# Découverte et extraction d'arguments de relations n-aires corrélés dans les textes

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

# Découverte et extraction d'arguments de relations n-aires corrélés dans les textes

Soumia Lilia Berrahou[*,**], Patrice Buche[**], Juliette Dibie[***], Mathieu Roche[*,****]

[*]LIRMM - 860, rue de Saint Priest, 34095 Montpellier, France
berrahou@lirmm.fr,
[**]INRA - UMR IATE - 2, place Pierre Viala, 34060 Montpellier, France
[***]AgroParisTech/INRA - UMR MIA - Université Paris-Saclay, 75005 Paris, France
[****]CIRAD - TETIS - 500, rue J.F. Breton, 34093 Montpellier, France

**Résumé.** Dans cet article, nous présentons une méthode hybride combinant des approches de fouille de données et des analyses syntaxiques afin de découvrir et extraire automatiquement des données dans les textes. Ces données sont modélisées sous forme de relations n-aires représentées dans une Ressource Termino-Ontologique (RTO). La relation n-aire relie un objet étudié (e.g. un emballage) à ses caractéristiques sous forme d'arguments (e.g. son épaisseur). Dans les textes, les arguments de l'objet étudié sont quantitatifs, associés à leurs attributs, une valeur numérique et une unité de mesure, à extraire pour peupler l'ontologie de nouvelles instances. La méthode proposée repose sur la découverte de relations implicites d'expression des arguments dans les textes en utilisant les motifs et règles séquentiels puis, sur l'intégration de relations syntaxiques d'intérêt dans les motifs découverts afin de construire des patrons linguistiques d'extraction d'arguments corrélés. Les expérimentations ont été menées sur un corpus du domaine des emballages et consistent à extraire les résultats expérimentaux de perméabilités des emballages alimentaires.

## 1 Introduction

Les publications scientifiques, disponibles à partir de bibliothèques spécialisées en ligne, sont une source d'information précieuse à exploiter et analyser par les experts du domaine pour, par exemple, paramétrer des modèles d'aide à la décision (Guillard et al., 2015). Le nombre d'articles publiés et disponibles en ligne est toujours grandissant. Aujourd'hui, le défi n'est pas de trouver l'information mais d'être en mesure de l'identifier et l'extraire automatiquement, notamment dans la perspective du développement de l'open access, en prenant en compte la complexité des données textuelles. Cette problématique rassemble plusieurs domaines depuis plusieurs années dont le domaine du Traitement Automatique des Langues (TAL) et celui de l'Ingénierie des Connaissances (IC). En effet, identifier et extraire l'information pertinente se révèlent être des tâches complexes car la grande majorité des documents collectés est, en général, partagée en langage naturel. Le langage naturel, du fait de sa richesse et de sa variété est souvent difficile à appréhender. Un même mot revêt plusieurs significations, une même

information peut s'exprimer de multiples manières, souvent implicitement, générant des ambiguïtés difficiles à cerner automatiquement par les machines. Les premiers travaux en extraction d'information se sont portés, tout d'abord, sur l'extraction des entités nommées puis sur l'extraction de relations entre entités d'intérêt dans les textes, c'est-à-dire que l'ensemble des entités (ou arguments) sont connectées entre elles par la relation qui les associe. Dans le cadre des travaux menés sur l'extraction automatique de relations dans les documents, nous trouvons deux grands axes de recherche, l'extraction des relations binaires et l'extraction des relations n-aires.

Dans le domaine biomédical, de nombreux travaux se sont intéressés à l'extraction des relations binaires. Les premières approches se sont concentrées sur la découverte de co-occurrences et la construction manuelle de patrons d'extraction (Huang et al., 2004) entre entités en se situant dans des contextes linguistiques restreints. Les règles exprimées sous forme d'expressions régulières utilisant des mots spécifiques ou des catégories grammaticales ont permis la construction de patrons linguistiques ou syntaxiques (Hawizy et al., 2011), (Proux et al., 2000), (Hao et al., 2005), (Raja et al., 2013). Ces méthodes obtiennent des taux de précision entre 0.80 et 0.94 et des taux de rappel entre 0.6 et 0.86 mais la construction manuelle des règles nécessitent toutefois, un effort humain conséquent. Les travaux intégrant des approches d'apprentissage, e.g. machines à vecteurs de support (SVMs) (Minard et al., 2011) ont rapidement émergé (Rosario et Hearst, 2005), (Zhang et al., 2011), (Miwa et al., 2009), (Van Landeghem et al., 2009), en proposant des méthodes fondées essentiellement sur la classification (Rosario et Hearst, 2004). Les méthodes d'apprentissage obtiennent en général de meilleurs résultats avec des taux de précision entre 0.6 et 0.9 et des taux de rappel entre 0.21 et 0.89 mais obligent à constituer un ensemble de données annotées et restituent des modèles appris qui ne sont pas directement interprétables par l'utilisateur.

L'extraction des relations n-aires, c'est-à-dire des relations faisant intervenir plus de deux arguments (ou entités), est un problème plus complexe à résoudre. Tout d'abord, la relation n-aire fait intervenir plusieurs arguments de différents types et ces arguments, à regrouper dans la relation à identifier, peuvent se trouver dans une phrase mais également dans l'ensemble du document selon des formes d'expression variées, rendant la tâche d'extraction automatique particulièrement difficile. Les travaux de (Mcdonald et al., 2005) proposent d'identifier les relations n-aires se situant dans une phrase en classant les paires d'entités en relations binaires puis, en construisant la relation complexe à partir du graphe de relations constitué (sélection des cliques maximales dans le graphe). Les travaux suivants décomposent la problématique d'extraction des relations n-aires en trois étapes : (i) l'identification des entités de la relation puis (ii) la détection de l'élément déclencheur en utilisant des méthodes à base de dictionnaires, de règles en construisant des patrons à partir des arbres de constituants (Le Minh et al., 2011) ou, en utilisant des méthodes par apprentissage (Buyko et al., 2009), (Bui et Sloot, 2011), (Björne et al., 2009), (Zhou et al., 2014) pour déterminer si un mot de la phrase est déclencheur ou pas. Enfin, (iii) les patrons sont alors utilisés afin de relier les arguments autour de l'élément déclencheur en décomposant la problématique en plusieurs relations binaires mais avec une perte sensible de la précision.

Dans le cadre de nos travaux, nous nous interrogeons sur l'existence de relations implicites d'expression des arguments dans les textes afin d'en faciliter l'identification et la mise en relation. Dans ce contexte de recherche, nous nous intéressons au potentiel des techniques de fouille de données à faire émerger des règles implicites dans les données. Ces approches ont

déjà été proposées avec succès sur les textes en découvrant des relations sémantiques entre entités dans le but d'enrichir des ontologies (Di-Jorio et al., 2008), pour la découverte de patrons linguistiques sans utiliser de ressources externes (Béchet et al., 2012), (Cellier et al., 2015) ou encore l'utilisation des règles d'association et motifs séquentiels pour catégoriser des textes en proposant des règles compréhensibles et réutilisables par l'utilisateur (Jaillet et al., 2006), contrairement aux modèles d'apprentissage classiques non interprétables.
Dans cet article, nous nous intéressons à l'extraction de données modélisés sous forme de relations n-aires. Les instances de relations sont exprimées dans les textes et décrivent un objet étudié, représenté sous forme d'argument symbolique, à des arguments représentant les mesures effectuées sur cet objet. Ces arguments sont quantitatifs et sont associés à leurs attributs dans l'instance, une valeur numérique et une unité de mesure. La phrase (1) restitue un extrait d'instance de relation n-aire à identifier et extraire du texte : L'objet étudié est un emballage *polypropylene* et ses arguments quantitatifs *thickness, oxygen permeability, temperature* et *relative humidity (RH)* sont associés à leur valeur numérique et leur unité de mesure. La phrase (2) restitue un autre extrait d'instance de relation n-aire : L'objet étudié est un avion *A380-800* et ses arguments quantitatifs *capacité, rayon d'action, vitesse* sont associés à leur valeur numérique et leur unité de mesure.

