Exploiter un réseau lexico-sémantique pour la construction d'ontologie

Nadia Bebeshina-Clairet[1, 2] Sylvie Desprès [1]
(1) LIMICS, 74 rue Marcel Cachin, 93017 Bobigny France
(2) LIRMM, 860 rue de St Priest, 34095 Montpellier, France
clairet@lirmm.fr, sylvie.despres@univ-paris13.fr

RÉSUMÉ

Dans le présent article nous nous intéressons à l'exploitation des ressources de connaissance lexico-sémantiques dans le cadre de la construction et de l'enrichissement d'ontologie. En effet, la construction d'ontologie est souvent menée de manière descendante où ses concepts de haut niveau sont définis pour ensuite être spécifiés sur la base d'un consensus entre les experts humains. Nous explorons une technique d'utilisation des réseaux lexico-sémantiques (RLS) multilingue et monolingue pour enrichir une ontologie existante. Cette technique vise à réduire l'effort humain nécessaire à la localisation ou à l'enrichissement d'une ontologie.

ABSTRACT

Lexical Semantic Network Use for Ontology Building

In the present article, we focus on exploiting lexical semantic knowledge resources for ontology building and enhancement. The ontology building process is often a descending and manual process where the high level concepts are defined, then detailed by human experts. We explore a way of using a mutlilingual or monolingual lexical semantic network to make evolve or localize an existing ontology with a limited human effort.

MOTS-CLÉS : ontologie, réseau lexico-sémantique, inférence, ressource multilingue.

KEYWORDS: ontology, lexical semantic network, inference, multilingual resource.

1 Introduction

Les ressources termino-ontologiques actuelles sont des ressources de plus en plus riches en données. La réflexion sur l'acquisition de ces données est activement menée. Notamment, le projet NeOn [1], a été mis en œuvre pour définir un cadre méthodologique de construction d'ontologie permettant d'intégrer les connaissances variées dans le processus de construction, de construire des ontologies modulaires et de mettre les ontologies en réseau. Ce cadre méthodologique comprend les phases d'acquisition et de modélisation. Plusieurs scénarios y ont été définis dont, en particulier, *spécification → implémentation* (développement sans réutilisation des ressources existantes), *réutilisation des ressources non ontologiques et ontologiques* (leur intégration dans l'ontologie), fusion des ressources ontologiques, localisation etc.

Dans le cadre de notre approche, nous nous intéressons à l'exploitation des réseaux lexico-sémantiques

1. http://neon-project.org/nw/About_NeOn.html

(RLS) pour enrichir une ontologie existante ou accompagner la construction d'une ontologie. La motivation de notre démarche est d'explorer l'intérêt et les limites d'exploitation d'une ressource lexicale et sémantique pour la construction d'ontologie.

2 État de l'Art

La construction ontologique outillée par le TAL à partir des textes en langue naturelle est explorée depuis plus de 20 ans. Parmi les outils de recherche des candidats termes se sont distingués SYNTEX-UPERY (Bourigault, 2002), YATEA (Aubin & Hamon, 2006), BIOTEX (Lossio-Ventura *et al.*, 2014). Des architectures de développement des ontologies telles que *Terminae*, (Szulman, 2012) utilisent ce type de outils. L'outil Archonte (Charlet *et al.*, 2006) qui distingue les étapes de normalisation, formalisation s'inscrit également dans ce type de démarche ascendante. De même, parmi les outils proposés, se distinguent :

— les outils différenciés (outils qui prennent en compte la différence entre le terme et le concept d'ontologie). Dans le cadre de ce type d'outils, les unités terminologiques extraites des textes et organisées sous forme de réseau via un ensemble de relations hiérarchiques (hyperonymie) et d'équivalence (synonymie) servent à guider l'ontologue dans la construction ontologique. Ainsi, la construction d'ontologie passe par une structure intermédiaire, une termino-ontologie (ex. : *Terminae*, (Szulman, 2012)) ;

— les outils non différenciés qui n'introduisent pas de distinction entre terme et concept et se basent sur des mesures statistiques (fréquences, *tf-idf*) pour proposer des candidats-concepts soit par analyse des concepts formels (ACF, ex. (Mondary, 2011)) soit par des méthodes basées sur la connaissance (ex. *TextToOnto* [2]).

Comme remarqué par (Mondary, 2011), TextToOnto n'est adossé à aucune méthode de construction ontologique. De plus, les mesures statistiques prises comme appui tendent à favoriser le phénomène dit *long tail* où les mots peu fréquents ont peu d'impact sur les concepts pouvant être calculés par ces méthode (notamment les *hapax* [3]).

