Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées

Zoulikha Heddadji[*,**], Nicole Vincent[*]
Séverine Kirchner[**], Georges Stamon[*]

[*]Université René Descartes
45, rue des Saints Pères 75270 Paris CEDEX06
[**]CSTB
84, avenue Jean Jaurès Champs-sur-Marne
77421 Marne-la-Vallée CEDEX2
{zoulikha.heddadji, severine.kirchner}@cstb.fr
{nicole.vincent, Georges.Stamon}@math-info.univ-paris5.fr

Résumé. Nous présentons dans cet article un système informatique pour le traitement des plaintes en lien avec des situations de pollution domestique écrites en français. Après la construction automatique d'une base de scénarii de plaintes, un module de recherche apparie la plainte à traiter à la thématique de la plainte la plus similaire. Enfin, il s'agit d'assigner au problème courant la solution correspondante au scénario de pollution auquel est affectée la plainte pertinente. Nous montrons ici l'intérêt de l'introduction dans l'appariement des textes de l'aspect sémantique géré par un dictionnaire généraliste de synonymes et en quoi il n'est pas réalisable pour notre problème particulier de construire une ontologie.

1 Introduction

L'objectif de notre étude est de pouvoir semi-automatiser le processus de réponse aux plaintes exprimées en français, en langue naturelle et relatives à la pollution de l'air au sein des logements. Ces plaintes reflètent chacune un cas particulier, cependant elles abordent des problèmes communs que les experts aimeraient identifier de manière objective. Notre démarche est de construire de manière automatique des scénarii. Dans la première étape nous établissons un modèle de représentation et de recherche en ne négligeant pas l'aspect sémantique. Le choix de la ressource sémantique est guidé par l'étude du vocabulaire du corpus, il est présenté dans la partie suivante. Enfin, nous présentons l'évaluation de la qualité des partitions (scénarii) obtenues.

2 Modélisation de l'espace des plaintes

Par manque de place ici, nous ne pouvons rappeler de manière détaillée nos nombreuses positions pour formaliser les textes et pour définir les différentes mesures de similarité textuelle correspondantes. Néanmoins, nous pouvons noter que pour le traitement des textes

longs, les modèles vectoriels sont les plus fréquemment utilisés, tandis que pour la comparaison des textes courts le modèle booléen flouifié est le mieux adapté. De manière générale, la plainte comporte des informations concernant les symptômes de l'occupant, elle décrit également le logement et son environnement extérieur, ..etc. Nous avons traduit ces champs de conversation sous forme de modèles de balise (unités sémantiques) dans un document XML pour enregistrer les plaintes. Les balises que nous avons retenues pour le formalisme XML des plaintes sont: symptômes, habitat et environnement

2.1 Le modèle vectoriel étendu

Zargayouna et Salotti (2004) ont étendu le modèle vectoriel de Salton en définissant le nouveau poids des termes TF-ITDF adapté à la structure XML des documents. Un vecteur des poids des termes correspond à une unité sémantique (contenu d'une balise). La similarité entre deux unités sémantiques est calculée en fonction du cosinus de l'angle formé par les deux vecteurs correspondants.

2.2 Le modèle de recherche basé sur la proximité floue des termes

Le degré de proximité floue des termes de la requête permet d'évaluer le taux de densité des termes de la requête dans les textes. Il permet ainsi de classer les documents en fonction de leur pertinence par rapport à la requête. Mercier et Beigbeder (2004) calculent la pertinence relative μ des termes de la requête aux différentes positions x dans un document d comme suit:

$$\mu_t^d(\mathrm{x}) = Max_{i \in d^{-1}}\left(Max\left(\frac{k-|x-i|}{k}, 0\right)\right)$$

Le paramètre k caractérise le degré d'influence d'une occurrence d'un terme. La pertinence d'une requête booléenne disjonctive et/ou conjonctive est calculée en prenant respectivement le maximum et/ou le minimum des pertinences locales. Le score d'un document par rapport à une requête est calculé en agrégeant les pertinences relatives locales.

