Systèmes question-réponse et EuroWordNet

Christine Jacquin, Laura Monceaux, Emmanuel Desmontils

LINA, Laboratoire Informatique Nantes Atlantique
{christine.jacquin ; laura.monceaux ; emmanuel.desmontils}@univ-nantes.fr

Résumé

Pour améliorer l'efficacité des systèmes de recherche d'informations précises, l'utilisation de connaissances sémantiques est nécessaire. Cependant pour le français, les outils de connaissances sémantiques telles les thesaurus sur domaine ouvert ne sont d'une part pas très nombreux et d'autre part pas suffisamment complets. Dans cet article, nous expliquons premièrement, l'intérêt de l'utilisation de connaissances sémantiques pour un système de question réponse. Puis, nous présentons le thesaurus EuroWordNet, notamment ses limites et les améliorations que nous avons effectuées pour la base française dans un souci de le rendre plus satisfaisant pour notre application par l'ajout de relations inexistantes entre concepts et de définitions par le biais de l'encyclopédie Wikipedia (2006).

Mots-clés : thesaurus, système de question-réponse, similarité.

Abstract

In order to improve question answering systems, the use of semantic knowledge is essential. However for French, such knowledge is not easily available. Indeed, French thesauruses are scarse and under-developed. In this paper, we firstly explain the reason why semantic knowledge should be used in a question answering system. Then, we present the EuroWordnet thesaurus, focussing on its limitations and the improvements we have made. We add undefined relationships with the help of the English base by using Wikipedia (2006).

Keywords: thesaurus, question answering system, similarity.

1. Introduction

L'intérêt de l'utilisation de connaissances sémantiques dans les systèmes de question-réponse a largement été démontré notamment par des applications en langue anglaise. En effet, la mise en correspondance des termes de la question et de la réponse nécessite de telles ressources. (Crestan *et al*., 2004) ont montré que l'utilisation d'un thesaurus dans un système de question-réponse permet d'extraire un certain nombre de réponses que même la reformulation sur Internet, également utilisée, n'a pas détectée d'où la nécessité de l'utilisation d'un tel thesaurus. D'autres évaluations ont été réalisées sur l'apport de l'utilisation d'un thesaurus tel que la base lexicale WordNet (Miller, 1990), dans les systèmes de question-réponse. On note dans chacune de celles-ci une réelle amélioration des performances des systèmes évalués (Magnini *et al*., 2002). D'une part, les connaissances issues du thesaurus permettent d'améliorer la reformulation de la requête afin de sélectionner les documents les plus pertinents (Gonzalo *et al*., 1998). En effet, parfois il n'est pas facile de désambiguiser la question notamment à cause du contexte pauvre et il est plus intéressant de travailler au niveau des documents (Jacquemin, 2004) : le traitement sur les énoncés d'un document dispose d'un contexte plus large. D'autre part, on peut utiliser un thesaurus au niveau de la validation. En effet pour une question comme « Quel animal mi-

aule ? », on pourra vérifier que la réponse est bien un hyponyme du concept *animal*. Nous avons pu de plus constater que certaines réponses à des questions des campagnes d'évaluation TREC se trouvaient dans le thesaurus WordNet notamment dans les définitions associées aux concepts.

En anglais, le thesaurus WordNet est la plupart du temps le thesaurus utilisé pour réaliser cette correspondance entre termes de la question et la réponse. Pour le français, peu de ressources satisfaisantes sont disponibles. (Crestan *et al*., 2004) notent que la base EuroWordNet pour le français ne peut être considérée comme une base satisfaisante pour être utilisée dans un système de question réponse, d'où leur construction d'un thesaurus propre à leur attente.