Dans le cadre des approches historiques de construction ontologique basées sur le corpus, le terme *élément remarquable* a été introduit. Ce terme désigne à la fois les termes, expressions fréquentes et apparaissant dans un corpus donné et les connaissances tacites contenues dans les textes. Ces connaissances tacites sont principalement les relations sémantiques (notamment les relations de subsomption et des relations spécialisées) dont les indices peuvent apparaître dans un corpus donné. Quant aux éléments explicites, ils peuvent indiquer la présence d'un concept. L'inconvénient de ce type de définition d'*élément remarquable* est qu'elle repose sur les indices répertoriés pour une langue donnée notamment lorsqu'il s'agit d'utiliser des ensembles de patrons lexico-syntaxiques pour leur détection. Les critères quantitatifs sont privilégiés et il est difficile de qualifier ces éléments du point de vue sémantique de manière simple et portable entre les langues.

2. https://sourceforge.net/p/texttoonto/wiki/Home/

3. Fait de langue (mot, expression, construction) dont il n'existe qu'une seule occurrence dans un corpus donné.(Larousse, https://www.larousse.fr/dictionnaires/francais/hapax/39017.)

3 Ressources

RezoJDM (Lafourcade, 2007) est un réseau lexico-sémantique de connaissance générale construit par peuplonomie via le jeu JeuxDeMots [4] et les jeux annexes [5] depuis 2007. Cette ressource est un graphe orienté, typé et pondéré. À ce jour, RezoJDM contient 2,7 millions de termes représentés sous forme de nœuds du graphe et 240 millions de relations (arcs).

RLSM$_{\mathbf{PI}}$ est un réseau lexico-sémantique multilingue (français, anglais, russe, espagnol) avec pivot interlingue construit pour les domaines de la cuisine et de la nutrition. La conception de ce réseau s'est inspirée de la structure de RezoJDM en ce qui concerne l'utilisation de la structure de graphe Pour éviter les difficultés inhérentes à la construction d'un pivot artificiel dont notamment la nécessité aligner *N* sens simultanément, le pivot interlingue du RLSM$_{\text{PI}}$ est amorcé à la manière d'un pivot naturel (en utilisant l'édition anglais de DBNary (Sérasset, 2014) comme ensemble de données d'amorçage). Le pivot évolue ensuite vers un pivot interlingue de façon incrémentale. Les termes de ce réseau sont reliés uniquement via le pivot. A l'heure où nous écrivons, RLSM$_{\text{PI}}$ compte 821 781 termes et 2 231 197 relations.

MIAM (Després, 2016) [6] est une termino-ontologie modulaire pour l'univers de la cuisine numérique qui permet de fournir la connaissance nécessaire à l'élaboration de suggestions nutritionnelles généralistes. Le modèle de connaissance de l'ontologie regroupe les connaissances expertes sur les aliments, les transformations, les actions culinaires, les plats représentatifs de la tradition culinaire française, les recettes utilisées pour réaliser ces plats. Ces connaissances sont représentées sous forme de modules tels que Aliment, Sensoriel, Préparation, etc. Il s'agit d'une ontologie qui reflète dans une certaine mesure la tradition gastronomique française. Se fixer l'objectif de l'enrichir revient à exploiter le modèle pour d'autres langues et possiblement récupérer les éléments remarquables candidats qui reflètent les spécificités (par exemple, des recettes prototypiques) qui caractérisent les autres cultures.

4 Méthode proposée : immersion → projection

4.1 Synthèse

Notre méthode est construite autour de l'idée de projection d'un modèle donné sur un RLS multilingue ou monolingue afin d'en extraire une ressource médiatrice destinée à être utilisée par un expert humain pour la consultation ou la validation des éléments proposées par le système. Notre méthode diffère des méthodes traditionnellement utilisées pour la construction d'ontologie par sa définition de l' ***élément remarquable*** (ER) et par l'utilisation d'une ressource non ontologique pour accompagner la construction d'ontologie.

Nous proposons la définition suivante de l'élément remarquable : « Un ***élément remarquable*** est un terme, une relation ou une structure sémantique qualifiée et qualifiante. »

— *qualifié* se réfère à une possibilité de décrire cet élément de manière discrète (ex. énumération des relations sortantes typées). S'il s'agit d'un terme, il doit posséder un degré entrant

4. http://www.jeuxdemots.org
5. http://imaginat.name/JDM/Page_Liens_JDMv2.html
6. http://www-limics.smbh.univ-paris13.fr/ontoMIAM/

important (avoir un rôle conceptuel). En effet, si le terme d'un RLS a un degré entrant important, il participe à la définition et à la catégorisation d'un grand nombre de termes. S'il s'agit d'une relation, elle doit être contextualisée notamment via un mécanisme d'annotation [7]. S'il s'agit d'une structure [8], elle doit être repérée un nombre de fois suffisant dans le réseau. Le seuil est défini empiriquement et correspond à 3 occurrences d'une structure minimum.

