Une approche mixte pour la construction d'une ressource terminologique

Valentina Ceausu, Sylvie Despres

Une approche mixte pour 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, la méthodologie de construction de la ressource terminologique TERMINAE et les techniques de fouille de données textuelles sont utilisées simultanément. Les résultats issus de ces techniques sont utilisés pour contribuer à la construction de la ressource terminologique avec TERMINAE.

Mots-clés : Construction de ressource terminologique à partir de textes, TERMINAE, Techniques de fouille de données.

1 Introduction

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. Cette réflexion fait suite à un travail qui a consisté en la création d'une ressource terminologique pour l'analyse des accidents de la route à partir d'un corpus de procès verbaux des accidents survenus dans la région de Lille. Une ressource terminologique représente une structuration des termes spécifiques à un domaine particulier et 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.

Une ontologie de l'accidentologie élaborée à partir de connaissances expertes (textes du domaine, scénarios d'accidents rédigés par les chercheurs du domaine et entretiens réalisés auprès des chercheurs) existe (Després, 2002) et est implémentée à l'aide de Protégé 2000 (Chafai et Després, 2003). Le travail présenté a consisté à construire une ressource terminologique à partir de procès-verbaux d'accidents afin d'enrichir cette ontologie. La ressource ainsi enrichie doit permettre la comparaison de procès-verbaux d'accidents rédigés par les forces de l'ordre (gendarme ou police)

au moment où un accident survient et les scénarios d'accidents qui constituent le prototype de déroulement d'un ensemble d'accidents. Le format XML est utilisé pour exprimer les deux ressources et permet d'établir des correspondances entre elles. Il n'y a pas pour l'instant d'intégration de la ressource terminologique issue des PV dans l'ontologie de l'accidentologie. L'exploitation de ces deux ressources intervient dans la phase de conception d'une base de cas constituée des scénarios d'accidents et de l'élaboration des cas cibles à partir des procès-verbaux d'accidents d'un système de raisonnement à partir de cas conçu pour aider à l'amélioration du réseau urbain.

Dans ce contexte, la méthodologie de construction de la ressource terminologique TERMINAE (Biébow, Szulman, 2000) et des techniques de fouille de données (un algorithme de reconnaissance de patrons (Ceausu, 2003) et l'algorithme A PRIORI (Agrawal &Srikant, 1994)) sont utilisés simultanément. Les résultats issus des techniques de fouille de données textuelles contribuent à construire la ressource terminologique avec TERMINAE. L'ontologie de l'accidentologie implantée à l'aide de Protégé sert d'aide à l'élimination de certains regroupements et à la dénomination des regroupements sélectionnés.

Après avoir rappelé les arguments qui ont conduit à utiliser ces deux moyens de traitement de ressources textuelles, la méthodologie TERMINAE et ses limites sont discutées ainsi que les apports des techniques de fouille de données textuelles à l'élaboration de cette ressource. Le travail est situé par rapport aux résultats déjà obtenus dans ce domaine. Une proposition pour systématiser cette approche est faite dans le cadre de l'utilisation de TERMINAE.

2 Une approche mixte de construction de la ressource

Les ressources disponibles pour débuter ce travail consistaient en un corpus de procès-verbaux dont il fallait extraire des termes représentatifs du domaine. L'expérience acquise au cours de la construction de l'ontologie de l'accidentologie (Després, 2002) a abouti à la conviction de la nécessité d'automatiser certains des traitements à effectuer sur les termes extraits des textes étudiés. Les résultats obtenus par les techniques de fouille de données textuelles répondaient aux besoins identifiés.

Une approche mixte qui associe une méthodologie de construction de ressources terminologiques avec des techniques de fouille a été jugée pertinente puisque les résultats obtenus semblaient complémentaires. Cette approche se résume en l'utilisation : des outils avancés pour le traitement de la langue naturelle pour identifier et structurer les connaissances du domaine ; des techniques de fouille des textes pour développer des solutions particulières afin d'affiner et enrichir la ressource terminologique obtenue.