(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 $\mu$m thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 ° C and 0 % RH

(2) L'A380-800 a une capacité de 150 tonnes de transport, un rayon d'action de 15 400 kilomètres, ce qui lui permet de voler de New York jusqu'à Hong Kong sans escale, à la vitesse de 900 km/h jusqu'à 1012 km/h.

Ces instances sont complexes à identifier et extraire automatiquement des textes, d'une part parce que l'expression des arguments se fait rarement sur une seule phrase mais plus souvent sur plusieurs phrases et fréquemment de manière implicite, d'autre part du fait de la richesse du vocabulaire employé pour les décrire, et également du fait de la structure même des instances d'arguments quantitatifs qui varient par leurs attributs, i.e. la valeur numérique et l'unité de mesure, selon les mesures effectuées sur l'objet étudié.
Dans cet article, nous proposons une méthode combinant des approches de fouille de données et de relations syntaxiques pour extraire les instances des textes. Les approches de fouille de données sont utilisées pour découvrir les règles implicites d'expression des arguments dans les textes et ainsi mieux appréhender la structure variante des instances d'arguments. Le processus de découverte est fondé sur une nouvelle représentation des données s'appuyant sur les concepts définis dans une Ressource Termino-Ontologique de domaine (RTO). Les relations syntaxiques fondées sur des grammaires concises permettent de représenter les structures linguistiques des instances dans les textes. La combinaison des approches permet alors de construire des patrons d'extraction fondés à la fois sur des structures linguistiques et sur la structure des arguments liés par la relation n-aire.
L'article est structuré comme suit : la section 2 pose les définitions classiques de fouille de données, puis décrit, d'une part, la nouvelle représentation des données s'appuyant sur une RTO pour augmenter l'expressivité des données, d'autre part, le processus d'extraction des connaissances proposé afin d'extraire les règles et motifs séquentiels, et découvrir ainsi les

| Ville | Mois | items |
|---|---|---|
| Nîmes | 01/2011 | Humidité=Faible, Soleil |
| Montpellier | 02/2011 | Soleil |
| Nîmes | 02/2011 | Chaleur=Forte |
| Montpellier | 03/2011 | Humidité=Faible, Chaleur=Forte, Soleil |
| Nîmes | 04/2011 | Chaleur=Faible, Vent |

TAB. 1: Base de données $\mathcal{DB}$

| Ville | Séquence |
|---|---|
| Nîmes | <(Soleil Humidité=Faible)(Chaleur=Forte)(Chaleur=Faible Vent)> |
| Montpellier | <(Soleil)(Humidité=Faible, Chaleur=Forte, Soleil)> |

TAB. 2: Base de séquences $\mathcal{S}$ pour chaque ville

relations implicites d'expression des arguments dans les textes. La section 3 décrit notre proposition d'approche hybride qui prend en compte des relations syntaxiques spécifiques pour construire des patrons linguistiques à partir des motifs séquentiels découverts pour l'extraction d'arguments corrélés dans les textes. La section 4 présente dans un premier temps les résultats obtenus concernant les règles et motifs d'expression des arguments corrélés, puis, dans un second temps, les résultats d'extraction obtenus par approche hybride. La section 5 conclut sur les travaux menés et les perspectives envisagées.

# 2 Découverte d'arguments corrélés

Dans cette section, nous nous intéressons aux relations implicites d'expression des arguments dans les textes et nous proposons d'utiliser les approches de fouille de données pour les découvrir. Dans un premier temps, nous posons les définitions classiques de fouille de données utilisées dans l'article. Dans un second temps, nous présentons une nouvelle représentation des données, guidée par une RTO de domaine, afin d'augmenter l'expressivité des arguments recherchés. Finalement, nous détaillons le processus d'extraction de connaissances proposé pour la découverte de motifs fréquents d'arguments corrélés dans les textes bruts. Les principes et les fondements sont décrits dans la section qui suit.

## 2.1 Définitions

Les définitions suivantes sont détaillées dans (Agrawal et Srikant, 1995) et présentées à partir d'un exemple extrait de (Fabrègue et al., 2012). La base de données $\mathcal{DB}$ représentée dans le tableau 1 restitue l'ensemble des transactions. Chaque transaction représente un ensemble d'événements ou items ayant eu lieu dans des villes, à des dates différentes. Pour chaque ville, une séquence est générée. Les séquences sont illustrées dans le tableau 2.

**Définition 1** *(Séquence)*
*Soit $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ l'ensemble des items. Un itemset est un ensemble non vide, non or-*

*donné d'items noté ($I_1, I_2, ..., I_k$) où $I_j \in \mathcal{I}$. Une séquence est une liste ordonnée, non vide d'itemsets notée $< IS_1 IS_2 ... IS_p >$ où $IS_j \in \mathcal{IS}$, avec $\mathcal{IS}$ l'ensemble d'itemsets.*

L'extraction des motifs séquentiels consiste à rechercher l'ensemble des sous-séquences fréquentes extraites à partir de la base de séquences.

**Définition 2** *(Sous-séquence)*
*Une séquence A = $< IS_1 IS_2 ... IS_p >$ est une sous-séquence d'une autre séquence B = $< IS'_1 IS'_2 ... IS'_m >$ (A $\preceq$ B) si $p \leq m$ et s'il existe des entiers $j_1 < j_2 < ... < j_k < ... < j_p$ tels que $IS_1 \subseteq IS_{j_1}, IS_2 \subseteq IS_{j_2}, ..., IS_p \subseteq IS_{j_p}$.*

**Exemple 1**
*Considérons les séquences du tableau 2, la séquence S = $< (Soleil)(Chaleur = Forte) >$ est supportée par les séquences $S_{Nimes}$ et $S_{Montpellier}$. Nous avons $S \preceq S_{Nimes}$ et $S \preceq S_{Montpellier}$.*

Un motif séquentiel est une sous-séquence fréquente caractérisée par un support, représentant le nombre d'occurrences du motif dans $\mathcal{S}$. Seuls ceux ayant un support supérieur au support minimum $\theta$ sont extraits. Soit $\mathcal{M}$ l'ensemble des motifs extraits : $\forall M \in \mathcal{M}$, $Support(M) \geq \theta$.

**Définition 3** *(Support)*
*Une séquence $S \in \mathcal{S}$ supporte un motif M lorsque $M \preceq S$. Le support de M est le nombre de séquences de l'ensemble S dans lequel M est inclus. Soit $\mathcal{S}'$ un ensemble de séquences qui supportent M, $\mathcal{S}'$ = {$S_i \in \mathcal{S}$ tel que $M \preceq S_i$} et Support(M) = $|\mathcal{S}'|$.*

**Définition 4** *(Régle d'association)*
*Une règle d'association est une expression de la forme $X \longrightarrow Y$ où X et Y représentent deux ensembles disjoints d'itemsets, i.e., $X \cap Y = \emptyset$. La force d'une règle est mesurée par les paramètres de support et de confiance. Le support détermine le nombre de fois où la règle s'applique à l'ensemble des transactions, alors que la confiance exprime le nombre de fois où l'ensemble Y apparait dans les transactions comportant l'ensemble X.*

**Exemple 2**
*Considérons la règle {Soleil $\longrightarrow$ Humidité = Faible}. Sachant que le support de {Soleil, Humidité=Faible} est 2 et que le nombre total de transactions est 5, le support de la règle est 2/5=0.4. La confiance de la règle est obtenue en divisant le support de {Soleil, Humidité=Faible} par le support de {Soleil}. Sachant qu'il existe 3 transactions contenant Soleil, la confiance de la règle est 2/3=0.67.*

L'interprétation d'une règle séquentielle est donnée par (Fournier-Viger et al., 2012) et exprime que si un ensemble d'items X apparait dans certaines séquences, alors l'ensemble d'items de Y apparait en suivant l'ensemble X dans les mêmes séquences.

