TALN 2010, Montréal, 19–23 juillet 2010

Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique*

Eric Charton[1] Michel Gagnon[1] Benoit Ozell[1]
(1) École Polytechnique, 2900 boul. Edouard Montpetit, Montréal, Canada H3T 1J4
{eric.charton, michel.gagnon, benoit.ozell}@polymtl.ca

Résumé. L'étiquetage sémantique consiste à associer un ensemble de propriétés à une séquence de mots contenue dans un texte. Bien que proche de la tâche d'étiquetage par entités nommées, qui revient à attribuer une classe de sens à un mot, la tâche d'étiquetage ou d'annotation sémantique cherche à établir la relation entre l'entité dans son texte et sa représentation ontologique. Nous présentons un étiqueteur sémantique qui s'appuie sur un étiqueteur d'entités nommées pour mettre en relation un mot ou un groupe de mots avec sa représentation ontologique. Son originalité est d'utiliser une ontologie intermédiaire de nature statistique pour établir ce lien.

Abstract. Semantic labelling consist to associate a set of properties to a sequence of words from a text. Although its proximity with the named entity labelling task, which is equivalent to associate a class meaning to a sequence of word, the task of semantic labelling try to establish the relation between the entity in the text and it's ontological representation. We present a semantic labelling system based on a named entity recognition step. The originality of our system is that the link between named entity and its semantic representation is obtained trough the use of an intermediate statistical ontology.

Mots-clés : Étiqueteur sémantique, Entités nommées, Analyse sémantique, Ontologie.

Keywords: Semantic parser, Named entities, Semantic annotation.

1 Introduction

L'acquisition de masses de métadonnées depuis les contenus rendus disponibles sur le web a favorisé l'émergence de nouvelles formes d'applications sémantiques. Les thématiques des métadonnées disponibles sont larges : elles peuvent concerner des personnes, des objets, des lieux, mais aussi des notions plus abstraites telles que des concepts, des modalités, ou de simples données numériques (le chiffre d'affaire d'une entreprise, la date de naissance d'une personne). Ces représentations conceptuelles jouent un rôle essentiel dans les applications sémantiques du TAL. Elles sont notamment déployées dans deux activités d'automatisation bien distinctes que sont l'*analyse sémantique* et l'*étiquetage sémantique*. L'*analyse sémantique* consiste à attribuer un sens aux composants d'une phrase (la nature des événements verbaux, les circonstances de l'événement, les acteurs impliqués) (Zouaq & Gagnon, 2010) . L'activité *d'étiquetage sémantique* diffère de l'analyse, en associant des propriétés au mot ou au groupe de mots.

Ainsi, ces deux tâches, bien que très complémentaires et visant toutes deux à associer une méta-connaissance à des objets textuels, attribuent du sens sur des registres différents. Dans un composant de compréhension

(*) Ce travail s'inscrit dans le cadre du projet Gitan, soutenu par Unima Inc et Prompt Québec

d'un système de dialogue par exemple, il est aussi important de connaitre l'objectif sémantique (*je souhaite réserver dans l'hôtel Beau Soleil le mois prochain*, soit ***Action* :Réserver ;*Temps* :30j**), qui relève de l'analyse (les informations sont contenues dans le texte), que le détail des entités concernées (***EN* :Hôtel Beau Soleil ;*Propriétés* :types de chambres,prix des chambres**) qui relève d'une mise en relation d'une étiquette avec des connaissances externes (les informations sont extérieures au texte).

Nous proposons de compléter un système d'*étiquetage par entités nommées* (EEN) par un système *d'étiquetage sémantique* (ES). L'originalité du système présenté réside dans l'utilisation d'une ***ontologie de liaison*** pour relier l'entité dans son texte à son instance d'une ***ontologie descriptive***. Pour nos besoins expérimentaux, notre étiqueteur est relié à des ontologies existantes et normalisées.

L'article est structuré comme suit. Dans un premier temps nous examinons les propositions existantes de mise en relation d'une entité nommée avec sa description sémantique. Dans un second temps nous décrivons notre système et son *ontologie de liaison*. Dans un troisième temps nous procédons à une évaluation et mesurons la capacité de notre système à mettre en relation une instance ontologique avec un terme contenu dans un texte. Le cadre expérimental retenu est celui de la campagne EEN d'ESTER 2. Enfin, nous concluons cet article en commentant les résultats que nous avons obtenus.

