Identification des catégories de produits issus de catalogues publicitaires

Céline Alec [1], **Chantal Reynaud-Delaître** [1], **Brigitte Safar** [1], **Zied Sellami** [2], **Uriel Berdugo** [3]

1. *LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France*
   {celine.alec,chantal.reynaud,brigitte.safar}@lri.fr
2. *Linagora, 100 Terrasse Boieldieu - Tour Franklin, Paris - La Défense, France*
   zsellami@linagora.com
3. *Wepingo, 6 Cour Saint Eloi, Paris, France*
   uriel.berdugo@wepingo.com

RÉSUMÉ. *Nous proposons dans cet article une approche d'extraction d'informations, basée sur une ontologie, et appliquée à des documents issus de catalogues publicitaires. Les documents des catalogues sont des descriptifs de produits relativement pauvres. Les informations à extraire, ou annotations, concernent les catégories et les caractéristiques des produits considérés, répertoriées dans une ontologie de domaine. L'extraction d'informations concernant un produit consiste en un peuplement de cette ontologie, plus précisément le peuplement des concepts représentant ses catégories et ses caractéristiques. La relative pauvreté des descriptifs rend irréalisable un peuplement totalement automatique. Nous proposons donc une approche en deux étapes : (1) une première étape d'annotation semi-automatique qui porte sur un petit ensemble de documents ; (2) une deuxième étape qui annote l'ensemble des autres documents de façon entièrement automatique, en s'appuyant sur des mécanismes d'apprentissage automatique exploitant les résultats de la première étape. L'originalité de ce travail consiste en une approche incrémentale de raffinement des informations extraites. Le travail décrit a été appliqué sur des jeux de données réelles concernant des jouets.*

ABSTRACT. *In this paper, we propose an approach of information extraction, based on an ontology, and applied to documents from advertising catalogs. Documents are relatively poor descriptions of products. The information to be extracted, or annotations, concern the categories and features of the products, listed in a domain ontology. Thus, the information extraction about a product is actually an ontology population process, more precisely the population of concepts representing its categories and features. The poverty of the descriptions makes a fully automatic population impossible. We propose a two-step approach: (1) a first semi-automatic annotation step, which covers a small set of documents; (2) a second step, which annotates all other documents, in an entirely automatic way, based on machine learning mechanisms exploiting the results of the first step. The originality of this work relies on an incremental approach*

Revue d'intelligence artificielle – n° 5/2016, 557-578

to refine the extracted information. The work described has been applied on real data, in the toy domain.

MOTS-CLÉS : *extraction d'informations, peuplement d'ontologie, annotation sémantique, application dans le domaine du e-commerce.*

KEYWORDS: *information extraction, ontology population, semantic annotation, B2C application.*

DOI:10.3166/RIA.30.557-578

1. Introduction

Beaucoup de produits et d'informations sont disponibles aujourd'hui sur Internet, dans les applications B2C (*Business to Consumer*), mais le volume et la variété des sources complexifient l'accès, pour le consommateur, au produit souhaité. Les produits pouvant provenir de différents fournisseurs, avec des descriptions différentes et de qualité variable, il est crucial de trouver des techniques efficaces d'intégration de données, aussi automatiques que possible. Les technologies du web sémantique, et les ontologies en particulier, peuvent être considérées comme des solutions prometteuses pour résoudre ce type de problème. Une ontologie est une conceptualisation d'un domaine particulier (Gruber, 1993). Elle représente des concepts, des attributs et des relations entre les concepts.

Dans cet article, nous soutenons l'idée que les ontologies peuvent être utilisées comme support à l'intégration de connaissances issues de descriptions hétérogènes de produits, en facilitant l'extraction des connaissances explicitées dans les documents, et en jouant un rôle d'intermédiaire entre les souhaits des consommateurs et les produits présentés par les fournisseurs. Nous utilisons une ontologie spécifique, dans laquelle chaque concept dénote une catégorie ou une caractéristique de produits correspondant aux souhaits exprimés dans les recherches des consommateurs. Étant donné un descriptif de produit extrait d'un catalogue, notre approche consiste à identifier les concepts de l'ontologie dont le produit décrit est une instance.

L'appariement entre un élément d'un catalogue et les concepts d'une ontologie est lié au problème de peuplement d'ontologie. Bien que de multiples approches aient été proposées (Petasis *et al*., 2011), aucune n'a, à notre connaissance, été évaluée sur des instances ne disposant que de descriptions très pauvres et très peu contextualisées, alors que dans l'application que nous considérons, l'identification dans les textes des concepts pertinents de l'ontologie ne peut s'appuyer que sur très peu d'éléments. Nous proposons une approche permettant d'annoter des descriptions de produits de façon automatisée, les produits ainsi annotés étant ensuite introduits dans l'ontologie comme des instances des concepts les annotant. Les consommateurs pourront de ce fait, en utilisant les termes de l'ontologie pour expliciter leurs souhaits, accéder aux instances des concepts correspondant aux produits les satisfaisant.

L'originalité de notre approche repose sur sa capacité à générer et à progressivement raffiner des annotations, même à partir de descriptifs restreints et peu précis.

L'approche se décompose en deux étapes : dans la première étape, les annotations d'un sous-ensemble de documents sont établies par raffinements progressifs de façon automatique puis ces annotations sont validées/corrigées par l'intermédiaire d'une interface adaptée, par le concepteur du système qui doit être un expert du domaine. Cette première étape n'est donc que semi-automatique. Lors de la deuxième étape, des techniques d'apprentissage utilisent le premier ensemble d'annotations validées pour construire des classifieurs associés à chacun des concepts de l'ontologie. Ces classifieurs permettent d'annoter automatiquement le reste des catalogues. Cette approche est indépendante d'un catalogue particulier ou d'un domaine mais elle est plus particulièrement adaptée à des ontologies représentant des classifications de produits et leurs caractéristiques.

Notre travail est motivé par des besoins applicatifs spécifiques, dans le contexte d'une collaboration avec la start-up Wepingo [1] qui souhaite utiliser des technologies du web sémantique pour ses applications B2C. Les résultats que nous présentons sont ceux obtenus en travaillant avec l'ontologie de domaine et les catalogues de produits fournis par l'entreprise. Des expérimentations ont ainsi été faites dans le domaine des jouets.

Le reste de l'article est construit comme suit. Après avoir exposé le cadre de ce travail (section 2), nous faisons un rappel des travaux similaires (section 3). Nous détaillons notre approche (section 4) en présentant la génération progressive d'un premier ensemble d'annotations, leur validation par le biais de l'interface puis la phase d'apprentissage. Ces différentes étapes sont évaluées en section 5, puis nous concluons et énonçons quelques perspectives de travail (section 6).

2. Cadre de travail

Nous présentons dans cette section l'ontologie utilisée dans l'application et le type de documents qui doivent être annotés.

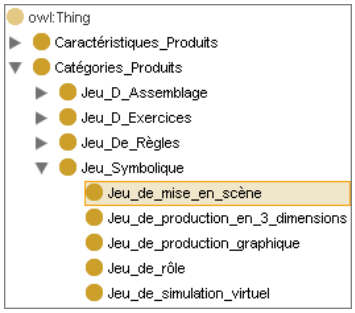
2.1. L'ontologie du monde des jouets

L'application de notre approche a porté sur le domaine des jouets. Comme support au système de recommandation, Wepingo a mis en place une ontologie des jouets (figure 1), basée sur la norme ESAR définie par des psychopédagogues (Garon *et al.*, 2002).

Cette norme identifie des catégories et des caractéristiques de jouets, en deux classifications indépendantes l'une de l'autre. Les catégories de jouets font référence au type de jouet (jeu de construction, jeu de hasard, ...) qu'elles organisent en 4 grandes catégories générales (Jeu_d_Exercice, Jeu_Symbolique, Jeu_d_Assemblage, Jeu_de_Règles d'où le nom ESAR) et les caractéristiques, aux valeurs éducatives transmises par un

1. www.wepingo.com

jeu (concentration, dextérité, ...) ou encore ses conditions d'utilisation (jeu coopératif, associatif, ...). Un exemple de catégorie (Jeu_de_mise_en_ scène) spécialisant la catégorie générale Jeu_Symbolique, est présenté en figure 1.



