Extraction de relations n-aires interphrastiques guidée par une RTO

Akila Ghersedine, Patrice Buche, Juliette Dibie, Nathalie Hernandez, Mouna Kamel

Extraction de relations n-aires interphrastiques guidée par une RTO

Akila Ghersedine[1,2], Patrice Buche[2], Juliette Dibie-Barthélemy[3]
Nathalie Hernandez[1], Mouna Kamel[1]

[1]*IRIT-IC3, Toulouse*
{hernande, kamel, ghersedine}@irit.fr
[2]*INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France*
LIRMM, CNRS-UM2, F-34392 Montpellier, France
Patrice.Buche@supagro.inra.fr
[3]*INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France*
Juliette.Dibie@agroparistech.fr

RÉSUMÉ. *Nous proposons dans cet article une méthode d'extraction d'instances de relations n-aires dans un texte guidée par une Ressource Termino-Ontologique (RTO) de domaine. Une RTO est une ressource comportant une composante conceptuelle (l'ontologie) et une composante terminologique (la terminologie), dans laquelle les termes sont distingués des concepts qu'ils dénotent. L'ontologie permet la modélisation de relations n-aires, reliant des arguments pouvant être des concepts symboliques et des quantités. La méthode proposée s'applique aux relations n-aires formulées de façon implicite dans le texte et dont les instances d'arguments peuvent être exprimées à travers différentes phrases du texte.*

ABSTRACT. *We propose in this paper a method to extract instances of n-ary relations in a text guided by an Ontological and Terminological Resource (OTR). An OTR is a resource composed of a conceptual component (the ontology) and a terminological component (the terminology) in which the terms are distinguished from the concepts they denote. The ontology allows n-ary relationships to be described between arguments which can be symbolic concepts and quantities. The method is dedicated to the extraction of n-ary relations which are implicit in the text and whose instances of arguments may be expressed in different sentences of the text.*

MOTS-CLÉS : *ontologies, extraction d'information, web sémantique, logique floue*

KEYWORDS: *ontologies, information extraction, semantic web, fuzzy logic*

CORIA 2012, pp. 179–190, Bordeaux, 21-23 mars 2012

1. Introduction

L'extraction d'information consiste à identifier de l'information bien précise dans un texte en langue naturelle et à la représenter sous forme structurée, dans le but de construire des bases de données, des ressources terminologiques ou ontologiques. Cela suppose une compréhension du texte, subordonnée à la forme du document, au support, à la langue ou encore au domaine. La phase d'analyse, dont le rôle est de mettre en relation les éléments du texte, peut être guidée par des connaissances, des ressources lexicales, sémantiques et conceptuelles adaptées aux documents et au domaine. La majeure partie des systèmes d'extraction ont été développés pour des domaines spécifiques et donc des langages de spécialité. Pour chacun d'eux, il s'agit de spécifier les informations recherchées, les ressources utilisées, ainsi que la tâche d'extraction. Il existe deux grandes familles de techniques d'extraction : l'approche linguistique et l'approche statistique. Les deux approches ont été utilisées dans de nombreux travaux. L'approche linguistique tend à utiliser des outils du TAL (analyseur syntaxique, analyseur morphologique, analyseur sémantique, etc.), et à définir un ensemble de patrons lexico-syntaxiques qui sont des règles décrivant une expression régulière, formée de mots et de catégories grammaticales. Ces patrons sont souvent porteurs d'un ou plusieurs marqueurs linguistiques ((Aussenac *et al.,* 2000), (Djioua *et al.,* 2006), (Khelif, 2006)). L'approche statistique met en oeuvre des algorithmes d'apprentissage, afin d'acquérir automatiquement des connaissances, à partir d'un corpus annoté manuellement (Bessières *et al.,* 2001). La combinaison de ces deux approches semble aujourd'hui prometteuse dans la mesure où les résultats fournis par les Machines à Support Vectoriel (SVM) donnent de meilleurs résultats que les approches prises individuellement (Burcu *et al.,* 2006), (Giuliano *et al.,* 2006). Ces méthodes sont généralement appliquées pour extraire les éléments pertinents qui se situent autour d'une relation sémantique portée par un prédicat, et connaître les rôles des arguments du prédicat (Khelif, 2006). Il peut également être nécessaire de représenter une relation sémantique n-aire ayant plus de deux arguments. Elle peut être exprimée autrement qu'à travers un schéma prédicatif : la tâche d'extraction des éléments appartenant alors à des structures plus complexes est confrontée à la résolution de difficultés linguistiques telles que le rattachement des compléments simples, les coréférences/anaphores (Pustejovsky *et al.,* 2002) ou encore l'implicite, et donc à des problématiques du TAL.