— *qualifiant* se réfère à la possibilité d'utiliser l'élément remarquable dans le cadre d'inférence endogène. S'il s'agit d'un terme, il doit avoir dans son voisinage des hyperonymes, des hyponymes et/ou des synonymes. Il doit également être aligné avec les termes dans les autres partitions d'une ressource multilingue ainsi qu'au niveau interlingue. S'il s'agit d'une relation, elle doit être non unique (il doit exister d'autres relations du même type réelles ou inférables dans le réseau). S'il s'agit d'une structure, ses termes et relations doivent être qualifiants.

Dans le présent article, nous détaillons

1. les expériences qui ont été menées sur la base d'un $RLSM_{PI}$ pour proposer les éléments remarquables candidats de type « classe »et « propriété » ;
2. les expériences qui ont concerné l'enrichissement d'une ébauche d'ontologie à partir du réseau lexico-sémantique du français RezoJDM (Lafourcade, 2007).

Ces expériences s'appuient sur la connaissance lexicale. Par conséquent les éléments proposés automatiquement par notre système le sont sans prétention à la validité ontologique. C'est à l'expert de valider ou invalider ces propositions.

4.2 Immersion

La projection d'un modèle ontologique sur un RLS commence par l'immersion de ce modèle. Le mécanisme de l'immersion s'appuie sur un ensemble de règles de mise en correspondance définies manuellement lors de nos expérimentations. Leur génération (semi)automatique peut être possible notamment pour les ontologies qui utilisent des vocabulaires entièrement standard [9] ce qui n'est pas le cas de l'ontologie MIAM pour laquelle un vocabulaire spécifique a été défini. L'algorithme d'immersion prend en entrée l'ontologie de référence et l'ensemble de règles et fournit à la sortie l'action d'inférer des termes et des relations dans le RLS.

Les règles de mise en correspondance mobilisent les notions de *classe d'ontologie* [10] et de *terme de réseau lexico-sémantique* [11]. Sous leur forme générale, elles stipulent : « Si *x* et *y* sont respectivement domaine et co-domaine d'une propriété à valeur objet *p* de l'ontologie de référence et *y* est sous-classe de *C*, alors *x* a une relation *R* avec *y* et *y* a une relation *is-a* avec C dans le réseau lexico-sémantique d'immersion. »

Pour l'expérience multilingue, cette règle peut être formulée car

1. pour chacune des 93 propriétés d'ontologie considérées, nous avons déterminé quel type de relation sémantique (ou, dans certains cas, quel ensemble de types) correspond à une

7. Ajout d'une méta-information à la relation du réseau.
8. Telle que chemin, sous-graphe etc.
9. C'est-à-dire, les ontologies qui utilisent des vocabulaires pré-existants tels que RDFS, FOAF, SKOS et dont la sémantique est accessible dans un format qui peut être lu par une machine sans ajustement spécifique.
10. **classe :** ensemble d'individus ayant les mêmes caractéristiques
11. **terme :** *((Terminologie* [12]*)* Désignation verbale d'un concept général dans un domaine spécifique.*(RLS)* Item lexical, (vocable, expression polylexicale). Le sens des termes est le sens lexical tel qu'il est observé au niveau des usages dans une langue donné.

propriété d'ontologie donnée. *Object Property* `aPourProduitInitial` correspond ainsi à la méronymie à la fois *partie-tout* et *substance* ;

2. nous avons préalablement mis en correspondance les étiquettes de l'ontologie à immerger et les termes du RLS par coïncidence (3 930 termes ; ex. : « poulet basquaise »,) ou par composition (4 135 termes, ex : « unité mesure capacité », terme utilisé pour la dénotation formelle du concept d'ontologie et ne faisant pas partie des expressions polylexicales courantes de la langue française).

Pour l'expérience monolingue, 115 descripteurs ont été automatiquement exprimés en français à partir des intitulés de leurs URI. Tous les termes à l'exception d'un étaient déjà présents dans le RLS utilisé pour cette expérience, RezoJDM. Le type de relation exploité a été r_carac (caractéristique typique). Cette relation a été annotée à partir des URI des propriétés : `aPourDescripteurBruit` donne, par exemple, *croûte* $\xrightarrow{r\_carac::bruit}$ *croustillant* (« *croûte* a comme caractéristique typique liée au bruit *croustillant* »).

