Enrichissement d'une RTO par l'ajout de termes spécialisés

Soumia Lilia Berrahou*,**, Ludovic Lebras*,**
Patrice Buche*,** Juliette Dibie-Barthélemy*** Mathieu Roche**

*INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
Soumialilia.Berrahou@lirmm.fr, Patrice.Buche@supagro.inra.fr
**LIRMM, CNRS, Université Montpellier 2, France
ludovic.lebras@gmail.com, Mathieu.Roche@lirmm.fr
***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'enrichissement d'une Ressource Termino-Ontologique (RTO) par l'ajout de termes extraits d'un corpus de documents textuels constitué dans le domaine d'application décrit par la RTO. 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. Nous nous intéressons à l'enrichissement d'une RTO permettant de modéliser des relations n-aires entre des données quantitatives expérimentales, où les arguments peuvent être des concepts symboliques ou des quantités caractérisées par des unités de mesure. La méthode proposée consiste à enrichir la terminologie associée aux concepts symboliques et la terminologie associée aux unités de mesure par de nouveaux termes extraits du corpus.

1 Introduction

Le travail présenté dans cet article se situe dans le cadre de l'identification de termes candidats spécialisés à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). La notion de RTO provient des travaux de (Reymonet et al., 2007), (Roche et al., 2009), (McCrae et al., 2011) et de (Cimiano et al., 2011) qui ont proposé d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la manifestation linguistique (le terme) et la notion qu'elle dénote (le concept). Nous nous intéressons à l'enrichissement d'une RTO permettant de modéliser des relations n-aires entre des données quantitatives expérimentales (Touhami et al., 2011), où les arguments peuvent être des concepts symboliques ou des quantités caractérisées par des unités de mesure. Etant donné un domaine d'application représenté par une RTO, nous proposons une méthode générique d'enrichissement de la partie terminologique de cette RTO par l'extraction de termes, permettant de dénoter des concepts symboliques ou des unités de mesure, à partir d'un corpus de données textuelles dans le domaine d'application étudié, l'extraction étant guidée par cette RTO. La méthode proposée permet plus précisément d'identifier des termes variants candidats dans un corpus de documents textuels pour enrichir la terminologie associée aux concepts

symboliques et aux unités de mesure définis dans une RTO. Cette méthode peut être appliquée à n'importe quelle RTO du moment qu'elle contient des concepts symboliques et des unités de mesure, qui ont été définies en s'appuyant sur le système international des unités de mesure (SI) (Thompson et Taylor, 2008). De telles RTO existent dans de nombreux domaines scientifiques où se côtoient fréquemment (1) des concepts symboliques, qui permettent de représenter des données non numériques, comme par exemple les objets d'étude dans le cas des données quantitatives expérimentales, et, (2) des concepts quantités caractérisées par des unités de mesure. Le corpus utilisé pour tester notre méthode d'identification de nouveaux termes candidats est extrait du domaine de la microbiologie prévisionnelle.

Nous présenterons, dans la section 2, la Ressource Termino-Ontologique (RTO) sur laquelle repose notre travail. Notre méthode d'identification de termes candidats extraits de texte pour enrichir la RTO sera présentée dans la section 3, avec en premier l'identification de termes pour enrichir la terminologie des concepts symboliques, puis l'identification de nouvelles unités de mesure. La section 4 présentera des premiers résultats expérimentaux. Enfin, nous conclurons et présenterons les perspectives de notre travail dans la section 5.

2 La ressource termino-ontologique

Dans cet article, nous étudions l'enrichissement d'une Ressource Termino-Ontologique (RTO) permettant de représenter des relations n-aires entre des données quantitatives expérimentales telle que définie dans (Touhami et al., 2011). Cette RTO, appelée dans la suite RTO naRyQ (n-ary Relations between Quantitative experimental data), joue un rôle fondamental au sein du système d'intégration de données ONDINE (ONtology-based Data INtEgration) (Buche et al., 2013). Ce système propose un processus complet de capitalisation et de modélisation de données et de connaissances s'appuyant sur une RTO qui permet d'enrichir des bases locales à partir de données extraites de documents scientifique. Il permet plus précisément d'acquérir, d'annoter et d'interroger des données extraites de tableaux trouvés dans des documents scientifiques (e.g. articles dans des revues, rapports), qui contiennent en général une synthèse des données quantitatives expérimentales publiées dans ces documents, ceci afin de pouvoir les exploiter et les traiter conjointement avec des données locales.

La Ressource Termino-Ontologique (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). Nous rappelons brièvement, pour les besoins de l'article, la composante conceptuelle, puis la composante terminologique de la RTO naRyQ.

2.1 La composante conceptuelle

La composante conceptuelle de la RTO naRyQ est composée de deux parties : une *ontologie noyau* qui permet de représenter des relations n-aires entre des données quantitatives expérimentales et une *ontologie de domaine* qui permet de représenter les concepts spécifiques à un domaine donné. La figure 1 présente un extrait d'une RTO naRyQ définie dans le do-

maine du risque alimentaire microbiologique étendu aux emballages [1], appelée dans la suite naRyQ_emb pour plus de simplicité.

