Extraction de connaissances à partir de textes structurés

Yannick Toussaint

LORIA - INRIA
BP 239
F-54506 Vandœuvre-lès-Nancy cedex
Yannick.Toussaint@loria.fr

RÉSUMÉ. *Cet article propose un schéma général d'*extraction de connaissances à partir de textes *et situe la* fouille de textes *comme une étape particulière d'un processus complexe. Notre position est que tout processus de fouille de textes doit nécessairement exploiter un modèle de connaissances et qu'il est essentiel d'extraire des textes des informations structurées auxquelles peut être associée une sémantique. De ce fait, nous nous intéressons tout particulièrement à la structure des textes, structure devant être prise dans un sens très général qui va d'une structuration physique (hiérarchique) à une structuration cognitive ou sémantique. Nous montrons comment ces différentes dimensions du document et du texte peuvent ou pourraient être prises en compte pour que le processus dans son ensemble soit incrémental, c'est-à-dire qu'il soit initialisé avec un ensemble de connaissances réduit qui augmente au fur et à mesure des boucles de traitement.*

ABSTRACT. *This paper proposes a global schema for* Knowledge Discovery in Texts *and presents Text Mining as a specific step of the overall process. We argue that any text mining process should rely on a knowledge model, and that it is crucial for the information extracted to be structured and semantically described. We investigate the different document dimensions and show how they contribute or could contribute to the process. Our goal is to define a process which is able to incrementally build upon a small of knowledge, augmenting it little by little at each processing loop.*

MOTS-CLÉS : *extraction de connaissances à partir de textes, fouille de textes, structure du document, traitement automatique de la langue, extraction d'information, modèle de connaissances.*

KEYWORDS: *knowledge extraction from texts, text mining, document structure, natural language processing, information extraction, knowledge model.*

DN – 8/2004. Fouille de textes, pages 11 à 34

1. Introduction

Ces dernières années, le terme de *fouille de textes* est apparu dans bon nombre de publications. Pourtant, la diversité des méthodes et des niveaux de traitement laisse penser qu'il n'y a pas véritablement de consensus sur la notion de fouille de textes. Nous proposons donc en premier lieu une définition de la fouille de textes en tant qu'étape particulière d'un processus plus général d'extraction de connaissances à partir de textes. Nous ne souhaitons pas prendre une position normative ou exclusive par rapport à tous ces travaux en fouille de textes. Bien au contraire, nous montrons que ces différents travaux contribuent à des niveaux différents au passage de grands volumes de textes à de la connaissance. Nous voulons notamment insister sur ce qui permet à ces travaux de contribuer à un objectif commun, faire en sorte qu'il y ait une trame commune pour extraire des textes une information plus précise, associée à une sémantique rigoureuse afin d'aider un expert à enrichir son modèle de connaissances ou à effectuer tout autre tâche de raisonnement comme la veille technologique.

L'article est structuré de la façon suivante. Nous proposons d'abord un bref historique et une définition de la fouille de textes. Nous soulignons les spécificités de la fouille de textes par rapport à la fouille de données et abordons de ce fait la notion de structure de documents. La section 4 propose un schéma général de l'extraction de connaissances à partir de textes. La section 6 présente comment les différents niveaux d'analyse des textes permettent de contribuer à cette tâche. Enfin, nous présentons quelques méthodes de fouille de textes en insistant sur le fait que le processus de fouille de textes doit nécessairement exploiter un modèle de connaissances.

2. Définition de la fouille de textes

La fouille de textes, traduit du terme anglais *text mining*, est apparue dans la deuxième moitié des années 90, en écho à des travaux réalisés depuis les années 80 sur des bases de données. En 1991, Piatetsky-Shapiro introduit comme titre de son ouvrage [PIA 91] le terme de *Knowledge Discovery from Databases*, abrégé par la suite en KDD et, dont l'équivalent français est *Extraction de Connaissances à partir de Bases de Données* (ECBD). Ce n'est que vers 1995 que l'usage des termes *Knowledge Discovery from Databases* et *Data Mining* se précise. L'extraction de connaissances à partir de bases de données désigne alors le processus global de découverte de connaissances qui permet de passer de données brutes à des connaissances alors que la fouille de données n'est qu'une étape de l'ECBD au cours de laquelle un modèle est construit.

Initialement proposée par [PIA 91], la définition de l'ECBD a été enrichie au cours du temps et celle proposée dans [FAY 96b] est maintenant consensuelle :

Définition 1 (L'extraction de connaissances) *L'extraction de connaissances à partir de bases de données est un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données.*

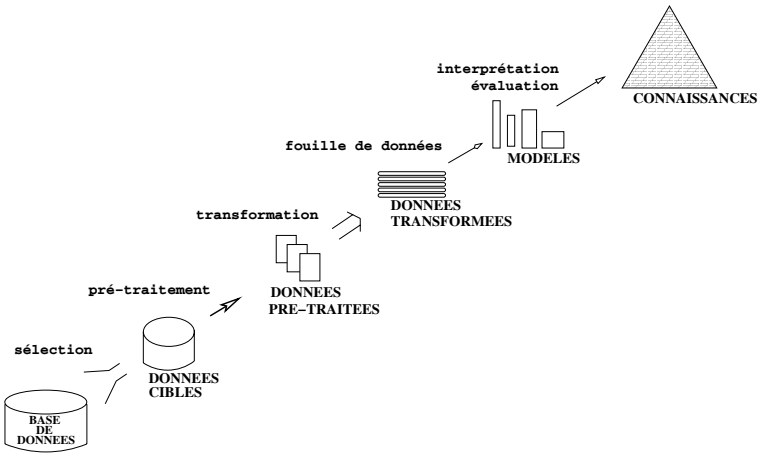

Figure 1. *Schéma global de l'ECBD d'après [FAY 96b]*

Comme l'explique ce dernier auteur, l'ECBD peut se décomposer en de nombreuses étapes plus ou moins complexes mais la figure 1 en donne une vision synthétique. Parmi les grandes étapes de l'ECBD, on peut distinguer :

– la **sélection** qui crée un ensemble de données à étudier ;

– le **prétraitement** qui vise à enlever le bruit et à définir une stratégie pour traiter les données manquantes ;

– la **transformation** où l'on recherche les meilleures structures pour représenter les données en fonction de la tâche ;

– la **fouille de données** : la fouille proprement dite et la définition de la tâche : classification, recherche de modèles... et la définition des paramètres appropriés ;

– l'interprétation et l'évaluation pendant laquelle les patrons extraits sont analysés. La connaissance qui en est ainsi extraite est alors stockée dans la base de connaissances.

La définition de l'ECBD montre qu'elle entretient des liens forts avec l'apprentissage. La différence provient de quatre points qui sont commentés dans [COR 02] dont les trois premiers nous intéressent ici très directement :

1) l'ECBD prend les données dans leur état brut et inclut donc un processus de nettoyage ;

2) L'ECBD fournit des outils capables de travailler sur des données mixtes, numériques et symboliques ;

3) l'ECBD construit des outils produisant une connaissance intelligible aux utilisateurs ;

4) l'ECBD utilise généralement des données qui ont été stockées dans une base de données (souvent relationnelle).

Les travaux sur les textes bénéficient donc de ces années d'expérience sur les bases de données. En 1995, Feldman [FEL 95] introduit le terme de *KDT : Knowledge Discovery in Texts* ou encore celui d'*association entre termes* [FEL 96]. D'autres auteurs utilisent également la notion de *tendance* (*trends in texts*, [LEN 97]).

Nous calquons donc notre définition de l'extraction de connaissances à partir de textes sur celle de l'extraction de connaissances à partir de bases de données :

Définition 2 (L'extraction de connaissances à partir de textes) *L'extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts.*

Par cette définition, nous adoptons également l'idée maîtresse du schéma global de la figure 1 dans lequel l'objectif de la fouille de textes est de construire un modèle du domaine ou base de connaissances. Nous nous proposons de préciser ce schéma pour y faire apparaître la dimension interactive mais également itérative du processus ainsi que pour introduire les différents travaux impliqués dans la fouille de textes, travaux à l'intersection de plusieurs domaines dont notamment celui de l'apprentissage, mais également celui du traitement automatique de la langue (TAL) ou encore, de la recherche d'information et de l'extraction d'information.

