Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Elaboration d'un corpus d'apprentissage à partir d'articles de recherche en chimie

Bénédicte Goujon
Thales R&T France, 1 avenue Fresnel, 91767 Palaiseau Cedex, France
benedicte.goujon@thalesgroup.com

RESUME

Dans le cadre d'un projet mené en 2021, un objectif consistait à extraire automatiquement des informations à partir d'articles de recherche en chimie des matériaux : des valeurs associées à des propriétés pour différents composants chimiques. Le travail présenté ici décrit les étapes de la construction du corpus textuel d'apprentissage, annoté manuellement par des experts du domaine selon les besoins identifiés dans le projet, pour une utilisation ultérieure par des outils d'extraction d'informations.

ABSTRACT

Here the title in English.
In a project conducted in 2021 for a chemistry consortium, an objective was dealing with the automatic extraction of the following specific information from research papers: numerical values associated to properties for various chemical components. The work presented here describes the steps for building the learning corpus manually annotated by domain experts according to the project specific needs, for a later use by extraction information tools.

MOTS-CLES : extraction d'informations, annotation manuelle, modélisation, articles de chimie
KEYWORDS: information extraction, manual annotation, modelling, chemistry papers

1 Introduction

Dans le cadre d'un projet mené en 2021 avec un consortium de chimie, un objectif consistait à extraire automatiquement les informations spécifiques suivantes à partir d'articles de recherche : des valeurs numériques associées à des propriétés pour différents composants chimiques. L'idée était de permettre ensuite des recherches dans la base de données regroupant les informations extraites, visant certains composants chimiques, certaines propriétés, ou avec des fourchettes sur des valeurs, pour repérer par exemple les valeurs obtenues sur une propriété ou pour identifier les composants chimiques ayant des valeurs recherchées sur une propriété, en proposant un lien avec le texte source afin de faciliter la validation des informations extraites.

Une quinzaine d'expert.es en chimie, avec des compétences et expertises variées, étaient disponibles pour participer aux différentes discussions et procéder à l'annotation manuelle des corpus. La plateforme INCEpTION (Klie et al., 2018), qui intègre l'outil d'annotation WebAnno, a été choisie pour permettre l'annotation manuelle d'extraits textuels par plusieurs annotateurs et annotatrices, en lien avec un modèle d'annotation partagé non modifiable devant être défini.

Alors que de nombreux travaux ciblent la génétique ou le biomédical (Islamaj et al. 2022) (Wei et al., 2016), la chimie des matériaux est pauvre en support pour la construction de corpus d'apprentissage annotés manuellement. Concernant les corpus annotés existants, il est difficile d'en trouver qui contiennent des entités de type Composant Chimique (Chemical) et des entités de type Propriété (Property). Le corpus bc5cdr[1], disponible sur HuggingFace, propose le type d'entité « Chemical », ainsi que « Disease », mais il ne couvre pas des textes en lien avec la chimie des matériaux comme visés ici. Un autre corpus, CHEMDNER (Krallinger et al., 2015)[2], contient des annotations de noms de composants chimiques, mais ce n'est pas non plus un corpus lié à la chimie des matériaux et les types d'entité utilisés (abbreviation, family, formula, identifier, multiple, systematic et trivial), qui correspondent à différentes formes textuelles des noms de composants chimiques, ne sont pas directement pertinents dans notre approche.

Le travail présenté ici aborde des étapes et réflexions ayant permis de construire un corpus d'apprentissage, annoté manuellement par des experts du domaine selon les besoins identifiés, pour une utilisation ultérieure par des outils d'extraction d'informations utilisant l'apprentissage symbolique de patrons linguistiques tels que STRASS (Goujon, 2021) ou l'apprentissage à base de réseaux de neurones comme les Transformers, en vue d'alimenter une base de données regroupant des informations issues de travaux de recherches ciblés.

