Découverte des patrons de connaissance grâce à la modélisation sémantique des phrases d'instructions

Nadia Bebeshina-Clairet, Sylvie Despres, Mathieu Lafourcade

Découverte des patrons de connaissance grâce à la modélisation sémantique des phrases d'instructions

Nadia Clairet* ** ***, Sylvie Despres**,
Mathieu Lafourcade*

*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), 161, rue Ada 34095 Montpellier Cedex 5 France
[nom]@lirmm.fr
** Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), 74 rue Marcel Cachin 93017 Bobigny cedex
[prénom].[nom]@univ-paris13.fr
*** Lingua et Machina, 7 Boulevard Anatole France, 92100 Boulogne-Billancourt
[nom]@lingua-et-machina.com

Résumé.

La recette de cuisine est depuis longtemps perçue comme une structure, un ensemble d'éléments coordonnés par les linguistes comme par les concepteurs d'ontologies. Cependant, la modélisation autour des entités et des relations qui sous-tendent la suite des instructions de préparation à partir de l'analyse de corpus demeure un défi car le texte de la recette comporte une grande part d'implicite, ne mentionne pas toujours les objets et les faits relevant d'une réalité extralinguistique qu'il est indispensable de cerner pour représenter la recette en tant que processus. Dans le présent article nous introduisons une méthode de découverte des patrons de connaissance grâce à l'analyse sémantique des textes. L'originalité de cette méthode réside dans l'utilisation d'un réseau lexico-sémantique de connaissance générale à la fois pour l'analyse et la modélisation des phrases d'instructions issues des recettes de cuisine.

1. Introduction

L'analyse sémantique des recettes de cuisine concentre les problématiques à la fois de TAL[1] et IA[2] telles que les modalités d'analyse du texte court, la transformation des instructions en actions, l'analyse et la reconstitution des séquences d'actions, la représentation des processus. Compte tenu de l'évolution des pratiques nutritionnelles, l'ingénierie des connaissances pour la cuisine numérique dans le respect d'interopérabilité des ressources est devenue un enjeu important tout en demeurant un sujet complexe.

Nous proposons une approche basée sur l'analyse textuelle et la modélisation qui reposent sur l'utilisation d'une grammaire locale et d'un réseau lexico-sémantique de connaissance générale pour le français. Cette approche nous permet, lors de l'analyse, de faire émerger les patrons de connaissance relatifs au domaine. En considérant une ontologie conçue pour la cuisine numérique comme référence, il est possible de situer ces patrons dans le cadre de cette ontologie compte tenu de certains champs d'application de celle-ci.

L'article sera organisé comme suit : nous présenterons l'état de l'art en ce qui concerne l'analyse sémantique des textes et la découverte des patrons de connaissance, puis nous détaillerons notre méthode et ses premiers résultats, nous proposerons une mise en perspective de la méthode décrite.

2. État de l'Art

2.1. Analyse des recettes de cuisine en tant que textes courts

Les différentes approches à l'analyse sémantique des recettes de cuisine se séparent en deux tendances. La première est guidée par la modélisation autour des ingrédients d'une recette et, plus largement, autour de l'aliment placé dans une perspective d'adaptation. La deuxième tendance est axée sur l'aspect procédural des recettes. Elles sont exploitées en qualité d'exemple permettant de raisonner sur les relations et les contraintes temporelles ainsi que d'expliciter les différentes problématiques liées à la modélisation et l'ordonnancement des tâches, à la transformation des instructions en actions. Certains travaux qui relèvent du projet Taaable[3] (Badra *et al.* (2008) inscrits dans le paradigme scientifique du raisonnement à partir des cas tels que Dufour-Lussier *et al. (*2014) se placent dans cette dernière perspective où

[1]Traitement automatique des langues.

[2]Intelligence Artificielle.

