Chapitre 4

Etat de l'art
Apprentissage ontologique
Extraction des connaissances à partir des textes

Introduction

Dans le cadre de notre recherche qui puise des sources textuelles « arabes », on peut se poser la question pourquoi les textes et pas autres choses ? Cette question trouve sa justification dans la définition même d'une ontologie citée par Grüber 1993, quand il fait intervenir la notion de consensus et de la conceptualisation partagée. En effet les textes sont très riches en connaissances et accumulent un vocabulaire partagé entre une grande communauté d'un domaine.

Notre problème est donc d'acquérir, à partir d'un texte, un ensemble de connaissances utiles pour la construction d'une ontologie. Il est question donc, de s'informer à partir de la langue écrite pour acquérir de la connaissance, dès lors que l'on veut y accéder à l'aide d'outils informatiques.

Dans cette partie, on présente un état de l'art d'extraction de connaissances à partir de textes. En ingénierie ontologique cet axe de recherche s'appelle « l'apprentissage ontologique » ou communément parlant « Ontology learning ». Ce chapitre va être sectionné en deux parties :

Dans une première partie, nous allons présenter un spectre illustrant aux utilisateurs différentes systèmes existants (liste non exhaustive) dans ce domaine ainsi que les moyens qui leurs permettrons de les comparer. Pour cela une présentation et un recensement sur toutes les techniques, stratégie, disciplines, approches et backgrounds agissants dans ce contexte seront étudiées. [**SHA02-a**] présente une classification en dimensions selon des critères pour faciliter les interactions entre ce domaine et les autres disciplines.

La seconde partie, est réservée pour la description d'un processus consensuelle de l'apprentissage ontologique (Ontology learning), ainsi qu'une classification des approches les plus importantes.

Partie I

Comparaison entre différents systèmes & approches

Une cinquantaine (50) de systèmes d'extraction de connaissances (d'apprentissage ontologique à partir des textes) issus des travaux récents de laboratoires, de conférences et de revues publiés, sont exploités par [SHA02-a] et choisi, parmi eux sept, systèmes les plus distingués pour ensuite relever leur différences dans un cadre de comparaison.

1. Les systèmes d'apprentissages ontologiques

L'apprentissage ontologique se réfère à l'extraction des éléments ontologiques (connaissances conceptuelles) à partir des textes et construit ensuite une ontologie avec ces éléments. La construction manuelle des ontologies est une tâche lourde et assez coûteuse en temps, chers, biaisé en fonction de leur développeur, non-flexible aux changements et spécifiques seulement aux objectifs tracés. L'automatisation de la construction d'ontologies élimine non seulement les coûts, mais aussi, il en résulte une meilleure ontologie correspondante à son application.

Beaucoup de systèmes, utilisant l'approche semi-automatiques d'apprentissage ontologique, ont attiré notre attention en vue de la préparation de notre état de l'art. Par exemple : Adaptiva, SOAT, OntoLearn, TextStorm, ASIUM, HASTI, DODDLE II, SVETLAN, SYNDICATE, TEXT-TO-ONTO, WEBGroup de systèmes →KB. Mais nous n'avons retenue que sept systèmes, de base, pour ce cadre de comparaison, les autres ne sont qu'une image des 7 systèmes modèles de notre études. Ainsi les systèmes retenues sont :
ASIUM, HASTI, DODDLE II, SVETLAN, SYNDICATE, TEXT-TO-ONTO, WEB→KB

Noms des Systèmes	Références	Caractéristiques
ASIUM	(Faure, et al., 1998; Faure & Poibeau, 2000)	Cadre d'apprentissage des verbes et les connaissances taxonomiques, basée sur l'analyse statistique de l'analyse syntaxique de textes en français. « **A**cquisition of **S**emant**I**c knowldge **U**sing **M**achine learning method »
DODDLE II	(Yamaguchi, 2001)	Outil de traitement pour apprendre les relations taxonomiques et non taxonomiques en utilisant de méthodes statistiques (analyse de co-occurrence), et l'exploitation des dictionnaires numérique (WordNet) et des textes spécifiques à un domaine.
HASTI	(Shamsfard 2003; Shamsfard & Barforoush, 2000; 2002a;b)	Apprentissage des mots, des concepts, des relations et des axiomes, dans les deux modes incrémental et non-incrémental, à partir d'un petit noyau (ou apprentissage à partir de zéro), en utilisant une approche hybride symbolique, les combinaisons logiques, basée sur la linguistique, basé sur des patrons, et des méthodes heuristiques.
SVETLAN'	(Chalendar & Grau, 2000)	Permet d'acquérir automatiquement des classes de noms par domaines sémantico-pragmatiques à partir de textes. Il regroupe des mots jouant le même rôle syntaxique par rapport à un même verbe, où seuls les mots les plus pertinents pour décrire le domaine sont retenus.
SYNDIKATE	(Hahn & Schnattinger, 1998; Hahn & Romacker, 2001; Hahn & Marko, 2002)	« SYNthesis of Distributed Knowledge Acquired from Texts » L'apprentissage progressif de mots, de concepts et de relations, est basé sur la compréhension du texte ou de la phrase, en utilisant deux sources, linguistique et conceptuelle « de qualité» des différentes formes d'éléments.
TEXT-TO-ONTO	(Maedche & Staab, 2000a; b; 2001)	Apprentissage des concepts et des relations à partir de données non structurées, semi-structurées et structurées, en utilisant une méthode multi-stratégie d'une combinaison de règles d'association, une analyse formelle de concepts et de Clustering.
WEB→KB	(Craven et al., 1998; 2000)	Combinaison de statistiques (bayésien) et de méthodes logiques (règles d'apprentissage FOL) pour apprendre les instances et l'extraction des règles à partir de documents du Web
Adaptiva	C. Brewster, F. Ciravegna, and Y. Wilks, 2002	Basée sur des modèles linguistiques et d'apprentissage automatique, mais avec un peu plus d'itérative et d'approche coopérative.
SOAT	T. Yamaguchi, 2000	Système hautement automatisé, apparemment très efficace. Quatre différentes relations entre les concepts sont extraites; catégorie, synonyme, d'attributs et d'événements. L'inconvénient, une préparation très lourde.
OntoLearn	A. Maedche et al. 2000,2001,2002, 2003	L'architecture OntoLearn se compose de trois phases principales : 1. Extraction terminologique 2. Interprétation Sémantique 3. Création d'une vue spécialisé de WordNet
TextStorm	A. Oliveira et al, 2001	Le système TextStorm analyse et étiquette un fichier texte, en utilisant WordNet, puis extraits des prédicats binaires à partir du corpus de texte. Les prédicats symbolisent une relation entre deux termes, extraite d'une phrase.