Enfin, puisque nous définissons un processus interactif, nous définissons l'utilisateur d'un système de fouille de textes comme étant un expert du domaine de spécialité et qui a besoin de répertorier, d'accéder ou de raisonner sur la connaissance pour une application donnée. Il est donc capable de valider ou d'invalider une information et de formuler des préférences quant à la structure finale des connaissances.

2.1. *Fouille de textes et fouille de données*

Acquérir des connaissances à partir de données ou à partir de textes, si l'on s'en tient à notre discussion précédente paraissent deux objectifs très proches qui exploitent des méthodes et outils, notamment au niveau de l'étape de fouille, eux aussi très semblables. Il existe pourtant des différences significatives entre ces deux processus. Ainsi, dans le cadre du traitement des données, le point 4 précédent introduit par [COR 02] dissimule une étape essentielle de modélisation des données brutes en une base de données. Cette modélisation, lorsqu'il s'agit de textes est beaucoup plus complexe, du moins si l'on souhaite avoir des résultats d'une qualité comparable aux travaux sur les données. Les textes font référence à des concepts, à des relations entre ces concepts et à des événements qu'il est parfois difficile d'identifier.

2.1.1. *La manipulation de concepts*

Une des caractéristiques premières des textes, notamment dans des domaines scientifiques ou techniques est qu'ils font très largement appel à des concepts. Le mot, unité sémantiquement trop pauvre, a été supplanté par la notion de termes que l'on peut associer à un concept dans une ontologie. Le concept appartient ainsi à une structure de connaissances où sont exprimées une partie de ses propriétés. Un certain nombre de ces propriétés sont reflétées dans les textes. Ainsi, il est fréquent de trouver dans les textes l'expression ou l'usage de relations méronimiques (*La mémoire du C840...*), synonymiques ou taxinomiques. Plus la dimension des données est importante, plus il sera difficile d'analyser une classification. Or compte tenu des variations dans la langue pour désigner un concept, il est nécessaire de regrouper des formulations différentes d'un même concepts (*taux de chômeurs / taux de chômage*) ou des concepts différents en un concept plus générique pour éviter une trop grande dispersion des données à fouiller.

2.1.2. *L'identification des événements ou des états*

L'identification des événements ou des états dans un texte est également une tâche difficile alors qu'elle est souvent implicite dans les bases de données car réalisée au préalable, éventuellement à la main, par l'analyste.

Ainsi une base de données où sont décrits des individus qui ont des propriétés telles que `blond`, `1,40 m`, ou `12 cm` décrit implicitement des états *avoir les cheveux de couleur blonde*, *avoir une taille de 1,40 m* ou encore *avoir un nodule de 1 cm de diamètre*. Dans le domaine des études en marketing, si un calcul de corrélations entre des caractéristiques de clients propose une règle d'association du type `moutarde` $\Longrightarrow$ `bière`, celle-ci peut être interprétée comme *tous les clients (d'une grande surface) qui achètent de la moutarde achètent aussi de la bière* ou encore *tous les clients (d'une brasserie) qui consomment de la moutarde consomment également de la bière*, suivant le contexte de l'expérience. Bien sûr, si les deux événements *acheter de la moutarde* et *consommer de la moutarde* sont présent dans la même expérience, alors, une phase de désambiguïsation des données amènera l'analyste à définir deux prédicats clairement distincts `ach_moutarde` et `cons_moutarde`. Ce qui guide l'analyste dans ce travail de désambiguïsation vient de sa connaissance de l'expérience qu'il a peut-être réalisée lui-même ou des fichiers de métadonnées associées aux données.

Dans le cadre du texte, la construction du sens d'un texte fait appel à de nombreuses étapes. Il est de plus en plus facile de se constituer une base de textes sur support électronique mais le contexte dans lequel ce texte a été produit, la finalité de son auteur, échappe généralement à celui qui le récupère. De ce fait, le texte doit être considéré comme autonome et c'est à l'intérieur du texte qu'il faut chercher les éléments nécessaires à son interprétation. Nous verrons que si de nombreux travaux utilisent les mots-clés ou les termes pour représenter le contenu d'un texte, ce n'est cependant qu'une approximation qui atteint très vite ses limites.

3. Les textes et leur structure

L'ECT telle que nous la définissons ne fait actuellement pas appel à la notion de document structuré. Cela s'explique en premier lieu par le fait que, notamment dans le traitement automatique de la langue (TAL), les travaux sur la structure des textes et les travaux sur la compréhension[1] ont souvent été menés en parallèle, sans réelle fusion. On doit y voir ici la difficulté à identifier comment la structure influe sur le contenu du texte et sur son interprétation, mais également la difficulté à proposer une notation formelle permettant de coder cette association structure-contenu. Une seconde difficulté vient de ce que cette notation doit être dotée d'une sémantique sur laquelle il doit être possible d'effectuer des raisonnements. C'est une condition essentielle pour alimenter un processus de fouille de textes. Les travaux que nous citons dans cette section font référence à deux types de travaux : d'une part ceux qui visent à associer contenu des textes et structure et d'autre part ceux qui montrent que la structure d'un texte influe sur l'interprétation de son contenu.

Une autre raison très simple à ce manque de structure est liée au fait que la très grande majorité des travaux en ECT exploitent des corpus de résumés. Nous verrons en section 4.1 pourquoi cet objet textuel suscite tant d'intérêt.

3.1. *Structuration des documents électroniques*

Les travaux sur la structure de documents remontent aux années 1960 dans le cadre des systèmes de production de documents. On peut dire en quelques mots que ces travaux étaient très orientés vers la production de documents structurés, leur édition, la mise en forme papier ou leur visualisation [FUR 89]. Cette notion de document structuré a évolué avec la définition de SGML, XML et, sur le plan plus spécifiquement textuel, avec les travaux de la TEI (*Text Encoding Initiative*) [SPE 94, SPE 02, TEI].

L'objectif de la TEI [ROL 96, IDE 94] est de rédiger des recommandations pour la préparation, la structuration des textes électroniques en vue de faciliter leurs échanges et leur utilisation en tant que ressources pour les industries de la langue. Elles exploitent le langage SGML et identifient les différents éléments textuels par un système de balises. Le spectre d'utilisation de ces recommandations est très large puisqu'elle sont destinées à être exploitées par des lexicographes, terminologues, ou encore pour des travaux de recherche en TAL ou en recherche d'information... Les jeux de balises proposés par la TEI sont modulaires et peuvent être combinés. On peut décomposer ces ensembles en un jeu de balises noyau, commun à tous les types de textes, un jeu de balises dédié à des types particuliers de textes et des jeux de balises additionnels eux aussi transversaux aux différents types de textes.

La TEI a le mérite d'expliciter la structure d'un document par la définition de titres, de divisions, de paragraphes... Cependant, dans une démarche de fouille de

1. Compréhension est à prendre dans un sens très général : les travaux visant à associer une représentation sémantique du contenu d'un énoncé.

textes où les corpus sont constitués de plusieurs centaines ou milliers de textes, il faut observer les écarts entre les recommandations et les pratiques. Ainsi, les articles scientifiques diffusés au format HTML sur internet dans le domaine de l'astronomie, présentent une très grande variabilité dans le codage de la structure, les titres pouvant être introduits dans un document par une balise <*head*> ou, dans un autre document, par une balise <*p*> associée à des changement de graisse ou de fontes. On observe, de plus, une très grande variabilité dans l'organisation des articles. A l'inverse, certaines revues (par exemple *Antimicrobial Agents and Chemotherapy*) dans le domaine de la microbiologie suivent des règles très strictes et une organisation très régulière : résumé, introduction, matériels et méthodes, résultats, discussion, remerciements et bibliographie.

Un second point intéressant dans la TEI est la notion de type de texte. La TEI couvre une grande variété de textes et propose notamment 8 types de texte différents dont la prose, la poésie, le théâtre, les dictionnaires, les bases terminologiques... A titre d'exemple, Ide et Veronis [IDE 95] se sont intéressés aux principes forts et récurrents pour la structuration des dictionnaires. Il soulignent notamment la forte hiérarchisation des définitions et, dans la mesure où une information peut être partagée par plusieurs niveaux, la notion de portée (*scope*) est également importante à prendre en compte. Nous retrouverons cette notion de portée sous une autre forme, dans la section suivante sur les cadres de discours.

Ide et Veronis s'accordent à dire que les dictionnaires sont probablement un des types de texte les plus facile à structurer. Pourtant, si ces travaux sont encourageants et semblent conforter l'idée qu'une certaine normalisation de la structure des documents est très bénéfique, trouver une DTD adaptée à un très grand nombre de dictionnaires s'avère être une tâche extrêmement complexe.