[3]http://wikitaaable.loria.fr/

l'adaptation des recettes est guidée par l'analyse des séquences d'actions basée sur les paires problème-solution extraites des textes et les usages typiques. Dans la tradition anglophone, un des projets de modélisation a été SOUR CREAM Tasse et Smith (2008) avec MILK, le langage de représentation des instructions basé sur la logique du premier ordre. Plus récemment, dans le cadre d'une approche de *shallow parsing*, Malmaud *et al.* (2014), en expérimentant sur le même corpus que celui du projet SOUR CREAM, CMU Recipe Database[4] ont proposé un système d'analyse des recettes de cuisine qui repose sur l'étiquetage en rôles sémantiques (*Semantic Rôle Labeling*) formalisé dans son ensemble en tant que processus décisionnel markovien partiellement observable à temporalité discrète. Dans le cadre de ce système, le contexte riche est maintenu autour des termes de la recette, elle est analysée comme un tout et non pas segment par segment. En tant que suite des instructions, la recette de cuisine s'inscrit dans la perspective d'interprétation des instructions en langue naturelle. L'analyse sémantique des instructions a donné lieu à plusieurs familles d'approches : représentation du texte en langage formel (ex. Chen et Mooney (2011) ; Zettlemoyer et Collins (2005)), alignement séquence par séquence entre le texte et les actions *cf.* Andreas et Klein (2015), estimation de procès linéaires (*linear policy estimation*) *cf.* Vogel et Jurafsky (2010). Tout comme le suivi des instructions de navigation chez les derniers auteurs, l'analyse sémantique des recettes de cuisine nécessite une approche ancrée dans le monde. Elle doit en même temps rendre compte de la transformation de l'état du monde, aspect déjà abordé du point de vue d'estimateurs de procès linéaire notamment en instanciant un processus décisionnel markovien pour représenter le domaine de l'action et en utilisant l'apprentissage supervisé par renforcement de la fonction de sélection entre les actions (qui peut couvrir une suite d'instructions entière). Les observations concernant l'environnement ont servi à former notamment chez Malmaud *et al*. (2014) précité le vecteur d'état latent qui contribue à remplir les rôles sémantiques au même titre que les éléments repérés dans le texte.

Parmi les difficultés liées à ces approches, il y a respectivement la difficulté à interpréter les quantités arbitraires, les changements de l'environnement (les instructions inconnues) ; la difficulté à intégrer la compositionnalité du langage ou des actions ; la nécessité de disposer des ressources spécifiques construites manuellement permettant de simuler l'environnement ou d'aligner les instructions et les actions.

2.2. Les patrons de connaissance.

Dans la littérature, il semble exister *a minima* quatre contextes majeurs d'utilisation du terme « patron ». Premièrement, il s'agit des patrons lexicaux dont l'utilisation est liée à la technologie à états finis *cf.* une série des patrons est conçue par les linguistes ou via une stratégie d'apprentissage automatique afin de rechercher des formes de surface qui correspondent à un type de contenu formel ou sémantique.

[4]http://www.cs.cmu.edu/~ark/CURD/

Deuxièmement, il s'agit des patrons *pour le raisonnement* nécessaires pour construire des ressources ontologiques. Troisièmement, les patrons *de raisonnement* évoquent le contexte de l'utilisation des ressources ontologiques et autres bases de connaissance structurées en forme de graphe pour des tâches de raisonnement via les mécanismes d'inférence et de déduction. Enfin, les patrons sémantiques tels que les patrons proposés par Ramadier *et al*. (2014) combinent les patrons lexicaux et les contraintes d'ordre sémantique (ce qui implique la présence d'une ressource de connaissance).

Initialement, les patrons de connaissance sont des patrons ciblés proposés par les experts de domaine, mais il est également possible de les découvrir à partir des données existantes, des ressources textuelles, des ontologies. Leur découverte vient compléter le paradigme d'ingénierie à base des patrons de conception d'ontologie car elle permet d'améliorer l'interopérabilité des ressources ontologiques parfois discordantes ainsi que de faire face à certains patrons émergents qui portent sur les signes définis comme littéraux par les différents vocabulaires contrôlés[5].