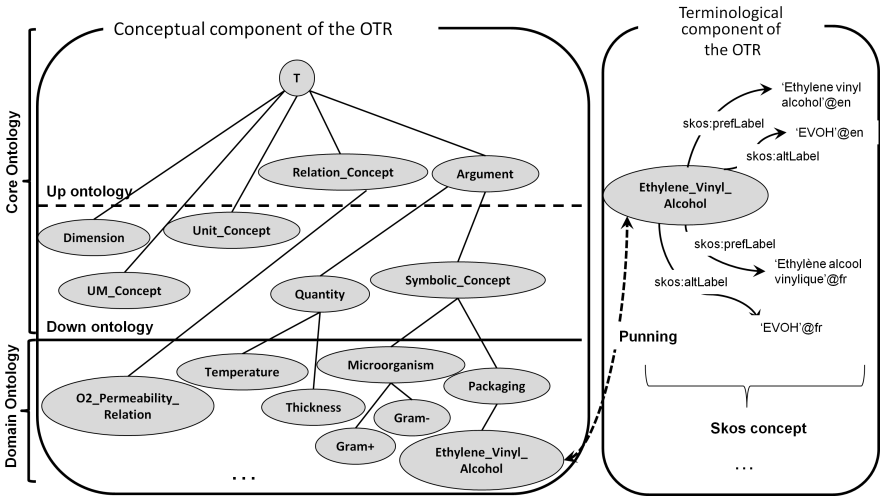

FIG. 1 – *Un extrait de la RTO naRyQ_emb dans le domaine du risque alimentaire microbiologique étendu aux emballages*

L'ontologie noyau est décomposée en 2 sous-parties (cf. figure 1) : une partie supérieure, appelée *ontologie noyau supérieure*, qui permet de représenter des relations n-aires entre n'importe quels arguments, et une partie inférieure, appelée *ontologie noyau inférieure*, qui permet de représenter des relations n-aires entre des données expérimentales quantitatives. Dans l'ontologie noyau supérieure, les concepts génériques *Relation_Concept* et *Argument* permettent de représenter respectivement les relations n-aires et leurs arguments. Dans l'ontologie noyau inférieure, les concepts génériques *Dimension*, *UM_Concept*, *Unit_Concept* et *Quantity* permettent de gérer les quantités et leurs unités de mesure. Le concept générique *Symbolic_Concept* permet, quant à lui, de représenter les autres arguments (i.e. les arguments non numériques) des relations n-aires entre des données quantitatives expérimentales. Ces autres arguments permettent de représenter les objets d'étude (e.g. produits alimentaires, microorganismes, emballages) et les données exprimées de manière qualitative (e.g. croissance/non croissance/mort d'un microorganisme).

L'ontologie de domaine contient les concepts spécifiques à un domaine d'application particulier. Ils apparaissent dans la RTO naRyQ comme des sous-concepts des concepts génériques de l'ontologie noyau. En OWL, tous les concepts sont représentés par des classes OWL, qui sont organisées hiérarchiquement à l'aide de la relation de subsomption *subClassOf* et sont deux à deux disjointes.

Nous présentons ci-dessous plus en détail les deux concepts qui nous intéressent pour l'enrichissement de la RTO naRyQ à savoir le concept générique *Symbolic_Concept* permettant

1. L'impact des emballages et des transferts de gaz est pris en compte sur la croissance des microorganismes dans la matrice alimentaire.

de représenter les concepts symboliques et le concept générique *Unit_Concept* permettant de représenter les unités de mesure.

2.1.1 Le concept générique *Symbolic_Concept* :

Un **concept symbolique**, sous-concept du concept générique *Symbolic_Concept*, est caractérisé par son label, défini dans la composante terminologique de la RTO naRyQ.

Exemple : La figure 2 présente un extrait de la version actuelle de la hiérarchie des concepts symboliques de la RTO naRyQ_emb.

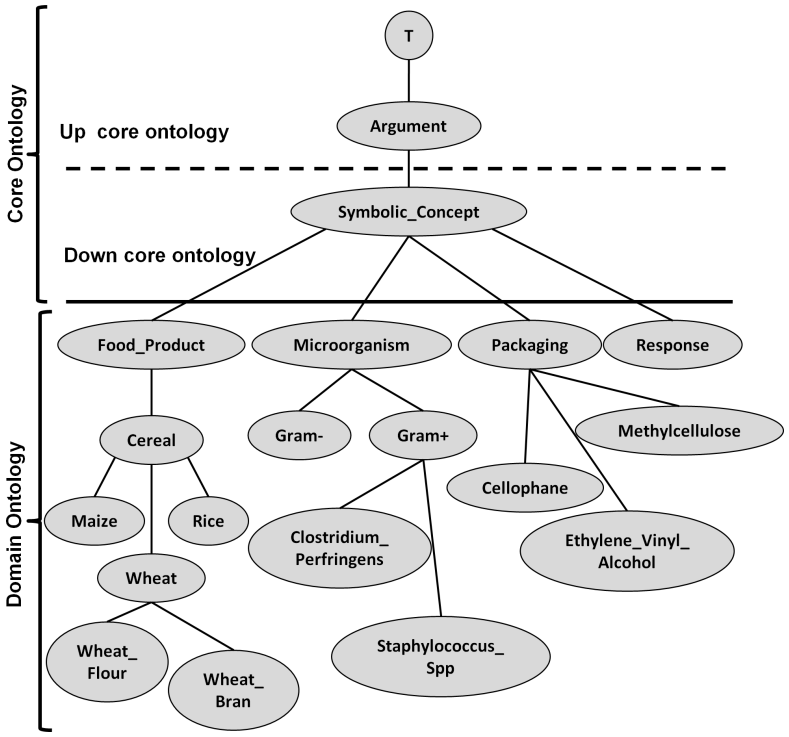

FIG. 2 – *Un extrait de la hiérarchie des concepts symboliques de la RTO naRyQ_emb*