3.2. *Vers une sémantique de la structure*

Un certain nombre de travaux se sont préoccupés d'associer une sémantique du contenu qui soit conditionnée par le contexte dans lequel l'énoncé est formulé. Comme nous l'avons déjà souligné, la structure de texte est beaucoup mieux exploitée dans une perspective de recherche d'information que dans une perspective de fouille de textes. [RUC 03] montre qu'il est possible de distinguer dans des textes des catégories d'information comme les objectifs, les méthodes, les résultats et les conclusions. Ces travaux exploitent ce qu'il est courant d'appeler la catégorisation de textes : (i) un certain nombre de classes sont prédéfinies, (ii) un ensemble de phrases constituant un jeu d'entraînement est constitué dans lequel les phrases sont assignées à une catégorie, (iii) un algorithme d'apprentissage caractérise chaque catégorie. Le système est alors en mesure d'assigner une catégorie à de nouveaux textes. Une sémantique n'a pas été véritablement assignée aux catégories comme on pourrait le souhaiter dans une perspective d'extraction de connaissances mais les performances affichées montrent qu'il est possible de rendre opérationnelle cette distinction. L'impact de tels travaux sur l'ECT n'a pas été étudié jusqu'à présent.

La notion de représentation du discours telle que la SDRT, par exemple, peut également avoir un impact sur les structures extraites des textes en vue de la fouille de textes. Nous pensons cependant que ces approches n'ont pas atteint une robustesse suffisante pour pouvoir être exploitées sur un grand nombre de textes réels. Nous introduirons tout de même ici la notion de cadre de discours dont on perçoit immédiatement l'impact en termes de fouille de textes.

En 1983, R. Martin décrit la notion d'univers du discours comme étant « l'ensemble des circonstances, souvent spécifiées sous forme d'adverbes de phrase, dans lesquelles la proposition peut être vraie ». [CHA 97] propose une typologie de marques de cohésion qui influent sur l'interprétation d'un énoncé. Il introduit la notion de cadre de discours qui délimite une certaine forme de portée à l'intérieur de laquelle des assertions peuvent être considérées comme vraies. Ces travaux ont été repris et poursuivis au sein du laboratoire Lattice et proposent de classer les expressions qui introduisent un cadre en plusieurs catégories : les cadres relationnels ou thématiques (*en ce qui concerne…*), les cadres organisationnels (*d'une part, d'autre part* ou les marqueurs de paragraphes), les cadres véridictionnels de type temporel (*En 1889…*), spatial (*Au Japon …*), représentationnel (*Dans le film de Luc…*), gnoséologique (*En chimie…*), médiatif (*Selon X…*)…

On imagine aisément, l'impact que ces différentes structures ont sur l'information que l'on peut extraire d'un texte et sur les limites dans la validité de ces informations. Jusqu'à présent, les travaux en fouille de textes n'intègrent pas cette dimension.

4. L'analyse du contenu

4.1. *Le statut privilégié des résumés*

Comme nous le montrent les sections précédentes, la notion de structure de textes intervient peu, actuellement, dans l'ECT, notamment parce que beaucoup de ces travaux exploitent des résumés d'articles scientifiques. L'objectif peut être de construire une base de connaissances, comme par exemple, l'identification dans la littérature de tous les gènes et de leurs interactions, ou encore de trouver des corrélations entre des phénomènes étudiés séparément dans des textes comme par exemple les réactions similaires de populations de bactéries différentes en présence de différents antibiotiques.

Le résumé est un texte court, très dense en termes, contenant normalement l'essentiel des résultats décrits dans l'article. Il y a donc peu de bruit (informations inutiles), peu de discussions et pas de contradiction à l'intérieur d'un même texte. Dans un résumé, la phrase est souvent longue, complexe, mais, de ce fait, plus « autonome » que dans un texte intégral, au sens où elle contiendra souvent la plupart des actants d'un événement alors que dans le texte intégral, la même information se trouve généralement distribuée sur plusieurs phrases. La prise en compte de la structure du discours apparaît donc généralement comme moins cruciale au niveau du résumé qu'au niveau du texte intégral. La phrase est également une unité mieux maîtrisée du point de vue du TAL que le niveau du discours.

Enfin, que ce soit à partir du web ou à partir de bases de données documentaires, il est actuellement relativement aisé de constituer un corpus électronique de résumés d'articles scientifiques sur un sujet donné. En revanche, les textes intégraux, disponibles au format Postscript ou Pdf, nécessitent une conversion en un texte Ascii structuré qui introduit souvent du bruit.

Dans la section suivante, nous proposons un schéma global de la fouille de textes sur lequel nous plaçons les différents domaines de recherche concernés par l'ECT : la recherche d'information, l'extraction d'information, la fouille de données (classification, règles d'association), la visualisation et l'interprétation de résultats.

4.2. *Schéma global de la fouille de textes*

L'ECT est à l'intersection de plusieurs domaines de recherche. En cherchant à présenter un schéma un peu plus détaillé de fouille de textes, il apparaît clairement que la représentation linéaire ne suffit pas et qu'il s'agit d'un processus itératif, incrémental dans lequel un même outil ou un même algorithme (comme la classification) peut être utilisée sur des mots comme sur des structures plus complexes d'objets. En réalité, les premières boucles du processus peuvent s'effectuer avec très peu de connaissances et sur des structures très simples avant de s'enrichir progressivement. Nous proposons donc un schéma général qui suit la figure 2 dans lequel un certain nombre d'étapes peuvent être court-circuitées pour permettre l'accès à des outils ou méthodes relevant des étapes suivantes.

L'apprentissage intervient presque à chaque étape et parfois sur des choses aussi simples que la constitution d'une liste de mots. On observe également que les connaissances sont utilisées dans le processus de fouille, ce qui est, pour nous, la garantie qu'au fur et à mesure que les connaissances s'enrichissent, le processus ne stagne pas en proposant des classifications toujours identiques mais au contraire qu'il permette l'identification de connaissances nouvelles.

5. Constitution de corpus

La constitution du corpus pour la fouille de textes est une étape essentielle dans laquelle les pratiques et les méthodes sont aussi importantes que les outils. Par outils, nous entendons bien sûr, les travaux classiques issus de la recherche d'information pour l'interrogation de bases documentaires et la sélection de documents tels que présentés dans [RIJ 79, SAL 89]. Mais il faut aussi mentionner les différentes méthodes de catégorisation, de classification et de cartographie qui sont toujours très utilisées pour découvrir des tendances dans certains domaines. L'approche statistique dite « des mots associés » a permis la réalisation d'outils comme LEXIMAPPE ou SDOC alors que de nouvelles approches exploitent des techniques neuronales [KOH 84, LAM 03].

Une caractéristique commune à la plupart de ces travaux est la robustesse des techniques mises en œuvre et leur capacité à traiter de très gros volumes de données ou

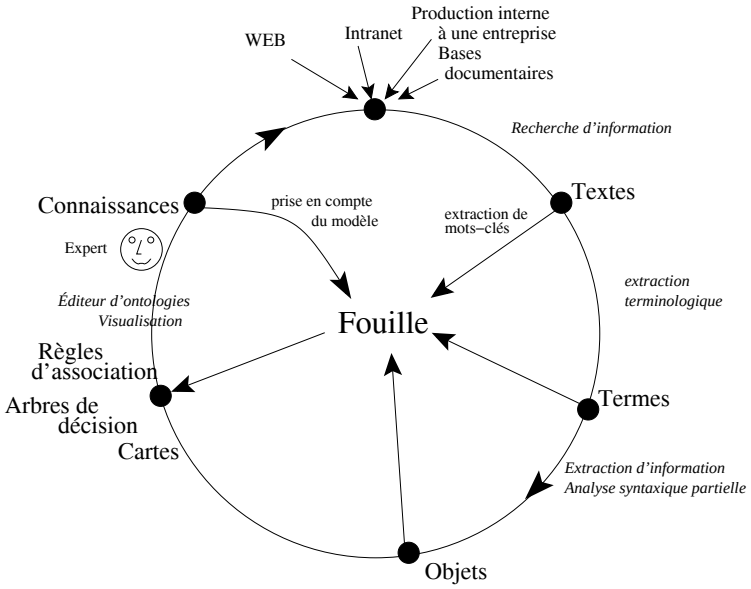

Figure 2. *Schéma global d'extraction de connaissances à partir de textes*

de textes. L'analyse se fait à un niveau très superficiel. L'unité couramment exploitée est le mot ou le mot-clé mais la notion de sens n'est pas vraiment pertinente au profit de la notion de pouvoir discriminant.

