Utilisation du Web Sémantique pour la gestion d'une liste de diffusion d'une CoP

Bassem Makni*, Khaled Khelif*
Rose Dieng-Kuntz*, Hacène Cherfi*

*INRIA Sophia Antipolis, 2004 route des Lucioles 06902, BP93 Sophia Antipolis - France
{bassem.makni, khaled.khelif, rose.dieng, hacene.cherfi}@sophia.inria.fr

Résumé. Cet article décrit une approche de création semi-automatique d'ontologies et d'annotations sémantiques à partir de messages électroniques échangés dans une liste de diffusion dédiée au support informatique. Les ressources sémantiques générées permettront d'identifier les questions fréquemment posées (FAQ) à travers une recherche guidée par cette ontologie.

1 Introduction

L'extraction d'informations à partir de messages électroniques (mails) n'a pas été très étudiée dans la communauté du TAL[1]. Ceci est dû principalement à la présentation informelle des mails et à leurs faibles apports d'informations. Cependant, les mails peuvent être parfois la principale source de connaissances pour une organisation ou une communauté de pratique (CoP). C'est le cas d'@pretic[2] qui est une association ouverte à tous les enseignants exploitant les TIC[3] en Belgique durant leurs interactions avec les apprenants pour préparer leurs leçons. La communication dans cette CoP se fait essentiellement par échanges de mails sur une liste de diffusion décrivant des problèmes rencontrés.
Dans le but de faciliter la navigation dans cette liste de diffusion et la recherche de solutions pour des problèmes déjà posés, nous proposons une approche de création d'annotations sémantiques pour cette liste, ces annotations reposant sur une ontologie qui est elle-même extraite en partie à partir du corpus de mails. La base d'annotations créée servira pour la navigation guidée par l'ontologie en s'appuyant sur le moteur de recherche sémantique CORESE (Corby et al., 2004). Dans ce qui suit, nous présentons l'ontologie @pretic puis nous présentons un scénario d'utilisation de cette ontologie avant de conclure.

2 Construction de l'ontologie @pretic

Afin de construire l'ontologie d'@pretic, nous optons pour une approche modulaire composée de quatre ontologies, chacune dédiée à une tâche particulière : (i) une ontologie pour les

[1] Traitement Automatique des Langues
[2] Association des professeurs exploitant les TIC en Belgique francophone : http://www.apretic.be/
[3] Technologies de l'information et de la communication

composants informatiques, (ii) une ontologie pour la description des mails, (iii) une ontologie qui décrit les membres de la CoP, et (iv) une ontologie pour la description des problèmes informatiques. L'ontologie @pretic est composée des modules suivants :

1. OntoPedia : Les composants informatiques ne peuvent pas être tous évoqués dans une liste de diffusion. Nous avons réutilisé une hiérarchie existante (WeboPedia) en développant un programme permettant de récupérer automatiquement cette encyclopédie en ligne, d'en extraire la hiérarchie de termes, les instances et de générer une ontologie en RDFS[4].
2. Oemail : elle permet de décrire les métadonnées des mails en définissant des concepts génériques (EmailMessage), des concepts plus spécifiques (ReplyMessage) ainsi que des relations sémantiques (auteur, date, destinataire, etc.)
3. O'CoP : proposée dans (Tifous et al., 2007), elle permet de décrire les CoP. Nous avons réutilisé une partie de O'Cop afin de décrire les membres d'@pretic.
4. L'ontologie des problèmes informatiques : elle constitue le coeur de l'ontologie @pretic et a pour but de fournir des concepts et des propriétés permettant de décrire les problèmes rencontrés par les membres de la CoP. Pour construire cette ontologie, nous nous sommes appuyés sur le corpus de mails fourni par la CoP et nous lui avons appliqué des techniques de TAL qui nous ont aidé à amorcer l'ontologie et ensuite à l'enrichir.

3 Construction de l'ontologie des problèmes

Le corpus fourni par la CoP était très dégradé ce qui nous a mené à effectuer une phase de nettoyage importante afin d'avoir des textes d'une qualité acceptable pour le traitement par des outils de TAL.

