Peuplement d'une base de connaissance par annotation automatique de textes relatifs à la cosmétique

Molka Tounsi, Cédric Lopez, Catherine Faron Zucker, Elena Cabrio, Fabien Gandon, Frédérique Segond

Peuplement d'une base de connaissance par annotation automatique de textes relatifs à la cosmétique

Molka Tounsi Dhouib[1], Cédric Lopez[2], Catherine Faron Zucker[1], Elena Cabrio[1], Fabien Gandon[1], Frédérique Segond[2]

[1] UNIVERSITÉ CÔTE D'AZUR, INRIA, CNRS, I3S, SOPHIA ANTIPOLIS, FRANCE
{dhouib,faron,cabrio}@i3s.unice.fr, fabien.gandon@inria.fr

[2] VISEO, R&D, GRENOBLE, FRANCE
{cedric.lopez,frederique.segond}@viseo.fr

Résumé : Dans cet article, nous proposons une approche pour construire une base de connaissances à partir de textes dans le domaine de la cosmétique. Il s'agit d'un cas particulier pour un domaine fixé du problème de l'extraction de relations à partir de textes. Dans le but de résoudre ce problème, nous proposons une approche semi-supervisée pour l'extraction des relations en utilisant parallèlement les méthodes suivantes : (i)l'extraction de relations basée sur la signature des propriétés, (ii) la construction de patrons d'extraction à partir des résumés présents dans les pages de DBpedia, (iii) l'annotation manuelle d'un ensemble de textes pour définir des patrons syntaxiques pour extraire les relations. Nous avons évalué notre approche sur deux types de corpus : (i) un premier corpus est composé d'articles de journaux spécialisés, tels que aufeminin.com et Cosmétique Hebdo, (ii) un deuxième corpus est constitué d'un ensemble de phrases collectées sur sur le Web. L'évaluation présentée dans cet article combine les résultats des trois méthodes.

Mots-clés : Base de connaissances, extraction de connaissances, reconnaissance d'entités nommées, extraction de relations, RDF, ontologies, TAL, cosmétique.

1 Introduction

Avec la croissance exponentielle de la quantité de données numériques et la diversification de leurs sources, les traitements automatiques des données deviennent chaque jour un enjeu plus crucial. La mise en place de tels traitements est particulièrement difficile lorsque les données sont hétérogènes, distribuées et peu structurées comme c'est le cas pour des textes libres de pages d'information existant sur le Web. Aussi, la grande masse de données textuelles publiées sur le Web a donné lieu à de nombreux travaux en Extraction d'Information dont un but est d'enrichir ces textes tout-venant par des méta-données structurées et fortement connectées afin d'exploiter ces liens et leur sémantique dans le traitement de l'information comme la recherche, la notification, l'agrégation, etc. L'extraction d'information (EI) consiste à extraire des informations pertinentes à partir des textes telles que des entités (par exemple, *personnes*, *organisations*, *dates*), et des relations (*née à*, *racheté par*, *est du type*, ..).

Dans le cadre du projet de recherche collaboratif SMILK [1] (Social Media Intelligence and Linked Knowledge), deux questions réciproques ont été soulevées : 1) Comment utiliser le Traitement Automatique du Langage Naturel (TALN) pour contribuer au développement du Web des données liées, 2) comment utiliser le web des données liées pour contribuer à la résolution

1. https ://project.inria.fr/smilk/fr/

de tâches du TALN. Au cœur de ces problématiques, nous nous focalisons sur la tâche d'extraction de relations entre deux entités à partir des textes en français. Par exemple, dans la phrase simpliste "La Vie est Belle contient du linalol", l'objectif est d'extraire la relation "contient" entre le parfum "La Vie est Belle" et le composant "Linalol". On peut dès lors structurer la connaissance sous forme de triplets RDF. D'un point de vue applicatif, porter l'intérêt sur les relations sémantiques permet d'enrichir les méta-données afin de procéder à une recherche plus précise, moins ambiguë.

