Peut-on évaluer les outils d'acquisition de connaissances à partir de textes ?

Haifa Zargayouna, Adeline Nazarenko

Peut-on évaluer les outils d'acquisition de connaissances à partir de textes ?

Haïfa Zargayouna, Adeline Nazarenko

LIPN, Université Paris 13 - CNRS UMR 7030,
99 av. J.B. Clément, 93440 Villetaneuse, France.
prenom.nom@lipn.univ-paris13.fr

Résumé. Malgré les années de recul et d'expériences accumulées, il est difficile de se faire une idée claire de l'état d'avancement des recherches en acquisition de connaissances à partir de textes. Le manque de protocoles d'évaluation ne facilite pas la comparaison des résultats. Nous développons, dans cet article, la question de l'évaluation des outils d'acquisition de terminologies et d'ontologies en soulignant les principales difficultés et en décrivant nos premières propositions dans ce domaine.

1 Introduction

Malgré les années de recul et d'expériences accumulées, il est difficile de se faire une idée claire de l'état d'avancement des recherches en acquisition de connaissances. Ces connaissances sont généralement utilisées comme ressources sémantiques pour des applications et le fait que ces travaux soient très orientés vers des applications ne facilite pas la comparaison des résultats.

Nous mettons l'accent sur deux types de ressources : les terminologies et les ontologies. Même si les premières sont souvent utilisées pour l'acquisition des secondes, nous considérons ici les deux problèmes d'acquisition séparément. Nous considérons que du point de vue de l'évaluation, il y a une parenté entre les deux tâches et qu'il est intéressant de faire le parallèle. Les difficultés sont les mêmes en effet. A ce jour, les outils d'acquisition terminologique et ontologique ont souvent été développés en fonction d'une application particulière sans qu'on sache apprécier leur portabilité à d'autres corpus (genre et taille), à d'autres domaines scientifiques ou techniques, et à d'autres tâches. On ne sait pas évaluer les outils terminologiques et ontologiques parce qu'on a du mal à définir ce que devraient être leurs résultats (qu'est-ce qu'une "bonne" terminologie ? sur quel critère comparer deux ontologies ?). Nous faisons l'hypothèse que les deux questions peuvent s'enrichir l'une l'autre : les solutions terminologiques peuvent aider à résoudre le problème ontologique et vice-versa[1].

Les deux premières parties de cet article présentent un état des lieux de l'évaluation et souligne les difficultés de l'évaluation des outils d'acquisition terminologique et ontologique. La troisième partie présente les pistes que nous suivons pour surmonter ces difficultés. Le volet terminologique du travail est plus avancé mais nous montrons qu'il éclaire le problème ontologique.

2 Premières expériences

Même si de nombreuses recherches ont porté sur l'acquisition de connaissances à partir de textes, il y a eu peu d'effort collectif pour définir un cadre d'évaluation adapté à ce type de travaux, à la différence de ce qui est fait dans d'autres sous-domaines du traitement automatique des langues. Nous présentons dans ce qui suit quelques expériences d'évaluation dans le cadre de campagnes d'évaluation ou plus ponctuellement au travers des applications.

2.1 Les campagnes d'évaluation

Les campagnes d'évaluation visent généralement à évaluer un ensemble de systèmes, sur une tâche clairement spécifiée et à partir d'un jeu de données commun, en classant les résultats obtenus par les différents systèmes ou en les comparant à une référence. Au final, on obtient les mesures de performances des systèmes et leur classement pour la tâche considérée.

[1] Nous nous intéressons à ces questions dans le cadre du programme Quaero (http ://www.quaero.org/modules/movie/scenes/home) qui considère l'évaluation de la contruction de terminologies et de l'acquisition d'ontologies comme deux tâches distinctes. Nous présentons dans cet article, les débuts de nos travaux effectués dans ce cadre.

2.1.1 Les campagnes d'évaluation des outils d'acquisition terminologique

L'acquisition de terminologie a déjà fait l'objet de campagnes d'évaluation qui ont permis de débroussailler les questions liées à l'évaluation des outils d'acquisition.

CESART c'est probablement la plus aboutie des campagnes dans le domaine de la terminologie computationnelle. Elle devait comporter initialement trois sous-tâches : l'extraction terminologique dans le but de construire des terminologies, l'indexation contrôlée et l'extraction de relations sémantiques entre termes (Mustafa el Hadi et al., 2006). Malheureusement CESART a réuni peu de participants avec quatre systèmes participant à la première tâche, aucun à la deuxième et seulement un pour la troisième. C'est le protocole élaboré pour la première tâche qui est le plus intéressant. Les résultats des extracteurs de termes sont évalués par comparaison avec une liste de termes établie par avance par des experts (en pratique dans CESART, ce sont des terminologies préexistantes qui ont été utilisées). Un corpus d'acquisition relevant du même domaine que la terminologie de référence est fourni aux systèmes comme données d'entrée, et ceux-ci doivent en extraire une liste de termes. Cette liste de termes candidats est comparée à la liste de référence. L'une des originalités de CESART a été de considérer la pertinence d'un candidat terme sur une échelle à cinq valeurs plutôt que comme une valeur booléenne. Un terme est considéré comme pertinent s'il apparaît tel que dans la terminologie de référence mais aussi, à un moindre degré, s'il est composé de mots qui relève du vocabulaire de la terminologie de référence. Pour chaque degré de pertinence, une mesure de précision spécifique est calculée. Cette campagne a également permis de mettre en évidence l'hétérogénéité des résultats fournis par les différents systèmes, du point de vue de la longueur des listes de candidats termes fournies. Les extracteurs ont été évalués sur des échantillons de 10 000 termes mais certains ont fourni 20 fois plus de candidats.

CoRRecT Bien que ce ne soit pas une campagne à proprement parler, CoRReCT propose un jeu de test et un protocole intéressants (Enguehard, 2003). L'objectif est d'évaluer la tâche de reconnaissance de termes en corpus, qui se rapproche de l'indexation contrôlée de documents. Les systèmes prennent en entrée un corpus et une terminologie relevant du même domaine et ils doivent indexer le corpus avec les termes de la liste fournie. Comme les termes de la références n'apparaissent pas toujours sous la même forme dans le corpus, les systèmes doivent souvent reconnaître les termes sous des formes variantes. CoRRecT a la particularité de reposer sur une construction incrémentale du corpus annoté de référence. Au départ le corpus de référence est le corpus brut qui est fourni aux participants mais il s'enrichit d'annotations de référence chaque fois qu'un nouveau système participe à l'évaluation. Lorsqu'un nouveau système soumet sa liste d'annotations, celles-ci confrontées avec l'état courant du corpus de référence. Les nouvelles propositions d'annotation sont évaluées manuellement et celles qui sont validées viennent enrichir le corpus de référence. Le corpus de référence comporte donc l'union des propositions d'annotations des différents systèmes une fois celles-ci validées. Les résultats du i^{eme} système sont comparés (en termes de précision et de rappel) avec l'union des résultats validés des i premiers systèmes. Malheureusement, là encore, seulement trois systèmes ont participé à CoRRecT.

