Combinaison de mesures lexicales et sémantiques pour l'extraction de données expérimentales dans des articles scientifiques

Martin Lentschat[*,**,***], Patrice Buche[***]
Juliette Dibie-Barthelemy[****], Mathieu Roche[**]

[*]Université de Montpellier.
[**]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier
[***]IATE. Montpellier SupAgro,
Campus Gaillarde 2 place Pierre Viala Bât 31 34060 Montpellier Cedex 01
[***]MIA Paris. AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 05

1 Introduction

Cet article présente une méthode pour représenter et mesurer la pertinence de données expérimentales extraites d'articles scientifiques. Dans le domaine étudié, les emballages alimentaires, le nombre de documents est réduit et ceux-ci contiennent un vocabulaire spécifique. Nous utilisons une Ressource Termino-ontologique (RTO) pour guider l'extraction, les approches par apprentissage n'étant pas adaptées à la taille du corpus. La RTO définit les entités d'intérêt et les décrits à travers un vocabulaire. Les informations recherchées sont liées aux relations de perméabilité et sont de deux types : symboliques (i.e. une expression lexicale) et quantitatives (i.e. une valeur numérique et son unité de mesure).

Les documents contiennent un grand nombre de faux-positifs dû à la présence d'informations n'étant pas liées à la perméabilité des emballages (par exemple, un nom d'emballage cité à titre de comparaison ou une température autre que le paramètre de contrôle de la mesure de perméabilité). Dans ce contexte, nous proposons ici une méthode complète et originale qui intègre une représentation multi-descripteurs des entités extraites permettant de calculer et combiner des scores de pertinence.

2 Méthode

Dans le cadre de ces travaux, nous nous appuyons sur la représentation (*SciPuRe*) (Scientific Publication Representation) (Lentschat et al., 2020a) qui utilise trois catégories de descripteurs afin de représenter les entités reconnues. Les descripteurs ontologiques indiquent le concept représenté et le concept générique dans la RTO. Les descripteurs lexicaux sont la manifestation de l'entité dans le texte, le terme dénotant l'entité et les termes utilisés pour la désambiguïser. Les descripteurs structurels situent l'entité dans le corpus à différents niveaux et

renseignent son contexte : la phrase, la fenêtre lexicale, sa section et l'article. Ces descripteurs sont impliqués dans des scores de pertinence et permettent d'identifier les éléments pertinents.

La mesure de pertinence sémantique utilisée, *conceptual distance*, reflète la spécificité du concept associée aux entités extraites en mesurant la distance entre son concept et son concept générique (Harispe, 2014). Les pertinences lexicales étudiées reposent sur la fréquence des termes et la discriminance exprimée par les modèles *tf-idf* (Salton, 1983) et *tf-icf* (Wang, 1983). Combiner des scores de pertinence permet ensuite de conjuguer les différents critères sur lesquels ils reposent. Des combinaisons linéaires et séquentielles de scores sont étudiées. Notons que la combinaison séquentielle de deux scores consiste à appliquer un ordonnancement à un ensemble d'entités préalablement filtrées selon un autre score.

3 Expérimentations

Nos approches sont évaluées sur un ensemble de 50 articles scientifiques collectées manuellement sur le site *ScienceDirect* (Lentschat et al, 2020b). L'extraction des entités présente un rappel de 0.85, avec peu de fluctuations selon les entités considérées. La précision de 0.41 est sujette à davantage de variations, avec une moyenne à 0.47 pour les instances symboliques et à 0.14 pour les quantitatives. Les résultats montrent que les scores sémantiques et ceux fondés sur la fréquence sont efficaces pour filtrer les entités symboliques. La combinaison linéaire de scores n'a pas offert de gains significatifs. Opérer une combinaison séquentielle permet au contraire d'améliorer la mesure de pertinence des entités symboliques en particulier pour les noms d'emballages. Les instances de concepts quantitatifs, qui présentent des scores de précisions plus bas, peuvent être filtrées par l'usage de scores de type icf utilisant les segments textuels identifiés dans les publications.

4 Conclusion

Les expérimentations réalisées dans cet article montrent que le nombre important de faux-positifs peut être réduit par l'utilisation de scores de pertinences sémantiques, lexicaux et leurs combinaisons. Ces scores peuvent être utilisés afin de filtrer les résultats et déterminer un compromis entre exhaustivité et validité.

Références

Salton, G et McGill, M.J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Harispe, S. (2014). *Knowledge-based semantic measures : From theory to applications*. PhD.

Lentschat, M., Buche, P., Dibie-Barthelemy, J. et Roche, M. (2020a). SciPuRe : a new representation of textual data for entity identification from scientific publications. In *Proceedings of the 10th WIMS conference*. pp. 220–226.

Lentschat, M. and Buche, P. et Menut, L. (2020b). Transmat gold standard. *CIRAD Dataverse doi :10.18167/DVN1/U7HK8J*.

Wang, D. et Zhang, H. (2010). Inverse-category-frequency based supervised term weighting scheme for text categorization *arXiv preprint arXiv :1012.2609*.