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 - UMR 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 informations dans les textes. Ces informations 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'identification 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 documents 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. 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 terme revêt souvent plusieurs significations, une même information peut s'exprimer de multiples manières, souvent implicitement, générant des ambiguïtés difficiles à cerner automatiquement par les machines.
Les travaux présentés dans cet article s'inscrivent dans la problématique d'identification et

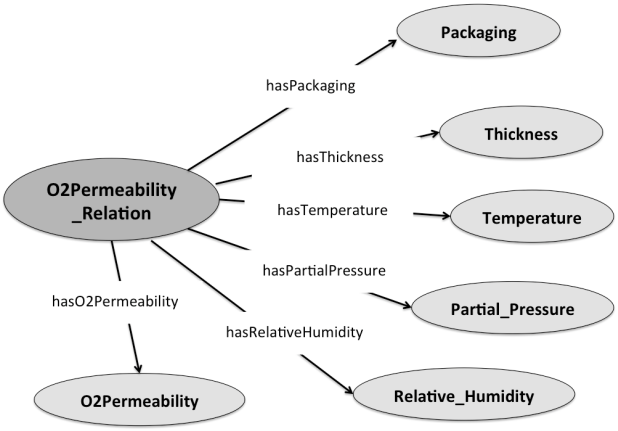

FIG. 1: La signature de la relation O2Permeability_Relation

d'extraction de données complexes dans les documents textuels. Notre méthode a pour objectif de proposer des outils intégrés dans un système d'annotation de données complexes d'intérêt. In fine, les données annotées et formatées sont intégrées dans des systèmes d'aide à la décision. Les données sont modélisées sous forme de relations n-aires représentées dans une Ressource Termino-Ontologique (RTO) naRyQ (n-ary Relations between Quantitative data) (Touhami et al., 2011) présentée dans la section 2. Un exemple de relation n-aire est donné dans la Figure 1 et représente la relation O2permeability_relation dans la RTO naRyQ_pack du domaine des emballages. Ce concept relation permet de représenter la perméabilité à l'oxygène d'un emballage dans des conditions expérimentales données par son épaisseur, son humidité relative, la pression partielle à l'oxygène et la température ambiante. Dans l'Exemple 1, la phrase (1) restitue un extrait d'une instance de la relation O2permeablity_Relation, dont les instances d'arguments à identifier et à annoter dans le texte sont soulignés. La phrase (2) restitue un autre extrait d'instance de relation n-aire dans le domaine de l'aviation.

Exemple 1

(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 $^\circ$ 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.

L'identification des instances de relations n-aires, c'est-à-dire des relations faisant intervenir plus de deux arguments (ou entités), se révèle être une tâche complexe à automatiser car l'expression de ses arguments est le plus souvent dispersée sur plusieurs phrases de l'article. Les données sont fréquemment exprimées de manière implicite et variante du fait de la richesse du vocabulaire employé pour les décrire. Elle diffère 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 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 l'instance de relation n-aire recherchée. Pour cela, nous proposons une méthode en deux étapes guidée par la RTO ; la première étape s'appuie sur les approches de fouille de données pour la découverte des règles d'expression des arguments dans le texte, à partir de l'extraction des motifs séquentiels ; la deuxième étape utilise l'analyse syntaxique pour enrichir les motifs séquentiels extraits et les utiliser pour identifier les arguments dans le texte. La RTO joue un rôle central dans la méthode proposée puisqu'elle en guide les étapes.
L'article est structuré comme suit. La section 2 rappelle quelques définitions associées à la RTO (Touhami et al., 2011) et utiles pour l'article et présente les étapes de la méthode proposée. La section 3 dresse l'état de l'art des recherches dans le domaine de l'extraction des données modélisées en relations binaires et n-aires. La section 4 décrit en détails la première étape de la méthode, fondée sur les approches de fouille de données. La section 5 décrit en détails la deuxième étape de la méthode, et en particulier l'approche hybride, fondée sur l'analyse syntaxique. La section 6 restitue et analyse les résultats obtenus au cours des expérimentations. La section 7 conclut sur les travaux menés et les perspectives envisagées.

2 Contexte et définitions préliminaires

2.1 Définitions associées à la RTO

Dans l'introduction, nous avons présenté les données d'intérêt, modélisées en relations n-aires dont les instances sont à identifier dans le texte. Ces relations n-aires sont représentées dans la RTO naRyQ. Un extrait de la RTO de domaine sur laquelle reposent les évaluations effectuées est représenté dans la Figure 2. La RTO, suivant la Définition 1, comporte deux composantes, une composante terminologique qui regroupe tous les termes du domaine, e.g. le nom des emballages, et une composante conceptuelle.
La composante conceptuelle de naRyQ est composée de l'ontologie noyau ou *core ontology* sur la figure, correspondant à la représentation générique des relations n-aires quelque soit le domaine d'application considéré, et de l'ontologie de domaine ou *domain ontology* sur la figure, correspondant à la représentation spécifique des concepts du domaine d'application donné. Dans l'ontologie noyau, plus précisément l'ontologie noyau supérieure ou up core ontology dans la figure, les concepts génériques *Relation_Concept* et *Argument* représentent respectivement les relations n-aires et leurs arguments. l'ontologie noyau inférieure ou down core ontology dans la figure, les concepts génériques *Dimension*, *UM_Concept*, *Unit_Concept* et *Quantity* permettent de gérer les quantités et leurs unités de mesure associées. Dans cette représentation, les unités de mesure sont représentées par des instances du concept générique *UM_Concept*. Les sous-concepts du concept générique *Symbolic_Concept* représentent les arguments non numériques des relations n-aires.
L'ontologie de domaine contient les concepts spécifiques du domaine d'application. Ils correspondent aux sous-concepts des concepts génériques de l'ontologie noyau.
Les définitions suivantes se rapportent à la représentation des relations n-aires.

