Extraction d'informations à partir de corpus dégradés

Fabrice Even[1], Chantal Enguehard

Institut de Recherche en Informatique de Nantes
Université de Nantes
Faculté des Sciences et des Techniques
2 rue de la Houssinière, BP 92208
44322 NANTES cedex 3, France
{even,enguehard}@irin.univ-nantes.fr

Résumé – Abstract

Nous présentons une méthode automatique d'extraction d'information à partir d'un corpus mono-domaine de mauvaise qualité, sur lequel il est impossible d'appliquer les méthodes classiques de traitement de la langue naturelle. Cette approche se fonde sur la construction d'une ontologie semi-formelle (modélisant les informations contenues dans le corpus et les relations entre elles). Notre méthode se déroule en trois phases : 1) la normalisation du corpus, 2) la construction de l'ontologie, et 3) sa formalisation sous la forme d'une grammaire. L'extraction d'information à proprement parler exploite un étiquetage utilisant les règles définies par la grammaire. Nous illustrons notre démarche d'une application sur un corpus bancaire.

We present an information extraction automatic method from poor quality specific-domain corpus (with which it is impossible to apply classical natural language methods). This approach is based on building a semi-formal ontology in order to modelise information present in the corpus and their relation. Our method happens in three stage : 1) corpus normalisation, 2) ontology building and 3) model formalisation in grammar. The information extraction itself is made by a tagging process using grammar rules. We illustrate our approach by an application working on a bank corpus.

Mots Clés – Keywords

Extraction d'information, modélisation, construction d'ontologie, corpus dégradés.
Information extraction, modelling, building ontology, poor quality corpus..

---

[1] Thèse financée par le Crédit Mutuel Loire-Atlantique Centre-Ouest dans le cadre d'un contrat CIFRE

1 Introduction

Cette recherche est directement issue de la volonté d'extraire des informations à partir d'un corpus dégradé (nombreuses abréviations, phrases asyntaxiques, ponctuation inexistante, etc.). L'extraction d'information est définie comme un processus en deux étapes : la modélisation des types d'informations recherchés et l'identification de leurs instances dans le corpus. La première étape utilise des sources de connaissances externes au corpus afin de bâtir une ontologie semi-formelle recouvrant une partie du domaine dans lequel s'inscrit le corpus. Seul le sous-domaine limité à la connaissance exprimée dans le corpus est effectivement modélisé. La deuxième étape s'appuie sur cette ontologie afin d'extraire les informations du corpus.

Après une brève présentation des méthodes de construction d'ontologie et leur confrontation à des corpus dégradés, nous détaillons les différentes étapes de notre démarche : la construction de l'ontologie, sa représentation formelle sous forme d'une grammaire, puis l'extraction des informations à l'aide de cette grammaire. Les résultats de l'extraction sont évalués et analysés.

Nous travaillons sur un corpus composé de textes issus du milieu bancaire. Il est composé de textes relatant des entretiens passés entre des clients et des employés. Notre but est d'extraire automatiquement certains types d'informations décrites informellement, comme les différents projets des clients ou un changement dans leur situation familiale. Les informations extraites viendront enrichir une base de données.

> Exemple : à partir du texte suivant,
>
> *« PRJ ACHT STUDIO EN 12 01 PARIS POUR 400 KFRCS »* [1]
>
> le processus doit repérer les projets d'achat du client et identifier ses paramètres : l'objet (studio), la date (12/01), le lieu (Paris) et le montant (400KF).