3.1 Nettoyage du corpus

La première phase de nettoyage consiste à extraire les mails sous un format XML, supprimer les «spams», restaurer les liens entre les messages d'origines avec leurs réponses et filtrer les signatures en concevant un algorithme de détection de signature par comparaison de messages. Bien que notre corpus provienne d'une communauté francophone il était trilingue (quelques messages en anglais et en flamand). Dans une seconde phase, nous avons utilisé l'outil TEXTCAT (Canvar et Trenkle., 1994) permettant de détecter la langue.
Une des origines majeures de la dégradation des textes des mails est la non accentuation. Cette erreur orthographique a un impact négatif sur l'extraction des candidats termes. Pour résoudre ce problème, nous avons utilisé l'outil REACC[5].

À l'opposé de corpus formels où le nettoyage fournit un corps de texte non bruité, les corps des mails contiennent beaucoup de bruit même après le nettoyage. Ce bruit n'a pas de forme générique : formes de salutations, remerciements, signatures non filtrées, etc. L'écriture de filtres pour éliminer ce bruit nécessite le parcours du corpus, la localisation des zones de dégradation et l'ajout d'un filtre pour chaque zone détectée. Or cette méthode manuelle est

[4] http://www.w3.org/TR/rdf-schema/
[5] http://rali.iro.umontreal.ca/Technologies/Reacc.fr.html

lente et les filtres écrits peuvent être valables seulement pour les zones examinées. Nous avons adopté une méthode de nettoyage semi-automatique dont le but est d'accélérer la détection du bruit. Dans cette approche, la chaîne de TAL est fermée par un retour sur le texte original.

3.2 Extraction des candidats termes

L'extraction des candidats termes a pour objectif d'extraire un maximum de termes significatifs afin de construire une ontologie assez riche et couvrir la majorité des problèmes informatiques. Pour cela, nous utilisons deux approches de TAL, à savoir : syntaxique par le biais de l'outil FASTR (Jacquemin, 1997) et syntaxico-statistique implémentée par l'outil ACABIT (Daille, 1994).

3.3 Amorçage et enrichissement de l'ontologie

Pour l'amorçage de l'ontologie des problèmes, nous considérons les candidats termes provenant de messages initiaux, c'est-à-dire les messages qui ouvrent une discussion et qui sont susceptibles de soulever un problème. Ces messages partagent une régularité syntaxique par rapport aux termes utilisés pour exprimer un problème. Cette régularité consiste à utiliser le mot « problème » suivi du composant informatique concerné par ce problème. Cette étude nous a menés à amorcer la construction de l'ontologie par la sélection des candidats termes ayant comme tête le mot « problème ».

```
Problème de réception - Problème de port - Problème de réseau
Problème de câblage - Problème de connexion - Problème de rapidité
Problème avec hotmail - Problème sur le disque  - etc.
```

La formalisation d'une partie de ces termes nous a permis d'avoir un premier schéma de l'ontologie qui a été validé par les formateurs de la CoP @pretic. Cependant, cet embryon d'ontologie, quoique intéressant en couvrant en grande partie des problèmes rencontrés, est assez générique et risque d'induire une ambiguïté lors de la génération des annotations.
Afin d'enrichir notre ontologie et la rendre de plus en plus spécifique, nous avons effectué une analyse manuelle de tous les candidats termes générés par les outils de TAL. La liste suivante montre des exemples de termes extraits par les deux outils et utilisés pour l'enrichissement de l'ontologie.

```
"lenteur de connexion", "manque de mémoire", "perte de donnée",
"retard dans la réponse","ordinateur contaminé",
"mémoire insuffisante", "manque d'efficacité", etc.
```

L'étude de cette liste de termes nous permet, dans une première étape, de :

- Détecter de nouveau termes significatifs permettant d'enrichir directement l'ontologie ("manque de mémoire", "lenteur de connexion", etc.).
- Détecter des relations de synonymie entre certains termes significatifs ("mémoire insuffisante" = "insuffisance de mémoire" ou "message infecté" = "infection de message"). Ces termes synonymes se traduiront par un même concept de l'ontologie.
- Faire émerger des régularités structurelles (i.e. patrons syntaxiques) dans une grande partie des termes ("lenteur de X", "perte de X", "difficulté de X", "retard de X", "manque de X", etc.), X étant un terme de l'ontologie des *Composants*.