NTCIR-TERMREC En 1999, La campagne japonaise d'évaluation en recherche d'information (NTCIR) a intégré une tâche d'acquisition terminologique qui se décomposait elle-même en trois sous-tâches distinctes : l'extraction de termes, l'extraction de mots-clefs et l'analyse de rôles (Kageura et al., 2000). Les résultats étaient évalués sur la base d'une comparaison avec une référence. Peu d'information est disponible sur cette expérience, mais l'organisateur lui-même s'est déclaré déçu par l'interprétation essentiellement quantitative qui a été faite des résultats. A noter également que les tâches terminologiques n'ont pas été reprises dans les éditions ultérieures de NTCIR.

2.1.2 Les campagnes d'évaluation des outils de construction d'ontologies

Avec l'essor des recherches visant à promouvoir le web sémantique, une réflexion sur l'évaluation des ontologies a également vu le jour, l'évaluation systématique des outils devant permettre d'atteindre un niveau homogène de qualité et de faciliter l'adoption des technologies du web sémantique par le monde industriel. L'atelier EON (Evaluation of Ontologies for the Web) a mis l'accent sur les langages du web sémantique proposés par le W3C et a cherché à mettre en place des procédures d'évaluation des ontologies adaptées aux besoins du Web sémantique et à la problématique d'hétérogénéité, d'évolutivité et d'incomplétude du web.

EON2002[2], associé à la conférence EKAW (European Conference on Knowledge Acquisition and Management), s'est interrogé sur la manière dont il convient d'évaluer les technologies liées aux ontologies et les environnement d'ingénierie des ontologies. Le protocole d'évaluation proposé comporte deux aspects principaux (Maynard et al., 2006) : (i) l'évaluation des caractéristiques syntaxiques et sémantiques de l'ontologie produite, (ii) l'évaluation technologique (passage à

[2]http ://km.aifb.uni-karlsruhe.de/ws/eon2002

l'échelle, allocation mémoire, interopérabilité, etc.). La compétition a réuni une douzaine de participants (dont le LIPN avec Terminae) mais il n'y a pas eu d'évaluation globale des systèmes en compétition, chaque participant présentant les limites et les atouts de son système.

Les ateliers suivants ont proposé une série d'expériences pour évaluer les outils de fusion et d'alignement d'ontologies (EON2004[3]), les outils d'annotation d'ontologies et les technologies de web services sémantiques (EON2008[4]). EON2006 a le mérite d'avoir cherché à comparer les approches d'évaluation sur 4 ontologies différentes. Il a débouché sur des propositions de méthodes et de métriques.

Au total, même s'il n'existe encore aucune méthode d'évaluation globale et intégrée, les ateliers EON ont mis l'accent sur l'importance et la difficulté de l'évaluation des ontologies et des outils d'ingénierie ontologique et ils ont permis de faire émerger des propositions (Brank, 2006; Sabou et al., 2006; Obrst et al., 2006).

2.2 Evaluation au travers des applications

Au-delà des tentatives d'évaluation centrées sur les outils d'acquisition, on a également cherché à évaluer l'apport de leurs résultats dans diverses applications. Cette approche de l'évaluation est d'autant plus importante qu'elle permet de mettre en évidence l'intérêt des terminologies ou des ontologies dans des applications diverses.

Terminologie pour l'analyse syntaxique Les résultats de l'acquisition de termes peuvent être exploités pour améliorer la qualité de l'analyse syntaxique. Elle permet notamment de réduire les ambiguïtés de rattachements prépositionnels fréquents dans les corpus spécialisés. Cette idée initialement introduite par Bourigault (1993) a été reprise et testée par Aubin et al. (2005) pour adapter le Link Grammar Parser (LGP) (Grinberg et al., 1995) au domaine de la biologie. Les premiers résultats ont montré que la connaissance des termes diminue considérablement le temps d'analyse du LGP qui reflète la complexité de la phrase. Le nombre d'analyses produites par phrase diminue de manière très significative avec l'introduction des termes qui permettent de simplifier la phrase et le nombre d'analyses erronées baisse de 40% environ. Même si ces résultats montrent l'intérêt de l'analyse terminologique pour l'analyse syntaxique, l'évaluation réelle de la terminologie au travers de l'analyse syntaxique est difficile à conduire : les résultats dépendent de l'analyseur utilisé et de ses caractéristiques, du sous-langage considéré dans lequel les termes peuvent avoir plus ou moins d'importance et il faudrait comparer l'apport de différentes terminologies pour avoir une idée un peu précise de la qualité de la liste des termes exploitée.

Terminologie dans les index de fin de livre Les outils d'analyse terminologique étant exploités pour construire des index de fin de livre, leur apport et leur qualité ont pu être évalués dans ce contexte. IndDoc est par exemple un outil d'aide à l'indexation qui repose sur des outils d'analyse terminologique. Il permet de construire des ébauches d'index à partir desquels les indexeurs peuvent travailler pour produire des index finaux. Dans le cas d'un outil interactif, l'évaluation vise à apprécier la charge de travail qui reste à la charge de l'indexeur qui doit retravailler l'ébauche d'index produite automatiquement. Les expériences d'évaluation présentées dans (Ait El Mekki et Nazarenko, 2006) mettent clairement en valeur l'apport de l'analyse terminologique (extraction de termes, tri des termes par ordre de pertinence, hiérarchisation de la terminologie produite) mais, là non plus, elles ne constituent pas des expériences complètes d'évaluation des outils d'analyse terminologique. Il faudrait pour cela comparer les résultats produits avec différents outils d'analyse terminologique, ce qui représente un travail d'intégration et de validation non négligeable.

Terminologie et traduction automatique Dans les domaines spécialisés, la traduction s'appuie depuis longtemps sur des ressources terminologiques bilingues que l'on cherche aujourd'hui à intégrer dans les systèmes de traduction automatique. Lors du projet CESTA sur l'évaluation des systèmes de traduction automatique, une tâche originale a été proposée aux participants en leur permettant d'adapter leur système au domaine spécifique de la santé (Hamon et al., 2007). La comparaison des résultats obtenus avant et après cet enrichissement terminologique met en lumière l'apport de la terminologie sur les systèmes de traduction automatique[5] : sur cinq systèmes évalués, trois ont obtenus des résultats légèrement meilleurs et deux autres ont nettement amélioré leurs performances. (Langlais et Carl, 2004) présente une expérience d'évaluation intéressante dans ce contexte. Les auteurs cherchent à mesurer l'apport d'une terminologie bilingue dans l'adaptation d'un système de traduction automatique générique. Ce système repose sur une approche statistique et les auteurs montrent comment la terminologie peut être prise en compte dans le modèle de langage : elle est utilisée comme un faisceau de contraintes pour élaguer l'espace des traductions possibles. Même si le bilan est fortement dépendant de

[3]Qui a donné l'amorce de la campagne OAEI (Ontology Alignment Evaluation Initiative).