2 Méthodes de construction d'ontologie

De nombreuses méthodes visent à construire des ontologies à partir de corpus. La plupart se fondent sur le contenu des textes pour construire l'ontologie, les textes sont alors la source principale de connaissance pour l'acquisition d'information (Nobécourt, 2000). L'ensemble des concepts du modèle ainsi que leurs relations sont exclusivement issus d'une analyse des textes, sans apport de connaissance extérieur. (Aussenac-Gilles et al., 2000) s'inscrivent dans cette démarche même s'il est reconnu que les textes peuvent ne pas constituer l'unique source de connaissance. Cette approche endogène se décompose en plusieurs étapes : l'extraction des termes se référant aux concepts de base (primitives conceptuelles (Nobécourt, 2000)), puis des relations lexicales qu'ils entretiennent (création d'une base terminologique (Aussenac-Gilles et al., 2000), (Lame, 2000)) afin de faire émerger les premières relations inter-concepts. L'étape suivante s'appuie sur l'analyse des relations sémantiques entre termes pour extraire de nouvelles relations entre les concepts ainsi que de nouveaux concepts afin d'aboutir à un réseau sémantique de concepts. Le réseau sémantique doit être validé par un expert du domaine afin de préciser les relations significatives (normalisation (Bouaud, Bachimont et al 1995)). Le résultat correspond à la définition de (Swartout et al., 1996) « Une ontologie est

une structure hiérarchique d'un ensemble de termes du domaine », soit une ontologie pouvant être représentée de manière formelle ou semi-formelle.

Les différentes étapes s'appuient sur des outils de TAL peu adaptés aux corpus de mauvaise qualité. La phase terminologique est envisageable après une correction des textes, mais l'extraction de relations lexicales et sémantiques se révèle peu efficace sur de tels corpus à cause de leur très faible qualité syntaxique et lexicale.

3 Modélisation du domaine

3.1 Prétraitements

Les textes de notre corpus sont si dégradés (incorrections typologiques ou orthographiques, emploi d'abréviations non normalisées) qu'il est nécessaire d'effectuer avant tout autre processus une correction et une normalisation du corpus. C'est à partir de ce corpus corrigé que nous procédons à la modélisation puis à l'extraction d'information à proprement parler. Cette normalisation porte sur le format des valeurs (nombres avec unités), les dates, les abréviations et la correction orthographique des fautes lexicales et typographiques. Ces prétraitements sont réalisés à l'aide d'expressions régulières écrites en fonction du corpus.

Exemple : l'extrait [1] devient : *« projet achat studio en 12/01 paris pour 400kf »* [2]

3.2 Construction du modèle

D'après Bachimont (Bachimont 2001), il n'existe pas de concepts indépendants du contexte ou du problème traité permettant de construire toute la connaissance d'un domaine. Une ontologie fonctionne comme un cadre théorique du domaine construit en fonction du problème traité. Le processus de modélisation décrit ici est fondé sur cette dernière définition. Nous construisons l'ontologie en nous appuyant sur les connaissances présentes dans le corpus et des connaissances externes au corpus (experts).

3.2.1 Définition de l'ontologie initiale : expression des informations recherchées

Les informations à rechercher, exprimées informellement, sont réécrites sous la forme de prédicats modélisés par des patrons d'informations. Cette tache doit être réalisée avec des experts du domaine ayant une bonne connaissance du corpus et sachant exprimer la nature des informations à extraire. Cette phase aboutit à la description d'un ensemble de hiérarchies de concepts constituant une première sous-ontologie (ontologie initiale). Cette ontologie met en jeu des relations argumentatives entre concepts (un concept est lié à un autre car il en est un argument). Les relations argumentatives respectent l'Attribute Consistency Postulate (Guarino 1992) : dans chaque relation argumentative, toute valeur d'un argument est une instance du concept correspondant à cet argument.

3.2.2 Définition de la terminologie

Elle est issue d'une part d'une étude terminologique des textes par le logiciel, ANA (Enguehard, Pantéra 1995), caractérisé par sa robustesse. Et d'autre part d'un ensemble de documents référençant la terminologie spécifique au domaine dans lequel nous trouvons une partie des termes potentiellement utilisés dans notre corpus.

3.2.3 Normalisation : Fusion de l'ontologie initiale et de la terminologie

A chaque terme ne correspond pas un concept de l'ontologie initiale. Il faut pouvoir relier les termes du corpus aux concepts définis dans cette ontologie. Un processus de normalisation est nécessaire. Il se déroule en trois étapes : 1) extension de l'ontologie initiale, 2) définition d'une semi-base terminologique et 3) unification des modèles ainsi définis.