Définition 1
La Ressource Termino-Ontologique RTO (ou OTR pour Ontological and Terminological Resource) est définie par le 6-uplet suivant :

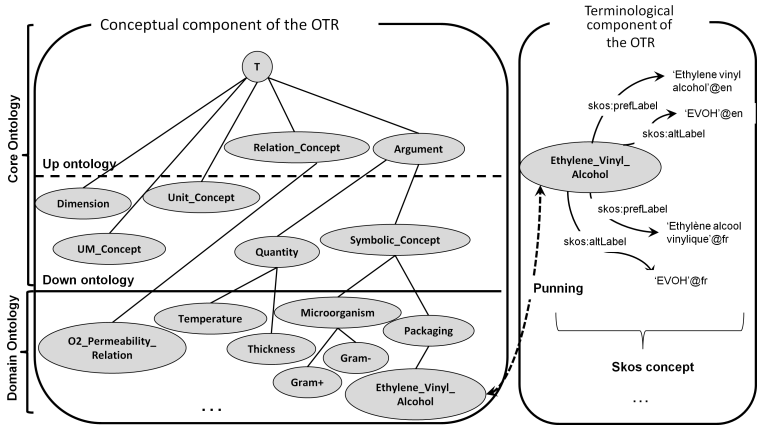

FIG. 2: Extrait de la hiérarchie de concepts de naRyQ du domaine des emballages

$OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *où*
C_{OTR} *est un ensemble de concepts de l'ontologie RTO,*
$C_{OTR} = C_{Rel} \cup C_{Qty} \cup CSymb$ *avec* C_{Rel} *l'ensemble des relations n-aires,* C_{Qty} *l'ensemble des quantités,* C_{Symb} *l'ensemble des concepts symboliques ;*
R *est l'ensemble des relations définies dans* $C_{OTR} \times C_{OTR}$ *;*
I *est l'ensemble d'instances avec* $I_{UM} \subset I$*, le sous-ensemble des instances représentant les unités de mesure ;*
V *est un ensemble de valeurs ;*
$\leq_o$ *est une relation de spécialisation définie dans* $(C_{OTR} \times C_{OTR}) \cup (R \times R)$ *;*
W_{oi} *est un ensemble de termes de la partie terminologique de la RTO où tous les termes* $w_i \in W_{oi}$ *dénotent soit un concept* $C_i \in C_{OTR}$ *soit une unité de mesure* $u \in I_{UM}$*.*

Une relation n-aire est alors représentée par un concept associé à ses arguments selon plusieurs relations binaires où l'un de ces arguments joue un rôle spécifique (e.g. le sujet ou l'objet). La Définition 2 formalise la représentation des relations n-aires entre des données quantitatives.

Définition 2
Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1,*
un **concept relation** $rel \in C_{OTR}$*,*
$\leq_o$*(rel, Relation_Concept) est défini dans* OTR *par l'ensemble de relations binaires* $r_j \in R$ *qui associent la relation n-aire à ses arguments, cet ensemble étant composé d'au moins deux relations binaires :*

$$\textit{Def}(rel) = \{r_j(rel, a_j) \mid r_j \in R,$$
$$(a_j \in C_{OTR} \wedge \leq_o (a_j, Argument))\},\ \textit{tel que}\ |\textit{Def}(rel)| \geq 2$$

Une relation n-aire est caractérisée par sa signature, i.e. l'ensemble de ses arguments.

Définition 3
Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1,*

la **signature** *signatureR :* $C_{OTR} \longrightarrow 2^{C_{OTR}}$ **d'un concept relation** $rel \in C_{OTR}$ *comme défini dans la Définition 1 est :*

$$\textit{signatureR}(rel) = \{(a_j \in C_{OTR} \wedge \leq_o (a_j, Argument)) \mid r_j(rel, a_j) \in Def(rel)\}$$

La signature de la relation n-aire O2Permeability_relation de la Figure 1 est définie par : signatureR(O2Permeability) = (Packaging, Thickness, Temperature, Partial_Pressure, Relative_Humidity, O2Permeability).
Nos propositions présentées dans cet article s'appuient sur ces définitions. Elles guident les étapes de notre méthode, dont nous présentons les grandes lignes dans la section suivante.

2.2 Présentation de la méthode

L'objectif de la méthode que nous proposons est de construire des patrons d'identification des arguments de la relation n-aire recherchée dans le texte, afin d'aider les experts de domaine à annoter des tableaux extraits d'articles scientifiques dans une plateforme d'annotation, la plateforme @Web [1] (Buche et al., 2013). Ces tableaux restituent des instances de relations recherchées mais qui sont fréquemment incomplètes. Les instances d'arguments manquantes dans les tableaux sont généralement présentes dans le texte. Les patrons permettraient alors de restituer les phrases dans lesquelles les arguments manquants sont identifiés et aideraient ainsi l'annotateur à compléter l'annotation de l'instance dans le tableau.
La méthode que nous proposons repose sur une approche hybride qui combine les approches

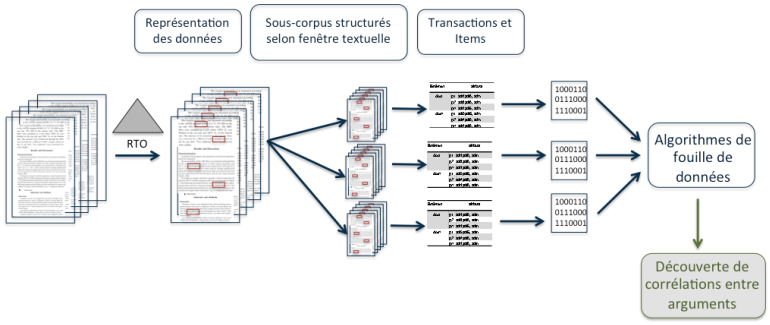

FIG. 3: Processus d'extraction de connaissances guidé par une RTO de domaine (KDP).