Suite à cette étude, l'intérêt de l'utilisation d'un thesaurus pour notre système de question-réponse pour le français nous paraît évident. Nous avons décidé d'utiliser le thesaurus EuroWordNet pour le français (même si celui ci à l'achat ne semble pas satisfaisant pour notre tâche) et de lui apporter certaines améliorations. En effet, après étude, il s'avère que la couverture d'EuroWordnet français est assez pauvre par rapport à la partie anglaise, tant au niveau du nombre de concepts, du nombre de relations entre concepts que de l'absence de définitions associées aux concepts. Dans cet article, nous présentons, d'une part, l'étude menée sur le thesaurus Eurowordnet français pour ajouter des relations entre concepts en s'appuyant sur la partie anglaise de la base, et d'autre part, le processus mis en place pour adjoindre des définitions aux concepts de la base française par le biais de l'encyclopédie multilingue Wikipedia (2006).

2. EuroWordNet : présentation et limites

EuroWordNet est un thesaurus multilingue comprenant plusieurs langues européennes comme le néerlandais, l'allemand, le français, l'italien, le tchèque, etc. La conception de ce thesaurus s'appuie sur le thesaurus WordNet en langue anglaise développé par l'Université de Princeton (Miller, 1990). Ce dernier s'inspire de travaux psycholinguistiques et informatiques sur la mémoire lexicale humaine. L'objectif de ce thesaurus est de fournir une recherche conceptuelle dans un dictionnaire et donc d'organiser l'information lexicale en terme de signification de mots. Les noms, verbes, adjectifs ou adverbes sont représentés par des synsets (ensemble de synonymes) qui sont reliés entre eux par des relations sémantiques comme la relation de subsomption (hyperonymie/hyponymie) ou la relation partie/tout (méronymie/holonymie). Chaque thesaurus monolingue est autonome et dispose de ses propres lexiques et relations. Mais chaque thesaurus est relié aux autres via un index interlangue (ILI). L'index sert donc de langage pivot pour pouvoir passer d'une langue à l'autre. Mais il faut garder à l'esprit que certains concepts existants dans une langue n'existent pas forcément dans une autre langue: la langue la mieux couverte au niveau des relations et des concepts est la langue anglaise. Il est à noter aussi que pour la langue française, nous ne disposons pas de définitions des concepts.

Comme la base WordNet, le thesaurus EuroWordNet, pour une utilisation en traitement automatique des langues, souffre de problèmes liés à sa conception même. Dans ce dernier, les liens entre les parties des discours et les indications sur le domaine, par exemple, sont quasiment inexistants. Dans la partie française, on a bien un début de structuration des concepts du domaine informatique, mais c'est assez restreint et surtout les concepts ne sont pratiquement pas liés les uns par rapport aux autres (c'est une hiérarchie assez plate). On peut également noter une polysémie importante (par exemple, le mot *break* est répertorié avec 63 sens) et une granularité trop fine qui ne rendent pas toujours facile son utilisation dans les applications de traitement automatique des langues (O'Hara *et al*., 1998 ; Hanks et Pustejovsky, 2005).

3. Améliorations de la base lexicale de EuroWordNet

La version disponible d'EuroWordNet à l'état brut n'est donc pas suffisamment satisfaisante pour être utilisable. Les premiers problèmes que nous avons tentés de résoudre reposent sur l'exploitation de cette base de données, la mise à jour des relations et l'ajout de définitions inexistantes dans EuroWordNet.

3.1. Une base de données utilisable

À l'achat d'EuroWordNet, on dispose d'un outil (Périscope) qui permet l'interrogation du thesaurus stocké dans une base de données sous un format propriétaire. Mais aucune bibliothèque (API) ne permet d'interfacer cette base avec une autre application informatique. On dispose seulement de fichiers textes formatés, qui contiennent les différents concepts, les synsets et les relations et qui au premier abord, peuvent sembler simples à traiter. Mais ces fichiers ne sont pas toujours cohérents, par exemple les occurrences des concepts ne sont pas toutes codées sous le même format (nombreuses erreurs de transcription, etc.).

Pour pouvoir utiliser la base EuroWordNet pour notre application, un travail long et fastidieux de récupération des données a donc été nécessaire. La base a été stockée dans une base de données sous PostgreSQL et une API en Java a été développée pour permettre notamment des interrogations sur la base.

À partir de ce travail, nous avons constaté un certain nombre de problèmes liés à la conception même du thesaurus EuroWordnet français.