Dans une deuxième étape, nous nous sommes inspirés du travail réalisé dans SAMOVAR[6] (Golebiowska et al., 2001), afin de proposer des règles heuristiques qui permettent d'alimenter de façon semi-automatique l'ontologie. Ces règles détectent des structures prédéfinies dans le texte et enrichissent l'ontologie par des candidats termes qui n'ont pas été nécessairement détectés par les outils de TAL. Ces règles sont écrites en syntaxe JAPE (Cunningham, 2002) et greffées dans le processus d'annotation par l'ontologie des composants.

```
{ term1.string == 'lenteur' }
{ term2.string == 'de' }
{ term3.cpt  == 'Composant' } =>
{ term = term1+term2+term3; term.cpt = 'Problème'}
```

3.4 Rattachement semi-automatique des concepts

À l'issue de ces phases de détection de termes révélateurs de problèmes, l'ontologie des problèmes informatiques ne montre pas de relations hiérarchiques entre les concepts. Par conséquent, nous avons conçu un algorithme de rattachement automatique pour lier chaque terme à un concept générique de l'ontologie *Problème* (Problème Matériel, Problème Logiciel, etc.). Pour chaque concept de l'ontologie *Problème*, nous générons une liste de concepts voisins (dans le même message) annotés par l'ontologie *Composants*. Nous avons choisi ensuite un ensemble de concepts pivots de l'ontologie *Composants* utilisés dans la majorité des discussions. Pour chaque liste obtenue, nous calculons la somme des distances sémantiques entre les concepts de cette liste et les concepts pivots. Nous calculons ces distances grâce au moteur de recherche sémantique CORESE. La catégorie retenue pour un terme est celle qui a la distance sémantique globale la plus petite. Par exemple le terme « lenteur du réseau » est rattaché à un « problème de modems » et le terme « cas d'infraction » est rattaché à un « problème de sécurité ». La méthode adoptée pour la construction de l'ontologie des problèmes est généralisable pour d'autres listes de diffusion, ainsi qu'à d'autres domaines. En effet :

- La première phase de nettoyage des messages est indépendante du domaine.
- Nous supposons que les termes utilisés pour poser un problème ne diffèrent pas d'un domaine à l'autre, ce qui rend la phase d'amorçage réutilisable.
- Des régularités structurelles dans les termes utilisés existent ou apparaissent au fur et à mesure au sein de la même communauté (i.e. notre cas et celui de SAMOVAR), ce qui facilite la tâche d'enrichissement de l'ontologie.

4 Exploitation des ontologies

4.1 Annotation sémantique

Nous avons développé un module d'annotation sémantique qui interroge le serveur d'annotation de KIM (Popov et al., 2004). Nous avons enrichi la plate-forme KIM par nos ontologies OntoPedia et l'*ontologie des problèmes* pour fournir pour chaque message les entités nommées détectés que nous sauvegardons en RDF[7].

[6] Système d'Analyse et de Modélisation des Validations des Automobiles Renault
[7] http://www.w3.org/RDF/

4.2 Navigation guidée par l'ontologie

L'ontologie @pretic a pour but de guider la recherche dans les questions fréquemment posées pour résoudre les problèmes rencontrés dans la CoP. Nous avons ainsi développé une interface Web (cf. figure 1) permettant une navigation dans les problèmes rencontrés et leurs réponses. L'interface web fournit une vision globale de l'ontologie à travers un graphe hyperbolique. La figure 1 montre un scénario d'utilisation dans lequel le membre cherche les messages annotées par *problème d'imprimante* (1), visualise le fil de discussion correspondant (2) et la réponse à son problème (3).

