Fouille de textes pour orienter la construction d'une ressource terminologique

Valentina CEAUSU, Sylvie DESPRES

Université René Descartes
CRIP5 – Equipe IAA – Groupe SBC
UFR Mathématiques et Informatique
45 rue des Saints-Pères
75006 PARIS
valentina.ceausu@math-info.univ-paris5.fr
sd@math-info.univ-paris5.fr

Résumé. La finalité de ce papier est d'analyser l'apport de techniques de fouille de données textuelles à une méthodologie de construction d'ontologie à partir de textes. Le domaine d'application de cette expérimentation est celui de l'accidentologie routière. Dans ce contexte, les résultats des techniques de fouille de données textuelles sont utilisés pour orienter la construction d'une ressource terminologique à partir de procès-verbaux d'accidents. La méthode TERMINAE et l'outil du même nom offrent le cadre général pour la modélisation de la ressource. Le papier présente les techniques de fouille employées et l'intégration des résultats des fouilles dans les différentes étapes du processus de construction de la ressource.

1 Introduction

La finalité de ce papier est d'analyser l'apport des techniques de fouille de données textuelles à une méthodologie de construction d'ontologie à partir de textes. Le domaine d'application de cette expérimentation est celui de l'accidentologie routière. Une ontologie du domaine a été élaborée à partir de connaissances expertes (Després, 2002). Le travail présenté concerne la construction d'une ressource terminologique à partir de procès verbaux d'accidents (PV) rédigés par les forces de l'ordre. Les deux ressources (ontologique et terminologique) seront exploitées dans un système de raisonnement à partir de cas ayant comme cas cible des procès verbaux et comme cas source des scénarios d'accidents.

Dans ce contexte sont utilisés simultanément : (a) un algorithme de reconnaissance de motifs qui engendre un ensemble de syntagmes nominaux et verbaux ; (b) l'algorithme Apriori pour affiner les syntagmes nominaux identifiés à l'aide de motifs ; (c) l'ontologie de l'accidentologie pour affiner les syntagmes verbaux et (d) la méthodologie TERMINAE de construction de la ressource terminologique (Biébow, Szulman, 2000). Après avoir présenté les techniques de fouille de textes utilisées, leur apport à l'élaboration de la ressource terminologique est discuté. En conclusion, les améliorations à apporter aux différentes techniques sont discutées.

2 Extraction des connaissances : techniques de fouilles

Une ressource terminologique est une structuration des termes spécifiques à un domaine particulier qui permet de créer une modélisation des connaissances du domaine. La modélisation des connaissances est spécifique à la tâche pour laquelle la ressource terminologique est construite. L'objectif de notre démarche est la construction d'une ressource terminologique à partir de textes en langue naturelle en adoptant une approche mixte : l'utilisation de la méthode et de l'outil TERMINAE tout en orientant le processus de construction par les résultats issus des fouilles des textes.

Le corpus est constitué d'environ 250 procès-verbaux (PV) d'accidents de la route survenus dans la région de Lille. Un PV est un document établi par les gendarmes ou les agents de police. Les PV de police ont préalablement été rendus anonymes par le logiciel PACTOL (Centre d'Etudes Techniques de l'Equipement (CETE) de Rouen). Un PV comprend des textes rédigés en langue naturelle (synthèse des faits, nature des faits, déclarations des impliqués etc.) et des rubriques correspondant à des variables concernant les lieux, les véhicules et les personnes impliquées.

Nous avons fait appel à des techniques de fouilles de textes afin de retrouver à partir d'un corpus, des termes du domaine et des relations entre les termes identifiés. Ceci dans la mesure où les termes représentent l'expression linguistique des concepts et forment des indicateurs privilégiés de la connaissance portée par les documents (Ville-Ometz et al., 2004).

Un module d'extraction des connaissances a été développé. Il utilise en entrée les résultats fournis par un étiqueteur syntaxique (Cordial ou TreeTagger) et extrait des connaissances en utilisant des modèles prédéfinis au niveau linguistique : les motifs. Un motif est un regroupement de catégories lexicales, par exemple (Nom, Nom) ou (Verbe, Préposition, Nom).