Les patrons de conception d'ontologie ont été introduits simultanément par Gangemi (2005) et Blomqvist (2005). Ils peuvent concerner la représentation des prédicats ternaires RDF ou OWL ou la modélisation logique des notions génériques d'ontologie comme *processus*, *événement, participation*, *rôle*<->*tâche*, *conception*<->*objet* etc. émergeant des domaines de spécialité et des contextes applicatifs variés. Les auteurs précités distinguent plus particulièrement les patrons de contenu également désignés comme patrons conceptuels. Contrairement aux patrons d'ingénierie des connaissances qui visent à résoudre les problèmes de conception d'ontologies pour le schéma OWL sans rapport avec une conceptualisation particulière, les patrons de contenu (patrons conceptuels) sont tournés vers les problèmes de conception concernant les classes et propriétés du domaine de spécialité qui peuplent une ontologie. En termes d'implémentation, cette méthode descendante décrite, outillée et documentée par Presutti *et al.* (2009) utilise comme point de départ « les cas d'utilisation générique » (*Generic Use Cases*) dont la découverte prend forme d'une série de questions. Le contexte axiomatique et taxonomique des patrons de contenu provient des ontologies de type *core* de référence. Cependant, les patrons conceptuels peuvent également être construits à partir des ressources informelles utilisés par les experts du domaine en utilisant des méthodes de re-ingeniérie. Le lecteur peut se tourner vers les écrits des auteurs précités pour plus de détails sur la mise en œuvre de ce type de patrons.

La conception des ontologies peut également suivre un chemin ascendant grâce aux patrons de connaissance découverts à partir des ressources aussi bien ontologiques que linguistiques.

Le processus décrit par Gangemi *et al*. (2011) concernant l'extraction des patrons de connaissance est celui d'identification des chemins dans une ressource en forme

[5]http://schema.org/

de graphe fournie en entrée (*cf.* le graphe de Wikipédia pour la découverte des patrons de connaissance encyclopédique). D'après ces auteurs, outre les ressources structurées en forme de graphe, la découverte des patrons de connaissance peut également s'appuyer sur l'analyse textuelle et notamment sur les méthodes de *shallow parsing* et, plus particulièrement, l'étiquetage en rôles sémantiques, sur l'analyse sémantique profonde (basée sur l'analyse syntaxique ou sur la Théorie de représentation du discours (TRD)). Notre approche s'inscrit dans ce paradigme d'extraction des patrons de connaissance à partir d'une ressource de connaissance et d'un ciblage grâce à un schéma d'analyse et de modélisation sémantique. Contrairement aux approches basées sur TRD (et la logique du premier ordre), notre approche nous permet d'aller au-delà de l'extraction des données faisant partie du texte en tant que percept.

3. Notre approche

3.1. Ressources utilisées, structures de connaissance

Notre approche à la découverte des patrons de connaissance à partir de l'analyse textuelle repose sur l'utilisation d'un réseau lexico-sémantique de connaissance générale et d'une ontologie du domaine. A partir d'un texte d'instruction issu d'une recette de cuisine, notre système produit par calcul un graphe d'analyse qui met en évidence les relations syntaxiques et sémantiques entre les mots (termes) du texte d'instruction fourni en entrée[6]. A partir de l'analyse sémantique de textes de spécialité, nous modélisons les entités du domaine et les relations sémantiques qui les caractérisent et les relient. Avant de décrire l'approche, il convient de détailler les ressources choisies.

La ressource lexico-sémantique que nous utilisons pour effectuer l'analyse sémantique des textes et la découverte des patrons de connaissance est le réseau sémantique du français Rézo JDM issu du projet JeuxDeMots Lafourcade (2007). Construit et constamment amélioré par externalisation ouverte (*crowdsourcing*) et grâce à l'entremise de GWAP (*Games with a Purpose*), le réseau JDM est un graphe orienté, typé et pondéré qui compte à ce jour 50M de relations divisées en plus de 100 types qui relient plus de 940K noeuds. Il est doté d'un moteur d'inférences des relations sémantiques *cf*. Zarrouk *et al*. (2014) et d'annotation en méta-informations, Ramadier *et al.* (2014). Le lecteur peut se tourner vers les publications de ces auteurs pour une description plus détaillée des techniques de peuplement et de valuation du réseau. Dans le cadre de l'utilisation du Rézo JDM, notre approche bénéficie du mécanisme de désambiguïsation basé sur la notion de raffinement (chaque raffi-