2.1.2 Le concept générique *Unit_concept* :

Les instances des quatre sous-concepts *Singular_Unit*, *Unit_Division_Or_Multiplication*, *Unit_Multiple_Or_SubMultiple* et *Unit_Exponentiation* du concept générique *Unit_Concept* permettent de représenter des unités de mesure. Un concept unité est caractérisé par son label, défini dans la composante terminologique de la RTO naRyQ, une dimension, instance du concept générique *Dimension*, et éventuellement des conversions.

Notre classification des unités de mesure repose sur le Système Internationnal des Unités [2]. Pour définir nos concepts unités, nous nous sommes inspirés de la modélisation des unités de mesure définies dans des ontologies existantes (OM [3], OBOE [4], QUDT [5], QUOMOS, ...). Nous en avons également définies de nouveaux pour les besoins de notre domaine d'application. Nous avons par exemple défini les concepts unités ppm [6] et CFU\per\gram [7].

Exemple : La figure 3 présente un extrait de la hiérarchie des concepts unités avec leurs instances dans la RTO naRyQ_emb.

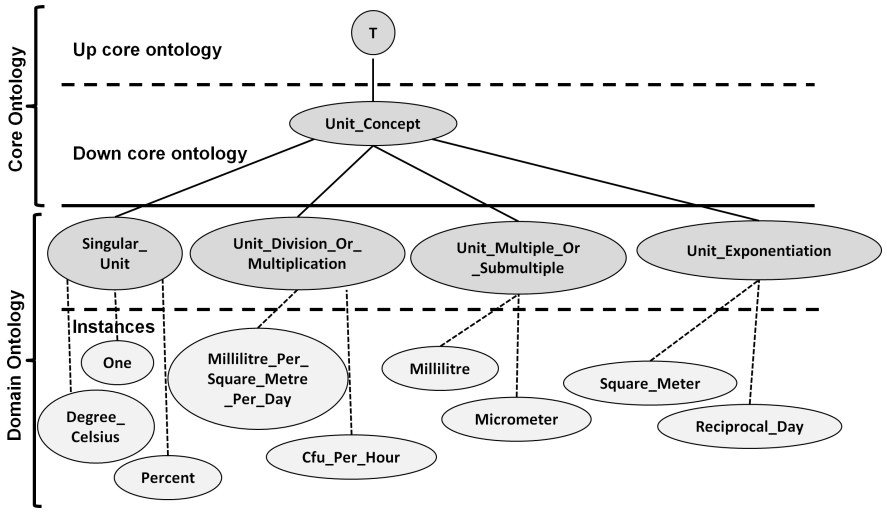

FIG. 3 – *Un extrait de la hiérarchie des concepts unités avec leurs instances dans la RTO naRyQ_emb*

2.2 La composante terminologique

La composante terminologique de la RTO naRyQ contient l'ensemble des termes du domaine étudié. Comme nous l'avons précisé plus haut, les sous-concepts du concept générique *Symbolic_Concept* ainsi que les instances du concept générique *Unit_Concept* sont chacun dénotés par au moins un terme de la composante terminologique. Chacun de ces sous-concepts ou instances est ainsi, dans une langue donnée, dénoté par un label préféré et éventuellement par un ensemble de labels alternatifs, qui correspondent à des synonymes ou des abréviations. Les

2. http ://www.bipm.org/en/si/
3. http ://www.wurvoc.org/vocabularies/om-1.8/
4. http ://marinemetadata.org/references/oboeontology
5. http ://www.qudt.org/
6. ppm, parts per million, est une unité de concentration souvent utilisée pour mesurer le niveau de polluants dans l'air, l'eau, les corps fluides, etc.
7. CFU\per\gram, Colony-Forming Units per gram, est utilisé pour mesurer le nombre de bactéries ou de fongiques viables en microbiologie.

labels sont associés à un concept ou une instance grâce aux propriétés SKOS de labellisation [8] (Simple Knowledge Organization Scheme), recommandées par le W3C. Par exemple, dans la figure 1, les termes anglais *Ethylene vinyl alcohol* et *EVOH* dénotent le concept symbolique *Ethylene_Vinyl_Alcohol*.

Dans la suite, l'enrichissement d'une RTO, telle que la RTO naRyQ décrite ci-dessus, consiste à trouver de nouveaux labels alternatifs, que nous appellerons termes, à associer aux concepts symboliques et aux unités de mesure définis dans la RTO.

3 Identification de termes candidats dans les textes

L'identification de termes candidats pour enrichir la terminologie associée à des concepts symboliques est effectuée en deux étapes : une étape d'extraction et une étape de filtrage, décrites ci-dessous.