Tableau 6 : Systèmes proposés et sélection du cadre de l'étude de comparaison

2. les six dimensions de comparaison

Un *framework* de comparaison est proposé par Shamsfard [SHA02-a], montrant ainsi les points qui font la différence entre une méthode et une autre. Ce cadre de comparaison réunit les caractéristiques et les techniques de plusieurs approches. (Voir figure 33)

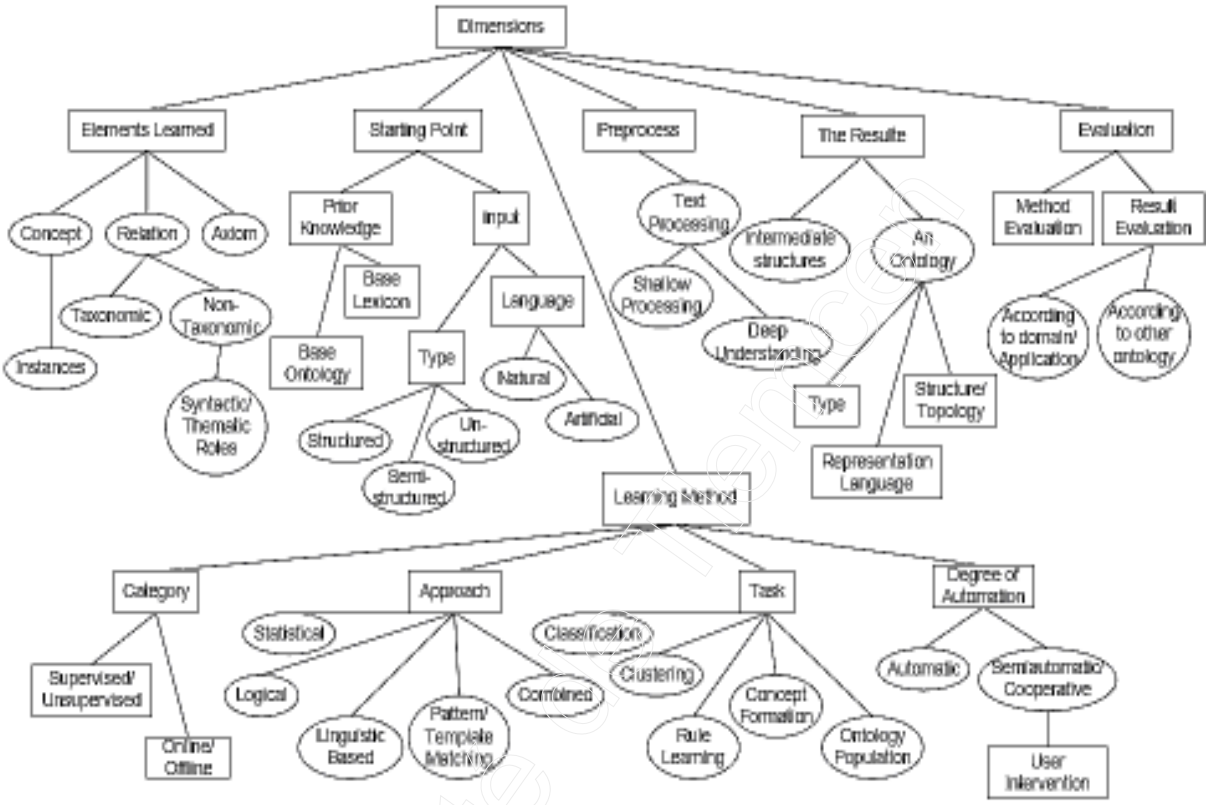

Fig. 33 : La taxonomie des six dimensions de comparaison

1. Les éléments à apprendre : Concepts, relations, axiomes, règles, instance, etc.).
2. Les sources d'apprentissages : Point de départ (textes, documents structurés, documents semi structurés, bases de données, dictionnaires, base de connaissances,…)
3. Le prétraitement : Traitement linguistique tels que la compréhension profonde ou peu profonde de traitement de texte.
4. La méthode d'apprentissage comprend :
   - Les catégories d'apprentissage : Supervisé, non supervisé {on line[1], off line[2]}
   - Les approches d'apprentissage : Statistique, symbolique, logique, basé sur la linguistique, *pattern matching*, méthodes hybrides,…

1 - L'apprentissage << on-line >> : Les exemples sont présentés les uns après les autres au fur et à mesure de leur disponibilité.
2 - L'apprentissage << off-line >> : toutes les données sont dans une base d'exemples d'apprentissage qui sont traités simultanément.

- Les tâches d'apprentissages : classification, clustering, apprentissage des règles, former des concepts, peuplement d'ontologie)
- Le degré d'automatisation : manuel, semi-automatique, coopératif, automatique.
- Le résultat : ontologie, structures intermédiaires, etc.
- Les méthodes d'évaluation : évaluation de la méthode d'apprentissage, évaluation de l'ontologie résultante.

2.1. Les éléments à apprendre

Les mots sont les principaux éléments lexicaux à apprendre. Mais comme principaux éléments ontologiques ce sont les concepts, relations et axiomes.