[6]Comme Lafourcade (2011), nous entendons par analyse sémantique « le calcul qui, à partir d'un texte, produit une structure 1) offrant un support pour traiter un certain nombre de phénomènes linguistiques, et/ou 2) fournissant une ou plusieurs solutions aux problèmes dus à ces mêmes phénomènes » (op.cit, p.4) .

nement correspond à un usage d'un terme). Le mécanisme repose sur un schéma triangulaire qui implique le terme à désambiguïser, son contexte et son raffinement dans le réseau. A ce jour 15K termes du réseau sont étiquetés comme polysémiques et raffinés en 44K usages (couverture de raffinement de 98%). Aucune distinction n'est faite entre les différentes formes de polysémie (ex. polysémie de spécialité) et le lien est maintenu entre un terme et tous ses raffinements. Ainsi, cette ressource peut s'inscrire dans le schéma de la découverte des patrons de connaissance.

En termes de types de connaissance[7] (exprimés par l'étiquette des relations), le Rézo JDM contient, outre les relations hiérarchiques (générique, spécifique, partie-tout), les relations grammaticales (partie du discours, nombre, genre etc.). La structure grammaticale de la langue est modélisée de manière exhaustive sous forme de sous-graphe avec des nœuds (objets) tels que *Ver :*, *Ver:PPas* (participe passé d'un verbe), *Nom :* etc . Les nœuds lexicaux (termes) se caractérisent via des relations typées *r_pos*, *r_nombre*, *r_genre* etc. qui les relient aux nœuds évoqués. Le réseau contient également des relations causales ainsi que de nombreux types de connaissance absents des autres ressources lexicales telles que WordNet Fillmore (1998) ou Wiktionnary. Il s'agit en particulier des relations thématiques (*lieu, instrument, manière* etc.), prédicatives (*agent, patient* etc.), des relations d'inspiration générativiste introduites par Pustejovsky (1991) en tant que partie intégrante de la structure de qualia (*rôle télique*, *rôle agentif* etc.).

La partie « alimentation, cuisine » du réseau a été adaptée à l'analyse des textes de cuisine et nutrition en intégrant les connaissances nutritionnelles, civilisationnelles, sensorielles, techniques. A ce jour, la partie alimentation/cuisine représente environ 30K termes (3 % du Rézo JDM). La palette des relations mobilisées par les nœuds lexicaux du domaine de cuisine/nutrition inclut les relations taxonomiques, les relations prédicatives (« battre les blancs en neige », <battre *r_patient* blancs>, les relations de lieu (*lieu>objet* et *lieu>action ex. « ajouter le sucre au mélange »,* <ajouter *r_lieu>action* mélange>), *manière (ex.* « battre vigoureusement les œufs », <battre *r_manière* vigoureusement>) *instrument (ex. «* remuer avec une cuillère », <remuer *r_instrument* cuillère>*)*, relations temporelles (*ex.* « lavez, épluchez les légumes »*,* <laver *r_successeur_temps* éplucher>), *rôle télique (ex.* « ajouter de la fécule pour épaissir »*,* <fécule *r_telic_role* épaissir>), *matière-substance* (*ex.* « le lait est riche en protéines », <lait *r_object>mater* protéine>), *quantité* (« prenez un verre de lait », <lait *r_quantifieur* verre>).

Compte tenu de la structure syntaxique des phrases d'instruction, notre approche s'intéresse à la structure prédicat-arguments et, plus largement, à la «monotonie» du champ argumental des verbes de transformation[8] à savoir la présence d'une série de relations transversales (et non uniquement hiérarchiques) récurrentes. Cette régularité de la recette de cuisine examinée sous l'angle syntaxique, mais aussi sémantique a

[7]Pour les termes monosémiques, il serait possible dire que les relations du Rézo JDM sont une mise en œuvre des fonctions lexicales

[8]Verbes qui renvoient à un « faire transformateur »

également été analysée par Colson (2010). Son étude relie quatre catégories d'arguments à des fonctions syntaxiques type que nous avons également remarqué lors de nos expérimentations. Cette analyse linguistique concerne un texte de recette extrait du Cuisinier françois *de La Varenne (1651)*. Ainsi, du point de vue diachronique, la recette de cuisine constitue un genre de texte instructionnel ancien et bien à part.