Le prémisses des règles de mise en correspondance reposent sur la contextualisation des relations du RLS. Cette dernière est possible à travers les ensembles d'hyperonymes et des relations sémantiques (voisinage des termes source et cible de la relation) ainsi que les méta-informations attachées à la relation (poids, annotation). Par exemple : *pétrir* $\xrightarrow{r\_object}$ *pâte* $\wedge$ *pétrir* $\xrightarrow{r\_isa}$ *technique de base* $\wedge$ *pâte* $\xrightarrow{r\_isa}$ préparation).

Lors de l'immersion dans un RLS, les étiquettes d'ontologie deviennent des termes RLS polysémiques. Ainsi, une fois immergée dans un RLS, l'étiquette *poulet* (utilisée pour la dénotation formelle dans le cadre de l'ontologie) acquiert plusieurs sens dont *poulet>viande*, *poulet>jeune coq*, *poulet>policier*.

4.3 Projection

4.3.1 Inférence dans le contexte d'un RLS

Dans le contexte multilingue du RLSM_{PI}, nous avons mis en place les algorithmes de découverte par inférence des éléments remarquables (ER) de type « classe/individu » et ceux de type « propriétés d'ontologie ». Lorsqu'il s'agit de découvrir des éléments de type « classe/individu », nous comparons les termes voisins dans une chaîne hiérarchique qui remonte vers un terme RLS correspondant à un concept MIAM de haut niveau. Pour les ER de type « propriété », l'un des éléments à comparer doit correspondre à une structure (souvent, relation annotée) qui représente une propriété de l'ontologie MIAM immergée au sein du RLSM_{PI}. Le schéma d'inférence a été celui par abduction. En présence de deux termes similaires. Dans le cadre de ce schéma, il s'agit de sélectionner un ensemble de termes similaires et à une terme T et de proposer les relations détenues par les termes similaires à T.

4.3.2 Découverte des ER de type « classe/individu »

Dans le cadre de l'ontologie MIAM, l'**axiome général** est de forme suivante :

```
<rdf:Description>
<rdf:type rdf:resource="&owl;AllDisjointClasses"/>
<owl:members rdf:parseType="Collection">
<rdf:Description rdf:about="&aliment;AB"/>
```

```
<rdf:Description rdf:about="&aliment;AppelationOrigine"/>
<rdf:Description rdf:about="&aliment;IGP"/>
<rdf:Description rdf:about="&aliment;LabelRegional"/>
<rdf:Description rdf:about="&aliment;LabelRouge"/>
<rdf:Description rdf:about="&aliment;STG"/>
<rdf:Description rdf:about="&aliment;SpecialiteOrigine"/>
</owl:members>
</rdf:Description>
```

Les axiomes concernent la disjonction entre les classes de MIAM qui garantit la consistance de l'ontologie. Afin de « traduire »les axiomes généraux de MIAM en termes de $RLSM_{PI}$, nous avons considéré les étiquettes des classes listées dans les axiomes afin de pouvoir identifier les différents critères sémantiques selon lesquels ces disjonctions auraient pu être faites. Une analyse manuelle d'un sous ensemble des axiomes a fait ressortir les catégories suivantes et faire le rapprochement entre les *axiomes* MIAM et les *types de relations* $RLSM_{PI}$:

- catégories basées sur l'appartenance, par exemple, appartenance à un label (agriculture bio, indication géographique protégée etc.) : *r_has_part* ;
- catégories basées sur la transformation *r_carac* (aliment découpé) ;
- catégories basées sur la composition : *r_matter* (aliment *à base de poisson*) ;
- catégories basées sur le type/appartenance à une catégorie : *r_hypo* (volaille type dinde).

Ce rapprochement a permis d'identifier les types de relations à considérer dans le cadre d'inférence. Le processus d'inférence des ER de type « classe d'ontologie » se décompose en validation de chaîne hiérarchique (figure 1) et proposition de candidats. Le calcul de validation des chaîne hiérarchiques a

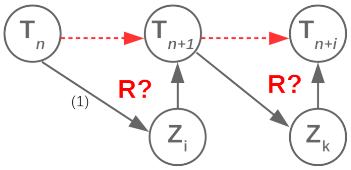

FIGURE 1 – Cas général de validation d'une chaîne hiérarchique.T_i sont des termes qui constituent la chaîne à valider. R correspond à une relation sémantique (non lexicale, non morpho-syntaxique, non ontologique) qui existerait entre les termes voisins de T_i dans la chaîne hiérarchique.