2 Étiquetage d'entités nommées versus étiquetage sémantique

La masse de connaissances inter-connectées et normalisées disponibles sur le web, par exemple à travers le réseau *LinkedData*[1] permet aujourd'hui de décrire avec précision les attributs de nombreux objets ou concepts. On pourrait envisager que la mise en relation de ces connaissances avec des mots dans un texte est le prolongement de la tâche d'EEN. Pourtant il existe une différence essentielle entre l'*entité nommée* (EN) dans son texte et sa représentation ontologique : affecter un taxon[2] à un objet textuel se fait en observant son contexte. L'information qui est présente autour de lui (les mots, la nature des lettres - majuscules, minuscules) permettent de caractériser sa nature. Il en va tout autrement pour une représentation sémantique de ce même objet qui fait intervenir une masse de caractéristiques qui sont justement absentes du texte. Avec la phrase *Un Airbus A380 a décollé de Roissy ce matin*, un système d'EEN correctement entrainé et configuré, sait que *Roissy* décrit un aéroport et non une ville ou une station de métro, que l'Airbus est un véhicule. Mais rien ne nous indique dans cette phrase la date d'ouverture de l'aéroport, le nom des compagnies qui l'utilisent ou les caractéristiques de l'avion.

2.1 Améliorations de l'arbre taxonomique d'étiquetage

Une première étape d'amélioration de la description sémantique des EN a consisté à améliorer les arbres taxonomiques d'étiquetage, principalement pour répondre à des tâches de Question Réponses. Si nous reprenons l'exemple plus haut *Un Airbus A380 a décollé de Roissy ce matin.*, nous obtenons selon la norme taxonomique d'EEN l'étiquetage suivant :

Un <ent=prod.vehicule>Airbus A380<ent/> a décollé de <ent=loc.fac>Roissy<ent/> ce matin.

Soit l'étiquette (*prod.vehicule*) attribuée à un avion, (*loc.fac*) à un aéroport. On peut envisager un accroissement de la granularité taxonomique (la qualité du lieu *loc.fac.airport*, ou du véhicule *prod.vehicule.plane*).

[1] LinkedData est un projet qui vise à exposer sur le Web des données structurées avec les technologies du Web sémantique, en particulier RDF, suivant les principes proposés par Tim Berners-Lee.

[2] Dans cet article, le terme *taxon* décrit une étiquette complète d'une norme d'étiquetage d'entité nommées : pers.hum, loc.fac, org.com sont des exemples de taxons.

Plusieurs auteurs se sont engagés dans cette voie. La proposition de (Sekine *et al.*, 2002) consistait à développer un arbre taxonomique de 150 types. Aujourd'hui, des arbres de 200 types et plus sont couramment utilisés (voir par exemple (Rosset *et al.*, 2007)) dans le cadre des campagnes d'évaluation (Turmo *et al.*, 2009). La diversité de nature des entités utiles conduit aussi des auteurs à exploiter des taxonomies dérivées de contenus encyclopédiques. Un des premiers systèmes d'étiquetage permettant de mettre en relation un objet textuel et sa catégorie taxonomique dans Wikipédia est présenté dans (Bunescu & Pasca, 2006). Dans cette expérience, sur les 59759 catégories de la version de Wikipédia utilisée, 2037 sont conservées. Les travaux de (Kazama & Torisawa, 2007) sont de nature similaire : ils consistent à exploiter, en utilisant le même processus d'extraction lexicale que Bunescu, un ensemble de 2000 labels de catégories issus de Wikipédia pour étiqueter le corpus de la campagne EEN de Conll (Tjong & Meulder, 2003). D'abord avec les 4 classes normalisées de cette campagne (Per, Loc, Org, Misc), puis en appliquant la totalité des classes. Les auteurs prennent néanmoins soin de préciser[3] que cette taxonomie étendue n'est pas forcément très fiable à l'usage. Finalement, bien que ces recherches conduisent à améliorer la granularité des classes attribuées à des entités - donc, de facto, à affiner la compréhension de leur sens - elles n'en demeurent pas moins des applications de classification de mots, selon un arbre taxonomique restreint, conçu le plus souvent souvent pour répondre aux besoins formulés dans les campagnes d'évaluation (Charton & Torres-Moreno, 2009). Et cette augmentation se heurte à deux limitations : elle risque de rendre la mise au point d'étiqueteurs d'EN très délicate[4] et un arbre taxonomique ne permet pas de représenter les prédicats d'ordre supérieurs permis par les standards de définition actuels des ontologies.

