Évaluation de classes sémantiques pour la construction d'ontologies

Sarra Ben Abbès, Haifa Zargayouna, Adeline Nazarenko

Évaluation de classes sémantiques pour la construction d'ontologies

Sarra Ben Abbès, Haïfa Zargayouna, Adeline Nazarenko

Laboratoire d'Informatique de l'université Paris-Nord
(LIPN) - UMR 7030
Université Paris 13 - CNRS
99, avenue Jean-Baptiste Clément -
F-93430 Villetaneuse, France
Email: prenom.nom@lipn.univ-paris13.fr

Résumé :
L'essor du Web Sémantique a permis la multiplication des ontologies en ligne, donnant un large choix de ressources mais soulevant par là-même le problème de leur évaluation. La diversité des motivations de construction des ontologies et la complexité des domaines de spécialité rend difficile l'évaluation de la "qualité" des ontologies. Nous mettons ici l'accent sur l'acquisition des classes sémantiques, une des sous-tâches de l'acquisition des connaissances. Nous définissons un cadre d'évaluation permettant de comparer la liste de classes sémantiques produite par un système d'acquisition avec une ontologie de référence où les concepts sont eux-mêmes représentés par un ensemble de termes ou de labels. L'originalité de notre approche consiste à tenir compte des cas d'appariement partiel entre les classes sémantiques et les concepts de référence et donc entre la sortie d'un système d'acquisition et une ontologie de référence. Les premiers résultats expérimentaux montrent l'intérêt de nos propositions.

Mots-clés : ontologie, concepts, termes, classes sémantiques, évaluation.

1 Introduction

Ces dernières années, un effort considérable a été accompli dans les domaines du partage et de la réutilisation des connaissances et notamment des ontologies. Celles-ci sont destinées à représenter la sémantique d'un domaine, mais tout travail de modélisation reflétant le point de vue de son auteur, la comparaison des ontologies produites reste difficile, même pour un domaine donné. L'évaluation des ontologies apparaît aujourd'hui comme une problématique incontournable (Maedche & Staab, 2002) : on veut pouvoir choisir entre plusieurs ontologies et évaluer les résultats fournis par les outils d'acquisition d'ontologies.

Il n'existe pourtant pas encore de cadre fédérateur permettant l'évaluation des ontologies, ce qui s'explique en partie par la complexité des structures ontologiques. Une autre difficulté tient à la variabilité de la référence ("gold standard") qui se révèle

généralement très dépendante du domaine modélisé et de l'utilisation prévue. Une dernière difficulté tient aux limites des mesures classiques telles que la précision et le rappel qui reposent sur des jugements de pertinence binaires alors qu'on veut pouvoir rendre compte d'une pertinence graduée, une classe pouvant ne correspondre que partiellement à un concept de la référence (Zargayouna & Nazarenko, 2010).

Pour remédier à ces problèmes, une solution consiste à décomposer le problème de l'évaluation des ontologies en sous-problèmes suivant la nature des données manipulées. Nous considérons ici la partie conceptuelle des ontologies, ou la T-Box, sans prendre en compte le niveau des instances. Une telle ontologie est donc composée d'un ensemble de concepts, de relations hiérarchiques entre ces concepts et d'un ensemble de rôles ou relations sémantiques. Lorsque les concepts sont représentés par des classes sémantiques de termes, on peut évaluer les concepts en tant que tels, indépendamment de leur définition formelle. C'est ce sur quoi nous mettons l'accent ici. Nous nous intéressons à l'évaluation des classes sémantiques et nous proposons un protocole et des mesures adaptés à cette tâche. Cela permet d'évaluer la sortie des systèmes d'acquisition de classes sémantiques dans la perspective de la construction des ontologies, même si ceux-ci ne produisent pas des ontologies complètement structurées. Cette évaluation repose sur la comparaison de l'ensemble des classes fournies par un système avec une ontologie de référence, dont les concepts sont eux-mêmes considérés comme des classes de termes mais organisés en hiérarchie. Nous proposons un protocole d'évaluation comparative permettant de mettre en correspondance les classes fournies par des systèmes d'acquisition et les concepts d'une ontologie de référence sur la base des termes qu'ils contiennent ou qui leur sont associés.