été effectué sur l'ensemble des termes correspondant aux 1 322 concepts concepts haut niveau de MIAM qui appartiennent au module *Aliment*. Au départ, nous avons obtenu 132 213 chaînes. Après filtrage par poids de la chaîne, cet ensemble a été réduit à 53 749 chaînes (40% de chaînes retenues par filtrage statistique). De nombreuses redondances existent à l'intérieur de cet ensemble car une chaîne peut contenir d'autres chaînes plus courtes. Par ailleurs, une chaîne hiérarchique peut n'être que partiellement validée. Le filtrage logique a permis d'aboutir à un ensemble de chaînes validées qui contient 9 600 chaînes (18% de chaînes retenues par rapport à l'ensemble de chaînes pré-validées statistiquement et 7% par rapport à l'ensemble des chaînes non filtrées).

Exemples de chaînes hiérarchiques après filtrage sémantique :

baguette complète→*pain complet*→*pain*→***ingrédient de recette de cuisine***→*aliment*

angélique→confiserie→bonbon La découverte des ER de type « classe/individus » est très productive quant à la proposition des individus. Suggestion des individus des classes de l'ontologie de référence :

baguette de campagne `subClassOf` pain de campagne

angélique `subClassOf` confiserie

truffe>chocolat `subClassOf` chocolat

pomme douce amère `subClassOf` pomme
pomme à cidre `subClassOf` pomme
sucre de pomme `subClassOf` confiserie

L'analyse et la validation des chaînes hiérarchiques constitue la partie la plus importante et coûteuse de l'algorithme de découverte des ER « classe/individu ». La complexité de l'algorithme dépend de l'importance du concept en train d'être traité et de la longueur des chaînes à filtrer. Le degré typé *r_isa* d_{isa} le plus élevé étant de 5 264 (pour le terme *aliment*) et la longueur maximale l des chaînes hiérarchiques obtenues étant égale à 9, la complexité dans le pire des cas serait $O(d_{isa}^{l}$ soit $O(5\ 264^{9}) = 3,103436942 \times 10^{33}$.

Le processus de découverte des ER de type« classe/individu » a pu être quantifié comme présenté dans le tableau 1.

#candidats	#valides	%valides	#nouveaux	%nouveaux
11 520	11 289	98%	4 741	42%

TABLE 1 – Découverte des ER « classe/individu » (module *Aliment* et au sous-graphe français). **#candidats** : nombre de relations candidates ; **#valides** : nombre de relations valides (après filtrage statistique et logique) ; **%valides** : taux de relations valides parmi l'ensemble des relations proposées ; **#nouveaux** : relations nouvelles parmi l'ensemble des relations valides proposées ; **%nouveaux** : taux de relations nouvelles parmi les relations valides.

4.3.3 Découverte des ER de type « propriété »

La découverte des ER de type « propriété » est utile dans le cadre de la localisation ou de passage à une ontologie multilingue. Dans le cas de notre expérience, chaque module de l'ontologie de référence possède sa propre hiérarchie de propriétés. Lors de l'immersion dans le RLSM_{PI}, les propriétés ont été exprimées en termes de relations sémantiques contextualisées via l'ajout d'une annotation. Le choix du type ou d'un ensemble de types de relations permet de distinguer les cas de figure suivants :

- *Object Properties* construites autour de la composition : `aPourProduitInitial` ;
- *Object Properties* basées sur les relations spatio-temporelles : `aPourMoisPrimeur`, `aPourRegion` ;
- *Object Properties* explicitant les caractéristiques : `aPourEtat`, `aPourArome` ;
- *Object Properties* procédurales : `aPourMethodeDeConservation` ;
- *Object Properties* qui peuvent être définies par un sous-graphe spécifique[13] : `aPourAlimentAmi`, `aPourQualificateur`.

L'ontologie à enrichir, MIAM compte 21 565 instances de *Object Properties*. Après l'immersion de l'ontologie de référence dans le RLSM_{PI}, il est possible de considérer, de manière quelque peu approximative, que nous disposons du même nombre d'instances de règle qui peuvent servir pour l'inférence des propriétés d'ontologie. Une approche naïve consisterait à mettre en place une simple inférence translingue par transfert. Cependant, il s'agit d'une technique peu productive et sujette aux erreurs dues à des problèmes d'alignement et de désambiguïsation. De plus, construite selon une méthode descendante par une communauté d'experts du domaine, l'ontologie de référence comporte un nombre variable d'instances par propriété. Les instances peuvent également être absentes pour

13. Nous entendons par sous-graphe, un sous-ensemble de termes du RLSM_{PI} connectés par des relations sémantiques.

une propriété définie. Une approche naïve ne ferait que "reconduire" ce déséquilibre.Pour affiner la méthode de découverte translingue des ER de type « propriété », nous avons opté pour une méthode par règles qui permet de sélectionner les ER candidats plus finement. L'algorithme se déroule en deux temps. D'abord, la validité de la règle pour la langue de départ est vérifiée. Puis, les structures similaires à la règle sont découvertes dans le RLS. Cette règle a la forme générale suivante :