Dans cet article, nous partageons notre expérience concernant l'extraction de relations en utilisant d'une part des outils et des techniques utilisées dans le domaine du TALN, d'autre part des ontologies publiées sur le Web des données ouvertes. Nos collaborations antérieures nous ont conduites à travailler dans le secteur du Luxe et plus précisément dans la Cosmétique, c'est pourquoi nous baserons nos expériences sur l'utilisation de ProVoc, décrite par (Lopez *et al.*, 2016), qui est un vocabulaire permettant de décrire les produits sur le Web, construit à partir de scénarios majoritairement issus de la Cosmétique. ProVoc se positionne comme une extension de l'ontologie GoodRelations, décrite par (Hepp, 2008) et se concentre sur une représentation fine des entités d'intérêts (gammes de produits, composants, créateurs, *etc.*) et des relations les reliant (`belongsToBrand`, `hasComponent`, `hasDesigner`), *etc*.

Dans la suite de l'article, nous décrivons les travaux antérieurs (section 2), puis le cadre général de notre approche (section 3). Dans la section 4, nous nous focalisons sur la tâche d'extraction de relations que nous évaluons dans la section 5. La dernière section met en avant nos conclusions et nous y indiquons nos perspectives.

2 Approche générale

L'approche générale d'extraction de connaissances que nous avons suivie est illustrée en Figure 1. Le système prend en entrée un texte non structuré, transmis à un premier module de reconnaissance d'entités nommées, puis à un module d'extraction des relations.

La reconnaissance des entités doit être en mesure de reconnaître les entités qui peuvent être typées avec des classes de ProVoc dont les labels sont en anglais, *i.e.* personnes, entreprises, divisions d'une entreprise, noms de marques, noms de produits, noms de gammes de produits, composants de produits. Cette reconnaissance s'effectue par des outils complémentaires :

— Renco décrit par (Lopez *et al.*, 2014), basé sur des règles lexico-syntaxiques, qui permet d'extraire à partir des textes français des produits, leurs ingrédients, gammes de produits, marques, divisions et groupes. Les ressources utilisées par Renco étant limitées pour les noms de produits cosmétiques, nous avons construit automatiquement un lexique complémentaire contenant 1130 noms de produits à partir de pages Wikipedia français telles que "Liste de parfums" (en utilisant le service Web MediaWiki).

— Holmes Semantic Solutions [2], basé sur une approche hybride (symbolique et statistique) pour reconnaître les noms de personnes.

Une fois les entités nommées annotées, la seconde étape consiste à extraire automatiquement les relations qui existent entre ces entités, en se basant sur un analyseur syntaxique et sur la base de connaissances DBpedia. L'extraction de relations est détaillée dans la section suivante.