Une étude des outils permettant d'élaborer une ressource terminologique a été faite. Dans une première approche l'idée était d'être complètement maître des outils qui permettraient de travailler en amont sur les textes. Une première tentative a été d'examiner GATE (General Architecture for Text Engineering) (Cunningham, 2002 ; http://gate.ac.uk/download/) afin de le coupler avec le logiciel WEKA (Witten & *al.*, 1999). L'absence de dictionnaire en langue française ayant un format exploitable par

GATE nous a conduit à abandonner cette idée. Le choix s'est alors porté sur TERMINAE auquel des techniques de fouilles de données ont été associées. L'intérêt du choix de TERMINAE est qu'il s'agit avant tout d'une méthodologie dont les étapes sont bien identifiées et qui par conséquent permet de repérer plus facilement les moments où une automatisation de certains traitements est utile.

2.1 Terminae

TERMINAE est une méthodologie de construction de ressources terminologiques à partir de l'étude de textes qui est supportée par un outil du même nom. Pour une application donnée, TERMINAE utilise des textes du domaine et assiste l'utilisateur dans sa démarche de modélisation de la ressource, du repérage des termes à la formalisation des concepts.

Les différentes étapes de construction de l'ontologie sont : (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 patrons 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.

Lors de l'utilisation de TERMINAE, en tant que la plateforme d'aide à la construction d'ontologies, nous avons été confrontées à des traitements manuels des termes qu'il nous a semblé utile d'automatiser. Dans l'étape de sélection des termes, la difficulté réside dans le choix des bons termes si l'on ne dispose de connaissances supplémentaires. Des indicateurs relatifs aux associations entre les termes peuvent aider au choix de ces derniers, l'utilisation de LINGUAE peut être améliorée si l'on dispose de relations déjà identifiées entre les termes. Les patrons les plus fréquents sont alors plus faciles à construire. L'élaboration des concepts terminologiques notamment dans la phase de normalisation peut être enrichie en utilisant l'ontologie de l'accidentologie (pour désigner le concept) et en utilisant des associations entre les termes pour compléter les concepts ainsi obtenus.

2.2 Les techniques de fouille

Le constat dressé à l'issue des premières utilisations de TERMINAE a conduit à rechercher des techniques répondant aux besoins précédemment identifiés. En outre, les travaux réalisés en accidentologie et la modélisation des connaissances du domaine centrée sur les actions ont montré la nécessité de disposer d'un outil qui permette de déterminer des patrons traitant des catégories lexicales telles que les verbes et les prépositions qui leur sont associées. Les techniques de fouille de données textuelles répondaient à ces besoins et sont utilisées pour contribuer à construire la ressource terminologique avec TERMINAE.

2.2.1 L'algorithme de reconnaissance des patrons

L'algorithme de reconnaissance des patrons est exécuté sur des résultats fournis par un analyseur syntaxique (Cordial pour cette version, depuis Treetager est utilisé). Un patron est un regroupement de catégories lexicales. L'algorithme s'applique au niveau de chaque phrase et identifie deux catégories de patrons : les patrons nominaux dont le premier terme est un nom et les patrons verbaux qui ont comme premier terme un verbe. L'ensemble des motifs (instanciation des patrons par les termes sélectionnés) constitue le résultat de l'exécution de l'algorithme.

L'algorithme de reconnaissance de patrons s'écrit :

```
Pour chaque phrase du corpus
(1)Eliminer les éléments de contexte (noms propres) et les
références (pronoms) ;
(2)Pour chaque mot de la phrase :
   Vérifier si sa catégorie lexicale avec les catégories
lexicales de ses voisins  peut engendrer un patron ;
      En cas affirmatif, construire un motif ;
      En cas négatif, analyser le mot suivant ;
(3)Ajouter les motifs découverts à l'ensemble des motifs ;
```

2.2.2 L'algorithme A PRIORI

La finalité de cette phase est de découvrir des relations entre certains des termes apparaissant dans les procès-verbaux. Les relations recherchées sont celles apparaissant entre des noms (par exemple, passage piéton).

Les règles d'association ont été appliquées à la fouille de textes (Feldman et *al*. 1998 ; Kodratoff, 1999). L'algorithme A PRIORI (Agrawal &Srikant, 1994) de découverte de règles d'association a été adapté au traitement des mots dans le cadre des travaux de (Maedche et Stabb, 2000) pour construire automatiquement une ontologie à partir de textes.