La classification peut être vue comme un moyen d'augmenter la cohésion d'un corpus. Ainsi, d'un corpus global sur l'agriculture, on peut distinguer des sous-ensembles de textes traitant plus particulièrement des questions liées à la croissance du maïs ou liées à la conservation et au transport du maïs. De même, en bibliométrie, la classification construit des réseaux d'auteurs, d'institutions ou de manifestations scientifiques qui peuvent être exploités pour contraindre ou augmenter un ensemble de textes initialement constitué.

Ce niveau d'analyse, bien qu'assez pauvre du point de vue sémantique permet néanmoins de structurer un domaine. Ces techniques sont utilisées en veille technologique ou en intelligence économique. Elles sont peu coûteuses à mettre en œuvre et polyvalentes. Certaines méthodes, comme les cartes de Kohonen multi points de vue [LAM 03] sont appréciées pour leur lisibilité. L'expert en charge de l'analyse se sert alors de classes comme d'un support pour formuler des phrases explicatives, voire des hypothèses qu'il pourra vérifier en accédant directement aux textes. L'échelle – le nombre de documents pris en compte et le nombre de classes générées – conditionne bien évidemment la qualité de cette analyse.

6. Vers l'extraction d'une information de plus en plus structurée

La recherche d'information apporte des outils robustes mais une part importante du travail d'analyse de l'information incombe à l'expert. Pour le guider davantage, un certain nombre de travaux visent à extraire une information plus précise et plus structurée. Le coût à payer, c'est la mise en œuvre de techniques issues de la linguistique de corpus, un temps de traitement plus long et, même si l'on reste dans des traitements partiels du texte, une moins bonne robustesse. Des boucles successives d'apprentissage permettent, là aussi d'améliorer progressivement les résultats.

Les travaux que nous présentons sur la terminologie (section 6.1) permettent d'identifier, à partir de critères linguistiques, le terme, qui peut être associé, dans un modèle de connaissances, au concept. Nous présentons ensuite des travaux permettant de relier ces concepts.

6.1. *Les bases de connaissances terminologiques*

Au début des années 1990, l'article de [SKU 91] a été repris en France par le groupe Terminologie et intelligence artificielle. Il s'est alors créé un véritable rapprochement entre la terminologie et l'intelligence artificielle autour de la notion de base de connaissances terminologique[2]. Du côté de l'intelligence artificielle, deux communautés se sont plus particulièrement impliquées : d'une part l'ingénierie des connaissances qui s'intéresse tout particulièrement à la construction d'ontologies et, d'autre part, le traitement automatique de la langue (TAL) qui recherche à travers la terminologie le moyen de relier les mots à des concepts mais qui apporte aussi des outils de traitement partiels et robustes. Du point de vue de la terminologie, ce rapprochement a entraîné une réflexion sur la construction du sens d'un texte dans laquelle la vision purement référentielle était inappropriée pour lui préférer une sémantique textuelle [CON 03].

Cette réflexion autour de la terminologie a coïncidé avec un regain d'intérêt pour la linguistique de corpus et l'arrivée d'outils de traitement de corpus performants, notamment, les étiqueteurs.

6.1.1. *La préparation du corpus*

6.1.1.1. L'étiquetage des textes

L'étiquetage morpho-syntaxique consiste à assigner à chaque mot d'un texte sa catégorie morpho-syntaxique (nom, verbe, adjectif...). L'intérêt de l'étiquetage est qu'il rend possible l'analyse d'un corpus du point de vue linguistique et non plus purement statistique, basé sur les fréquences d'association. A partir d'un texte étiqueté, il est en effet possible de caractériser ou de rechercher certaines structures syntaxiques, c'est-

2. Ces travaux ont été menés par le groupe terminologie et intelligence artificielle (TIA) du GDR-PRC I3.

à-dire, être capable de s'abstraire du mot pour passer à sa catégorie linguistique. Par exemple, la phrase (1) est étiquetée par (2) :

1) Two resistant strains were isolated after four rounds of selection.

2) Two/CD resistant/JJ strains/NNS :pl were/VBD isolated/VBN after/IN four/CD rounds/NNS :pl of/IN selection/NN ./.

Il existe plusieurs étiqueteurs morpho-syntaxiques pour le français, l'anglais et également pour d'autres langues européennes. D'autres outils, comme les analyseurs syntaxiques partiels (*chunker*) intègrent parfois leur propres étiqueteurs.

L'étiqueteur de Brill [BRI 92] ou le TreeTagger [SCH] sont des outils très couramment utilisés dans la communauté francophone. Suivant la nature du corpus et le domaine, les performances de ces outils vont de 95 % à 99 %. Ces outils exploitent des techniques d'apprentissage et leur performance dépend du jeu d'étiquettes, de la qualité et de la quantité du corpus d'entraînement. L'étiqueteur de Brill peut être complètement réentrainé (l'INaLF[3] l'a ainsi adapté au français) alors qu'il n'est possible que de compléter l'entraînement du TreeTagger pour l'adapter, par exemple à un nouveau domaine. Il faut souligner que ces outils sont robustes et qu'ils fournissent donc une réponse, même pour les mots qu'ils n'ont jamais rencontrés. Brill, par exemple, opère en utilisant des règles lexicales qui exploitent la morphologie du mot pour en prédire la catégorie.

6.1.1.2. La lemmatisation

L'étiquetage est souvent à la base de tous les travaux en linguistique de corpus mais il n'apporte pas vraiment de transformation à l'énoncé initial. Un certain nombre d'autres travaux permettent non seulement de repérer certains phénomènes mais aussi d'effectuer des regroupements motivés linguistiquement. Par rapport à la fouille de textes, l'idée est de pouvoir opérer des regroupements sémantiquement interprétables. La lemmatisation permet de ramener les mots fléchis à une forme canonique ou les verbes, à l'infinitif [NAM 00]. La morphologie dérivationnelle qui s'introduit progressivement dans le traitement de la terminologie puis de la fouille de textes vise à identifier un sens construit à partir de la base d'un mot et d'un ensemble de suffixes et à proposer une glose qui, par la suite pourrait être formalisée. Par exemple, le système DeriF [NAM 04] produit : benladenisation, NOM → [[[Benladen NPR] is(er) VERBE] tion NOM] (benladenisation/NOM, benladeniser/VERBE, Benladen/NPR) : : « action ou résultat de benladeniser ».
Pour les mots composés, DeriF calcule des relations lexicales. Si *galstralgie/NOM* est fourni en entrée, il produit :
gastralgie : synonyme de gastrodynie (maladie)
gastralgie : voir aussi gastrite, gastrose, entéralgie, hépatalgie (maladie)
gastralgie : sous-type de abdominalgie (maladie)

3. Institut national de la langue française qui est devenu maintenant sur Nancy l'ATILF (Analyse et traitement informatique de la langue française).

6.1.2. *L'extraction terminologique*

Définition 3 (Le terme) *Un terme est une unité syntaxique composée d'un mot ou d'un groupe de mots considérée comme une notion importante dans le domaine de spécialité étudié et qui peut être associée à un concept dans une base de connaissances.*

Il existe plusieurs type d'outils pour extraire une terminologie d'un corpus et le choix d'un type doit s'effectuer en lien avec la finalité du processus de fouille. Au même titre que l'on distingue l'indexation libre de l'indexation contrôlée, nous distinguons l'extraction terminologique libre de l'extraction contrôlée. Il faut souligner qu'aucune de ces méthodes ne permet de construire une base de connaissances ou une ontologie. Ce n'est là qu'une des étapes du processus global.