2. http ://www.ho2s.com/fr/

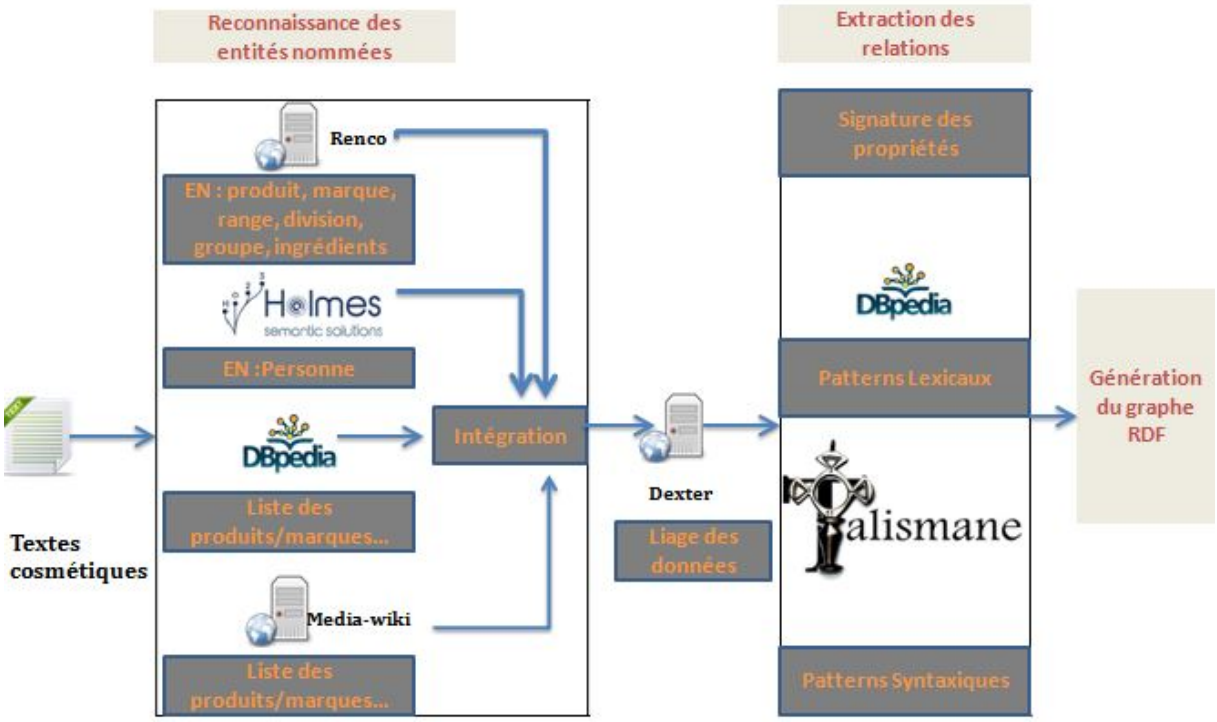

FIGURE 1 – Processus général d'extraction de connaissances à partir de textes

3 Extraction de relations

Dans cette section, nous faisons un tour d'horizon des travaux antérieurs menés sur la tâche d'extraction de relations puis nous décrivons notre méthode.

3.1 Positionnement

La tâche d'extraction des relations consiste à détecter des liens sémantiques qui existent entre différentes entités. Les différentes approches existantes peuvent être classées en trois catégories :

— Une première approche, supervisée, consiste à considérer le problème de l'extraction de relations comme un problème de classification, et à utiliser un classifieur linéaire, auquel il faut fournir un ensemble d'exemples positifs et un ensemble d'exemples négatifs pour l'apprentissage. Par exemple, les méthodes dites *feature-based* et *kernel-based* décrites dans (Nebhi, 2013) adoptent cette approche.

— Une deuxième approche, semi-supervisée, permet au système d'apprendre itérativement des patrons et des instances à partir d'un nombre réduit d'instances : les patrons sont

Propriété	Domaine	Co-domaine
belongsToBrand	pv:ProductOrServiceRange gr:ProductOrService	gr:Brand
belongsToDivision	gr:Brand	pv:Division
belongsToGroup	pv:Division	gr:BusinessEntity
hasComponent	gr:ProductOrService	pv:Component
hasFragranceCreator	gr:ProductOrService	foaf:Person
hasRepresentative	rdf:Resource	foaf:Person

TABLE 1 – Liste des propriétés à extraire

utilisés pour extraire des nouvelles relations à partir de l'ensemble de données, qui sont ajoutées à l'ensemble des exemples. Cette opération est répétée jusqu'à ce qu'aucune nouvelle relation ne puisse être apprise à partir de l'ensemble de données. Les systèmes DIPRE, Snowbal, supervision distance, BOA décrits par (Kumar & Manocha, 2007) adoptent cette approche.

— Une troisième approche, non supervisée, est la génération des patrons d'extraction. Elle est décrite dans (Nebhi, 2013). Cette méthode consiste à collecter des paires de mots avec les chaînes de caractères (string) qui les séparent, pour calculer la cooccurrence de termes ou pour générer les patrons. RdfLiveNews est un système décrit dans (Gerber *et al.*, 2013) qui adopte cette approche pour générer une base de connaissances RDF à partir de textes.

Dans notre travail, nous combinons deux approches pour extraire les relations : (i) une approche semi-supervisée qui consiste à extraire des relations dans les textes en utilisant les propriétés définies dans des ontologies, (ii) une approche supervisée qui consiste à extraire des relations à l'aide de patrons syntaxiques définis manuellement en analysant un corpus de textes dit d'entrainement, au préalable annoté également manuellement dans ce but.