Nous exploitons également la valeur adjectivale du participe passé des verbes que nous observons notamment au niveau de la relation sémantique *caractéristique* ex."oignons ciselés" permet de restituer l'action antérieure "ciseler les oignons". Nous explorons également les structures sans marqueurs syntaxiques explicites ou avec connecteurs inexpressifs, ex. « mélange sucre ricotta » .

Par certains points, notre schéma d'analyse est similaire à celui propose par (Clark *et al.*, 2014) dans le cadre d'élaboration d'un système intelligent de question-réponse. On retrouve d'autres similarités avec notre approche dans le paradigme de réécriture des graphes et, plus particulièrement, de la méthode GrGen de Geiss *et al. (*2006). Il s'agit notamment du plan de recherche structure en deux temps : *lookup* et extension[9]. La différence importante entre ces approches et la nôtre est l'utilisation du graphe de connaissance générale. Cette ressource nous permet d'intégrer le schéma interprétatif directement a l'intérieur du schéma d'analyse. L'interprétation sémantique contribue à l'analyse sémantique et *vice versa*.

Parmi les approches similaires à l'analyse sémantique des textes basées sur une ressource de connaissance générale, sont à remarquer l'analyseur sémantique conceptuel de Rajagopal *et al*. (2013) basé sur les parties du discours et le réseau ConceptNet Liu *et al*. (2004) l'analyseur de Poria *et al*. (2014) qui utilisent également l'ontologie ConceptNet et les relations de dépendance syntaxique pour l'analyse sémantique des opinions.

La découverte des patrons de connaissance est une démarche qui implique la présence d'une ou plusieurs ontologies de référence afin de permettre de vérifier la consistance des patrons. Dans notre cas, l'ontologie modulaire pour la cuisine numérique développée par le laboratoire LIMICS Despres (2014) représente cette référence. Elle a été conçue selon la méthodologie NéON[10], avec la participation directe des experts du domaine et dans l'objectif de constituer un standard. Le module «aliment» est le module central de l'ontologie, il est structuré autour de 6 groupes de caractéristiques qui correspondent aux champs applicatifs de l'ontologie : les descripteurs aliment, les caractéristiques sensorielles, géographiques, nutritionnelles, de qualité, d'approvisionnement. C'est dans le cadre de ce module et plus particulièrement, en ce qui concerne les descripteurs aliment, que nous nous plaçons pour découvrir des patrons de connaissance. Nous nous intéressons également aux caractéristiques nutritionnelles (notamment dans le cadre de patron de substitution en cas d'allergies alimentaires).

[9]« exploration et recopie » selon Lafourcade (2011), op.cit, p.226.

[10]http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project

3.2. Méthode et implémentation

En termes de méthode, notre approche à l'analyse des textes repose sur l'implémentation d'une grammaire locale qui impose des contraintes sur le réseau JDM à la fois d'ordre syntaxique et sémantique. Une telle approche basée sur un réseau lexico-sémantique « multi-facette » permet d'éviter la dissociation des strates grammaticale et lexicale du sens. Cette dissociation parfois prématurée peut mener à des erreurs de désambiguïsation, le succès uniquement partiel ou à l'échec de l'analyse.

Notre schéma d'analyse se présente comme suit ;

Entrée : texte brut (une phrase d'instruction).

Sortie : graphe d'analyse (triplets au format *« x r_relation/poids y », cf. 5:r_carac-1/100:24*).

Etape 0. Le texte fourni en entrée est représenté sous forme de chaîne où chaque terme, mais aussi chaque événement comme début et fin de la phase, ponctuation est représenté sous forme de nœud relié à ses voisins par une relation symétrique successeur/prédécesseur.