1. L'ontologie initiale est révisée avec l'aide d'experts du domaine. De nouveaux concepts complètent les hiérarchies (on parlera d'ontologie initiale étendue).

2. Grâce aux mêmes experts, et en nous appuyant sur des documents propres au domaine, nous spécifions à partir de la terminologie un autre ensemble de concepts, les concepts de base. Récursivement de nouveaux concepts sont définis par héritage à partir des concepts de base. Le résultat est un ensemble de petites hiérarchies ayant chacune un ancêtre unique dont les derniers descendants sont des concepts de base. Ces hiérarchies sont normées, c'est-à-dire organisées de manière systématique, car chaque père se décompose en fils selon un critère unique. Elles respectent également le critère de rigidité de Guarino (Guarino 2000).

3. Les deux processus précédents donnent d'une part une ontologie liée au problème traité et un ensemble de sous-hiérarchies directement liées à la terminologie du texte. Nous procédons à l'unification des sous-hiérarchies et de l'ontologie initiale étendue en analysant si des concepts ancêtres (pères) des sous-hiérachies sont déjà présents dans l'ontologie ou si une relation peut être définie avec des concepts de cette ontologie.

Nous obtenons une modélisation du domaine couvrant l'ensemble des concepts qui nous intéressent pour la recherche d'information. Cette modélisation est décrite par un schéma relationnel formalisé par un ensemble de graphes orientés. Elle établit une ontologie décrite de manière semi-formelle car indépendante d'un langage de représentation (Barry et al. 2001).

3.3 Représentation formelle

Nous représentons la modélisation précédemment obtenue sous la forme d'une grammaire, afin de la rendre utilisable. Comme indiqué en 3.2, l'ontologie met en jeu deux types de relations entre concepts : des relations hiérarchiques et des relations "argumentatives". Dans cette formalisation les relations hiérarchiques sont appelées relations constitutives car nous nous plaçons ici du point de vue de la base de chacune des hiérarchies et non pas de celui de son sommet. Lorsqu'une relation hiérarchique existe entre un concept père A et un concept fils B, on dit que B constitue A (on part du fils pour remonter jusqu'au père puis au père du

père et ainsi de suite). Une telle relation peut s'apparenter à une relation d'hyperonymie (A est un hyperonyme de B) ou à l'inverse d'hyponymie (B est un hyponyme de A). Lorsque l'on a une relation de type argumentative entre deux concepts C et D, le concept D est un argument (dont le type dépend de la relation) du concept C. La grammaire doit donc pouvoir rendre compte de ces deux types de relations. Aussi nous définissons pour celle-ci deux types de règles : les règles constitutives et les règles prédicatives.

3.3.1 Règles constitutives

Un concept A est défini par un ensemble de règles Def(a) impliquant des termes ou des concepts. Ces règles sont dites constitutives car le concept A est défini (constitué) par ces règles. Quel que soit X appartenant à Def(A), X ne peut définir un autre concept que A. Ces règles s'écrivent A ::= Def(A) et sont de trois types : les règles sélectives, les règles conjonctives et les règles disjonctives.

Les règles sélectives sont du type A ::= B1 | B2 | … . L'opérateur « | » est équivalant au OU exclusif : la valeur de l'expression B1 | B2 est soit B1, soit B2 mais pas les deux.

Exemple : *<VEHICULE> ::= <AUTO> | <MOTO> | <DIV_VEHI>*

Les règles conjonctives sont du type A ::= B1 + B2 + … . L'opérateur « + » est non commutatif et correspond à la conjonction classique : la valeur de l'expression B1 + B2 est exclusivement "B1 ET B2".

Exemple : *<P_AUTO> ::= <DC_PRET> + <VEHICULE>*

Les règles disjonctives sont du type A ::= B1 v B2 v … . L'opérateur « v » équivaut à la disjonction classique : la valeur de l'expression B1 v B2 est soit B1, soit B2, soit les deux.