3.1 Identification de termes candidats à associer aux concepts symboliques

3.1.1 Extraction des termes.

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Nous ne traiterons pas ici les approches d'aide à la structuration et au regroupement conceptuel des termes qui sont détaillés dans les travaux de (Bourigault et al., 2004). Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques et/ou linguistiques. Le système TERMINO de (David et Plante, 1990) est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie d'une analyse fondée sur une grammaire. Les travaux de (Smadja, 1993) (XTRACT) s'appuient sur une méthode statistique. XTRACT extrait, dans un premier temps, les candidats binaires (composés de deux mots) situées dans une fenêtre de dix mots. Les candidats sélectionnés sont ceux qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les candidats plus généraux (composés de plus de deux mots) contenant les candidats binaires trouvés à la précédente étape. ACABIT de (Daille et al., 1998) effectue une analyse linguistique afin de transformer les groupes nominaux en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Contrairement à ACABIT qui est fondé sur une méthode statistique, LEXTER (Bourigault, 1993) et SYNTEX (Bourigault et Fabre, 2000) s'appuient essentiellement sur une analyse syntaxique afin d'extraire (après décomposition des candidats en têtes et expansions) la terminologie du domaine.

D'autres systèmes plus adaptés à notre problématique tel que FASTR (FAst Syntactic Term Recognizer) (Jacquemin, 1999) permettent d'extraire des termes de base (en utilisant des règles linguistiques) mais également des termes variants. Ce système reste aujourd'hui largement utilisé dans des plateformes de fouille de textes (Lux-Pogodalla et al., 2010). Par exemple, afin d'extraire des termes variants, le système FASTR s'appuie, entre autres, sur trois méta-règles :

8. http ://www.w3.org/TR/skos-reference/

– *Insertion* : ajout d'un modifieur (adjectif, adverbe, etc.) ou d'un déterminant au terme de référence. Par exemple, *sterilized fortified milks* est une variante par insertion de *sterilized milks*.
– *Permutation* : permutation entre têtes et expansions (avec l'ajout d'un mot fonctionnel de type préposition). Par exemple, *milk proteins* est une variante par permutation de *proteins of milk*.
– *Coordination* : insertion d'une coordination et d'un mot (tête ou expansion) au terme de référence. Par exemple *fresh and ripened cheeses* est une variante pour *fresh cheeses* et *ripened cheeses*.

Dans nos travaux nous utiliserons le système FASTR de deux manières différentes :

– **Extraction contrôlée** : A partir de termes de références issus d'une RTO, l'extraction de termes variants présents dans un corpus spécialisé sera effectuée. Cette terminologie variante peut aisément être utilisée pour enrichir la partie terminologique de la RTO avec des termes permettant de dénoter un concept symbolique défini dans la RTO.
– **Extraction libre** : Les termes extraits dans les textes sur la base des règles linguistiques de FASTR peuvent être proposés aux experts comme terminologie candidate dans le but d'enrichir une RTO. Dans ce cas, le nombre important de termes candidats retournés par le système nécessite l'utilisation de filtrage statistiques et/ou sémantiques. Ces derniers sont présentés dans les paragraphes suivants.

3.1.2 Filtrage statistique

Pour filtrer et proposer aux experts les termes candidats les plus pertinents issus de l'extraction libre, nous proposons d'appliquer la mesure $TF\text{-}IDF$ qui donne un poids plus important aux termes discriminants (Salton et McGill, 1983). D'autres pondérations, telle que Okapi qui prend en compte la taille des textes, peuvent également être appliquées (Claveau, 2012). En outre, dans ces travaux, nous proposons d'appliquer également un filtrage statistique fondé sur la fréquence des termes dans le corpus traité (TF – *Term Frequency*). Bien que de telles pondérations statistiques soient, en général, pertinentes dans le domaine de la fouille de textes, il semble intéressant de les coupler à un filtrage sémantique que nous décrivons dans le paragraphe suivant.

3.1.3 Combinaison filtrage statistique / filtrage sémantique

Le principal problème lié à l'extraction libre de la terminologie tient au fait que nous obtenons un nombre considérable de termes issus des données textuelles. À titre d'exemple, avec 121 documents scientifiques traités, 61 607 termes candidats ont été extraits. Afin de filtrer ce résultat issu de FASTR, nous proposons dans un premier temps de classer les candidats potentiels à l'enrichissement par ordre décroissant de TF et/ou $TF\text{-}IDF$. Un tel classement permet de retrouver facilement les candidats potentiels les plus pertinents à l'enrichissement de la RTO, les candidats ayant les meilleurs scores étant classés en premier. Mais ceci ne résout pas le problème lié au nombre considérable de termes issus de la phase d'extraction. Nous avons alors appliqué une règle de filtrage sémantique sur les termes candidats, qui consiste à ne conserver que les termes extraits "sémantiquement proches" de termes définis dans la

RTO, c'est-à-dire les termes extraits composés d'un ou plusieurs mots qui sont en commun avec des termes définis dans la RTO. Différentes expérimentations nous ont permis de fixer le filtrage à deux mots en commun afin de restreindre au maximum le nombre de termes candidats retournés. La liste finale des candidats potentiels obtenus par filtrage statistique et par filtrage sémantique est ainsi présentée à l'expert pour être évaluée, afin de décider pour chaque candidat s'il sera ou non utilisé pour enrichir la partie terminologique de la RTO.

