EA 3083), Université de Lyon
{sabine.loudcher, julien.velcin, jean-hugues.chauchat}@univ-lyon2.fr

**Laboratoire EVS (UMR 5600), Université de Lyon
{max.beligne, isabelle.lefort}@univ-lyon2.fr

Si les recherches sur les métaphores en Géographie ne sont pas nouvelles, il n'existe pas d'étude de grande ampleur sur cette question. Pourtant les recherches et les réflexions existantes menées de façon qualitative sur des petits corpus ont montré que les métaphores constituent une entrée particulièrement intéressante pour réfléchir sur la scientificité de la Géographie. Dans le contexte des Humanités Numériques, un projet de recherche regroupant des chercheurs en Géographie, Informatique et Linguistique a donc identifié l'intérêt d'étudier ce trope sur un large corpus et cherche par conséquent à détecter automatiquement des métaphores dans des textes de Géographie. Cette communication présente un premier retour d'expérience de l'application de la méthode de Heintz et al. (2013). Il s'agit d'une première étape de travail dont les résultats sont pour l'instant mitigés. Par conséquent, l'objectif est de présenter le processus de recherche en insistant sur les choix qui ont été faits et sur leurs conséquences permettant de mieux comprendre les résultats obtenus et d'envisager des améliorations à venir.

La métaphore est un trope qui peut être défini comme un système de correspondances partielles entre un domaine source et un domaine cible. Le choix de la méthode de détection automatique de métaphores résulte d'un état des lieux sur la question. Le travail de Roy et al. (2006) présentant des évolutions diachroniques de métaphores conceptuelles dans des corpus textuels a été identifié comme le plus proche des attentes des géographes. Pourtant, la volonté de travailler sur un large corpus avec des thématiques diverses conduit à vouloir automatiser le processus de recherche. Dans ce cadre, le travail de Heintz et al. (2013) utilisant un modèle de thématiques latentes LDA (*Allocation de Dirichlet Latente*) est choisi comme répondant le mieux aux objectifs attendus. Il permet de travailler sur un large spectre de domaines sources de métaphores avec comme seule contrainte d'établir pour chaque domaine source quelques mots représentatifs. Concernant le domaine cible des métaphores, il est décidé dans un premier temps de cibler un seul domaine, celui de la Géographie.

La méthode choisie s'appuie sur l'utilisation de la méthode LDA sur un large corpus, ici la moitié des articles de Wikipédia choisis de manière aléatoire, pour extraire 100 thématiques. Les domaines sources choisis par les spécialistes sont ensuite articulés à ces thématiques par l'intermédiaire de mots représentatifs. Dans chaque phrase, la présence de chaque domaine source et du domaine cible est calculée par l'intermédiaire des fréquences d'apparition des mots dans les thématiques. Si une phrase contient le domaine cible et un domaine source sous représenté dans l'article (car la méthode fait l'hypothèse qu'une forte représentation est souvent le

signe d'une utilisation littérale), la phrase est annotée par l'algorithme comme métaphorique. Pour évaluer les résultats trouvés, il a fallu construire un corpus de référence car sur le sujet spécifique des métaphores en Géographie, il n'existait pas de ressource déjà constituée. Les métaphores ont été annotées manuellement par deux personnes de manière indépendante en suivant la méthode *Metaphor Identification Procedure Group* (2007).

Le retour d'expérience est d'autant plus intéressant que la méthode a été testée avec différentes valeurs pour le seuil déterminant à partir de quand un domaine peut être considéré comme faisant partie d'une phrase. Une approche tout d'abord purement quantitative a conduit à baisser ce seuil de manière importante pour faire augmenter les résultats de détection. Malheureusement, l'approche qualitative finale montre que ces choix n'étaient pas judicieux. Une analyse plus fine montre que la méthode d'annotation manuelle choisie est maximale au sens où elle tend à annoter beaucoup de phrases comme métaphoriques. En baissant les seuils, on a donc trouvé plus de phrases métaphoriques mais pour de mauvaises raisons. En effet, les concepts sources retenus par l'algorithme sont souvent inexacts et beaucoup de faux positifs sont présents.

A partir de ce constat, en amont, il s'agit de mieux comprendre ce qui s'est passé par rapport au travail initial de Heintz *et al.* qui obtient de meilleurs résultats. Ces derniers ont au contraire maintenu des seuils hauts mais leur méthode d'évaluation est différente. En effet, ils sélectionnent les métaphores les mieux notées par l'algorithme (obtenues avec des seuils de détection élevés) mélangées avec un ensemble de phrases détectées comme non métaphoriques. Ensuite, des annotateurs vérifient si les phrases sont métaphoriques ou pas. Cette sélection en amont du corpus à évaluer permet d'augmenter significativement les résultats, notamment le taux de rappel. En rentrant ainsi dans le détail d'évaluations différentes, ce retour d'expérience attire l'attention sur ce que peuvent cacher les chiffres et provoquer des erreurs importantes.

En aval, il s'agit d'explorer des pistes pour améliorer les résultats actuels. En remontant les seuils, la méthode de Heintz *et al.* ne peut pas être utilisée pour trouver toutes les métaphores. Une utilisation plus ciblée sur un certain type de métaphore, que l'on peut qualifier de "filée sur une même phrase", est sûrement plus adéquate. Cette piste pose le problème de la définition de critères pour une telle catégorie afin d'évaluer l'algorithme sur cette tâche. Pour finir d'autres pistes importantes de recherche, comme une meilleure précision du concept cible ou l'augmentation du nombre de thématiques, sont des résultats importants de ce retour d'expérience afin d'augmenter à l'avenir l'efficacité de cette méthode de détection automatique.

Références

- Group, P. (2007). Mip : A method for identifying metaphorically used words in discourse. *Metaphor and symbol* 22(1), 1–39.
- Heintz, I., R. Gabbard, M. Srinivasan, D. Barner, D. S. Black, M. Freedman, et R. Weischedel (2013). Automatic extraction of linguistic metaphor with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pp. 58–66. Citeseer.
- Roy, T., S. Ferrari, et P. Beust (2006). Étude de métaphores conceptuelles à l'aide de vues globales et temporelles sur un corpus. *Actes de Traitement Automatique des Langues Naturelles (TALN), Leuven, Belgique 1*, 580–589.