2.2 Etiquetage sémantique

Une entité décrite sémantiquement dans une ontologie a pour particularité d'être associée à un ensemble de propriétés décrites par des prédicats de premier ordre (par exemple le triplet *{sujet :nom de ville ;prédicat : population ;objet :valeur numérique}*) et par des graphes lorsque les propriétés sont d'ordre supérieur (exemple, une relation entre toutes les villes de plus de 100 000 habitants). L'instance de DBpedia représentant l'aéroport de Roissy contient à elle seule une centaine d'informations complémentaires sous forme de triplets RDF[5]{sujet, prédicat, objet} (des coordonnées géographiques, des noms de compagnies). La particularité de *l'étiquetage sémantique* est qu'il consiste donc en une mise en relation d'un terme avec un graphe, alors que *l'étiquetage d'une entité nommée* relève de la classification de ce terme. La perspective de transformer directement un étiqueteur d'EN en étiqueteur sémantique est donc très hypothétique pour le domaine applicatif que nous décrivons.

En revanche, on peut envisager un module de complément sémantique d'un système d'EEN qui soit en mesure de relier un objet textuel dans son document avec une description formelle et conceptuelle de cet objet. On retrouve cette idée dans le système SemTAG (Dill *et al.*, 2003). L'idée de SemTAG est de mettre en relation automatiquement un terme contenu dans un document, avec sa description conceptuelle, via une URI[6]. Le problème du repérage et de la désambiguisation de l'entité dans le texte et de sa mise en

[3] *"These category labels seem to be usefull, altough there is no guarantee that the extracted category label is correct for each candidate"*.

[4] Par exemple en augmentant le nombre de micro-classes. Les micro-classes rendent la mise au point de classifieurs numériques difficiles pour des raisons théoriques mais aussi pratiques, lorsque le phénomène à modéliser est trop rare pour apparaitre suffisamment dans un corpus d'apprentissage.

[4] http ://DBpedia.org/page/Paris-Charles_de_Gaulle_Airport

[5] Un triplet RDF est un prédicat normalisé par le W3C pour les applications dîtes du Web Sémantique.

[6] Un URI, de l'anglais Uniform Resource Identifier, soit Identifiant Uniforme de Ressource, est une courte chaîne de caractères identifiant une ressource sur un réseau, par exemple le web.

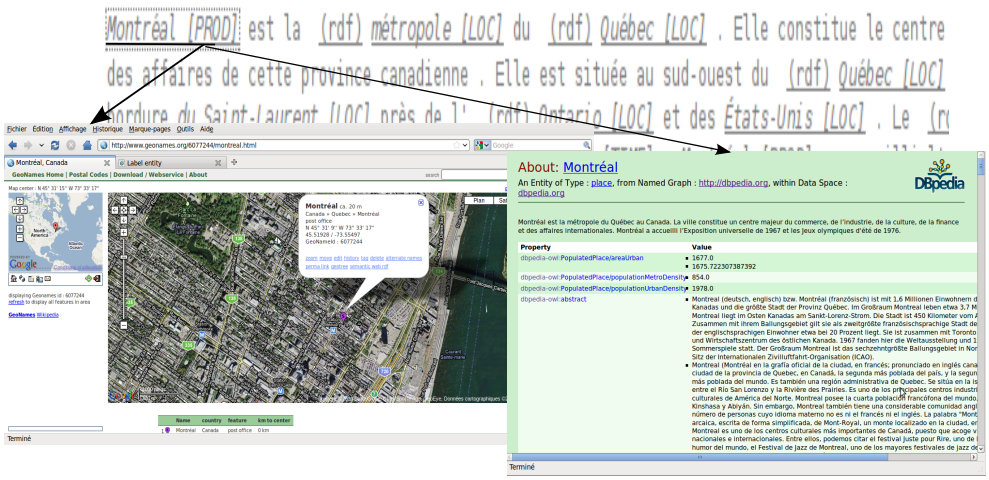

FIG. 1 – Le système proposé : l'entité nommée est associée à ses descriptions dans une ou plusieurs ontologies du réseau LinkedData. Ici DBpedia et Géonames sont reliés à l'entité nommée Montréal du texte.

relation avec une ontologie, est résolu en une seule fois par un algorithme qui combine une mesure de similarité cosinus et un calcul de probabilité bayésienne entre les mots contextuels de l'entité candidate et des mots contextuels probables conservés dans l'ontologie. Ce système diffère de celui que nous présentons car il hybride la tâche d'EEN et celle d'étiquetage sémantique, en conservant les mots contextuels dans la fiche ontologique. Mais les mots contextuels associés aux objets décrits ne sont que très rarement présents dans une ontologie standard. Ce modèle n'est donc pas adaptables aux nouvelles données OWL[7] qui apparaissent régulièrement. Ailleurs, (Kiryakov *et al.*, 2004) argumente de manière convaincante, en présentant la plateforme KIM, que le processus d'annotation sémantique fait nécessairement entrer en jeu un dispositif d'EEN, une ontologie (ou une base de données), et des normes taxonomiques. Pour cet auteur, la fonction d'extraction de l'entité dans son texte ne peut être réalisée avec des éléments ontologiques. D'autres systèmes plus récent adoptent une architecture à plusieurs niveaux qui fait entrer en jeu une première étape d'EEN, comme le projet ALVIS (Nazarenko *et al.*, 2006). Face à cette architecture dominante d'étiqueteurs sémantiques associés à un étiqueteur d'EN qui semble progressivement se dessiner, des alternatives sont proposées. Le système de (Brun & Hagège, 2009) ne cherche plus à établir un lien entre entité et ontologie, mais à relier entre elles plusieurs EN pour créer une information très proche de celle que l'on retrouve dans un triplet RDF.