3.2 Identification de termes candidats à associer aux unités de mesure

Les unités de mesures peuvent être dénotées de plusieurs manières différentes dans les documents scientifiques et subir de fortes variations terminologiques. Par exemple, l'unité de mesure *amol/(m.s.Pa)* définie dans la RTO naRyQ_emb peut être présente dans les documents scientifiques sous différentes formes (*amol/m.sec.Pa*, *amol/m.s.Pa*, ...). Dans nos travaux, nous proposons d'extraire de telles variations dans les documents afin d'enrichir une RTO. Une telle extraction ne peut pas reposer sur des méthodes d'extraction utilisant des patrons linguistiques "classiques". Nous proposerons dans cet article de nouvelles mesures de similarité entre chaînes de caractères spécifiques à notre problématique.

3.2.1 La distance de Damerau-Levenshtein pour comparer des chaînes de caractères

Contrairement aux variations terminologiques décrites dans la sous-section 3.1 qui reposaient sur des règles linguistiques rigoureuses, les unités de mesures suivent des règles de variations spécifiques, telles que :

- Insertion de caractères spéciaux (parenthèses, points, etc.).
  Par exemple, *amol/m.s.Pa* et *amol/(m.s.Pa)*
- Substitution de caractères.
  Par exemple, *cm3.um/(m2.d.kPa)* et *cm3.μm/(m2.d.kPa)*

Les situations observées ci-dessus nous ont alors encouragé à utiliser, puis adapter des mesures de proximité lexicale afin de mesurer la similarité entre chaînes de caractères. (Cohen et al., 2003) compare différentes approches extraites de la littérature pour résoudre la similarité entre deux chaînes. Plusieurs communautés (statistiques, des bases de données et intelligence artificielle) se sont intéressées à cette problématique. Deux approches se distinguent : (i) les mesures considérant la chaîne de caractères comme une séquence de caractères (Damerau, 1964) (Winkler, 1999) ; (ii) les mesures considérant la chaîne comme un sous-ensemble de mots (ou tokens) (Baets et Meyer, 2005), telles que la similarité de *Jaccard*, pour lesquels la mesure est calculée en tenant compte d'un éventuel désordre et/ou de l'absence de certains tokens. Des mesures hybrides ont également émergé et combinent ces deux premières approches en obtenant des résultats intéressants (Monge et Elkan, 1996). D'autres, enfin, incluent une information complémentaire statistique tel que le $TF\text{-}IDF$ pour obtenir un score de similarité combiné (Cohen et al., 2003).

Dans nos travaux, nous avons tout d'abord appliqué la distance (appelée D_c) de Damerau-Levenshtein (Damerau, 1964) qui repose sur le calcul du coût minimum des opérations pour transformer une chaîne a en une chaîne b :

- Substitution d'un caractère de a en caractère de b.
  Par exemple, *deci* $\rightarrow$ *deca* : substitution de i en a

- Ajout dans b d'un caractère de a.
  Par exemple, *h* $\rightarrow$ *hr* : ajout d'un r pour l'unité de mesure des heures
- Suppression d'un caractère de a.
  Par exemple, *mole* $\rightarrow$ *mol* : suppression du e
- Transposition de 2 caractères de a.
  Par exemple, *litre* $\rightarrow$ *liter* : transposition du r et du e

De plus, la distance D_c peut être normalisée en utilisant l'approche de (Maedche et Staab, 2002) :

$$SM_{D_c}(u1, u2) = max[0; \frac{\min(|u1|, |u2|) - D_c(u1, u2)}{min(|u1|, |u2|)}] \in [0; 1]$$

Exemple : Pour passer de l'unité de mesure *amol/(m.s.Pa)* à *amol/m.sec.Pa*, 2 suppressions (caractères "(" et ")") et 2 ajouts ("e","c") sont nécessaires. La distance D_c entre ces deux unités est donc de 4 et on obtient la distance de similarité suivante :

$$SM_{D_c}(amol/(m.s.Pa), amol/m.sec.Pa) = max[0; \frac{13 - 4}{13}] = 0.69$$

3.2.2 Une nouvelle distance pour comparer des chaînes de caractères

Dans nos travaux, nous avons souhaité définir une granularité plus importante des éléments des chaînes de caractères propres aux unités de mesures. En effet, la distance de base de Damerau-Levenshtein effectue une comparaison caractère par caractère. Une telle granularité semble trop fine dans notre cas, les termes candidats sont restitués en grand nombre du fait de la trop fine granularité de comparaison, ce qui fait considérablement chuter la précision ainsi obtenue. Nous avons alors choisi d'adopter une autre stratégie de comparaison, qui repose sur une comparaison par bloc de caractères, méthode plus appropriée dans le contexte des unités de mesure (par exemple, *Pa*, *cm3*, etc.). Dans ce cas, les séparateurs de ces blocs seront des caractères spéciaux spécifiques aux unités de mesures : */*, *.*, *(*, *)*, etc. Une fois cette segmentation effectuée, le calcul de la distance de Damerau-Levenshtein est effectué sur la base de la comparaison du coût des opérations de transformation d'un ensemble de blocs en un autre ensemble de blocs. Afin d'évaluer la pertinence de notre méthode, nous avons choisi d'attribuer, dans un premier temps, le même coût aux différentes opérations. Nous appellerons cette nouvelle distance fondée sur les blocs, D_b.