Nous posons le cadre de l'évaluation modulaire dans lequel nous avons choisi de nous placer (section 2) avant de présenter un protocole et des mesures adaptés à l'évaluation des classes sémantiques (section 3). La section 4 apporte une première validation de notre approche et notre positionnement par rapport aux travaux existants est présenté dans la section 5.

2 Evaluation modulaire

Pour nous affranchir de la complexité intrinsèque des d'ontologies, nous proposons une évaluation par niveau en considérant les concepts, leur structure en hiérarchie et les relations sémantiques comme différents artefacts à évaluer :

Classes sémantiques On considère comme classes sémantiques les regroupements des termes que fournissent souvent les outils d'acquisition comme ébauches de concepts (par ex. ASIUM (Faure & Nédellec, 1998)). Les classes sémantiques sont ensuite comparées aux concepts de la référence. Une classe sémantique est définie par un identifiant (le nom de la classe) et un ensemble de termes. Nous faisons l'hypothèse que les concepts de la référence sont eux aussi associés à des ensembles de termes qui les caractérisent.

Par souci de clarté, nous parlons, dans cet article, de "classes sémantiques" pour désigner les sorties des outils d'acquisition et de "concepts" pour parler de la référence.

Relations hiérarchiques Ces relations structurent l'ensemble des concepts d'un domaine. Il s'agit selon les cas de relations de subsomption (généralisation / spécification), de relations de partie à tout (agrégation / composition) ou d"autres relations.

Rôles Sont définies sous cette forme toutes les relations sémantiques entre concepts d'une ontologie autres que les relations hiérarchiques.

Nous considérons ces trois aspects de manière distincte pour les évaluer indépendamment les uns des autres, ce qui permet d'évaluer les systèmes qui proposent des classes sémantiques sans construire une ontologie complète.

Dans cet article, nous mettons l'accent sur l'évaluation des classes sémantiques. Nous ne tenons donc pas compte de l'éventuelle organisation hiérarchique fournie par les systèmes, laquelle est destinée à être évaluée séparément. Nous verrons, en revanche, que l'évaluation des classes sémantiques tient compte de la structure hiérarchique de la référence.

3 Propositions

Cette section présente le protocole d'évaluation pour l'évaluation des classes sémantiques. La procédure d'évaluation prend en entrée une liste de classes sémantiques et une ontologie de référence et fournit en sortie une mesure de pertinence pour chacune des classes sémantiques et au regard de l'ontologie de référence (voir figure 1).

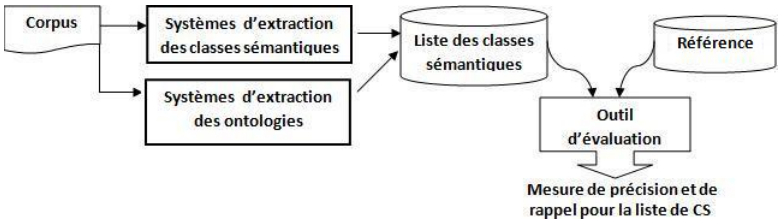

FIGURE 1 – Protocole d'évaluation des classes sémantiques : comparaison à une référence

3.1 Mise en correspondance entre les classes et les concepts

L'évaluation des classes sémantiques revient à établir une *correspondance globale* entre la liste des classes et les concepts de référence. Une classe sémantique CS correspond à un concept de référence CR si et seulement si CS possède au moins un terme en commun avec CR. Autrement dit, une classe sémantique CS parfaite est une classe qui a exactement les mêmes termes qu'un concept de la référence CR. Par contre, une classe nulle est une classe qui n'a aucun terme en commun avec aucun concept de référence. Formellement, si TCS est l'ensemble des termes de la classe sémantique et TCR l'ensemble des termes du concept de la référence, la correspondance entre CS et CR est parfaite quand TCS = TCR et nulle quand $TCS \cap TCR = \emptyset$.

3.2 Prise en compte des correspondances partielles

Entre ces deux extrêmes, il existe beaucoup de situations intermédiaires mais nous souhaitons éviter de considérer comme bruitées toutes les classes qui ne sont pas en correspondance parfaite avec un concept de la référence. Pour cela, nous proposons d'adapter les sorties des systèmes à la pertinence, une opération d'ajustement qui change le nombre de classes de la sortie des systèmes et en modifie les contours en fonction de la référence. Il s'agit de maximiser les points de correspondance en considérant que la granularité de la description fournie par la référence est arbitraire. Mettre en correspondance les sorties des systèmes directement avec la référence pénaliserait les systèmes qui ont choisi une granularité de description différente (plus fine ou plus grossière, avec respectivement plus et moins de classes).