2.1.1. Les termes

Bien que la majorité des méthodes utilisent des sources lexicales existantes (**Text-To-Onto** [MAE00-a]; **DODDLE II** [YAM01], [KIE00], [BOR97]). D'autres systèmes soutirent par eux-mêmes la connaissance lexicale relative aux termes, comme le cas pour **SyndiKate** [HAH01] et **HASTI** [SHA02-b].

- **SyndiKate :** Utilise une hiérarchie des différentes classes des mots pour être capable de prédire la catégorie syntaxique du mot entrée, et par la suite déduire toutes les informations grammaticales qui en découlent.
- **HASTI :** Traite la phase morphologique et les catégories syntaxiques des mots avant de passer à la phase sémantique.

2.1.2. Les concepts

Un concept peut être [COR00] :

- Une définition d'un objet abstrait ou concret, élémentaire ou composé, réel ou virtuel.
- Une description d'une tâche,
- Une description d'une fonction,
- Une description d'une action,
- Une description d'une stratégie,
- D'un processus de raisonnement, etc.

Une ontologie est représentée sous forme d'une taxonomie avec les nœuds comme concepts. Ces derniers peuvent être prélevés à partir des sources d'entrées ou bien crées au cours d'un processus de raffinement via d'autres concepts.

2.1.3. Les instances

On peut trouver des systèmes qui se limitent simplement par l'enrichissement des classes de l'ontologie, cette technique est nommée « ***peuplement d'ontologie*** ». Dans cette catégorie on a le système : **WEB→KB**. [CRA00] et [SUR00]

2.1.4. Les relations entre concepts

Les relations se manifestent en deux classes : Taxonomiques et non taxonomiques.

- *Relations Taxonomiques :*

  Les ontologies sont organisées autour d'une taxonomie qui utilise les relations généralisations/spécialisations et engendre les deux type d'héritages : simple et multiple. La relation d'hyponymie « is-a » est la relation de base pour la hiérarchie. Citons des exemples de ces systèmes : **SyndiKate** [HAH01] et **HASTI** [SHA02-**b**], **DODDLE II** [YAM01], [TOD00], [AGI00], [SUR00], [HEY01], [CAR99], [DEI01], [SUN02] et [SPO02].

- *Relations Non taxonomique* :

  Ce sont toutes les relations qui excluent la relation « is-a ». On peut donc citer :

  La méronymie – la synonymie – l'antonymie – attribute-of – la possession – la causalité, ou autres. Plusieurs systèmes raisonnent avec ces types de relations :
  HASTI [SHA02-b], **Texte-to-Onto** [MAE00a], [AGI00] et [GAM02]

2.1.5. Les axiomes

Les axiomes sont utilisés pour modéliser les phrases toujours vraies. Ils sont très utiles afin de formaliser les contraintes contenues dans une ontologie, la vérification de son exactitude ou de déduire de nouvelles informations [FAR96]. Peu de système utilise l'apprentissage de part sa complexité, néanmoins le système **HASTI** apprend les axiomes dans des situations limitée, il transforme les axiomes explicites décrits à l'aide des phrases conditionnelles et quantifiées du langage naturel en des axiomes exprimés à l'aide de KIF (Knowdlge Interchange Format). Des travaux sont en cours pour étendre **HASTI** afin qu'il soit capable d'apprendre les axiomes implicites.

2.1.6. Les Méta-connaissances

Les méta-connaissances sont des connaissances ontologiques primitives qu'un système essaie d'acquérir (règles pour extraire des instances, modèles de connaissances, etc.),

pour essayer par la suite de l'exploiter dans l'extraction des connaissances ontologiques. Finkelstein et Morin [FIN99] proposent une approche pour apprendre des patrons lexico-syntaxiques pour extraire des connaissances à partir des textes. Par contre **WEB→KB** [CRA00] apprend des règles pour extraire des instances à partir des textes.

2.2. Les sources d'apprentissages

La question posée dans cette dimension est « *A partir de quoi l'ontologie va t'elle apprendre ?* ». La plupart des approches soutiennent l'idée d'acquisition à partir des connaissances déjà présentes (afin de les réutiliser) ou bien d'enrichir par de nouveaux éléments, à partir d'autres sources d'entrées (documents, Web,…). La qualité, et la quantité de la connaissance déjà existante et qui va être réutilisé, sa structure, son type, et le langage de la deuxième source d'entrée diffèrent d'un système à un autre.

2.2.1. Les sources réutilisables (Ontologie de base)

Les connaissances de base essentielles varient selon le type et le volume dans les différentes approches. Les connaissances préalables peuvent être présentées en linguistique (lexical, grammatical, modèles, etc.) ou sous forme de ressources ontologiques (l'ontologie de base). Beaucoup de projet utilise une base de connaissance lexicale (Lexicon) pour traiter des textes comme dans [KIE00], ou à des Ontologies comme Wordnet ou EuroWordNet dans **Text-To-Onto** [MAE00-b], **SyndiKate** [HAH01] et **DODDLE II** [YAM01], [WAG00], [AGI00], [TER01]. Le volume de ces sources diffère d'une approche à une autre. Le système **HASTI** [SHA02-b], démarre le processus à partir d'un noyau presque vide, dans [BRE01] à partir d'une esquisse d'ontologie ou d'un petit ensemble de mots représentant les concepts de haut niveaux [HWA99] ou bien encore d'une ontologie générique telle que **CYC** dans [LEN90].

2.2.2. Les entrées

Les sources d'entrée varient selon le type et la langue.

a. Type :

<u>*Données structurés*</u> :

- *Kashyap* extrait les connaissances à partir des schémas de base données. [KAS99]
- *Suryanto* le fait à partir d'une base de connaissances (*database schemata*). [SUR00]
- *William*, par contre, à traves une ontologie existante. [WIL00]

Mais les approches qui réutilisent WordNet couvrent la littérature.

Données Semi-structurés

C'est parce que le web est immensément riche en source d'informations, que plein de concepteurs se sont hâtés vers les documents HTML, XML et DTDs (*Documents Type Definition*) : par exemple **WEB→KB** [CRA00] et [KAV02]. Les dictionnaires aussi sont considérés comme sources d'entrées semi structuré.