6.1.2.1. L'extraction terminologique libre

L'extraction terminologique libre peut se faire suivant une stratégie ascendante ou descendante. La stratégie ascendante consiste à rechercher dans les textes des collocations entre mots, à les organiser suivant des indices statistiques pour identifier ceux qui seraient potentiellement des termes. C'est une approche robuste dans la mesure où il est possible de rechercher les collocations de mots, de mots étiquetés ou de mots étiquetés et lemmatisés suivant la nature des prétraitements effectués. Elle peut, de ce fait, être peu dépendante de la langue. ANA [ENG 95] a été développé suivant ce principe. L'approche descendante est un peu plus contrainte puisqu'elle se base sur un certain nombre de patrons linguistiques qui peuvent potentiellement délimiter des termes. Ainsi Nomino [DAV 90] travaille-t-il à partir de patrons du type *Nom Prep Nom*, *Nom* de *Nom* ou encore *Nom Adjectif*. Enfin, plusieurs travaux combinent une approche descendante avec une approche ascendante, ou réciproquement. Il s'agit par exemple des outils ACABIT [DAI 94], Lexter [BOU 94] ou Xtract [SMA 93]. Dans tous les cas, l'extraction libre ne donne pas des termes mais des candidats-termes : ces outils privilégient, en effet, un silence faible (peu de termes qui auraient dû être proposés sont oubliés) au détriment du bruit qui est assez élevé (beaucoup de candidats termes ne sont pas des termes).

6.1.2.2. L'extraction terminologique contrôlée

Le principe de l'extraction terminologique contrôlée est de chercher à reconnaître les termes appartenant à un référentiel connu : thésaurus ou simple liste de termes. Contrairement à l'extraction libre, cette approche génère assez peu de bruit mais plus de silence. Pour réduire ce silence et rechercher des termes qui n'apparaissent pas toujours sous une forme strictement identique à la forme initiale, les travaux de [JAC 94] exploitent la notion de variante linguistique des termes. Cette notion de variante a d'ailleurs également été intégrée dans l'extraction libre. L'idée est que bien que certains termes apparaissent sous une forme variante, le concept qui est désigné par l'expression reste identique. On cherche donc des variantes qui préservent le sens du terme initial. L'expression comme « acte de terrorrisme » est ainsi considérée comme une va-

riante du terme « acte terrorriste ». Pour préserver le sens, FASTR (outil de reconnaissance des termes et de leurs variantes) fait appel à des métarègles de transformations linguistiques qui supposent que le corpus initial soit étiqueté et lemmatisé. Certaines dérivations morphologiques peuvent aussi être prises en compte ainsi qu'un certain typage sémantique, tout dépend de la qualité de la phase de préparation des textes. On se réfèrera à [JAC 01] ou à [DAI 02] pour une analyse plus détaillée de la notion de variation et des différents types de variation. La variation, suivant le domaine et le type de corpus représente entre 15 et 30% des occurrences des termes.

Le terme est une unité peu exploitée en recherche d'information. D'une part, le coût de préparation des textes est plus élevé que pour le mot. D'autre part, la prise en compte des termes au lieu des mots engendre une baisse des fréquences et suppose de travailler dans un espace de plus grande dimension. Le gain est faible, voire négatif pour la recherche d'information. En fouille de textes, le travail au niveau du terme peut être très positif. Il est bien sûr nécessaire d'identifier les termes parmi l'ensemble des candidats-termes, ce qui est réalisé généralement, à la main, par un expert. L'occurrence des termes dans des classes est beaucoup plus explicite et précise que l'occurrence des mots. Dans le cas de la veille technologique, par exemple, l'expert est mieux à même de verbaliser le contenu des classes. De plus, la prise en compte des formes variantes des termes et le regroupement des variantes autour de la forme initiale réduit la dispersion, et réduit donc la complexité des processus de fouille, que ce soit en classification ou en extraction de règles d'association tant sur le plan du calcul que sur le plan de l'interprétation par un expert. C'est actuellement la base de travail de l'Unité Recherche et Innovation de l'INIST[4] qui mène régulièrement des expérimentations en veille technologique en utilisant la plateforme ILC d'indexation terminologique développée en collaboration avec l'équipe Orpailleur du LORIA.

6.1.2.3. L'extraction de relations entre termes

Les relations entre les termes sont très importantes pour la construction d'une base de connaissances. Les systèmes SEEK [JOU 93] ou COATIS [GAR 98] travaillent à partir de marqueurs de relations définis *a priori*. Les travaux comme Promethee [MOR 99], se rapprochent davantage de techniques d'apprentissage : si une relation ρ a été identifiée entre deux termes $\mathtt{t}_1$ et $\mathtt{t}_2$, la recherche de toutes les phrases où ces deux termes apparaissent permet d'identifier les différents marqueurs de cette relation. Dans un contexte un peu différent, [SEB 02] utilise la programmation logique inductive pour l'apprentissage de relations lexicales et leur caractérisation. Les résultats d'un telle méthode sont très attrayants.

L'extraction de relations entre termes peut être rapprochée des phases d'apprentissage en extraction d'information (voir section 6.2). En effet, tous ces travaux identifient des notions puis cherchent des marqueurs permettant de les retrouver en corpus. Ces deux méthodes se distinguent cependant sur deux points. Tout d'abord, la recherche de relations entre termes est issue de travaux en linguistique alors que l'ex-

4. Institut de l'information scientifique et technique, UPS 0076 du CNRS.

traction d'information est davantage issue de la recherche d'information. De ce fait, et cela constitue le second point, les marqueurs de relation doivent avoir une validité linguistique alors que les éléments d'identification en extraction d'information doivent en priorité être performants en termes de rappel et de précision avant d'être linguistiquement pertinents.

6.1.2.4. Les limites de la représentation par les termes

L'extraction terminologique peut être considérée actuellement comme un processus robuste ayant acquis une bonne maturité. Cependant, du point de vue de l'ECT, un certain nombre de difficultés subsistent. La sélection des termes parmi les termes-candidats nécessite un travail important. L'extraction de relations est également un travail long qui reste *ad hoc* et dépendant du domaine.

Dans le cadre de la veille technologique, l'interprétation au niveau du terme est bien meilleure qu'au niveau du mot mais on constate qu'une très grosse part de travail revient à l'expert du domaine. Ainsi, dans [CHE 02], nous proposons une méthode de veille technologique basée sur l'extraction de règles d'association [5] sur les termes extraits des textes de façon contrôlée à partir d'une nomenclature de référence. Notre évaluation montre que l'expert retrouve, bien l'expression dans les règles d'association des connaissances qu'il a du domaine, mais il est, en revanche, très difficile pour lui d'identifier de nouvelles tendances ou des travaux marginaux constituant une avancée importante dans le domaine. Les limites que nous avons identifiées pour cette approches sont les suivantes :

– la représentation d'un texte par un ensemble de termes est trop pauvre. Ainsi, dans une phrase comme :
« When maintained under nonselective conditions, neither the aadA mRNA nor the AadA protein were detected in these subclones »
le terme mRNA sera identifié au même titre que dans la phrase
« Identification of an ABC transporter gene that exhibits mRNA level overexpression in fluoroquinolone-resistant Mycobacterium smegmatis ».
Pourtant, dans la première phrase, la négation montre bien que mRNA n'est pas présent dans cette expérience alors qu'il l'est dans la seconde. La représentation par un ensemble de termes ne permet pas de refléter cette différence qui peut, par la suite engendrer des erreurs d'interprétation. Seul l'accès aux textes pendant la phase de d'analyse des résultats du processus de fouille permet de se rendre compte de cette différence.

De plus, dans des domaines comme la chimie ou la biologie, les conditions d'expérience (concentration, températures. . .) sont très importantes et ne sont pas reflétées correctement par une telle représentation.

– La corrélation de termes reflète généralement l'existence d'un lien sémantique entre eux. Cependant, ces liens sont plus souvent de nature ontologique que le reflet d'une expérience particulière ;

5. Nous donnerons une définition formelle des règles d'association en section 7.1 p. 29.

– enfin, l'indexation contrôlée ne permet pas de repérer de nouveaux termes et limite donc la portée de la veille à des notions déjà connues.

L'extraction d'information que nous présentons dans la section suivante, s'inscrit, à l'origine, dans une perspective moins linguistique et plus statistique mais l'évolution actuelle de ces travaux tend à gommer cette différence.

6.2. *L'extraction d'information*

Extraire des textes une information plus précise et plus structurée est donc une étape incontournable pour faire faire un pas significatif à des activités comme la veille technologique. Les contraintes à prendre en compte sont d'une part, de pouvoir représenter le contenu de gros volumes de textes pour en analyser ou en comparer le contenu, d'autre part, de définir un niveau de représentation adapté aux textes, au domaine et à la nature des raisonnements que l'on souhaite effectuer sur ces textes.