de fouille de données et l'analyse syntaxique pour construire les patrons d'intérêt. La première étape de la méthode proposée, détaillée dans la section 4 et représentée dans la Figure 3, propose de s'appuyer sur les approches de fouilles de données, et plus particulièrement l'extraction des motifs séquentiels, pour la découverte de relations implicites d'expression, ou règles de corrélations, entre les arguments dans les textes (e.g. découverte de co-occurrences). La découverte de ces motifs repose sur une nouvelle représentation des données pour en augmenter l'expressivité en s'appuyant sur le niveau conceptuel de la RTO. Ces motifs sont définis dans

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

la suite de l'article comme des OSP (Ontological Sequential Patterns) et permettent de faciliter l'identification et la mise en relation des arguments dans l'instance de relation recherchée.
Dans la deuxième étape de la méthode, détaillée dans la section 5 et représentée dans la Fi-

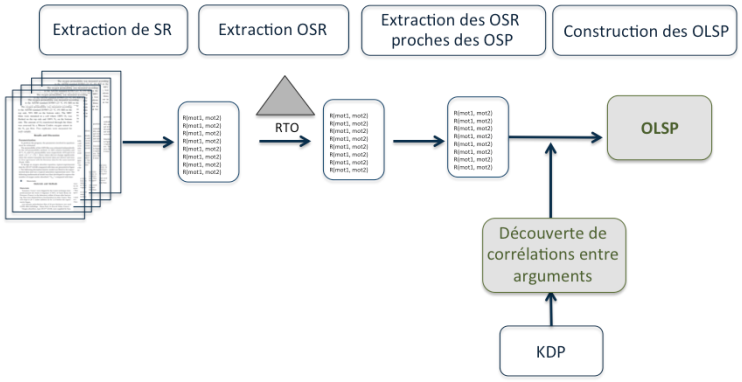

FIG. 4: Approche hybride

gure 4, nous utilisons les motifs extraits (OSP), restituant les règles de corrélations découvertes entre arguments et, nous proposons l'approche hybride pour construire les patrons d'identification des arguments dans les textes. Pour la construction des patrons, définis dans la suite comme des OLSP (Ontological Linguistic Sequential Patterns), l'approche hybride s'appuie également sur la RTO et utilise les structures linguistiques fréquentes, définis dans la suite comme des OSR (Ontological Syntactic Relations) associant les arguments dans les textes pour enrichir les motifs. Notre méthode contribue à la problématique de l'extraction des relations binaires et n-aires, dont nous dressons l'état de l'art dans la section suivante.

3 État de l'art

Dans cette section, nous présentons et discutons les travaux menés dans le domaine d'extraction de données textuelles modélisées sous forme de relations binaires et de relations n-aires. Nous nous sommes prioritairement rapprochés des travaux traitant l'extraction de données quantitatives notamment dans le domaine biomédical.

Extraction des relations binaires. Les approches proposées sont essentiellement fondées sur l'identification d'entités co-occurrentes, soit en appliquant des patrons lexico-syntaxiques d'extraction construits manuellement (Huang et al., 2004; Hawizy et al., 2011; Hao et al., 2005; Raja et al., 2013), soit en utilisant les méthodes d'apprentissage supervisé (Minard et al., 2011; Rosario et Hearst, 2005; Zhang et al., 2011; Van Landeghem et al., 2009), déterminant par classification si les entités identifiées sont en relation. Les attributs utilisés comportent par exemple, des informations lexico-syntaxiques ou sémantiques de la phrase dans laquelle les entités sont identifiées. Ces méthodes ne peuvent pas résoudre notre problématique d'identification et extraction d'arguments de relations n-aires car elles ne s'appliquent que dans des contextes restreints, limités à une seule phrase. En revanche, notre méthode est adaptée à l'identification et

l'extraction des relations binaires en s'affranchissant des constructions manuelles de patrons, ou encore de la constitution fastidieuse de corpus annotés. La méthode hybride permet à la fois de découvrir les règles implicites d'expression des entités fondées sur la connaissance du domaine et, d'intégrer des relations syntaxiques d'intérêt pour la construction de patrons d'extraction d'arguments corrélés (de 2 à plusieurs) dans les textes.

Extraction des relations n-aires. Différentes méthodes sont proposées mais pour lesquelles le processus d'identification et extraction se décompose généralement en trois étapes : (i) l'identification des entités de la relation en utilisant des ressources externes, puis (ii) la détection de l'élément déclencheur en utilisant des méthodes à base de dictionnaires, de règles définies dans des patrons à partir des arbres de constituants (Le Minh et al., 2011) ou encore fondées sur des heuristiques (Minard et al., 2010), 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 de ce fait, en perdant sensiblement en précision. Le processus reste particulièrement complexe à mettre en œuvre lorsque les entités s'expriment de manière implicite, avec un choix de l'élément déclencheur de la relation qui n'est pas trivial et, un regroupement des entités autour de l'élément déclencheur qui s'avère souvent impossible lorsque les arguments sont très éloignés dans le document. Notre approche propose de s'affranchir de ces étapes et s'appuyer sur le processus d'extraction de connaissances, guidé par l'ontologie de domaine pour découvrir les règles implicites d'expression des arguments associés dans la relation, y compris l'élément déclencheur, en extrayant les motifs séquentiels.