Données Non structurés

Trop complexe, ce type de sources, pour extraire de la connaissance : elle peut être du texte en langage naturel comme le projet **HASTI** [SHA02-a], **SVETLAN** [HAH01] et [HEY01] ou bien on épuise à partir des textes du Web comme **Text-To-Onto** [MAE00-b] et [TOD00].

b. Langage :

Les sources d'entrées peuvent être des textes en langages naturels comme l'anglais dans **DODDLE II** [YAM01], [WAG00], [TER01], l'allemand dans **SyndiKate** [HAH01] et [HEY01], le Français dans **ASIUM** [FAU98], **SVETLAN** [CHA00] et [TOD00], le persan dans **HASTI** [SHA02-a], et aussi dans d'autres langages artificiels XML dans **Text-To-Onto** [MAE00-b] ou **RDF** dans [DEI01].

2.3. Le Prétraitement

La question posée dans ce contexte est : « Quels sont les outils à utiliser pour transformer ces entrées en une structure exploitable ? ».

Dans la catégorie des entrées textuelles, le premier traitement a fortiori est le traitement linguistique. De plus, La compréhension profonde des textes ralentisse le processus de construction de l'ontologie, mais elle permet de fournir des relations spécifiques entre les concepts, alors que les techniques peu profondes pourraient fournir des connaissances génériques sur les concepts [AGI00]. Notons que beaucoup de systèmes préfèrent les techniques du *Shallow text processing* qui engendre des techniques telles que le *tokeninzing*[1] ***P**art **O**f **S**peech tagging*[2] *(PoS)* et les analyses syntaxiques. Le système **Text-To-Onto** [MAE00-b] utilise SMES (Saarbrücken Message Extraction System) pour traiter les textes

1 - Tokeninzing : Il s'agit du processus permettant de marquer les différentes sections d'une chaîne de caractères. En effet, un ordinateur n'est pas capable seul de déterminer quels sont les mots d'une phrase ; il n'y voit qu'une chaîne de caractères. Un processus de tokenization consisterait donc à séparer ces mots, selon les espaces. [http://fr.wikipedia.org/wiki/Analyse_lexicale].

2 - En linguistique, l'**étiquetage grammatical** (*POS tagging* : *part-of-speech tagging* en anglais) est le processus qui consiste à associer aux mots d'un texte leur fonction grammaticale, grâce à leur définition et leur contexte (c'est-à-dire leur relation avec les mots adjacents dans un terme, une phrase ou un paragraphe).

allemands, **ASIUM** [FAU98] utilise *Sylex*[1] pour les textes Français, **SynDiKATe** utilise compréhension profonde pour extraire des connaissances ontologiques du texte, *InfoSleut*[2] [HWA99] fait appel un simple marqueur *Part Of speech (PoS) tagger* pour parfaire une analyse syntaxique peu profonde. Par contre HASTI [SHA02-a] utilise le système Petex qui est un traitement de texte Persan. Quand aux approches qui manipulent les Databases et les bases de connaissances, ont recours à la discipline du DATA Mining.

2.4. Les méthodes d'apprentissages

On se pose la question suivante : « Quels sont les méthodes d'extractions de connaissances ? » Comme réponses directe, on peut dire qu'il existe plusieurs méthodes selon les approches les plus simples (statistiques) aux plus complexes (logiques), comme elles peuvent être supervisées ou non supervisées. Beaucoup de systèmes diffèrent de part leurs approches méthodologiques ou par leurs tâches de réalisations. On peut dire alors que chaque approche apprend en réalisant une tâche bien précise, comme la classification, ou le *clustering*.

2.4.1. Approches d'apprentissage

Les approches de l'apprentissage ontologique peuvent être statistiques ou symboliques (basé sur la logique, la linguistique, et celles qui utilisent les patrons, chacune de ces dernières peut être combinée avec des techniques heuristiques). Les approches hybrides ne sont pas excluent, il donne un profit maximum de chacune des deux premières méthodes citées ci-dessus.

a. L'approche statistique

L'analyse statistique est appliquée sur les données extraites à partir des entrées. **Web→KB** [CRA00] utilise une analyse statistique nommée « *Bag-of-Words* » pour classer les pages web. Wagner, [WAG00] exploite une modification de l'algorithme de Li & Abe (1996) pour l'acquisition des concepts préférés dans la phase de sélection et localise le niveau

1 - **Sylex** : Un outil permettant l'affichage multilingue de mots, de phrases ou de sous phrase, avec leur contexte dans des textes déjà traduits. Cet outil doit permettre aux traducteurs de rechercher des exemples de traductions ainsi qu'aux réviseurs de vérifier la traduction d'un mot, d'une phrase ou d'une tournure de phrase dans son contexte. **[**http://www.issco.unige.ch/en/research/projects/sylex/intro.html**]**

2 - **InfoSleuth** : est un système multi-agents qui peut être configuré pour exécuter différentes activités dans le cadre de la gestion d'information dans un environnement distribué ainsi qu'un système pour la recherche coopérative d'informations dans des bases de données distribuées.
http://www.limsi.fr/~jps/enseignement/examsma/2005/2.applications/parties/Rechercheinfo.htm

de généralisation appropriée dans l'ontologie. Tandis que **Text-To-Onto** [MAE00-b], Heyer [HEY01] et **DODDLE II** [YAM01] utilise l'analyse statistique de cooccurrence des données pour apprendre des relations conceptuelles à partir de textes. [BIK99] utilise les chaîne de Markov cachées (HMM, Hidden Markov Chain) pour localiser et étiqueter les noms et les entités numériques. Notons que les approches statistiques peuvent opérer sur des mots isolés ou sur des mots dans leurs contextes.