Ce processus d'*ajustement* est similaire à celui proposé dans (Nazarenko & Zargayouna, 2009) pour l'extraction de termes. Il repose sur deux opérations fondamentales. La première consiste à *éclater* les classes sémantiques qui correspondent à plusieurs concepts de l'ontologie de référence. La figure 2 montre un exemple d'éclatement où une classe CS1 d'une sortie partage des termes avec trois concepts différents (CR1, CR2 et CR3) de la référence. Il n'y a pas de correspondance parfaite mais plusieurs correspondances partielles. Dans ce cas, nous proposons de décomposer la classe CS1 en trois sous-classes qui partagent chacune des termes avec un seul concept de la référence. S'il y a des termes de CS1, comme t5, qui n'apparaissent dans aucun concept de la référence, ils constituent du bruit et sont maintenus dans toutes les sous classes issues de l'éclatement. La seconde opération consiste au contraire à *regrouper* les classes sémantiques qui correspondent à un même concept de la référence. La figure 2 montre trois classes CS2, CS3 et CS4 qui partagent toutes des termes avec le même concept CR4. Les termes des trois classes sont alors regroupés en une seule classe CS'234, qui correspond au concept CR4 de la référence.

En pratique, la mise en correspondance est établie par un calcul matriciel et, une fois cette correspondance établie, on transforme les classes sémantiques pour avoir une relation de correspondance binaire. On peut alors calculer la pertinence de chaque classe par rapport au concept qui lui correspond, et partant, celle de la sortie ajustée par rapport à la référence. Nous distinguons dans ce qui suit les classes de la sortie initiale d'un système (SC_i) et les classes de la sortie transformées (CS'_j).

3.3 Pertinence d'une classe

Nous distinguons trois mesures de pertinence d'une classe selon qu'elle est transformée ou non et selon la nature de la transformation opérée. Nous nous inspirons des mesures classiques de précision et de rappel [1] qui ont le mérite d'être génériques et faciles à interpréter (Martin *et al.*, 2004).

1. Rappelons que :
P = $\frac{|S \cap R|}{|S|}$
R = $\frac{|S \cap R|}{|R|}$
où $|S \cap R|$ est le nombre des éléments pertinents retournés par le système, $|S|$ est le nombre total d'éléments retournés par le système et $|R|$ le nombre total d'éléments dans la référence.

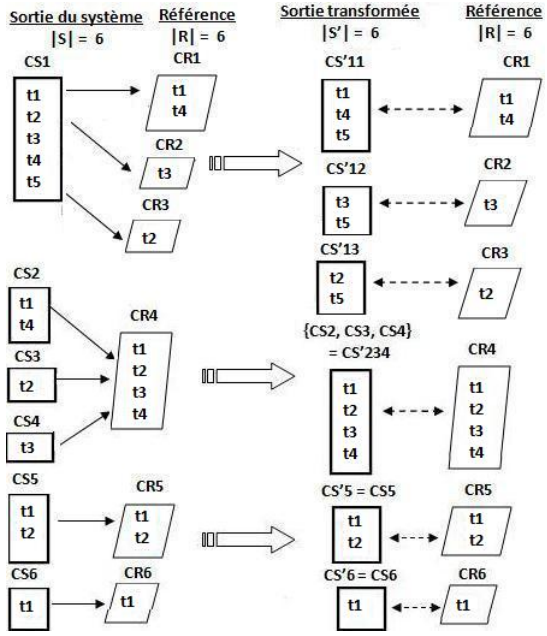

FIGURE 2 – Exemple d'éclatement, de regroupement et d'une correspondance parfaite

3.3.1 Cas des classes non transformées

Une classe pouvant n'être qu'en cas de correspondance partielle avec le concept avec lequel elle a été mise en correspondance, nous proposons une mesure de pertinence graduée qui repose sur le nombre de termes partagés entre la classe et le concept correspondant :

$$P(CS, CR) = \frac{\text{nombre de termes pertinents d'une classe CS}}{\text{nombre total de termes de la classe CS}}$$
$$R(CS, CR) = \frac{\text{nombre de termes pertinents d'une classe CS}}{\text{nombre total de termes du concept de la référence CR}}$$
$$pertinence_s(CS', CR) = FM(CS, CR) = \frac{2 * \text{P(CS, CR)} * \text{R(CS, CR)}}{\text{P(CS, CR)} + \text{R(CS, CR)}}$$

où CS' est une classe non transformée de la sortie ($CS = CS'$), CR un concept de la référence, $P(CS, CR)$, la précision de CS par rapport à CR et $R(CS, CR)$, le rappel de CS par rapport à CR et FM, la F-mesure calculée à partir de P et R.