Etape 1. Le texte est représenté sous forme d'un graphe de dépendances. La détection des unités polylexicales est faite grâce au Rézo JDM (approche à base des n-grammes et de listes des unités polylexicales issus de la ressource, complétée par des règles spécifiques), une approche qui peut être caractérisée comme basée sur un dictionnaire ;

Etape 2. Augmentation du graphe de dépendances grâce à la projection du Rézo JDM sur le texte. Conformément aux contraintes portées par les différentes règles de la grammaire locale, il s'agit de récupérer les relations sortantes d'un certain type dans le réseau JDM pour constituer un graphe d'analyse. Nous considérons les chaînes (succession des arcs) dont la longueur est égale à 3. Nous utilisons également la transitivité de certaines relations (*synonyme*, *is-a*, *hyperonyme*), qualité utilisée également dans le cadre de propagation des relations dans le Rézo JDM[11]. Une fois le graphe local d'analyse construit, les règles relatives à la découverte des entités peuvent s'appliquer.

Etape 3. Réification des entités d'analyse, reconstitution de la séquence des actions qui correspond à la suite des instructions, émergence des patrons de connaissance. La découverte des patrons de connaissance repose sur la détection des entités pouvant être définies comme des entités d'alignement. Ces entités sont prédéfinies et spécifiques à l'analyse des processus de préparation en cuisine. Dans la version actuelle de notre grammaire, nous distinguons des entités comme *ingrédient*, *ensemble d'ingrédients* (ingrédients qui subissent la même action), *ustensile*, *quantité*, *action*, *état*, *préparation*. Les patrons de connaissance émergents sont des sous-graphes du graphe d'analyse qui mettent en évidence les relations hiérarchiques et les rôles sémantiques jugés pertinents. Quant à la reconstitution de la séquence d'actions (ordre

[11]Zarrouk *et al.*, op cit.

implicite des actions), il s'agit d'un travail en cours qui exploite les ressources (*ustensile*, *ingrédient simple ou complexe*, *lieu*) mobilisées par une action de préparation en tenant compte de leur transformation. On observe, outre les analogies à la fois formelles et sémantiques (comme dans le cas d'adjectif/participe passé désignant à la fois l'état présent et l'action précédente, le changement des rôles sémantique récurrent. Ainsi, un nœud-patient d'une action N (ex « œuf », « sucre ») disparaît au profit d'un nœud-lieu d'action N+1 (« pâte », « appareil>cuisine ») combiné avec une relation de méronymie *has-part* ou avec une relation *matière-substance* vis-à-vis du patient de l'action N.

En pratique, dans la grammaire hors contexte implémentée, nous utilisons un formalisme de travail où les relations sont notées r _type _de_relation et où les variables sont précédées du signe $ (ex. $x, $entité, etc). Leur instanciation correspond soit à des signes terminaux soit à des éléments de la structure syntaxique (GN, GV, GNPREP etc.).

Les règles pouvant être formalisées en logique du premier ordre comme *x, Rb(x;B)) Rc(x;C)* sont exprimées comme suit (exemples) :

- (*huile d'olive*) : $x de $y & $x *r_isa* substance & $x *r_carac* comestible & $y *r_isa* aliment => $x de $y *r_isa* substance

- (découverte de la relation *lieu de l'action) :* $x == GV: & $x r_patient $y & $y r_isa substance => $x r_lieu_action $y

- (propagation) : $x *r_lieu*_action $y & $x *r_head*[12] $z => $z r_lieu_action $y

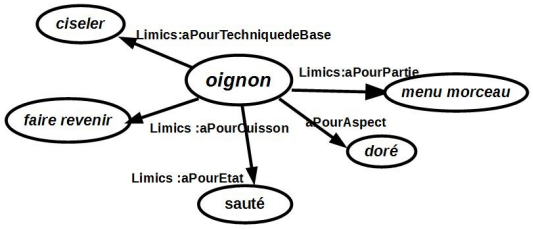

FIG. 1 – *Représentation du patrons de connaissance émergeant en termes d'ontologie de référence, phrase fournie en entrée :« faites revenir les oignons ciselés »)*

Une fois que les entités et les relations découvertes, il est possible d'exprimer les patrons de connaissance émergents en termes d'ontologie de référence afin de vérifier la pertinence de la connaissance extraite des textes vis-à-vis de la construction d'ontologie du domaine.