Ce travail se situe dans le cadre de l'identification d'instances de relations n-aires formalisées au sein d'une Ressource Termino-Ontologique (RTO). Plus spécifiquement, la méthode propose des perspectives de recherche originales dans la mesure où elle s'applique aux relations n-aires qui peuvent être formulées de façon implicite dans le texte et dont les arguments peuvent être exprimées à travers différentes phrases du texte. Ce cas de figure est fréquent, notamment dans les articles scientifiques où les conditions expérimentales sont décrites à travers différentes phrases d'une section du texte, sachant que ces conditions sont directement liées aux résultats présentés et discutés dans une autre section. L'extraction d'instances de relations n-aires a pour objectif d'augmenter la connaissance formalisée dans la RTO par un peuplement de celle-ci

avec de nouvelles instances. La méthode d'extraction d'instances de relations n-aires proposée dans ce papier repose sur une RTO. Dans l'esprit des liens interphrastiques de (Sager *et al.,* 1980), cette RTO contient les ressources qui permettent d'unir des segments de phrases du texte de manière à ce qu'elles 'forment un discours' plutôt qu'une 'séquence aléatoire' de phrases. Nous présenterons, dans la section 2, la Ressource Termino-Ontologique (RTO) sur laquelle repose notre travail. Notre méthode d'extraction d'instances de relations n-aires sera présentée dans la section 3, avec en premier la description de l'identification des instances des arguments d'une relation n-aire donnée, puis l'identification et l'extraction des instances de cette relation. La section 4 présentera les premiers résultats expérimentaux. Enfin, nous conclurons et présenterons les perspectives de notre travail dans la section 5.

2. La Ressource Termino-Ontologique (RTO)

Notre approche repose sur l'utilisation d'une Ressource Termino-Ontologique (RTO) dédiée à la tâche d'annotation de relations n-aires dans les documents (Touhami *et al.,* 2011). Une RTO est une ressource comportant une composante conceptuelle (l'ontologie) et une composante terminologique (la terminologie), dans laquelle la manifestation linguistique (le terme) se distingue de la notion qu'elle dénote (le concept). La composante conceptuelle de la RTO est composée de deux parties principales : une partie générique, communément appelée ontologie noyau, qui permet de représenter la structure de l'ontologie dédiée à la tâche d'annotation de relations n-aires, et une partie spécifique, communément appelée ontologie de domaine, qui dépend du domaine étudié. La figure 1 présente la partie générique de la RTO et sa spécialisation dans le domaine de la microbiologie prévisionnelle. L'ontologie noyau se compose de trois catégories de concepts génériques : *Dimension*, *UM_Concept* et *T_Concept*. Nous ne présenterons ici que le concept générique *T_Concept* qui nous intéresse pour ce papier. Le concept générique *T_Concept*, pour *Terminological Concept*, généralise les concepts *Simple_Concept*, *Unit_Concept* et *Relation*, auxquels sont associés un ou plusieurs termes dans la composante terminologique. Le concept générique *Simple_Concept* regroupe les concepts *Quantity* et *Symbolic_Concept*. Un concept symbolique est caractérisé par sa hiérarchie de spécialisation (par exemple *Gram+* et *Gram-* sont des spécialisations de *Microorganism*). Une quantité est caractérisée par un ensemble d'unités de mesure, sous-concepts du concept générique *Unit_Concept*, et éventuellement un domaine de valeurs. La propriété *hasUnitConcept* permet de relier une quantité à ses unités de mesure. Par exemple, le concept *Relative_humidity* peut être exprimé dans l'unité '%' et a pour domaine de valeurs [0, 100]. Le concept *Relation* permet de représenter des relations n-aires entre des concepts simples. La signature d'une relation est définie par un domaine et un co-domaine. Le domaine et le co-domaine peuvent être composés d'un ou plusieurs concepts simples. Les concepts du domaine sont appelés *arguments d'accès*, ceux du co-domaine *argument résultat*.