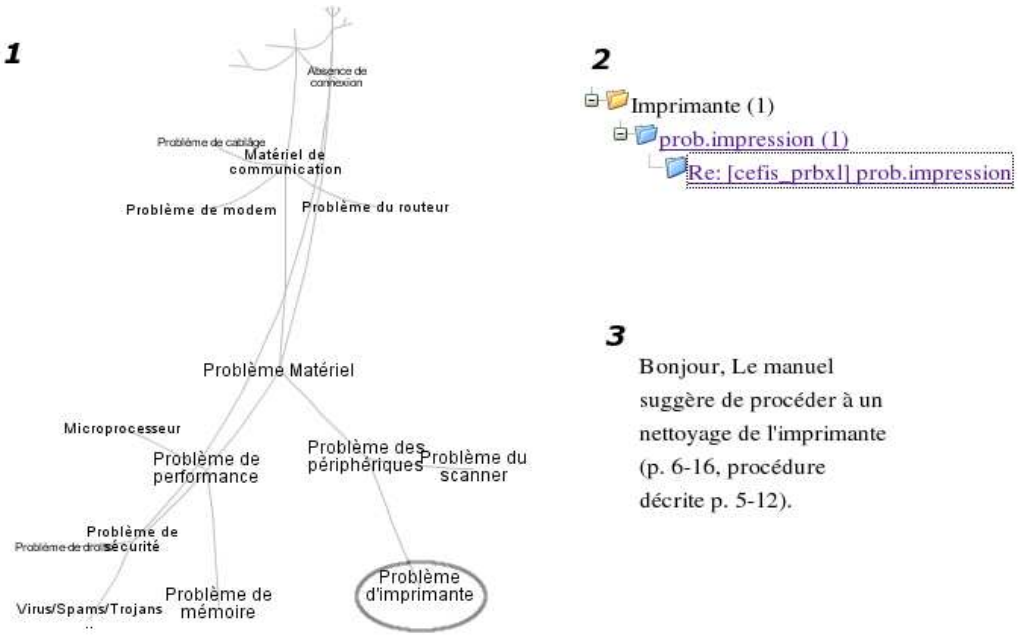

FIG. 1 – *Navigation et recherche guidée par l'ontologie.*

5 Conclusion et perspectives

Dans cet article, nous avons présenté une approche pour la génération semi-automatique d'une ontologie et d'annotations sémantiques à partir d'une liste de diffusion de support informatique. La finalité de ce travail est la proposition d'une FAQ sémantique utilisant des ontologies, des annotations sémantiques et des raisonnements offerts par un moteur de recherche sémantique (i.e. CORESE). L'approche proposée s'inspire en partie de celle de (Golebiowska et al., 2001) et plus généralement de celle de (Aussenac-Gilles et al., 2000), proposées pour la construction d'ontologies à partir des textes.
Deux des originalités de ce travail consistent en (i) l'utilisation d'un corpus très dégradé, à savoir le corpus de mails, et (ii) l'utilisation effective de deux approches et outils correspondants de TAL, ce qui à notre avis rend les résultats plus riches.

Le caractère informel des messages échangés sur la liste de diffusion a rendu les tâches de nettoyage et d'extraction très prenantes et différentes de celles présentées dans (Even et Enguehard, 2002) où les connaissances sont extraites à partir de textes dégradés mais plus au moins formels. Le nettoyage de mails a été aussi traité dans (Tang et al., 2006). Enfin, notons que l'ontologie des problèmes a été validée par les membres de la CoP @pretic ; ils seront aussi impliqués dans la génération semi-automatique de la FAQ.

Remerciements

Nous remercions le projet IST PALETTE pour le financement de ce travail et nos collègues de l'université de Liège pour la validation de l'ontologie.

Références

Aussenac-Gilles, N., B. Biébow, et S. Szulman (2000). Corpus analysis for conceptual modelling. In *EKAW'2000 workshop 'Ontologies and texts'* , pp. 13–20.

Canvar, W. et J. Trenkle. (1994). N-gram-based text categorization. *3rd Annual Symp. on Document Analysis and Information Retrieval*.

Corby, O., R. Dieng-Kuntz, et C. Faron (2004). Querying the semantic web with the corese engine. *3rd Annual Symp. on Document Analysis and Information Retrieval*, 705–709.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities 36*, 223–254.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*. Ph. D. thesis, Université de Paris VII.

Even, F. et C. Enguehard (2002). Extraction d'information à partir de corpus dégradés. In *Actes, (TALN 2002)*, Volume 1, pp. 105–114.

Golebiowska, J., R. Dieng-Kuntz, O. Corby, et D. Mousseau (2001). Building and exploiting ontologies for an automobile project memory.

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. *HDR, Université de Nantes, Nantes.*.

Popov, B., A. Kiryakov, D. Ognyanoff, D. Manov, et A. Kirilov (2004). Kim a semantic platform for information extraction and retrieval. *Nat. Lang. Eng. 10*(3-4), 375–392.

Tang, J., H. Li, Y. Cao, Z. Tang, B. Liu, et J. Li (2006). Email data cleaning. Technical Report MSR-TR-2006-16, Microsoft Research (MSR).

Tifous, A., A. E. Ghali, R. Dieng-Kuntz, A. Giboin, C. Christina, et G. Vidou (2007). An ontology for supporting communities of practice. In *K-CAP '07*, Whistler, BC, Canada, pp. 39–46. ACM.

Summary

In this paper we describe a methodology for semi-automatic creation of an ontology along with the subsequent annotation base extracted from a mailing list belonging to computer assistance community. This study raises many original issues which are unusual for NLP techniques because it starts from an email-list corpora. The challenging annotation extraction process from this email-list will feed a frequently asked questions (FAQ).