Exemple : *<PERSONNE> ::= <NOM> v <PRENOM>*

3.3.2 Règles prédicatives

Ces règles décrivent les concepts-prédicats (appelés aussi prédicats), c'est-à-dire les concepts mettant en jeu une ou plusieurs relations argumentatives. Elles définissent un prédicat P par un descripteur et un objet. Le descripteur est un concept unique (un concept ne pouvant être le descripteur de plus d'une règle). L'objet est un concept pris parmi un certain nombre de concepts possibles, ceux ci étant déduits de la modélisation.

Ces règles peuvent également s'accompagner d'un ou plusieurs arguments supplémentaires dits options. Ceux ci lorsqu'ils sont valués donnent plus d'informations sur le prédicat P mais ne sont ni nécessaires, ni suffisants pour le définir.

Les règles prédicatives s'expriment de la forme P ::= (descripteur = D ; objet = O1 | O2 | O3 … ; option 1 = A1 | A2 | A3 … ; option 2 = B1 | B2 … ; ….).

Exemple : *<ACHAT> ::= (*
descripteur = <DC_ACHAT> ;
objet = <IMMOBILIER> | <VEHICULE> | <PROD_BANCAIRE> ;

date = <DATE> ;
montant = <SOMME>)

4 Moteur d'extraction

Le moteur d'extraction s'appuie sur la grammaire modélisant le domaine (la grammaire issue de l'ontologie). Il comprend quatre phases : l'alimentation d'une base de règle, deux étiquetages successifs et le recueil des informations alimentant la base de données.

La base de règles se compose de deux sous-ensembles de règles (les règles constitutives et les règles prédicatives) déduites de la grammaire. L'étiquetage constitutif du corpus s'appuie sur les règles constitutives de la base (chaque terme est étiqueté par le concept qui lui est associé, et chaque concept est lui-même étiqueté en fonction de sa description dans la grammaire). Le second étiquetage s'appuie sur les règles prédicatives pour instancier les prédicats. Après ce double étiquetage, le processus de recueil d'information est directement opérationnel. Son résultat alimente la base de données.

4.1 Construction de la base de règles

Les règles de la base sont déduites de la grammaire, directement pour les règles prédicatives, après transformation pour les règles constitutives.

- les règles sélectives permettant d'identifier les concepts de base à partir des termes sont transformées en ensemble de règles simples du type A ::= terme (règles sélectives terminales).

- les règles sélectives sur les concepts sont transformées en un ensemble de règles simples A ::= B. Les règles conjonctives et disjonctives sont adaptées afin de se conformer à la syntaxe du corpus. Cette adaptation consiste à adjoindre à ces règles une notion d'ordre (l'opérateur + n'étant pas commutatif) et de proximité. L'ensemble de ces règles transformées forme l'ensemble des règles conceptuelles.

4.2 Etiquetage constitutif

L'étiquetage constitutif repère dans le texte les différents termes puis les concepts en appliquant récursivement les règles constitutives. A chaque fois qu'un concept est repéré, il est marqué par une balise. Certains concepts très spécifiques (dont la syntaxe répond à un formalisme connu) comme les dates, les montants ou les sommes sont traités au préalable. Ensuite l'étiquetage se déroule en deux étapes : 1) l'étiquetage des termes et 2) la propagation des concepts.

1. Les règles sélectives terminales sont appliquées sur le corpus. Lorsque toutes les règles sont appliquées, tous les termes reconnus par la grammaire sont étiquetés par des concepts (exemple 3).

2. La propagation des concepts permet de détecter de nouveaux concepts. La règle A ::= B permet de repérer le concept A en rajoutant les balises de A à celles

identifiant B. Les règles conceptuelles sont appliquées sur le corpus autant de fois qu'il est possible de le faire. Le processus s'arrête lorsque plus aucune d'entre elles n'est applicable. A ce point, le corpus est entièrement étiqueté par les règles constitutives (exemple 4).