2 La construction d'un corpus d'extraits textuels à annoter

Le consortium de chimie s'est intéressé aux propriétés mécaniques telles que la résistance à la traction (« tensile strength ») de composants chimiques en lien avec le polymère acrylonitrile butadiène styrène ou ABS. Un premier corpus contenant 43 documents a été récupéré de bases d'articles de recherche en ligne par différents experts. Les documents obtenus, au format pdf, font en moyenne 15 pages avec un minimum de 9 pages et un maximum de 26 pages.

Afin de filtrer cet ensemble textuel pour ne fournir aux personnes devant annoter que des phrases potentiellement porteuses d'informations recherchées (valeurs sur des propriétés), une experte a parcouru chaque document initial pour en extraire des sous-paragraphes. Résultant de cette étape, un corpus de 89 extraits d'articles au format texte a été obtenu. Chaque extrait, contenant entre 2 et 10 phrases, a fait l'objet d'un nettoyage manuel afin d'obtenir des textes non bruités (sans sauts de ligne inutiles ou références bibliographiques). Les tableaux des textes sources, contenant de nombreuses valeurs recherchées, n'ont pas été retenus pour les annotations manuelles. Ils ont fait l'objet d'annotations automatiques, via des requêtes (proches de règles symboliques) portant sur le contenu textuel des en-têtes de lignes et colonnes, avec l'outil I2E[3] de Linguamatics utilisé en parallèle du travail présenté ici. Voici trois sous-extraits pour illustrer nos remarques, tirés de (Dul et al., 2018) (Aw et al. 2018) (Verbeteen et al., 2021) :

[1] https://huggingface.co/datasets/tner/bc5cdr
[2] https://huggingface.co/datasets/bigbio/chemdner
[3] https://www.linguamatics.com/products/i2e

- Extrait 1 : « … for ABS/graphene composites. Interestingly enough, these composites show elastic modulus of about 7362 MPa and tensile strength of about 44 MPa. »

- Extrait 2 : « 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. For CABS/ZnO line samples, the tensile strength improved 6.3% to 10.31 MPa when infill density changed from 50% to 100%. »

- Extrait 3 : « The set at v_p = 5 mm/s has the lowest elastic modulus of E = 1.9–2.0 GPa, while samples fabricated at v_p = 20 mm/s have values around 2.2 GPa. »

En analysant les extraits textuels obtenus, on a pu observer que peu de phrases portent explicitement toutes les informations recherchées, et que certaines phrases contiennent plusieurs informations simultanément, ce qui, dans les deux cas, pénalise l'efficacité des annotations manuelles. En effet, d'une part, le recours aux coréférences est assez fréquent dans les articles assez longs, notamment pour améliorer la lisibilité des textes, comme avec « these composites » dans l'extrait 1. Or, dans notre contexte, le recours à des extraits textuels et à des experts du domaine sans expertise en langage naturel nous ont amenés à ne pas gérer l'annotation des coréférences. D'autre part, certaines phrases comparent des valeurs obtenues, dont les détails sont présentés dans des tableaux, en se focalisant sur les améliorations obtenues (augmentations ou diminutions), comme dans l'extrait 2 avec « the tensile strength improved 6.3% », où la valeur initiale n'est pas explicitée. Au final, l'annotation manuelle et l'annotation automatique visée ont peu de chance d'être suffisantes pour l'extraction de toutes les informations visées, et devraient nécessiter des relectures et complétions.