3 Système proposé

L'idée principale que nous proposons dans cet article consiste à relier une EN dans son texte avec sa représentation sémantique. Notre contribution consiste à séparer les deux tâches d'EEN et d'ES, en introduisant un degré de connaissance intermédiaire que nous appelons ***ontologie de liaison***, entre l'étiqueteur d'entités nommées et l'ontologie. La nature de cette *ontologie de liaison* est à la fois statistique et conceptuelle. Nous appellerons les instances de l'*ontologie de liaison* des ***métadonnées*** pour les distinguer des ***instances*** de ***l'ontologie descriptive***. La *nature statistique* des *métadonnées* indique qu'elles représen-

[7]OWL est le format de description d'ontologies complexes regroupant des triplets RDF, défini par le W3C dans le cadre du Web dît Sémantique.

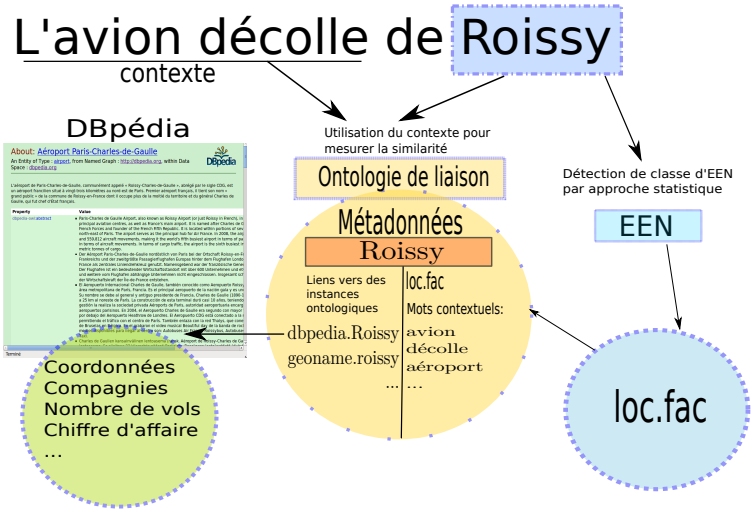

FIG. 2 – Schéma de principe de l'étiqueteur sémantique proposé

tent chacune un objet et lui associent des informations telles que des sacs de mots, des poids pour ces mots, des réseaux lexicaux et non des attributs sémantiques. Ces informations permettront d'établir un lien entre une entité détectée dans un texte et sa représentation ontologique. Chaque *métadonnée* est extraite depuis une fiche de encyclopédique de Wikipédia. Elle contient les mots qui composent le texte de cette fiche avec leurs poids $Tf.Idf$ calculé sur l'ensemble du corpus, et un ensemble de *formes de surfaces* (c'est à dire les écritures possibles) : par exemple le lieu *Aéroport Charles de Gaulle* peut aussi être écrit *Roissy Charles de Gaulle*, *Roissy*, *CDG*. Les *métadonnées* sont enrichies d'un lien vers une instance de l'*ontologie descriptive*, c'est à dire représentant un objet par les attributs descriptifs de sa nature (la hauteur d'un immeuble, la population d'une ville, etc). Dans le cadre expérimental de cet article, nous avons relié chaque *métadonnées* avec son instance correspondante dans DBpedia (Auer *et al.*, 2007). Mais le principe de notre système est que *l'ontologie de liaison* doit être considérée comme un pivot permettant de relier l'EN dans son texte à n'importe quelle source ontologique de description la mieux appropriée (DBpedia pour une personne, ou Bio2RDF pour un gène, par exemple). Nous voyons plusieurs avantages à cette architecture : elle permet tout d'abord de mettre en valeur les standards émergents (RDF, OWL) proposés par le W3C en matière de description ontologiques. Elle permet ensuite de séparer les tâches d'EEN et d'ES, qui relèvent d'une compréhension différente du texte, mais se complètent : le système d'EEN est efficace pour localiser l'entité dans le texte mais pauvre en informations, le système d'ES est performant pour attribuer de l'information à l'EN localisée mais est mal adapté à la détection.