L'extrait [2] devient après l'étiquetage des termes :

<DC_PROJET>projet</DC_PROJET>
<DC_ACHAT>achat</DC_ACHAT>
d'
<APPARTEMENT>studio</APPARTEMENT>
en
<DATE>12/01</DATE>
<VILLE>paris</VILLE>
pour
<SOMME>400kf</SOMME> [3]

Exemple 3

L'extrait [3] devient après propagation des concepts :

<DC_PROJET>projet</DC_PROJET>
<DC_ACHAT>achat</DC_ACHAT>
d'
<IMMOBILIER><APPARTEMENT>studio</APPARTEMENT></IMMOBILIER>
en
<DATE>12/01</DATE>
<LIEU><VILLE>paris</VILLE></LIEU>
pour
<SOMME>400kf</SOMME> [4]

en appliquant les règles : <IMMOBILIER> ::= <APPARTEMENT>
<LIEU> ::= <VILLE> | <PAYS> | <REGION>

Exemple 4

4.3 Etiquetage prédicatif

Les règles prédicatives repèrent les instances des prédicats décrits dans la grammaire. Elles sont appliquées sur le corpus en recherchant pour chaque descripteur de concept trouvé, un des concepts objets possibles du prédicat mis en jeu. On cherche à instancier les prédicats jusqu'à ce qu'il soit impossible de le faire en procédant de la façon suivante. Le texte est parcouru de gauche à droite. Lorsqu'un descripteur de prédicat est reconnu, on recherche un argument objet (concept ou prédicat) valable pour ce prédicat avant le prochain descripteur de prédicat non traité. Si c'est le cas l'objet est valué, puis le système cherche à valuer les éventuels autres arguments et passe au prochain descripteur du texte. Sinon ce descripteur est laissé de coté et le système passe au suivant. Cela est fait jusqu'à l'extrémité du texte. S'il reste des descripteurs non traités, c'est à dire décrivant des prédicats dont l'objet n'a pu être valué, le processus est répété à partir du début du texte. Cette opération est réitérée jusqu'à ce qu'il ne reste plus de descripteurs à traiter ou que ceux qui restent ne peuvent plus l'être. Dans ce dernier cas ils seront marqués comme décrivant des instances de prédicats vides (sans objet). Un prédicat peut être objet d'un autre prédicat. Dans ce cas les arguments du dernier prédicat sont valués lorsque cela est possible par ceux du premier prédicat.

L'extrait [4] devient :

<PROJET_1><DC_PROJET>projet</DC_PROJET></PROJET_1>
<PROJET_1 ARG=OBJET>
<ACHAT_1><DC_ACHAT>achat</DC_ACHAT></ACHAT_1>
</PROJET_1 ARG=OBJET>
d'
<ACHAT_1 ARG=OBJET>
<IMMOBILIER><APPARTEMENT>studio</APPARTEMENT></IMMOBILIER>
</ACHAT_1 ARG=OBJ>
en
<PROJET_1 ARG=DATE>
<ACHAT_1 ARG=DATE><DATE>12/01</DATE></ACHAT_1 ARG=DATE>
</PROJET_1 ARG=DATE>
<PROJET_1 ARG=LOCALISATION>
<ACHAT_1 ARG=LOCALISATION>
<LIEU><VILLE>paris</VILLE></LIEU>
</ACHAT_1 ARG=LOCALISATION>
</PROJET_1 ARG=LOCALISATION>
pour
<PROJET_1 ARG=MONTANT>
<ACHAT_1 ARG=MONTANT><SOMME>400kf</SOMME></ACHAT_1 ARG=MONTANT>
<PROJET_1 ARG=MONTANT>

D'où les instances suivantes
des prédicats ACHAT et PROJET :

<ACHAT 1> [
DESCRIPTEUR=achat
OBJET=studio
DATE=12/01
LOCALISATION=paris
MONTANT=400kf
]

<PROJET 1> [
DESCRIPTEUR=projet
OBJET=<ACHAT 1>
DATE=12/01
LOCALISATION=paris
MONTANT=400kf
]

Exemple 5

Les instanciations de prédicats se traduisent par un balisage du texte (exemple 5). Le descripteur est étiqueté par la référence du prédicat, c'est à dire son nom et un numéro d'instance (pour les cas ou un même prédicat peut être instancié plusieurs fois dans le texte). Chaque concept argument d'un prédicat est balisé par la référence du prédicat dont il est argument et par son type (Objet, Date, …).

