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