Le système que nous proposons commence par étiqueter les EN contenues dans un texte libre, selon un procédé statistique. Il utilise pour cela une implémentation d'un système de détection d'EN à base de CRF[8]. Puis il met en relation les formes de surface correspondant aux EN détectées dans le texte, avec l'*ontologie de liaison*.

Pour finir, un algorithme de calcul de similarité associe à chaque EN localisée dans le texte, sa *métadonnée* correspondante exacte et désambiguïsée dans *l'ontologie de liaison*, ce qui permet d'établir un lien direct entre l'EN et son *instance* de l'*ontologie descriptive* (voir figure 2).

[8]Le CRF ou *Champ Conditionnel Aléatoire* est un modèle de graphe non orienté dans lequel chaque arrête représente une variable aléatoire et ou chaque sommet représente une relation de dépendance entre deux variables aléatoires. Ce modèle est bien adapté à l'identification de séquences de texte.

3.1 Construction de l'ontologie de liaison

Considérons le corpus Wikipédia qui est composé de plusieurs éditions linguistiques[9]. Chaque édition linguistique de Wikipedia contient des articles composés de mots, et qui sont identifiés par un titre non ambigu (par exemple *Paris (France)* ou *Paris (Texas)*). Ces articles peuvent être reliés à des pages d'*homonymie* qui listent tous les articles correspondant à un même nom, à des pages de *redirections* (qui associent à un article un nom alternatif, par exemple *Ville lumière→Paris*). Ils peuvent également être reliés via des liens dits *Interwikis* à leurs équivalents dans d'autres éditions linguistiques. La construction d'une *métadonnée* consiste à collecter toutes ces informations pour construire une représentation lexicale et statistique d'un article. Une description formelle de l'algorithme de construction d'une *métadonnée* E d'après un article encyclopédique D peut être résumée comme suit[10].

– Nous appellerons le corpus Wikipédia C, une édition linguistique C^l, les articles d'une édition linguistique de Wikipedia D. Les propriétés des articles seront décrites par $D = (D.t, D.w, D.l)$ telles que $D.t$ contient le titre, $D.w$ est la collection des mots contenus dans D, et $D.l$ est un ensemble de liens qui relie D à d'autres pages de C. Les liens de $D.l$ correspondent aux liens de D vers des pages de *synonymes, d'homonymie, de redirection* de C^l ou vers C (un lien Interwiki vers l'équivalent de D dans une autre édition linguistique de Wikipédia).
– La *métadonnée* E possède les propriétés $E = (E.t, E.w, E.r, E.rdf)$. On considère que E et D sont en relation si et seulement si $E.t = D.t$. $E.w$ contient des duplets composés de tous les mots de $D.w$ et de leur poids $Tf.Idf$ calculés d'après le corpus C^l. $E.r$ contient l'ensemble des *formes de surfaces* obtenues en appliquant des heuristiques d'extractions de ces formes aux pages reliées à D par $D.l$.

L'attribut $E.rdf$ de la métadonnée contient un lien vers le point d'entrée au format RDF d'une instance d'ontologie disponible sur le réseau *LinkedData*[11]. L'adoption de ce format pour E permet de rendre l'*ontologie de liaison* universelle et capable de mettre en relation une EN avec n'importe quel descriptif *LinkedDdata*. Pour les besoins de cette expérience $E.rdf$, est défini d'après $D.t$ en utilisant le fichier de correspondance[12] entre DBpedia et Wikipedia, que nous appellons R. Pour une instance de DBpedia B, R décrit tous les $B.t = D.t$. Nous savons que de $E \rightarrow D$ donc $E.t = D.t$, en conséquence $E.rdf \rightarrow B.t$.

3.2 Modules d'étiquetage et de mise en relation

Considérons une norme taxonomique d'étiquetage composé d'un ensemble L de labels l. Considérons une phrase S contenant une séquence de mots $s_{1...n}$. Dans un premier temps le système d'EEN recherche pour tout s_n ou suite de mots $s_{[n,m]}$, un label $l \in L$. On intitule fs la *forme de surface* composée du mot ou de la suite de mots étiquetés par l. Nous savons que EN (entité nommée) est équivalente à un duplet composé de l et de fs. Notre système fonctionne selon un processus en trois étapes successives que nous intitulons EEN, REN et LEN :

– 1 EEN : Détection et étiquetage de EN dans le texte en utilisant un étiqueteur.
– 2 REN : Identification dans l'*ontologie de liaison* de l'entrée correspondante à la *forme de surface* fs associée à EN, par un algorithme de mesure de similarité cosinus appliqué au contexte textuel $s_{1...n}$ de fs dans S.
– 3 LEN : Établissement du lien entre l'entité dans le texte et son instance dans DBpedia en utilisant la référence contenue dans une des *metadonnées* de l'*ontologie de liaison*.