On appelle le modèle basé sur les mots isolés : Model *bag-words* ou *unigram*. Il ignore la séquence dans laquelle le mot apparaît. Les méthodes qui traitent les mots indépendamment de leurs séquences fond appel aux règles bayésiennes, elles sont appelées *naïve bayes* comme l'approche **Web→KB** présenté par Craven et al. 2000 dans [CRA00]. Ce système classifie les documents web par la méthode naïve bayes modifiée, en construisant un modèle probabiliste pour chaque classe de document web, pour classifier ensuite chaque nouvelle page web dans la classe la plus probable à contenir les mots qui décrivent celle-ci.

La seconde classe des méthodes s'intéresse aux mots dans leurs séquences. D'une autre façon, l'identité sémantique d'un mot se reflète dans sa distribution dans des contextes différents, de sorte que le sens d'un mot est représenté en termes de mots qui lui sont co-occurrents et la fréquence des cooccurrences, c'est l'idée adoptée par ***Maedche*** [BIK99]. Quand la fréquence de deux mots ou plusieurs est élevée lors d'une construction bien définie alors celle-ci est appelée collocation. L'apprentissage par co-occurrence et collocation sont plus usités dans les méthodes statistiques.

b. L'approche logique

Plusieurs méthodes logiques sont utilisées pour extraire des connaissances à partir des entrées. Citons parmi ces approches : **ILP** (Inductive Logique Programming), le clustering basé sur FOL (First Order Logic) et l'apprentissage propositionnel basé sur la logique ; tous utilisent la déduction ou l'induction et présentent le résultat sous forme de propositions logiques de premier ordre ou d'ordre supérieur. **HASTI** [SHA02-**a**] profite de la déduction logique et des règles d'inférences pour produire de nouvelles connaissances à partir d'autres connaissances déjà présentes. Par contre le système **Web→KB** [CRA00] et [BOW00] sont basés sur l'induction des hypothèses à partir des observations (exemples) et assemble de nouvelles connaissances à partir des expériences. **ILP** (Inductive Logique Programming) se positionne au croisement de la programmation logique et l'apprentissage inductif. **FOL** est le système de ILP le plus réussi et le plus apprécié, il est repris par quelques systèmes d'apprentissage ontologique, comme **Web→KB**.

c. Les approches linguistiques

Ces approches sont beaucoup plus usitées dans la construction des ontologies à partir des textes. Parmi ces méthodes linguistiques, citons à titre d'exemple, l'analyse syntaxique d'**ASIUM** [FAU98], l'analyse morpho-syntaxique dans [ASS97], le modèle lexico-analyse syntaxique de [FIN99], le traitement sémantique de **HASTI** et la compréhension du texte utilisées par **SynDiKATe**. Toutes ces méthodes sont exploitées dans le but d'extraire des connaissances essentielles pour construire des ontologies à partir de textes en langage naturel.

Prenons un détour pour voir la méthode utilisée par ***Assadi et al***, dans [ASS97][1], il effectue une analyse morpho-syntaxique partielle pour extraire "des termes candidats" à partir de textes techniques. Ensuite l'ingénieur de connaissances, assisté par un outil de classification automatique de *clustering*, construit les champs conceptuels du domaine. Le résultat de cette analyse est un graphe constitué de syntagmes nominaux (Phrases nominales). Tout terme complexe ou composé est à son tour décomposé en deux parties : la tête et l'expansion, tous deux liées aux termes candidats complexes dans le graphe terminologique. Le graphe sera ensuite utilisé par l'analyseur conceptuel pour construire l'arbre de l'ontologie.

Quand à **ASIUM** [FAU98][2], acquiert des connaissances sémantiques à partir de textes techniques analysés syntaxiquement par le parseur *Sylex* qui fourni comme résultat des cadres syntaxique. Les adjectifs et les « mots vides » sont retirés et il n'est retenu que les « tête » des prépositions et les compléments. Un clustering conceptuel sur mots « tête » qui apparaissent dans des régularités syntaxiques. Les mots « têtes » qui apparaissent avec les <*verb*> ((<*preposition*>/<*function*>) <*headword*>) sont regroupés dans un même concept.

Exemple : **ASIUM**

« Amine voyage en bateau »

<Voyager> <Sujet> <Amine>
<en> <bateau>

Et on peut avoir :

<Voyager> <Sujet> <Bedro>
<en> <train>

<Voyager> <Sujet> <Fethi>
<en> <Voiture>

Dans l'exemple précédent **ASIUM** va créer deux *clusters* de base pour le verbe « Voyager » :

1. (Voiture, Train) est associé au verbe : « voyager » + « en » et
2. (Voiture, avion) est associé au verbe : « Conduire » + « Objet ».

SynDiKATe [HAH01], utilise les techniques de compréhension de textes pour extraire de la connaissance. Comme résultat de l'analyse syntaxique, un graphe de dépendance avec comme

1 - http://www.ldc.upenn.edu/acl/P/P97/P97-1066.pdf
2 - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2735&rank=1

nœud les mots et ses arcs les relations conceptuelles en mappant les mots avec les concepts qui leurs sont équivalent dans la base de connaissance (utilisé comme background).

HASTI [SHA02-a], exploite à la fois l'analyse morpho-syntaxique et sémantique des textes en entrée pour extraire des connaissances lexicales et ontologiques. L'analyse morpho-syntaxique prédit les caractéristiques des mots inconnus et crée des structures de phrase, qui indiquent les rôles thématiques dans la phrase. L'analyse sémantique complète les emplacements vides dans la structure de phrase et amène le processus d'extraction de connaissance conceptuelle à utiliser des modèles sémantiques.

d. Les approches basées sur les Patrons (Pattern matching approaches)

Largement utilisé dans le domaine d'extraction d'information et sont également hérités du domaine d'apprentissage ontologique. Dans ces méthodes une recherche sera effectuée sur les entrées (en général du texte) des mots clés, des patrons qui indiquent certaines relations (par exemple, *hyponymie*). On distingue une variété de patrons (*templates*): syntaxique, sémantique, à but général ou spécifique pour extraire les différents éléments d'une ontologie.