[4]http ://sws-challenge.org/wiki/index.php/EON-SWSC2008

[5]Ceci en dépit d'un protocole d'évaluation difficile : la campagne a été réalisée sur une durée relativement courte et sur un corpus d'adaptation de faible volume, environ 20 000 mots.

la manière dont la terminologie est prise en compte et du système de traduction automatique dans laquelle elle est exploitée, le bilan est globalement positif : les auteurs font état d'un accueil favorable des traducteurs et d'une amélioration significative du NIST[6]. Ils montrent également que cet impact dépend assez fortement de la couverture de la ressource terminologique.

Ontologie et recherche d'information La prise en compte de l'ontologie dans un système de recherche d'information peut intervenir essentiellement à deux niveaux : à l'étape d'indexation des documents et des requêtes et au moment de recherche elle-même, c'est-à-dire, lors de l'appariement requêtes-documents (Zargayouna, 2005). Les ontologies utiles en recherche d'information sont des ontologies lexicales ou temino-ontologies. La majorité des travaux utilisent WordNet en utilisant les liens lexicaux (des synonymes) et conceptuels (hiérarchiques ou méronymiques) (Baziz et al., 2003). Même si on voit bien en quoi les ontologies peuvent être utiles pour l'expansion ou le raffinement sémantique des requêtes, l'apport des ontologies dans les systèmes de recherche d'information reste difficile à déterminer. Les expérimentations effectuées sont généralement difficilement reproductibles, ce qui rend plus difficiles les évaluations comparatives. Selon les cas, les résultats sont liés à un domaine spécialisé et donc peu généralisables ou ils restent de portée limitée.

2.3 Discussion

La problématique de l'évaluation n'a pas le même poids dans les communauté d'acquisition terminologique et ontologique. Les campagnes d'évaluation en terminologie ont souffert du nombre restreint de participants, ce qui s'explique sans doute en partie par un déficit d'intérêt pour l'évaluation. Du côté de l'acquisition d'ontologie, la problématique de l'évaluation est reconnue comme importante en revanche. Les ateliers EON ont rassemblé des chercheurs autour de cette problématique et abouti à la proposition de protocoles et de métriques, ce qui constitue un point de départ intéressant.

Dans les deux cas, il reste difficile de faire émerger un cadre fédérateur. Le faible succès des campagnes en terminologie s'explique aussi par le fait que les systèmes qui avaient été conçus dans des perspectives différentes ont eu du mal à "entrer dans le cadre" de l'évaluation. L'atelier EON n'a pas permis de faire émerger un protocole d'évaluation global. Les travaux autour de l'évaluation des ontologies contournent souvent cette difficulté en proposant d'évaluer les ontologies par des critères intrinsèques, le plus souvent formels (pour vérifier par exemple la cohérence ou la consistence), sans que la corrélation de ces critères formels avec ceux de qualité globale de l'ontologie soit prouvée.

Il est également important d'évaluer les outils d'acquisition dans leur contexte applicatif et de mesurer leur valeur ajoutée dans ce cadre. Il s'agit alors de mesurer la différence de qualité de l'application elle-même selon qu'elle incorpore ou non des outils d'acquisition de connaissances. Ce type d'évaluations est indéniablement intéressant mais, il est difficile à mettre en oeuvre : on a du mal à dissocier ce qui relève de la qualité des ressources utilisées et du fonctionnement de l'application.

3 Difficultés

Nous avons décrit les expériences d'évaluation, dans le cadre de campagnes ou ponctuellement. Il en sort que ces expériences n'ont pas encore permis de faire émerger un cadre fédérateur d'évaluation comme c'est le cas dans d'autres disciplines. Nous présentons dans cette section quelques éléments qui expliquent, à notre sens, la difficulté de mise en œuvre d'un tel cadre pour les outils d'acquisition de connaissances.

3.1 Une tâche d'acquisition difficile à définir

Complexité des artefacts produits La principale difficulté tient à la nature même des terminologies et des ontologies. Ce sont des artefacts complexes. Considérons le cas des terminologies. Les termes eux-mêmes sont souvent des unités "complexes", composés de plusieurs mots, de longueur très variable et obéissant à des règles de variation multiples en corpus. Ensuite une terminologie ne se réduit souvent pas à une liste de termes, aussi complexes soient-ils. Des relations existent entre ces termes, et ces relations sont elles-mêmes diverses, depuis la variation morphologique (*véhicule d'occasion, véhicules d'occasion*) jusqu'aux relations de synonymie (*voiture d'occasion, automobile d'occasion*) ou d'hyperonymie (*véhicule d'occasion, véhicule*). La complexité même de ces artefacts constitue un frein à leur évaluation. On ne peut pas évaluer à la fois la qualité des termes extraits et des différentes relations entre ces termes. De même, il est difficile dans le cas d'une ontologie d'évaluer à la fois la qualité de ses concepts, sa structuration hiérarchique (granularité, densité, etc.), sa cohérence et sa consistence.

[6]Score utilisé comme critère de qualité en traduction automatique

Importance du rôle de l'application L'application pour laquelle la terminologie ou l'ontologie ont été développées joue aussi un rôle dans les contours de ces artefacts que l'on cherche à produire ainsi que sur les critères de qualité qui peuvent être exigés. Par exemple, toutes les terminologies ne doivent pas être évaluées selon les mêmes critères : si on a besoin d'une liste de termes bien formés pour la présenter à l'utilisateur (*logement étudiant* plutôt que *logement de l'étudiant*), de simples associations de mots statistiquement pertinentes comme *étudiant-logement* suffisent quand il s'agit de mesurer le poids des termes dans un document. Cette variété dans les critères de qualité selon l'application empêche d'avoir un cadre global qui pourrait s'appliquer à tout type de terminologie ou ontologie.

Place de l'interaction Dans la mesure où le résultat des outils d'acquisition de connaissances est dépendant de l'application, du domaine considéré, du niveau de qualité attendu et du point de vue adopté, le processus d'acquisition est rarement vu comme un processus entièrement automatique. Le plus souvent les outils d'acquisition sont des outils d'aide à l'acquisition qui intègre la participation du terminologue ou de l'ontologue dans le processus de construction de la ressource sémantique. Cette part d'interaction nécessite, au moment de l'évaluation, de départager ce qui relève du système d'acquisition et la part de travail manuel.