[9] fr.wikipedia.org, en.wikipedia.org, par exemple, sont des éditions linguistiques différentes et forment des sous corpus indépendants au sein de l'ensemble Wikipédia.

[10] On pourra se reporter à (Charton, 2009) pour une description détaillée

[11] Par exemple http ://dbpedia.org/data/Spain.rdf est le *point d'entrée* de l'instance *Espagne* de DBpedia)

[12] wiki.DBpedia.org/Downloads34 fichier intitulé *Links to Wikipedia articles*

Etiqueteur d'entités nommés (EEN)

Pour les besoins d'un système complet d'étiquetage sémantique, l'étiqueteur d'entités nommées (EEN) peut être n'importe quel système capable de localiser fs et de lui attribuer une étiquette l. Dans le démonstrateur[13] qui découle de la présente communication, le système utilisé est une implémentation du système d'EEN robuste utilisé par le LIA[14] lors de la campagne ESTER 2(Galliano *et al.*, 2009). On s'en remetra à (Béchet & Charton, 2010) pour une description détaillée de EEN.

Mise en relation de l'EN avec *Ontologie de Liaison* (REN)

REN est le système de mise en relation de fs avec sa *metadonnée* représentante dans l'*ontologie de liaison*. REN utilise les duplets {mots,$Tf.Idf$} contenus dans $E.w$ d'une *métadonnée* E afin de calculer le degré de similarité entre $E.w$ et le contexte textuel de fs dans S.

Le système de détection est divisé en 3 algorithmes intitulés $REN.1$, $REN.2$ et $REN.3$:

1. $REN.1$: La fonction $Rm = f_{synsets}(fs)$ recherche dans les *métadonnées* tous les éléments $E.r = fs$ et retourne une liste de E dans Rm. Nous considérerons que si $|Rm| = 1$, Rm propose une unique *metadonnée* candidate E_c correspondant à EN (pas d'ambiguité de sens).
2. $REN.2$: La fonction $f_{simcos}(Rm, S)$ calcule la similarité cosinus entre les s de S et tout les $E.w$ des E de Rm. Considérant que la fonction de similarité cosinus $cos(E.w, S)$ existe, nous pouvons déterminer le score de l'entité candidate E_c de Rm en mesurant le cosinus de l'angle entre les vecteurs de poids des mots correspondant à $E_c.w.Tf.Idf$ de Rm et les $s.Tf.Idf$ de S. Nous obtenons alors une liste triée contenant toutes les *métadonnées* candidate E_c pour $EN \in S$ avec leur score de similarité.
3. $REN.3$: Détermine la meilleure *métadonnée* correspondant à EN en utilisant la formule $\hat{E} = argmax_{E_c}\ score(cos(E_c.w, S))$. Si $|Rm| > 1$, un ensemble de *métadonnées* homonymes sont disponibles : $REN.3$ recherche alors $\hat{E}$.

On observe les cas suivants dans lesquels aucun lien sémantique n'est établi entre EN et E. C'est le cas si $REN.1$ n'identifie pas de *métadonnée* candidate pour fs_m (la forme de surface étiquetée dans le texte n'existe pas dans les métadonnées). C'est également le cas quand nous appliquons un seuil de détection d à$REN.3$. Seuls les résultats de $d < argmax_{E_c}\ score(cos(E_c.w, S.s))$ seront alors retenus pour établir le lien entre EN et E. Ce dispositif de seuil permet notamment de rejeter les propositions d'étiquetage de *lien sémantique* peu fiables dues à un score final de similarité cosinus faible.

Établissement du lien sémantique(LEN)

LEN est la dernière étape triviale qui consiste à établir le lien entre l'entité dans le texte et son instance dans DBpedia en associant $E.rdf$ de la *metadonnée* $\hat{E}$ de l'*ontologie de liaison* avec EN de S. Ce qui permet de relier EN à son *instance* dans l'*ontologie descriptive*

4 Expériences et résultats

Nous proposons pour évaluer notre système d'*étiquetage sémantique* d'associer à des EN détectées dans un fichier de transcription de dialogues radiophoniques un lien vers leurs *instances* contenues dans une *ontologie descriptive*. C'est l'établissement de ce lien qui caractérise le principe de l'*étiquetage sémantique*.