3.2 Processus d'extraction des relations

Dans le cadre de cette étude, nous avons utilisé six propriétés définies dans le vocabulaire ProVoc, listées dans le tableau 1 avec l'indication de leurs domaines et co-domaines correspondant aux types des entités nommées concernés par ces relations. Dans les sous-sections suivantes, nous décrivons nos approches semi-supervisée et supervisée.

3.2.1 Extraction de propriétés basée sur des ontologies

La première approche d'extraction repose sur (i) les signatures des relations telles que définies dans les ontologies ProVoc et GoodRelations, (ii) des patrons lexicaux générés à partir de DBpedia.

3.2.1.1 Extraction de relations basée sur la signature des propriétés

ProVoc et GoodRelations fournissent les signatures de chaque propriété à partir desquelles des règles d'extraction sont automatiquement générées. Par exemple, la propriété belongsTo-

`Division` implique que le sujet doit être de type `Brand` (resp. `Division`) et que l'objet doit être du type `Division` (resp. `Brand`). La règle suivante est ainsi générée :

belongsToDivision : (sujet=Brand ET objet=Division) OU (sujet=Division ET objet=Brand)

Ces règles sont mises en oeuvre en utilisant Renco qui permet de nous fournir le typage des EN, et appliquées sur trois types de relations `belongsToBrand`, `belongsToDivision` et `belongsToGroup`. Nous avons défini une règle pour chacune des relations. L'application de ces règles sur le texte annoté par les outils de reconnaissance d'entités nommées revient à projeter les signatures de propriétés sur le texte pour repérer les entités ayant pour types respectifs le domaine et co-domaine d'une propriété recherchée. L'existence d'une phrase contenant deux types différents d'entités nommées implique que nous avons forcément une relation entre ces deux entités nommées.

3.2.1.2 Extraction des relations basée sur des patrons lexicaux à partir de DBpedia

En nous inspirant de l'approche proposée par (Gerber & Ngomo, 2011) nous avons construit des patrons lexicaux (sans utilisation de la syntaxe) à partir de l'analyse des résumés des produits cosmétiques décrits dans DBpedia en l'interrogeant avec des requêtes SPARQL. Nous avons écrit une requête SPARQL pour interroger la page DBpedia dont le nom est "Liste de parfums", le but étant d'extraire des informations concernant les parfums tels que leurs noms, leurs marques et le résumé présenté pour chaque parfum.

```
select  ?parfum_name ?brand_name ?abstract
where {
       ?parfum_list rdfs:label "Liste de parfums" @fr.
       ?parfum_list dbpedia-owl:wikiPageWikiLink ?parfum.
       ?parfum rdfs:label ?parfum_name.
       ?parfum prop-fr:marque ?brand.
       ?brand rdfs:label ?brand_name.
       ?parfum dbpedia-owl:abstract ?abstract.
       FILTER (LangMatch (lang(?abstract), "fr" ))
       FILTER (LangMatches (lang(?parfum_name), "fr"))
       FILTER (LangMatches (lang(?brand_name), "fr"))
 }
```

A partir de ces trois principales informations, nous avons considéré les patrons lexicaux comme les éléments textuels se situant entre deux entités respectant la signature d'une propriété donnée. En analysant par exemple les informations relatives au parfum "Allure Homme", nous obtenons que ce parfum appartient à la marque "Chanel" et que son résumé est "Allure Homme est un parfum masculin de Chanel, créé par Jacques Polge et sorti en 1999". Nous avons considéré que les éléments lexicaux qui se trouvent entre le nom de parfum et le nom de la marque représentent le patron lexical « **est un parfum masculin de** » . Nous avons obtenu neuf patrons lexicaux.

3.2.2 Extraction de relations basée sur des patrons syntaxiques

En complément des approches lexicales présentés ci-avant, nous avons développé une approche basée sur les relations de dépendances syntaxiques ce qui a pour avantage de ne pas s'en tenir exclusivement aux éléments lexicaux se situant entre les entités. En particulier, les dépendances syntaxiques entre le verbe, le sujet et l'objet permettent d'assurer la cohésion de ces trois éléments. De même, la présence d'une préposition dans un syntagme déterminatif, comme dans le cas de "Coco Noir de Chanel" permet d'identifier une relation d'appartenance (*belongsTo*). Dans le cadre de cette étude, nous avons utilisé l'analyseur syntaxique Talismane décrit dans (Urieli, 2013), un des rares analyseurs pour le français sous licence GPL qui permette le repérage et l'étiquetage des dépendances syntaxiques entre les mots.