2.3 Le modèle vectoriel sémantique

Le poids sémantique SemW du terme t au sein d'une balise b d'un document d au niveau du vecteur sémantique définit par Zargayouna et Salotti (2004) correspond à la somme de son poids TF_ITDF et les poids des termes qui lui sont proches sémantiquement.

$$SemW(t,b,d) = TF-ITDF(t,b,d) + \left(\sum_{i..n} Sim_{ZS}(t,t_i)\, TF-ITDF(t_i,b,d)\right) \Big/ n$$

2.4 Notre modèle de recherche

La mesure de Mercier-Beigbeder ne tient pas compte de la sémantique. Dans (Heddadji et al., 2007) nous augmentons cette mesure de manière à prendre en considération des liens de similarité latents entre documents et à l'adapter au formalisme XML.

$$\mu_t^d(x) = Max_{i \in d^{-1}(Sim(t))}(Max(\frac{(k - |x - i|)\ Sim(ti, t)}{k}, 0))$$

Un seuil de similarité est nécessaire pour délimiter l'ensemble des termes sémantiquement pertinents par rapport à t. Nous fixons cette limite à l'ensemble des termes dont le degré de similarité avec t est au-delà du score de similarité de ce dernier avec le terme auquel est rattachée la balise correspondant à l'unité sémantique où son occurrence apparaît. Dans le cas d'un corpus annoté en XML, les similarités citées calculent des appariements locaux. Une agrégation des similarités locales est nécessaire pour généraliser ces mesures au niveau «document».

3 Génération de la sémantique

Il est généralement reconnu que l'emploi d'une ontologie apporte une solution élégante au problème de la gestion de la sémantique. Les plaintes sont formulées par des particuliers. De nombreuses marques de produits et autres sigles sont utilisés, le vocabulaire est très vivant, d'où l'interrogation de la possibilité de construire une ontologie.

3.1 Le vocabulaire des plaintes

Nous avons étudié le vocabulaire des plaintes dont nous disposons. L'expérimentation que nous avons menée a eu lieu sur un corpus de 655 documents formulant des plaintes provenant de 4 organismes différents. Pour étiqueter les textes des plaintes, nous avons utilisé l'outil Tree-tagger adapté au français. Les sigles, abréviations et autres acronymes en relation avec le domaine de la pollution inconnus de Tree-Tagger sont récapitulés dans un fichier dédié. Les termes retenus dans ce fichier sont substitués automatiquement par leurs synonymes ou par des termes plus génériques compréhensibles par l'étiqueteur. La compréhension de la langue naturelle est spécifiée formellement par la notion d'ontologie. Dans des domaines précis la communauté professionnelle s'accorde autour d'une «ontologie métier». Nous analysons ici l'évolution du vocabulaire de nos plaintes qui sont en rapport avec la pollution intérieure uniquement.

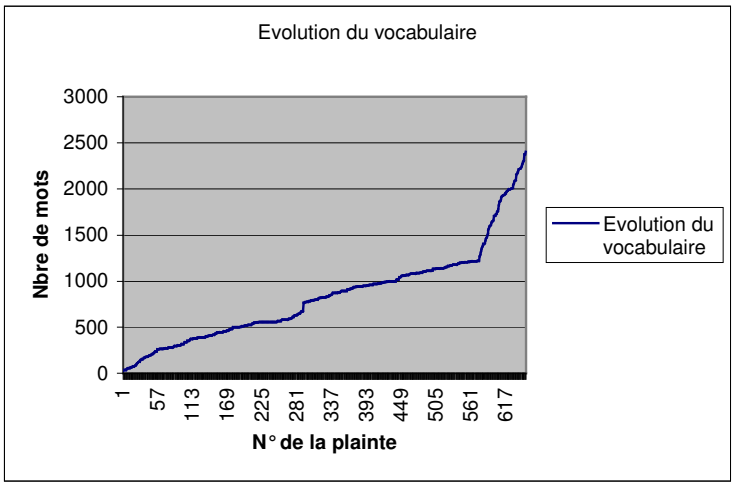

FIG. 1 – Evolution du nombre de mots en fonction du nombre de plaintes analysées.