Absence de référence stable Au-delà de l'hétérogénéité formelle des artefacts produits, on observe également une grande hétérogénéité d'un point de vue sémantique, ce qui constitue une troisième difficulté pour l'évaluation. Pour un même domaine, on peut produire plusieurs terminologies ou ontologies différentes qui ne décrivent pas le domaine modélisé avec la même granularité, qui ne reflètent pas le même point de vue sur le domaine et qui peuvent même traduire des parti pris terminologiques et ontologiques différents. Le fait de partir de corpus permet de restreindre quelque peu l'espace de ces choix mais seulement partiellement. Ceci se traduit par l'absence de ressources standard utilisables pour l'évaluation. Autant de facteurs d'hétérogénéité qui sont difficiles à isoler et qui compliquent donc un peu plus l'évaluation : il n'y a pas de référence stable mais au contraire un grand choix de "solutions possibles".

Il est en effet assez difficile de disposer d'un référentiel établi pour un domaine et une application donnés. La construction de ces ressources est coûteuse et l'évaluation d'outil d'acquisition automatique ne serait pas aussi cruciale s'il était aisé de les construire manuellement. De plus, en admettant qu'on dispose d'un référentiel construit manuellement, le silence dans les résultats peut être dû à un manque dans le corpus d'extraction ou à une faille dans les systèmes d'acquisition.

Le référentiel peut être produit de différentes manières. On peut réutiliser une ressource existante mais elle risque de n'être que partiellement liée au corpus d'acquisition. On peut demander à un ou des experts de procéder à l'acquisition manuelle des connaissances à partir du même corpus d'acquisition qui est fourni aux systèmes. Cette seconde solution a l'avantage de la fiabilité mais elle est coûteuse. Une troisième approche consiste comme dans CoRRecT à fusionner la validation des résultats de différents systèmes. Même si elle manque d'exhaustivité, cette approche permet de comparer les résultats des différents systèmes entre eux et est *a priori* moins coûteuse que la précédente.

La nature même des connaissances, le lien à la langue et à une nécessaire modélisation explique la variation des références (même pour un domaine spécifique). Il est donc nécessaire de prendre en compte l'ensemble des solutions possibles, comme c'est le cas par exemple pour la traduction automatique.

3.2 Diversité des protocoles

Nous présentons dans la figure 1, différents scénarios d'évaluation.

Le premier consiste à comparer les sorties du système à une référence. Même si les sorties ainsi produites ne sont généralement pas utilisées sans être validées ou filtrées, il est important de pouvoir évaluer les sorties de manière indépendante. Elles offrent un bon point de comparaison entre différents outils.

Le second scénario vise à évaluer l'interaction. Quand les terminologies ou ontologies sont considérées comme ressources autonomes, elles font généralement l'objet d'un travail de validation par un expert avant d'être publiées. Ce type d'évaluation permet de mesurer l'effort fourni par l'expert pour aboutir à un produit final. Cette évaluation est évidemment liée à la qualité de l'interface de validation qui peut jouer un rôle important en facilitant le travail de l'expert et à la problématique de l'usage[7] mais nous mettons ici l'accent sur les opérations de validation elles-mêmes que l'on peut interpréter comme une distance d'édition.

Le troisième scénario vise à évaluer au travers d'une application. Comme nous l'avons souligné dans la section 2.2, une telle évaluation sert à quantifier la "plus-value" du produit (brut ou final) et à mesurer son apport à l'application. On a vu plus haut, cependant, que ce type d'évaluation n'est pas aisée car on a du mal à savoir si on est en train d'évaluer l'apport du produit ou la manière dont il est pris en compte dans le système.

[7] Des protocoles d'évaluation centrés sur l'usage sont proposés dans (Mustafa El Hadi et Chaudiron, 2007)

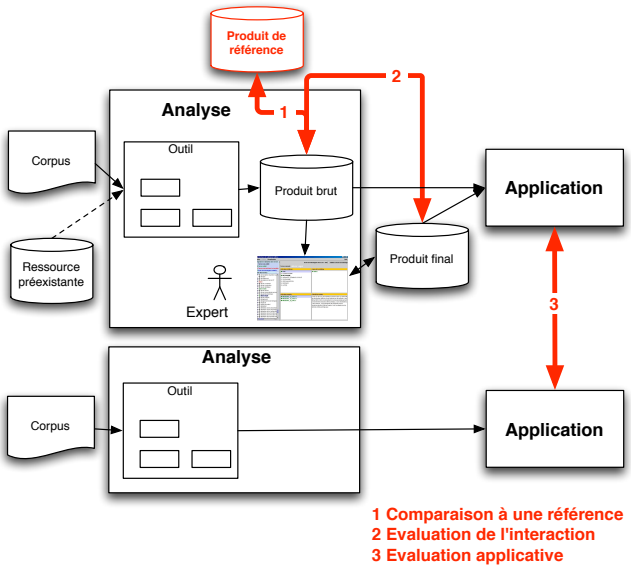

FIG. 1 – *Place des outils d'acquisition*

Les deux premiers scénarios d'évaluation permettent une évaluation comparative (entre deux ressources) qui est particulièrement adaptée aux évaluations de type "boîte noire" dans lesquelles on s'intéresse uniquement aux résultats fournis par les logiciels et non pas à leur mode de production. Ce type d'évaluation peut être représenté comme un triplet $< S, Corr, R) >$ où S est le produit à évaluer, R est une référence (indépendante ou construite à partir de la validation de S) et $Corr$ est une fonction de correspondance qui permet de mesurer la proximité entre les deux ressources. Les résultats de cette fonction de correspondance servent de base pour le calcul des différentes mesures d'évaluation.

Nous préconisons dans ce qui suit ce type d'évaluation en détaillant les difficultés inhérentes à ce cadre.

3.3 Limites des mesures actuelles

Les fonctions de correspondance entre le résultat produit et le résultat attendu doivent prendre en compte les caractéristiques de ces derniers. Les mesures les plus répandues pour rendre compte des résultats des systèmes sont très classiques : le rappel et la précision. Ce sont des mesures faciles à calculer, qui ont aussi le mérite d'être faciles à interpréter quel que soit le domaine de recherche où elles sont appliquées. Elles permettent de rendre compte du bruit et du silence des résultats produits par rapport à un résultat de référence.