3 La construction du modèle d'annotation

Des ontologies existent pour couvrir le domaine très large de la chimie, telles que ChEBI (Chemical Entities of Biological Interest) (Hastings, 2016), qui contient 46 000 entrées. Ces modèles très détaillés ne sont pas adaptés à l'annotation manuelle pour l'extraction d'informations spécifiques. D'une part, il n'est pas envisageable de proposer pour l'annotation manuelle des modèles contenant des milliers de concepts, dont la structure et le contenu devraient être maîtrisés pour la production d'annotations manuelles homogènes et de bonne qualité. De plus, les relations répertoriées dans ChEBI ne couvrent pas l'association de valeurs numériques à des propriétés de composants chimiques mais principalement l'organisation des concepts : « is a », « has part »… Enfin, de nombreux noms de composants chimiques cités dans les articles de recherche ne correspondent pas à des entités chimiques préexistantes mais font juste référence à des compositions spécifiques testées dans les expériences décrites, tels que « ABS/graphene » (extrait 1) ou « ABS/ZnO » (extrait 2).

Dans ce contexte, nous avons choisi de définir notre propre modèle d'annotation, le plus simple possible, centré sur le besoin visé. Initialement, nous avons proposé le modèle suivant : un type d'entité central « Chemical » et deux types complémentaires « Property » et « Value » complétés par des relations de type « has_property » et « has_value ». Cependant, la confrontation avec les premiers extraits textuels a rapidement fait remonter certaines limites. En effet, peu de phrases sont réellement centrées sur le nom explicite du composant chimique, avec une valeur et un nom de propriété. Ainsi dans la phrase 1, le nom du composant chimique n'est pas précisé. L'objectif étant l'extraction de valeurs sur les propriétés, il nous a ensuite semblé pertinent de centrer le modèle sur les entités de type Value, mais cette modélisation n'a pas été validée par certains experts, et nous avons pu observer que certaines valeurs n'étaient pas toujours explicitement données dans les textes (voir extrait 3). Enfin, les conditions d'obtention des valeurs, telles que les conditions de

températures par exemple, sont très importantes pour expliquer des différences de valeurs. Nous avons dû ajouter le nouveau type d'entité « Condition », associé à des valeurs numériques (extrait 2 : « infill density of 50% ») ou non (extrait 2 : « line » correspond à la mise en forme du matériau). Ce type d'entité regroupe des informations très variées, allant de températures à des noms d'appareils de mesure utilisés.

L'annotation manuelle de textes permet de tracer des relations binaires orientées entre deux entités, or dans le besoin visé ici plusieurs relations différentes peuvent être définies entre les quatre types d'entités, puisqu'une valeur correspond à une propriété pour un composant chimique avec une ou plusieurs conditions et valeurs associées. Afin de simplifier la tâche d'annotation, et pour éviter d'avoir différentes annotations acceptables pour une même phrase, nous avons retenu les quatre types de relations suivants pour l'annotation manuelle, plutôt centrés autour du concept « Property », avec entre parenthèses le ou les types d'entités source possibles suivis du type d'entité cible : Is_property_of(Property, Chemical), Has_value([Property, Condition], Value), Measured_at([Property,Value], Condition) et With_condition(Chemical, Condition).

4 La préparation à l'annotation manuelle

Les chimistes disponibles pour réaliser l'annotation manuelle du corpus n'ont jamais eu à effectuer ce type de tâche. Un tutoriel a été mis en place et leur a été présenté afin de diriger leur façon d'annoter les extraits d'articles avec l'outil INCEpTION. Voici quelques commentaires qui ont été exprimés pour la production des annotations manuelles et leur exploitation :

- Une mention d'entité se compose d'un ou plusieurs mots consécutifs d'une même phrase, et ne peut être recouverte par une autre. En chimie, un nom de matériau pouvant être formé des noms de ses composants, comme dans « ABS/ZnO », le choix a été fait d'annoter l'expression complète comme « Chemical », l'identification de ses sous-composants « ABS » et « ZnO » étant transférée à des post-traitements. Quand la phrase ne contient qu'une expression peu spécifique, comme « these composites » dans l'extrait 1, cet élément doit être annoté comme un composant chimique. Dans l'extrait 3, l'annotation de « the set at v_p = 5 mm/s » est délicate, l'expression désignant un échantillon de type « Chemical » tout en contenant des informations de type « Condition ». Concernant les valeurs, nous avons fait le choix de fusionner dans « Value » la valeur numérique et l'unité, comme « 44 MPa » ou « 50% », afin de simplifier l'annotation manuelle. Dans certains cas (extrait 2), l'unité n'est précisée qu'une fois pour plusieurs valeurs et doit être récupérée via des post-traitements. Par ailleurs, les valeurs complexes telles que « 1.9–2.0 GPa » (extrait 3) doivent être annotées comme « Value » et gérées plus finement avec des post-traitements.