L'ensemble des définitions posées, nous présentons dans la section suivante le principe de représentation des données guidée par la RTO afin d'augmenter leur expressivité dans les textes et exploiter plus efficacement l'extraction des règles et motifs par les algorithmes de fouille de données.

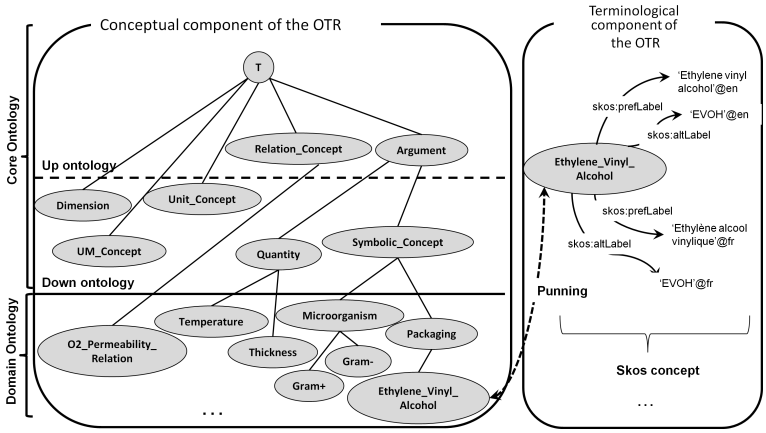


FIG. 1: Extrait de la RTO de domaine des emballages alimentaires

## 2.2 Représentation des données

**Fouiller des contextes proches des arguments.** Comme nous l'avons défini précédemment, les données d'intérêt sont modélisées en une relation n-aire où sont représentés un argument symbolique (i.e. objet étudié) et ses arguments quantitatifs (i.e. les mesures effectuées sur ou représentant l'objet étudié) associés à leurs attributs, une valeur numérique et une unité de mesure. Les relations n-aires sont représentées dans une RTO de domaine. Un extrait de la RTO de domaine (Touhami et al., 2011) sur laquelle reposent les évaluations effectuées est représenté dans la figure 1. La RTO comporte deux composantes : une composante terminologique qui regroupe toute la terminologie du domaine, e.g. les noms des emballages alimentaires, et une composante conceptuelle qui restitue les relations hiérarchiques entre les concepts de domaine et génériques selon la relation de subsomption.

Dans notre travail, nous définissons l'unité de mesure (Berrahou et al., 2013) associée aux arguments quantitatifs comme un descripteur pertinent dans le texte afin de définir des contextes favorables à la découverte des arguments recherchés de la relation n-aire. À partir de ce descripteur, nous proposons deux contextes textuels pertinents de recherche dans les textes :

**Définition 5** *(la phrase pivot)*
*La phrase pivot est définie comme la phrase où au moins une unité référencée dans la RTO est identifiée*

**Définition 6** *(La fenétre textuelle)*
*La fenêtre textuelle, notée $f_{sn}$ est définie comme l'ensemble des phrases composé de la phrase pivot et des n phrases précédentes et/ou des n phrases suivantes, où n correspond à la dimension de la fenêtre. Le sens de recherche dans les phrases, noté s, est représenté par le signe - en considérant les phrases précédentes, par le signe + en considérant les phrases suivantes et par le signe $\pm$ en s'appuyant sur les phrases précédentes et suivantes.*

**Augmenter l'expressivité des arguments.** L'objectif de notre travail est d'extraire des instances d'arguments ayant des formes d'expression variées dans les textes. Les formes d'expression varient dans les textes du fait de la richesse du vocabulaire employé mais également du fait de la structure même de l'instance de la relation. En effet, une instance de relation est caractérisée par des valeurs numériques qui changent fréquemment selon les mesures effectuées sur l'objet étudié. Ces variations ne permettent pas d'appliquer efficacement le processus d'extraction des connaissances fondé sur le caractère fréquent des données. Pour remédier à cette problématique, nous proposons une nouvelle représentation des données, guidée par la RTO de domaine afin d'augmenter l'expressivité des données d'intérêt dans les textes. La nouvelle représentation s'appuie sur la terminologie du domaine identifiée dans les textes pour représenter les concepts correspondant aux arguments de la relation n-aire. Dans cette représentation, les arguments de la relation n-aire peuvent être des arguments symboliques ou des arguments quantitatifs. L'expression des arguments quantitatifs, ayant très peu de variabilité dans les textes est représentée par le concept générique de la RTO, i.e. $< quantity >$. Le concept générique $< Symbolic\_Concept >$ représente l'argument symbolique pour lequel il existe une grande variabilité d'expression dans les textes. Nous avons choisi de représenter les sous-concepts du concept générique $< Symbolic\_Concept >$, plus précis et représentatifs des objets étudiés. Dans les expérimentations effectuées, nous travaillons sur un corpus du domaine des emballages, nous avons donc choisi le sous-concept $< packaging >$ pour représenter la notion d'emballage, qui est l'objet étudié recherché dans les textes. Enfin, les valeurs numériques, variant sensiblement dans les textes, sont représentées en considérant l'unité de mesure qui les qualifie et selon les sous-concepts quantité que ces mêmes unités dénotent pour augmenter leur expressivité dans les textes. Dans l'exemple 3, la valeur numérique *64* est suivie de l'unité *µm* qui est rattachée au concept *thickness*. La présence de l'unité permet donc de qualifier la valeur numérique pour en augmenter le sens, i.e. $< numvalthick >$ pour *thickness*, $< numvaltemp >$ pour *temperature*, $< numvalrh >$ pour *relative humidity*, $< numvalperm >$ pour *permeability*. Nous choisissons ensuite de représenter les unités de mesure simplement par le concept générique *um* pour *unit_Concept*. Le principe de représentation des données proposée est illustré dans l'exemple 3. Dans (1), l'expressivité des arguments soulignés est augmentée en utilisant les concepts de domaine et génériques de la RTO, comme cela est illustré dans (2).