L'analyse sémantique des phrases d'instructions permet également de découvrir les patrons émergents de substitutions d'ingrédients en cas d'intolérance alimentaire ou encore de suggérer une manière de représenter une séquence d'actions (sachant

[12]La relation *r_head* correspond à un lien de constituance. (Lafourcade, 2011, op.cit., p.229).

que pour valider cette représentation une autre ontologie de référence sera nécessaire).

A ce jour, nous avons appliqué notre méthode à un corpus de 1 500 phrases d'instructions. En termes de modélisation des séquences d'actions, conformément à un protocole d'évaluation manuel faute de système de base référent pour le français les actions ont été correctement modélisées dans 57 % des cas (tous les nœuds et relations attendus ont été construits par le système), dans 28 % des cas la modélisation a été partielle, enfin dans 28 % des cas le système a échoué ce qui est dû à l'incomplétude de notre ressource. Nous avons extrait 1500 relations typées r_patient, 505 caractéristique , 378 manière, 420 successeur temporel, 168 partie-tout, 84 quantificateur, 462 lieu-objet, 336 lieu-action, 25 instrument. En termes de modélisation, nous avons obtenu 966 entités « action » ; 70 « état » ; 15 « événement » (ex. « le lait frémit »), 420 « ensemble d'ingrédients », 63 « préparation ».

4. Conclusion et perspectives

L'approche de la découverte des patrons de connaissance grâce à l'analyse sémantique des textes basée sur l'utilisation d'un réseau lexico-sémantique de connaissance générale que nous avons introduite offre des possibilités intéressantes. Cependant, cette approche a également des limites dues à la couverture et la précision de la ressource utilisée. La connaissance générale est fournie par les contributeurs non spécialistes ce qui explique un certain nombre d'incohérences pour la tâche d'analyse. Ainsi, dans Rézo JDM, il existe la relation <faire revenir *r_conséquence* cuit>, or, l'aliment à l'issue de cette opération ne peut pas être considéré comme « cuit ». Plus généralement, les techniques de base en cuisine repérées grâce à l'analyse sémantique nécessitent d'être validées via une fonction de distance polyaxiale (multidimensionnelle), qualifiée (en quoi la technique consiste-t-elle ?) et qualifiante (pourquoi s'agit-il d'une technique de base ?). Les perspectives de notre travail sont l'implémentation d'un algorithme de « traduction » des patrons de connaissance découverts en format RDF ainsi que l'élaboration d'un algorithme « d'alignement » vis-à-vis de l'ontologie de référence. Une autre piste intéressante est l'introduction de la dimension multilingue à la fois dans la représentation des connaissances lexico-sémantiques et dans l'analyse des textes et la découverte des patrons de connaissance. En effet, en incluant cette dimension, chaque élément de connaissance trouvé dans une langue peut profiter aux autres langues. Des connaissances conceptuelles indépendantes de langues peuvent ainsi être repérées.

References

Andreas, Jacob and Klein, Dan (2015). Alignment-based compositional semantics for instruction following. In *EMNLP,* 10 p.

Badra, F., Bendaoud, R., Bentebibel, R., Champin, P. P.-A., Cojan, J., Cordier, A. et al. (2008). TAAABLE:Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *9th European Conference on Case-Based Reasoning - ECCBR 2008*, 5239, 219–228.

Blomqvist, Eva and Sandkuhl, P. (2005). Patterns in ontology engineering : Classifcation of ontology patterns. In *ICEIS 2005, Proceedings of the Seventh International Conference on Enterprise Information Systems,* Miami, USA, May 25-28, 413-416.

Chen, David L. and Raymond Mooney, Raymond J. (2011). Panning for Gold: Finding Relevant Semantic Content for Grounded Language. In *Learning Symposium on Machine Learning in Speech and Language Processing (MLSLP),* 6 p.

Colson, M. (2010). Le vocabulaire culinaire comme système, la recette de cuisine comme genre. Etude linguistique du Cuisinier françois de La Varenne" In *Les Journées de linguistique, Québec, Canada.,* 4 p.

Das, D., Chen, D., Martins, A., N. Schneider, and Smith, N. (2014). Frame-semantic parsing. In *Computational Linguistics,* 82 p.