3.2.3 Une nouvelle méthode d'identification d'unités de mesure variantes

La nouvelle méthode proposée qui consiste à retrouver les variants d'un terme u dénotant une unité de mesure définie dans une RTO, suit les étapes suivantes :

- Segmenter chaque phrase des documents en mots m_i (chaînes de caractères séparées par des "blancs"),
- Segmenter chaque mot m_i par blocs blm_{ij} (chaînes de caractères séparées par des séparateurs d'unités de mesure),
- Segmenter u par blocs blu_j,

- Si $SM_{D_c}(blm_{ij}, blu_j) \geq K1$ [9], alors considérer les blocs blu_j (de l'unité de mesure) et blm_{ij} (du candidat issu des textes) comme identiques pour la suite des opérations,
- Mesurer $SM_{D_b}(u, m_i)$,
- Sélectionner m_i si $SM_{D_b}(u, m_i) \geq K2$, seuil qui sera discuté dans la section 4.

Ainsi, après application de notre méthode, nous pouvons proposer m_i comme nouveau terme variant de u dans la RTO. Notons que nous pouvons ajouter une étape intermédiaire à notre méthode consistant à effectuer une comparaison "approchée" de chaque bloc en considérant que deux blocs sont identiques si leur mesure de proximité SM_{D_c} est proche de 1 (seuil $K1$).

Exemple : Prenons l'exemple de la comparaison des deux unités *amol/(m.sec.Pa)* et *mol/(m.s.Pa)* et appliquons trois mesures, la mesure de base de Damereau SM_{D_c} décrite précédemment, $SM_{D'_c}$ sans considérer les séparateurs et la mesure SM_{D_b}. Nous obtiendrions respectivement :

$$SM_{D_c}(amol/m.sec.Pa, mol/(m.s.Pa)) = max[0; \frac{13-5}{13}] = 0.61$$

$$SM_{D'_c}(amol/m.sec.Pa, mol/(m.s.Pa)) = max[0; \frac{10-3}{10}] = 0.7$$

$$SM_{D_b}(amol/m.sec.Pa, mol/(m.s.Pa)) = max[0; \frac{4-2}{4}] = 0.5$$

La mesure SM_{D_b} est plus discriminante et permet ainsi de filtrer de manière plus efficace les unités candidates.

4 Expérimentations

4.1 Protocole expérimental

La méthode d'enrichissement de la RTO décrite dans la section 3 a été expérimentée sur un corpus de 121 articles scientifiques de la littérature internationale dans le domaine de la conception d'emballages alimentaires biodégradables. La récupération des documents est gérée au sein du système ONDINE décrit en section 2, qui récupère des articles scientifiques de la base de données FSTA (Food Science and Technology Abstracts) [10] à l'aide de requêtes à base de mots-clés. Les articles collectés au format HTML, ont été pré-traités afin d'éliminer les balises HTML. Le nombre de termes de la RTO utilisés pour les expérimentations est respectivement de 460 termes pour le concept symbolique *Food Product*, 186 pour *Microorganism*, 140 pour *Packaging* et 3 pour *Response*. En ce qui concerne les unités de mesure, le nombre de termes utilisés est respectivement 84 pour le concept *Singular_Unit*, 174 pour *Unit_Division_Or_Multiplication*, 78 pour *Unit_Exponentiation* et 83 pour *Unit_Multiple_Or_Submultiple*. Les résultats ont été évalués en calculant la précision des résultats, c'est-à-dire le nombre de termes jugés pertinents par l'expert pour l'enrichissement de la RTO par rapport à l'ensemble des termes candidats restitués. Dans ce protocole, le rappel ne

9. En posant, $K1 = 1$ les chaînes de caractères propres aux blocs devront être strictement identiques. Les expérimentations seront menées avec cette valeur.

10. http ://www.ifis.org/fsta/

peut pas être évalué car nous ne connaissons pas de manière exhaustive l'ensemble des termes pertinents du corpus.

4.2 Résultats de l'identification de termes candidats à associer aux concepts symboliques

4.2.1 Enrichissement par des termes variants à partir d'une extraction contrôlée

Pour guider l'extraction de termes variants, nous avons découpé les concepts symboliques en quatre grandes classes correspondant aux quatre sous-concepts directs du concept générique *Symbolic_Concept* présentés dans la figure 2. Ce découpage a été réalisé afin de faciliter la tâche d'évaluation des résultats par les experts du domaine. Ainsi un paquet de termes a été associé à chaque concept *Food_Product*, *Microorganism*, *Packaging* et *Response* qui contient les termes dénotant ces concepts et l'ensemble de leurs sous-concepts.

Concept Symbolique	**Nbr de termes variants**	**précision**
Food Product	65	72%
Microorganism	5	0%
Packaging	17	88%
Response	3	100%

TAB. 1 – *Précision évaluée à partir de la terminologie obtenue par extraction contrôlée.*

Étant donné que nous traitons un corpus sur les emballages alimentaires biodégradables, peu de termes sont associés aux concepts symboliques *Microoorganism* et *Response*. La valeur beaucoup plus faible du nombre de termes associés au concept symbolique *Packaging* par rapport à *Food_Product* s'explique par le fait que le concept *Packaging* contient trois fois moins de termes que *Food Product*. Le tableau 1 montre des résultats satisfaisants en terme de pertinence pour ces deux concepts, qui sont davantage représentés dans le corpus de documents scientifiques utilisé.

4.2.2 Enrichissement par des termes variants à partir d'une extraction libre

Les différents résultats de filtrage sémantique à partir des quatre concepts symboliques de base sont présentés dans le tableau 2.