Extraction de données textuelles par la fouille de données. Les techniques de fouille de données ont déjà été proposées avec succès pour traiter les données textuelles, par exemple pour découvrir des relations sémantiques entre entités, ou pour combiner des informations syntaxiques et sémantiques dans le but d'enrichir et maintenir des ontologies (Di-Jorio et al., 2008; Bloehdorn et al., 2005), ou encore pour catégoriser des textes en proposant des règles compréhensibles et réutilisables par l'utilisateur (Jaillet et al., 2006). Les objectifs de ces travaux restent néanmoins assez éloignés de notre problématique d'extraction des arguments de relations n-aires. En revanche, nous nous rapprochons des travaux de (Béchet et al., 2012) sur la découverte de patrons linguistiques en utilisant les motifs séquentiels. Néanmoins, ces travaux se penchent sur la problématique d'extraction de phrases exprimant un jugement ou sentiment et ne s'adaptent pas à notre problématique de relations n-aires. (Cellier et al., 2015) proposent d'utiliser les motifs séquentiels pour l'extraction de relations entre entités nommées (gènes), donc de relations binaires exprimées dans le contexte restreint de la phrase. Nos travaux proposent également de s'appuyer sur la fouille de données, en étant guidé par la connaissance du domaine, pour définir des patrons fondés sur les règles implicites d'expression découvertes, entre 2 à plusieurs arguments de la relation recherchée et, sur plusieurs phrases du document.

4 Découverte d'arguments corrélés

Dans cette section, nous présentons en détails la première étape de la méthode proposée, représentée dans la Figure 3. La première étape, fondée sur un processus d'extraction de connaissances guidé par la RTO, repose sur quatre sous-étapes.

4.1 Première sous-étape : une nouvelle représentation des données

Nous proposons une nouvelle représentation des données pour augmenter l'expressivité des données d'intérêt dans le texte en s'appuyant sur la RTO de domaine. En effet, l'objectif de notre travail est d'extraire des instances d'arguments ayant des formes d'expression variées dans les textes. Elles 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.
La nouvelle représentation repose sur l'ontologie noyau où les arguments symboliques et quantitatifs (i.e. quantity) sont distincts. Nous proposons la Définition 4 pour augmenter l'expressivité des arguments symboliques en les représentant par leurs concepts correspondants, sous-concepts du concept générique $< Symbolic\_Concept >$. Par exemple, dans nos expérimentations conduites sur le corpus du domaine des emballages, nous choisissons le sous-concept $< Packaging >$ représenté dans la RTO par la relation de spécialisation (Packaging, Symbolic_Concept) pour représenter l'objet étudié dans le texte (i.e. l'emballage).

Définition 4 *(Représentation du concept symbolique)*
Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1 ;*
$\forall$ *t, un terme du texte, tel que* $\exists\ w_i \in W_{oi}$ *qui dénote* $C_i \in C_{Symb}$*, et sim(*w_i*, t)=1 où sim est une mesure de similarité, et* $\exists\ C_j$ *tel que* $C_i \leq_o C_j$*, et* $C_j \in signatureR(rel)$ *avec* $rel \in C_{Rel}$*, alors t est annoté par* $C_j \in C_{Symb}$ *dans la nouvelle représentation des données.*

Nous proposons également la Définition 5 pour augmenter l'expressivité des données quantitatives fondée sur l'expressivité des valeurs numériques en utilisant les unités de mesure qui leur sont associées. En effet, la valeur numérique représente l'information pertinente recherchée que nous souhaitons découvrir et qui varie fréquemment dans les textes, en fonction des différentes mesures appliquées sur l'objet étudié. Les valeurs numériques sont associées à leurs unités de mesure qui, dans la RTO sont elles-mêmes associées à des sous-concepts spécifiques du concept générique $< Quantity >$ selon la relation $hasUnit(hasUnit \in R)$. Par exemple, l'unité de mesure $^\circ\ C$ est associée dans la RTO naRyQ au sous-concept quantitatif $< Temperature >$. Nous utilisons ces sous-concepts de $< Quantity >$ pour représenter les valeurs numériques. Nous proposons également de représenter et généraliser les arguments quantitatifs avec le concept générique $< Quantity >$ et l'unité avec $< um >$ pour $< Unit\_Concept >$.

Définition 5 *(Représentation des concepts quantitatifs)*
Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1 ;*
$\forall\ t_u$*, un terme d'unité dans le texte défini par* $\exists\ w_i \in W_{oi}$ *tel que* w_i *dénote* $i \in I_{um}$ *et* $sim(w_i, t_u) = 1$ *où* $i \in hasUnit(C_i^u)$ *avec* $hasUnit \in R$*,* $C_i^u \in C_{Qty}$ *et* $C_i^u \in signatureR(rel)$ *avec* $rel \in C_{Rel}$*,*
$\forall\ v_i$*, une valeur dans le texte associée à* t_u*,*
$\forall\ t_i$*, un terme dans le texte tel que* $\exists\ w_i \in W_{oi}$ *qui dénote* C_i^u *défini précédemment et* $sim(w_i, t_i) = 1$*,*
v_i *est annoté par* C_i^u *dénoté par* $< numvalC_i^u >$*,* t_i *par* $< Quantity >$ *et* t_u *par* $< um >$ *dans la nouvelle représentation des données.*

Le principe de la représentation des données est illustré dans l'Exemple 2 : Dans la phrase (1), nous souhaitons augmenter l'expressivité des arguments soulignés en utilisant les Définitions 4 et 5. La phrase (2) correspond à la nouvelle représentation des données de la phrase (1). Cette phrase contient une instanciation de la relation n-aire O2_Permeability, qui représente la perméabilité à l'oxygène de l'emballage dans des conditions expérimentales données, définie par une épaisseur (64 μm), une température (23 ° C) et une humidité relative (0%). Plus précisément, la valeur numérique *64* est associée à l'unité de mesure *μm* qui est associée au concept $< Thickness >$. Ainsi, $< numvalthick >$ permet d'annoter la valeur 64. Le terme "thickness" est simplement représenté par $< Quantity >$ et l'unité par $< um >$ pour $< Unit\_Concept >$.