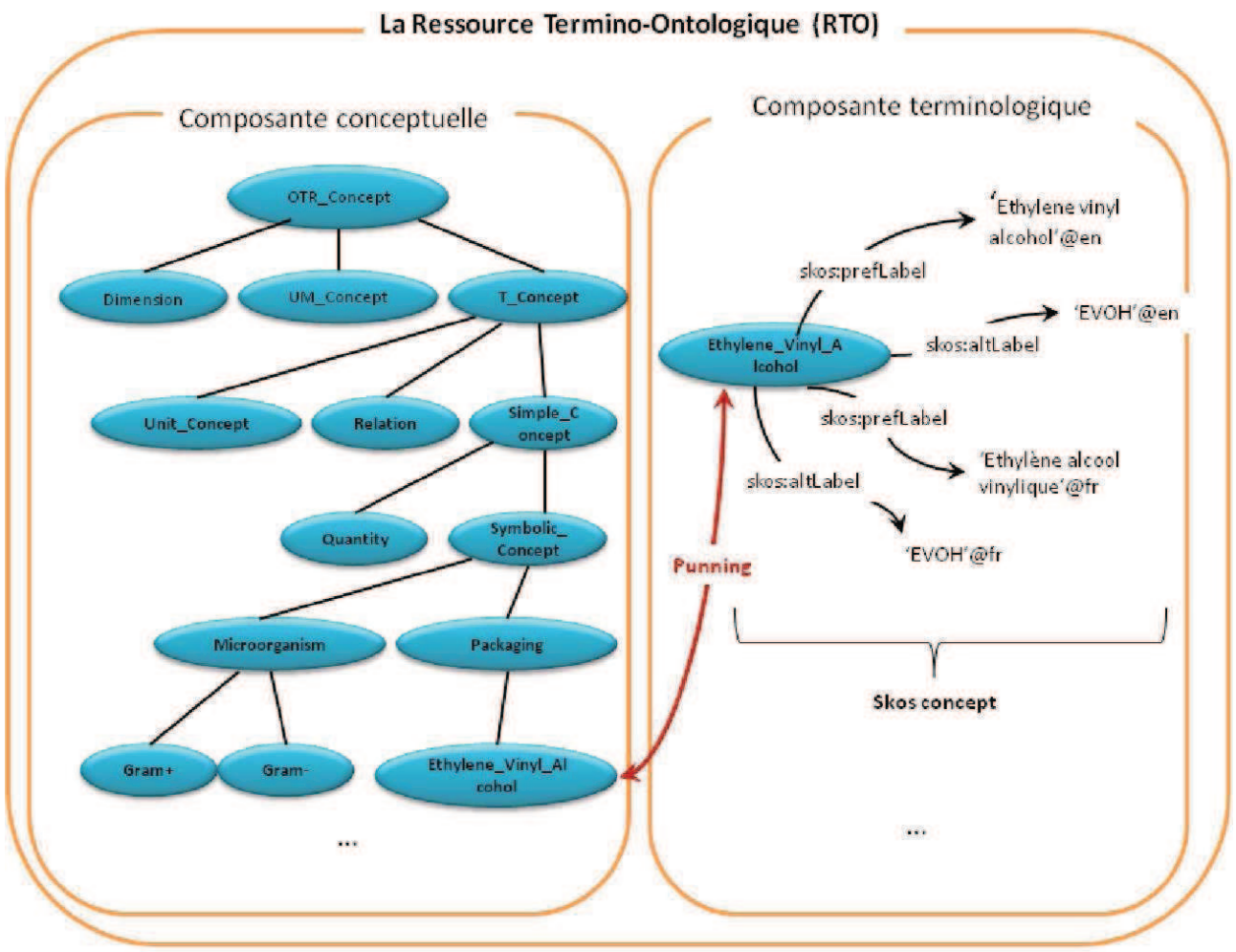

Figure 1. *La RTO dans le domaine de la microbiologie prévisionnelle.*

La partie spécifique de la RTO contient les concepts spécifiques au domaine d'application étudié. Ils apparaissent dans la RTO comme sous concepts des concepts génériques de l'ontologie noyau.

Exemple 1 *Dans le domaine de la microbiologie prévisionnelle, le concept relation spécifique* O2_permeability_Relation *présenté dans la figure 2 permet de représenter une relation n-aire. Ses arguments d'accès sont les concepts relatifs au nom de l'emballage, à son épaisseur, à la température et à l'humidité relative de l'environnement. L'argument résultat représente la perméabilité de l'emballage à l'oxygène dans ces conditions expérimentales.*

La composante terminologique représente la terminologie de la RTO : elle contient l'ensemble des termes du domaine étudié. Au moins un terme de la composante terminologique est associée à chaque sous concept du concept générique *T_Concept*. Par exemple, dans la figure 1, les termes *Ethylene vinyl alcohol* et *EVOH* sont associés au concept *Ethylene_Vinyl_Alcohol*. Chaque sous-concept du concept générique *T_Concept* est, dans une langue donnée, caractérisé par un label préféré et éventuellement par un ensemble de labels alternatifs, qui correspondent à des synonymes ou des abréviations.