Afin de développer les patrons syntaxiques, nous avons construit un jeu d'apprentissage constitué de 58 phrases issues de différents journaux tels que Cosmétique Hebdo, Cosmétique Mag. Ce corpus contient 55 relations de type `hasComponent`, 27 relations de type `hasFragranceCreator` et 24 relations de type `hasRepresentative`. Une étape de pré-traitement a consisté à annoter les résultats issus de la reconnaissance des entités nommées (cf. Section 2) de telle façon à ce que Talismane en tienne compte. Le type des EN reconnues est également indiqué à Talismane en utilisant l'attribut `comment` de son API.

A partir du jeu d'apprentissage nous avons défini manuellement 30 patrons pour extraire la relation `hasRepresentative`, 22 patrons pour extraire la relation `hasFragranceCreator` et 63 patrons pour extraire la relation `hasComponent`. Considérons par exemple la phrase suivante : « La ligne s'anime également en mai avec une édition limitée Summer, rafraîchie d'ananas, création d'Anne Flipo et de Carlos Benam (IFF), dans un flacon orange (EdT vapo 50 et 100 ml, 42 et 48) ». Cette phrase contient une relation de type `hasComponent` entre `Summer` et `ananas` et deux relations de type `hasFragranceCreator` entre `Summer` et `Anne Flipo` et entre `Summer` et `Carlos Benam`.

Concrètement, pour extraire la relation `hasComponent` et son objet, nous avons écrit la règle syntaxique 1. Cette règle consiste à extraire un token dont la partie du discours (POS) est un participe passé (VPP) ayant une relation de type modificateur ou dépendance, suivi d'un token dont le POS est "préposition et déterminant" ou seulement préposition ayant une relation de dépendance de type syntagme prépositionnel, suivi d'un token de type nom propre ayant une relation de préposition avec son prédécesseur et qui représente une entité nommée de type "Component". Pour extraire la relation `hasFragranceCreator` et son objet, nous avons écrit les règles syntaxiques 2 et 3. L'extraction du sujet attachée à une propriété est effectuée en suivant le même principe : pour identifier le sujet, nous avons implémenté 22 patrons. Par exemple, pour extraire le sujet des propriétés extraites de la phrase ci-dessous, nous avons défini la règle 4.

1 "SI depRel(,VPP)=(mod ou dep) ET depRel(VPP,(P+D ou P))=de obj ET depRel((P+D ou P),NPP)=prep ET type (NPP)=ingredient".
2 "SI depRel(0,VPP)=mod ET depRel(VPP,NC)=(mod ou suj ou prep) ET depRel(NC,(P ou P+D))=dep ET depRel((P ou P+D),NPP)=prep ET type (NPP)=PER
alors (NPP)=FragranceCreator"
3 "SI depRel(0,VPP)=mod ET depRel(VPP,NC)=(mod ou suj ou prep) ET depRel(NC,(P ou P+D))=dep ET depRel((P ou P+D),NPP)=prep ET type NPP =PER + isCoordination"
4 "Si depRel(0,V)=root ET depRel(V,P)=mod ET depRel(P,NC)=(prep ou obj) ET de-

pRel(NC,NPP)=mod ET type(NPP)=product"

Le résultat du processus est un ensemble de triplets RDF décrivant d'une part chaque entité reconnue avec son type et son label, et d'autre part chaque relation extraite avec son sujet et son objet. Par exemple :

```
smilk:Txt1 schema:about skp:Summer ; rdfs:comment "....." .

skp:Summer a gr:ProductOrService ; rdfs:label "Summer" ;
           pv:hasComponent skc:ananas ;
           pv: hasFragranceCreator skf:Anne_Flipo ;
           pv: hasFragranceCreator skf:Carlos_Benam .

skc:ananas a pv:Component ; rdfs:label  "ananas" .

skf:Anne_Flipo a foaf:Person ; rdfs:label  "Anne Flipo" .

skf:Carlos_Benam a foaf:Person ; rdfs:label "Carlos Benam" .
```