Néanmoins, ces mesures ne sont pas applicables telles quelles pour juger de la qualité des connaissances parce qu'elles font l'hypothèse que la pertinence est une notion binaire (oui/non), ce qui dans la réalité n'est pas le cas. Cette pertinence binaire relève d'une logique du "tout ou rien" et ne tient pas compte du fait qu'un résultat s'approche plus de la solution qu'un autre. Toutes les erreurs ne devraient pas compter autant : certaines erreurs sont flagrantes, d'autres plus discutables, d'autres encore sont apparentées et donc plus facilement identifiables. Ce fait est d'autant plus problématique que la référence elle-même ne peut être unique comme expliqué plus haut. Les mesures de rappel et précision ne permettent ni de discriminer les résultats entre eux (différence entre un mauvais et un moins mauvais), ni de mesurer l'effort requis pour corriger une sortie.

On retrouve ce problème dans d'autres domaines de recherche telles que la recherche d'information structurée et l'alignement d'ontologies. Dans le cadre de la campagne INEX [8], il a été proposé dès le départ de graduer le jugement de pertinence en le quantifiant selon deux nouveaux critères (spécificité et exhaustivité) (Fuhr et al., 2002).. La campagne a constitué un excellent vivier de définition des mesures, telles que la mesure EPRUM (*expected precision-recall with user modelling*) proposée par Piwowarski et Dupret (2006), même si les mesures ne sont pas encore stabilisées (Fuhr et al., 2007). La question de l'évaluation des résultats des systèmes d'alignement d'ontologies est posée dans le cadre de la campagne OAEI[9]. Une solution proposée dans Ehrig et Euzenat (2005) est d'approximer la référence par similarité. Nous verrons dans la section 4.2 que nous adoptons une solution similaire pour l'évaluation de l'acquisition de terminologies.

[8] Initiative for the Evaluation of XML Retrieval
[9] Ontology Alignment Evaluation Initiative

4 Premières propositions

Nous proposons dans ce qui suit des solutions aux difficultés présentées plus haut. Nous proposons d'abord de décomposer le processus d'acquisition en fonctionnalités élémentaires pour mieux délimiter les objets d'étude. Nous proposons ensuite de définir des mesures qui tiennent compte de la variation des références. Il nous paraît aussi essentiel de mettre en œuvre une méta-évaluation pour vérifier l'adéquation des mesures aux tâches à évaluer.

Nous détaillons les deux premières propositions en rapportant les travaux effectués dans le cadre du programme Quaero sur l'acquisition des terminologies.

4.1 Découper pour mieux observer

Un des enjeux de l'évaluation des outils d'acquisition consiste à décomposer le processus d'acquisition global en fonctionnalités élémentaires. Il s'agit d'identifier, au-delà de la diversité des outils d'acquisition et de la complexité des résultats qu'ils produisent, et en faisant abstraction à la fois des méthodes utilisées par les outils et des applications pour lesquelles ils ont été conçus, d'identifier ce qu'ils ont en commun et de les découper en fonctionnalités élémentaires et indépendantes. Ces fonctionnalités ne fournissent pas nécessairement des résultats utilisables en tant que tels mais elles offrent des points de comparaison entre les outils. Ce ne sont pas forcément des sorties standard des systèmes à évaluer mais peu importe dès lors qu'on peut extraire des résultats de leurs sorties standard (voir figure 2).

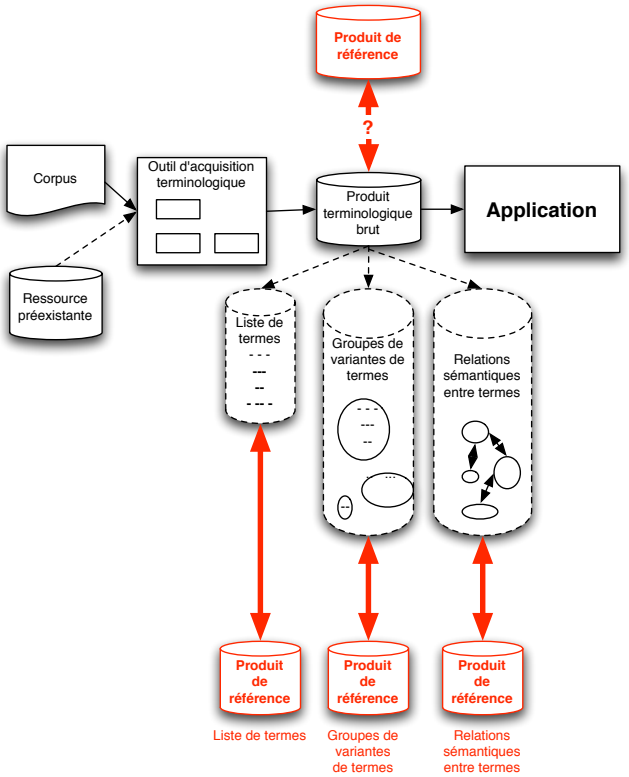

FIG. 2 – *Découpage et évaluation par fonctionnalité.*

Pour l'acquisition de terminologie, nous avons fait ce travail de "décomposition en facteurs premiers", première étape de la définition d'un protocole d'évaluation. En nous inspirant des sous-tâches définies dans les campagnes précédentes, nous avons proposé de distinguer trois fonctionnalités élémentaires génériques et d'évaluer les outils d'acquisition terminologiques selon ces trois dimensions (Zargayouna et al., 2007).

L'extraction terminologique représente la capacité d'un système à établir une liste de termes simples et complexes à partir d'un corpus d'acquisition. La liste plate des termes extraits du corpus est rarement utilisée en tant que telle dans les applications mais la plupart des outils d'acquisition terminologique passent par une étape d'extraction. Cette étape du processus d'acquisition constitue donc un point d'observation privilégié pour l'évaluation. L'évaluation, à ce niveau, peut consister à comparer la sortie d'un ou plusieurs systèmes avec une terminologie de référence, mais aussi à comparer

la sortie brute d'un système avec cette même sortie une fois validée. Nous avons pris le parti de dissocier l'extraction terminologique à proprement parler du tri des termes extraits, les critères de pertinence essentiellement dépendants de l'application cible dans laquelle la terminologie doit être exploitée.

Le calcul de variation terminologique consiste à regrouper les familles de termes qui sont des variantes les uns des autres, c'est-à-dire à établir des classes d'équivalence de termes. Pour s'en tenir à des fonctionnalités élémentaires, nous dissocions le fait de construire la classe (ou d'identifier des relations d'équivalence entre termes) et le choix du représentant de la classe, le terme canonique. Les systèmes qui font du calcul de variation (en lien avec l'extraction de termes, l'expansion de thesaurus ou l'indexation contrôlée) n'ont pas tous besoin d'identifier un terme canonique. Évaluer le calcul de variation va consister à comparer les relations d'équivalence proposées par les systèmes entre eux ou par rapport à une référence établie au préalable.