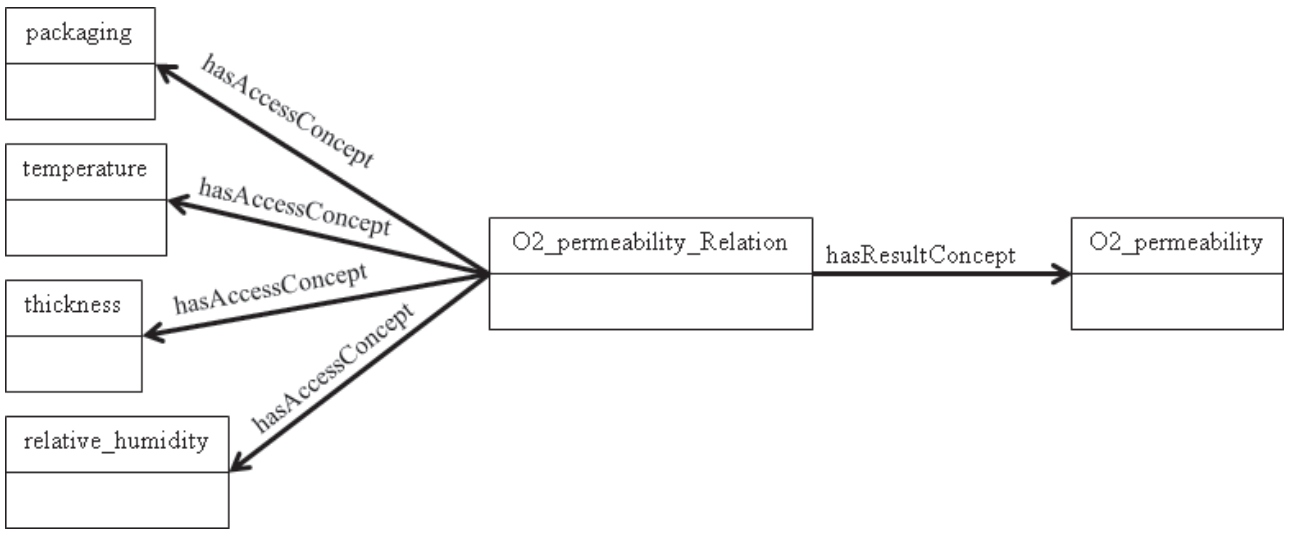

Figure 2. *Un exemple de relation n-aire dans le domaine de la microbiologie*

L'objectif de ce travail est d'identifier, au sein d'articles scientifiques, des instances de relations n-aires définies dans la RTO. Une étude de corpus a montré que les instances de relations n-aires sont le plus souvent exprimées, d'une part, de façon implicite et, d'autre part, à travers plusieurs phrases, voire plusieurs paragraphes.

Exemple 2 *Considérons le paragraphe* "The permeability of the LDPE film was estimated independently by the cell permeability method. At 100% relative humidity and 20 °C, O2 permeability was 1078 amol x m-1 x s-1 x Pa-1" *issu de l'article (Charles* et al., *2003). Conformément au schéma de la relation* O2_Permeability_Relation *de la figure 2, nous identifions dans ce paragraphe une instance de cette relation (bien qu'aucun marqueur lexical associé à l'expression de cette relation ne soit présent), grâce aux instances d'arguments* LDPE *(instance du concept symbolique* Packaging*),* 100% relative humidity *(instance de la quantité* Relative_Humidity*),* 20 °C *(instance de la quantité* Temperature*), et* 1078 amol x m-1 x s-1 x Pa-1 *(instance de la quantité* O2_Permeability*).*

Identifier une instance de relation n-aire revient alors à identifier une instance de chacun de ses arguments et à relier ces instances entre elles. La méthode que nous proposons ici pour établir ces liens ne vise pas à résoudre les phénomènes linguistiques d'anaphore ou de coréférence, mais consiste à identifier un argument pivot pour la relation et à localiser dans le texte toutes les instances d'arguments partageant le même contexte que chaque instance de l'argument pivot. Cette méthode est décrite en détail dans la section suivante.

3. Méthode d'identification des instances de relations n-aires

Notre méthode d'extraction de relations n-aires se décompose en trois étapes : l'identification des instances des arguments d'une relation n-aire donnée (section 3.1),

puis l'identification de l'argument pivot (section 3.2) et enfin l'identification et l'extraction des instances de la relation n-aire elle-même (section 3.3).

3.1. *Identification des instances des arguments*

L'identification des instances des arguments d'une relation n-aire se décompose en deux étapes : (1) identification des instances de concepts symboliques et de quantités (descendants de *Simple_Concept*) ; (2) identification des valeurs et de l'unité de mesure (*Unit_concept*) associés à une instance de quantité (*Quantity*).

3.1.1. *Identification des instances de concepts symboliques et de quantités*

Un lexique est construit à partir de l'ensemble des termes associés, dans la composante terminologique de la RTO, aux spécialisations de *Simple_Concept*. Pour pouvoir identifier les instances de concepts symboliques et de quantités (sous-concepts de *Simple_Concept*), le lexique ainsi construit est projeté sur le corpus.

Exemple 3 *Dans le paragraphe de l'exemple 2,* LDPE *(appartenant au lexique) est reconnu comme une instance du concept symbolique* Packaging.

3.1.2. *Identification des valeurs et de l'unité de mesure associés à une instance de quantité*

L'identification des valeurs et de l'unité de mesure associés à une instance de quantité est faite à l'aide de patrons lexico-syntaxiques. Un patron lexico-syntaxique décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte répondant à ce format (Aussenac *et al.,* 2000). Nos patrons sont construits à partir de la propriété *hasUnit-Concept*, définie dans la RTO, qui permet de relier une instance de concept simple *Quantity* (précédemment identifiée dans le corpus) à ses unités de mesure (instances de sous-concepts du concept *Unit_Concept*). Ces patrons identifient comme valeur d'une instance de *Quantity* toute expression composée d'un nombre (ou d'un intervalle de nombres) suivi d'une unité, co-occurant dans une même phrase avec un terme désignant une quantité (i.e. un sous-concept de *Quantity*). Pour extraire ces instances, les patrons sont projetés sur le corpus qui a été préalablement traité par des outils de TAL pour la segmentation, l'analyse lexicale et l'étiquetage grammatical.

Exemple 4 *Dans l'expression "*where RH changed from 0 to 100%*", l'expression "*from 0 to 100%*" sera considérée comme une valeur d'une instance du concept* Relative_Humidity *(intervalle [0-100]), sous-concept de* Quantity*, parce qu'elle co-occurre avec le terme* RH *désignant le concept* Relative_Humidity.