4 Mise en oeuvre et évaluation

4.1 Corpus et base de connaissance

Nous avons construit deux corpus pour tester notre approche :

— Le **corpus L'Oréal** est issu de l'outil Factiva et contient des articles journalistiques contenant la mention L'Oréal. Ce corpus est composé de 392 phrases issues de différents journaux tels que aufeminin.com, Cosmétique Hebdo, Cosmétique Mag. Ces phrases peuvent contenir ou non des relations ProVoc. Ce corpus contient 84 relations de type `belongsToBrand`, 79 relations de type `hasComponent`, 38 relations de type `hasFragranceCreator` et 44 de type `hasRepresentative`.

— Le **corpus Web** est issu d'une extraction manuelle de phrases issues du Web par le biais du moteur de recherche Google. Nous avons cherché le nom d'un parfum dans Google, par la suite nous avons choisi les dix premiers liens après les annonces tels que des liens vers des sites commerciaux ou des blogs cosmétiques. Ce corpus est composé de 119 phrases. Il contient 69 relations de type `hasFragranceCreator`, 62 relations de type `hasComponent`, 81 relations de type `hasRepresentative` et 43 relations de type `belongsToBrand`.

Notre système a permis de générer 325 triplets à partir du *corpus L'Oréal*, et 988 triplets à partir du *corpus Web*. La pertinence de ces triplets est évaluée dans les sections suivantes.

4.2 Évaluation de l'extraction des entités nommées

Nous avons réalisé une première évaluation de l'extraction des EN, nous avons considéré un sous-ensemble du *corpus du Web* composé de 25 phrases et nous avons cherché à évaluer l'extraction des entités nommées de type Product, Brand, Personne et Component. Les phrases contiennent 31 entités de type Product, 19 entités de type Brand et 20 entités de type Component.

Type EN	Rappel	Précision	Outils
Product	0.53	0.89	Renco
	0.66	0.91	Renco + Wikipedia
	0.17	0.5	DBpedia Spotlight
Brand	0.47	0.9	Renco
	0.73	0.91	Renco + Wikipedia
	0.61	0.84	DBpedia Spotlight
Component	0.55	0.73	Renco
	0.68	0.92	DBpedia Spotlight

TABLE 2 – Évaluation de l'extraction des entités nommées

Nous avons constaté que Holmes permet d'extraire les entités nommées de type Personne avec une valeur de rappel et de précision de 1. Pour les autres types d'EN, nous avons comparé la précision et le rappel de l'extraction d'EN avec Renco, avec Renco et les listes d'EN extraites de Wikipedia et avec DBpedia Spotlight [3], un outil de référence dans l'état de l'art de l'annotation sémantique de textes, qui repose sur DBpedia. Le tableau 2 présente les résultats obtenus. Sans surprise, pour l'extraction d'EN de type Product ou Brand, le rappel et la précision obtenus en complémentant Renco d'une liste d'EN extraite de Wikipedia sont meilleurs qu'en utilisant Renco seul. La valeur moyenne du rappel obtenue peut être expliquée par l'absence de certaines EN dans les pages Wikipedia exploitées et une tokenisation incorrecte de certaines EN par Renco. Par exemple, l'EN "La vie est belle" est considérée par Renco comme quatre tokens au lieu d'un seul. Comparés aux résultats obtenus avec DBpedia Spotlight, notre approche obtient de meilleures valeurs de rappel et de précision. Cependant, pour les EN de type Component, pour lesquelles une liste d'EN n'a pas été extraite de Wikipedia, DBpedia Spotlight obtient de meilleurs résultats.

4.3 Évaluation de l'extraction des relations

Le protocole de l'évaluation consiste à faire des expérimentations sur les deux corpus et dans le cas du corpus construit à partir de textes du Web à raffiner l'évaluation en considérant un sous-ensemble où la détection des entités nommées a été manuellement validée et au besoin corrigée ou complétée.