3.2. Mise à jour des relations sémantiques

En étudiant le thesaurus à partir de l'API développée, nous avons pu constater que certaines relations étaient manquantes. Par exemple, le terme *citronnier* existe bien dans le thesaurus français mais il est orphelin. En effet, il n'est relié à aucun concept, alors qu'il aurait dû être relié par la relation d'hyperonymie/hyponymie au concept *arbre fruitier* qui existe bien dans le thesaurus.

Dans un premier temps, notre objectif a consisté en l'utilisation de la propriété de bilinguisme d'EuroWordNet (passage d'une langue à une autre via l'index interlangue) pour tenter d'ajouter des relations existantes dans une langue et non dans une autre. Nous avons bien sûr présent à l'esprit que, d'un point de vue théorique, cette démarche n'est pas satisfaisante, car d'une langue à une autre un concept ne se décline pas forcément de la même manière. Mais d'un point de vue pratique, nous avons quand même adopté ce point de vue tout en vérifiant manuellement le bien fondé de l'ajout. Nous avons d'ailleurs validé toutes les mises à jour proposées par le processus.

Le traitement mis en place pour la relation d'hyperonymie/hyponymie est détaillé dans l'algorithme ci-dessous. Nous utilisons la propriété de transitivité de cette relation pour effectuer les mises à jour.

- **pour** chaque concept *c* d'EuroWordNet défini dans une langue source (ici français) **faire**
  - récupérer le synset *s* correspondant dans la langue cible (ici anglais)
  - **si** *s* existe **alors**
    - déterminer l'hyperonyme *h* le plus proche du synset *s* qui a un équivalent dans la langue source
    - **si** *h* existe et *h* n'est pas un hyperonyme de *c* dans la langue source **alors**

      – ajout d'un lien entre *h* et *c* dans la langue source
    – **finsi**
  – **finsi**
– **finpour**

En utilisant l'algorithme décrit ci-dessus, nous avons pu rajouter 70 relations d'hyperonymie manquantes, ce qui est en fait peu sachant que le nombre de relations d'hyperonymie en français est de 22757 dans EuroWordnet (71958 en anglais). Ceci montre que pour les concepts existant, les concepteurs du thesaurus français ont accompli sur ce type de relations un travail assez complet.

Pour les autres relations qui sont transitives, nous avons réitéré le même processus. Pour celles qui ne supportaient pas cette propriété, nous avons simplement lié un concept à un autre si les deux concepts existaient dans la langue source et si la relation trouvée en langue cible n'existait pas en langue source. Ainsi pour la relation de méronymie, 218 relations ont été rajoutées sur les 1418 présentes au départ (17530 en anglais). D'ailleurs, il faut noter qu'en étudiant ce résultat, nous avons constaté de nombreuses incohérences présentes dans les fichiers textes fournis au départ. Pour les autres relations (telles que near-antonym, has-subevent, etc.), 562 relations ont été ajoutées aux 1408 existantes dans le thesaurus français (69920 en anglais).

	Concepts	**Relation d'hyperonymie**	**Relation de méronymie**	**Autres relations**
Langue anglaise	91143	71958	17530	69920
Langue française	22737	22757	1418	1408
Ajout de relation en langue française		70	218	562

Tableau 1. Résultats concernant la mise à jour des relations dans EuroWordnet français

Nous nous apercevons donc que peu de relations de type hyperonymie/hyponymie ont été ajoutées, mais par contre un nombre conséquent de relations d'autre type ont pu être définies. Donc le thesaurus initial était assez complet concernant la relation de subsomption. La solution pour augmenter le nombre de ces relations est donc soit de les acquérir à partir de textes comme (Morin, 1999), ou soit de se servir de ressources externes pour rajouter des concepts et ensuite de réutiliser le traitement que nous avons explicité au début de ce paragraphe pour définir de nouvelles relations.