Dans la figure 2, la classe $CS5$ est par exemple en correspondance parfaite avec un concept $CR5$ de la référence. On a donc :

$$pertinence_r(CS'5, CR5) = FM(CS5, CR5) = 1.$$

3.3.2 Cas des classes obtenues par transformation

Appliquer la même mesure de pertinence pour les classes obtenues par transformation ne tiendrait pas compte du fait que la sortie a été ajustée à la référence. Une classe obte-

nue après ajustement de la sortie (qu'il s'agisse d'un regroupement ou d'un éclatement) doit être considérée comme moins pertinente que la même classe qui n'aurait pas eu besoin d'être ajustée.

Dans les cas de *regroupement*, on a une classe CS' obtenue à partir de différentes classes CS_i et nous proposons de calculer la pertinence de CS' par rapport à CR en prenant la moyenne [2] de la pertinence des classes CS_i non transformées, ce qui revient à pénaliser les classes obtenues par regroupement :

$$pertinence_r(CS', CR) = \frac{\sum_{i=1}^{|X|} FM(CS_i, CR)}{|X|}$$

où CS_i sont des classes sémantiques retournées par un système qui sont regroupées en une seule classe CS', CR le concept de l'ontologie de référence correspondant à ces différentes CS_i et $|X|$ le nombre de classes CS_i qui ont été regroupées en CS'.

Dans l'exemple de CS2, CS3 et CS4 (figure 2), la pertinence de (CS'234, CR4) est calculée comme suit :

$$FM(CS2, CR4) = 0,66$$
$$FM(CS3, CR4) = FM(CS4, CR4) = 0,4$$
$$pertinence_r(CS'234, CR4) = \frac{0.66+0.4+0.4}{3} = 0.48.$$

Dans les cas *d'éclatement*, on a plusieurs classes CS'_i transformées, obtenues à partir d'une même classe CS initiale et mises en correspondance avec différents concepts CR_i de la référence. On veut tenir compte de la proximité dans la référence des différents concepts CR_i sur lesquels se fait l'éclatement. S'ils sont proches, la pertinence est considérée comme meilleure que s'ils sont éloignés. Cette proximité est calculée par rapport à un concept pivot qui sert à un point fixe pour les classes obtenues par éclatement. Nous calculons la pertinence des classes obtenues par éclatement comme suit :

1. On recherche le concept p de la référence ayant avec la classe CS initiale la FM la plus élevée : il est destiné à jouer le rôle de concept pivot.
2. Pour chaque classe CS' obtenue par éclatement, on calcule sa pertinence par rapport au concept de la référence CR avec lequel elle est mise en correspondance à partir de la pertinence initiale de CS et de la distance que CR entretient avec p de manière à pénaliser les éclatements qui dispersent le plus avec des classes initiales. Nous introduisons la mesure de (Wu & Palmer, 1994) [3] qui calcule la similarité entre deux concepts dans une hiérarchie, pour calculer la similarité entre le concept pivot p et un concept de la référence CR : où C est le plus petit ancêtre commun de p et CR et où les fonctions $depth(X)$ et $depth_C(X)$ retournent respectivement le nombre d'arcs séparant X de la racine et le nombre d'arcs séparant X de la racine en passant par C. On obtient ainsi la mesure de pertinence suivante pour les classes obtenues par transformation d'une classe dont le concept pivot est p :

$$pertinence_e(CS', CR) = FM(CS, CR) * Sim(p, CR)$$

2. D'autres mesures comme le minimum ou le maximum peuvent être envisagées mais elles n'ont pas encore été testées.

3. D'autres mesures sont envisageables comme la mesure de similarité de Jaccard, de Cosine, etc. mais seule la similarité de Wu et Palmer a été testée jusqu'à présent.

où CS, CS' et CR sont respectivement une classe sémantique fournie par un système, une classe obtenue par éclatement et un concept de la référence. p est le concept de la référence ayant la FM la plus élevée avec la classe sémantique CS dont CS' est issue par éclatement.