4.4 Recueil des informations

Les concepts ainsi que leurs relations apparaissent clairement dans le corpus grâce aux balises. Lors de la phase d'extraction, il suffit de spécifier les concepts à rechercher. Les balises permettent de localiser ces concepts ainsi que leurs différents arguments. Ces informations nourrissent une base de données dont les tables correspondent aux prédicats de la grammaire.

5 Résultats

Le corpus est constitué d'un million d'enregistrements. Chaque enregistrement est issu d'un entretien d'un employé d'une agence de la banque avec un client. Il se présente sous la forme d'une ligne constituée d'un entête numérique et d'un champ texte. L'entête contient le numéro d'identifiant et la date de l'enregistrement. Le champ texte contient le texte saisi par le banquier rendant compte de l'entretien. Le nombre de mots de ce champ varie d'un enregistrement à l'autre : de quelques mots à plus d'une trentaine. Avant toute analyse, ce champ est soumis à un traitement pour le mettre en conformité avec les règles de la CNIL. L'extraction terminologique avec ANA (Enguehard, Pantéra 1995) définit 15000 candidats-termes. Après écrémage (rassemblement des candidats-termes et élimination des parasites), il reste 1300 termes. Sur les 350 termes qu'ils contiennent, les documents terminologiques ont fourni 200 nouveaux termes supplémentaires.

5.1 Méthode d'évaluation

Le but de la recherche est d'extraire automatiquement les évènements concernant les clients, c'est-à-dire leurs projets et les refus de proposition (de leur part ou de celle de la banque). Le résultat est un ensemble d'instances du prédicat recherché. Comme ces prédicats ont plusieurs arguments, nous définissons 3 degrés de validité en fonction de la manière dont sont valués ces arguments :

1. Une instance d'un prédicat est dite valide si les arguments valués le sont par les bonnes valeurs. Le taux de validité est le nombre d'instances valides par rapport au nombre d'instances trouvées ;

2. Une instance valide est dite totalement valide si tous ses arguments sont valués, et partiellement valide si au moins un de ces arguments n'est pas valué ;

3. Une instance partiellement valide est dite incomplète lorsqu'au moins un argument n'est pas valué à cause d'un manquement du processus, et complète lorsque la totalité des arguments non valués est due à l'absence des informations correspondantes dans le texte.

5.2 Expérimentation

Notre première expérimentation porte sur un échantillon de 4000 enregistrements pris au hasard dans le corpus. Il s'agit d'extraire les projets présents dans cet échantillon. Les résultats trouvés ont été validés manuellement par les experts qui ont aligné l'échantillon avec les instances de la table PROJET créées par notre application. D'après les experts, 265 projets sont présents dans cet échantillon dans lequel 253 instances sont détectées. Nous obtenons les résultats décrits par la figure 1.

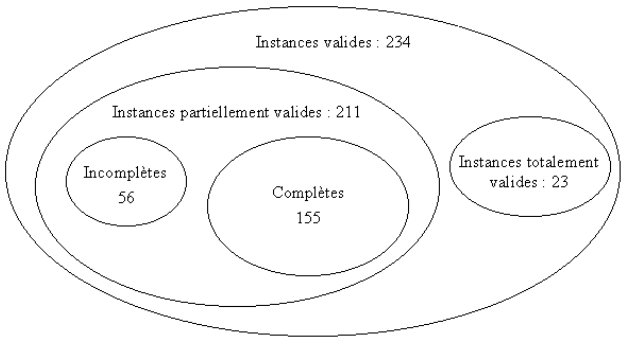

Figure 1 : Résultats

5.3 Analyse des résultats

Le taux de couverture n'est pas significatif car de nombreux enregistrements ne contiennent pas de projets (95 %). Les taux de rappel (correspondant au nombre d'instances trouvées par rapport à celle présentes et égal à 95,4 %) et de validité (92,5 %) sont très satisfaisants mais nombre d'instances ne sont que partiellement valides (90 % des projets valides). 73,4 % de celles-ci le sont en raison de l'incomplétude du corpus qui ne peut nous délivrer toutes les informations mais le reste est du au processus. Une analyse des enregistrements dans lesquels ces cas sont observés fait apparaître deux principaux problèmes :

- un très grand nombre de dates sont présentes sous une forme littérale de manière plus ou moins floue (au prochain semestre, dans quelques mois) et ne sont pas détectées.