3.3. Ajout des définitions

Dans de nombreuses applications de traitement automatique des langues, et en particulier dans celles de type question-réponse, la prise en compte des définitions des termes lors de la recherche de la réponse à une question est très utile. Cette stratégie est d'ailleurs largement exploitée par les systèmes de ce type dédiés à la langue anglaise, qui utilisent les définitions de Wordnet (Moldovan *et al.*, 2003 ; Saggion *et al.*, 2004). En effet, ces définitions issues de thesaurus permettent de guider le processus de recherche de la réponse, voire même peuvent contenir la réponse. Malheureusement, un problème majeur se pose pour la langue française, car il existe peu ou pas de ressources de type thesaurus permettant l'accès à de telles définitions.

3.3.1. Wikipedia

Depuis quelques années, une heureuse initiative a vu le jour qui permet le développement sous forme de wiki d'une encyclopédie libre sur Internet : Wikipedia (2006). Dans ce contexte, des personnes peuvent contribuer en écrivant des pages relatives à des sujets divers et peuvent les intégrer à l'encyclopédie [1]. Cette source d'information qui évolue tous les jours est d'ailleurs devenue une base d'étude pour diverses communautés scientifiques (inex, 2005) (wiqa, 2005).

Lorsqu'une requête composée d'un terme (simple ou complexe) est soumise à Wikipedia, la réponse est constituée d'une page qui peut être de différentes natures :
– S'il n'y a qu'un sens répertorié dans l'encyclopédie relatif au terme, une page que nous nommons *page de définition* s'affiche. On peut, moyennant un filtrage adéquat du code HTML, récupérer le premier paragraphe de la page qui peut être considéré comme une définition du terme pour un sens donné.
– Si le terme se décline sous plusieurs sens, une page que nous nommons *page homonyme* s'affiche avec une énumération de tous les sens répertoriés dans l'encyclopédie et pour chaque sens un lien hypertexte qui renvoie à une page de définition relative à ce sens donné.
– Si le terme n'est pas dans Wikipedia une page spécifique s'affiche.

Une autre particularité intéressante de Wikipedia est qu'elle est une encyclopédie multilingue et qu'à l'aide de liens hypertextes, un internaute peut naviguer d'une langue à une autre (si le lien a été mis en place). Typiquement, lorsque le terme « bicycle » est soumis à Wikipedia anglais, une page relative à ce terme s'affiche et il est possible de passer sur les pages similaires en d'autres langues (français, allemand, turc, etc.). Il est à noter que pour les langues autres que anglaises, le lien vers la page en anglais existe le plus souvent, par contre, l'inverse n'est pas toujours vérifié !

3.3.2. Extraction de définition en langue française

Notre première idée s'est basée sur l'exploitation de Wikipedia et d'EuroWordnet en langue française pour extraire des définitions. Pour chaque concept d'Eurowordnet français, la liste des synsets correspondants est fournie à Wikipedia et toutes les pages relatives à ces synsets sont récupérées. Ensuite, la définition la plus probable est déterminée via la mise en oeuvre d'une mesure de similarité. Mais, il s'est avéré que la couverture d'EuroWordnet français n'était pas suffisante. De ce fait, de nombreux termes contenus dans les définitions n'étant pas dans le thesaurus, le critère de discrimination des définitions n'étaient pas applicables et conduisaient à des résultats erronés. En effet, certaines définitions se voyaient rétrogradées à une mauvaise place, simplement parce que peu, voire aucun terme contenu dans la définition n'étaient présents dans le thesaurus. Nous avons alors décidé d'exploiter le multilinguisme de Wikipedia et d'EuroWordnet. Nous présentons dans ce qui suit le processus mis en place ainsi que des résultats obtenus. En s'appuyant, d'une part sur l'observation que la couverture d'Eurowordnet et de Wikipedia en langue anglaise est beaucoup plus élevée que celle en langue française et d'autre part sur la possibilité du passage d'une langue à une autre, nous avons finalement extrait de Wikipedia des définitions en langue anglaise qui ont permis d'accéder à celles en langue française.

[1] Il faut bien sûr garder à l'esprit le problème de la validité de l'information. Mais voir, l'article de la revue Nature (Butler *et al*., 2005) qui montre que les articles de Wikipedia sont quasiment aussi pertinents que ceux de l'encyclopédie Britannica !