Exemple 2

(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 $< Quantity >$.

Avec notre proposition, les données d'intérêt sont représentées selon leurs concepts dans l'ontologie, leur octroyant une expressivité plus pertinente par rapport au contexte de recherche des arguments de la relation n-aire. Au cours de la deuxième sous-étape, nous proposons de définir des contextes textuels pertinents de recherche par rapport à ces mêmes arguments.

4.2 Deuxième sous-étape : Constitution des sous-corpus

Nous proposons de constituer plusieurs sous-corpus, à exploiter par les algorithmes de fouille de données, construits à partir du corpus initial en découpant les données textuelles selon plusieurs fenêtres textuelles. Cette proposition permet de fouiller des contextes proches des arguments. Pour cela, nous avons choisi l'unité de mesure (Berrahou et al., 2013) associée aux arguments quantitatifs comme descripteur pertinent dans le texte pour définir des contextes favorables à la découverte des arguments recherchés. À partir de ce descripteur, nous proposons deux contextes textuels pertinents de recherche dans les textes.

Définition 6 *(La phrase pivot)*
La phrase pivot est définie comme la phrase où au moins une unité référencée dans la RTO est identifiée.

Définition 7 *(La fenêtre 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.

Les fenêtres textuelles permettent de constituer les différents sous-corpus pertinents. Ainsi, ils constituent à la fois des contextes de recherche proches des arguments de la relation n-aire, tout en étendant les possibilités d'extraction des arguments à plusieurs phrases.

4.3 Troisième sous-étape : Constitution des transactions et items

Dans la sous-étape précédente, nous avons constitué plusieurs sous-corpus à partir des fenêtres textuelles, construits pour permettre une recherche de l'expression des arguments dans des contextes textuels pertinents, c'est-à-dire proches des arguments de la relation n-aire recherchée. Dans cette sous-étape, nous proposons de préparer ces sous-corpus pour l'étape de fouille de données, c'est-à-dire constituer pour chaque sous-corpus, représentant une fenêtre textuelle donnée, l'ensemble des transactions et items pour l'extraction des motifs. L'ensemble des transactions est obtenu selon la Définition 8 et l'ensemble d'items selon la Définition 9.

Définition 8 *(Transaction)*
Une transaction est définie par un ensemble de phrases selon la fenêtre textuelle définie.

Exemple 3
Dans une fenêtre textuelle $f_{\pm 1}$, chaque transaction considérée correspond à un ensemble de phrases composé de la phrase pivot, de sa phrase précédente et de sa phrase suivante.

Définition 9 *(Item)*
Un ensemble d'items I_n d'intérêt est défini comme l'ensemble des n-termes ou concepts voisins des concepts identifiés dans la nouvelle représentation des données guidée par la RTO.

Exemple 4
Considérons la phrase (2) de l'Exemple 2, si nous choisissons de sélectionner les 1-termes plus proches voisins du concept identifié <packaging>, nous obtenons un ensemble d'items composé de {<packaging>, polypropylene, trays, films}.

Dans la section suivante, nous proposons d'appliquer pour chaque fenêtre textuelle, représentée selon son ensemble de transactions et d'items, les algorithmes de fouille de données.

4.4 Quatrième sous-étape : La fouille de données

Dans cette dernière sous-étape du processus d'extraction de connaissances, nous nous intéressons aux relations implicites d'expression des arguments dans les textes, fondées sur la nouvelle représentation conceptuelle des données dans le texte, guidée par la RTO. Nous proposons d'utiliser les approches de fouille de données pour les découvrir. Posons, dans un premier temps, les définitions fondamentales (Agrawal et Srikant, 1995) de fouille de données sur lesquelles reposent nos propositions associées à notre méthode.
Soit $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ l'ensemble des items. Un itemset est un ensemble non vide, non ordonné 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. 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 $IS_1 \subseteq IS_{j_1}, IS_2 \subseteq IS_{j_2}, ..., IS_p \subseteq IS_{j_p}$. **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 θ sont extraits. Soit $\mathcal{M}$ l'ensemble des motifs extraits : $\forall M \in \mathcal{M}$, $Support(M) \geq \theta$.
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., $X \cap Y = \emptyset$. L'"intérêt" et la "qualité" d'une règle sont évaluées par les mesures 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. 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.
Nous proposons d'adapter ces définitions à notre contexte de recherche, pour lequel l'extraction des motifs s'appuie sur la nouvelle représentation des données, fondée sur le niveau conceptuel de la RTO. Nous proposons ainsi de définir les OSP (Ontological Sequential Patterns) qui sont des motifs séquentiels restituant les arguments corrélés, représentés selon leur niveau conceptuel dans la RTO. Nous définissons tout d'abord la notion de séquence guidée par la RTO selon la Définition 10, sur laquelle repose l'extraction des OSP selon la Défintion 11.