Cette fonction est cependant plus difficile à définir que la précédente dans la mesure où différentes relations d'équivalence peuvent être prises en compte. Certains auteurs distinguent différents niveaux de variation (variation morphologique, syntaxique ou sémantique) mais ce critère est lié aux méthodes de repérage en corpus plutôt qu'à la nature du résultat. En tenir compte biaiserait l'évaluation. Il faut au contraire tenir compte de la nature de la relation obtenue qui doit refléter une équivalence sémantique. On dira que deux termes sont les variantes l'un de l'autre s'ils sont synonymes. On considérera donc comme bien formée la classe de termes suivante : *expression de gène, expression génique, gènes exprimés* et *production de gène*, dès lors qu'un terminologue connaissant la biologie pose que *production de gène* et *expression de gène* peuvent être employés l'un pour l'autre.

L'extraction de relations terminologiques est moins souvent présente dans les outils d'acquisition. Elle consiste à élaborer un réseau de relations sémantiques entre termes ou classes de termes. Il peut s'agir de différents types de relations : relations hiérarchiques ou relations plus spécialisées. La diversité des relations considérées, leur dépendance au domaine d'application et les différences de granularité des descriptions sémantiques produites font que cette troisième fonction est sans doute la plus difficile à évaluer.

Pistes pour l'acquisition d'ontologies Nous préconisons la même démarche pour les outils d'acquisition d'ontologies. Nos premières propositions s'orientent vers la décomposition en : (i) acquisition de classes sémantiques, (ii) structuration hiérarchique de l'ontologie, (iii) extraction de relations sémantiques ou rôles et (iv) formalisation[10]..

4.2 Définir des mesures et des protocoles pour chaque tâche

Si l'on adopte le découpage proposé ci-dessus, il faut donc, pour l'évaluation des outils d'acquisition terminologique, proposer des métriques pour chaque sous-tâche considérée séparément. Nous mettons ici l'accent sur la première de ces sous-tâches, l'extraction de termes.

Il est essentiel que les mesures tiennent compte des caractéristiques des produits à évaluer mais, en même temps, ces mesures doivent autant que possible être indépendantes des méthodes qu'elles visent à évaluer et elles doivent permettre d'évaluer chaque sous-tâche indépendamment les unes des autres, une fois une décomposition en tâches établie.

Pour l'extraction de termes, nous définissons des métriques qui tiennent compte de la variation des références et du fait que la pertinence est plus graduée que booléenne. Il est aussi important d'avoir des mesures simples et adaptables. De ce point de vue les mesures de rappel et de précision sont bien connues et communément admises. C'est pourquoi nous proposons d'adapter les mesures classiques de précision et rappel rappelées ci-dessous plutôt que d'en adopter de nouvelles :

$$precision = \frac{|S \cap R|}{|S|}$$

$$rappel = \frac{|S \cap R|}{|R|}$$

où $|S \cap R|$ est le nombre d'éléments pertinents retournés par le système, $|S|$ est le nombre d'éléments retournés par le système et $|R|$ le nombre d'éléments dans la référence.

La tâche d'extraction produit une liste de termes plate, dite "de sortie" (S), à comparer à une liste de termes de référence (R). La référence peut être de longueur variable, même pour un domaine et un corpus donnés : cela dépend de la granularité de la description terminologique choisie. La liste produite par le système peut elle-même être de taille très hétérogène comme l'ont montré les expériences de CESART. Elle peut comporter des termes pertinents (présents dans

[10]Précisons que nous excluons du champ de l'évaluation ontologique tout ce qui concerne la population des ontologies par extraction des entités nommées.

la référence), des termes non pertinents (considérés comme du bruit) et des termes proches des termes de la référence sans être strictement identiques. Ces derniers sont à considérer comme "presque bons" ou comme redondants si le terme exact figure aussi dans la ressource proposée. En pratique, la liste produite par le système d'extraction peut comporter des variantes de termes mais pour ne pas interférer avec le calcul de la variation qui fait l'objet de la deuxième sous-tâche, nous la considérons comme de la pseudo-redondance ici. Nous voulons pouvoir quantifier les phénomènes suivants :

- un système parfait ($S = R$) doit avoir une valeur de qualité maximale (si on raisonne en rappel et précision, la valeur doit être à 1).
- un système qui renvoie une liste de termes avec variantes ($S = R \cup Var(R)$[11]) ne doit pas être pénalisé ou faiblement.
- un système qui renvoie une liste de non termes ($S \cap R = \emptyset$) doit avoir la valeur minimale (valeur proche de zéro).
- la qualité d'un système S doit augmenter quand il se rapproche globalement de la référence.

Considérons le cas où $R = \{$*base de données*$\}$ et les listes résultats suivantes : $S_1 = \{$*base de données, bases de données*$\}$, $S_2 = \{$*bases de données*$\}$, $S_3 = \{$*base de données, langage de requête*$\}$. On souhaite que S_1 et S_2 soient considérés comme sensiblement de même qualité mais que celle de $S3$ soit moindre au regard de R.

Prendre une mesure de pertinence graduée Nous proposons de reprendre les mesures de rappel et précision classique mais en considérant une mesure de pertinence graduée. Cette fonction de pertinence $Pert(S, R)$ doit vérifier

$$|S \cap R| \leq Pert(S, R) \leq min(|S|, |R|)$$

et rendre compte d'une proximité globale entre S et R fondée sur un calcul de similarités individuelles entre les éléments de la sortie et ceux de la référence. Nous définissons pour cela une similarité entre deux éléments de S et R, $sim(e_s, e_r)$ par :

$$sim(e_s, e_r) = 1 - dist(e_s, e_r)$$

où $dist$ est une distance terminologique du type de celle proposée par El Mouedden (2008). Cette distance repose sur la distance d'édition qui permet de calculer de manière homogène la distance entre termes simples et la distance entres termes complexes comme l'a souligné Tartier (2004). Notre approche diffère cependant de cette dernière parce que nous ne voulons pas faire appel à des outils de calcul linguistique, lesquels pourraient introduire des biais dans l'évaluation. La distance entre termes complexes repose sur le même principe de l'alignement et de la distance optimale, mais on fait cette fois les opérations de transformation sur les mots plutôt que sur les caractères. On calcule l'alignement optimal qui permet d'atteindre un terme en partant de l'autre, en ajoutant à chaque fois le coût de l'opération faite. Les opérations d'ajout et de suppression ont le même coût égal à 1, tandis que le coût de la substitution varie selon la nature des mots en question : il est égal à la distance entre termes simples séparant les deux mots mis en correspondance. La distance finale est égale au coût optimal de l'ensemble des opérations divisé par le nombre de mots du plus long terme. Cette normalisation par la longueur du plus long terme permet de dire que *base de donnée relationnelle avancée* est plus proche de *base de donnée relationnelle* que ce dernier ne l'est de *base de données*.