Concept Symbolique	**Nbr de termes extraits et filtrés**	**précision**
Food Product	186	77%
Microorganism	2	0%
Packaging	79	59%
Response	1	0%

TAB. 2 – *Précision évaluée à partir de la terminologie obtenue par extraction libre.*

Les résultats du tableau 2 montrent que les termes extraits et associés aux concepts les plus représentés, à savoir *Packaging* et *Food Product*, sont globalement pertinents (respectivement 59% et 77%). Le contenu du corpus ciblé sur les emballages alimentaires biodégradables explique que l'on retient très peu de candidats pour le concept *Microorganism* et *Response*. Le nombre différent de termes contenus dans la RTO entre *Food Product* et *Packaging* peut expliquer la différence du nombre de candidats pour ces deux concepts.

La mesure statistique $TF\text{-}IDF$ est utilisée comme premier filtre afin d'obtenir une liste suffisamment exhaustive de termes spécialisés pertinents. Afin d'affiner la pertinence à l'enrichissement de la terminologie de la RTO, cette liste est filtrée à nouveau en utilisant un critère sémantique comme décrit en section 3.1, puis fournie à l'expert pour validation. Le filtrage sémantique permet de ne proposer qu'un nombre de termes restreint aux experts (moins de 300 candidats dans notre cas). Notons que nos expérimentations ont montré que les termes classés sur la base de la fréquence (TF) étaient en général mieux positionnés qu'en appliquant une pondération $TF\text{-}IDF$. Ce résultat fondé sur un calcul de taux de couverture (termes extraits également présents dans la RTO) mériterait d'être approfondi à partir d'un jeu de données plus conséquent.

Après avoir évalué notre méthode d'identification de termes variants dans les textes pour enrichir la terminologie associée aux concepts symboliques définis dans une RTO, la section suivante décrit les expérimentations menées pour enrichir la terminologie associée aux unités de mesures.

4.3 Résultats de l'extraction de termes candidats à associer aux unités de mesure

Dans un premier temps, afin d'évaluer les distances de similarité décrites en section 3.2, nous avons effectué quelques expérimentations sur un jeu de données fourni par les experts. Ces expérimentations synthétisées dans le tableau 3 montrent que 81% des termes extraits sont identifiées comme étant des variations pertinentes avec un seuil de similarité ($K2$) de 0.75 avec la mesure SM_{D_b} que nous avons proposée. Ce taux de similarité est ramené à 72% avec des mesures plus classiques (SM_{D_c}). Le tableau 3 montre aussi que les valeurs de SM_{D_c} signifiant une possible similarité sont en général amplifiée avec SM_{D_b}. Ainsi, notre nouvelle mesure de similarité fondée sur la notion de blocs retourne une similarité parfaite (égale à 1) dans 45% des cas.

Dans un deuxième temps, nous avons expérimenté les deux mesures de similarité sur le corpus d'articles scientifiques. Avec notre mesure SM_{D_b}, nous avons pu réaliser une évaluation sur l'intégralité du corpus. En effet, en fixant le seuil de SM_{D_b} à la valeur $K2 = 0.6$, seuls 11 termes candidats sont retournés. Parmi ces 11 termes, 7 ont été jugés pertinents pour ajout dans la RTO. La mesure SM_{D_b}, fondée sur l'analyse de blocs, se révèle bien adaptée à l'enrichissement terminologique pour les unités de mesure complexes qui sont des combinaisons d'unités simples. Elle extrait un petit nombre de termes candidats à évaluer dont plus de la moitié sont pertinents pour ajout dans la RTO. Il est donc tout à fait envisageable d'utiliser cette mesure sur des corpus de documents plus importants.

Avec la mesure de similarité classique SM_{D_c}, l'expérimentation n'a été évaluée que sur 14 articles. En effet, le nombre de termes candidats obtenus est très élevé (564 candidats). Les 100 premiers candidats, triés par similarité décroissante, ont été évalués et seulement 5 ont

Écriture rencontrée	Ecriture attendue	D_c	SM_{D_c}	D_b	SM_{D_b}
10e10 (cm3.m-1.sec-1.Pa-1)	10e10.cm3.m-1.sec-1.Pa-1	3	0.875	0	1
10e-14(cm3/m.s.Pa)	10e-14.cm3/(m.s.Pa)	2	0.894	0	1
10e-16cm3.cm/cm.cm2.s.Pa	(10e-16cm3.cm)/(cm2.s.Pa)	6	0.76	1	0.8
10e18 (mol.m/Pa.sec.m2)	10e18.mol.m/(Pa.sec.m2)	3	0.869	0	1
10e-5cm3/m2.day.Pas	10e-5cm3/(m2.day.Pas)	2	0.90	0	1
amol.m-1.s-1.Pa-1	amol.s-1.m-1.Pa-1	2	0.88	1	0.75
amol/m.s.Pa	amol/(m.s.Pa)	2	0.846	0	1
amol/m.sec.Pa	amol/(m.s.Pa)	4	0.69	1	0.75
ml/m2/day	mL/m2/24h/bar	7	0.22	3	0
cm3.um/m2.d.kPa	cm3.μm/(m2.d.kPa)	4	0.77	1	0.8
mol/m/s/Pa	moles.m.m-2.s-1.Pa-1	13	0.35	4	0.2

TAB. 3 – *Similarités lexicales sur la base de l'utilisation de SM_{D_c} et SM_{D_b}.*

été jugés pertinents pour ajout dans la RTO. Les termes pertinents extraits correspondent à des unités de mesure simples (m^2, cm^2, ...). Ils sont donc différents de ceux obtenus avec la mesure SM_{D_b}. Cette mesure est donc plus difficile à utiliser puisque, d'une part, elle requière l'évaluation d'un grand nombre de termes candidats et d'autre part, sur l'échantillon évalué, seulement 5% des termes candidats ont été jugés pertinents.