On remarquera que la valeur associée à une instance de quantité peut être de nature imprécise (par exemple, être représentée par un intervalle de valeurs). Dans la suite de ce travail, nous avons fait le choix de représenter une donnée imprécise comme un ensemble flou (Zadeh, 1978). Ce choix est justifié par le fait que la collection

d'instances de relations n-aires extraites des textes est interrogée par un système d'interrogation flexible à base de préférences également représentées par des ensembles flous (Buche *et al.,* in press). Ce choix permet d'une part d'utiliser un cadre homogène, la théorie de la logique floue, pour représenter à la fois des données imprécises et des préférences (Dubois *et al.,* 1997) et d'autre part de s'appuyer sur la théorie des possibilités pour comparer les données imprécises aux préférences exprimées dans les requêtes. La notion d'ensemble flou est une extension de la notion d'ensemble classique. Dans un ensemble flou F, les éléments peuvent appartenir partiellement à F avec un degré d'appartenance compris entre 0 (éléments qui n'appartiennent pas à F) et 1 (éléments qui font complétement partie de F). Le degré d'appartenance d'un élément $x \in X$ pour l'ensemble flou F est noté $\mu_F(x)$. Le support d'un ensemble flou F défini sur un domaine X est l'ensemble (ordinaire) des éléments $x \in X$ tels que $\mu_A(x) > 0$. Le noyau d'un ensemble flou F défini sur un domaine X est l'ensemble (ordinaire) des éléments $x \in X$ tels que $\mu_A(x) = 1$. Un ensemble flou défini sur un domaine de valeur continu est appelé par la suite *CFS* (pour *continuous fuzzy set*). Pour représenter une donnée imprécise, nous utiliserons plus précisément la notion d'ensemble flou trapézoidal, qui est un cas particulier de *CFS*, décrit par son support $sup = [min_{sup}, max_{sup}]$ et son noyau $ker = [min_{ker}, max_{ker}]$ (voir (Buche *et al.,* in press) pour plus de détails).

Exemple 5 *A partir de l'expression "*where RH changed from 0 to 100%*", on extrait une instance du concept* Relative_Humidity *associée à l'intervalle [0-100]. Cet intervalle est représenté par le CFS ayant pour support* $sup = [min_{sup} = 0, max_{sup} = 100]$ *et pour noyau* $ker = [min_{ker} = 0, max_{ker} = 100]$.

3.2. *Identification de l'argument pivot*

A l'issue des deux étapes décrites ci-dessus, le texte du document a été annoté avec des instances de concepts symboliques, des instances de quantités et éventuellement leurs valeurs numériques et unités de mesure associées. Avant de continuer l'analyse du document, nous proposons d'évaluer la pertinence du document pour la relation n-aire cible. Si le texte contient au moins une instance de l'argument résultat de la relation cible, alors le document est considéré comme potentiellement pertinent et son analyse peut continuer.

Pour construire une instance de la relation n-aire cible, il faut relier les instances des arguments qui composent sa signature. Afin d'établir un lien sémantique entre les instances de ses arguments, nous proposons de définir la notion d'*argument pivot*. L'argument pivot est l'argument d'accès de la relation ciblée qui co-occurre le plus fréquemment en corpus avec un des arguments résultats de cette même relation. En effet, pour des considération "sémantico-pragmatiques", l'expression d'un argument nécessite de le re-situer dans un contexte et donc de le lier sémantiquement à au moins un autre argument de la relation. Ainsi à chaque relation n-aire de la RTO est associée un argument pivot. Cet argument pivot sera déterminé de manière semi-automatique, avec validation par l'utilisateur.

Exemple 6 *Soit l'extrait de texte suivant : "*The permeability of the **LDPE** film was estimated independently by the cell permeability method. **O2 permeability** ranged from 77 to 1970 amol.m-1.s-1.Pa-1.*". Dans cet extrait, une instance de l'argument résultat* O2_Permeability *de la relation n-aire* O2_permeability_Relation *est trouvée dans la deuxième phrase. Dans un contexte de taille 2, l'instance du concept symbolique* Low_Density_Polyethylene_Film *(dénoté dans le texte par l'abréviation* LDPE*), spécialisation du concept symbolique* Packaging*, est candidat pour être l'argument pivot.*

3.3. *Identification et extraction des instances d'une relation n-aire*

L'identification d'une instance de l'argument résultat étant indispensable pour reconnaitre une instance de la relation n-aire cible, nous considérons que chaque instance d'argument résultat identifiée donnera lieu à la création d'une instance de la relation. Une instance de relation est représentée par un graphe, appelé *graphe résultat*, qui est composé de : (i) la référence du document, (ii) l'instance de l'argument résultat et (iii) l'instance de l'argument pivot trouvée dans le contexte de l'instance de l'argument résultat (cf. sous-section 3.2).

Exemple 7 *Le graphe résultat RDF associé à l'instance d'argument résultat de l'exemple 6 est présenté dans la figure 3. L'instance de la relation n-aire* O2Permeability_Relation *ayant pour URI* $O2PRel1$ *est extraite de l'article référencé (Charles* et al.*, 2003). Elle est liée à l'instance du concept symbolique* Low_Density_Polyethylene_Film *considéré comme argument pivot. Elle est également liée à l'instance d'argument résultat* O2Permeability *ayant pour URI* $O2P1$*. De plus, l'argument résultat étant une quantité, son instance est liée à sa valeur numérique de nature imprécise par l'instance de CFS ayant pour URI CFS1.*

Ce graphe résultat correspond à un "fragment" de l'instance de la relation n-aire cible, que nous allons maintenant chercher à compléter avec les instances des autres arguments de la relation trouvées dans le texte. Nous ne nous intéresserons dans la suite qu'aux instances de l'argument pivot qui sont été associées à une instance de l'argument résultat, cette association étant matérialisée par le graphe résultat décrit ci-dessus. Pour chaque instance d'arguments d'accès de la relation n-aire cible trouvée dans le contexte d'une instance de l'argument pivot, on construit un graphe, appelé *graphe argument d'accès*, composé de : (i) la référence du document, (ii) l'instance de l'argument pivot, (iii) l'instance de l'argument d'accès et (iv) la valeur associée, dans le même contexte, à l'instance de l'argument d'accès si cet argument est une quantité. Le contexte d'une instance de l'argument pivot correspond à une fenêtre composée de m phrases après l'instance, m étant défini de manière empirique.