**Exemple 3**

*(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 µm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 ° C and 0 % RH.*

*(2) Eight apple wedges were packaged into polypropylene $< packaging >$ trays and wrap-sealed using a 64 $< numvalthick >$ µm $< um >$ thickness $< quantity >$ polypropylene $< packaging >$ film with a permeability to oxygen $< quantity >$ of 110 $< numvalperm >$ $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ $< um >$ at 23 $< numvaltemp >$ ° C $< um >$ and 0 $< numvalrh >$ % $< um >$ RH $< quantity >$.*

Ces propositions nous permettent de définir plusieurs contextes textuels à exploiter en s'appuyant sur une nouvelle représentation des données qui en augmente sensiblement l'expressi-

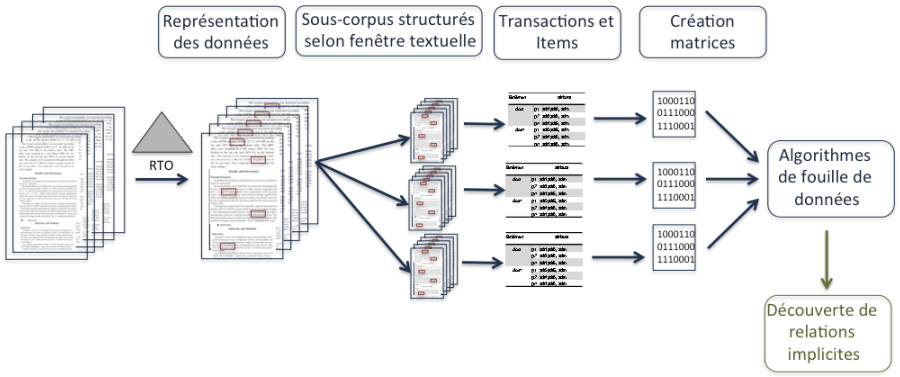


FIG. 2: Processus d'extraction de connaissances guidé par une RTO de domaine.

vité dans les textes. Dans la section suivante, nous détaillons le processus d'extraction pour la découverte de motifs fréquents d'arguments corrélés dans les textes.

## 2.3 Processus d'extraction de connaissances guidé par la RTO

Dans cette section, nous détaillons le processus d'extraction des connaissances illustré dans la figure 2 et qui repose sur trois étapes essentielles :

La première étape correspond à la représentation des données, précédemment décrite dans la section 2.2. Dans la deuxième étape, nous proposons de représenter le corpus initial en sous-corpus selon plusieurs fenêtres textuelles à explorer par les algorithmes de fouille de données. Dans ces nouveaux sous-corpus, chaque transaction correspond à un ensemble de phrases défini selon la fenêtre textuelle évaluée. Par exemple, le sous-corpus représenté selon la fenêtre $f_{\pm 1}$ propose d'explorer l'ensemble des transactions qui sont constitués à partir de l'ensemble des phrases composé de la phrase pivot et de la phrase précédente et suivante de cette même phrase pivot. Ensuite, nous sélectionnons un ensemble d'items associés. Cet ensemble est formé par les *n-terms* voisins des concepts identifiés dans notre processus de représentation des données. En reprenant l'instance représentée en (2) de l'exemple 3, si nous choisissons de sélectionner les *1-term* proche du concept <*packaging*> représenté, nous obtenons l'ensemble d'items composé de <*packaging*>, *polypropylene, trays, films*. En sélectionnant ainsi les items, notre proposition permet d'exploiter les données dans un contexte toujours proche de l'expression des arguments dans les textes. La dernière étape correspond à l'étape de fouille de données. Chaque fenêtre textuelle évaluée est transformée en matrice à explorer par les algorithmes pour la découverte des motifs fréquents d'arguments corrélés dans les textes. Au cours de cette étape, notre objectif consiste à extraire les règles et motifs fréquents les plus pertinents selon une séquence précise d'apparition dans les phrases définie par la fenêtre textuelle, e.g. l'extraction d'un motif restituant une corrélation fréquente entre le concept $< packaging >$ et son épaisseur représentée par sa valeur numérique *numvalthick* selon une fenêtre $f_{\pm 1}$ signifie que les deux arguments sont séparés dans un contexte maximal de 3 phrases.

La fin du processus de découverte permet d'extraire un ensemble de règles et motifs séquen-

tiels exprimant les relations entre les arguments dans les textes. Ces motifs restituent essentiellement les relations à un niveau conceptuel et représentent des motifs d'expression génériques. Dans la suite, nous proposons une approche hybride fondée sur l'extraction de relations syntaxiques afin d'enrichir et d'étendre les motifs séquentiels découverts avec des catégories grammaticales et des mots spécifiques pour construire des patrons linguistiques. Comme nous le décrivons dans la section suivante, les catégories grammaticales et mots pertinents de la relation syntaxique sont ceux qui restent proches de l'expression des arguments des relations n-aires recherchées.

# 3 Approche hybride fondée sur les relations syntaxiques

Dans la section précédente, nous avons montré l'intérêt des approches de fouille de données à découvrir des relations implicites entre les arguments des relations n-aires dans les textes, en fondant notre approche sur une nouvelle représentation des données. Dans cette section, nous présentons les principes essentiels de l'analyse syntaxique et justifions le choix des relations syntaxiques pour enrichir les motifs séquentiels découverts pour l'extraction d'arguments corrélés dans les textes.

## 3.1 Principe de l'analyse syntaxique

L'analyse syntaxique repose sur un ensemble de règles de syntaxe formant une grammaire formelle. La structure grammaticale restituée par l'analyse donne alors précisément la façon dont les règles de syntaxe sont combinées dans le texte et révèle ainsi les structures syntaxiques utilisées en langage naturelle. L'analyse syntaxique est utilisée essentiellement pour générer un étiquetage grammatical des phrases ou l'arbre syntaxique ou syntagmatique, souvent utilisé par les linguistes. Dans cette dernière représentation, les phrases sont décomposées selon une structure en arbre, où chaque mot est représenté par le constituant qui le définit, e.g. préposition, nom, verbe, et chaque groupe de mots est représenté par un syntagme, e.g. syntagme nominal, verbal, prépositionnel. Ces représentations structurelles peuvent être utilisées pour rechercher les règles et motifs séquentiels fondés sur les structures syntaxiques fréquentes. Nous avons choisi de ne pas utiliser l'analyse syntaxique d'emblée sur les textes avant de générer les séquences candidates par les algorithmes de fouille de données, en considérant les avantages et inconvénients des deux approches. Ces aspects sont présentés dans le tableau 3.

| | Avantages | Inconvénients |
|---|---|---|
| Fouille de données | - Exhaustivité<br>- Gère de grandes quantités de données | - Motifs et règles en grand nombre<br>- Fondée uniquement sur des critères statistiques |
| Analyse syntaxique | - Grammaire concise<br>- Relation entre mots | - Couverture insuffisante des grammaires<br>- Ambiguïtés non levées trop nombreuses<br>- Analyse sémantique des connaissances difficile<br>- Approche générique impossible<br>- Analyse limitée à la phrase |

TAB. 3: Principaux avantages et inconvénients des approches de fouille de données et de l'analyse syntaxique

En effet, ces analyses sont réduites uniquement à la phrase, ce qui limite sensiblement les perspectives d'exploration des textes et, ne permet pas de prendre en compte l'analyse sémantique des données. De plus, un des inconvénients des approches de fouille de données est l'effet exponentiel des algorithmes lors de la génération des candidats. Nous avons considéré que d'effectuer d'emblée une analyse syntaxique sur les textes aggraverait cet aspect et rendrait la phase de validation d'autant plus inconfortable.

## 3.2 Choix des relations syntaxiques

Notre approche utilise les capacités exploratoires des algorithmes sur les données pour extraire les motifs et règles fréquents fondés sur la connaissance des données (i.e. la RTO) pour découvrir les corrélations entre les arguments dans les textes. Néanmoins, comme nous l'avons présenté dans la section 2, les motifs et règles restitués permettent uniquement d'appréhender les relations que partagent les arguments dans les textes. Les relations syntaxiques présentent un avantage incontestable, celui d'extraire des relations de dépendance entre paire de mots de la phrase. Ces relations sont facilement comprises et efficacement exploitées sans besoin d'expertise linguistique. Elles proposent une analyse alternative aux classiques représentations structurelles des phrases présentées dans la section 3.1. Ces relations correspondent à une analyse plus appropriée à l'enrichissement des motifs car nous sommes en mesure, à la fois, de gérer le nombre et la qualité des structures linguistiques choisies pour enrichir les motifs. Les relations se présentent sous forme de triplets de dépendance associant un rôle grammatical à une paire de mots. Le premier mot du triplet est considéré comme le régent et le deuxième mot comme le dépendant. L'exemple 4 restitue un type de relations possibles. Les noms *oil* et *price* modifient, selon le rôle nominal, le régent qui est également un nom. L'analyseur est associé à un ensemble de règles grammaticales sur lesquelles reposent la majorité des structures linguistiques.