A ce titre, on cite un travail très émergeant, sur le *Pattern matching*, celui réalisé par Hearst (1992) [HER92], qui présente quelques patrons lexico-syntaxiques sous la forme d'expressions régulières pour extraire les relations d'*hyponymie* / d'*hyperonymie* à partir de textes, parmi ces patrons on dicte :

NPSuchas {*NP*,}* (*and* | *or*) *NP*

NP {, *NP*}* {,} (*or* | *and*) *NP*

NP {,} *including* {, *NP*}* (*or* | *and*) *NP*

Ainsi, grâce au troisième parton cité par exemple, les relations d'hyponymie suivante ont était détecté :

```
All common-law countries, including
Canada and England ...
```
$\Longrightarrow$ *hyponym("Canada", "common-law country"),*
hyponym("England", "common-law country")

HASTI [SHA02-a], est un autre système qui utilise des patrons lexico-syntaxique et sémantiques (templates) pour extraire à partir de textes, les relations taxonomiques et non taxinomiques comme hyponymie, méronymie, les rôles thématiques, valeurs des attributs («has-prop» relation) et d'autres relations et axiomes. Un exemple de son modèle lexico-syntaxique est le modèle de l'exception pour extraire les hyponymies :

{*all* | *every*} NP_0 *except* NP_1 {*(and* | ,) *NPi*}*… (*i>1*), *implies* (*sub-class NPi NP0*) (*i* ≥*1*)

Un autre travail est fait par **Sundblad** (2002) [SUN02] dans lequel certains patrons linguistiques sont utilisés pour extraire les relations d'hyponymie et méronymie à partir des questions d'un corpus telle que :

Who is/was X?
What is the location of X?
What is/was the X of Y?
How many X are in/on Y?

Heyer et al, (2001) [HEY01] a proposé deux patrons pour extraire les prénoms et les relations à partir des phrases. Les patrons peuvent être de nature générale, tels que ceux proposées par *Hearst*, **HASTI** *et* **Sundblad** ou spécifiques à un domaine d'application tels que ceux utilisés par *Assadi* (1999) pour extraire des connaissances à partir des textes de planification du réseau électrique. Mais d'autre part, les patrons peuvent être définie manuellement [SUN02], [GAM02] ou peut être extrait (semi) automatique comme dans **Promethee** [FIN99], **AutoSlog**-*TG* [RIL96] et **Crystal** [SOD95].

e. Les approches heuristiques

Outre les approches vues précédemment, l'approche heuristique peut être aussi utilisée. En d'autres termes les méthodes heuristiques ne sont pas indépendantes, elles sont plutôt utilisées pour appuyer et compléter les autres approches. On peut citer quelques exemples dans cet axe de recherche, l'approche **Texte-To-Onto** ; **HASTI** ; **Infosleuth** [HWA99] **et [**GAM02**].**

f. Stratégie d'apprentissage Multiples

La plupart des systèmes, qui apprennent plus d'un type d'éléments d'ontologie, combinent différentes approches. Elles appliquent de multiples stratégies d'apprentissage pour apprendre les différentes composantes de l'ontologie. Elles utilisent différents algorithmes d'apprentissage tels que **Texte-To-Onto** qui utilise des règles d'association et une analyse formelle de concepts ainsi que des techniques de clustering, **WEB → KB** combinant règle d'apprentissage de FOL avec les règles d'apprentissage bayésien, **HASTI** quand à lui, appliquant une combinaison de méthodes logiques, linguistiques basé sur des patrons heuristique et *A.Termier* et al, dans [TER01] combine le clustering statistiques et le clustering sémantique des mots et des documents.

2.4.2. Les tâches d'apprentissage

Les méthodes d'apprentissage peuvent être classées en fonction de la tâche qu'ils accomplissent. La tâche effectuée dans [SUR00] est la classification, dans **HASTI**

[SHA02-a], et **ASIUM** [FAU98] nous retrouvons le clustering, par contre dans Texte-To-Onto [MAE00-b] et dans [RIC92] l'apprentissage des règles et dans **Web→KB** [CRA00] et une analyse conceptuelle formelle dans [BRE01].

La tâche d'apprentissage peut être utilisée pour extraire des connaissances d'entrée ou pour affiner une ontologie. Ci-dessous nous allons parcourir la tâche de clustering, l'une des tâches les plus employés appliqués dans l'apprentissage des ontologies.

- ***Le clustering conceptuel***

Les méthodes de clustering (voir [BIS00]) se distinguent par quatre facteurs adoptés par *Maedeche* [MAE02] : mode du clustering, direction du clustering, mesure de similarité, et la stratégie de traitement.

→ *Le mode de clustering* **:**

Online Vs Offline : Le clustering Online effectue un regroupement au fur et à mesure on parle d'un clustering incrémentale. Par contre Offline l'effectue périodiquement.

Hiérarchique Vs Non Hiérarchique : dans le premier les clusters obtenus sont reliés par des relations hiérarchiques, mais dans le second ces relations sont absentes

Simples Vs multiple : En « Multi-clustering chaque concept peuvent être regroupé en plusieurs clusters ou en d'autres termes, dans le graphe orienté généré par le clustering, chaque nœud peut avoir de nombreux parents et/ou de nombreux enfants.

→ *La direction du clustering :*

Le clustering hiérarchique peut être fait dans l'une des directions suivante : du haut en bas, du bas en haut, ou bien Middle-out (une combinaison des deux).

→ *Les mesures de similarité :*

Les algorithmes de clustering utilisent les mesures de similarité pour calculer la distance sémantique de deux clusters (classes). Dans la littérature, deux grands types de similarité ont été abordés [EAG96] :

-« Similarité Sémantique» (appelée aussi similarité paradigmatique or substitutionnelle)

-« Semantic relatedness » (appelée aussi similarité syntagmatique)

→ *Stratégie de traitement :*

Pour calculer la similarité entre deux clusters, nous pourrons utiliser la stratégie de lien unique « signle link » dans laquelle la similitude entre deux clusters est

la similarité entre deux objets les plus proches en eux. Une autre stratégie est celle de la liaison complète « complete link » dans lequel la similarité de deux clusters est la similarité de leurs deux membres les plus dissimilaires. La troisième stratégie est la similarité moyenne du cluster « Average group » dans laquelle la similarité est la similarité moyenne entre les membres.