On retient pour chaque terme de S le terme de la référence avec qui il a une similarité maximale, ce qui permet de définir la pertinence d'un terme de la manière suivante :

$$pert(e_s) = max_{e_r \in R}(sim(e_s, e_r)) \; si \; max_{e_r \in R}(sim(e_s, e_r)) > \sigma$$

$$pert(e_s) = 0 \; sinon$$

où σ est un seuil de similarité en-deçà duquel on considère que deux termes ne sont pas comparables[12].

Adapter formellement la sortie à la référence Dans la mesure où la référence ne peut pas être considérée comme un absolu, il serait artificiel de comparer directement la sortie du système avec la référence. On favoriserait trop le système qui aurait "par hasard" fait les mêmes choix de granularité de description que la référence et on risquerait d'avoir des résultats d'évaluation trop dépendants du type de référence adopté. Nous proposons donc de transformer la sortie pour trouver sa correspondance maximale avec la référence, ce qui revient à adapter la sortie à la référence. La précision et le rappel sont calculés sur la sortie transformée plutôt que de la sortie brute.

De fait, comme plusieurs termes de la sortie peuvent correspondre au même terme de la référence, on regroupe dans certains cas les termes de la sortie. Il s'agit donc de calculer les mesures de précision et rappel non pas directement sur S mais sur une partition de S qui est définie relativement à R. Nous définissons cette partition $\mathcal{P}(\mathcal{S})$ telle que toute partie

[11] Où $Var(R)$ est un ensemble de variantes de termes de R.

[12] Nous le fixons arbitrairement à 0,5 mais ce seuil pourrait être fixé en fonction de la distance des termes de la référence entre eux.

p de $\mathcal{P}(\mathcal{S})$ correspond soit à un ensemble des termes de S qui ont une similarité maximale strictement positive avec un même terme de R, soit à un terme singleton. On peut dès lors définir la pertinence d'une partie p de $\mathcal{P}(\mathcal{S})$[13] :

$$pert(p) = max_{e_s \in p}(pert(e_s))$$

Les mesures de rappel et de précision terminologiques se définissent alors comme suit :

$$T - precision = \frac{Pert(S,R)}{|\mathcal{P}(\mathcal{S})|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{S})}(pert(p))}{|\mathcal{P}(\mathcal{S})|}$$

$$T - rappel = \frac{Pert(S,R)}{|R|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{S})}(pert(p))}{|R|}$$

Nous pouvons vérifier que dans le cas d'un système parfait $T - precision = T - rappel = 1$ et que dans le cas d'un système qui ne retournerait que du bruit les valeurs de $T - precision$ et $T - rappel$ tendent vers 0. La figure 3 montre ce qu'on obtient comme mesures de précision et de rappel pour les trois sorties mentionnées ci-dessous.

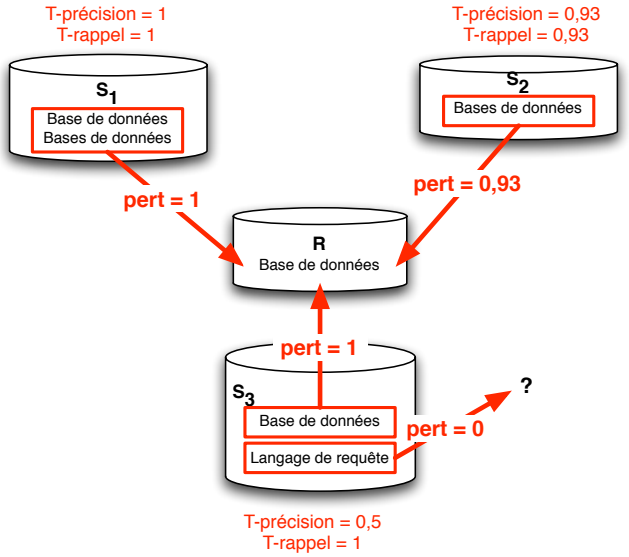

FIG. 3 – *Exemples de mesures de $T - precision$ et $T - rappel$.*

4.3 Méta-évaluer les métriques proposées

Une bonne évaluation nécessite que les efforts d'évaluation eux-même soient évalués. Il est nécessaire de vérifier l'évaluation pour détecter les problèmes tels que le biais, les erreurs techniques, les utilisations erronées (Stufflebeam, 1974).

Après avoir bien défini les spécifications des protocoles d'évaluation, la méta-évaluation consiste en une étape de vérification (comme en génie logiciel) qui revient à s'assurer que les spécifications ont été bien respectées. En d'autres termes, la méta-évaluation consiste à se doter de moyens pour vérifier qu'on est bien en train d'évaluer ce qui doit être évalué, de vérifier les biais, etc. En traduction automatique, la méta-évaluation a ainsi permis de bien cerner les limites des mesures utilisées. Les études reportées dans (Koehn et al., 2006) ont montré que la mesure Bleu sous-estime la qualité des systèmes à base de règles. Timimi (2007) rapporte un exemple de méta-évaluation qui a été proposé dans le cadre de la campagne d'évaluation CESART. L'étude s'intéresse au cadre global du protocole d'évaluation *i.e* choix du corpus de tests, statut de l'expert, etc.

Nous sommes actuellement en train de définir un cadre de méta-évaluation des mesures que nous avons présentées dans la section précédente. On peut facilement vérifier les critères de cohérence des mesures en vérifiant les bornes inférieures et bornes supérieures Popescu-Belis (1999). Nous voulons éprouver les mesures proposées expérimentalement sur des jeux de tests choisis pour vérifier le comportement des mesures dans certains cas particuliers. Ces mêmes tests devraient être exécutés sur des volumes significatifs, par exemple en introduisant progressivement du bruit représentatif et en suivant son influence sur les résultats.

[13] Des variantes de cette formule sont à étudier expérimentalement.

L'étude de corrélation des références, dans le cas où on dispose de plusieurs références ou plusieurs experts, peut être intéressante à mettre en regard avec les résultats des mesures.

L'évaluation étant un processus itératif, la méta-évaluation permet aussi de faire évoluer les mesures et les protocoles.