Concept (*Label*)	Jeu de mise en scène
Définition	Jeu de faire semblant dans lequel le joueur est le metteur en scène. Il réalise des scénarios élaborés dans le but de reproduire des thèmes particuliers, des scènes précises, des événements, des métiers, etc. Ces formes de jeux exigent de pouvoir mettre en scène les accessoires pertinents au contexte ou à la situation représentée.
Exemples (Ex)	playmobil, marionnette, figurine, ...

Figure 1. L'ontologie ESAR et le concept Jeu_de_mise_en_scène

L'ontologie ESAR est définie par l'ensemble de ses composants $O_{ESAR} = (C_{ESAR}, L_{ESAR}, H_{ESAR}, Att_{ESAR}, A_{ESAR}, I_{ESAR})$. C_{ESAR} est l'ensemble de concepts, composé de 33 catégories et de 129 caractéristiques. Le lexique L_{ESAR} est composé d'un ensemble d'entrées lexicales pour les concepts et est muni d'une fonction de référence $F : 2^L \mapsto 2^C$, qui associe des ensembles de concepts à des ensembles d'entrées lexicales. Le lexique est composé de deux sous-ensembles de termes : $Label$ et Ex. Chaque concept $c \in C_{ESAR}$ est associé à au moins un label. Ex regroupe des exemples représentatifs de certains concepts feuilles, i.e. certains des concepts les plus précis de l'ontologie (cf figure 1). On notera $L_{ESAR}(c)$ l'ensemble des termes du lexique L_{ESAR} dénotant le concept c. H_{ESAR} est l'ensemble des relations de subsomption entre les concepts. Att_{ESAR} est l'ensemble des attributs caractérisant les concepts, restreint, dans cette ontologie à l'unique attribut *définition*. L'ensemble des axiomes A_{ESAR} est initialement vide. Aucune relation du domaine ne décrit les liens entre catégories et caractéristiques et ces deux classifications comportent très peu de relations de subsomption. L'ensemble des instances des concepts I_{ESAR} est aussi initialement vide et notre objectif est de le peupler avec les instances de produits décrits dans les catalogues.

2.2. Les documents à annoter

Les documents (notés *Corpus*) sont des fiches décrivant un jouet par son label, sa marque, sa description, qui est le plus souvent un texte court et non contextualisé, et sa catégorie. La catégorie représentée ici ne recouvre pas les mêmes informations que celle de l'ontologie. Dans les documents, la catégorie varie beaucoup suivant le vendeur. Elle peut être très générale ("Jouet", "Jeux"), comme très spécifique ("HABA cubes et perles à assembler", "Briques"), parfois difficilement interprétable ("Bosch", "Couleurs unies"). Un exemple de descriptif de jouet est présenté figure 2. Les formes

```
<jouet>
    <nom>PLAYMOBIL 4221 Ambulanciers / blessé / véhicule</nom>
    <marque>PLAYMOBIL</marque>
    <categorie>PLAYMOBIL La vie en ville</categorie>
    <description>"Set de jeu Ambulanciers / blessé / véhicule (PLAYMOBIL n Grad 4221) sur
    le thème ""Les sauveteurs"". L'ambulance doit foncer :  un garçon est tombé à vélo et
    s'est blessé au genou. Une fois sur place, les secouristes visualisent la situation.
    Ils apportent les premiers soins au garçon et l'installent sur la civière. Le blessé
    est transporté à l'hôpital pour être examiné. La jambe pourrait bien être cassée !
    Caractéristiques :  - Gyrophare (fonctionne avec 2 piles CR 2032 V fournies) - Le
    toit est amovible, les portes peuvent être ouvertes - Pieds de la civière rabattable,
    dossier ajustable  - Dimensions (L x P x h) : 27 x 13 x 15 cm - Figurines : - 1 homme,
    1 femme, 1 robot humanoïde - Accessoires : matériel d'intervention médicale, 1
    téléphone portable, 1 civière, 1 plâtre, 1 ambulance et de nombreux autres accessoires"
    </description>
</jouet>
```

Figure 2. Exemple de descriptif de jouet

et les contenus de ces descriptions sont très éloignés des définitions des concepts de la norme ESAR.

3. État de l'art

L'extraction d'informations est un vaste champ du web sémantique. Grâce aux ontologies, des informations d'un domaine particulier peuvent être extraites à partir de documents. Les méthodes d'extraction diffèrent selon la richesse de l'ontologie. Nous nous focaliserons ici sur les méthodes adaptées aux ontologies légères, i.e. où les différents concepts ne sont reliés que par des relations de subsomption, comme dans notre contexte. Le lecteur intéressé par les méthodes s'appuyant sur des ontologies plus riches, pourra par exemple consulter les travaux cités dans (Petasis *et al.*, 2011).

Annoter un document avec une ontologie consiste à rechercher dans celui-ci les fragments de texte mentionnant des concepts ou des instances de concepts appartenant à l'ontologie puis à associer ces mentions aux concepts considérés. Les méthodes d'annotation sémantique peuvent être classées en deux catégories (Reeve, 2005) : la première regroupe les méthodes basées sur des patrons linguistiques, qui peuvent être soit découverts automatiquement soit manuellement définis; la deuxième regroupe celles basées sur un apprentissage automatique (Manning, Schütze, 1999), utilisant soit des modèles statistiques pour repérer des entités dans les textes soit de l'induction. Nous nous focalisons ici sur les approches linguistiques.

Divers travaux d'annotation et d'extraction d'informations ont été proposés sur des domaines spécifiques. Beaucoup de ces outils, comme KIM (Popov *et al.*, 2004) ou SOFIE (Suchanek *et al.*, 2009) extraient des groupes nominaux spécifiques correspondant à des entités nommées, i.e. des noms de personnes, de lieux, d'organisations,..., repérables grâce à des grammaires formelles associées à des modèles statistiques et répertoriées dans des bases de connaissances ou des "gazeteers" (ressource terminologique) (Bontcheva *et al.*, 2004). D'autres approches linguistiques comme (Barriere, Agbago, 2006) essaient d'extraire des entités nommées ou d'autres éléments en uti-

lisant éventuellement des ressources sémantiques additionnelles telles que des glossaires, des dictionnaires ou des bases de connaissances.

L'identification d'instances qui ne sont pas des entités nommées est beaucoup plus délicate car aucune base ne répertorie a priori l'ensemble des instances à reconnaître et encore moins les expressions linguistiques qui leur sont associées. Ces ensembles d'instances et la terminologie propre au domaine doivent donc être recueillies pour construire l'ensemble des termes qui doivent être recherchés correspondant à un domaine particulier. Par exemple, (Amardeilh, Damljanovic, 2009) prétraitent l'ensemble des termes présents dans les différentes ressources d'une ontologie (classes, instances, propriétés, valeurs de propriétés) pour en extraire un ensemble de lemmes à partir desquels est constituée la ressource terminologique associée à cette ontologie.

D'autres approches exploitent la structure du document à annoter. Par exemple, dans (Amardeilh *et al*., 2005), la structure d'un document est représentée sous la forme d'un arbre conceptuel dont chaque nœud est mis en correspondance avec un concept de l'ontologie via des règles définies manuellement. De même, (Aussenac-Gilles *et al*., 2013) définissent des règles d'extraction en exploitant la structure hiérarchique exprimée par les marqueurs typo-dispositionnels (police gras, italique, symbole de ponctuation ':') au sein d'un ensemble de fiches de même format.

