Traçabilité de l'information, de l'extraction à l'exploitation

Claire Laudy*, Charlotte Jacobé De Naurois*, Bénédicte Goujon*

*THALES, 1 avenue Agustin Fresnel, 91767 Palaiseau, France
<prénom>.<nom>@thalesgroup.com

1 Une chaîne intégrée de l'extration à l'exploitation

Afin de fournir un support à la prise de décision, nous proposons une chaîne fonctionnelle permettant d'extraire des informations à partir de textes et de les aggréger au sein d'un réseau d'informations sémantiques. Pour illustrer notre approche, nous proposons un exemple d'extraction d'informations précises (noms de composants chimiques et valeurs sur les propriétés associées) à partir d'articles scientifiques. Dans cet exemple, nous nous concentrons sur l'extraction et la fusion d'information concernant la molécule ABS/ZnO et ses propriétés.

Phrase1 : *The tensile strength for ABS/ZnO line samples were 23.3, 24.19, and 28.24 MPa for the infill density of 50%, 75%, and 100%, respectively.*

Phrase2 : *The tensile strength for ABS/ZnO rectilinear samples were 20.21, 20.32, and 22.19 MPa for the infill density of 50%, 75%, and 100%.*

La première étape d'extraction d'informations est réalisée par le module d'annotation de la plateforme STRASS (fig. 1). La plateforme STRASS est centrée sur l'apprentissage de patrons linguistiques à partir de textes annotés manuellement par un expert métier. Elle vise l'annotation automatique de textes et l'export des informations extraites (Goujon (2021)). En sortie de ce module, des graphes composés de nœuds entités isolés ou liés par une relation sont générés.

FIG. 1 – *Annotation de la Phrase1 de notre exemple illlustratif.*

Cet ensemble de petits graphes est aggrégé par fusion avec InSyTo. InSyTo est une bibliothèque d'algorithmes de manipulation de graphes conceptuels qui peuvent être combinés afin de fournir des fonctions avancées (Laudy (2017)). En sortie de ce second module, l'ensemble des informations extraites des textes initiaux est aggrégé au sein d'un graphe d'informations.

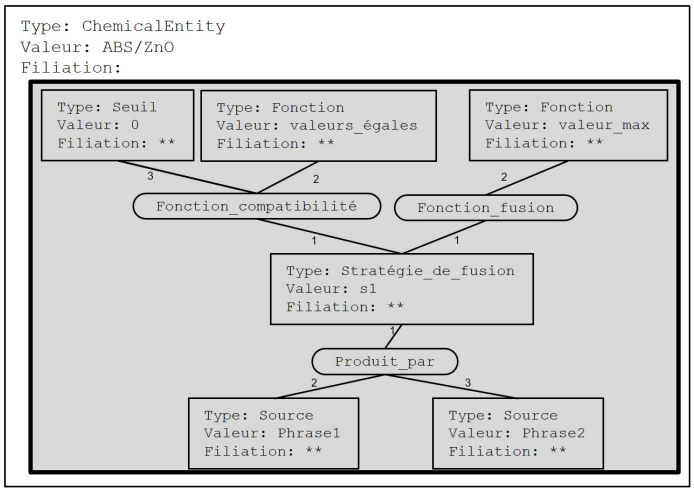

FIG. 2 – *Concept ABS/ZnO après fusion tracée des 2 phrases de l'exemple.*

2 Tracer la fusion d'informations

La fusion d'InSyTo est augmentée d'une capacité de traçabilité, afin de garder un lien entre les unités d'informations agrégées au sein du graphe et les textes sources. Si la traçabilité est souvent une fonction disponible dans les systèmes d'extraction d'informations issues de texte, une fois cette information extraite, il est plus rare que les liens des informations unitaires vers leur source soient conservés au fil des transformations.

Notre approche de la traçabilité est basée sur l'utilisation de graphes conceptuels imbriqués afin d'exprimer, pour chaque composant élémentaire de l'information, un graphe de filiation qui sauvegarde l'ensemble de l'"historique" de l'élément d'information tout au long de ses évolutions. Les graphes conceptuels imbriqués sont une extension des graphes conceptuels basiques (Chein et Mugnier (2008)). Ils sont utilisés afin de fournir différents niveaux de connaissance liés aux concepts d'un graphe. Alors que les concepts et les relations liées à un concept fournissent des informations contextuelles externes sur le concept, des informations internes sur le concept peuvent être fournies sous forme de graphe, imbriqué dans le concept.

Pour suivre toutes les fusions de concepts, nous proposons d'ajouter un *graphe de filiation imbriqué* à l'intérieur de chaque nœud concept qui permet de reconstruire l'ensemble du processus de fusion appliqué aux données (fig. 2).

Références

Chein, M. et M.-L. Mugnier (2008). *Graph-based Knowledge Representation : Computational Foundations of Conceptual Graphs*. Springer.

Goujon, B. (2021). Extraction d'informations spécifiques à partir de textes avec peu de textes d'apprentissage. In *TextMine*.

Laudy, C. (2017). Rumors detection on social media during crisis management. In *ISCRAM*.