Exemple 8 *Soit l'extrait de texte suivant : "*Low density polyethylene package film of 50 μm **thickness** was used.*". Le graphe RDF associé à l'instance d'argument*

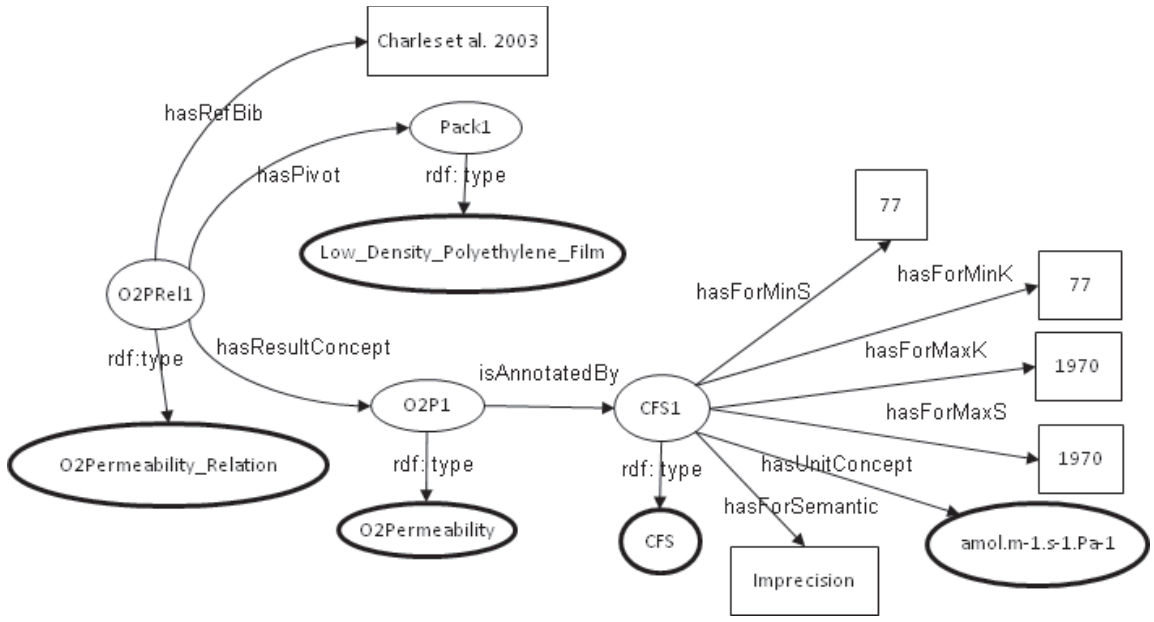

Figure 3. *Un exemple de graphe résultat*

d'accès Thickness *est présenté dans la figure 4. L'instance de la relation n-aire* O2Permeability_Relation *ayant pour URI* $O2PRel2$ *est extraite de l'article référencé (Charles* et al., *2003). Elle est liée à l'instance du concept symbolique* Low_Density_Polyethylene_Film *considéré comme argument pivot. Elle est également liée à l'instance d'argument d'accès* Thickness *ayant pour URI* $Thick1$. *De plus, l'argument d'accès étant une quantité, son instance est liée à sa valeur par l'instance de CFS ayant pour URI CFS2.*