Les travaux cités précédemment relèvent directement du domaine de l'extraction d'informations et de l'annotation de documents. Notre objectif est sensiblement différent puisqu'il s'agit plutôt d'interpréter un document décrivant un produit comme un tout, en essayant de le rapprocher de la définition d'un concept de l'ontologie. Quand ce sera possible, le produit décrit pourra être considéré comme une instance du concept concerné. D'autres travaux, a priori plus éloignés des tâches d'extraction ou d'annotation, sont ainsi intéressants pour notre problématique bien qu'ils ne soient pas réalisés dans ce contexte précis. Ainsi, l'objectif de (Kessler *et al*., 2012) est de vérifier l'adéquation entre des candidatures à des offres d'emploi (CV et lettres de motivation) et les offres d'emploi considérées, c'est-à-dire évaluer la proximité entre la description d'un élément général (une offre d'emploi ou un concept d'une ontologie) et celles d'éléments plus spécifiques (des candidatures ou des instances de concept). Après avoir été soumis à différents traitements, tous les documents manipulés sont représentés par des vecteurs qui sont ensuite comparés en utilisant des combinaisons de diverses mesures de similarité (cosinus, Minkowski, ...) afin de classer les candidatures. De plus, pour être sûr de ne pas écarter trop vite une candidature, on évalue aussi sa similarité avec le vecteur représentant l'offre d'emploi enrichie des candidatures jugées pertinentes par un recruteur.

Enfin, dans (Béchet *et al*., 2012), l'objectif est de peupler automatiquement une structure hiérarchique de concepts décrivant des services hôteliers, en s'appuyant sur un premier ensemble d'instances identifié par un expert. Les différents services de chaque hôtel, définis par chaque hôtelier avec son propre vocabulaire doivent être comparés aux instances initiales. Un service sera considéré comme une instance du concept correspondant à l'instance dont il est le plus proche suivant un calcul de similarité basé sur les n-grammes.

Dans notre contexte, il faut annoter des descriptions de produits comme des instances de concepts. Pour cela, il faut identifier des instances de concepts qui ne sont pas des entités nommées, au sein de documents non structurés et sans aucune homogénéité. Un certain nombre des techniques présentées précédemment, en particulier celles qui s'appuient sur la recherche d'entités nommées ou de régularités dans la structure des documents sont donc complètement inadéquates. L'approche consistant à évaluer directement la proximité entre la description d'un concept et celle d'une instance est, elle aussi, inapplicable car les définitions des concepts sont très éloignées des descriptions des produits et leur rapprochement avec des mesures de similarité a été tenté mais n'a donné aucun résultat. En revanche, nous allons utiliser une méthode vectorielle de ce type, pour rapprocher les descriptions de produits, des descriptions d'un premier ensemble d'instances annotées, suivant ainsi l'approche proposée dans (Béchet *et al.*, 2012). Pour l'annotation de ces premières instances, nous nous sommes inspirés des travaux qui s'appuient sur des termes préalablement identifiés dans des ressources adaptées au domaine (Amardeilh, Damljanovic, 2009) ou dans la composante terminologique d'une Ressource Termino-Ontologique (RTO), i.e. où l'ontologie a été explicitement enrichie d'informations lexicales (Reymonet *et al.*, 2007). Ces premières annotations devront ensuite être validées par un expert avant d'être utilisées comme base d'exemples pour les classifieurs.

4. Proposition d'une approche de peuplement d'ontologie

L'approche de peuplement de l'ontologie consiste à générer une base de connaissances $BC(O, I, W)$ à partir de l'ontologie O avec $W : 2^I \mapsto 2^C$, une fonction *membre* qui associe, à des ensembles d'instances appartenant à I, les ensembles de concepts de C dont ils sont membres.

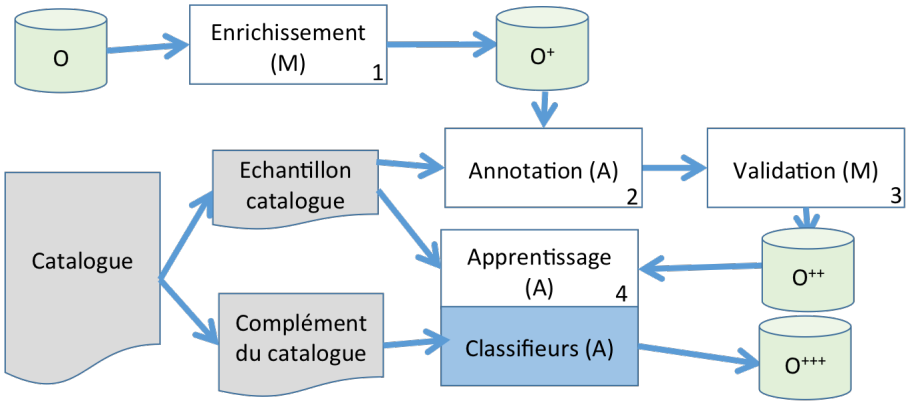


Figure 3. L'approche proposée : (A) automatique, (M) manuel

Un peuplement automatique même partiel n'est possible que si l'ontologie contient les formes linguistiques associées aux concepts dont nous voulons reconnaître des

instances. Notre proposition (cf. figure 3) consiste donc, dans un premier temps, à enrichir l'ontologie O de connaissances complémentaires (étape 1 sur la figure 3). Cette étape peut être vue comme une phase préliminaire. L'ontologie enrichie (O^{+}) est ensuite utilisée pour annoter de façon semi-automatique un échantillon de documents (étapes 2 et 3) et peupler partiellement l'ontologie avec les instances ainsi annotées (O^{++}). Enfin, des techniques d'apprentissage automatique (étape 4) exploitent ces premières annotations et construisent les classifieurs qui seront utilisés sur l'ensemble du corpus de documents à annoter pour peupler complètement l'ontologie (O^{+++}).

L'approche d'annotation proposée comporte donc trois phases qui seront successivement décrites : la phase préliminaire d'enrichissement de l'ontologie (section 4.1), la phase 1 qui regroupe l'annotation automatique d'un échantillon de documents (section 4.2) et la validation manuelle de ces annotations (section 4.3), puis la phase 2 d'annotation du corpus complet de documents (section 4.4) basée sur l'application de techniques d'apprentissage automatique. Ces différentes étapes, appliquées sur le domaine des jouets, sont détaillées dans les sections suivantes.

4.1. Enrichissement de l'ontologie ESAR

L'ontologie O_{ESAR} telle qu'elle nous a été communiquée initialement était très limitée. La composante terminologique L_{ESAR} était très pauvre et ne contenait pas l'ensemble des termes nécessaires pour reconnaître les catégories de jouets évoquées dans les descriptions et encore moins ceux permettant de reconnaître les caractéristiques. De plus, à part un nombre limité de liens de subsomption, aucune relation du domaine ne décrivait les liens entre les différents concepts que ce soit entre les différentes catégories, les différentes caractéristiques ou entre ces deux types de concepts. Des connaissances complémentaires nécessaires au processus d'annotation ont donc été ajoutées par des experts maîtrisant la norme ESAR mentionnée en section 2.1., pour enrichir les composantes terminologique L_{ESAR} et axiomatique A_{ESAR}. Il est clair que ces différents ajouts ont demandé un travail important aux experts. Ce travail est difficile à quantifier exactement, et nous souhaitons ici surtout qualifier les différents types de connaissances introduites de façon à pouvoir y faire référence dans la suite et préciser comment elles sont utilisées.

4.1.1. Enrichissement de L_{ESAR}

Pour enrichir L_{ESAR}, la composante terminologique de l'ontologie, des exemples extraits de ressources externes ont été ajoutés aux exemples de Ex. Ces ajouts correspondent d'une part, à des noms de jouets provenant d'un site internet dédié [2] et ayant utilisé la classification ESAR pour décrire ses jouets et, d'autre part, à une liste de sports provenant du site Wikipedia.