Définition 10 *(OS (Ontological Sequence))*
Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1 ;*
Soit $IO = \{IO_1, IO_2, ..., IO_m\}$ *un ensemble d'items* IO_j *où* $IO_j \in W \times W_{oi}$ *ou* $IO_j \in C_{OTR}$*. Un itemset est un ensemble non vide, non ordonné d'items dénoté par*
$(IO_1, IO_2, ..., IO_k)$ *où* $IO_j \in IO$*. OS est un ensemble non vide, non ordonné d'itemsets dénoté par*
$< IOS_1 IOS_2 ... IOS_p >$ *où* $IOS_j \in IOS$*, avec* IOS *l'ensemble d'itemsets.*

L'extraction des OSP consiste alors, à partir de cet ensemble de OS, à rechercher les sous-séquences fréquentes caractérisées par leur support, qui représente le nombre d'occurrences d'un motif dans l'ensemble OS.

Définition 11 *(OSP (Ontological Sequential Patterns))*
$OS_A = < IOS_1 IOS_2 ... IOS_p >$ *est une sous-séquence fréquente d'une autre séquence* OS_B*.*
$OS_B = < IOS'_1 IOS'_2 ... IOS'_m >$*, (*$OS_A \preceq OS_B$*) si* $p \leq m$ *et s'il existe des entiers* $j_1 < j_2 < ... < j_k < ... < j_p$ *tel que* $IOS_1 \subseteq IOS'_{j_1}, IOS_2 \subseteq IOS'_{j_2}, ..., IOS_p \subseteq IOS'_{j_p}$*.*
Soit un support minimum θ*, OSP est défini comme l'ensemble des sous-séquences fréquentes de* OS *tel que* $Support(OSP) \geq \theta$*.*

Nous rappelons que l'extraction de nos OSP s'effectue dans un contexte de recherche défini selon des fenêtres textuelles. Pour représenter cette notion d'extraction selon la fenêtre considérée, nous proposons d'adopter la nomenclature OS_{f_n} qui suggère que l'extraction du motif OSP est supportée dans l'ensemble des séquences OS considérées dans la fenêtre f_n. À la fin du processus d'extraction, nous obtenons automatiquement un ensemble de OSP représentant des règles de corrélations d'arguments dans les textes, fondées sur les différents niveaux conceptuels de la RTO.

Exemple 5
Considérons l'OSP <(packaging)(numvalthick um)> supporté par $OS_{f_{\pm 1}}$*. Cet OSP, extrait dans les*

séquences de la fenêtre textuelle $f_{\pm 1}$ permet de découvrir une corrélation de l'expression du concept emballage ***packaging*** *de la RTO du domaine des emballages avec la représentation de la valeur d'épaisseur de l'emballage* ***thickness****. Ce motif nous permet de découvrir que dans les textes, l'objet étudié (ici l'emballage) et son épaisseur (valeur numérique et unité d'épaisseur) s'expriment fréquemment ensemble dans une fenêtre textuelle maximale de $f_{\pm 1}$.*

Cette découverte de corrélations pertinentes permet de comprendre de quelle manière les arguments sont associés entre eux dans les textes. De plus, la fenêtre textuelle permet de définir l'espace de recherche (ici un ensemble maximal de 3 phrases) de ces arguments dans le texte.
Dans la suite de notre travail, nous proposons une approche hybride fondée sur l'extraction de relations syntaxiques afin d'enrichir et d'étendre les OSP extraits avec des catégories grammaticales et des termes spécifiques pour construire des patrons linguistiques plus précis pour l'identification des données dans un contexte étendu de phrases.

5 Approche hybride fondée sur les relations syntaxiques

Dans cette section, nous présentons, dans un premier temps, les principes essentiels de l'analyse syntaxique, en particulier le choix des relations syntaxiques pour enrichir les OSP extraits. Dans un deuxième temps, nous présentons chaque sous-étape de notre approche hybride représentée dans la Figure 4 pour la construction des patrons d'identification d'arguments corrélés dans les textes.

5.1 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. 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.
Les relations syntaxiques proposent une analyse alternative aux classiques représentations structurelles des phrases en restituant les relations de dépendance des mots de la phrase. Elles sont facilement comprises et efficacement exploitées sans besoin d'expertise linguistique. Pour

l'enrichissement des motifs, elles nous permettent à la fois de gérer le nombre et la qualité des structures linguistiques choisies. Les relations se présentent sous forme de triplets de dépendance associant un rôle grammatical à une paire de mots. Par exemple, la structure *Oil prices futures* est représentée par les triplets *NN(futures, oil)* et *NN(futures, price)* où NN correspond à la forme nominale. L'analyseur syntaxique est associé à un ensemble de règles grammaticales sur lesquelles reposent la majorité des structures linguistiques.
Dans nos travaux, nous choisissons de sélectionner uniquement les relations syntaxiques d'intérêt pour l'identification 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 de découvrir de manière exhaustive 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.

5.2 Principe de l'approche hybride

Dans cette section, nous cherchons donc à extraire les relations syntaxiques (SR pour Syntactic Relations) fréquentes et pertinentes, proches des arguments des relations n-aires recherchées. Ces SR 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'identification d'arguments corrélés dans les textes. La Figure 4 représentant l'approche hybride comporte 4 sous-étapes.

1. La première sous-étape consiste à extraire des SR en utilisant un analyseur syntaxique sur l'ensemble du corpus. L'analyseur analyse chaque phrase du corpus et retourne toutes les SR du corpus. Par exemple, la SR *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 SR 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.
2. La deuxième sous-étape concerne l'extraction des SR proches de la RTO. Tous les termes référencés dans la RTO et qui dénotent les concepts du domaine sont utilisés afin d'identifier les SR les plus pertinentes, i.e. toutes les SR contenant au moins un terme dénotant un concept de la RTO sont extraites selon la Définition 12.