4.3.1 Évaluation de l'extraction de relations sur le corpus des articles du journaux

Le tableau 3 montre les résultats obtenus sur le corpus des articles du journaux. Plusieurs raisons peuvent expliquer la faiblesse de ces résultats :

— L'impossibilité de définir toutes les règles syntaxiques qui peuvent exprimer les relations à extraire.
— L'utilisation de règles syntaxiques seulement et l'absence de règles lexicales peuvent limiter l'approche de l'extraction des relations en diminuant la valeur de la précision. Pour

3. http ://demo.dbpedia-spotlight.org/

Propriétés	Méthode d'extraction	Rappel	Précision
belongsToBrand	signature des propriétés	0.17	0.46
hasComponent	Patrons syntaxiques	0.06	0.25
hasFragranceCreator	Patrons syntaxiques	0.21	0.20
hasRepresentative	Patrons syntaxiques	0.11	0.18

TABLE 3 – Évaluation de l'extraction de relations sur les articles de journaux spécialisés

Propriétés	Méthode d'extraction	Rappel	Precision
belongsToBrand	signature des propriétés	0.4	0.4
hasComponent	Patrons syntaxiques	0.2	1
hasFragranceCreator	Patrons syntaxiques	0.43	0.45
hasRepresentative	Patrons syntaxiques	0.3	0.52

TABLE 4 – Évaluation de l'extraction de relations sur un corpus d'articles sélectionnés sur le Web

certaines relations, une même règle syntaxique peut conduire à extraire deux relations différentes, comme par exemple `hasFragranceCreator` et `hasRepresentative`. Nous citons par exemple la phrase suivante « Monica Bellucci reste l'égérie de Rouge Dior sous l'objectif de Tyen. ». En réalité, cette phrase représente la relation `hasRepresentative` entre le produit Rouge Dior et l'égérie « Monica Belluci », alors que notre approche permet d'extraire en plus une relation de type `hasFragranceCreator` entre les mêmes entités.

— La non détection de certaines EN. Dans notre approche, nous utilisons DBpedia et Wikipedia pour détecter les entités nommées de type produit ou marque dans le cas où Renco ne peut pas les identifier, mais Wikipedia et DBpedia peuvent ne pas contenir toutes les marques et noms du parfums. Par exemple les parfums « Spicebomb » et « the essence » n'y figurent pas. De plus, l'expression du nom d'un produit différemment dans le texte et dans Wikipedia ou DBpedia empêche la détection. Par exemple, un même produit est représenté par « 1 Million » dans le corpus de textes et par « one Million » dans Wikipedia dont le token "one" est reconnu comme un nom propre et pas un nombre.

4.3.2 Évaluation de l'extraction de relations à partir d'articles sélectionnés sur le Web

Le tableau 4 présente les résultats de l'extraction des relations sur le corpus construit en sélectionnant des articles sur le Web. Les mêmes raisons citées dans l'évaluation des articles des journaux peuvent aussi expliquer les résultats de cette évaluation. De plus, nous avons supposé dans notre approche qu'une règle syntaxique doit impérativement contenir un verbe, alors que ce corpus contient des phrases sans verbe comme par exemple « Theo James, nouvel ambassadeur des parfums Hugo Boss. »

Afin d'évaluer la qualité de l'extraction de relations indépendamment de la qualité de la reconnaissance d'EN, nous avons considéré 45 phrases du corpus collecté sur le Web avec Google,

Propriétés	Méthode d'extraction	Rappel	Précision
belongsToBrand	signature des propriétés	0.625	0.5
hasFragranceCreator	Patrons syntaxiques	0.57	0.31
hasRepresentative	Patrons syntaxiques	0.32	0.69

TABLE 5 – Évaluation de l'extraction de relations sur un sous-ensemble du corpus d'articles sélectionnés sur le Web où les EN sont bien reconnues

pour lesquelles la reconnaissance des EN était correcte. Le tableau 5 présente les résultats de cette évaluation.