La version de l'algorithme présenté a été adaptée à notre problème. Nous travaillons à partir d'une phrase d'où sont extraites des relations restreintes grâce aux patrons qui ont été identifiés dans le domaine de l'accidentologie.

Formellement, soient :

$M = \{m_1, m_2, \ldots m_k\}$ l'ensemble des mots

$T = \{t_1, t_2, \ldots, t_p\}$ l'ensemble des transactions. Une transaction est une phrase du corpus.

Une règle d'association est une relation (X=>Y), où X et Y sont des regroupements de mots.

$$X = \{x_1, x_2, ..., x_n \mid x_i \in M\}$$
$$Y = \{y_1, y_2, ..., y_n \mid y_i \in M\}$$

Dans notre contexte, nous avons utilisé deux formes restreintes d'associations :

Forme (1) : (X=>Y) où

$$X = \{x \mid x \in M\}, \; Y = \{y \mid y \in M\} \qquad (1)$$

Forme (2) : (X=>Y), où

$$X = \{x \mid x \in M\}, \; Y = \{y_1, y_2 \mid y_1, y_2 \in M\} \qquad (2)$$

Les patrons définis nous permettent la construction des associations correspondant aux deux formes. Par exemple, un patron (Nom, Nom) engendre des associations ayant la forme (1) (X=conducteur, Y=véhicule) ; un patron (Nom, Preposition,Nom) crée une association de la forme (2) (X=ceinture, y=de securité).

L'algoritme utilise deux indicateurs afin d'estimer la pertinence d'une association.
Considerons l'association :

$$(X \Rightarrow Y)(X, Y \subset T)$$

Le support représente le pourcentage des transactions contenant les deux items ($X \cup Y$).

$$Support \quad (X \Rightarrow Y) = \frac{|\{t_i \mid (X \cup Y) \subset t_i\}|}{|\{t_i\}|}$$

Un seuil au dessus duquel le motif est considéré comme fréquent est fixé. Les valeurs faibles du support correspondent à des associations rares que nous considérons comme accidentelles. La confiance correspond au pourcentage des transactions contenant les deux items ($X \cup Y$) calculé par rapport à l'ensemble des transactions contenant le premier item X.

$$Confiance \quad (X \Rightarrow Y) = \frac{|\{t_i \mid (X \cup Y) \subset t_i\}|}{|\{t_i \mid X \subset t_i\}|}$$

Lorsque la confiance vaut 1, la règle est dite exacte et un seuil minimal est fixé pour engendrer que des règles dont la confiance est comprise entre ce seuil et 1.

L'utilisation des seuils pour le support et la confiance nous a permis d'identifier les associations pertinentes. Un motif ayant un support et une confiance supérieurs aux seuils imposés représente une règle d'association. D'autres mesures de qualité aidant à l'interprétation des règles (Cherfi et *al*., 2003) pourront par la suite être utilisées.

L'algorithme utilisé s'écrit :

```
Identifier l'ensemble des associations (à l'aide des patrons)
A PRIORI :
 (2)Pour chaque association:
       calculer le support ;
       calculer la confiance ;
(3) éliminer les associations ayant une confiance et un
support inférieurs aux seuils imposés ;
```

Parmi les règles identifiées on retrouve des relations entre termes (X=véhicule, Y=conducteur) identifiées grâce à la forme (1) et des termes du domaine (X= voie, Y= de contournement) qui correspondent à la forme (2).

2.3 Apport des techniques de fouilles à l'utilisation de TERMINAE

Les techniques de fouilles de données sont utiles dans les différentes étapes de l'utilisation de TERMINAE. Dans la phase d'élaboration de la ressource linguistique, les règles d'associations constituent des indicateurs pour sélectionner les termes dans la liste fournie par Syntex. Les motifs obtenus à partir du module des transactions sont utiles pour l'utilisation de LINGUAE. Dans la phase de normalisation, les associations apparaissant dans le texte qui ont été identifiées à l'aide de l'algorithme APRIORI et qui ne figurent pas dans la liste des termes de Syntex sont ajoutées afin de compléter les classes de termes associées au concept formel défini. Les classes sémantiques de verbes permettent de définir les rôles associés au concept dans la phase de la construction de l'ontologie avec TERMINAE

3 La réalisation

Dans ce paragraphe, après avoir défini le corpus à partir duquel la ressource terminologique est établie, les différentes étapes des traitements réalisés sont décrites.