   Définition 12 *(OSR (Ontological Syntactic Relation))*
   SR est un ensemble de relations syntaxiques ;
   W est un ensemble de termes de la phrase ;
   Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ de la Définition 1 ;

   OSR est définie comme l'ensemble des triplets $sr(w_1, w_2)$ avec $sr \in SR$, où $w_1 \in W \times W_{oi}$ et $w_2 \in W \times W_{oi}$.
3. La troisième sous-étape consiste à extraire les OSR proches des OSP. Ainsi, à partir de l'ensemble des 0SR, nous nous intéressons principalement 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*.

4. La quatrième sous-étape propose de construire les OLSP (Ontological Linguistic Sequential Patterns). Toutes les OSR restituant les arguments corrélés découverts dans l'étape de fouille de données sont utilisées pour enrichir les OSP afin d'obtenir des patrons linguistiques d'identification définis comme suit.

   Définition 13 *(OLSP (Ontological Linguistic Sequential Pattern))*
   Considérons $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *de la Définition 1 ;*
   OSR *un ensemble de triplets selon la Définition 12 ;*
   OSP *un ensemble de motifs fréquents selon la Définition 11 ;*
   Pour chaque $osr \in OSR$ *représentée par* $sr(w_1, w_2)$*, pour chaque* $osp \in OSP$ *défini comme une sous-séquence fréquente dénotée par* $< IOS_1 IOS_2 ... IOS_p >$ *où* IO_j *peut être dénoté par* w_1 *ou* w_2 *,*
   OLSP est défini comme l'ensemble des sous-séquences fréquentes osp *enrichies par les* osr*.*

   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 SR 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), NN(film ou films, HPMC)*. Le terme *HPMC* de la relation dénote le concept *packaging* dans la RTO, on conserve le concept dans la relation. Les règles de génération des structures linguistiques sont intégrées dans l'analyseur utilisé et sont illustrées dans l'Exemple 6.

   Exemple 6
   Génération :
   - NN(thickness, film ou films) => film/films thickness
   - NN(film ou films, HPMC) =>(packaging) film/films
   Hybridation avec motif séquentiel <(packaging)(numvalthick um)> :
   - <(packaging) film/films thickness (numvalthick um)>

Après validation, les OLSP obtenus peuvent être appliqués sur les textes pour identifier les phrases comportant la séquence d'arguments corrélés recherchés.
Par exemple, la phrase *mango films thickness was 0.17 ± 0.02 mm* est identifiée en utilisant l'OLSP <(packaging) film/films thickness (numvalthick um)>.

6 Expérimentations

Nous avons mené des expérimentations sur un corpus constitué de plus de 35 000 phrases extraites à partir de 115 documents, 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

OSP en utilisant le processus d'extraction de connaissances fondé sur la nouvelle représentation des données. Puis, après une étape de validation, nous construisons les OLSP d'arguments corrélés pour leur identification dans le texte.

6.1 Extraction des OSP

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, $f_{\pm 2}$). Le nombre de transactions varie selon la fenêtre textuelle évaluée de 5 000 à 35 000 phrases. Le nombre d'items 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 tels 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 est un des algorithmes de référence à l'état de l'art car il permet d'extraire un ensemble concis de motifs sans redondance ni perte d'information, en exploitant la propriété de fermeture des motifs. Cette propriété permet d'élaguer l'espace de recherche en évitant l'extraction de sous-séquences redondantes ayant le même support. Pour la découverte des règles séquentielles, nous utilisons 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. Le premier critère permet de sélectionner uniquement les OSP comportant au moins un argument des relations n-aires recherchées et représentées dans la RTO. Le second critère permet d'extraire les OSP issus de l'intersection de plusieurs fenêtres étudiées, en se basant toujours sur le critère de support de ces OSP.
Résultats quantitatifs. Le nombre de OSP extraits varie en fonction des critères de sélection appliqués. Par exemple, nous obtenons plus de 52 000 règles et OSP 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 OSP extraits.
Résultats qualitatifs. Le Tableau 1 restitue un ensemble d'OSP extraits avec le processus d'extraction des connaissances fondé sur la nouvelle représentation des données que nous proposons. Ces OSP correspondent aux motifs obtenus à la fin du processus d'extraction avant hybridation avec l'analyse syntaxique. Nous constatons que les OSP et règles extraits sont plus expressifs et proches des instances d'arguments recherchés, ce qui montre l'intérêt de la nouvelle représentation des données proposée, guidée par la RTO de domaine. De plus, les résultats mis en évidence dans le tableau montrent plusieurs relations implicites d'expression des arguments dans les textes, révélant plusieurs co-occurrences d'arguments dans les textes, que nous n'aurions pas pu déterminer sans l'étape d'extraction des connaissances. Nous découvrons que l'expression de l'argument symbolique *packaging* et de l'argument quantitatif *thickness* apparaissent fréquemment ensemble dans le texte et que cette corrélation se manifeste dans une fenêtre textuelle maximale de $f_{\pm 1}$, c'est-à-dire dans un ensemble de trois phrases ;

Fenêtre textuelle	**OSP et règles**	Support
$f_{\pm 1}$	<**(packaging)(numvalthick um)**>	0.5
	<(film thickness)(rh)>	0.1
	<**(packaging)(permeability)**>	0.6
f_0	<(pressure)(water permeability)>	0.05
	<(oxygen permeability)(pressure)>	0.05
$\bigcap f_n$	packaging => numvalthick	>0.05
	temperature => numvalrh	
	<**(numvaltemp)(numvalrh%)**>	
	<(packaging)(numvalthick)>	
	<**(packaging)(numvaltemp °c)**>	
	packaging => temperature numvalrh	

TAB. 1: Extrait de OSP et règles découverts - $\bigcap$ critère d'intersection des fenêtres

Nous découvrons l'expression co-occurrente, dans la même phrase, des arguments quantitatifs *temperature* et *relative humidity* ; Nous découvrons que les quatre arguments précédemment cités apparaissent fréquemment dans un contexte proche, également dans un contexte de trois phrases ; Nous découvrons que l'argument *packaging* pourrait jouer le rôle de déclencheur de l'instance de relation car il apparait fréquemment dans de nombreux OSP restituant les corrélations précédentes. Il peut donc être utilisé pour regrouper les arguments dans l'instance de relation recherchée, même si ceux-ci sont éloignés dans le document puisque le déclencheur est présent dans chaque motif permettant l'identification de 2 ou plusieurs arguments.