2.4.3. Le degré d'automatisation

La phase d'acquisition de connaissances peut être manuelle, semi-automatique ou tout simplement automatique. Comme notre problématique souligne une construction semi-automatique, nous allons donc nous pencher sur les approches d'un certain degré d'automatisation.

HASTI [SHA02-a] et [WAG00] utilise des outils d'acquisition automatique, tandis que **Text-To-Onto** [TOD00] préfère des outils semi-automatique. Mais il existe aussi des systèmes utilisant des méthodes coopératives comme **HASTI**, **ASIUM**.

Dans les systèmes semi-automatiques et de coopération, le rôle des utilisateurs varie selon la méthode adoptée. Il peut proposer une ontologie initiale, et de valider ou modifier les différentes versions proposées par le système [BRE01] ou de sélectionner des patrons dans les relations entre classes [SUR00] ou bien de contrôler les niveaux de généralité et étiqueter nouveaux concepts tel que **ASIUM**, et confirmer les décisions du système comme dans **HASTI**.

2.5. Les Résultats

Cette dimension est concernée par le résultat du processus d'apprentissage et répond à la question : « que va-ton construire et quelles sont ses fonctions ? »

On assiste à des systèmes qui construisent effectivement des ontologies, tandis que d'autres ne font qu'aider et guider l'utilisateur, un expert ou un autre système pour produire une ontologie. Autrement parlant, il existe des systèmes d'apprentissages ontologiques autonomes et d'autres qui sont simplement des modules effectuant une tâche pour aboutir à un ensemble de données intermédiaires qui sera utilisé pour construire l'ontologie. **DODDLE II** [YAM01], **SVETLAN** [CHA00] et [MOI00] sont classés dans la seconde catégorie car la structures initiales pour construire l'ontologie est déjà existante.

Le résultat peut être classé selon trois critères : le type de l'ontologie, sa structure et enfin le langage de représentation de l'ontologie.

2.6. L'évaluation

Il existe jusqu'à maintenant deux approches pour évaluer les méthodes d'apprentissages. L'évaluation des méthodes d'apprentissages : Cette tâche n'est pas assez simple et nous dirons aussi que cette tâche est non triviale, c'est pour cela qu'elle est moins prise en compte dans la littérature, car elle vise à mesurer la justesse des techniques d'apprentissage.

L'évaluation des ontologies résultantes : Ainsi, c'est la méthode la plus courante pour évaluer un système d'apprentissage ontologique. Elle consiste à évaluer (partiellement) leurs ontologies résultantes avec l'une des méthodes suivantes :

- Méthode cité dans [MAE01-b], comparer plusieurs ontologie pour un domaine.
- Méthode de comparaison d'ontologie selon les applications dans lesquelles elles sont utilisées.

En effet, plusieurs approches et systèmes proposent leur propre environnement de tests et d'évaluations, en fonction de leurs applications et le domaine choisi. La majorité des systèmes sont évalués par le calcul du *Recall* et *Precision.* Le Recall est calculé en divisant le nombre de concepts extraits valides acquis sur le nombre total des concepts existants dans d'échantillons d'entrée. Quand à la *Précision* c'est le résultat de la division du nombre de concepts extraits valides sur le nombre total des concepts extraits. Mais ces calculs ne donnent pas une vision réelle de comparaison entre les différentes approches car chacune d'elles peut utiliser des entrées de domaines différents.

Parti II

Apprentissage Ontologique

Techniques et Approches

1. Introduction

Ontologie Learning : Pouvons nous se poser la question légitime si la roue n'est pas réinventée, par la question suivante : « Est-ce que ***l'apprentissage ontologique*** n'est pas simplement une réédition des notions et des techniques déjà existantes sous un nouveau nom ? ».

La réponse est assurément « Non ». Bien que les objectifs d'acquisition[1] de connaissances et de l'apprentissage ontologique[2] (à partir du texte) sont certainement comparables. Les recherches sur les ontologies sont devenues de plus en plus répandues dans la communauté informatique. Les ontologies sont utilisées dans de nombreux domaines tels que le web sémantique, les moteurs de recherche, le traitement du langage naturel, l'ingénierie des connaissances, l'extraction et la recherche d'information, les systèmes multi-agents, le e-commerce, la modélisation qualitative des systèmes physiques, la conception de base de données, les sciences de l'information géographique et les bibliothèques numériques.

2. Classification des sources d'apprentissage

Dans la plupart des cas, il existe déjà des sources de connaissances différentes qui peuvent être incorporés dans un processus d'ingénierie ontologique. Ces sources d'information peuvent être des documents, des bases de données, des taxonomies, des sites web, des applications et d'autres choses.

La question est de savoir comment extraire les connaissances incorporées dans ces sources automatiquement, ou du moins semi-automatiques, et la reformuler dans une ontologie. Alexandre Maedche [MAE03], présente une classification des différentes approches, du domaine d'apprentissage ontologique, selon le type d'entrées : « sources d'apprentissage ».

---

1 - L'essentiel de cette technique c'est qu'elle permet l'acquisition des connaissances explicite, implicitement contenue dans les données (textuelles).

2 - Dans l'apprentissage ontologique, il existe toutefois, un certain nombre d'aspects nouveaux et innovateurs permettent de le distinguer parmi beaucoup de travaux antérieurs d'acquisition des connaissances.