**Exemple 4**
*"Oil prices futures"*
*NN(futures, oil) et NN(futures, price)*

Dans nos travaux, nous choisissons de sélectionner uniquement les relations syntaxiques d'intérêt pour l'extraction des arguments corrélés. En résumé, l'approche hybride, présentée dans la section suivante, propose de combiner, d'une part, les approches de fouille de données qui permettent d'extraire exhaustivement les relations implicites que partagent les arguments dans les textes, et d'autre part, en utilisant des relations syntaxiques spécifiques pour en appréhender la structure linguistique dans les textes.

## 3.3 Principe de l'approche hybride

Dans cette section, nous cherchons donc à extraire les relations syntaxiques (RS) fréquentes et pertinentes, proches des arguments des relations n-aires recherchées. Ces RS révèlent l'utilisation de catégories grammaticales et de termes spécifiques dans l'expression des arguments dans les textes que nous identifions comme pertinents à l'enrichissement des motifs génériques découverts dans l'étape de fouille de données pour la construction de patrons linguistiques d'extraction d'arguments corrélés dans les textes.

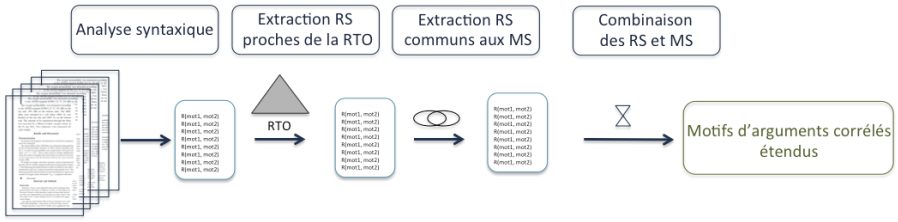


FIG. 3: Approche hybride

La figure 3 montre que l'approche hybride que nous proposons repose sur 4 étapes :
**L'extraction des RS** consiste à utiliser un analyseur syntaxique sur l'ensemble du corpus. L'analyseur analyse chaque phrase du corpus et retourne toutes les RS du corpus. Une RS est un triplet de dépendance de chaque paire de mots de la phrase. Le triplet restitue la classe grammaticale qui relie la paire de mots. Par exemple, la RS *NN(thickness, film)* montre que les mots *thickness* et *film* peuvent être associés dans les textes selon la catégorie grammaticale spécifique *Noun* (i.e. "Nom") et peuvent être représentés selon la structure linguistique *film thickness* dans les textes. Notre proposition consiste à extraire l'ensemble des RS retournant des rôles grammaticaux et des termes proches des arguments recherchés, en s'appuyant sur la RTO de domaine pour identifier les relations les plus pertinentes.
**Extraction des RS proches de la RTO.** Tous les termes référencés dans la RTO et qui dénotent les concepts du domaine sont utilisés afin d'identifier les RS les plus pertinentes, i.e. que toutes les RS contenant au moins un terme dénotant un concept de la RTO sont extraites. Ces RS sont ensuite classées en fonction de leur fréquence d'apparition dans le corpus pour conserver les structures les plus fréquentes.
**RS proches des motifs.** Ensuite, de cet ensemble de RS pertinentes, car proches de la RTO et fréquentes dans les textes, nous nous intéressons également à celles qui restituent des catégories grammaticales des termes employés pour l'expression des arguments corrélés, découverts dans le processus d'extraction de connaissances. Par exemple, prep_of(thickness, LDPE) montre que les mots *thickness* et *LDPE* sont liés selon un rôle grammatical spécifique, le groupe prépositionnel *prep_of*. Dans la relation, *LDPE* représente un terme dénotant le concept de la RTO *packaging*. Cette relation peut également être utilisée pour enrichir le motif fréquent découvert et corrélant les arguments *packaging* et *thickness*. Ce type de RS est particulièrement intéressant à exploiter afin de construire des patrons linguistiques et séquentiels d'arguments corrélés dans les textes.
**Hybridation des motifs avec les RS.** Toutes les RS restituant les arguments corrélés découverts dans les règles et motifs séquentiels sont utilisées pour enrichir les motifs séquentiels. Le principe de l'approche hybride consiste à joindre les RS avec les motifs séquentiels sur les termes ou concepts communs.
Considérons le motif <(packaging)(numvalthick um)> extrait du corpus des emballages restituant la corrélation entre l'emballage et son épaisseur. Nous cherchons à appréhender plus précisément la structure linguistique associée à ce motif séquentiel. Le motif séquentiel guide

l'extraction des RS puisque, à partir de la corrélation, nous extrayons celles qui vont restituer les arguments recherchés et la corrélation, e.g. *NN(thickness, film ou films) ou NN(film ou films, thickness), NN(LDPE, film ou films) ou NN(film ou films, HPMC), prep_of(thickness, LDPE)*. Les termes *LDPE et HPMC* dans la relation dénotent le concept *packaging* dans la RTO, on conserve le concept dans la relation. Ces relations syntaxiques sont transformées, à partir des règles grammaticales de dépendance associées à l'analyseur, comme décrit dans la section 3.2, en autant de structures linguistiques reliant les paires de mots. Les règles de génération des structures adoptées sont illustrées dans l'exemple 5

**Exemple 5**
*Génération :*
*- NN(thickness, film ou films) => film films thickness*
*- NN(film ou films, thickness) => thickness film films*
*- NN(LDPE, film ou films) => film films (packaging)*
*- NN(film ou films, HPMC) =>(packaging) film films*
*- prep_of(thickness, LDPE) => thickness of (packaging)*
*Hybridation avec motif séquentiel <(packaging)(numvalthick um)> :*
*- <(packaging) film films thickness (numvalthick um)>*
*- <thickness of (packaging) film films (numvalthick um)>*
...

Nous obtenons les patrons linguistiques construits dans la séquence définie par le motif séquentiel. Après validation, ces patrons linguistiques peuvent être appliqués sur les textes pour extraire les arguments corrélés recherchés, e.g. *mango <packaging> films thickness was 0.17 ± 0.02 <numvalthick> mm <um>* ou *Thickness of resulting starch <packaging> films ranged from 199.6 ± 22.6 to 271.4 ± 581 <numvalthick>* $\mu m < um >$.

La section suivante détaille les expérimentations menées et les résultats obtenus sur le corpus d'évaluation.

# 4 Expérimentations

Nous avons mené des expérimentations sur un corpus de 115 documents scientifiques, rédigés en texte brut et en anglais, du domaine des emballages alimentaires. La première étape de la méthode proposée consiste à extraire l'ensemble des règles et motifs séquentiels en utilisant le processus d'extraction de connaissances fondé sur la nouvelle représentation des données, dans ce cas des données expérimentales. Puis, après une étape de validation des motifs pertinents découverts, nous utilisons l'approche hybride pour construire des patrons linguistiques d'extraction d'arguments corrélés en utilisant des RS spécifiques.