6.2 Construction des OLSP à partir des OSP

Résultats quantitatifs des SR. Pour extraire les relations syntaxiques de notre corpus de plus de 35 000 phrases, 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 SR. Nous avons réduit cet ensemble en utilisant les termes référencés dans la RTO. Nous constituons ainsi, un sous-ensemble de OSR d'intérêt proches des arguments de la relation n-aire recherchée. De ce nouvel ensemble, nous conservons uniquement les OSR qui restituent les corrélations d'arguments dans les textes et que nous avons découvertes dans les OSP. Les catégories grammaticales et termes restitués sont ajoutés aux OSP sélectionnés afin de construire des OLSP. Le nombre de OSR finalement conservées pour l'enrichissement des motifs est de 6 600.
Résultats quantitatifs d'identification d'arguments corrélés par approche hybride. Les expérimentations ont été menées en trois 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 (relations O2permeability, CO2permeability, H2Opermeability). Les résultats sont restitués 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. Dans un premier

temps, nous avons utilisé uniquement l'analyse syntaxique afin d'évaluer si les SR sont suffisamment expressives pour identifier les arguments dans le texte. Rappelons que l'analyse syntaxique restitue les SR uniquement au niveau de la phrase et qu'elle ne restitue au mieux que l'association de deux arguments de la relation n-aire du fait de la structure en triplet des SR. Parmi les 100 premières SR les plus fréquentes, seuls 2 arguments quantitatifs (*temperature* et *thickness*) apparaissent (au rang 23 et 43) sans restituer de corrélations. En utilisant ces SR, nous obtenons une précision de 0.3 et un rappel de 0.1 pour l'identification de l'argument *thickness*, et une précision de 0.2 et un rappel de 0.04 pour l'identification de l'argument *temperature*. Ces premiers résultats montrent que considérer uniquement les SR les plus fréquentes dans le texte, sans tenir compte des corrélations découvertes dans les motifs et sans s'appuyer sur la RTO pour le filtrage sémantique pertinent, celles-ci restituent très peu d'instances d'arguments. Dans un second temps, nous avons effectué les évaluations avec les OSP génériques sans hybridation qui permettent de découvrir les corrélations entre arguments. Les résultats, restitués dans le Tableau 2, montrent que ces OSP fondés essentiellement sur la représentation conceptuelle des arguments identifient de manière exhaustive les différentes corrélations d'arguments dans le texte mais la validation reste encore "bruitée", caractérisée par une faible précision. En combinant les SR extraites par approche hybride et les OSP, c'est-à-dire les SR qui s'appuient sur la RTO et tiennent compte des corrélations d'arguments découvertes dans les OSP, pour la construction des OLSP, la qualité de l'identification des arguments de la relation n-aire est sensiblement améliorée. Les résultats montrent en effet que cette approche permet d'identifier de manière tout aussi exhaustive et sans bruit excessif pour la validation des instances d'arguments obtenues.

Type d'évaluation	**OSP**			**OLSP**		
	Précision	*Rappel*	*F-Mesure*	*Précision*	*Rappel*	*F-Mesure*
Évaluation générale	0.5	0.8	0.6	**0.7**	0.8	0.7
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. 2: Évaluation de l'identification des arguments corrélés dans les textes

7 Conclusion

Dans cet article, nous proposons une approche hybride combinant les techniques de fouille de données et les relations syntaxiques pour l'identification 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 identifier 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 (Ressource Termino-Ontologique). Dans ce processus, nous proposons également d'évaluer plusieurs fenêtres textuelles pour la découverte des OSP (Ontological Sequential Patterns) pertinents concernant les arguments recherchés dans une séquence de phrases définie par la fenêtre. Puis, dans une seconde étape, nous proposons d'extraire les relations syntaxiques OSR (Ontological Syntactic Relations) pertinentes à l'enrichissement des OSP pour la construction des OLSP (Ontological Linguistic Sequential Patterns), patrons d'identification qui combinent plusieurs niveaux d'abstraction (terme, catégorie grammaticale et concept) pour une identification plus efficace des arguments dans les textes. Au cours des expérimentations menées, les patrons ont permis l'identification 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 identifiées. 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 OLSP d'arguments corrélés dans une plateforme d'annotation existante, @Web [2] (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 OLSP permettraient alors d'identifier les phrases dans 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.

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

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.

Bloehdorn, S., P. Cimiano, A. Hotho, et S. Staab (2005). An ontology-based framework for text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 20*(1), 87–112.

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 ProceedinIN 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.

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.

Minard, A.-L., B. Grau, et A.-L. Ligozat (2010). Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données. In *Journées internationales d'analyse statistique des données textuelles (JADT)*.

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.

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.

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 (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. In *OTM Conferences (2)*, pp. 662–679.

Van Landeghem, S., Y. Saeys, B. De Baets, et Y. Van de Peer (2009). Analyzing text in search of bio-molecular 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. 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.