`property=aPourEtatPhysique` *(nom de la propriété d'ontologie de référence)*
`source=?s` *(ensemble de termes qui constitue le domaine de la propriété)*
`reltype=r_carac` *(type de relation choisi pour encoder la propriété)*
`target=?o` *(ensemble de termes qui constitue le co-domaine de la propriété)*
`source_isa=aliment,préparation culinaire` *(ensemble d'hyperonymes source)*
`target_isa=état physique,état` *(ensemble d'hyperonymes cible)*
`annotation=int:physical state` *(méta-information sur la relation)*
`source_features=OUTGOING/r_pos/int:Noun`
(relations entrantes et/ou sortantes qui caractérisent les termes de l'ensemble source)
`target_features=OUTGOING/r_pos/int:Adj`
(relations entrantes et/ou sortantes qui caractérisent les termes de l'ensemble cible)

Si la règle à permis de détecter suffisamment de structures dans la langue de départ (au minimum 2 structures), elle est considérée comme valide et peut générer un objet qualifiant. Grâce à cet objet, dans un second temps, des structures candidates sont détectées au sein du $RLSM_{PI}$ au niveau des autres partitions. Cette démarche permet de capter la différence de granularité entre les différentes langues (cultures).

Le mécanisme de découverte par règles laisse apparaître les cas de figure suivants :

- relation sémantique éventuellement annotée (cas des propriétés telles que `aPresenceLactose`, `aPresenceGluten`, `aTeneurLipide`;
- patron spécifique composé de relation annotée dont les extrémités sont des termes enrichis (termes et leurs hyperonymes ainsi que éventuellement des relations sémantiques);
- structure plus complexe pour les propriétés procédurales (par exemple, `aPourRemplacant`, `aPourAlimentAmi`).

Pour les propriétés pour lesquelles il est possible d'obtenir les ER sur la base des relations sémantiques simples (notamment, *Data Properties*), nous avons obtenu les résultats présentés dans le tableau 2 dans l'état actuel du $RLSM_{PI}$.

A titre d'exemple, l'information retournée par l'algorithme par règles a la forme suivante :
M : жаркое aPourProduitDiscriminant подливка (ragoût aPourProduitDiscriminat la sauce)
R : ragoût aPourProduitInitial vegetable (produce)
M correspond au cas de figure où l'inférence a été obtenue par correspondance, en exploitant le pivot interlingue. R indique que le mécanisme d'inférence a utilisé les raffinements des termes (sens d'usage). Dans

#DP	**#triplets MIAM**	**#élém $RLSM_{PI}$**	**filtrage**	**%cand** (aug. potentielle)
aTeneurLipide	0	4 741	3 271	-
aPrésenceLactose	2 593	530	408	+16%
aPrésenceGluten	289	820	762	+263%

TABLE 2 – Découverte des ER de type Data Property à partir des relations sémantiques non annotées.

le cadre des éléments de type *Object Property* dont la découverte s'appuie les règles plus complexes, notre méthode a permis de fournir les résultats listés dans la table 3.

m	**prop**	**#trip**	**en**	**fr**	**es**	**ru**	***filt***	***%aug***
A	aPourProduitInitial	2 031	292	1 208	203	2 245	3 039	+149%
A	aPourEtatPhysique	543	30	29	10	53	85	+16%
A	aPourForme	39	77	78	5	37	132	+338%
A	aPourLabel	114	15	11	3	1	29	26%
A	aPourMethodeDeConservation	115	94	101	13	156	309	+269%
A	aPourMois	116	117	221	23	28	116	+288%
A	aPourRegion	289	98	71	2	57	216	+75%
A	aPourProduitConstituant	98	256	302	143	103	570	+582%
A	aPourProduitInitialAromatisant	41	94	147	12	567	259	+633%
P	aPourTypeDeCuisson	23	155	124	80	285	686	+2 986%
P	aPourDomaineCulinaire	82	112	92	120	1 313	1276	+1 557%
P	aPourDecoupe	82	82	78	56	77	272	+332%
S	aPourSaveur	752	51	78	47	98	232	+31%
S	aPourDescripteurBruit	119	67	80	10	6	159	+134%
S	aPourCouleur	233	192	451	59	423	911	+391%
S	aPourAspectSurface	176	40	35	12	52	101	+58%
S	aPourSensationToucher	54	84	77	21	12	155	+287%
-	Total	5388	2384	3960	937	4953	9531	+177%

TABLE 3 – Résultats approche par règles. **m** correspond au nom du module (*Aliment* (A), *Préparation* (P), *Sensoriel* (S)), **prop** - propriété, **#trip** - triplets qui correspondent à une propriété MIAM, **en**, **fr**, **es**, **ru** sont des contributions des différentes partitions en termes ER Object Property. **filt** - nombre d'éléments après filtrage (statistique, logique, validation, manuel sommaire par un non expert), **%aug** correspond à l'augmentation potentielle que représentent les ER acquis par le processus.