3.3.3. Processus général

La processus mis en place est détaillé dans l'algorithme ci-dessous:

- **pour** chaque concept en langue cible **faire**
  - **si** l'équivalent en langue anglaise existe dans EuroWordnet **alors**
    - soumettre chaque élément du synset du concept à Wikipedia anglais
    - récupérer les pages de définitions
    - extraire les définitions potentielles
    - **pour** chaque définition potentielle
      - calculer la similarité entre la définition dans EuroWordnet anglais et la définition potentielle
    - **finpour**
    - déterminer la définition potentielle la plus probable.
    - **si** la page qui contient la définition potentielle sélectionnée existe dans la langue cible **alors**
      - extraire la définition de la page en langue cible
      - mise à jour de la définition du concept en langue cible
    - **finsi**
  - **finsi**
- **finpour**

L'objectif est donc d'extraire les définitions de Wikipedia anglais relatives aux concepts d'EuroWordnet anglais et ensuite, si la page et le lien existent, de récupérer les définitions en langue française dans Wikipédia français. Pour un concept donné, afin de déterminer la définition anglaise la plus probable, nous utilisons une mesure qui calcule la similarité entre la définition du concept dans EuroWordnet anglais et les définitions potentielles dans Wikipedia anglais. Les définitions sont représentées par un ensemble de termes (syntagmes nominaux simples[2]) qui leurs appartiennent. Calculer la similarité entre 2 définitions, revient donc à calculer la similarité entre deux ensembles de termes. Nous nous sommes inspirés des travaux de (Halkidi *et al.*, 2003) qui ont défini une mesure de similarité entre deux ensembles de concepts qui s'appuie sur la mesure de Wu et Palmer (1994). Cette dernière mesure est donnée par l'équation (1) où c est le concept le plus proche qui subsume c_1 et c_2 (en nombre d'arcs), $depth(c)$ est le nombre d'arc qui sépare c de la racine de la hiérarchie, et $depth_c(c_i)$ avec i élément de $\{1, 2\}$ est le nombre d'arc qui sépare c_i de la racine de la hiérarchie en passant par c.

$$s_{wp}(c_1, c_2) = \frac{2 * depth(c)}{depth_c(c_1) + depth_c(c_2)} \quad (1)$$

La mesure de similarité de Wu et Palmer prend donc en compte le concept le plus proche qui subsume deux concepts (ce qui les rapproche) tout en normalisant ensuite le calcul par ce qui les différencie.

Cette mesure est moins performante que la mesure de Resnik (1999) mais plus que la traditionnelle mesure « edge counting » qui calcule le nombre minimal d'arc qui sépare deux concepts en passant par le concept le plus proche qui les subsume. Mais, dans notre contexte la mesure de Resnik (1999) n'est pas facilement applicable, car celle-ci demande des évaluations de fréquences sémantiques de concepts sur corpus que nous n'avons pas à notre disposition (ceci demande un étiquetage sémantique manuel de gros corpus). (Halkidi *et al.*, 2003) ont étendu la

[2] Les syntagmes nominaux simples sont acquis après avoir étiqueté et lemmatisé les définitions.

mesure de Wu et Palmer pour calculer la similarité entre deux ensembles *C* et *C'* de concepts de respectivement *n* et *m* éléments. L'équation relative a cette mesure est donnée ci-dessous:

$$sim_{Concept}(C, C') = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n} \max_{j\in[1,m]} s_{wp}(c_i, c'_j) + \frac{1}{m}\sum_{j=1}^{m} \max_{i\in[1,n]} s_{wp}(c'_j, c_i)) \quad (2)$$

Dans notre cas, nous déterminons la similarité entre 2 définitions qui sont représentées par 2 ensembles de termes. Nous définissons d'abord la similarité $sim(t, t')$ entre deux termes *t* et *t'* comme suit:

Soient *C* et *C'* les ensembles de concepts que représentent respectivement *t* et *t'* et qui comportent *n* et *m* éléments (en d'autre termes *C* et *C'* sont respectivement les ensembles des sens possibles de *t* et de *t'*).

$$sim(t, t') = \max_{i\in[1,n]j\in[1,m]} s_{wp}(c_i, c'_j) \quad (3)$$

La similarité entre deux termes est donc la similarité au sens de Wu et Palmer maximale entre les concepts qui peuvent être représentés par ces termes.