Prenons l'exemple de CS1 (figure 2) qui est éclatée en trois classes CS'11, CS'12 et CS'13 correspondant respectivement à CR1, CR2 et CR3. Dans ce cas, le concept pivot est CR1 car c'est celui qui a la FM la plus élevée avec CS1 :

$$FM(CS1, CR1) = 0.57$$
$$FM(CS1, CR2) = 0.33$$
$$FM(CS1, CR3) = 0.33$$

Si on suppose $Sim(CR1, CR2) = 0.5$, la pertinence de $CS'12$ par rapport à $CR2$ est :

$$pertinence_e(CS'12, CR2) = FM(CS1, CR2) * Sim(CR1, CR2) = 0.33 * 0.5 = 0.165.$$

3.4 Pertinence d'un ensemble de classes

La pertinence globale de la sortie est calculée comme suit :

$$\text{P} = \frac{\sum_{i=1}^{|S'|} \sum_{j=1}^{|R|} pertinence(CS'_i, CR_j)}{|S'|}$$
$$\text{R} = \frac{\sum_{i=1}^{|S'|} \sum_{j=1}^{|R|} pertinence(CS'_i, CR_j)}{|R|}$$

où $|S'|$ et $|R|$ sont respectivement le nombre de classes de la sortie ajustée et le nombre de concepts de la référence.

$$pertinence(CS'_i, CR_j) = \begin{cases} pertinence_s & \text{si } CS'_i \text{ n'est pas transformée} \\ pertinence_r & \text{si } CS'_i \text{ est obtenue par regroupement} \\ pertinence_e & \text{si } CS'_i \text{ est issue d'un éclatement} \end{cases}$$

4 Méta-évaluation de l'approche proposée

Les efforts d'évaluation devant eux-mêmes être évalués (Stufflebeam, 1974), nous avons testé nos mesures expérimentalement pour vérifier que leur comportement correspond bien à nos spécifications. Nous présentons deux expérimentations qui comportent chacune un jeu de test et une ontologie de référence. La première expérimentation montre le comportement des mesures proposées sur un ensemble de classes sémantiques construites à partir des synsets de WordNet. La deuxième expérimentation va plus loin et montre les résultats obtenus dans l'évaluation de petites ontologies construites par des étudiants.

Ces expérimentations reposent sur un outil d'évaluation qui a été implémenté en Java sur la base des propositions précédentes. Il prend en entrée un ensemble de classes sémantiques S présenté dans un format textuel défini ou des ontologies en format

OWL [4] et une ontologie de référence R en format OWL et calcule la pertinence de S par rapport à R.

4.1 Analyse du comportement des mesures

Nous avons construit manuellement une petite ontologie de référence à partir des synsets de WordNet qui traitent du domaine de transport. Elle est constituée de 6 concepts qui regroupent un total de 17 termes (à chaque concept est associé un ou plusieurs termes). Nous avons également réalisé un jeu de test présentant des cas de regroupement et d'éclatement pour tester le comportement de nos mesures aux limites. Ce jeu de test artificiel propose plusieurs classifications de termes différentes, chacune étant supposée représenter une sortie de système d'acquisition de classe sémantique.

Nous avons produit une série de variantes à partir d'une première sortie Sp correspondant parfaitement à l'ontologie de référence en ajoutant du bruit sous la forme de 6 termes ne figurant pas dans la référence ($+b$) ou du silence en supprimant 6 termes initiaux ($+s$). Le bruit peut être réparti sur l'ensemble des classes de la sortie ($+br$), concentré sur une classe existante ($+bc$) ou regroupé dans une classe supplémentaire ($+bC$). Aucune des sorties de type $Sp + / - x$ n'ayant besoin d'être transformée pour être mise en correspondance avec R, nous avons également considéré des sorties nécessitant des regroupements (Sr comporte 17 classes contenant chacune un terme) et des éclatements (Se ne propose qu'une seule classe contenant les 17 termes de la références) et leur variantes bruitées ($+b$) et silencieuses ($+s$). Nous avons enfin considéré une sortie Ser combinant des cas de regroupements et d'éclatements : Ser propose 9 classes dont la première contient 9 termes correspondant à 3 concepts de la référence et les 8 autres ne contiennent qu'un terme et correspondent ensemble aux 3 autres concepts de la référence. Une dernière sortie Sn ne contient aucun terme de la référence.