L'augmentation potentielle peut être importante car pour certaines propriétés l'ontologie de départ compte peu d'instances. De plus, la proposition des ER candidats est faite sur l'ensemble des langues couvertes par le RLSM_{PI}.

4.3.4 Vers une proposition automatique des ER de type « propriété »

Afin d'étoffer l'expérience de la découverte des ER, nous nous sommes intéressés à la possibilité de suggérer de nouvelles pseudo propriétés aux experts humains lorsqu'il s'agit d'une ontologie en cours de construction. Nous avons considéré SensoMIAM [14], un module de MIAM et exploité RezoJDM. SensoMIAM contient des ensembles de descripteurs sensoriels des aliments tels que :

`DescripteurTact` = {*astringent, fibreux, filandreux, granuleux, grumeleux, lisse, nerveux*}

`DescripteurSubstance`= {*aéré, bouillant, dense, épais, fin*}

`DescripteurSensationTrigeminale` = {*fraîcheur, fraîcheur mentholée, pétillant, piquant*} etc.

Pour proposer de nouveaux descripteurs, nous avons pu procéder de façon suivante :

- vérifier la convergence entre les ensembles des termes de RezoJDM par type de descripteur (*arôme*, *aspect*, *toucher*) qui correspondent aux listes des individus des classes énumérées (calcul de l'intersection) et **proposer des individus-descripteurs supplémentaires** ;
- identifier les aliments possédant les caractéristiques listées dans les descripteurs sensoriels, annoter les relations correspondantes [15] et récupérer leurs relations sémantiques contextualisées par rapport au domaine de spécialité de l'ontologie à enrichir.

14. SensoMIAM, `http://www-limics.smbh.univ-paris13.fr/sensoMIAM/`

15. Ainsi, dans *pomme* $\xrightarrow{r\_carac::bruit}$ *croquante*, le terme bruit vient contextualiser la relation.

Pour calculer de ER candidats-descripteurs, nous avons procédé procédé comme suit. Si l'ensemble des relations sémantiques sortantes d'un terme source le relie à un type d'aliment et s'il possède un ensemble de caractéristiques partagées avec d'autres termes avec un générique ≈ "aliment", le terme-cible de la relation typée r_carac non couvert par l'ébauche d'ontologie peut être proposé en tant que descripteur potentiel pour être relié à la classe correspondante. La relation *r_carac* considérées dans ce cas est annotée. Sa réification (point rouge sur la figure 2) permet de représenter la relation sous forme d'un terme (noeud).

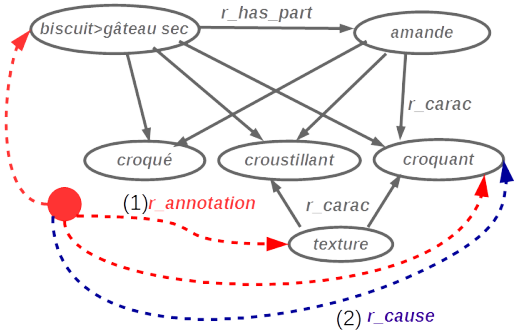

FIGURE 2 – Schéma et exemple de détection des structures pour la proposition automatique des descripteurs et des ER de type « propriété ». (1) annotation de la relation. (2) suggestion des éléments d'une nouvelle propriété.

L'expérience a permis de proposer les descripteurs suivants (exemples) :

`DescripteurArome`= {*violette, sucré-salé, persillé, miellé, musqué, moutardé, limoneux, saumuré, vinaigré*}

`DescripteurTact` = {*pâteux, écailleux, spongieux, spumeux, floconneux, velouté*}

`DescripteurSubstance`= {*sirupeux, laiteux, frémissant*}

Nous avons pu proposer automatiquement et valider semi-automatiquement 342 ER-candidats descripteurs.