L'allure asymptotique de la courbe de la *FIG. 1* est une preuve de l'insuffisance d'une éventuelle ontologie gérant la sémantique car nous ne connaissons pas le vocabulaire utilisé qui reste ouvert en l'état actuel de la base des plaintes. Afin de permettre l'usage de la langue naturelle dans la description des plaintes, il est nécessaire de construire un réseau conceptuel de façon à comprendre la langue française. A date, il n'existe pas d'ontologie universelle en français dans laquelle on pourrait retrouver de manière exhaustive les termes du langage naturel et qui puisse servir de base à un système de recherche implémentant la sémantique. Le résultat de notre étude conduit à considérer l'utilité des dictionnaires électroniques des synonymes pour le contrôle de la sémantique tout en assurant une couverture la plus exhaustive possible du lexique des plaintes.

3.2 DICTIONNAIRE et codage des mots

Nous avons utilisé le dictionnaire électronique des synonymes du laboratoire CRISCO de l'université de CAEN baptisé DICTIONNAIRE et qui regroupe les synonymes de 48 881 mots (vedettes) (Manguin, 2004). En plus de la connaissance sémantique que nous offre DICTIONNAIRE, nous l'utilisons en tant que vocabulaire de base. Nous utilisons les vedettes de DICTIONNAIRE en tant que primitives vectorielles caractérisant les textes et permettant de les apparier à l'aide du module de recherche implémentant les modèles vectoriels. La proximité sémantique entre deux termes A et B du dictionnaire correspond au rapport entre le nombre de synonymes communs et le nombre total de ces synonymes (indice de Jaccard). En principe, les formes fléchies d'un terme partagent le même sens (combien-même elles ne sont pas de la même catégorie grammaticale). Le dictionnaire des synonymes est insuffisant (le taux de similarité entre « pollution » et « polluer » est de 0 dans la base sémantique inférée par DICTIONNAIRE) et un système de codage des termes est nécessaire pour la gestion de la sémantique entre termes du même code. Pour coder les mots, nous avons choisi d'utiliser une heuristique, certes imparfaite, mais qui permet de rapprocher des

termes ayant la même racine. L'heuristique d'Enguehard (Enguehard ,1992) qui définit le code d'un terme comme étant la sous-chaîne des premières lettres jusqu'à l'obtention de deux voyelles non consécutives nous a paru simple à mettre en œuvre. Ayant fait ses preuves dans d'autres applications (serradura et al., 2002) nous avons appliqué ce principe pour apparier les termes issus du même code. Pour chaque paire de mots de DICIONNAIRE de même code on a effectué un échange de synonymes avec une influence de ½. Le degré de similarité entre les éléments de ces paires est de 0,5. Cette définition du lien sémantique entre termes nous permet maintenant de pouvoir comparer sémantiquement deux plaintes.

4 Comparaisons

4.1 Evaluation de la qualité des partitions obtenues

Pour définir les scénarii nous choisissons de réaliser une partition de l'ensemble des plaintes. Pour ceci nous utilisons l'algorithme des nuées dynamiques dans lequel nous avons choisi pour noyau itératif la plainte qui minimise la disparité autour d'elle. Ce travail a été réalisé parallèlement par 3 experts du CSTB (Centre Scientifique et Technique du Bâtiment) sur 100 documents d'entraînement. Les experts se sont entendus sur l'existence de 3 thématiques traitant chacune d'un phénomène de pollution ***isolé***. Le but de cette étude est d'évaluer d'une part les performances des modèles de représentation et de recherche, et d'autre part la qualité de la synthèse automatique de la base de plaintes en un ensemble de scénarii type de pollution. In fine, la solution affectée au scénario dont appartient la plainte la plus pertinente sera assignée au problème courant. Pour évaluer la qualité de nos différentes classifications, nous avons calculé le rapport entre la distance inter-classes et la distante intra-classe. De plus, nous avons comparé les partitions automatiques avec les partitions des experts en utilisant l'indice de Rand-corrigé.