3.1 Le corpus

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 langage libre (synthèse des faits, nature des faits, déclarations des impliqués et selon le cas des témoignages) et des rubriques

correspondant à des variables concernant les lieux, les véhicules et les personnes concernées.

Synthèse des faits

D'après les déclarations, le véhicule FORD nø XXXXXX conduit par M XXXXXXXXXXXXXXXoulevard de la République venant de la Place Marcel Sembat et se dirige vers le Pont d'Issy. Arrivé à la hauteur du passage piéton implanté au 74 du dit boulevard, il aperçoit un piéton Me XXXXXXXXXXXXXXXr ce dit passage, arrêté au milieu de la chaussée. Il s'arrête puis redémarre et entre en collision au niveau de son avant gauche avec Me XXXXXXXXXXXXXXX sa traversée en courant. Me XXXXXXXXXXXXXXXXXXêtre engagée sur le passage piétons et ne pas avoir vu arriver le véhicule FORD. Le piéton traverse des nø pairs vers les nø impairs et de gauche à droite par rapport au sens de progression du véhicule. Suite au choc, le piéton chute sur la chaussée. Me XXXXXXXXXXXXXXXnt blessée, est transportée par les S.P. à l'hôpital. A.PARE. Non admise.

Personnes concernées

Identification conventionnelle	Concernée à titre de	Age	Entendue	Incapacité prévue	Prise de sang	Résultat
A	Conducteur	53	X			
Y01A	*Piéton*	*64*	*X*	*INC*		

Nature des faits :

Accident corporel de la circulation entre un véhicule particulier et un piéton. Le piéton légèrement blessé. Transporté par les S.P. à l'hôpital. Non admis.

CIRCONSTANCES : D'après les déclarations, A circule boulevard de la République venant de la Place Marcel Sembat et se dirige vers le Pont d'Issy. Arrivé à la hauteur du passage piétons implanté au 74 du dit boulevard, A aperçoit Y engagé sur ce dit passage arrêté au milieu de la chaussée. A s'arrête puis alors qu'il repart Y finit sa traversée en courant et entre en collision avec A au niveau de son avant gauche puis Y chute sur la chaussée. Y traverse des num XXXXXXXXXXXXXXXXX les num XXXXXXXXXXXXXXXXX de gauche à droite par rapport au sens de progression de A. Y dit s'être engagée sur le passage piétons et ne pas avoir vu A arriver.

Fig. numéro I - Un extrait de PV PACTOL

3.2 Les étapes de la construction de la ressource terminologique

L'étape *de pré - traitement* repose sur l'utilisation des résultats fournis par les logiciels Syntex et Cordial après le traitement du corpus. Ces résultats sont utilisés dans la chaîne de traitement par TERMINAE et le module des techniques de fouille de données.

3.3 Le module des techniques de fouilles de données

Le *module des techniques de fouille de données* permet d'identifier des connaissances capables d'enrichir la ressource terminologique créée avec TERMINAE.

Il a été développé en Java et utilise en entrée les résultats fournis par Cordial. Les recherches sont par conséquent effectuées uniquement au niveau de la phrase. Il fait intervenir deux sous modules : la génération des associations ; les affinages.

Le module de *« génération des associations »* permet de définir des patrons et engendre un ensemble de regroupements des mots correspondant aux patrons définis. Les associations sont engendrées automatiquement et correspondent aux deux catégories de patrons. Les deux catégories d'associations obtenues correspondent aux syntagmes nominaux et verbaux fournis par Syntex. Toutefois, les patrons utilisés permettent un affinage des résultats obtenus avec Syntex.