L'extraction d'information apparaît comme un intermédiaire entre la représentation par ensemble de termes et la représentation sémantique issu du TAL. En effet, la sémantique dans le TAL s'intéresse peu au contexte ou à la finalité, mais s'interroge sur la représentation théorique d'une phrase ou d'un discours. Construire de telles représentations est assez coûteux alors que le processus de fouille obligera à une simplification des structures, voire un éclatement de ces structures sémantiques. Prenons un exemple. Dans un corpus de microbiologie, l'expert s'intéresse aux phénomènes de mutation génétique des bactéries en résistance aux antibiotiques. Les notions déterminantes pour lui sont, par exemple, des structures du type prédicatif comme `résistance`(`bactérie1`, `antibiotique1`) ou `mutation`(`gène1`, `type1`) et éventuellement d'autres éléments reflétant les conditions d'expérience. L'extraction de corrélations « éclatées » entre `bactérie1`, `gène1`, `type1`... sont moins explicites pour l'expert que les corrélations de structures prédicatives. A l'inverse, une structure sémantique qui intègre tous les éléments du texte ne lui sera pas très utile non plus. La recherche de corrélation est donc guidée par la nature des textes et par le type de notions manipulées. L'extraction d'information est donc bien appropriée à cet objectif.

L'extraction d'information est un domaine qui s'est beaucoup développé à l'initiative de la DARPA vers la fin des années 1980 grâce à l'organisation régulière des conférences d'évaluation MUC (*Message Understanding Conference*). Elle consiste à rechercher dans des textes en langue naturelle des informations à propos de classes d'entités et de relations prédéfinies et de stocker cette information dans un modèle ou dans une base de données. Cela peut se voir aussi comme le fait de compléter un formulaire ou de remplir une base d'information structurée à partir de texte libre. Une synthèse a été publiée récemment dans le cadre de la thèse de T. Poibeau [POI 03]. Si nous reprenons l'exemple présenté dans [NED 01] et donné en figure 3, la phrase en gras dans le texte suivant :dans le résumé est représentée par une structure donnée en table 1.

Le développement d'un système d'extraction d'information repose sur trois étapes

UI : 99175219[...]
Abstract :[..] It is a critical regulator of cot genes encoding proteins that form the spore coat late in development. Most cot genes, and the gerE gene, are transcribed by sigmaK RNA polymerase. **Previously, it was shown that the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK**. Here, we show that GerE binds near the sigK transcriptional start site, [..]

Figure 3. *Exemple de résumé d'une notice bibliographique*

Interaction :	**Type :**			
	Agent :	GerE protein		
	Target :	**Expression :**	**Source :**	SigK gene
			Product :	sigmaK protein

Tableau 1. *Le formulaire correspondant à la phrase ci-dessusau texte de la figure 3*

[NED 01] : (i) il faut identifier les fragments de textes pertinents, *i.e.* contenant une information ; (ii) définir la structure de représentation de l'information, puis (iii) développer les règles permettant d'identifier l'information puis remplir la structure proposée. Le principe sous-jacent à l'extraction d'information est de décomposer l'identification d'informations, parfois complexes, en des sous-problèmes simples. Ainsi, les grandes tâches qui sont classiquement identifiées sont [HUM 00] :

– reconnaissances d'entités nommées : noms de personnes, d'organisations, expressions temporelles comme les dates ou les heures, unités de mesures...

– résolution des coréférences : pronoms, la désignation par la fonction (Premier Ministre), abréviations et variantes orthographiques (béta-lactamase, $\beta-$lactamase)...

– extraction de propriétés : localisation géographique, association entre une personne et ses fonctions...

– identification de relations : interaction entre des protéines, personne employée par une société...

– identification des événements : les événements sont des relations complexes telle que la description d'une attaque terroriste dans laquelle doivent être identifiés les terroristes, les victimes, les lieux, la date...

En 1998, les performances évaluées au cours de la campagne de la septième conférence *MUC 7* sont données par le tableau 2 [GAI 02].

Du simple remplissage de formulaire, les travaux de [NED 01] montrent bien qu'il est possible d'extraire une information structurée et sémantiquement forte. En effet, initialement, un formulaire est une structure sémantiquement faible : il est composé de champs et ces champs sont remplis par des valeurs qui ne sont autres que des chaînes de caractères. Pourtant, pour identifier ces valeurs dans les textes, le système fait généralement appel à un typage des objets (lieux publics, personnalités...) qui n'apparaît plus dans la représentation finale en formulaire. La structure proposée par [NED 01]

Tâche	rappel(%)	précision(%)
entités nommées	92	95
coréférence	56,1	68,8
entités-propriétés	86	87
relation	67	86
événements	42	65

Tableau 2. *Performances évaluées au cours de la conférence MUC 7*

fait évoluer cette représentation vers une structure objet ou encore une structure de concepts dans laquelle les étiquettes de champs sont des relations et les valeurs sont des concepts (des classes) ou des instances.

Quelques expériences ont été menées notamment par [NAH 00, NAH 02] pour associer extraction d'information et fouille en exploitant des règles d'association. L'originalité de ce travail vient de ce que les entités extraites des textes sont associées à un champ du formulaire et c'est ce couple (champ, entité) qui est donné en entrée au processus de fouille. Au lieu d'extraire des règles d'association qui portent simplement sur des termes extraits des textes, comme par exemple `HTML`, `DHTML` $\rightarrow$ `XML`, les règles se présentent sous la forme (`HTML` $\in$ `language`), (`DHTML` $\in$ `language`) $\rightarrow$ (`XML` $\in$ `language`) ou encore (`ODBC` $\in$ `application`) $\rightarrow$ (`JSP` $\in$ `language`) où `HTML`, ...`JSP` sont des termes extraits des textes et `langage`, `application`... sont des champs.

7. Le processus de fouille

Les processus mis en œuvre pour la fouille dans les textes exploitent très largement les méthodes et techniques mises en œuvre plus globalement sur des données. Nous ne souhaitons pas présenter l'ensemble de ces méthodes puisque certains ouvrages en dressent un panorama assez large et en font de bonnes présentations [COR 02, HAN 01]. Nous nous intéressons ici à quelques travaux qui portent plus spécifiquement sur le texte et nous mettons l'emphase sur un principe qui nous est cher : la fouille de données, et peut-être plus encore la fouille de textes, ne peut se faire qu'en s'appuyant sur des connaissances du domaine. Autrement dit, pas de fouille sans modèle de connaissances. Ainsi, l'enrichissement progressif du modèle de connaissances par les itérations successives de la boucle d'ECT permet d'extraire de nouvelles connaissances et d'enrichir de nouveau le modèle. Les règles d'association ainsi que certaines méthodes de classication symbolique comme les treillis de Galois en analyse formelle des concepts peuvent être plus facilement associées à un modèle de connaissances que certains autres méthodes numériques.

7.1. *Les règles d'association*

De par leur lisibilité, les règles d'association constituent une méthode de fouille attractive. Nous donnons brièvement une définition des règles d'association puis mentionnons quelques travaux sur leur utilisation en fouille de textes.

7.1.1. *Définition formelle des règles d'association*

Nous introduisons les règles d'association dans le contexte de la fouille de textes. Chaque texte est représenté par un ensemble de concepts qui modélise son contenu.

Définition 4 (Règle d'association) *Soit $\mathcal{T} = \{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ l'ensemble fini non vide de textes. Soit $\mathcal{C} = \{\mathtt{c}_1, \ldots, \mathtt{c}_n\}$ l'ensemble des concepts modélisant les textes $\mathcal{T}$.$\mathcal{T}$ et $\mathcal{C}$ sont liés par une relation $R \subseteq \mathcal{T} \times \mathcal{C}$.*

*Une règle d'association est une implication de la forme $\mathtt{B} \Longrightarrow \mathtt{H}$ où $\mathtt{B}$ est la prémisse (*body*), et $\mathtt{H}$ est la conclusion (*head*) avec $\mathtt{B} \subseteq \mathcal{C}$, $\mathtt{H} \subseteq \mathcal{C}$ et $\mathtt{B} \cap \mathtt{H} = \emptyset$.*

Si $\mathtt{B} = \{\mathtt{c}_1, \ldots, \mathtt{c}_p\}$ est l'ensemble des concepts de la prémisse d'une règle d'association r_j et $\mathtt{H} = \{\mathtt{c}_{p+1}, \ldots, \mathtt{c}_q\}$ l'ensemble des concepts de la conclusion de r_j, $\mathtt{B} \Longrightarrow \mathtt{H}$ signifie que tous les textes de $\mathcal{T}$ contenant les concepts $\mathtt{c}_1, \mathtt{c}_2, \ldots, \mathtt{c}_p$ contiennent aussi les concepts $\mathtt{c}_{p+1}, \mathtt{c}_{p+2}, \ldots, \mathtt{c}_q$ avec une certaine probabilité $\mathtt{P}$.