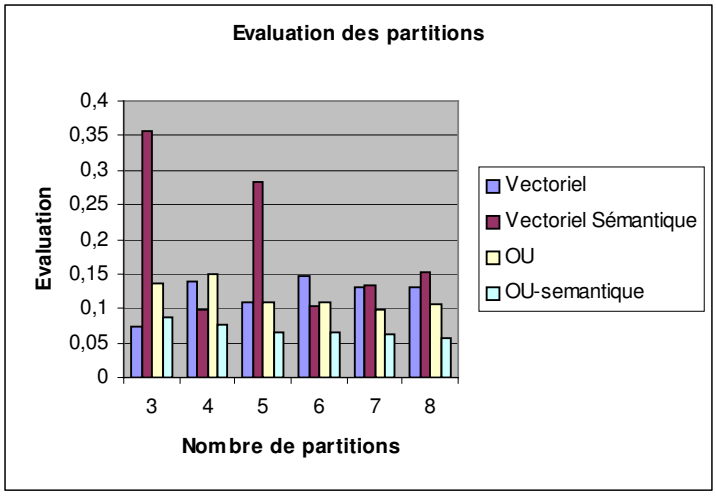

FIG. 2 – *Evaluation des partitions effectuées sous les différents systèmes d'appariement.*

Le modèle vectoriel sémantique et le modèle flou sémantique donne de meilleures classes quand l'espace de données est partitionné en 3 clusters. En effet, nous percevons dans *FIG. 2* que les modèles sémantiques partitionnent moins strictement que les systèmes directs dans le cas où le nombre de classes est supérieur à 3. D'un autre coté, les experts ont constaté l'existence de 3 classes de plaintes disjointes. Ce qui explique que le meilleur score de partition soit au niveau du graphe à la position k=3 pour les modèles sémantiques. Ailleurs, la partition sous les systèmes sémantiques constituent des clusters moins denses, ce qui explique la flexion enregistrée aux partitions supérieures.

5 Conclusion

Nous souhaitions à partir de la série d'études présentée appuyer l'idée de la correspondance existante entre le raisonnement des experts basé sur des faits et le raisonnement de nos systèmes de recherche basé sur les termes et leur sémantique. Pour améliorer encore les résultats, il nous semble pertinent de réaliser une classification floue des plaintes pour mettre en évidence des cumuls de thématiques dans une plainte particulière.

Références

Enguehard, C. (1992). ANA, Apprentissage Naturel Automatique d'un réseau sémantique. Thèse de doctorat.

Heddadji, Z. et N. Vincent et G. Stamon et S. Kirchner (2007). Extension sémantique du modèle de similarité basé sur la proximité floue des termes. RNTI, (EGC'2007).

Manguin, J-L.(2005). Regroupements de synonymes par indices de similitude : exemple avec l'adjectif ancien. Dans le Cahier de Lexicologie, n° 86.

Mercier, A. et M. Beigbeder (2005). Application de la logique floue à un modèle de recherche d'information basé sur la proximité. Dans les Actes LFA 2004, 231–237.

Serradura, L et M. Slimane et N. Vincent (2002). Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées. Inforsid, 215-228.

Zargayouna, H et S. Salotti (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Dans Actes de la conférence IC'2004.

Summary

Nowday, indoor air complaints are left without answers. This is caused by several reasons, in particular because the expertise in indoor air domain is new and rare. We propose in this paper a new approach to resolve the indoor air complaints. This approach is composed of three functions. First, a set of classes is automatically built and each classe corresponds to a collection of complaints having the same theme.Second, the research module seek for the most similar complaint stored in a compact base to the new problem. The third function atribute to the new complaint the solution established and juged avalable for the scenario of pollution represented by the set of documents in which belong the most similar complaint.