Les mesures de pertinence de ces différentes sorties sont présentées dans le tableau 3. On constate sans surprise que les sorties parfaite et nulle figurent en tête et queue de classement et que toutes les autres sorties se situent dans l'intervalle. La mesure proposée permet donc une évaluation assez fine des sorties des systèmes. L'ajout de bruit ou de silence fait baisser la F-mesure, quelle que soit le type de sortie considéré mais moins fortement quand cela se combine avec des transformations (séries Sr et Se). Sur ce jeu de test, on a généralement $P = R = FM$ parce que presque toutes les sorties transformées contiennent le même nombre de classes que la référence ; seules les sorties de type $+bC$ proposent une classe de bruit en plus. On constate par ailleurs que les opérations de transformation (sorties de types Sr et Se) font baisser la F-mesure ($Sr < Sp$ et $Se < Sp$) sans la dégrader complètement ($Sr \neq Sn$ et $Se \neq Sn$). On note un écart important entre Sr et Se qui place la série des sorties Se derrière celles de la série Sr mais cela tient à la structure de l'ontologie de référence [5]. La sortie qui fait à la fois l'objet d'un éclatement et d'un regroupement (Ser) occupe une position intermédiaire.

4. L'outil transforme les concepts de l'ontologie en classes sémantiques qui serviront dans le protocole.

5. Des jeux de test faisant varier la taille et la structure de l'ontologie de référence permettraient de mettre en valeur cet effet.

Sortie	P	R	FM	*Classement*
Sp	1	1	1	1
Sp+s	0,77	0,77	0,77	4
Sp+br	0,84	0,84	0,84	3
Sp+bc	0,69	0,69	0,69	5
Sp+bC	0,86	1	0,92	2
Sr	0,54	0,54	0,54	6
Sr+s	0,52	0,52	0,52	7
Sr+br	0,50	0,50	0,50	8
Sr+bc	0,50	0,50	0,50	8
Sr+bC	0,46	0,54	0,50	10
Se	0,20	0,20	0,20	13
Se+br	0,15	0,15	0,15	15
Se+s	0,19	0,19	0,19	14
Ser	0,46	0,46	0,46	10
Sn	0	0	0	16

TABLE 1 – Résultats de la première expérience

4.2 Comparaison avec les mesures de précision et rappel classiques

A partir d'un corpus textuel du domaine de volley-ball, nous avons construit manuellement une petite ontologie de référence comportant 64 concepts (où chaque concept contient un ou plusieurs termes) et nous avons comparé à cette ontologie prise comme référence, trois ontologies construites par des étudiants à partir du même corpus textuel. Ces trois ontologies ($S1$, $S2$ et $S3$) sont présentées dans le tableau 2. L'évaluation manuelle de ces trois ontologies a donné le classement suivant : $S2$, $S1$ et $S3$.

Sorties	Nb. total de classes	Nb. de classes en corresp. parfaite	Nb. de classes en corresp. partielle	Nb. de classes en corresp. nulle
S1	63	26	19	18
S2	64	29	16	19
S3	67	23	15	29

TABLE 2 – Sorties des systèmes

Dans cette deuxième expérience, nous avons cherché à comparer nos mesures d'évaluation avec les mesures de précision et rappel classique. Les mesures classiques reposent sur un jugement binaire : les classes qui ne correspondent pas parfaitement à des concepts de la référence sont jugées non pertinentes [6].

Les résultats figurent dans le tableau 3. On constate que le changement de mesure n'a pas d'effet sur le classement global, ce qui est cohérent puisque nos mesures restent des mesures de rappel et de précision même si elles reposent sur des jugements de

6. Pour $S1$, on a par exemple $Precision = 26/63$ et $Rappel = 26/64$.

pertinence gradués. En revanche, on observe un écart de l'ordre de 20% entre notre FM et la FM classique. Cet écart est essentiellement dû à l'augmentation de la précision, dans le cas présent, mais cela tient au fait que, sur cette expérience, l'adaptation des sorties à la référence repose essentiellement sur des opérations de regroupement : quand il y a correspondance partielle entre les sorties des étudiants et la référence, les étudiants ont eu tendance à fournir des classes plus petites que celles de la référence. Le rappel est plus sensible aux opérations d'éclatement.