- des arguments d'un prédicat décrit par un descripteur X peuvent se trouver avant lui dans le texte ou en dehors de la fenêtre de recherche et ne sont pas pris en compte.

Nous travaillons actuellement au règlement de ces deux problèmes. Pour le 2ème nous proposons de modifier le processus d'étiquetage prédicatif pour détecter les arguments présents à gauche d'un descripteur de prédicat. Quant aux dates, nous introduisons la notion de date floue associée à un degré.

6 Conclusion et Perspectives

Nous avons décrit une méthode permettant d'extraire des informations à partir de textes de qualité dégradée, sur lesquels sont inopérants les processus d'extraction d'information

existants. Cette approche s'appuie sur la construction d'une ontologie, guidée par la nature des informations à rechercher dans les textes. Nous obtenons avec cette approche de très bons résultats avec une couverture très importante des informations présentes dans chaque enregistrement. De plus même quand celle-ci est minimale, le système permet de repérer les enregistrements dans lesquels est présente une information. Cette méthode peut s'appliquer à d'autres corpus d'autres domaines. La modélisation du domaine et les prétraitements sont liés aux textes, mais le reste du processus est générique et ne nécessite pas de modification du système. Une expérimentation en ce sens est en cours sur des corpus concernant un domaine proche de celui de notre corpus actuel (textes de nature similaire d'une autre banque).

Références

Aussenac-Gilles N., Biébow B., Szulman S. (2000), Corpus analysis for conceptual modelling, Proceeding of *EKAW'2000,* 13-20.

Aussenac-Gilles N., Bourigault D., Condamines A., Gross C. (1995), How can knowledge acquisition benefit from terminology ?, Proceedings of *EKAW'95*.

Bachimont B. (2001), Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle, Actes d'*IC 2001*, 349-368

Barry C., Cormier C., Kassel G., Nobécourt J. (2001), Evaluation de langages opérationnels de représentation d'ontologies, Actes d'*IC'2001*, 309-327.

Biébow B., Szulman S. (1998), Une approche terminologique pour catégoriser les concepts d'une ontologie, Actes d'*IC'98*, 51-58.

Bouaud J., Bachimont B., Charlet J., Zweigenbaum P. (1995), Methodological Principles for Structuring an Ontology, Proceedings of *IJCAI-95*

Enguehard C., Pantéra L. (1995), Automatic Natural Acquisition of a Terminology, *Journal of quantitative linguistics*, Vol. 2, n°1, pp.27-32.

Guarino N., Welty W. (2000), A Formal Ontology of Properties, Proceedings of the *ICAI-00 Workshop on Applications of Ontologies and Problem-Solving Methods*, 12/1-12/8.

Guarino N. (1992), Concepts, Attributes and Arbitrary Relations : Some Linguistic and Ontological Criteria for Structuring Knowledge Bases, *Data & Knowledge Bases Engineering*, 8(2) : 249-261.

Lame G. (2000), Knowledge acquisition from texts towards an ontology of French law, Proceedings of *EKAW'2000*, 53-62.

Nobécourt J. (2000), A method to build formal ontologies from texts, Proceedings of *EKAW'2000*, 21-27.

Nestorov S. & al. (1997), Representative objects : concise representation of semistructured, hierarchical data., Proceedings of *International Conference on Data Engineering*, 79-90.

Riloff E. (1996), Automatically Generating Extraction Patterns from Untagged Text, Proceedings of *Thirteenth National Conference on Artificial Intelligence*, 1044-1049.

Soderland S. (1997), Learning Text Analysis Rules for Domain-specific Natural Language Processing, Ph.D. thesis, University of Massachusetts, Amherst.