La *« génération des regroupements »* permet la définition et l'identification des motifs. Un ensemble de regroupements de mots correspondant aux motifs définis est engendré automatiquement. Deux catégories de motifs ont été définies : les motifs nominaux ayant comme premier terme un nom et les motifs verbaux dont le premier terme est un verbe. Les relations conceptuelles associées aux motifs nominaux traduisent par exemple des liens d'hyponymie (Hearst, 1992 ; Morin, 1999), celles associées aux motifs verbaux portent sur les propriétés entre les concepts qui seront traduites comme des rôles. L'algorithme de reconnaissance des motifs s'applique au niveau de chaque phrase et identifie les instances des motifs définis. L'ensemble des regroupements obtenus (Fig. I) constitue le résultat de l'exécution de l'algorithme. Un regroupement peut représenter : une construction verbale ***{venir de, tourner sur droite}*** ; des termes du domaine ***{balise de priorité, priorité du passage}*** ; une relation entre des termes du domaine ***{(propriétaire, véhicule) ; (passager, véhicule)}*** ; des regroupements sans contenu sémantique qui constituent du bruit ***{(c, véhicule) ; (venir de 3o6)}***. Le nombre des regroupements obtenus est important (environ 44000).

```
(Nom, Nom  ;- fait, circonstance)
(Nom, Préposition, Nom ; - usager, de, route)
(Verbe, Préposition, Adjectif ; - circuler, sur, gauche)
```

Fig. I - *Exemples de motifs et des regroupements associés*

A ce stade, des affinages sont nécessaires pour permettre l'exploitation des connaissances extraites. Les procédures d'affinage sont spécifiques à chaque catégorie de regroupements.

L'affinage des syntagmes nominaux est effectué par l'application de l'algorithme APRIORI, le recours à l'ontologie de l'accidentologie permet de préciser l'ensemble des syntagmes verbaux.

2.1 Affinage des résultats obtenus

Les règles d'association sont employées en fouille de données et constituent de bons indicateurs pour identifier les régularités dans des grands volumes de données. En fouille de textes, les règles extraites peuvent être interprétées comme des cooccurrences de termes dans les textes et par conséquent refléter des liens sémantiques entre les termes. Dans le domaine de l'ingénierie ontologique (Maedche et Staab, 2000), les règles d'association ont été utilisées pour découvrir des relations non taxinomiques entre des concepts en utilisant une hiérarchie de concepts comme connaissance de base. L'algorithme APRIORI tel qu'il est utilisé par (Maedche et Staab, 2000) a été adapté à notre problème. Il aide à l'élimination des regroupements accidentels et produit un ensemble de regroupements contenant des termes du domaine (usager de route) et des relations entre les termes du domaine (conducteur, véhicule). La génération de l'ensemble des motifs nominaux fait partie intégrante de l'algorithme APRIORI. Nous travaillons à partir d'une phrase d'où sont extraits des regroupements grâce aux motifs qui ont été définis. Une transaction est une phrase du corpus. On définit une règle d'association par une relation R : (X=>Y), où X (prémisse de la règle) et Y (conclusion de la règle) sont des regroupements de mots. Nous avons utilisé deux formes restreintes de règles d'associations : la forme (R1) restreinte à deux mots (1 mot en prémisse, 1 mot en conclusion) ; la forme (R2) restreinte à trois mots (1 mot en prémisse, une association à 2 mots en conclusion). Les motifs précédemment définis permettent la construction de règles d'association correspondant aux deux formes. Un motif (Nom, Nom) engendre des associations ayant la forme (1) (X = conducteur, Y = véhicule) ; un motif (Nom, Préposition, Nom) crée une association de la forme (2) (X = ceinture, y = de securité). La forme R1 permet de retrouver des relations entre termes et les concepts intervenant en accidentologie peuvent être retrouvés grace à la forme R2.

Deux mesures de qualité, le *support* et la *confiance* sont utilisées pour ordonner les règles extraites selon leur pertinence pour la modélisation. Le *support* de R : (X=>Y) représente le pourcentage des phrases contenant les termes de $(X \cup Y)$ (dans notre application, {x, y} ou {x, y_1,y_2}). La *confiance* correspond au pourcentage de phrases contenant les regroupements X et Y $(X \cup Y)$ calculé par rapport à l'ensemble des phrases contenant le regroupement X. Elle mesure le degré de validité d'une règle. Lorsque la confiance vaut 1, la règle est dite totale, dans le cas contraire elle est dite partielle. Des seuils sont définis pour les mesures de qualité pour éliminer les règles triviales : minsup pour le support minimal et minconf pour la confiance minimale. Les valeurs du support inférieures à minsup correspondent à des associations rares que nous considérons comme accidentelles (bruit). Les valeurs v du support telles que (minsup<v<1) indiquent le plus souvent des concepts génériques du domaine.

