Extraction et reconstitution de relation n-Aires issues d'articles scientifiques en domaine expérimental guidées par une Ressource Termino-Ontologique

Martin Lentschat*,**, Patrice Buche*, Juliette Dibie***, Mathieu Roche**

*IATE, Univ Montpellier, INRAE, Institut Agro, Montpellier 34060, France
**TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, France
***Univ. Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris, France

1 Contexte de contribution

Le but de ce travail est l'extraction d'informations expérimentales dans le domaine des emballages alimentaires et leur formalisation dans des relations n-Aires. Les difficultés d'extraction de ces relations proviennent de la dispersion des instances d'arguments dans l'ensemble des documents et de la multiplicité des instances à mettre en relation. Nous comparons différentes méthodes s'appuyant sur une nouvelle représentation multi-descripteurs des relations n-Aires partielles présentes dans les tableaux des articles (STaRe - Scientific Table Representation) pouvant être complétées par les instances d'arguments issues du texte.

Notre méthode de reconstruction des relations n-Aires est guidée par une Ressource Termino-Ontologique (RTO) et repose sur des représentations multi-descripteurs des instances d'arguments issues du texte (SciPuRe - Scientific Publication Representation) (Lentschat et al., 2022) et des instances de relations n-Aires partielles issues des tableaux (STaRe - Scientific Table Representation, e.g. Table 1). STaRe comprend plusieurs types de descripteurs : (i) *ontologiques* qui classifient la relation selon le formalisme de la RTO, (ii) *structurels* qui donnent le contexte du tableau et à sa place dans le document et (iii) *lexicaux* qui indique la forme textuelle des instances d'arguments composant la relation.

Les relations obtenues sur la base de ces représentations sont reconstituées selon trois stratégies exploitant les descripteurs de STaRe et de SciPuRe. L'*approche structurelle* exploite la structure des articles. L'*approche fréquentiste* recherche les co-occurrences fréquentes. L'*approche par plongements lexicaux* utilise les scores de similarité donnés par des modèles de langage.

2 Résultats et discussions

Dans l'évaluation des approches de complétion des instances de relations partielles, nous considérons deux aspects : le filtrage des instances d'arguments candidates et la sélection des candidats. Nos expérimentations réalisées sur un jeu de données dédié (Lentschat et al., 2021) ont montré que les mesures proposées permettent d'améliorer à la fois la précision et le rappel

	Descripteur	Valeur
ONTOLOGIQUE	Relation	H2O_Permeability_Relation
	Result_Argument	***SciPure***
	Arguments	Target: H2O_Perm. | Node: H2O_Perm. | Original_Value: $1.27 * 10^2$ $cm^3 mm^{-2} s^{-1} bar$ | Attached_Value: *Water Perm.* Packaging | Chitosan | *Chitosan films* | *Chitosan films* Method | ∅ | ∅ | ∅ R_H | ∅ | ∅ | ∅ Temperature | Temperature | $25°C$ | *Temp.* ($°C$) Thickness | ∅ | ∅ | ∅
STRUCT.	Table	*Table 3*
	Caption	*Water permeability of tested packaging at* $25°C$
	Segment	*Results and Discussion*
	Document	*Barrier properties of chitosan coated polyethylene*
	DOI	*10.1016/j.memsci.2012.02.037*

TAB. 1: Exemple de représentation STaRe d'une instance de relation n-Aire partielle

pour l'identification des relations n-Aires. L'approche fréquentiste montre les meilleurs résultats ($f - score = .48$) en sélectionnant un unique candidat. Lorsque le travail est assisté par un expert qui sélectionnera l'instance d'argument parmi différents candidats proposés, l'approche par plongements lexicaux apparaît comme la meilleure avec des modèles BERT ($f - score = .59$ et $.65$ avec respectivement trois et cinq candidats proposés). Dans le cadre demandant une intervention de l'expert pour trier parmi dix candidats, l'approche structurelle se distingue avec un f-score de $.74$.

Références

Lentschat, M., P. Buche, J. Dibie-Barthelemy, et M. Roche (2022). Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications. *Int. J. Intelligent Information and Database Systems 15*(1), 78–103.

Lentschat, M., P. Buche, L. Menut, et R. Guari (2021). TRANSMAT tables data - CIRAD Dataverse. https://doi.org/10.18167/DVN1/GCZBC9.

Summary

We present different approaches to construct n-Ary relation instances from scientific articles in experimental domains (i.e. food packaging) driven by a Termino-Ontological Resource dealing with the food packaging domain. Our method starts with the identification of partial n-Ary relations in the document tables and seeks to complete them with argument instances from text of the articles. Our approach is based on a new multi-feature representation (i.e. STaRe - Scientific Table Representation) and the n-Ary relation reconstruction based on text-mining methods.