## 4.1 Motifs séquentiels

**Constitution des sous-corpus.** À partir du corpus des emballages alimentaires, nous organisons plusieurs sous-corpus selon les fenêtres textuelles évaluées (e.g. le sous-corpus $f_0$, $f_{\pm 2}$). Nous appliquons le processus d'extraction des connaissances tel que nous l'avons décrit

dans la section 2.3 et nous obtenons un ensemble de matrices à explorer. Le nombre de transactions varie selon la fenêtre textuelle évaluée de 5 000 à 35 000 phrases. Le nombre d'items constitués varie également selon la fenêtre de 2 000 à plus de 10 000 items.

**Choix des algorithmes.** De nombreux algorithmes existent à l'état de l'art tel que Apriori (Agrawal et Srikant, 1994), Spade (Zaki, 2001) et PrefixSpan (Pei et al., 2001). Les expérimentations ont été menées en utilisant l'algorithme Clospan (Yan et al., 2003) pour extraire les motifs séquentiels. Clospan implémente l'algorithme PrefixSpan, le plus efficace à l'état de l'art tout en permettant de découvrir un ensemble concis de motifs sans redondance ni perte d'information. Pour la découverte des règles séquentielles, nous avons employé l'algorithme CMRules (Fournier-Viger et al., 2012).

**Critères de sélection.** La génération d'une grande quantité de motifs et règles est une problématique connue en fouille de données, rendant la tâche de validation particulièrement inconfortable. Ainsi, la mesure du support aide à réduire le nombre de motifs et règles aux plus pertinentes à conserver. La confiance a été fixée au maximum (i.e. confiance = 1) afin d'être sélectif sur les règles les plus sures.

Au-delà de ces mesures classiques, nous proposons deux critères de sélection supplémentaires fondés à la fois sur des critères statistiques et sémantiques. Le premier critère permet de sélectionner uniquement les règles et motifs comportant au moins un argument des relations n-aires recherchées et représentées dans la RTO. Le second critère permet d'extraire les règles et motifs issus de l'intersection de plusieurs fenêtres étudiées.

**Résultats quantitatifs.** Le nombre de motifs et règles extraits varie en fonction des critères de sélection appliqués. Par exemple, nous obtenons plus de 52 000 règles et motifs à partir du sous-corpus $f_{\pm 2}$ selon un minimum de support de 0.5 et selon le critère de présence d'au moins un argument des relations n-aires recherchées. Lorsqu'on ajoute le critère de sélection d'intersection des fenêtres, représenté par le symbole $\bigcap f_n$ dans le tableau, nous réduisons l'ensemble autour de 1 000 règles et motifs extraits.

**Résultats qualitatifs.** Nous avons dans un premier temps effectué une série d'évaluations en

| Fenêtre textuelle | Règles et motifs | Support |
|---|---|---|
| $f_{\pm 1}$ | <(packaging)(numvalthick um)><br><(numvalthick)(films)><br><(film)(mm)(thickness)><br><(film thickness)(rh)><br><(packaging)(quantity)(permeability)><br><(packaging)(permeability)> | 0.5<br>0.5<br>0.1<br>0.1<br>0.5<br>0.6 |
| $f_0$ | <(pressure)(water permeability)><br><(oxygen permeability)(pressure)><br><(thickness)(films)(observed)><br><(thickness)(water)(films)> | 0.05<br>0.05<br>0.05<br>0.07 |
| $\bigcap f_n$ | packaging => numvalthick<br>temperature => numvalrh<br><(numvaltemp)(numvalrh%)><br><(packaging)(numvalthick)><br><(packaging)(numvaltemp °c)><br><(packaging) => temperature numvalrh ><br>packaging => quantity numvaltemp °c | >0.05 |

TAB. 4: Extrait de règles et de motifs découverts

utilisant le processus d'extraction des connaissances sans représenter les données expérimentales pour en augmenter l'expressivité. Nous obtenons un faible ensemble de motifs (comparativement à l'autre méthode), i.e. autour de 500. Les motifs extraits sont pauvres en information et ne comportent pas d'arguments corrélés (i.e. très peu d'itemsets restitués). Aucun des motifs ne restitue de valeurs numériques alors que celles-ci ont un sens particulièrement important dans l'instance de relation recherchée.
Le tableau 4 restitue un extrait des motifs et règles séquentiels extraits avec le processus d'extraction des connaissances fondé sur la nouvelle représentation des données expérimentales que nous proposons. Nous constatons que les motifs et règles extraits sont plus expressifs et proches des instances d'arguments recherchés, ce qui montre l'intérêt de la nouvelle représentation des données expérimentales proposée, guidée par la RTO de domaine. Puis, les résultats montrent que les règles et motifs extraits nous permettent de découvrir plusieurs relations implicites d'expression des arguments dans les textes.

1. Une première règle restituée dans plusieurs motifs, e.g. $<$*(packaging)(numvalthick um)*$>$, montre que l'argument *packaging* et l'argument quantitatif *thickness* apparaissent fréquemment ensemble dans le texte et que cette corrélation se manifeste dans une fenêtre textuelle maximale de $f_{\pm 1}$ ;
2. Une seconde règle, e.g. *temperature* $=>$ *numvalrh*, montre que si l'argument quantitatif *temperature* est présent, nous trouvons fréquemment l'argument quantitatif *relative humidity* qui lui est associé. Ces deux arguments sont fréquemment corrélés dans la même phrase, i.e. $f_0$ ;
3. Une troisième règle, e.g. $<$*(packaging)* $=>$ *temperature numvalrh* $>$, montre que l'argument *packaging* se retrouve fréquemment dans le même contexte que les arguments quantitatifs *temperature* et *relative humidity*. Ainsi, 4 arguments des relations n-aires recherchées sont corrélés dans le texte, i.e. *packaging, thickness, temperature* et *relative humidity*, et cette corrélation se fait fréquemment dans un contexte proche, i.e. dans une fenêtre maximale de $f_{\pm 1}$ ;
4. Une quatrième règle, e.g. $<$(pressure)(water permeability)$>$, montre que les arguments quantitatifs *partial pressure* et *permeability* sont souvent associés dans la même phrase ;
5. Une dernière règle est particulièrement intéressante car elle suggère que l'argument *packaging* pourrait jouer le rôle de déclencheur de l'instance de relation car il apparait fréquemment dans de nombreux motifs et règles restituant les corrélations précédentes, e.g. $<$*(packaging)(permeability)*$>$, $<$*(packaging)(numvaltemp* $^\circ c$*)*$>$. Le terme dénotant le concept *packaging* peut donc être utilisé pour regrouper les arguments dans l'instance de relation recherchée.

## 4.2 Construction de patrons linguistiques à partir des motifs séquentiels

**Résultats quantitatifs et qualitatifs d'extraction des RS.** Pour extraire les relations syntaxiques de notre corpus de 115 documents, nous avons utilisé l'analyseur syntaxique en anglais de Stanford (Klein et Manning, 2003), le plus efficace à l'état de l'art avec un taux de précision de 86%. L'analyseur de Stanford est associé à une cinquantaine de règles grammaticales définissant les relations syntaxiques communément utilisées dans la langue anglaise (de Marneffe et al., 2006). Après analyse, nous obtenons un ensemble constitué de plus de 50

000 RS. Ensuite, nous avons réduit cet ensemble en utilisant les termes référencés dans la RTO. Nous constituons ainsi, un sous-ensemble de RS d'intérêt proches des arguments de la relation n-aire recherchée. Finalement, de ce nouvel ensemble, nous conservons uniquement les RS qui restituent les corrélations d'arguments dans les textes, découvertes dans les motifs séquentiels. Les catégories grammaticales et termes restitués sont ajoutés aux motifs séquentiels sélectionnés afin de construire des patrons linguistiques. Le nombre de RS finalement conservées et d'intérêt pour l'enrichissement des motifs est de 6 600 RS. En termes de résultats qualitatifs, les RS les plus pertinentes sont celles restituant les catégories nominales, les prépositions et les conjonctions. D'autres RS montrent également que la relation entre la valeur numérique et son unité de mesure est toujours de type numérique. D'autres encore restituent des verbes particulièrement intéressants car ils appartiennent au groupe des verbes de type expérimental, soit décrivant un contexte de description expérimentale ou d'analyse expérimentale tels que les verbes *conduct, prepare, measure, pack, compare, increase, decrease*.