Dans une première version de l'implantation de l'algorithme, des valeurs arbitraires étaient affectées aux seuils minsup et minconf. Une seconde version permet une intervention de l'expert pour sélectionner des valeurs appropriées pour les deux seuils parmi les valeurs du support et de la confiance. Parmi les règles retenues à l'issue de cette étape d'affinage, des termes composites du domaine tel que (X= voie, Y= de contournement) sont observées et différents types de relations (Fig. II) sont mises en évidence. Après un recours à l'ontologie de l'accidentologie, les relations fonctionnelles seront étiquetées par des verbes.

Relation identifiée	***Type de la relation***
fourgonnette, véhicule	est-un
automobile, véhicule	est-un
volant, véhicule	partie-de
véhicule, propriétaire	fonctionnelle
conducteur, véhicule	fonctionnelle
conducteur, camion	cas particulier de la relation précédente

Fig. II - *Relations conceptuelles découvertes*

Le traitement des syntagmes verbaux est réalisé sur deux plans : l'identification de classes des verbes dans l'ensemble des regroupements verbaux et l'utilisation de l'ontologie de l'accidentologie pour affiner les classes identifiées. Une classe de verbes contient l'ensemble des regroupements constitués à partir d'un verbe. Chaque classe de verbes contient deux catégories de regroupements (Fig. III): celles à deux termes correspondant à un motif « verbe, préposition » (par exemple, diriger vers) et celles à trois termes engendrées par un motif « verbe, préposition, syntagme nominal » (par exemple, diriger vers bretelle). Les regroupements à trois termes sont obtenus à partir des regroupements à deux termes en ajoutant une extension qui correspond à la fonction grammaticale de complément.

```
diriger vers; diriger sur; diriger dans
diriger vers square; diriger vers opéra; diriger vers esplanade
```

Fig. III - *Extrait de la classe « diriger »*

Les résultats obtenus font apparaître, à l'intérieur de chaque classe, un nombre réduit de regroupements à deux termes auxquelles correspondent un nombre assez important d'extensions possibles qui conduisent à des regroupements à trois termes. Cependant les regroupements à trois termes sont à un niveau de granularité trop fin pour être exploité. Pour pallier cet inconvénient, nous avons recours à l'ontologie de l'accidentologie.

Un regroupement à trois termes est composé d'un verbe, d'une préposition et d'un terme qui relève d'un concept du domaine. Des listes (Fig. IV) des termes associés aux extensions sont constituées. Une intervention manuelle est nécessaire pour associer chaque liste à un concept de l'ontologie.

```
Liste « voie » :(autoroute, avenue, boulevard, bretelle)
Liste « lieu » :(commune, domicile, école, garage, gare)
```

Fig. IV - *Liste de termes associée à un concept*

L'utilisation de l'ontologie réduit ainsi le nombre des regroupements à trois termes (Fig. V). Elle élimine également le bruit en permettant de supprimer les regroupements dans lesquels figurent des termes parasites comme « diriger_vers_12 » ou des regroupements comme « diriger_par_sapeur » qui ne relève du sens du terme dans le contexte étudié (diriger au sens de commandement n'est pas

le sens commun en accidentologie). Toutefois, si ce traitement réduit le nombre de regroupements à trois termes, il risque d'éliminer des syntagmes valides si les listes construites sont incomplètes.

```
diriger vers Direction; diriger vers Lieu ; diriger vers Voie
```

Fig. V - *Classe « diriger » affinée avec un recours à l'ontologie*

3 Orientation du processus de construction

Afin de clarifier la présentation, les étapes prônées par la méthodologie TERMINAE sont rappelées : (a) la sélection des termes obtenues à l'issue d'un traitement réalisé par l'analyseur syntaxique de corpus Syntex (Bourigault & Fabre 2000) ; (b) l'étude dans le corpus des occurrences d'un terme et de ses relations lexico-syntaxiques à l'aide de motifs grâce au module Linguae ; (c) l'établissement d'une fiche terminologique où sont définis les différents sens du terme ; (d) chaque sens du terme est ensuite considéré et normalisé relativement au corpus, à l'application considérée, au point de vue choisi ; cette normalisation définit un concept terminologique ; (e) la construction d'une ontologie formelle qui pourra être validée et permettra des inférences est élaborée.