Nous avons par ailleurs ajouté de nouveaux éléments terminologiques : des signes linguistiques (SL) et des signes linguistiques complexes ($SLcomp$). Les signes lin-

2. http://www.jeuxrigole.com/liste-des-jeux.html

guistiques sont des termes ou expressions évocateurs de concepts (par exemple musical et parlant pour le concept Jeu_sensoriel_sonore). Un terme considéré de façon isolée pouvant ne pas être suffisant pour différencier les concepts évoqués, nous avons introduit les signes linguistiques complexes, représentés sous la forme de patron : terme ET [NON] terme ET [NON] terme ..., pour faciliter la désambiguïsation.

Par exemple, il existe deux types de jeux de domino : les dominos numérotés de 1 à 6, que les joueurs doivent associer (jeu_d'_association), et les dominos de construction, sans signes distinctifs, où le jeu consiste à construire un parcours avec des dominos debout, puis à pousser le premier qui va pousser les autres (jeu_de_construction). L'utilisation de signes complexes permet de différencier ces deux jeux de domino : le jeu de construction sera évoqué par la présence conjointe des termes domino et construction alors que le jeu d'association le sera par la présence du terme domino et l'absence du terme construction.

Les exemples et les signes linguistiques étant des connaissances de nature différente, nous les avons distingués dans la représentation, mais le processus d'annotation les exploite de la même façon. Après enrichissement, $L_{ESAR} = \{Label \cup Ex \cup SL \cup SLcomp\}$.

4.1.2. Enrichissement de A_{ESAR}

Les axiomes introduits dans A_{ESAR} sont de deux types.

1) Le premier type correspond à des connaissances fiables qui s'appliquent avec un degré de probabilité élevé. Nous pourrions trouver des jouets très particuliers sur lesquels ces connaissances ne s'appliqueraient pas, mais ils constitueraient des cas exceptionnels. Ces axiomes sont exprimés sous la forme de règles propositionnelles de deux types. Il s'agit :

– d'expressions d'incompatibilités entre concepts, donnant la priorité à l'un d'eux. Ces règles sont de la forme : SI concept A ET concept B ALORS NON concept A, (30 règles). Par exemple, selon leur définition cf figure 1, les Jeu_de_mise_ en_scène, s'appuient sur l'usage de figurines et d'accessoires de taille réduite adaptée aux figurines, qui peuvent être mis en scène par le joueur. Même si un des accessoires cités dans la description du jouet (comme un vélo, un ballon, un arc) est aussi évocateur d'un autre type de jeu (Jeu_Sportif, Jeu_d_adresse), les annotations correspondantes ne doivent pas êtres prises en compte et seule l'annotation Jeu_de_mise_ en_scène doit être considérée.

– d'expressions de dépendances de concepts traduisant des inclusions ou des relations manquantes entre concepts. Ces règles sont de la forme : concept A IMPLIQUE concept B, (95 règles). Par exemple, "Jeu _associatif_et_coopératif IMPLIQUE Jeu_associatif" ou "Différenciation_de_texture IMPLIQUE Discrimination_tactile".

2) Le deuxième type d'axiome correspond à des connaissances heuristiques permettant de relier les deux types de concept de l'ontologie, i.e., les catégories et les caractéristiques, et d'inférer des caractéristiques potentielles à partir des catégories. Une partie de ces règles dites d'*implications potentielles* a été automatiquement géné-

rée à partir des exemples et des signes linguistiques partagés par les catégories (Cat) et les caractéristiques (Car), comme suit :

$\forall\, cat_i \in Cat, \forall\, car_k \in Car$,
Si $\exists v \in L_{ESAR}(cat_i)$ tel que $v \in L_{ESAR}(car_k)$,
alors $creer\ la\ regle :\ cat_i \underset{potentiellement}{\Rightarrow} car_k$.

Par exemple, la catégorie Jeu_d'adresse implique potentiellement la caractéristique Coordination_œil-main car ces deux concepts partagent l'exemple Toupie. 72 caractéristiques sur 129 ont été associées à au moins une catégorie et l'ensemble des règles obtenues a été ensuite complété manuellement (476 règles). Par exemple, la caractéristique Concentration a été introduite par les experts comme impliquée potentiellement par toutes les catégories de Jeu_d_Assemblage ou de Jeu_de_Règles (16 règles) et la caractéristique Causalité_Sensori_Motrice l'a été pour tous les Jeu_d_Exercice (8 règles).

Les différents exemples de règles introduites en complément par les experts et présentées ici, montrent qu'à part les expressions d'incompatibilités entre concepts, qui sont des connaissances propres à l'approche d'annotation que nous proposons et qui peuvent être trouvées en analysant les exemples, les autres règles correspondent à des connaissances du domaine qui auraient dû être présentes dans l'ontologie initiale.

4.2. Phase 1 : annotation initiale d'un échantillon de documents représentatifs du domaine

La génération des annotations est une chaîne de traitements qui vise à trouver le maximum possible d'annotations candidates exactes pour un jouet donné (catégories comme caractéristiques). Elle est composée de 3 étapes :

1. l'établissement d'un premier ensemble d'annotations candidates définissant le contexte d'interprétation d'un jouet ;
2. la recherche d'incohérences qui détecte, au sein du contexte d'interprétation, les annotations incompatibles et effectue un choix parmi elles ;
3. la complétion qui complète l'ensemble des annotations candidates en prenant en compte les relations d'implication entre concepts.

4.2.1. Génération d'un premier ensemble d'annotations

La génération d'annotations des fiches jouets s'appuie, pour chaque concept c, sur l'ensemble $lemme(c)$ des lemmes du lexique du concept c considéré $L_{ESAR}(c)$. De même, on garde pour chaque jouet j appartenant au $Corpus$, l'ensemble $info(j)$ composé des lemmes des informations disponibles sur le jouet, i.e. son nom, sa marque, sa catégorie et sa description :

$\forall c \in C_{ESAR}, lemme(c) = lemmatisation(L_{ESAR}(c))$

$\forall j \in Corpus, info(j) = lemmatisation\{Nom(j) \cup Marque(j) \cup Cat(j) \cup Desc(j)\}$

La génération des annotations est une opération de recherche d'inclusion de mots qui consiste à rechercher si un élément de $lemme(c)$ d'un concept c apparaît dans l'ensemble des informations d'un jouet j, et dans ce cas, à annoter le jouet j par le concept c (catégorie ou caractéristique) considéré :

$\forall j \in Corpus, \forall c \in C_{ESAR}$,
Si $\exists v \in lemme(c)$ tel que $v \in info(j)$ alors j `instanceOf` c.

Pour le traitement des signes linguistiques complexes, et qui contiennent des négations, on appelle *Termes négatifs* les termes précédés de NON, *Termes positifs* les autres termes et on considère qu'un jouet j contient un signe linguistique complexe slc d'un concept si

$\forall\, tp \in TermesPositifs(slc), \forall\, tn \in TermesNegatifs(slc),$
$tp \in info(j)\ et\ tn \notin info(j)$

Les premières annotations produites dans cette phase définissent le contexte d'interprétation d'un jouet j, comme suit: $Ctxt(j) = \{c \mid j$ `instanceOf` $c\}$.