Nous définissons de même la similarité entre un concept *c* et un terme *t* qui est l'étiquette de *n* concepts par:

$$sim_{conceptTerme}(c, t) = \max_{i\in[1,n]} s_{wp}(c, c_i) \quad (4)$$

Nous définissons ensuite la similarité entre 2 ensembles de termes *T* et *T'* qui ont un cardinal respectif de *n* et *m* termes par:

$$Sim(T, T') = \frac{1}{n+1}(\sum_{i=1}^{n} \max_{j\in[1,m]} sim(t_i, t'_j) + \max_{i\in[1,p]j\in[1,m]} sim_{conceptTerme}(h_i, t'_j)) \quad (5)$$

Nous avons donc repris le calcul de (Halkidi *et al.*, 2003) et nous avons d'une part remplacé la similarité $s_{wp}(c_i, c'_j)$ de Wu et Palmer qui s'applique sur deux concepts par la similarité $sim(t_i, t'_j)$ entre deux termes, et d'autre part, nous ne prenons plus en compte dans le calcul de manière symétrique les similarités entre les ensembles de termes mais nous privilégions le premier ensemble de termes dans le calcul. En pratique, le premier ensemble de terme correspond aux termes représentant la définition dans EuroWordnet anglais, et l'autre ensemble de termes est celui qui représente une définition potentielle de Wikipedia. Nous avons fait ce choix car d'une part, la définition issue d'EuroWordnet étant le plus souvent concise et d'autre part, celles potentielles issues de Wikipedia étant souvent plus développées, la prise en compte exclusive dans le calcul de la similarité des concepts de la première définition et de ceux d'une seconde définition potentielle discrimine le résultat obtenu. En effet, nous calculons via notre mesure de similarité, le degré de recouvrement des concepts de la définition d'Eurowordnet et des concepts d'une définition potentielle de Wikipedia. Dans ce contexte, prendre aussi en compte dans l'évaluation le degré de recouvrement inverse conduirait à avoir des valeurs calculées moins discriminantes[3]. Il est à noter, que nous prenons aussi en compte dans le calcul les hyperonymes h_i du concept dont nous traitons la définition, mais nous nous restreignons aux

[3] Du fait de la différence de taille entre les deux ensembles de concept, le degré de recouvrement de l'ensemble des concepts d'une définition potentielle de Wikipedia et de l'ensemble des concepts de la définition d'Eurowordnet est souvent faible.

hyperonymes qui sont éloignés de plus de trois arcs par rapport à une des racines de la hiérarchie de concepts. Les définitions d'EuroWordnet anglais sont quelquefois assez sommaires, l'ajout de certains hyperonymes proche du concept (en exceptant ceux trop généraux) aide à déterminer la bonne définition dans Wikipedia qui souvent contient les termes relatifs à ces hyperonymes.

3.3.4. Les résultats

Nous avons lancé notre processus sur 1000 concepts de EuroWordnet qui sont présents à la fois dans la base anglaise et française. Les résultats sont donnés dans le tableau 2. Les concepts sont principalement relatifs à des objets concrets et des êtres vivants. Ce choix a été fait d'une part parce que, le plus souvent dans les systèmes de question-réponse, c'est ce type de concept qui est l'objet des questions (question définitionnelle), d'autre part, il ne faut pas oublier que Wikipedia est une encyclopédie, donc elle contient peu de définitions relatives à des concepts abstraits et surtout proches des racines de la hiérarchie de concept d'EuroWordnet. Nous avons d'ailleurs fait une expérimentation sur des concepts relatifs à des actes (défense, sauvegarde, etc.), nous avons obtenu de mauvais résultats car ces termes abstraits ne sont pratiquement jamais définis de manière brute dans Wikipedia, mais ils sont déclinés suivant les différents domaines et leurs différents usages dans différents domaines. Il faudrait pour ces termes plutôt que des connaissances encyclopédiques avoir des connaissances plus de type dictionnaire.