Despres, Sylvie. (2014). Construction d'une ontologie modulaire pour l'univers de la cuisine numerique. In Catherine Faron-Zucker. *IC - 25emes Journees francophones d'Ingenierie des Connaissances*, May 2014, Clermont-Ferrand, France. 27-38.

Dufour-Lussier, V., Le Ber, F., Lieber, J., & Nauer, E. (2014). Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems*, *40*, 153–167.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. 423 p.

Gangemi, Aldo (2005).Ontology design patterns for semantic web content. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *The Semantic Web - ISWC 2005*, *4th International Semantic Web Con-ference, ISWC 2005,* Galway, Ireland, November 6-10, 2005, Proceedings, volume 3729 of Lecture Notes in Computer Science, 262-276. Springer.

Geis, R., Veit Batz, G., Grund, D., Hack, S., and Szalkowski, A. (2006). GrGen: A fast SPO-based graph rewriting tool*, ICGT 2006,* A.Corradini et al. (eds.), *Springer,* 383-397.

Kamp, Hans. et Reyle, Uwe (1993). *From Discourse to Logic*. *Introduction to Modeltheoretic Semantics of Natural Language*, Formal Logic and Discourse Representation Theory Kluwer, Dordrecht, 59-139.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th international symposium on natural language processing.*

Lafourcade M. (2011). *Lexique et analyse sémantique de textes – structures, acquisitions, calculs, et jeux de mots,* HDR, Université Montpellier II, 298 p.

Liu, H. and Singh, P. (2004). Conceptnet - a practical commonsense reasoning tool-kit. In *BT technology journal*, 22(4):211–226.

Malmaud, J., Wagner, E. J., Chang, N., & Murphy, K. (2014). Cooking with Semantics. In ACL Proceedings, 33–38.

Nuzzolese, Andrea Giovanni, Gangemi, Aldo, Presutti, Valentina, Ciancarini, Paolo (2011). Encyclopedic Knowledge Patterns from Wikipedia Links. In C.Welty, A. Harith, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist, editors. *The Semantic Web -- ISWC 2011: 10th International Semantic Web Conference*, Bonn, Germany, October 23-27, 2011, Proceedings, Part I, 520-536. Springer.

Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., and Howard, N. (2014). Dependency-Based Semantic Parsing for Concept-Level Text Analysis. In *Computational Linguistics and Intelligent Text Processing Volume 8403 of the series Lecture Notes in Computer Science,* 113-127.

Presutti, Valentina, Daga, Enrico, Gangemi, Aldo, Blomqvist, Eva (2009). eXtreme design with content ontology design patterns. In *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516,* 83-97.

Pustejovsky, James (1991). The generative lexicon. In *Journal Computational Linguisticsolume 17 Issue 4,* 409-441, MIT Press Cambridge, MA, USA.

Rajagopal, D., Cambria, E.,Olsher,D., and Kwok, K. (2013). A graph-based ap-proach to commonsense concept extraction and semantic similarity detection. In*Proceedings of the 22nd international conference on World Wide Web companion*, 565–570. International World Wide Web Conferences Steering Committee.

Tasse, D. and Smith, N.(2008). *SOUR CREAM: Toward Semantic Processing of Recipes*. Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA.

Vogel, Adam and Jurafsky, Dan. (2010). Learning to Follow Navigational Directions. In *Proceedings of ACL-2010*, Uppsala, Sweden, 10 p.

Zettlemoyer, L.S., Collins, M. (2005). Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 9 p.

Zarrouk, M., Lafourcade, M. adn Joubert, A. (2013). Inference and reconciliation in a lexical-semantic network.*14th International Conference on Intelligent Text Processing and Computational Linguistic (CICLING-2013),* 13 p.

Abstract

The cooking recipe has long been considered as a system by linguists and ontology engineers. However, the corpora based cooking ingredient & instructions' modelling still remains a challenging issue as the cooking recipe text contains a lot of impliciteness. It doesn't mention extralinguistic objects and facts which has to be grasped in order to represent the recipe as a process. In the present article, we introduce a method of knowledge pattern discovery based on deep semantic parsing. The originality of our approach stems from the use of a large general knowledge lexical semantic network for cooking instruction analysis and modelling.