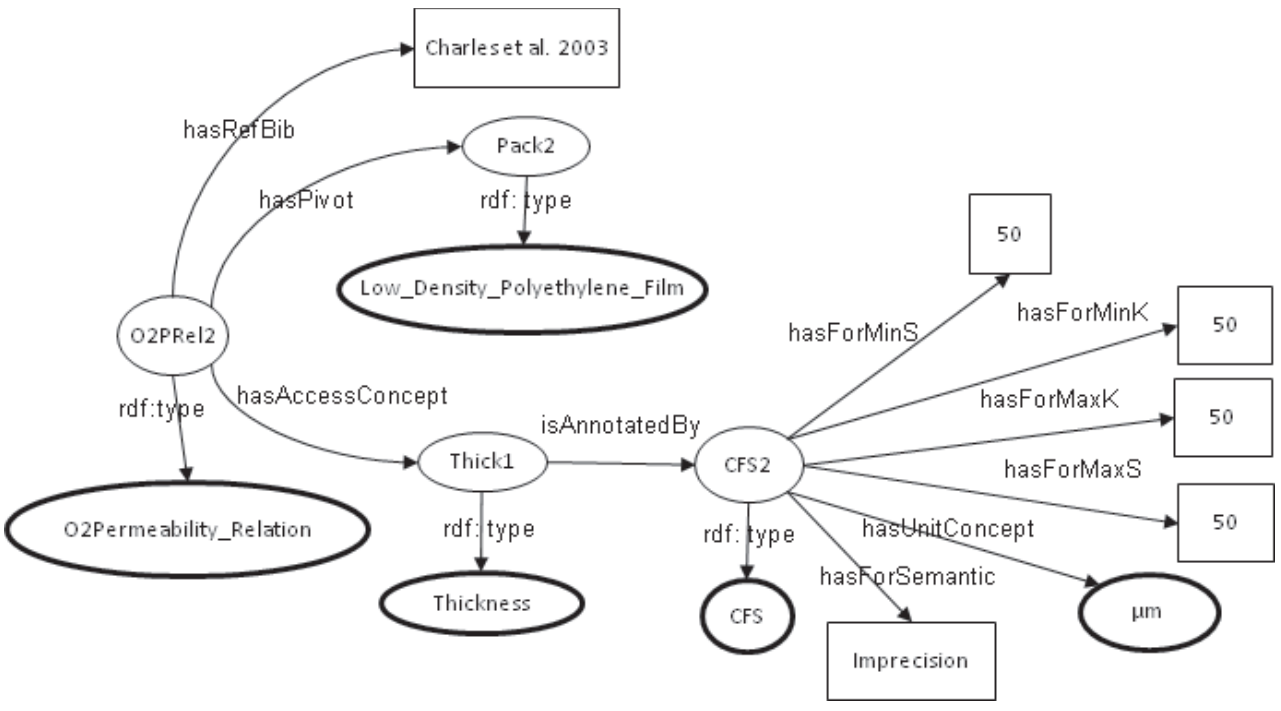

Figure 4. *Un exemple de graphe argument d'accès*

Pour pouvoir construire une instance de relation n-aire, il faut fusionner les différents graphes générés. Cette fusion repose sur la référence du document et des instances de l'argument pivot. Ainsi chaque graphe résultat est fusionné avec tous les graphes ayant la même référence de document et contenant une instance du même argument pivot d'une relation n-aire cible. A l'issue de cette fusion, tous les arguments de la relation n-aire ne sont pas forcément renseignés. Une étape de validation manuelle permet à l'ontologue d'éventuellement compléter l'instance de la relation n-aire.

Exemple 9 *A partir du graphe d'argument résultat de l'exemple 7 et du graphe d'argument d'accès de l'exemple 8, on obtient un graphe fusionné. Dans celui-ci, l'instance de l'argument pivot est également une instance d'un argument d'accès de la relation, ici une instance de l'argument d'accès sous-concept de* Packaging. *Le graphe fusionné est également constitué d'une instance de l'argument d'accès* Thickness *(cf figure 4) et d'une instance de l'argument résultat* O2Permeability *(cf figure 3).*

4. Expérimentation

Nous avons évalué notre approche sur un corpus composé de 11 articles scientifiques écrits en langue anglaise et traitant de microbiologie prévisionnelle, en nous basant sur la RTO présentée figure 1. Les documents ont été annotés manuellement dans un premier temps, puis analysés automatiquement selon l'approche proposée. L'évaulation consiste à comparer les annotations produites dans les deux cas, pour la recherche d'instances de la relation *O2_permeability_Relation* représentée dans la figure 2. Chacun des articles contient au moins une instance d'un argument de la relation (i.e. on y retrouve une occurrence d'instances des concepts *packaging*, *temperature*, *thickness*, *relative_humidity*, *02_Permeability*) mais seulement 8 d'entre eux contiennent au moins une instance de la relation. L'évaluation porte sur chaque étape de l'approche proposée.

4.1. *Identification des instances des arguments*

Nous avons, dans un premier temps, comparé les instances d'arguments identifiées manuellement avec celles extraites selon l'approche présentée dans la section 3.1.
Comme le montrent les valeurs présentées dans le tableau 1, nous obtenons des taux de précision relativement intéressants. Les occurrences incorrectement identifiées correspondent principalement à des problèmes d'identification des unités. Par exemple, pour exprimer l'unité de *02_Permeability* les auteurs emploient des notations variées telles que $cm3.\mu m/cm2.d.kPa$, $cm3.\mu m/m2.d.kPa$, $cm3.\mu$*m.m-2.d-1.kPa-1*, $cm3.\mu m/m2.d.kPa$, $cm3.um/m2.d.kPa$, $cm3/\mu m/cm2/d/kPa$, etc. L'ensemble des patrons lexico-syntaxiques doit pouvoir prendre en compte tous ces cas.
Les taux de rappel plus bas s'expliquent par le fait que la RTO que nous avons utilisée est incomplète. Nous avons en effet constaté que certains types d'emballages et cer-

taines unités permettant de mesurer la température, l'humidité relative et l'épaisseur ne sont actuellement pas représentés dans la RTO.

argument	rappel	précision
packaging	0.64	0.81
temperature	0.55	0.45
thickness	0.51	0.81
relative_humidity	0,69	0.82
02_Permeability	0.88	0.85

Tableau 1. *Rappel et précision pour l'extraction des instances d'arguments*

4.2. *Identification de l'argument pivot*

L'argument pivot de la relation *O2_permeability_Relation* est *Packaging*. Il est un bon candidat pour être argument pivot car il apparait très souvent dans le contexte de l'argument *02_permeability*. Le pourcentage d'instances de l'argument *packaging* cooccurant avec une instance de 02_permeability est en effet de 83% alors que pour les autres arguments candidats *temperature*, *thickness* et *relative_humidity*, les pourcentages sont respectivement de 61%, 58% et 43%.