	1 Sens	**<= 5 Sens**	**> 5 Sens**	**Non Existence**	**Total**
Etiquette(s) dans Wikipedia anglais	584	157	203	56	1000
Etiquette(s) dans Wikipedia français	486	124	112		722
Erreur de détermination de définition ou similarité faible (< 0,5)	56	12	22		90

Tableau 2. Résultats d'expérimentation concernant 1000 concepts présents dans EuroWordnet anglais et français

Nous voyons que pour 1000 concepts présents dans EuroWordnet français et anglais, 56 parmi ces derniers n'ont aucune de leurs étiquettes dans Wikipedia. C'est peu, mais ceci montre la bonne couverture de Wikipedia en langue anglaise qui augmente tous les jours. 58 % des concepts (584), ne disposent que d'un sens (c'est à dire que pour ces concepts via l'ensemble de leur étiquette, on n'accéde sur Wikipedia qu'à une et une seule page de définition). Par contre, pour 10 % d'entre eux (56), la similarité calculée entre la définition d'EuroWordnet anglais et la définition de Wikipedia est faible (< 0,5). Après analyse, nous avons constaté que ces problèmes proviennent du fait d'une part, que certaines définitions d'EuroWordnet sont très sommaires et non complètes et d'autre part, qu'une définition n'est en fait représentée que par les syntagmes nominaux simples qu'elle contient. Dans certains cas, il faudrait aussi s'intéresser aux verbes et aux adjectifs. Par exemple, ceci est important dans le cas des concepts relatifs à des actes où un acte peut s'exprimer via des substantifs ou via des formes verbales.

15 % des concepts (157) ont entre 1 et 5 sens et pour environ 8 % (12) d'entre eux, le processus ne leur a pas attribué la bonne définition. 20 % des concepts (203) ont plus de 5 sens qui leur sont associés dans Wikipedia (certains en ont plus de 20) et pour environ 11 % d'entre eux (22), le processus ne leur a pas attribué la bonne définition. Les résultats erronés proviennent soit des mêmes problèmes rencontrés pour le cas d'une unique définition, soit proviennent de la non présence de la définition dans l'ensemble des déclinaisons du terme proposé.

En ce qui concerne l'extraction de la définition en langue française, nous pouvons voir que plus le nombre de sens est élevé plus le nombre des définitions présentes dans des pages en langue française diminue. Ceci s'explique par le fait que certaines définitions sont en fait directement extraites des *pages d'homonymes* en langue anglaise (aucune page de définition n'a encore été écrite), et l'équivalent de ces pages n'existe bien évidemment pas en langue française.

4. Conclusion

Dans cet article, nous avons présenté des travaux relatifs à l'amélioration d'EuroWordnet dans le but d'être utilisé dans des applications de type question-réponse en langue française. En effet, ce thesaurus présente une faible couverture au niveau du nombre de concepts et du nombre de relations par rapport à sa version en langue anglaise. Les définitions relatives aux concepts ne sont pas non plus disponibles en langue française. Nous avons d'une part, en utilisant la propriété de bilinguisme d'EuroWordnet et de passage d'une langue à une autre, cherché à augmenter le nombre de relations en langue française. Par rapport aux concepts existant en langue française, nous avons peu ajouté de relation de subsomption. Ceci signifie donc que pour l'ensemble des concepts présents, le travail des concepteurs a été assez approfondi et exhaustif. Par contre, nous avons ajouté un nombre conséquent d'autres relations (méronymie, etc.) qui n'étaient pas présentes dans le thesaurus au départ. Pour augmenter le nombre de ces relations, il faudra donc à présent, soit les acquérir à partir de textes comme (Morin, 1999), ou soit se servir de ressources externes pour rajouter des concepts dans la version française et ensuite, en appliquant le traitement que nous avons explicité dans la partie 3.2 de cet article, définir de nouvelles relations en s'appuyant sur celles déjà présentes dans la version anglaise.Nous avons aussi présenté un processus d'acquisition de définitions à partir de Wikipedia. Il s'appuie sur la propriété de multilinguisme d'EuroWordnet et de Wikipedia, et sur une mesure de similarité entre définitions. Pour améliorer les résultats nous travaillons, à l'heure actuelle, sur l'exploitation des indices définitoires (Pearson, 1998) pour aider le processus à choisir la bonne définition lorsque plusieurs définitions potentielles avec un coefficient de similarité proche sont proposées. Nous utiliserons ces résultats dans la prochaine campagne d'évaluation CLEF dédiée à la langue française. Nous pouvons aussi envisager une participation à la campagne multilingue (français-anglais). Nous avons aussi commencé des travaux sur l'ajout de concepts dans EuroWordnet français qui s'appuient aussi sur Wikipedia.