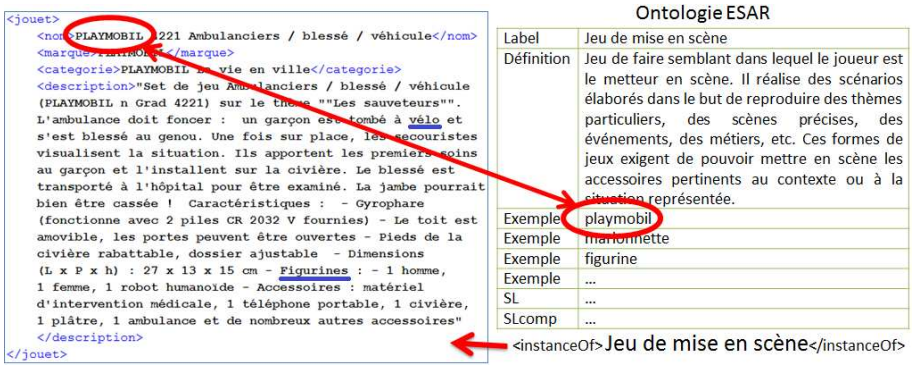


Figure 4. Exemple d'annotation

Par exemple, le descriptif du jouet de la figure 4 contient le terme playmobil qui est un exemple du concept Jeu_de_mise_en_scène. Ce jouet est donc annoté avec le concept Jeu_de_mise_en_scène [3]. De même, le terme vélo permet de l'annoter comme Jeu_moteur et le terme figurines permet de rajouter la catégorie jeu de mise en scène et les caractéristiques Créativité_expressive, Reproduction_de_rôles et Reproduction_d'évènements.

Ce contexte est ensuite plus facile à analyser par les étapes suivantes que le contenu non structuré des descriptions textuelles. Pour mettre en œuvre les étapes suivantes,

3. Remarquons que le fait de trouver un terme exemple d'un concept dans le nom du jouet ne suffit pas à le classer directement et définitivement comme instance du concept. Par exemple, le jouet Playmobil pirates interactif, qui serait également annoté comme un jeu_de_mise_en_scène, devra finalement être reconnu comme une instance de jeu_de_simulation_virtuelle.

nous avons introduit différents ensembles de règles, chaque ensemble s'appliquant sur les résultats obtenus à la phase précédente.

4.2.2. Phase de recherche d'incohérences

La phase de recherche d'incohérences est un processus de raffinement dont le but est de détecter et d'éliminer des concepts erronés du contexte d'interprétation d'un jouet. Cette phase vise donc à améliorer la **précision** des résultats. Elle consiste à appliquer sur le contexte les règles d'incompatibilité introduites au cours de l'enrichissement. En effet, le contexte peut contenir des concepts multiples dont certains doivent être éliminés en présence d'autres. À l'issue de cette phase, on obtient un ensemble d'annotations A_1 tel que $A_1(\text{j}) \subset Ctxt(j)$.

Par exemple, le jouet de la figure 4 a été précédemment annoté comme Jeu_moteur car sa description contenait le terme vélo, alors qu'il ne s'agit pas ici d'un vrai vélo mais d'un vélo miniature associé à une figurine donc représentatif d'un Jeu_de_mise_en_scène. Dans ce contexte précis, l'annotation Jeu_moteur n'est pas adaptée et il est plus facile de s'en rendre compte en la confrontant avec l'annotation Jeu_de_mise_en_scène également présente dans le contexte qu'en cherchant à interpréter finement la description du jouet. L'application de la règle d'incompatibilité r_1 : SI Jeu_de_mise_en_scène ET Jeu_moteur ALORS NON Jeu_moteur permet de supprimer l'annotation inadaptée.

4.2.3. Phase de complétion

La phase de recherche d'incohérences vise à augmenter la précision des annotations et permet à la phase de complétion de s'appuyer sur des données les plus sûres possibles [4]. La complétion cherche à améliorer le **rappel** en exploitant toutes les inclusions entre concepts, qu'elles soient exprimées dans l'ontologie initiale ou enrichie. Elle permet d'identifier des annotations additionnelles non retrouvées lors de la phase de génération d'un premier ensemble d'annotations. À l'issue de cette phase, on obtient un ensemble d'annotations A_2 tel que $A_1(\text{j}) \subset A_2(\text{j})$.

Par exemple, connaissant les implications "Endurance IMPLIQUE Jeu_sportif" et "Jeu_sportif IMPLIQUE Jeu_moteur", un jouet annoté avec le concept Endurance sera par complétion également annoté par les concepts Jeu_sportif puis Jeu_moteur.

La figure 5 montre une application des phases de recherche d'incohérences et de complétion toujours sur le même exemple du jouet de la figure 4. La phase de recherche d'incohérences a supprimé l'annotation Jeu_moteur en appliquant la règle r_1, puis la phase de complétion ajoute les annotations Jeu_symbolique, Création_inventive et Imitation_différée.

4. Dans notre contexte, effectuer la phase de complétion avant la phase de recherche d'incohérences risquerait d'introduire beaucoup de concepts impliqués erronés mais pas forcément incohérents, qui ne seraient donc pas forcément tous détectés lors de la phase de recherche d'incohérences.

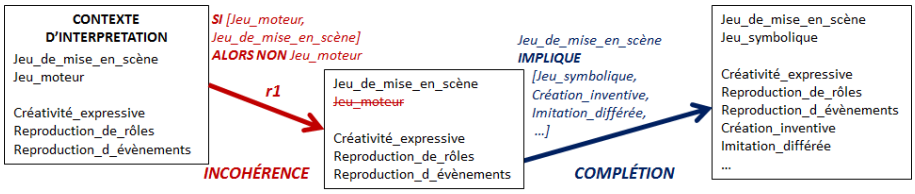


Figure 5. Phases de recherche d'incohérences et de complétion sur le jouet de la figure 4

Ces phases s'appliquent indifféremment aux catégories et caractéristiques mais dans la pratique elles ne permettent de trouver que peu d'annotations de caractéristiques car celles-ci font référence à des notions abstraites pour lesquelles les signes linguistiques sont limités. De ce fait, des étapes de raisonnement supplémentaires sont nécessaires pour déduire plus d'annotations de caractéristiques à partir des annotations de catégories trouvées dans les étapes précédentes. Le processus est basé sur deux heuristiques s'appuyant sur les annotations de catégories déjà reconnues.

La première heuristique consiste à identifier les caractéristiques communes aux jouets déjà annotés et validés par l'utilisateur et qui sont de la même (des mêmes) catégorie(s) que le jouet en cours d'étude. Ainsi, si un jouet j est d'une catégorie A et que tous les jouets de catégorie A précédemment classés partagent un ensemble de caractéristiques E, alors l'outil propose également d'annoter le jouet j avec les caractéristiques E, en plus de celles issues du processus de génération des annotations. À l'issue de cette phase, on obtient un ensemble d'annotations A_p tel que A_2(j) $\subset$ A_p(j). Cet ensemble A_p est l'ensemble des **annotations proposées** par défaut.

La deuxième heuristique est l'application des règles dites d'implication potentielle présentées en section 4.1 qui associent à une catégorie ses caractéristiques potentielles (i.e. qui semblent dans les exemples être impliquées par la catégorie). Ces règles ne sont pas des règles certaines mais leur application permet d'identifier pour un jouet j, un ensemble supplémentaire d'annotations de caractéristiques nommé A_s pour **annotations suggérées**.

Par exemple, comme le jouet de la figure 5 est annoté par la catégorie Jeu_de_mise_en_scène, les annotations suivantes sont suggérées : Jeu_associatif, Jeu_coopératif, Jeu_individuel, Raisonnement_pratique, Répétition_par_essais_et_erreurs, etc.

Les annotations dites proposées sont bien plus fiables que les annotations suggérées, c'est pourquoi nous les avons représentées en deux ensembles distincts qui vont être exploités différement par l'interface de validation présentée en Section 4.3. Ainsi les annotations suggérées vont être utilisées comme un filtre pour exclure toutes les autres caractéristiques, i.e., celles qui ne sont ni proposées ni suggérées pour le jouet considéré.

4.3. Phase 1 : Validation manuelle des annotations générées pour l'échantillon

Les différentes annotations, générées automatiquement pour les jouets de l'échantillon dans la phase précédente, doivent pouvoir être confirmées ou modifiées par le concepteur qui doit aussi pouvoir introduire d'éventuelles annotations manquantes. Cette étape de validation manuelle est importante car la qualité des résultats de l'application dépend de la qualité des résultats de la phase d'apprentissage automatique qui elle-même dépend complètement de la qualité des annotations des jouets de l'échantillon. Cette tâche étant longue au vu du nombre de concepts à prendre en compte (33 catégories et 129 caractéristiques), un outil d'aide à la validation des annotations a été conçu. Cet outil est muni d'une interface graphique qui présente les différentes annotations générées pour un jouet et facilite les saisies du concepteur.