Pour accompagner l'élaboration de nouvelles « propriétés » d'ontologie, nous avons défini 3 pseudo « propriétés » de test : *aPourComposantFlaveur*, *aPourComposantToucher* et *aPourComposantAspect*. Pour les peupler, nous avons utilisé la *généralisation méronymique* [16]

Exemples des propriétés de test proposées :

- *veau Orloff* `aPourComposantFlaveur` *lard* à partir de {*gras, viande*} ;
- *gratin savoyard* `aPourComposantAspect` *fromage* à partir de {*gratiné, gras, brûlé*} ;
- *sauce* `aPourComposantAspect` *graisse végétale* à partir de {*gras, fluide,nappant*} ;
- *baclava* `aPourComposantToucher` miel *à partir de* {*collant, parfumé, fondant*} ;

Nous avons pu proposer automatiquement 1 709 ER « propriété » valides validation automatique : la contrainte sur le co-domaine de la pseudo « propriété » est remplie, l'intersection entre l'ensemble des relations typées *r_carac* de l'aliment-tout et celles de composant-partie est de cardinalité suffisante).

Conclusion

Nous avons décrit une façon d'exploiter les ressources lexico-sémantiques pour enrichir et construire les ontologies. Cette démarche prend toute son importance dans le contexte multilingue où la proposition des ensembles de « classes »folksonomiques partagées appuieraient la construction collaborative multilingue entre les experts. Parmi les difficultés liées à notre méthode se trouve la disponibilité des ressources de type RLS sémantiquement riches et libres de droits.

16. Il s'agit de généraliser à un tout certaines caractéristiques de ses parties. Par exemple, "tout aliment $r\_isa \approx$ *pain* qui a une croûte est caractérisé par un bruit croquant" etc.

Références

ALLAN K. (2001). *Natural Language Semantics*. Blackwell.

AUBIN S. & HAMON T. (2006). YaTeA - version 2006. YaTeA (Yet Another Term ExtrActor) aims at identifying and extracting noun phrases which are potential terms (*i.e.* term candidates). Each term candidate is syntactically analysed in order to identify head and modifier components. YaTeA can integra.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. p. 24–27.

CHARLET J., BACHIMONT B. & JAULENT M.-C. (2006). Building medical ontologies by terminology extraction from texts : An experiment for the intensive care units. *Computer in Biology and Medicine*, **36**(7-8), 857–870.

DESPRÈS S. (2014). Construction d'une ontologie modulaire pour l'univers de la cuisine numérique. In *IC 2014 : 25es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 25th French Knowledge Engineering Conference), Clermont Ferrand, France, May 12-16, 2014.*, p. 27–38.

DESPRÈS S. (2016). Construction d'une ontologie modulaire. application au domaine de la cuisine numérique. *Revue d'Intelligence Artificielle*, **30**(5), 509–532.

DONG Z., DONG Q. & HAO C. (2010). Hownet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics : Demonstrations*, COLING '10, p. 53–56, Stroudsburg, PA, USA : Association for Computational Linguistics.

GAILLARD E., LIEBER J. & NAUER E. (2015). Improving ingredient substitution using formal concept analysis and adaptation of ingredient quantities with mixed linear optimization. In *Computer Cooking Contest Workshop*, Frankfort, Germany.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, p.7̃, Pattaya, Chonburi, Thailand.

LAFOURCADE M. (2011). *Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots. (Lexicon and semantic analysis of texts - structures, acquisition, computation and games with words)*.

LASSILA O. & MCGUINNESS D. (2001). *The role of frame-based representation on the Semantic Web*. Rapport interne, Knowledge Systems Laboratory Report KSL-01-02, Stanford University, Stanford (USA).

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014). BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation. In *ISWC : International Semantic Web Conference*, volume CEUR-WS.org of *Posters & Demonstrations*, p. 157–160, Riva del Garda, Italy.

MONDARY T. (2011). *Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels.* Theses, Université Paris-Nord - Paris XIII. Equipe RCLN.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, **193**, 217–250.

RAMADIER L. (2016). *Indexation and learning of terms and relations from reports of radiology*. Theses, Université de Montpellier.

ROCHE C. (2007). Le terme et le concept : fondements d'une ontoterminologie. In *TOTh 2007 : Terminologie et Ontologie : Théories et Applications*, p. 1–22, Annecy, France. 22 pages.

SÉRASSET G. (2014). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, p.≃-. To appear.

SPEER R. & HAVASI C. (2012a). Representing general relational knowledge in conceptnet 5.

SPEER R. & HAVASI C. (2012b). Representing general relational knowledge in conceptnet 5. In *LREC Proceedings*.

SZULMAN S. (2012). Logiciel Terminae - Version 2012. TERMINAE est une plateforme d'aide à la construction de ressources termino-ontologiques à partir de ressources textuelles.

TCHECHMEDJIEV A. (2016). *Semantic Interoperability of Multilingual Lexical Resources in Lexical Linked Data*. Theses, Université Grenoble Alpes.

ZARROUK M. (2015). *Endogeneous Consolidation of Lexical Semantic Networks*. Theses, Université de Montpellier.