4.3. *Identification des instances de la relation n-aire*

Nous avons finalement évalué la capacité de notre approche à détecter les instances de relation en comparant celles extraites à partir de l'approche présentée dans la section 3.3 et celles identifiées manuellement. Lorsque l'on considère l'ensemble des instances de relations extraites, elles sont toutes au moins partiellement correctes. Une instance de relation est dite partiellement correcte si au moins 3 des 5 arguments appartenant à la signature de la relation le sont effectivement. Le taux de rappel n'est pas très élevé (0,6 à la fois pour les instances de la relation n-aire *O2_permeability_Relation* exactes et partiellement correctes) car les problèmes soulevés dans la section 4.1 nous empêchent d'extraire toutes les instances d'arguments. Cependant, il est important de noter que le taux de précision est intéressant : 0,7 (resp. 1) pour les instances de la relation n-aire *O2_permeability_Relation* exactes (resp. partiellement correctes). Nous pensons qu'enrichir la RTO nous permettra de mieux extraire les arguments et d'obtenir un meilleur rappel lors de l'extraction des instances de relation.

5. Conclusion

Nous avons proposé dans cet article une méthode d'identification d'instances de relations n-aires formalisées au sein d'une Ressource Termino-Ontologique (RTO). Cette proposition ouvre des perspectives de recherche originales dans la mesure où elle s'applique aux relations n-aires formulées de façon implicite dans le texte et dont les instances d'arguments peuvent être exprimées à travers différentes phrases du texte.

A notre connaissance, cette proposition est une contribution originale à l'état de l'art. Les premiers résultats expérimentaux, même s'ils ont été réalisés sur un petit corpus de documents, sont encourageants. Avant de réaliser des études sur un corpus plus important, de nouveaux patrons seront implémentés notamment pour permettre une meilleure extraction des instances d'arguments dans lesquelles apparaissent des intervalles de valeurs. Par ailleurs, nous avons constaté que les termes dénotant certains concepts comme les emballages ou les unités de mesure peuvent être présents dans le corpus sous différentes formes. Nous envisageons donc d'enrichir la RTO en utilisant des méthodes d'extraction des termes variants (Daille *et al.,* 1996) afin d'améliorer l'extraction des instances d'arguments.

6. Bibliographie

Aussenac N., Seguela P., « Les relations sémantiques : du linguistique au formel », *Cahiers de grammaire, Numéro spécial sur la linguistique de corpus*, vol. 25, p. 175-198, 2000.

Bessières P., Nazarenko A., Nédellec C., « Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques », *Proceedings of CIDE*, p. 12-20, 2001.

Buche P., Dibie-Barthelemy J., Ibanescu L., Soler L., « Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource », *IEEE TKDE*, in press.

Burcu B., Osman U. S., « Functional Classification of G-Protein Coupled Receptors, Based on Their Specific Ligand Coupling Patterns », *Proceedings of EvoWorkshop*, p. 5-11, 2006.

Charles F., J. S., Gontard N., « Active Modified Atmosphere Packaging of Fresh Fruits and Vegetables : Modeling with Tomatoes and Oxygen Absorber », *Journal of Food Science : Food Engineering and Physical Properties*, vol. 68, n° 5, p. 1736-1742, 2003.

Daille B., Habert B., Jacquemin C., Royauté J., « Empirical observation of term variations and principles for their description », *Terminology*, vol. 3, n° 2, p. 197-258, 1996.

Djioua B., Garcia-Flores J., Blais A., Desclés J.-P., Guibert G., Jackiewicz A., Priol F. L., Nait-Baha L., Sauzay B., « EXCOM : an automatic annotation engine for semantic information », *Proceedings of FLAIRS 2006*, p. 33-40, 2006.

Dubois D., Prade H., « The three semantics of fuzzy sets », *Fuzzy Sets and Systems*, vol. 90, p. 141-150, 1997.

Giuliano C., Lavelli A., Romano L., « Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature », *Proceedings of ECAI*, p. 7-15, 2006.

Khelif K., Web sémantique et mémoire d'expériences pour l'analyse du transcriptome, PhD thesis, 2006.

Pustejovsky J., Castaño J., Zhang J., « Robust Relationnal Parsing over Biomedical Literature : Extracting Inhibit Relations », *Proceedings of PSB*, p. 362-373, 2002.

Sager J. C., Dungworth D., MacDonald P. F., *English Special Languages*, Wiesbaden, Brandstetter Verlag, 1980.

Touhami R., Buche P., Dibie-Barthélemy J., Ibanescu L., « An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables », *International Conference ODBASE, OTM Workshops 2011*, vol. 7045, LNCS series, p. 662-679, 2011.

Zadeh L., « Fuzzy sets as a basis for a theory of possibility », *Fuzzy Sets and Systems*, vol. 1, p. 3-28, 1978.