5 Conclusion

Peut-on évaluer les outils d'acquisition de connaissances à partir de textes ? Notre réponse est positive, mais il faut prendre en compte toutes les difficultés liées au domaine. Nous avons étayé notre réponse par des propositions concrètes dans le cadre de l'évaluation de l'acquisition terminologique. Nous proposons de nous affranchir dans un premier temps d'une vision globalisante des ressources produites pour décomposer la tâche d'acquisition et ses résultats en sous-tâches élémentaires et en résultats partiels. Il faut mettre l'accent sur les composantes de base qui sont communes à la plupart des outils d'acquisition de connaissance. Nous proposons également d'adapter à notre cadre les métriques existantes en essayant d'être le plus fidèle possible à cette part de variation caractéristique des systèmes qui manipulent la langue ou qui produisent des modèles de connaissances. Ces mesures sont implémentées et des expérimentations sont en cours pour valider nos propositions. Nous soulignons pour finir l'importance de la méta-évaluation pour la mise au point des métriques.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme Quaero,financé par OSEO, l'agence française pour l'innovation. Nous remercions Mourad El Moueddeb, Olivier Hamon et Laurent Audibert pour leurs discussions et contributions à ce travail.

Références

Ait El Mekki, T. et A. Nazarenko (2006). An application-oriented terminology evaluation : the case dof back-of-the-book indexes. In R. Costa, F. Ibekwe-SanJuan, S. Lervad, M.-C. L'Homme, A. Nazarenko, et H. Nilsson (Eds.), *Proceedings of the Workshop "Terminology Design : Quality Criteria and Evaluation Methods" (TerEval) associated with the Language Ressource and Evaluation Conference (LREC)*, Genova, Italy, pp. 18–21.

Aubin, S., A. Nazarenko, et C. Nédellec (2005). Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, Borovets, Bulgaria, pp. 89–93.

Baziz, M., N. Aussenac-Gilles, et M. Boughanem (2003). Désambiguisation et expansion de requêtes dans un sri, etude de l'apport des liens sémantiques. *Revue des Sciences et Technologies de l'Information (RSTI) série ISI 8*(4), 113–136.

Bourigault, D. (1993). An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings of the 6th European Chapter of the Association for Computational Linguistics (EACL'93)*, pp. 81–86.

Brank, J. (2006). Gold standard based ontology evaluation using instance assignment. In *Proceedings of the Evaluation of Ontologies for the Web (EON 2006) Workshop*.

Ehrig, M. et J. Euzenat (2005). Relaxed precision and recall for ontology matching. In *K-Cap 2005 workshop on Integrating ontology, Banff (CA)*, pp. 25–32.

El Moueddeb, M. (2008). Définition d'un protocole d'évaluation des outils d'analyse terminologique. Master's thesis, Université Paris-Nord, France.

Enguehard, C. (2003). Correct : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, Nancy, pp. 339–345.

Fuhr, N., N. Gövert, G. Kazai, et M. Lalmas (Eds.) (2002). *Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Schloss Dagstuhl, Germany, December 9-11, 2002*.

Fuhr, N., J. Kamps, M. Lalmas, S. Malik, et A. Trotman (2007). Overview of the inex 2007 ad hoc track. In *INEX*, pp. 1–23.

Grinberg, D., J. Lafferty, et D. Sleator (1995). A robust parsing algorithm for link grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies*.

Hamon, O., A. Hartley, A. Popescu-Belis, et K. Choukri (2007). Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA. In *Proceedings of MT Summit XI*, Copenhagen, Danmark.

Kageura, K., T. Fukushima, N. Kando, M. Okumura, S. Sekine, K. Kuriyama, K. Takeuchi, M. Yoshioka, T. Koyama, et H. Isahara (2000). Ir/ie/summarisation evaluation projects in japan. In *LREC2000 Workshop on Using Evaluation within HLT Programs*, pp. 19–22.

Koehn, P., N. Bertoldi, O. Bojar, C. Callison-Burch, A. Constantin, B. Cowan, C. Dyer, M. Federico, E. Herbst, H. Hoang, C. Moran, W. Shen, et R. Zens (2006). Factored translation models. In J. H. University (Ed.), *CLSP Summer Workshop Final Report WS-2006*.

Langlais, P. et M. Carl (2004). General-purpose statistical translation engine and domain specific texts : Would it work ? *Terminology 10*(1), 131–152.

Maynard, D., W. Peters, et Y. Li (2006). Metrics for evaluation of ontology-based information extraction. In *Proceedings of the Evaluation of Ontologies for the Web (EON 2006) Workshop*.

Mustafa El Hadi, W. et S. Chaudiron (2007). L'évaluation des outils d'acquisition de ressources terminologiques : problèmes et enjeux. In *Actes de la Conférence TOTh (Terminologie et Ontologie : Théories et Applications)*, pp. 163–179.

Mustafa el Hadi, W., I. Timimi, M. Dabbadie, K. Choukri, O. Hamon, et Y. Chiao (2006). Terminological resources acquisition tools : Toward a user-oriented evaluation model. In *Proceedings of the Language Resources and Evaluation Conference (LREC'06)*, Genova, Italy, pp. 945–948.

Obrst, L., T. Hughes, et S. Ray (2006). Prospects and posiibilities for ontology evaluation : The view from ncor. In *Proceedings of the Evaluation of Ontologies for the Web (EON 2006) Workshop*.

Piwowarski, B. et G. Dupret (2006). Evaluation in (xml) information retrieval : expected precision-recall with user modelling (eprum). In *SIGIR*, pp. 260–267.

Popescu-Belis, A. (1999). L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures. *Langues (Cahiers d'études et de recherches francophones) 2*(2), 151–162.

Sabou, M., V. Lopez, E. Motta, et V. Uren (2006). Ontology selection : Ontology evaluation on the real semantic web. In *Proceedings of the Evaluation of Ontologies for the Web (EON 2006) Workshop*.

Stufflebeam, D. L. (1974). Metaevaluation. In *Occasional Paper Series, Kalamazoo MI : Western Michigan University Evaluation Center*.

Tartier, A. (2004). *Analyse automatique de l'évolution terminologique : variations et distances*. Ph. D. thesis, Université de Nantes.

Timimi, I. (2007). Peut-on faire confiance aux outils de terminologie ? ou l'évaluation entre un souci de normalisation et une complexité de modélisation. In *Actes du Colloque Terminologie et Ontologie Théorie et Applications (Toth)*.

Zargayouna, H. (2005). *Indexation sémantique de documents XML*. Ph. D. thesis, Uniersité Paris-Sud, France.

Zargayouna, H., O. Hamon, et A. Nazarenko (2007). Evaluation des outils terminologiques : état des lieux et propositions. In *Actes des 7èmes rencontres Terminologie et Intelligence Artificielle*.

Summary

A large effort has been devoted to the development of knowledge acquisition tools, but it is still difficult to assess the progress that have been made.. The lack of well-accepted evaluation protocols and data leads to a gap in comparing results. In this article, we raise the question of evaluation for acquisition from text of terminologies and ontologies. We underline the major difficulties and describe first propositions we have made so far in this context.