La figure 6 montre une première version de l'interface. Elle est pour l'instant très simple. Dans la partie supérieure gauche, on peut voir la liste des jouets qui doivent être annotés. Le premier jouet, qui apparait comme sélectionné sur la figure, est le jouet Playmobil déjà étudié dans les figures 4 et 5. Quand un jouet est sélectionné, son nom, sa marque et sa description apparaissent dans la partie supérieure droite. La partie inférieure de l'interface affiche deux listes de concepts, classés par ordre alphabétique : les catégories dans la fenêtre de gauche, puis les caractéristiques dans celle du centre.

Pour le jouet sélectionné, toutes les **annotations proposées** dans A_p, que ce soient les catégories ou les caractéristiques, apparaissent cochées. Ce sont les annotations par défaut. De plus, dans la fenêtre des caractéristiques, seules les **annotations suggérées** A_s pour le jouet sélectionné sont cochables, i.e., toutes les caractéristiques qui ne sont pas suggérées pour ce jouet sont grisées. L'idée est que le concepteur qui recherche d'éventuelles caractéristiques manquantes n'aie qu'à vérifier celles qui sont cochables sans se pré-occuper des autres. En effet, si une caractéristique ne fait pas partie des caractéristiques proposées et qu'elle n'apparait pas non plus parmi l'ensemble des caractéristiques suggérées déduites par les axiomes pour ce jouet, il est fort probable qu'elle ne soit pas correcte pour le jouet considéré. Elle n'est donc pas mise en avant et apparaît grisée.

L'interface de l'outil d'annotation est dynamique. Quand le concepteur coche un concept, la phase de complétion est automatiquement activée. Cela signifie que dès qu'un nouveau concept est coché, que ce soit une catégorie ou une caractéristique, les règles de complétion s'appliquent, et d'éventuels nouveaux concepts sont proposés (cochés). De plus, quand le concepteur coche une nouvelle catégorie, la première heuristique s'applique et de nouvelles caractéristiques sont elles aussi proposées et cochées. La seconde heuristique s'applique à son tour et de nouvelles caractéristiques sont suggérées (dégrisées donc cochables).

A l'inverse, quand le concepteur décoche une catégorie, la seconde heuristique met à jour ses inférences, i.e., toutes les caractéristiques suggérées qui dépendaient du fait que cette catégorie soit cochée, redeviennent grisées alors que les autres restent

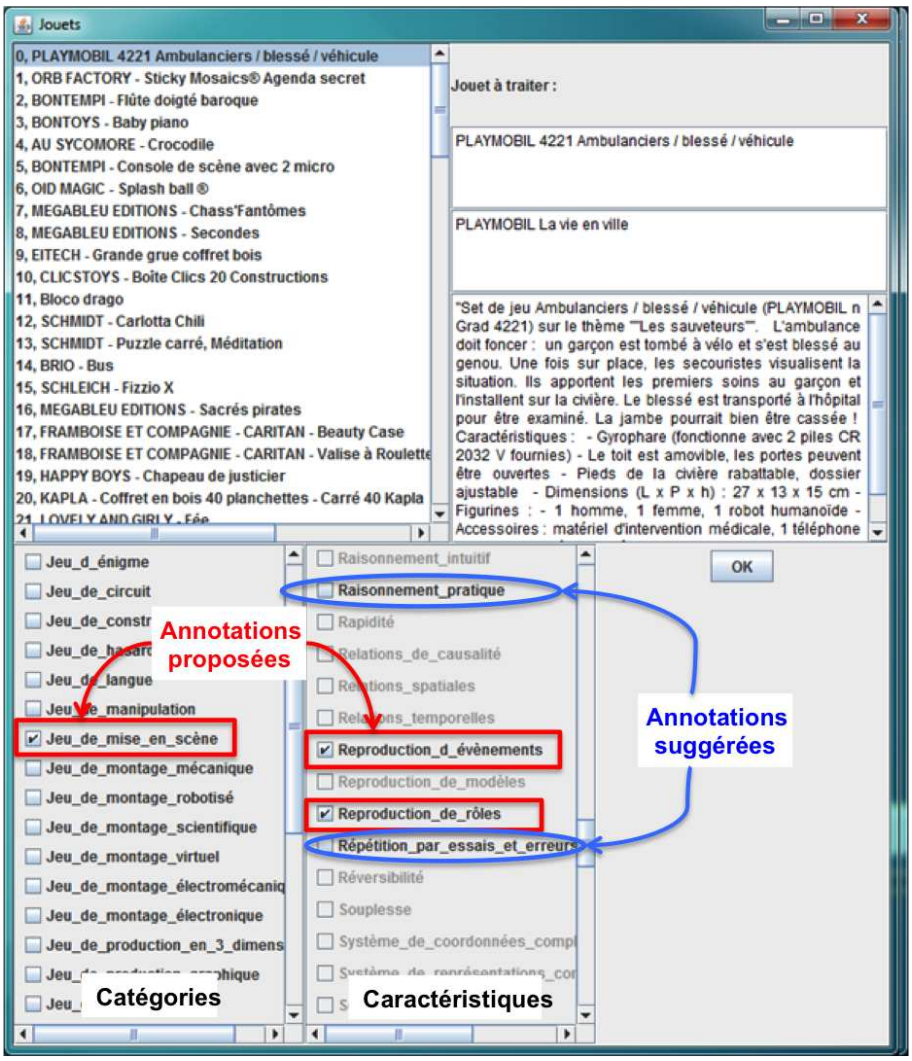


Figure 6. Interface de l'outil d'aide à l'annotation

inchangées. Par contraste, quand le concepteur décoche une caractéristique, rien ne se passe.

Notons que les règles d'incompatibilité ne sont pas appliquées dans cette phase dynamique. En effet, ces règles expriment des connaissances fiables avec une forte probabilité, mais il peut exister des jouets sur lesquels elles ne devront pas s'appliquer. Même si une règle de ce type existe qui dit que SI concept A ET concept B ALORS NON concept A, on peut imaginer qu'il existe un jouet exceptionnel qui doive être annoté à la fois par A et B. Le concepteur doit être libre de cocher les concepts qu'il

souhaite et ne pas être totalement contraint par l'interface. Les règles d'incompatibilité qui pourraient bloquer certains ensembles d'annotations ponctuellement souhaitables ne doivent donc pas être ré-appliquées ici.

De la même façon, le concepteur doit pouvoir supprimer une annotation obtenue par application d'une règle de complétion. Par exemple, s'il existe une règle SI concept A ALORS concept B, et si le concepteur coche le concept A, alors le concept B est automatiquement coché mais le concepteur peut le décocher.

Une fois que toutes les annotations des jouets de l'échantillon ont été validées par le concepteur, ces jouets sont ajoutés à l'ensemble I_{ESAR} comme des instances de chacun des concepts qui les annotent.

4.4. Phase 2 : Annotation du corpus complet par apprentissage basé sur l'échantillon

La section précédente décrivait la première phase de l'approche. Cette phase ayant été utilisée pour annoter un échantillon représentatif de 316 jouets, cette section présente maintenant le modèle d'apprentissage supervisé qui exploite l'échantillon pour construire les classifieurs permettant lors de la deuxième phase, d'annoter de nouveaux jouets (i.e. n'appartenant pas à l'échantillon) qui seront eux-aussi ajoutés à I_{ESAR}.

Nous avons utilisé le classifieur linéaire LIBLINEAR (Fan *et al*., 2008), basé sur SVM (Cortes, Vapnik, 1995), et conseillé notamment pour la classification de documents (Hsu *et al*., 2003). Pour chaque concept c_i, nous avons construit un classifieur SVM qui prédit pour un jouet donné si celui-ci doit être annoté par le concept c_i considéré ou pas. Nous avons donc construit 162 modèles SVM, un pour chaque concept de l'ontologie.