(Nom, Nom ;- ***fait, circonstance*** *)*
(Nom, Préposition; - ***arrivée, sur)***
(Verbe, Nom; - ***voir, personne)***
(Nom, Préposition, Nom; - ***usager, de, route)***
(Verbe, Préposition; - ***venir, de)***
(*Verbe, Préposition, Nom; -* ***circuler, sur, chaussée)***
(Verbe, Préposition, Adjectif; - ***circuler, sur, gauche)***

Fig. numéro II - Exemples de patrons et des instances associées

Chaque regroupement peut représenter : une construction verbale ***(venir de, tourner sur droite)*** ; une construction nominale ***(balise de priorité, priorité du passage)*** ; une relation entre les termes du domaine ***(propriétaire, véhicule ; passager, véhicule)*** ; des associations sans contenu sémantique (bruit) ***(c, véhicule ; venir de 3o6)***. Le nombre des regroupements obtenus est important (environ 44000). A ce stade des affinages sont nécessaires pour permettre l'exploitation de la ressource.

Le module des « *affinages* » réalise deux traitements qui portent sur les syntagmes nominaux et les syntagmes verbaux.

Le **traitement des syntagmes nominaux** a été développé à l'aide de l'algorithme APRIORI qui identifie les règles d'associations. Une règle d'association ***(ex. conducteur, camion)*** représente une implication au sens de la cooccurrence des mots dans un texte. Les seuils imposés aux indicateurs **support et confiance** nous permettent d'éliminer les associations de mots s'avérant peu représentatives (le bruit).

0,119	*0,46*	*origine, de_alerte*
0,119	*0,29*	*heure, de_origine*
0,115	*0,32*	*mn , nature*
0,106	*0,48*	*fonction, service*
0,977	*0,32*	*suite, choc*
0,598	*0,29*	*service, central*
0,594	*0,49*	*central, accident*
0,557	*0,34*	*conducteur, véhicule*
0,552	*0,24*	*hauteur, num*
0,543	*0,28*	*direction, de_rue*

Fig. numéro III - Confiance et support pour les règles d'association

Parmi les résultats des traitements des syntagmes nominaux on trouve des relations de différents types (hiérarchique, synonymie, non taxinomiques) qui vont enrichir la ressource créée avec TERMINAE.

Relation	***Type***
véhicule, fourgonnette ; véhicule, automobile ;	*est-un*
véhicule, propriétaire ;	*non-taxinomique*
volant, véhicule ;	*partie-de*
conducteur, véhicule ;	*non-taxinomique*
conducteur, camion ;	*cas particulier relation précédente*

Fig. numéro IV - Relations conceptuelles découvertes

Le **traitement des syntagmes verbaux** est réalisé sur deux niveaux : l'identification de classes des verbes dans l'ensemble d'associations et l'utilisation de l'ontologie de l'accidentologie afin de regrouper les arguments des verbes.

Une classe de verbes contient l'ensemble des regroupements obtenus à partir d'un verbe. Chaque classe de verbes contient deux catégories de regroupements : celles à deux termes (diriger vers) et celles à trois termes (diriger vers bretelle). Les regroupements à trois termes sont obtenus à partir des regroupements à deux termes en ajoutant une extension. Cette extension correspond à la fonction grammaticale de complément

diriger vers
diriger sur
diriger dans

Fig. numéro V - 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.

diriger vers square,
diriger vers opéra,
diriger vers esplanade,
diriger vers hauteur

Fig. numéro V I - Exemple de regroupements à trois termes

Un regroupement à trois termes est composé d'un verbe, d'une préposition et d'un terme qui relève d'un concept du domaine. Une liste des termes associés à une

extension de niveau 2 (verbe+préposition) est constituée. Une intervention manuelle est nécessaire pour associer la liste à un concept de l'ontologie.

Liste voie *(autoroute, avenue, boulevard, bretelle, carrefour, chemin, esplanade, périphérique, route, voie, etc.)* ***Liste lieu*** *(citadelle, commune, domicile, école, garage, gare, habitation, mairie, manège, opéra, parc, port, porte, square, usine, etc.)*

Fig. numéro VII - Exemple de liste de terme associée à un concept

L'utilisation de l'ontologie réduit ainsi le nombre des regroupements à trois termes. 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.