4.4 Synthèse

Le rappel et la précision sur le corpus des textes du Web sélectionnées avec Google sont meilleurs que ceux correspondants sur le corpus des articles de journaux. Cela peut être expliqué par la richesse des structures grammaticales dans les articles des journaux comparativement aux phrases collectées sur le Web qui expriment en général une représentation simple du produit avec donc une forme syntaxique plus ou moins régulière. Par exemple, dans le corpus du Web, la relation `hasComponent` est généralement exprimée avec un lien direct entre le parfum et le composant comme la phrase « Gentelmen Only Absolute contient notamment de la canelle, du safran et de la muscade ». Alors que dans les articles des journaux, cette relation est exprimée d'une manière plus compliquée comme « Alors que l'Agence nationale de sécurité du médicament et des produits de santé déconseille pour les enfants les produits contenant plus de 4 g de phénoxyéthanol par kilo et que la recommandation du Comité scientifique pour la sécurité des consommateurs est de ne pas dépasser 2,48 g de propylparaben par kilo, l'UFC relève des taux de 2,68 g de propylparaben et de 8,02 g de phénoxyéthanol dans le gel douche Nivea Water Lily Oil »

Le rappel et la précision de la reconnaissance des relations sur le sous-ensemble du corpus extrait du Web où la reconnaissance des entités nommées a été validée sont significativement supérieurs à ceux obtenus sur l'ensemble du corpus avec une reconnaissance d'EN automatique. En d'autres termes, la qualité de la reconnaissance des EN influe significativement sur celle de l'extraction des relations et donc des performances de l'approche dans son ensemble.

5 Conclusion

Nous avons proposé dans cet article une approche d'extraction de relations à partir de textes dans le domaine de la cosmétique, qui combine une approche semi-supervisée utilisant les signatures des propriétés déclarées dans l'ontologie du domaine Provoc, avec une approche supervisée utilisant des patrons syntaxiques définis manuellement en analysant un corpus de textes préalablement annoté manuellement. Le résultat est la production d'une base de connaissances formalisée en RDF. Nous avons évalué notre approche sur deux corpus différents, un corpus d'articles de journaux spécialisés dans le domaine de la cosmétique, similaire au corpus analysé pour produire manuellement les règles syntaxiques d'extraction, et un corpus de textes du Web sélectionnées avec Google.

Les résultats obtenus montrent bien l'intérêt d'une approche mixte dès lors qu'un domaine particulier est ciblé, et la corrélation qui existe entre qualité de l'extraction des propriétés et qualité de la reconnaissance d'entités nommées. Les pistes naturelles d'amélioration de cette approche sont d'une part l'amélioration de la reconnaissance d'entités nommées et d'autre part l'ajout d'une dimension lexicale et sémantique aux patrons syntaxiques. Une autre perspective de ce travail est de comparer les résultats ainsi obtenus avec ceux que l'on obtiendrait avec des outils d'apprentissage automatique statistique.

Références

GERBER D., HELLMANN S., BUHMANN L., SORU T., USBECK R. & NGOMO A.-C. N. (2013). Real-time rdf extraction from unstructured data streams. In *International Semantic Web Conference*, p. 135–150 : Springer.

GERBER D. & NGOMO A.-C. N. (2011). Bootstrapping the linked data web. In *1st Workshop on Web Scale Knowledge Extraction@ ISWC*, volume 2011.

HEPP M. (2008). Goodrelations : An ontology for describing products and services offers on the web. In *International Conference on Knowledge Engineering and Knowledge Management*, p. 329–346 : Springer.

KUMAR K. & MANOCHA S. (2007). Constructing knowledge graph from unstructured text. *Self*, **3**, 4.

LOPEZ C., NOORALAHZADEH F., CABRIO E., SEGOND F. & GANDON F. (2016). Provoc : une ontologie pour d'écrire des produits sur le web. In *IC2016 : 27es Journees francophones d'Ingenierie des Connaissances*.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P. & DINI L. (2014). Generating a resource for products and brandnames recognition. application to the cosmetic domain. In *LREC*, p. 2559–2564.

NEBHI K. (2013). A rule-based relation extraction system using dbpedia and syntactic parsing. In *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064*, p. 74–79 : CEUR-WS. org.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail-Toulouse II.