Pour représenter les jouets d'une manière vectorielle, nous avons testé plusieurs représentations de type sac-de-mots (Salton, McGill, 1986) : le monde est décrit avec un dictionnaire de mots et un jouet est représenté par un vecteur de la même taille que le dictionnaire de mots choisi. Chaque élément du vecteur représente un mot. Nous avons testé une représentation sac-de-mots binaire (1 pour la présence du mot dans le descriptif du jouet et 0 pour son absence) et une représentation TF-IDF. Le dictionnaire utilisé est composé des lemmes des mots issus des descriptifs des jouets. Pour chaque représentation vectorielle testée, nous avons pris en compte différents sous-ensembles des attributs des jouets.

Nous avons aussi appliqué une *stop-list* de mots à ne pas prendre en compte (entre autres les nombres, pronoms, prépositions, déterminants, abréviations et conjonctions) que nous appelons *stop-list* de base. Nous proposons aussi une *stop-list* plus élaborée, paramétrable par le concepteur, pour éventuellement ajouter d'autres catégories grammaticales à ne pas prendre en compte. Des compléments d'informations sont donnés dans la partie applicative section 5.2.

La représentation vectorielle des jouets et la création des modèles SVM est entièrement automatique. Une fois les paramètres définitifs choisis, tous les jouets du catalogue sont insérés automatiquement dans I_{ESAR}.

5. Évaluation de l'approche

Dans cette section, nous évaluons les deux phases de l'approche de façon indépendante, d'une part, la phase d'annotation semi-automatique et d'autre part, la phase d'apprentissage. Nous définissons également un protocole expérimental afin d'évaluer la précision des instances introduites dans l'ontologie par l'approche.

5.1. Évaluation du processus de génération d'annotations

Protocole expérimental. Pour évaluer la qualité des annotations proposées par l'outil d'annotation, un sous-échantillon de 100 jouets a été construit de manière aléatoire et annoté manuellement. Seules les catégories de jouets ont été considérées ici car les annotations de caractéristiques sont difficiles à établir, que ce soit manuellement ou par l'outil. Les annotations proposées par l'outil pour cet échantillon ont ensuite été confrontées aux annotations manuelles.

Résultats. Le tableau 1 montre l'amélioration de la précision et du rappel apportée par les différentes étapes d'enrichissement et de raffinement. On remarque que l'amélioration la plus significative vient de l'introduction de nouveaux exemples et des signes linguistiques. Dans la confrontation des résultats, nous avons considéré comme faux un jouet annoté par plusieurs catégories dont l'une au moins était erronée. En revanche, une annotation partielle mais correcte est considérée comme juste. De ce fait, les règles de complétion ne modifient pas le résultat alors qu'en fait, elles introduisent de nombreuses annotations.

L'analyse des résultats montre que notre méthode atteint une précision satisfaisante même si le rappel est assez limité. Avoir une précision élevée pour les annotations proposées est très important. Cela signifie que le travail de validation manuelle du concepteur sera minimisé car peu de fausses annotations devront être supprimées parmi celles proposées. Le fait que le rappel reste relativement faible malgré l'ajout de nouveaux exemples et de signes linguistiques dans la composante terminologique reflète le fait qu'un tel enrichissement n'est encore pas suffisant et qu'il doit être complété.

Tableau 1. Précision, Rappel et F-mesure du processus d'annotation

Étape	Précision	Rappel	F-mesure
Avant enrichissement	0,38	0,20	0,26
Exemples + signes linguistiques ajoutés	0,87	0,55	0,68
Signes linguistiques complexes	0,88	0,59	0,71
Détection incohérences (+ complétion)	0,94	0,64	0,76

5.2. Évaluation de la phase d'apprentissage automatique

Protocole expérimental. Pour évaluer la partie apprentissage de l'approche, nous nous sommes concentrés sur le concept jeu_de_mise_en_scène. En effet, il y a de très nombreux exemples positifs et négatifs de ce concept dans le corpus, ce qui n'est pas le cas des autres concepts. Un échantillon de jouets (316 jouets), extrait d'un catalogue particulier (Toys'R'Us) et annoté avec l'outil, constitue l'ensemble d'apprentissage. Pour l'ensemble de test, nous avons repris le même catalogue privé des jouets de l'échantillon (595 jouets) et annoté avec l'outil ces différents jouets, uniquement en terme de jeu_de_mise_en_scène ou non. Ainsi, nous construisons un modèle sur un échantillon de jouets d'un catalogue et nous observons le taux d'erreur sur les autres jouets de ce catalogue. Parmi les 36 modèles testés, nous avons opté pour celui qui génère le plus faible taux d'erreur (soit le modèle n° 12b obtenu avec les paramètres en gras italique sur le tableau 2, et qui génère 2,52 % d'erreur). Nous avons ensuite appliqué ces mêmes paramètres pour l'apprentissage de chacun des 162 modèles SVM créés (un par concept de l'ontologie).

Tableau 2. Taux d'erreur d'annotation de l'ensemble de test pour les jeux de mise en scène

			Taux d'erreur		
N°	C	Descriptif	Représentation	Stop-list (a)	**Stop-list (b)**
...	...	...	...	...	...
10	10	LMC	TF-IDF	6,72 %	6,72 %
11	10	LMCD	Binaire	3,87 %	4,87 %
12	***10***	***LMCD***	***TF-IDF***	3,03 %	***2,52 %***
13	100	LM	Binaire	9,41 %	9,41 %
14	100	LM	TF-IDF	9,75 %	9,75 %
...	...	...	...	...	...

Résultats. Le tableau 2 montre un extrait des pourcentages d'erreur obtenus avec chacun des 36 modèles testés pour le classifieur de jeu_de_mise_en_scène sur l'ensemble de test du catalogue Toys'R'Us. Dans ce tableau, le paramètre **C** du classifieur modélise le coût de violation des contraintes. Autrement dit, plus C est grand, plus on impose que les données soient sûres (non bruitées). Le paramètre **descriptif** correspond aux éléments pris en compte dans le vecteur parmi les différents attributs d'un jouet (label L, marque M, catégorie C, description D). Le paramètre **représentation** correspond à la méthode de représentation vectorielle utilisée (binaire ou TF-IDF). Les expérimentations ont été conduites avec deux stop-lists : la stop-list (a) qui correspond à la stop-list de base et la **stop-list (b)** qui supprime en plus les adverbes.

L'ensemble d'apprentissage étant représentatif du catalogue Toys'R'Us tout entier, il l'est donc aussi de l'ensemble de test. Autrement dit, les jouets de l'ensemble de test sont assez proches d'au moins un jouet de l'ensemble d'apprentissage ce qui explique nos bons résultats.

5.3. Évaluation de la phase de peuplement

Protocole expérimental. Pour résumer, notre approche est composée de deux phases : la première repose sur la génération automatique d'annotations pour un ensemble de jouets, ces annotations étant ensuite validées et/ou corrigées par le concepteur de façon à ce qu'elles soient correctes à la fin de cette première phase. La deuxième phase permet, grâce à de multiples classifieurs, d'annoter un jouet avec tous les concepts dont il est une instance et donc de peupler l'ontologie. Nous avons besoin de valider ces dernières annotations.

Pour ce faire, nous avons fait annoter (en tant que jeu_de_mise_en_scène ou non) par le modèle SVM trouvé, un ensemble de 100 jouets issus d'un autre catalogue (Jeux et Jouets en folie) et pris de manière à être le plus hétérogène possible. Soulignons que ces jouets sont très différents de ceux du catalogue Toys'R'Us et que les deux catalogues n'ont aucun jouet en commun. On peut donc s'attendre à ce que le modèle d'apprentissage, basé uniquement sur un ensemble représentatif du premier catalogue, ne soit pas très performant sur ces données.