- Une relation est binaire et relie deux entités qui appartiennent à une même phrase. Cela permet d'obtenir un premier ensemble d'informations. Des post-traitements de type fusion seront apportés d'une part pour relier des informations issues de différentes phrases, d'autre part pour réattribuer les bonnes valeurs aux informations multiples sous forme de listes comme dans la phrase 2 : « 24.19 » correspond à « 75% » de la condition « infill density ».

5 Les résultats obtenus

Après l'annotation de chaque extrait textuel par un ou plusieurs annotateur(s), 288 extraits annotés ont été obtenus, certains étant peu annotés et sans forcément d'homogénéité entre les annotateurs. Pour quantifier les résultats obtenus, dans un sous-ensemble de 58 extraits distincts bien annotés, contenant environ 270 phrases, on a obtenu environ 380 mentions de « Chemistry » et de « Property » et environ 600 mentions de « Condition » et de « Value ». Côté relations, on a observé environ 200 mentions de « has_value » et « with_condition », et environ 40 mentions de « is_property_of » et « measured_at ».

Des post-traitements ont été ajoutés afin d'enrichir les annotations manuelles avec les unités complétant les valeurs via une relation « has_unit » par exemples, et pour améliorer l'extraction d'informations suite à des annotations automatiques.

Références

Aw Y. Y., Yeoh C. K., Idris M. A., Teh P. L., Hamzah K. A., Sazali S. A. (2018). Effect of Printing Parameters on Tensile, Dynamic Mechanical, and Thermoelectric Properties of FDM 3D Printed CABS/ZnO Composites. Materials (Basel). 2018 Mar 22;11(4):466.

Dul S, Fambri L, Pegoretti A. (2018) Filaments Production and Fused Deposition Modelling of ABS/Carbon Nanotubes Composites. Nanomaterials (Basel). 2018 Jan 18;8(1):49. doi: 10.3390/nano8010049.

Goujon B. (2021). Extraction d'informations spécifiques à partir de textes avec peu de textes d'apprentissage, in *TextMine2021*, Montpellier.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2015). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. DOI: 10.1093/nar/gkv1031

Islamaj R., Leaman R., Cissel D., Coss C., Denicola J., Fisher C., Guzman R., Gokal Kochar P., Miliaras N., Punske Z., Sekiya K., Trinh D., Whitman D., Schmidt S., Lu Z. (2022). NLM-Chem-BC7: manually annotated full-text resources for chemical entity annotation and indexing in biomedical articles, Database, Volume 2022, baac102. DOI: 10.1093/database/baac102

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. (2018): The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA.

Krallinger, M., Rabal, O., Leitner, F. et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 7 (Suppl 1), S2 (2015). DOI: 10.1186/1758-2946-7-S1-S2

Verbeeten, W.M.H.; Arnold-Bik, R.J.; Lorenzo-Bañuelos, M. (2021). Print Velocity Effects on Strain-Rate Sensitivity of Acrylonitrile-Butadiene-Styrene Using Material Extrusion Additive Manufacturing. Polymers 2021, 13, 149. DOI: 10.3390/polym13010149

Wei C.-H., Peng Y., Leaman R., Davis A. P., Mattingly C. J., Li J., Wiegers T. C. , Lu Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task, Database, Volume 2016, 2016, baw032. DOI: 10.1093/database/baw032