Les valeurs obtenues par nos mesures de précision (autour de $0,85$) sont plus conformes à l'appréciation intuitive que pourrait faire un ingénieur de la connaissance que les mesures classiques : même s'il y a des retouches à faire sur les classes fournies par les étudiants, dire que seule la moitié est bonne (précision classique inférieure à $0,45$) est un jugement sévère qui ne rend pas compte de l'intérêt que présentent ces classes.

Sorties	P	Précision classique	R	Rappel classique	FM	FM classique
S1	0,85	0,41	0,48	0,4	0,61	0,41
S2	0,86	0,45	0,49	0,45	0,62	0,44
S3	0,83	0,34	0,39	0,36	0,53	0,34

TABLE 3 – Résultats de la deuxième expérience

5 Travaux existants et discussion

Nos propositions s'inscrivent dans le cadre de l'évaluation d'ontologies, notre objectif étant de proposer à terme une évaluation globale des ontologies fondée sur l'évaluation de leurs composantes. Plusieurs méthodologies ont été proposées pour l'évaluation d'ontologies (Hartmann *et al.*, 2005; Brank *et al.*, 2005). Elles évaluent selon les cas les ontologies (i) selon des critères intrinsèques propres à qualifier une ontologie, nous citons par exemple les travaux de Fox *et al.* (1997); Welty & Guarino (2001); Lozano-Tello & Gómez-Pérez (2004); Burton-Jones *et al.* (2005) (ii) selon un critère particulier tels que l'adéquation par rapport à un concept clé, la densité, (Ding *et al.*, 2004; Alani & Brewster, 2005; Alani *et al.*, 2006), (iii) dans le cadre d'une application (Porzel & Malaka, 2004; Patel *et al.*, 2004; Brewster *et al.*, 2004) ou (iv) par rapport à une référence (Maedche & Staab, 2002; Brank, 2006). C'est plus précisément dans la dernière catégorie que nous nous positionnons. Les travaux de Maedche & Staab (2002) proposent des mesures de similarité entre deux ontologies, en tenant compte de l'aspect lexical ou des instances ainsi que des structures des ontologies. Nos propositions se distinguent par la prise en compte du caractère relatif de la référence en adaptant les sorties du système. Nous dissocions également, dans l'évaluation des classes, l'aspect terminologique de l'aspect structurel. L'intérêt de nos propositions est de faire une évaluation du niveau terminologique des concepts, ce qui permet d'évaluer des systèmes autres que les outils d'acquisition d'ontologies.

Poibeau *et al.* (2002) proposent une évaluation des classes sémantiques en utilisant des mesures classiques de précision et rappel. Les mesures sont calculées par rapport aux classes définies préalablement par un expert. Ce travail présente les résultats par classe et ne permet pas d'avoir une vision globale sur la pertinence de l'acquisition. Notre proposition permet de calculer une mesure globale qui rend compte de la pertinence de la totalité des classes. De plus le processus d'acquisition de classes sémantique est évalué dans une perspective d'extraction d'information, le "gold standard" est une liste plate de classes. Nous prenons en compte dans nos propositions la structure de l'ontologie de référence pour mesurer la pertinence des classes sémantiques dans une perspective de construction d'ontologies.

6 Conclusion et perspectives

Nous avons décomposé le problème de l'évaluation des ontologies en différents sous-problèmes, nous avons mis l'accent sur l'évaluation des classes sémantiques et proposé un protocole d'évaluation comparative permettant de mettre en correspondance les classes fournies par des systèmes d'acquisition et les concepts d'une ontologie de référence sur la base des termes qu'ils contiennent ou qui leur sont associés.

La difficulté de cette comparaison vient de ce que les classes peuvent ne correspondre que partiellement aux concepts de la référence. Dans certains cas, la correspondance optimale suppose d'éclater certaines classes en différentes sous-classes qui sont mises en correspondance avec des concepts différents. Dans d'autre cas, il faut au contraire regrouper plusieurs classes pour les mettre en correspondance avec un unique concept. Ces opérations d'éclatement et de regroupement permettent en réalité d'ajuster la sortie des systèmes à l'ontologie de référence, ce qui se justifie puisque la granularité de la description de l'ontologie de référence est pour partie arbitraire.