Résultats. Nous avons vu en 5.2 que quand l'ensemble d'apprentissage était représentatif de l'ensemble de test, nous obtenions un taux d'erreur de 2,52 %, ce qui est très faible. Les 100 jouets extraits du catalogue Jeux et Jouets en folie étant très différents de ceux de l'ensemble d'apprentissage, nous ne pouvons pas espérer obtenir un aussi faible taux d'erreur. Le tableau 3 montre les résultats obtenus sur ces 100 jouets avec le modèle n° 12b retenu. Parmi les 31 jeu_de_mise_en_scène, 15 ont bien été étiquetés comme tel. Aucun jouet n'a été étiqueté comme jeu_de_mise_en_scène alors qu'il ne l'était pas. On obtient donc 100 % de précision. Le taux d'erreur est élevé, 16 %, et nous pouvons voir que les erreurs viennent des faux négatifs car le rappel n'est qu'à 50 %. Comme nous l'avions dit, le rappel est faible car l'échantillon d'apprentissage, basé sur le premier catalogue, n'est pas représentatif des jouets du second catalogue. Ces résultats nous semblent donc très satisfaisants et nous pouvons supposer qu'en agrandissant l'échantillon d'apprentissage avec des jouets du deuxième catalogue, nous obtiendrions un meilleur rappel.

Tableau 3. Résultats sur 100 jouets de "Jeux et Jouets en folie"

Résultats	
Taux d'erreur	16 %
Précision	100 %
Rappel	48,39 %
F-Mesure	65,22 %

6. Conclusion et perspectives

Dans cet article, nous avons proposé une approche originale pour associer des produits décrits dans des catalogues aux concepts d'une ontologie de domaine. Cette approche a été testée sur l'univers des jouets. Elle répond ainsi à une problématique de

peuplement automatisé d'ontologie. Son originalité est d'une part, la génération itérative des annotations et, d'autre part, la complémentarité entre les phases automatiques et semi-automatiques. Ainsi, l'approche est optimisée afin de réduire au minimum le travail de l'expert. Néanmoins le travail de celui-ci est nécessaire car la faible qualité des descriptifs de produits ne permet pas à une approche automatique d'être performante.

Les premiers résultats d'annotation des produits par leurs catégories sont prometteurs. En revanche, les caractéristiques évoquant des notions abstraites rarement utilisées dans les descriptifs, les signes linguistiques évocateurs sont plus rares et les annotations sont plus difficiles à établir.

La partie apprentissage a bien fonctionné sur les jeux de mise en scène même si ces types de jeux sont difficiles à reconnaître. Par exemple, un humain peut lire la description d'un tracteur sans comprendre s'il s'agit d'un tracteur miniature (jeu_de_mise_en_scène) ou d'un tracteur à pédales (non jeu_de_mise_en_scène). Étant donné cette difficulté pour un humain, nous estimons qu'un tel concept n'était pas simple à traiter d'une façon automatique.

L'approche exige du concepteur un travail de validation des annotations d'un échantillon représentatif de la diversité des produits. Cette tâche est manuelle et peut sembler lourde mais elle est limitée dans le temps car elle n'est à faire qu'une seule fois (modulo quelques ajustements pour prendre en compte les articles nouveaux). Le reste de l'approche est entièrement automatique.

Nous envisageons plusieurs perspectives à ce travail. Tout d'abord, trouver une solution mieux adaptée au traitement des caractéristiques. Ensuite, il faudrait utiliser une ressource externe qui permettrait d'ajouter des signes linguistiques de manière automatique et d'aider le concepteur à définir les axiomes. Nous agrandirons l'échantillon afin de tenir compte des jouets de tous les catalogues. On pourrait aussi envisager d'améliorer la partie automatique en testant d'autres méthodes d'apprentissage (Bayes, Perceptron Multi-Couches, ...) et d'autres formes de représentations vectorielles (tenant compte des synonymes par exemple). Plutôt que d'utiliser un classifieur, on pourrait tester une méthode plus proche de celle proposée par (Kessler *et al.*, 2012), consistant à comparer un vecteur représentant plusieurs instances d'un concept donné avec un vecteur représentant un jouet à classer. Enfin, comme cette approche est indépendante du domaine et reproductible avec des connaissances adaptées, il serait intéressant de l'appliquer à d'autres domaines, tels que les cadeaux en général ou les voyages, comme souhaite le faire Wepingo.

Remerciements

Nous remercions la société Wepingo qui a financé ce travail dans le cadre du projet PORASO.

Bibliographie

Amardeilh F., Damljanovic D. (2009). Du texte à la connaissance : annotation sémantique et peuplement d'ontologie appliqués à des artefacts logiciels. In F. L. Gandon (Ed.), *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 157-168. Hammamet, Tunisie, PUG.

Amardeilh F., Laublet P., Minel J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *Proceedings of the 3rd international conference on Knowledge Capture (K-CAP)*, p. 161–168. New York, NY, USA, ACM.

Aussenac-Gilles N., Kamel M., Comparot C., Buscaldi D. (2013, juillet). Construction d'ontologies à partir de pages web structurées. In R. Troncy (Ed.), *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 1–17. Lille, France, AFIA.

Barriere C., Agbago A. (2006). Terminoweb: a software environment for term study in rich contexts. In *Proceedings of the 2005 international conference on terminology, standardization and technology transfer*, p. 103–113.

Béchet N., Aufaure M.-A., Lechevallier Y. (2012, mai). Construction et peuplement de structures hiérarchiques de concepts dans le domaine du e-tourisme. In *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 475-490. Chambéry, France. Consulté sur http://hal.archives-ouvertes.fr/hal-00746719

Bontcheva K., Tablan V., Maynard D., Cunningham H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, vol. 10, n° 3/4, p. 349–373.

Cortes C., Vapnik V. (1995, septembre). Support-Vector Networks. *Machine Learning*, vol. 20, n° 3, p. 273–297.

Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, vol. 9, p. 1871–1874. (Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear)

Garon D., Filion R., Chiasson R. (2002). *Le système ESAR: guide d'analyse, de classification et d'organisation d'une collection de jeux et jouets*. Editions ASTED.

Gruber T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, vol. 5, n° 2, p. 199-220.

Hsu C.-W., Chang C.-C., Lin C.-J. (2003). *A Practical Guide to Support Vector Classification*. Rapport technique. Department of Computer Science, National Taiwan University. Consulté sur http://www.csie.ntu.edu.tw/~cjlin/papers.html

Kessler R., Béchet N., Roche M., Moreno J. M. T., El-Bèze M. (2012). A Hybrid Approach to Managing Job Offers and Candidates. *Information Processing and Management*, vol. 48, n° 6, p. 1124-1135.

Manning C. D., Schütze H. (1999). *Foundations of statistical natural language processing*. Cambridge, Massachusetts, The MIT Press.

Petasis G., Karkaletsis V., Paliouras G., Krithara A., Zavitsanos E. (2011). Ontology Population and Enrichment: State of the Art. In *Knowledge-driven multimedia information extraction and ontology evolution*, p. 134-166.

Popov B., Kiryakov A., Ognyanoff D., Manov D., Kirilov A. (2004, septembre). KIM – a Semantic Platform for Information Extraction and Retrieval. *Natural Language Engineering*, vol. 10, n° 3-4, p. 375–392.

Reeve L. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 acm symposium on applied computing*, p. 1634–1638. ACM Press.

Reymonet A., Thomas J., Aussenac-Gilles N. (2007). Modélisation de Ressources Termino-Ontologiques en OWl. In F. Trichet (Ed.), *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 169-181. Grenoble, France, Cepadues.

Salton G., McGill M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA, McGraw-Hill, Inc.

Suchanek F. M., Sozio M., Weikum G. (2009). SOFIE: a Self-Organizing Framework for Information Extraction. In *World Wide Web Conference (WWW)*, p. 631-640. Madrid, Spain, ACM.