Dans l'étape (a), la sélection des termes est orientée par les résultats des règles d'associations. La liste des termes composites validés par l'algorithme APRIORI constitue une aide au moment de la sélection des candidats termes et permet d'éliminer plus rapidement ceux qui ne sont pas pertinents pour le travail effectué. Dans l'étape (b) les relations identifiées par l'algorithme APRIORI interviennent en permettant d'étudier les occurrences d'un terme dans le corpus. Elles améliorent également l'utilisation du module LINGUAE de traitement des relations. Dans l'étape (d) des concepts terminologiques modélisés sont enrichis en intégrant des termes découverts par APRIORI. L'étape (e) consiste en la modélisation des concepts et des rôles au niveau formel. Les concepts issus de l'étape (d) servent de base pour la création des concepts formels. Les rôles décrivent les relations entre les concepts. Les syntagmes verbaux constituent de bons indicateurs pour les modéliser. Les rôles sont créés à partir des classes de verbes et la structure de la ressource terminologique est affinée par l'intégration des relations issues des travaux sur les syntagmes nominaux.

4 Conclusion et perspectives

D'un point de vue pratique, notre démarche facilite la construction d'une ressource terminologique en automatisant une partie des traitements sur les textes. Les premiers résultats obtenus (extraction des termes du domaine, construction de classes de verbes et émergence de relations générique et fonctionnelle) permettent d'orienter le processus de construction. Ils offrent une aide pour le choix des termes à sélectionner et constituent des indicateurs relatifs aux associations entre les termes. L'élaboration des concepts terminologiques dans la phase de normalisation est enrichie par le recours à l'ontologie de l'accidentologie qui sert d'aide à l'élimination de certains regroupements.

Des algorithmes qui engendrent des regroupements à d'autres niveaux que la phrase comme le paragraphe ou la proposition vont être intégrés. La définition de motifs syntaxiques capables d'engendrer des regroupements plus pertinents est aussi envisageable. Pour la validation de l'ensemble de syntagmes nominaux des mesures de qualité fondées sur un modèle de connaissances

(Cherfi et *al*., 2003) pourront être utilisées. Le recours à l'ontologie pour associer les termes de la ressource terminologique aux concepts doit en partie être automatisé. L'utilisation de techniques de fouilles de textes pour permettre de repérer des propriétés structurelles et fonctionnelles de chaque terme fait également partie des perspectives étudiées.

Références

Biébow B., Szulman S., TERMINAE : A linguistic-based tool for the building of a domain ontology11th European Workshop, Knowledge Acquisition, Modeling and Management (EKAW' 99), Dagstuhl Castle, Germany, 26-29 Mai, 1999, p. 49-66.

Bourigault D. & Fabre C., Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaires, n° 25, 2000, Université Toulouse - Le Mirail, pp. 131-151.

Cherfi H., Toussaint, Y (2002) – Adéquation d'indices statistiques à l'interprétation de règles d'association. In Actes JADT 2002 : 6èmes journées internationales d'Analyse statistique des Données Textuelles.

Cherfi H., Napoli A., Toussaint Y., Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association, CAP'2003.

Després S., "Contribution à la conception de méthodes et d'outils pour la gestion des connaissances", Habilitation à Diriger des Recherches en Informatique, Université René Descartes, décembre 2002.

Faure, D., Nedellec, C., (1998) A corpus-based conceptual clustering method for verb frames and ontologies acquisition. *LREC workshop on adapting lexical and corpus resources to sublanguages and applications, Granada, Spain* .

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O. Rajman, M., Schler Y., Zamir O. (1998) Text Mining at the term level. LNAI: Principle of Data Mining and Knowledge Discovery, 1510(1), 65-73.

Hahn, U., Schnattinger, K., (1998) Towards text knowledge engineering. Proc. of AAAI'98, pages 129-144.

Hearst, M.A., 1992 Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France.

Kodratoff Y. (1999) Knowledge Discovery in Texts: A definition, and Applications. In LNAI. Proc.of the 11th Int'lSymp.ISM'99, volume 1609, p. 16-29, Warsaw:Springer.

Maedche, A., Staab, S. (2000) Mining ontologies from text. In Knowledge Acquisition, Modeling and Management, 12th International Conference, EKAW 2000, pages 189-202.

Morin, E., 1999 Automatic acquisition of semantic relations between terms from technical corpora. In Proc. of the Fifth International Congress on Terminology and Knowledge Engineering - TKE'99.

Séguéla P., (1999) " Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés ", in Actes de TIA'99 (Terminologie et Intelligence Artificielle), Nantes, Terminologies Nouvelles n°19, pp 52-60.

Srikant, R., Agrawal, R, (1997) Mining generalized association rules. In " Future Generation Computer Systems », pages 161—180.

Toussaint Y., Simon A., Cherfi H. (2000) – Apport de la fouille de données textuelles pour l'analyse de l'information. In Actes de la conférence IC'2000, Ingénierie des connaissances, pp. 335-344, Toulouse, France.

Ville-Ometz, F., Royauté, J., Zasadzinski A., Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrôlée. In Actes de CIFT 2004, pp.89-101.