**Résultats quantitatifs d'extraction d'arguments corrélés par approche hybride.** Les expérimentations ont été menées en deux temps sur un échantillon de 11 articles scientifiques extraits de notre corpus et qui restituent 87 instances d'arguments recherchés dans 3 types de relations n-aires différentes, i.e. la relation de perméabilité à l'oxygène, la relation de perméabilité au dioxyde de carbone et enfin la relation de perméabilité à l'eau. Dans un premier temps, nous avons testé les motifs séquentiels génériques sans hybridation avec les relations syntaxiques sélectionnées, puis, dans un deuxième temps, en utilisant les motifs construits par approche hybride. Les résultats sont restitués dans le tableau 5. Les patrons ont été appliqués directement sur les textes du sous-corpus $f_{\pm 1}$, où 4 arguments corrélés recherchés sont découverts, i.e. *packaging, thickness, temperature et relative humidity (RH)*. Dans cette première phase d'évaluation, nous avons testé 6 motifs séquentiels fréquents que nous avons enrichis avec 4 catégories grammaticales *nom, numérique, préposition et conjonction* et termes restitués par les RS les plus pertinentes et d'intérêt pour la construction des patrons linguistiques. Nous obtenons finalement un ensemble de 50 patrons linguistiques et motifs séquentiels à appliquer sur les textes pour l'extraction d'arguments de relations n-aires corrélés dans les textes. Les résultats sont donnés dans le tableau 5 selon les mesures de précision, qui évalue la qualité de la méthode, de rappel, qui mesure l'exhaustivité des résultats extraits, et de F-mesure, qui reflète un compromis entre qualité et exhaustivité des résultats extraits. Les résultats sont restitués selon les différentes corrélations découvertes. Nous avons testé les motifs et patrons restituant l'argument *packaging* étudié et l'argument quantitatif *thickness* qui lui est associé dans les textes. Puis, les motifs et patrons restituant l'argument *packaging* associé à l'argument *temperature* et l'argument *packaging* associé à l'argument *relative humidity*. Enfin, nous avons testé les motifs restituant 3 ou 4 arguments corrélés dans le texte. Les résultats montrent que les patrons linguistiques obtenus par approche hybride augmentent sensiblement la précision des extractions de 0.4 à 0.7 pour les corrélations de *packaging* et *thickness* et de 0.3 à 0.7 pour les arguments *packaging et temperature* et *packaging et relative humidity* sans avoir d'impact négatif sur le rappel. Les résultats concernant l'extraction de 3 à 4 arguments corrélés dans le texte montrent une légère amélioration de la précision avec les patrons linguistiques mais montrent également la force des motifs à découvrir de manière précise les corrélations dans les textes avec une F-mesure de 0.6.

| Type d'évaluation | **Motifs séquentiels** | | | **Patrons linguistiques et séquentiels** | | |
|---|---|---|---|---|---|---|
| | *Précision* | *Rappel* | *F-Mesure* | *Précision* | *Rappel* | *F-Mesure* |
| Évaluation générale | 0.5 | 0.8 | 0.6 | **0.7** | 0.8 | 0.6 |
| *packaging* and *thickness* | 0.4 | 0.9 | 0.5 | **0.7** | 0.9 | 0.8 |
| *temperature* and *relative humidity* | 0.3 | 0.9 | 0.4 | **0.8** | 0.7 | 0.7 |
| n $>$ 2 arguments corrélés | 0.6 | 0.6 | 0.6 | **0.7** | 0.6 | 0.6 |

TAB. 5: Évaluation de l'extraction des arguments corrélés dans les textes

# 5 Conclusion

Dans cet article, nous proposons une approche hybride combinant les techniques de fouille de données et les relations syntaxiques pour l'extraction de données complexes dans les textes. Les données complexes sont des instances de relations n-aires qui associent un objet étudié (e.g. l'emballage) à ses caractéristiques, i.e. les mesures effectuées sur ou associées à cet objet étudié selon différents arguments quantitatifs. L'expression de ces instances varie fréquemment dans les textes du fait de la richesse du vocabulaire employé pour les décrire, mais également du fait de la structure même des instances d'arguments quantitatifs qui varient par leurs attributs, i.e. la valeur numérique et l'unité de mesure, selon les mesures effectuées sur l'objet étudié. La méthode proposée pour extraire ces instances repose sur un processus d'extraction des connaissances fondé sur une nouvelle représentation des données, guidée par une RTO de domaine. Cette nouvelle représentation permet d'augmenter l'expressivité des arguments des relations n-aires recherchées en s'appuyant sur le niveau conceptuel exprimé dans la RTO. Dans ce processus, nous proposons également d'évaluer plusieurs fenêtres textuelles pour la découverte des motifs et règles pertinents concernant les arguments recherchés dans une séquence de phrases définie par la fenêtre. La première étape permet de découvrir les règles et motifs d'expression d'arguments corrélés dans les textes. Puis, dans une seconde étape, nous proposons d'extraire les relations syntaxiques pertinentes à l'enrichissement de ces motifs découverts pour la construction de patrons linguistiques et séquentiels qui combinent plusieurs niveaux d'abstraction (mot, catégorie grammaticale et concept) pour une extraction plus efficace des arguments dans les textes. Au cours des expérimentations menées, les patrons ont permis l'extraction d'instances restituant de 2 à 4 arguments de relations n-aires corrélés dans les textes.

Dans une prochaine évaluation, l'approche hybride doit être appliquée sur un nouveau corpus (corpus de bioraffinerie) dans lequel les données, modélisées en relations n-aires et représentées dans une RTO, doivent être extraites. De la même manière, le processus d'extraction de connaissances reposera sur la nouvelle représentation des données proposée, en utilisant les concepts génériques et de domaine, et en utilisant les unités de mesure du domaine pour représenter les valeurs numériques de l'instance recherchée, puis sur la construction des patrons linguistiques en utilisant les relations syntaxiques pertinentes et d'intérêt.

Une autre perspective concerne l'intégration des patrons linguistiques et séquentiels d'extraction d'arguments corrélés dans une plateforme d'annotation existante, @web [1] (Buche et al., 2013), de tableaux extraits des articles . Ces tableaux restituent des instances de relations recherchées mais fréquemment incomplètes, avec des instances d'arguments manquants dans le tableau et présents dans le texte. Les patrons permettraient alors d'identifier les phrases dans

1. http ://www6.inra.fr/cati-icat-atweb/Web-platform

lesquelles les arguments manquants sont identifiés et aideraient ainsi l'annotateur à compléter l'annotation de l'instance dans le tableau.

**Remerciements :** Le travail de recherche ayant mené aux résultats présentés dans cet article a reçu le soutien du labex NUMEV et du projet 3BCAR IC2ACV.

# Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, San Francisco, CA, USA, pp. 487–499. Morgan Kaufmann Publishers Inc.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In Proceedings of the Eleventh International Conference on Data Engineering, ICDE '95, Washington, DC, USA, pp. 3–14. IEEE Computer Society.

Béchet, N., P. Cellier, T. Charnois, et B. Crémilleux (2012). Discovering linguistic patterns using sequence mining. In Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I, CICLing'12, Berlin, Heidelberg, pp. 154–165. Springer-Verlag.

Berrahou, S. L., P. Buche, J. Dibie-Barthélemy, et M. Roche (2013). How to extract unit of measure in scientific documents ? In KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013, pp. 249–256.

Björne, J., J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, et T. Salakoski (2009). Extracting complex biological events with rich graph-based feature sets. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task, BioNLP '09, Stroudsburg, PA, USA, pp. 10–18. Association for Computational Linguistics.

Buche, P., S. Dervaux, J. Dibie-Barthélemy, L. Soler, L. Ibanescu, et R. Touhami (2013). Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. Revue d'Intelligence Artificielle 27(4-5), 539–568.

Bui, Q.-C. et P. M. A. Sloot (2011). Extracting biological events from text using simple syntactic patterns. In Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11, Stroudsburg, PA, USA, pp. 143–146. Association for Computational Linguistics.

Buyko, E., E. Faessler, J. Wermter, et U. Hahn (2009). Event extraction from trimmed dependency graphs. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task, BioNLP '09, Stroudsburg, PA, USA, pp. 19–27. Association for Computational Linguistics.

Cellier, P., T. Charnois, M. Plantevit, C. Rigotti, B. Crémilleux, O. Gandrillon, J. Kléma, et J. Manguin (2015). Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. J. Biomedical Semantics 6, 27.

de Marneffe, M.-C., B. MacCartney, et C. D. Manning (2006). Generating typed dependency parses from phrase structure parses. In IN PROC. INT'L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC, pp. 449–454.

Di-Jorio, L., S. Bringay, C. Fiot, A. Laurent, et M. Teisseire (2008). Sequential patterns for maintaining ontologies over time. In On the Move to Meaningful Internet Systems : OTM 2008, OTM 2008

Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II, pp. 1385–1403.

Fabrègue, M., A. Braud, S. Bringay, F. L. Ber, et M. Teisseire (2012). Including spatial relations and scales within sequential pattern extraction. In Discovery Science - 15th International Conference, DS 2012, Lyon, France, October 29-31, 2012. Proceedings, pp. 209–223.

Fournier-Viger, P., U. Faghihi, R. Nkambou, et E. M. Nguifo (2012). Cmrules : Mining sequential rules common to several sequences. Knowl.-Based Syst., 63–76.

Guillard, V., P. Buche, S. Destercke, N. Tamani, M. Croitoru, L. Menut, C. Guillaume, et N. Gontard (2015). A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. Computers and Electronics in Agriculture (111), 131–139.

Hao, Y., X. Zhu, M. Huang, et M. Li (2005). Discovering patterns to extract protein ?protein interactions from the literature : part ii. Bioinformatics 21, 3294 ?3300.

Hawizy, L., D. Jessop, N. Adams, et P. Murray-Rust (2011). ChemicalTagger : a tool for semantic text-mining in chemistry. Journal of cheminformatics 3(1), 17.

Huang, M., X. Zhu, D. G. Payan, K. Qu, et M. Li (2004). Discovering patterns to extract protein-protein interactions from full biomedical texts. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, JNLPBA '04, Stroudsburg, PA, USA, pp. 22–28. Association for Computational Linguistics.

Jaillet, S., A. Laurent, et M. Teisseire (2006). Sequential patterns for text categorization. Intell. Data Anal. 10(3), 199–214.

Klein, D. et C. D. Manning (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, Stroudsburg, PA, USA, pp. 423–430. Association for Computational Linguistics.

Le Minh, Q., S. N. Truong, et Q. H. Bao (2011). A pattern approach for biomedical event annotation. In Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11, Stroudsburg, PA, USA, pp. 149–150. Association for Computational Linguistics.

Mcdonald, R., F. Pereira, S. Kulick, S. Winters, Y. Jin, et P. White (2005). Simple algorithms for complex relation extraction with applications to biomedical ie. In In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05, pp. 491–498.

Minard, A.-L., A.-L. Ligozat, et B. Grau (2011). Multi-class svm for relation extraction from clinical reports. In G. Angelova, K. Bontcheva, R. Mitkov, et N. Nicolov (Eds.), RANLP, pp. 604–609. RANLP 2011 Organising Committee.

Miwa, M., R. Sætre, Y. Miyao, et J. Tsujii (2009). A rich feature vector for protein-protein interaction extraction from multiple corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1 - Volume 1, EMNLP '09, Stroudsburg, PA, USA, pp. 121–130. Association for Computational Linguistics.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefixspan : Mining sequential patterns by prefix-projected growth. In Proceedings of the 17th International Conference on Data Engineering, Washington, DC, USA, pp. 215–224. IEEE Computer Society.

Proux, D., F. Rechenmann, et L. Julliard (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. In P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, et H. Weissig (Eds.), ISMB, pp. 279–285. AAAI.

Raja, K., S. Subramani, et J. Natarajan (2013). Ppinterfinder - a mining tool for extracting causal relations on human proteins from literature. Database 2013.

Rosario, B. et M. A. Hearst (2004). Classifying semantic relations in bioscience texts. In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rosario, B. et M. A. Hearst (2005). Multi-way relation classification : Application to protein-protein interactions. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Stroudsburg, PA, USA, pp. 732–739. Association for Computational Linguistics.

Touhami, R., P. Buche, J. Dibie-Barthélemy, et L. Ibanescu (2011). An ontological and terminological resource for n-ary relation annotation in web data tables. On the Move to Meaningful Internet Systems : OTM 2011, 662–679.

Van Landeghem, S., Y. Saeys, B. De Baets, et Y. Van de Peer (2009). Analyzing text in search of biomolecular events : A high-precision machine learning framework. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task, BioNLP '09, Stroudsburg, PA, USA, pp. 128–136. Association for Computational Linguistics.

Yan, X., J. Han, et R. Afshar (2003). Clospan : Mining closed sequential patterns in large databases. In D. Barbará et C. Kamath (Eds.), SDM. SIAM.

Zaki, M. J. (2001). Spade : An efficient algorithm for mining frequent sequences. Mach. Learn. 42(1-2), 31–60.

Zhang, H., M. Huang, et X. Zhu (2011). Protein-protein interaction extraction from bio-literature with compact features and data sampling strategy. In 4th International Conference on Biomedical Engineering and Informatics, BMEI 2011, Shanghai, China, October 15-17, 2011, pp. 1767–1771.

Zhou, D., D. Zhong, et Y. He (2014). Biomedical relation extraction : From binary to complex. Comp. Math. Methods in Medicine 2014.

# Summary

In this paper, we present a hybrid method based on datamining approaches and syntactic relations to automatically discover and extract relevant data found in plain text. We use a domain Ontological and Terminological Resource (OTR) which represents relevant data modelled as n-ary relations. N-ary relation links a studied object (e.g. a packaging) with its features as several arguments (e.g. its thickness). Our work focuses on extracting those arguments in texts in order to populate the OTR with new instances. The method relies on discovering implicit rules concerning the expression of arguments in texts using sequential pattern mining and sequential rules, and on integrating specific syntactic relations in the discovered sequential patterns to construct linguistic sequential patterns of correlated arguments in texts. We have made concluding experiments on a corpus from food packaging domain where relevant data to be extracted are experimental results on packagings.