[13]Consulter www.nlgbase.org/perl/nlgetiq.pl

[14]Ce système a été mis au point au LIA par Frédéric Béchet avec la contribution du premier auteur de cet article

NE	Test ESTER 2	métadonnées	Couverture (%)
Pers	1096	483	44%
Org	1204	764	63%
Loc	1218	1017	83%
Prod	59	23	39%
Total	3577	2287	64%

TAB. 1 – Mesure de couverture des entités contenues dans les *métadonnées* en regard du corpus d'évaluation ESTER 2.

Nous utilisons le corpus d'évaluation EEN de la campagne ESTER 2 pour mener ces expériences. La campagne d'évaluation ESTER 2 (Galliano *et al.*, 2009) a été organisée en 2009 conjointement par l'association Française de la communication parlée (AFCP) et la DGA, avec la collaboration de l'agence de distribution de ressources ELDA. Cette campagne était divisée en trois tâches, segmentation, transcription et étiquetage d'entités nommées. La tâche d'EEN était elle même divisée en 4 sous tâches, la première consistant à étiqueter un système corrigé manuellement (absence d'erreur de transcription) dit *ref trancript*. Les trois autres tâches recouraient à des systèmes contenant des erreurs de transcriptions d'importance croissante afin de mesurer la robustesse des étiqueteurs d'EN. Nous utilisons comme support de nos expériences le corpus de test manuellement corrigé de la première sous-tâche, la question de la robustesse d'EEN sortant du champs de cet article.

Nous complétons la tâche d'EEN d'ESTER 2 en évaluant la capacité de notre système à associer correctement une identité sémantique à une entité nommée déjà détectée par un système d'EEN. Ceci revient à affecter à une EN reconnue dans le corpus de test d'étiquetage d'ESTER un lien vers sa description ontologique dans DBpedia. Nous vérifions plusieurs aspects de notre système. En premier lieu que confronté à un homonyme, le système d'étiquetage sémantique relie correctement l'EN à son *instance ontologique* exacte, via *l'ontologie de liaison* grâce à l'algorithme de désambiguïsation par mesure de similarité cosinus. En second lieu, la capacité de la constante de seuil d permette bien de rejeter les propositions de lien sémantiques non pertinents.

Prise en compte du taux de couverture

On doit aussi vérifier que lorsqu'il n'existe pas de données sémantiques à relier à une EEN le système exploite correctement sa constante de seuil et ne propose aucun lien. En effet, la couverture d'une ressource ontologique ne peut être complète : certaines entités présentes dans le texte du corpus ESTER 2 ne seront pas représentés dans *l'ontologie de liaison* (principe des **mots hors vocabulaires** ou **OOV**). Ces mêmes entités, présentes dans le texte mais absentes de la connaissance ontologique, peuvent pourtant être correctement étiquetées lors de l'étape d'EEN qui agit par inférence d'après un contexte, et peut donc étiqueter correctement un groupe de mots sans le connaitre au préalable. Pour ces EN, il existe une impossibilité matérielle à être correctement associées à un lien sémantique puisque ce dernier n'existe pas. Pour évaluer ce degré d'impossibilité nous avons calculé le taux de couverture des objets connus dans les *métadonnées* de *l'ontologie de liaison* pour toutes les EN présentes dans le corpus dévaluation d'ESTER 2. Nous obtenons le résultat indiqué par le tableau 1.

Méthode

La mesure de précision et de rappel est réalisée d'aprés un complément d'étiquetage de la référence du corpus dévaluation d'ESTER 2, indiquant pour chaque EN son lien vers les *métadonnées*. Ce complément a été réalisé par méthode semi-automatique. Notre évaluation consiste à mesurer dans un premier temps pour les 64% d'entités du corpus de test EEN ESTER 2 possédant une représentation correspondante dans l'*ontologie de liaison* la capacité du système à établir un lien entre une *métadonnée* de l'ontologie de liaison et la forme de surface de l'EN dans son texte. C'est à dire le bon fonctionnement de l'algorithme REN. Dans un second temps, nous évaluons la capacité de REN avec la constante de seuil d à rejeter les EN qui ne possèdent pas de représentation ontologique.

4.1 Résultats

Les résultats obtenus sont indiqués dans le tableau 2. Les résultats de mise en relation d'une entité déjà correctement étiquetée, avec sa représentation dans DBpedia, en utilisant la ressource *ontologique de liaison* sont précis à 95%. Ce résultat indique que la méthode de calcul de similarité cosinus désambiguïse correctement l'entité dans son contexte. Les résultats de mise en relation d'une entité dans un texte ouvert, c'est à dire dans lequel toutes les représentations ontologiques ne sont pas forcément disponibles, laisse apparaître une légère baisse de précision et un rappel affaibli de 1 à 5% pour les trois entités principales (Pers, Org, Loc). L'affectation d'un lien sémantique à l'entité Prod est particulièrement difficile en milieu ouvert. Ce dernier cas peut être considéré comme particulier, étant à la fois sous-représenté dans le corpus ESTER 2 et déjà connu comme présentant une difficulté d'apprentissage[15].