5 Conclusion et perspectives

Nous avons proposé dans cet article une méthode d'enrichissement terminologique d'une RTO, et plus particulièrement de ses concepts symboliques et de ses unités de mesure, par l'ajout de termes extraits d'un corpus de documents textuels constitué dans le domaine d'application décrit par la RTO. Les méthodes d'extraction de termes candidats proposées s'appuient et/ou étendent des méthodes à base de règles linguistiques et des mesures de similarités de l'état de l'art en les combinant avec une RTO. L'évaluation expérimentale est encourageante car d'une part, le nombre de termes candidats proposé à l'expert pour évaluation n'est pas très élévé, et d'autre part, la proportion de nombres de termes pertinents obtenus est tout à fait satisfaisante. Dans un avenir proche, nous réaliserons donc des expérimentations sur des corpus de textes plus importants. Une autre perspective de ce travail est de proposer une méthode semi-automatique d'ajout des termes candidats pertinents dans la RTO.

6 Remerciements

Le travail de recherche ayant mené aux résultats présentés dans cet article a reçu le soutien de l'Agence Nationale de la Recherche (ANR) dans le cadre du projet ALIA MAP'OPT et du labex NUMEV.

Références

Baets, B. D. et H. D. Meyer (2005). Transitivity-preserving fuzzification schemes for cardinality-based similarity measures. *European Journal of Operational Research 160*(3), 726 – 740. Decision Analysis and Artificial Intelligence.

Bourigault, D. (1993). Analyse syntaxique locale pour le repèrage de termes complexes dans un texte. *T.A.L. 34*(2), 105–118.

Bourigault, D., N. Aussenac-Gilles, et J. Charlet (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas . *Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies, M. Slodzian (Ed.) 18*(1/2004), 87–110.

Bourigault, D. et C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires 25*, 131–151.

Buche, P., J. Dibie-Barthélemy, L. Ibanescu, et L. Soler (2013). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Trans. Knowl. Data Eng. 25*(4), 805–819.

Cimiano, P., P. Buitelaar, J. McCrae, et M. Sintek (2011). LexInfo : A declarative model for the lexicon-ontology interface. *J. Web Sem. 9*(1), 29–51.

Claveau, V. (2012). Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf (vectorization, okapi and computing similarity for nlp : Say goodbye to tf-idf). In *Proceedings of the Joint Conference JEP-TALN-RECITAL*, pp. 85–98.

Cohen, W. W., P. Ravikumar, et S. E. Fienberg (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, Volume 47.

Daille, B., E. Gaussier, et J. Langé (1998). An evaluation of statistical scores for word association. In *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation : Selected Papers, CSLI Publications*, pp. 177–188.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM 7*(3), 171–176.

David, S. et P. Plante (1990). De la nécessité d'une approche morpho syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Quebec*, Volume 3, pp. 140–154.

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), p. 341–348.*

Lux-Pogodalla, V., D. Besagni, et K. Fort (2010). Fastkwic, an "intelligent" concordancer using fastr. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *EKAW (Knowledge Engineering and Knowledge Management)*, pp. 251–263.

McCrae, J., D. Spohr, et P. Cimiano (2011). Linking lexical resources and ontologies on the se-

mantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I*, ESWC'11, Berlin, Heidelberg, pp. 245–259. Springer-Verlag.

Monge, A. E. et C. Elkan (1996). The field matching problem : Algorithms and applications. In *Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, pp. 267–270.

Reymonet, A., J. Thomas, et N. Aussenac-Gilles (2007). Modelling ontological and terminological resources in OWL DL. In *OntoLex 2007 - Workshop at ISWC07*, Busan, South-Korea.

Roche, C., M. Calberg-Challot, L. Damas, et P. Rouard (2009). Ontoterminology - a new paradigm for terminology. In J. L. G. Dietz (Ed.), *KEOD*, Funchal, Madeira, Portugal, pp. 321–326. INSTICC Press.

Salton, G. et M. McGill (Eds.) (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics 19*(1), 143–177.
English

Thompson, A. et B. N. Taylor (2008). Guide for the use of the international system of units (SI).

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 *International Conference ODBASE, OTM Workshops 2011*, Volume 7045, pp. 662–679. Lecture Notes in Computer Science series.

Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Summary

In this article, we focus on term extraction from scientific documents and present a novel method in order to enrich an Ontological and Terminological Resource (OTR). The OTR is composed of a conceptual part (Ontology) and a terminological part (Terminology), where terms are separated from the concepts they are related to. Scientific documents contain experimental data results that can be represented in N-ary Relations used to model a domain of knowledge in the OTR. An instance of N-ary relation links a studied object (e.g. a packaging) with its features (e.g. thickness, O2 permeability). Object and features are organized in symbolic concepts (e.g. packaging) and quantities (e.g. thickness) characterized with units of measure. We aim at extracting new terms of symbolic concepts and terms of units in order to enrich the OTR.