3.3.1 Utilisation des résultats pour la construction de la ressource avec TERMINAE

Dans la phase d'élaboration de la ressource linguistique, la sélection des termes est orientée par les résultats des règles d'associations. La liste des regroupements de termes obtenus par le module de génération des associations 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. Certaines classes de Terminae sont enrichies en intégrant des motifs découverts.

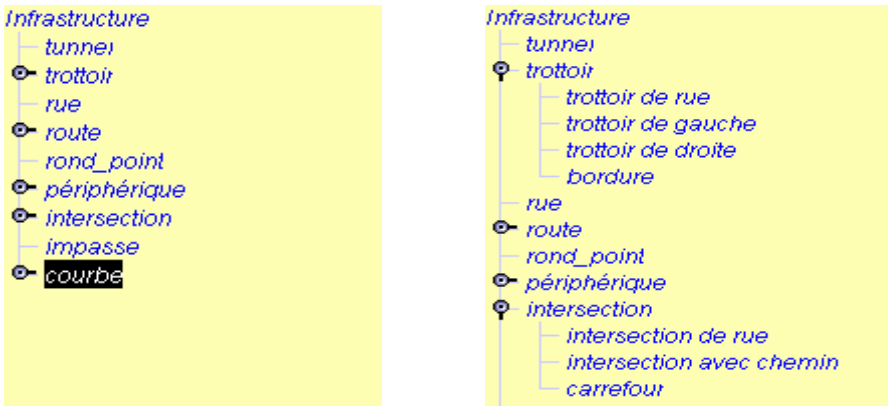

Fig. numéro V III - Enrichissement de la classe « Infrastructure »

Certaines instances identifiées par Syntex contiennent des informations relatives à des contextes particuliers. Grâce au module de pré-traitement, les résultats sont plus génériques, ayant un meilleur niveau conceptuel ***(SE DIRIGER vers la Commune de Wahagnies(résultat Syntex) : diriger_vers_lieu (résultat fouilles)***

Les motifs obtenus à partir du module des transactions sont utiles pour l'utilisation de LINGUAE. Les patrons fournis par les classes de verbe sont plus puissants pour déterminer les instances de relation. Les classes des verbes identifiées par le module d'affinages peuvent être utilisées pour modéliser les relations avec TERMINAE au sein de la ressource terminologique. Syntex identifie les verbes représentatifs du domaine, mais le nombre d'instances correspondant à un seul verbe reste limité. Cette limitation s'avère un facteur capable d'influencer la modélisation des relations entre concepts. Certaines relations ne pourront être mise en évidence à cause du nombre réduit des instances associées à un verbe.

4 Travaux connexes

L'objectif de notre démarche était la construction d'une ressource terminologique à partir de textes en langage libre.

Les travaux menés par Alexander Maedche et son équipe à l'université de Karlsruhe (2000) sont à l'origine de ce travail. Maedche & *al*. propose une solution générale permettant la construction d'ontologies à partir de corpus de domaine : l'extraction des connaissances est automatisée et la modélisation des connaissances découvertes est réalisée par un module semi–automatique. La solution que nous avons adoptée permet à la fois une validation manuelle des connaissances et une découverte automatique. Dans les deux cas, la construction de l'ontologie est un processus semi-automatique et cyclique. Les résultats obtenus par Maedche & *al*. (cf. *supra*) sont plus fins, car les travaux sont menés au niveau de la proposition. La création des associations selon des critères syntaxiques permet d'obtenir des associations plus pertinentes. Les ressources intégrées par l'équipe allemande sont génériques (dictionnaire générique et un dictionnaire du domaine), tandis que nous nous appuyons sur une ontologie de l'accidentologie. L'approche de Maedche & *al*. (cf. *supra*) est plus flexible et permet un choix plus large des ressources et offre plusieurs méthodes de pré-traitement. Elle travaille à un niveau plus fin et utilise des heuristiques propres aux sources utilisées. Notre approche est limitée par les modalités réduites de pré-traitement, les sources fournies par Cordial permettant seulement le traitement au niveau de la phrase. L'intégration de l'ontologie des accidents déjà existante constitue un avantage dans notre cas, l'approche allemande intégrant des ressources génériques. Les deux approches offrent des solutions spécifiques à la tâche et aux ressources utilisées et ont l'avantage de prendre en compte les relations non–taxinomiques identifiées à l'aide du même algorithme. La solution allemande ignore toute approche orientée sur les verbes, alors que nous avons adopté une démarche mixte afin de modéliser de manière complète les relations entre concepts.