NE	[TEP]	Précision	Rappel	[TEC]	Précision	Rappel
Pers	483	0,96	0,96	1096	0,94	0,91
Org	764	0,93	0,91	1204	0,96	0,90
Loc	1017	0,97	0,94	1218	0,94	0,92
Prod	23	0,96	0,60	59	0,61	0,50
Total	2287	0,95	0,93	3577	0,94	0,9

TAB. 2 – [TEP] Résultats de l'étiqueteur sémantique appliqué uniquement aux étiquettes de référence de ESTER 2 ayant une correspondance dans les *métadonnées* - [TEC] Résultats avec application au corpus d'évaluation de ESTER 2 complet.

5 Conclusion et perspectives

Nous avons présenté un système d'étiquetage sémantique (SE) utilisable pour compléter un système d'étiquetage par entités nommées (EEN). Ce système [16] utilise des *métadonnées* de nature statistique contenues dans une *ontologie de liaison* construite d'après la ressource encyclopédique Wikipédia. Une mesure de similarité cosinus a été mise en oeuvre pour établir un lien entre une entité nommée (EN) étiquetée dans sa phrase et sa représentation ontologique standardisée dans la ressource DBpedia. Cette méthode nous a permis d'associer à 94% des entités nommées (EN) du corpus d'évaluation de la campagne ESTER 2, connues dans Wikipédia, un lien vers leur description ontologique dans DBpedia. Nous envisageons maintenant d'intégrer dans un système complet, un étiqueteur d'EN associé à l'étiqueteur sémantique présenté ici, complétés par un un analyseur sémantique.

[15] Voir les résultats détaillés de la campagne ESTER sur ce sujet.

[16] Utilisable en ligne sur www.nlgbase.org/perl/nlgetiq.pl.

Références

AUER S., BIZER C., KOBILAROV G., LEHMANN J. & IVES Z. (2007). DBpedia : A Nucleus for a Web of Open Data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, p. 11–15 : Springer.

BÉCHET F. L. & CHARTON E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. In *ICASSP 2010*, Dallas : ICASSP.

BRUN C. & HAGÈGE C. (2009). Semantically-Driven Extraction of Relations between Named Entities. In *Cicling 2009*, Mexico : Cicling.

BUNESCU R. & PASCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6.

CHARTON E. (2009). Combinaison de contenus encyclopédiques multilingues pour une reconnaissance d'entités nommées en contexte. In *Recital*, number 1, p. 24–26.

CHARTON E. & TORRES-MORENO J. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Taln 2010*, volume 1, p. 24–26 : TALN.

DILL S., EIRON N., GIBSON D., GRUHL D., GUHA R., JHINGRAN A., KANUNGO T., RAJAGOPALAN S., TOMKINS A., TOMLIN J. & OTHERS (2003). SemTag and Seeker : Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, p. 186 : ACM.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *European Conference on Speech Communication and Technology*, p. 2583–2586 : Interspeech 2010.

KAZAMA J. & TORISAWA K. (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 698–707.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics : Science, Services and Agents on the World Wide Web*, **2**(1), 49–79.

NAZARENKO A., NEDELLEC C., ALPHONSE E., AUBIN S., HAMON T. & MANINE A. P. (2006). Semantic annotations in the Malvis Project. In *International Workshop on Intelligent Information Access*, Helsinky.

ROSSET S., GALIBERT O., ADDA G. & BILINSKI E. (2007). The LIMSI participation to the QAst track. In *Working Notes of CLEF Workshop, ECDL conference*, number System 1, p. 414–423 : Springer.

SEKINE S., SUDO K. & NOBATA C. (2002). Extended named entity hierarchy. In *Proceedings of the LREC-2002 Conference*, p. 1818–1824 : Citeseer.

TJONG E. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *In CoNLL*.

TURMO J., COMAS P., ROSSET S., GALIBERT O., MOREAU N., MOSTEFA D., ROSSO P. & BUSCALDI D. (2009). Overview of QAST 2009. In *Proceedings of the CLEF 2009 Workshop on Cross-Language Information Retrieval and Evaluation*.

ZOUAQ A. & GAGNON M. (2010). Semantic Analysis using Dependency-based Grammars and Upper-Level Ontologies. In *CICLING 2010* : CICLING.