Le support de r_j est le nombre de textes contenant les concepts de $\mathtt{B}$. La confiance de r_j est le rapport entre le nombre de textes contenant l'ensemble des concepts $\mathtt{B} \cup \mathtt{H}$ ($\{\mathtt{ci}_1, \ldots, \mathtt{ci}_p, \ldots, \mathtt{ci}_q\}$) et le nombre de textes contenant $\mathtt{H}$ ($\{\mathtt{ci}_1, \ldots, \mathtt{ci}_p\}$). Ce rapport est défini par la probabilité conditionnelle $\mathtt{P}(\mathtt{B}|\mathtt{H})$. Le support et la confiance sont deux mesures associées aux règles d'association [AGR 96] et exploitées par les algorithmes d'extraction de règles pour en réduire la complexité. Deux valeurs de seuils sont alors définies $\sigma_\mathtt{s}$ pour le support minimal et $\sigma_\mathtt{c}$ pour la confiance minimale.

7.1.2. *Définition de mesures de qualité pour les règles*

Le plus grand handicap de l'extraction des règles d'association comme méthode de fouille est lié au très grand nombre de règles générées. En effet, le choix de valeurs faibles pour les deux seuils $\sigma_\mathtt{s}$ et $\sigma_\mathtt{c}$ peut produire plusieurs centaines de milliers de règles. De nombreux travaux, issus des statistiques, se sont donc intéressés à définir des indices statistiques pour l'évaluation de la qualité des règles. Parmi ces indices, on peut citer l'intérêt, la dépendance, la conviction ou l'étonnement [GRA 01, CHE 02, AZE 02].

S'il ne se dégage pas de valeurs de seuil parfaitement claires pour ces différents indices ni une méthodologie de sélection des règles très précise, il apparaît tout de même que l'accès selon des valeurs décroissantes de l'intérêt ou de la dépendance permet à un expert d'accéder en priorité à des règles qu'il juge intéressantes [CHE 02]. Cette analyse nous conforte dans l'idée qu'une méthode de sélection statistique, fondée sur la distribution des données qui ont servi à l'extraction des règles n'est pas suffisante.

Cela accrédite notre thèse selon laquelle nous devons intégrer un modèle de connaissances dans notre processus.

7.2. *Construction et prise en compte d'un modèle du domaine*

Un certain nombre de travaux s'intéressent à la prise en compte d'un modèle de connaissances et à son enrichissement. Nous proposons, par exemple [JAN 03], d'introduire un modèle de connaissances purement hiérarchique pour filtrer les règles d'association triviales, *i.e.* les règles qui ne font que valider des connaissances hiérarchiques déjà établies. Ainsi si le modèle de connaissances établit que `a est − un b`, ce qui est interprété formellement comme tous les individus appartenant à la classe des `a` appartient aussi à la classe des `b`, alors une règle d'association `a` $\Longrightarrow$ `b` est considérée comme triviale. La démarche s'appuie sur la définition d'un modèle de connaissances probabiliste et propose de calculer une valeur de vraisemblance définie sur $[0, 1]$ pour chaque règle, valeur qui est d'autant plus forte que la règle est triviale. Une règle non triviale (avec une valeur de vraisemblance faible) peut devenir triviale (avec une valeur de vraisemblance forte) après enrichissement du modèle de connaissances.

D'autres travaux exploitent directement un modèle de connaissances symbolique et nous n'en citerons que quelques-uns. [SRI 95] propose un algorithme de calcul de règles d'association sachant que les propriétés sont structurées en hiérarchie. Ces travaux ont été repris par [MAE 00b, MAE 00a] pour la construction d'ontologies à partir de textes et l'identification de relations entre concepts. [HAH 98] utilise un modèle de connaissances et des patrons lexico-syntaxiques pour acquérir des relations hiérarchiques. On peut également citer ASIUM [FAU 98] qui permet d'acquérir des connaissances hiérarchiques et des patrons de sous-catégorisation par un système d'apprentissage interactif.

Ces méthodes et outils ne permettent cependant pas de construire directement une ontologie, ou un modèle de connaissances. Cette dernière étape fait nécessairement appel à un expert. Le passage du texte à une représentation formelle peut se faire grâce à des outils d'édition d'ontologie comme Terminae [BIé 99] ou SMES [MAE 00a]. Le rôle de ces outils est important car l'extraction automatique ne permet pas à elle seule d'obtenir une structuration formelle des connaissances ni même de faire des choix dans la structuration qui sont dépendants de l'application pour laquelle l'ontologie est construite.

8. Conclusion

Dans cet article, nous dressons un panorama des travaux autour de l'extraction de connaissances à partir de textes. Nous proposons un schéma en boucle pour définir cette activité qui fait appel à des domaines de recherches qui étaient jusqu'à présents relativement disjoints. Malgré tout, il est impossible de faire figurer sur un tel schéma chacune des phases d'apprentissage et toutes les ressources produites par ces

différentes phases car, comme nous l'avons vu, le parcours à l'intérieur d'une même boucle est peu contraint. Nous avons insisté sur une des spécificités de l'ECT comparé à l'ECBD, à savoir, la prise en compte du document et du texte dans ses diverses dimensions. Nous avons, de ce fait, pris le risque de faire un résumé un peu trop rapide des problèmes de fouille proprement dite et de représentation des connaissances. Malgré tout, nous insistons sur le fait que ce n'est que par la prise en compte de connaissances du domaine que les processus de fouille de textes peuvent faire émerger des connaissances nouvelles ou qu'ils peuvent aider l'expert dans sa tâche d'interprétation. Nous soulignons donc la nécessité de faire évoluer les processus qui extraient l'information pour que celle-ci soit plus précise et associée à une sémantique formelle qui puisse être exploitée par des outils de fouille.

9. Bibliographie

[AGR 96] AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A., « *Fast Discovery of Association Rules* », chapitre 12, p. 307 - 328, AAAI Press / The MIT Press, 1996.

[AND 89] ANDRE J., FURUTA R., QUINT V., Eds., *Structured documents*, The Cambridge Series on electronic publishing, 1989.

[AZE 02] AZE J., KODRATOFF Y., « A Study of the Effect of Noisy Data in Rule Extraction Systems », *Proc. of EMCSR 2002 : 16th European Meeting on Cybernetics and Systems Research*, Vienna, April 2002, 6 pages.

[BIé 99] BIÉBOW B., SZULMAN S., « TERMINAE : A linguistic-based tool for the building of a domain ontology », *11° European Workshop, Knowledge Acquisition, Modeling and Management (EKAW 99), Dagstuhl Castle, Germany*, 1999, p. 49-66.

[BOU 94] BOURIGAULT D., LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquistion des connaissances à partir de textes, Thèse d'Université, Ecole des Hautes Etudes en Sciences Sociales, juin 1994.

[BRI 92] BRILL E., « A Simple Rule-Based Part of Speech Tagger », ANLP-ACL, Ed., *Third Conference on Applied Natural Language Processing*, avril 1992.

[CHA 97] CHAROLLES M., « L'encadrement du discours : univers, champs, domaines et espaces », *Cahiers de Recherche linguistique, Landisco, URA 1035 du CNRS et Univ. de Nancy*, vol. 6, 1997, p. 1-73.

[CHE 02] CHERFI H., TOUSSAINT Y., « Interprétation des règles d'association extraites par un processus de fouille de textes », *13ème Congrès francophone AFRIF-AFIA de Reconnaissance des Formes et d'intelligence Artificielle - RFIA'02, Angers, France*, vol. 3, Jan 2002, p. 975-983.

[CON 03] CONDAMINES A., « Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques », Thèse d'habilitation à diriger des recherches, Equipe de Recherche en Syntaxe et Sémantique, Université de Toulouse II - Le Mirail, 2003.

[COR 02] CORNÉJUOLS A., MICLET L., *Apprentissage artificiel*, Eyrolles, 2002.

[DAI 94] DAILLE B., Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse d'Université, Université de Paris VII, Computer Communication and Vision, TALANA, février 1994.