Des travaux centrés sur les verbes ont également été étudiés dans la mesure où ils sont centraux dans la modélisation du domaine de l'accidentologie. Faure & Nedellec (1998) ont développé un environnement interactif appelé ASIUM « Acquisition of SemantIc knowledge Using Machine learning method ». ASIUM part de schémas de sous-catégorisation de verbes (structures prédicatives) pour « apprendre » une ontologie à partir de textes analysés syntaxiquement. Les travaux de Wiemer & Hastings, (1998) portent sur l'identification de la signification des verbes inconnus à l'aide du contexte d'occurrence du verbe. Le système CAMILLE utilise WordNet pour intégrer des connaissances et engendre des hypothèses concernant la signification des verbes. Les hypothèses sont formulées selon des critères linguistiques. La solution proposée par Byrd et Ravin (1999) identifie des relations et leur attribue des noms à l'aide de patrons syntaxiques particuliers. Une approche similaire est adoptée par Caméléon (Séguéla, 1999) qui utilise une base de marqueurs spécifiques pour identifier les relations d'hyponymie et de méronymie. Caméléon est aussi capable d'enrichir sa base de marqueurs. Le système Prométhée (Morin, 1999) offre une solution pour la structuration des unités terminologiques. En apprentissage, Prométhée extrait des patrons lexico-syntaxiques caractéristiques d'une relation sémantique. Les patrons ainsi identifiés sont utilisés dans la phase de structuration pour identifier des relations entre unités terminologiques. Dans notre approche, les relations correspondent aux patrons définis de manière générale et sont validées et nommées par un expert. Le système utilise un ensemble de patrons prédéfinis qui ne sera pas enrichi. Les solutions orientées sur les verbes présentent l'inconvénient d'identifier uniquement les relations caractérisées par les verbes. Notre approche offre une solution pour découvrir des relations supplémentaires à l'aide des règles d'associations.

5 Conclusion

TERMINAE permet de construire une ressource terminologique à partir des termes obtenus à la suite d'un traitement linguistique effectué par Syntex. Les techniques de fouilles de données permettent d'identifier des termes génériques et spécifiques du domaine et les relations liant les termes. Des relations non – taxinomiques entre les termes du domaine ont également été identifiées. L'ontologie de l'accidentologie sert d'aide à l'élimination de certains regroupements et à la dénomination des regroupements sélectionnés.

Les résultats obtenus sont complémentaires. La construction de la ressource terminologique est fastidieuse et les regroupements se font manuellement. L'avantage est la méthode qui sous-tend les étapes de l'élaboration de la ressource. Les techniques de fouille fournissent des résultats qui sont directement exploitable pour la structuration des termes dans TERMINAE. Il est guidé par les connaissances déjà acquises, il reste néanmoins des problèmes de seuil.

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

Chafai N., Després, Rapport de Magistère MIAIF, 2003.

Cunningham H., GATE, a General Architecture for Text Engineering. Computing and the Humanities, Vol. 36, pp. 223-254, 2002.

Ceausu V., Elaboration d'une ressource terminologique à partir de PV d'accidents. Rapport de DEA MIASH. Université René Descartes, 2003.

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.

Klein, M., Fensel, D., Harmelen, F., Horrocks, I., (2001) The relations between ontologies and XML schemas. In Computer and Information Science.

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, pages189-202.

Maedche, A., Schnurr, H.-P., Staab, S., Studer, R. Representation language –neutral modeling of ontologies (2000) . In U. Frank, editor, *Proceedings of the German Workshop "Modellierung-2000". Koblenz, Germany*.

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.

Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., and Cunningham, S.J. (1999) "Weka: Practical machine learning tools and techniques with Java implementations" *Proc ICONIP/ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences*, 192-196, Dunedin, New Zealand, November.