Références

BUTLER D., HOGAN J., HOPKIN M., PEPLOW M. et SIMONITE T. (2005). « Article en ligne de la revue Nature ». In *http://www.nature.com/news/2005/051212/full/438900a.html*.

CRESTAN E., LEMAIRE E. et DE LOUPY C. (2004). « Ressources pour un système de Question/Réponse ». In P. Blache (éd.), *Actes de TALN 2004 (Traitement automatique des langues naturelles)* : LPL. ATALA, Fès, Maroc, p. 19–21.

GONZALO J., VERDEJO F., CHUGUR I. et CIGARRAN J. (1998). « Indexing with WordNet synsets can improve text retrieval ». In *proceeding of the workshop on Usage of Wordnet for NLP, ACL-98*.

HALKIDI M., NGUYEN B., VARLAMIS I. et VAZIRGIANNIS M. (2003). « Thesus: Organising Web Document Collections based on Semantics and Clustering ». In A. transactions on information systems (éd.), *Journal on Very Large Databases, special Edition on the Semantic Web*.

HANKS P. et PUSTEJOVSKY J. (2005). « A Pattern Dictionnary for Natural Language Processing ». In *Revue Française de Linguistique Appliquée, numéro spécial sur les dictionaires*.

INEX (2005). « http://inex.is.informatik.uni-duisburg.de ».

JACQUEMIN B. (2004). « Analyse et expansion des textes en question-réponse ». In *7ème journées internationales d'Analyse statistique des Données Textuelles*. JAGT.

MAGNINI B., NEGRI M., PREVETE R. et TANEV H. (2002). « Mining knowledge from repeated co-occurrences : Diogene at Trec 2002 ». In *proceedings of the 11th Text REtrieval Conference*.

MILLER G. (1990). « Wordnet: an online lexical database ». In *International journal of lexicography*, volume 3. p. 235–312.

MOLDOVAN D., PASCAL M., HARABAGIU S. et SURDEANU M. (2003). « Performance issues and error analysis in an open-domain question answering system ». In A. transactions on information systems (éd.), *Journal on Very Large Databases*, volume 21. p. 133–154.

MORIN E. (1999). « Extraction de liens sémantiques entre termes à partir de corpus de textes techniques ». In *Thèse en informatique de l'université de Nantes*.

O'HARA T., MAHESH K. et NIREMBURG S. (1998). « Lexical Acquisition with WordNet and the Mikrokosmos Ontology ». In S. Harabagui (éd.), *COLING-ACL conference: Use of WordNet in Natural Language Processing Systems*. p. 94–101.

PEARSON J. (1998). In J. B. P. Compagny (éd.), *Terms in context*.

RESNIK P. (1999). « Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language ». In *Journal of artificial intelligence research*, volume 11. p. 95–130.

SAGGION H., GAIZAUSKAS R., HEPPLZ M., ROBERTS I. et GREENWOOD M. (2004). « Exploring the Performance of Boolean Retrieval Strategies for Open Domain Question Answering ». In *proceeding of the Workshop on Information Retrieval for Question Answering*. SIGIR.

WIKIPEDIA (2006). « http://en.wikipedia.org/wiki/ ».

WIQA (2005). « http://ilps.science.uva.nl/WiQA ».

WU Z. et PALMER M. (1994). « Verb semantics and lexical selection ». In *the 32nd annual meeting of the association for computational linguistics*. p. 133–138.