[DAI 02] DAILLE B., Découvertes linguistiques en corpus, Thèse d'Université, Mémoire d'Habilitation à Diriger des Recherches en Informatique, Université de Nantes, 2002.

[DAV 90] DAVID S., PLANTE P., « De la nécessité d'une approche morpho-syntaxique dans l'analyse des textes », *Intelligence artificielle et sciences cognitives au Québec*, vol. 3, n° 3, 1990, p. 140-154.

[ENG 95] ENGUEHARD C., PANTERA L., « Automatic natural acquisition of terminology », *Journal of Quantitative Linguistics*, vol. 2, n° 1, 1995, p. 27-32.

[FAU 98] FAURE D., NEDELLEC C., « A corpus-based conceptual clustering method for verb frames and ontology acquisition », *LREC Workshop on adaptating lexical and corpus resources to sublangages and applications, Granada, Spain*, 1998.

[FAY 96a] FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

[FAY 96b] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., « *From Data Mining to Knowledge Discovery* », chapitre 1, Fayyad et al. [FAY 96a], 1996.

[FEL 95] FELDMAN R., DAGAN I., « Knowledge Discovery in Texts », *Proceedings of the First International Conference on Knowledge Discovery*, 1995, p. 112-117.

[FEL 96] FELDMAN R., HIRSH H., « Mining Associations in Text in the Presence of Background Knowledge », *Knowledge Discovery and Data Mining*, 1996, p. 343-346.

[FUR 89] FURUTA R., « *Concepts and Models for Structured Documents* », chapitre 1, Andre et al. [AND 89], 1989.

[GAI 02] GAIZAUSKAS R., « An Information Extraction Perspective on Text Mining : Task, Technologies and Prototype Applications », Euromap Text Mining Seminar, 2002.

[GAR 98] GARCIA D., Analyse automatique des textes pour l'organisation causale des actions, réalisation du système COATIS, Thèse d'Université, Thèse d'informatique, Université Paris IV, 1998.

[GRA 01] GRAS R., KUNTZ P., COUTURIER R., GUILLET F., « Une version entropique de l'intensité d'implication pour les corpus volumineux », BRIAND H., GUILLET F., Eds., *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, vol. 1 de *1-2*, Nantes, France, January 2001, Éditions Hermès, p. 69-80.

[HAH 98] HAHN U., SCHNATTINGER K., « Towards text knowledge engineering », *Proc. of AAAI'98*, 1998, p. 129-144.

[HAN 01] HAN J., KAMBER M., *Data Mining : Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers,, 2001.

[HUM 00] HUMPHREYS K., DEMETRIOU G., GAIZAUSKAS R., « Two applications of information extraction to biological science journal articles : enzyme interactions and protein structures », *Proceedings of the Pacific Symposium on Biocomputing (PSB-2000)*, Honolulu, Hawaii, USA, January 2000, p. 505-516.

[IDE 94] IDE N., VÉRONIS J., « Validation of the TEI Guidelines in the Framework of an EU Initiative », *SGML Europe'94*, Montreux (Suisse), 1994.

[IDE 95] IDE N., VÉRONIS J., « Encoding dictionaries », *Computers and the Humanities*, vol. 29, n° 2, 1995, p. 167-179.

[JAC 94] JACQUEMIN C., ROYAUTÉ J., « Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework », *Proceedings of the 17th ACM SIGIR Conference on*

Research and Development in Information Retrieval, 1994, p. 1058-1063.

[JAC 01] JACQUEMIN C., *Spotting and Discovering Terms through NLP*, MIT Press, Cambridge MA, 2001.

[JAN 03] JANETZKO D., CHERFI H., TOUSSAINT Y., NAPOLI A., « Knowledge-based Selection of Association Rules for Text Mining », *Conf. European Conference on Artificial intelligence (ECAI 04)*, Valencia, Espagne, 2003.

[JOU 93] JOUIS C., Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système SEEK, Thèse d'Université, Thèse d'informatique, EHESS, Paris, 1993.

[KOH 84] KOHONEN T., *Self-Organization and Associative Memory*, Springer Verlag, 2 édition, 1984.

[LAM 03] LAMIREL J., TOUSSAINT Y., SHEHABI S. A., « A Hybrid Classification Method for Database Contents Analysis », *The 16th International FLAIRS Conference - FLAIRS 2003, St. Augustine, Florida*, May 2003.

[LEN 97] LENT B., AGRAWAL R., SRIKANT R., « Discovering Trends in Text Databases », HECKERMAN D., MANNILA H., PREGIBON D., UTHURUSAMY R., Eds., *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD*, AAAI Press, 1997, p. 227-230.

[MAE 00a] MAEDCHE A., STAAB S., « Discovering Conceptual Relations from Text », *European Conference on Artifical Intelligence (ECAI 2000)*, 2000.

[MAE 00b] MAEDCHE A., STAAB S., « Mining Ontologies from Text », *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, 2000.

[MOR 99] MORIN E., « Automatic Acquisition of semantic relations between terms fromtechnical corpora », *5th International Congress on terminology and Knowledge engineering, TKE'99*, 1999.

[NAH 00] NAHM U., MOONEY R., « A mutually beneficial integration of data mining and information extraction », *Actes de la Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, 2000, p. 627-632.

[NAH 02] NAHM U. Y., MOONEY R. J., « Text Mining with Information Extraction », *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002, p. 60-67, Stanford, CA.

[NAM 00] NAMER F., « Flemm : Un analyseur Flexionnel du Français à base de règles », JACQUEMIN C., Ed., *TAL, numéro spécial "Traitement automatique des langues pour la recherche d'information"*, Hermes, 2000.

[NAM 04] NAMER F., « Automatiser l'analyse morpho-sémantique non affixale : le système DériF », *Cahiers de Grammaire (à paraître), Toulouse*, vol. 28, 2004.

[NED 01] NEDELLEC C., VETAH M., BESSIÈRE P., « Sentence Filtering for Information Extraction in Genomics, a classification problem », RAEDT L. D., SIEBES A., Eds., *PKDD 2001, LNAI 2168*, p. 326 - 337, Springer-Verlag, 2001.

[PIA 91] PIATETSKY-SHAPIRO G., FRAWLEY W., *Knowledge Discovery in Databases*, AAAI Press, 1991.

[POI 03] POIBEAU T., *Extraction automatique d'information*, Hermès, Lavoisier, 2003.

[RIJ 79] RIJSBERGEN C. V., *Information Retrieval*, Butterworths & Co., 1979.

[ROL 96] ROLE F., « TEI : Text Encoding Initiative », *numéro spécial des Cahiers de Gutenberg*, vol. 24, 1996, page 190.

[RUC 03] RUCH P., CHICHESTER C., COHEN G., CORAY G., EHRLER F., GHORBEL H., MÜLLER H., PALLOTTA V., « Report on TREC 2003 Experiment : Genomic Track », *Conférence TREC 2003, Gaithersburg, Maryland, November 18-21, 2003.*, 2003.

[SAL 89] SALTON G., *Automatic Text Processing*, Addison Wesley Publishing Company, 1989.

[SCH] SCHMID H., http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

[SEB 02] SEBILLOT P., « Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues », 2002, Habilitation à diriger des recherches, Université de Rennes 1.

[SKU 91] SKUCE D., MEYER I., « Terminology and Knowledge Engineering : Exploring a Symbiotic Relationship », *6th International Workshop on Knowledge Acquisition for Knowledge-based Systems*, 1991.

[SMA 93] SMADJA F., « Retrieving collocations from texts : Xtract », *Computational Linguistics*, vol. 19, n° 1, 1993, p. 143-177, Special Issue on Using Large Corpora.

[SPE 94] SPERBERG-MCQUEEN C., BURNARD L., *Guidelines foir Electronic Text Encoding and Interchange*, Text Encoding Intiative, Chicago and Oxford, 1994.

[SPE 02] SPERBERG-MCQUEEN C., BURNARD L., Eds., *Guidelines for Text Encoding and Interchange*, Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford, 2002, ISBN 0-952-33013-X, see also http ://www.tei-c.org/ (TEI Home) or http ://www.tei-c.org/P4X/ latest version of the guidelines.

[SRI 95] SRIKANT R., AGRAWAL R., « Mining Generalized association rules », *Proceedings of the VLDB'95*, 1995, p. 407-419.

[TEI] TEI, « Text Encoding Initiative », consulter http ://www.tei-c.org/ ou http ://www.tei-c.org/P4X/ pour la dernière version en ligne.