Nous avons proposé une mesure de pertinence d'un ensemble de classes par rapport à une référence qui tient compte de cet ajustement et montre des résultats encourageants.

Ce travail vise à permettre de procéder à des évaluations d'outils d'acquisition de classes sémantiques dans le cadre du programme Quaero. Nous avons ici mis l'accent sur la qualité de la classification proposée en négligeant le fait que la correspondance entre termes peut elle-même n'être que partielle (un terme de la sortie est parfois une variante de celui de la référence). Nous nous proposons donc d'intégrer dans les mesures une distance entre termes comme proposé dans (Nazarenko & Zargayouna, 2009).

Références

ALANI H. & BREWSTER C. (2005). Ontology ranking based on the analysis of concept structures. In *K-CAP*, p. 51–58.

ALANI H., BREWSTER C. & SHADBOLT N. (2006). Ranking ontologies with aktiverank. In *ISWC'06*.

BRANK J. (2006). Gold standard based ontology evaluation using instance assignment. In *EON 2006 Workshop*.

BRANK J., GROBELNIK M. & MLADENIC D. (2005). A survey of ontology evaluation techniques. In *SiKDD*.

BREWSTER C., ALANI H. & DASMAHAPATRA A. (2004). Data driven ontology evaluation. In *LREC'04*.

BURTON-JONES A., STOREY V. C., SUGUMARAN V. & AHLUWALIA P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data Knowledge Engineering*.

DING L., FININ T., JOSHI A., PAN R., COST S. R., PENG Y., REDDIVARI P., DOSHI V. & SACHS J. (2004). Swoogle : a search and metadata engine for the semantic web. In *CIKM '04*, p. 652–659 : ACM Press.

FAURE D. & NÉDELLEC C. (1998). Asium : Learning subcategorization frames and restrictions of selection. In *ECML 98*.

FOX M. S., BARBUCEANU M., GRUNINGER M. & LIN J. (1997). An organization ontology for enterprise modelling. In *ICEIMT'97* : Springer.

HARTMANN J., SPYNS P., GIBOIN A., MAYNARD D., CUEL R., SUAREZ-FIGEROA M.-C. & SURE. Y. (2005). Methods for ontology evaluation. EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB Deliverable D1.2.3 (WP 1.2).

LOZANO-TELLO A. & GÓMEZ-PÉREZ A. (2004). Ontometric : A method to choose the appropriate ontology. *Journal of Database Management. Special Issue on Ontological analysis, Evaluation, and Engineering of Business*, p. 1–18.

MAEDCHE A. & STAAB S. (2002). Measuring similarity between ontologies. In *EKAW'02*.

MARTIN A. F., GAROFOLO J. S., FISCUS J. C., LE A. N., PALLETT D. S., PRZYBOCKI M. A. & SANDERS G. A. (2004). Nist language technology evaluation cookbook. In *LREC'04*.

NAZARENKO A. & ZARGAYOUNA H. (2009). Evaluating term extraction. In *RANLP'09*.

PATEL C., SUPEKAR K., LEE Y. & PARK E. (2004). Ontokhoj : a semantic web portal for ontology searching, ranking and classification. In *ACM Web Information. and Data Management*.

POIBEAU T., DUTOIT D. & BIZOUARD S. (2002). évaluer l'acquisition semi-automatique de classes sémantiques. In *TALN 2002*.

PORZEL R. & MALAKA R. (2004). A task-based approach for ontology evaluation. In *Proc. of ECAI 2004 Workshop on Ontology Learning and Population*.

STUFFLEBEAM D. (1974). Meta-evaluation. *Occasional Paper Series, Kalamazoo MI : Western Michigan University Evaluation Center*.

WELTY C. & GUARINO N. (2001). Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering*, **39**(1), 51–74.

WU Z. & PALMER M. (1994). Verb semantics and lexical selection. In *ACL*, p. 133–138.

ZARGAYOUNA H. & NAZARENKO A. (2010). Evaluation of textual knowledge acquisition tools : a challenging task. In *LREC 2010*, p. to appear, Malta : ELRA.

Remerciements

Ce travail a été réalisé en partie dans le cadre du programme Quaero, financé par OSEO, l'agence française pour l'innovation.