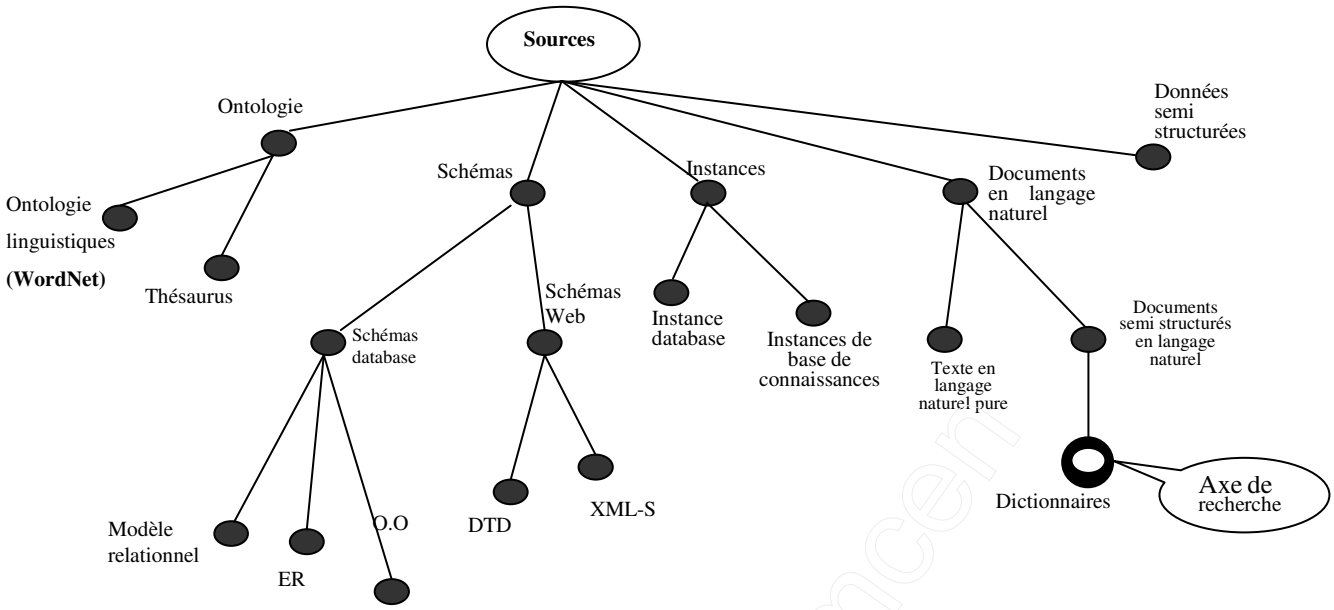

Fig. 34. Classification de *Maedche* : Sources d'apprentissages

La vision de Maedche était que chaque type de source faisait intervenir des traitements et des techniques de transformation différentes qui dictaient dès lors des réutilisations spécifiques. Nous distinguons les approches d'apprentissage à partir de textes :

- de dictionnaires [HER92], [JAN99]
- de bases de connaissances [SUR01],
- de schémas semi structurés [DEI01], [DOA00], [PAP02].
- et de schémas relationnels [JOH94], [KAS99], [RUN02].

Plusieurs techniques sont mises en jeu dans l'apprentissage d'ontologies comme les patrons lexico-syntaxiques, l'extraction basée sur les règles d'association, l'extraction basée sur le *clustering*, l'extraction basée sur le calcul des fréquences et l'extraction basée sur des techniques hybrides.

3. Un Processus d'apprentissage consensuel

D'un point de vue l'ensemble des méthodes énoncées par [RIN04], on peut distinguer six étapes suivantes dans un processus d'apprentissage d'ontologies à partir de textes (qui sont d'une certaine façon ou d'une autre, commun à la plupart des méthodes publiées) :

- Collection, sélection et prétraitement d'un corpus (textes) approprié (outils TAL).
- Découvrez les ensemble des mots (candidats-termes) et expressions équivalentes.
- Validation de l'ensemble (établir des concepts) avec l'aide d'un expert du domaine.

- Découvrir des ensembles de relations sémantiques en concepts.
- Validation des relations et extension des définitions des concepts à l'aide d'un expert du domaine.
- Créer une représentation formelle.

Il ne faut pas croire, que seulement les termes, les concepts et les relations entres eux qui sont importantes, mais aussi le sens des « *gloss* » et la formalisation (axiomes) des concepts ou des relations. Comment mener à bien ces étapes ? Une multitude de réponses peuvent être données. De nombreuses méthodes nécessitent l'intervention humaine avant que le déroulement réel du processus (étiquetage des candidats-termes - apprentissage supervisé, compilation/adaptation d'un dictionnaire sémantique ou des règles de grammaire d'un domaine,...) [RIN04].

Les méthodes non supervisées n'ont pas besoin d'étape préliminaire - cependant, ils ne donnent pas d'assez bon résultats, et le corpus peut empêcher l'utilisation de certaines techniques : par exemple, méthodes d'apprentissage automatique nécessitent un corpus suffisamment large - donc, certains auteurs utilisent l'Internet comme une source supplémentaire. Certaines méthodes nécessitent un prétraitement d'un corpus (par exemple, l'ajout de balises ou étiquette de position, l'identification de la terminaison d'une phrase, ...) indépendant de la langue. Encore une fois, il existe diverses manières d'exécuter ces tâches. Ainsi, de nombreux outils d'ingénierie linguistique ne peuvent être misent en faveur.

4. Méthodes d'extraction des termes (lexicaux)

4.1. Extraction des futurs concepts

L'extraction des termes (futur concept) est une opération pré-requise pour tout apprentissage d'ontologie à partir des textes. Elle implique des niveaux avancés de traitements linguistiques. Les concepts ne sont en général qu'un ensemble de termes. Les termes sont des mots ou suite de mots susceptible d'être retenus comme des entrées (terme, concept) dans une ontologie. Tous les nouveaux travaux convergent vers l'extraction de cette entité. On distingue les méthodes linguistiques basées sur des règles syntaxiques, les méthodes statistiques basées sur les fréquences de séquences et les méthodes hybrides.

Plusieurs modèles sont issues de ces 3 approches. Par exemple la méthode du dictionnaire qui s'appuie sur une ressource externe qui retienne les mots et expressions figées voir semi-figées susceptibles d'être rencontrées dans un texte du domaine, ils sont les plus

utilisées dans l'identification des concepts. La méthode des cooccurrences permet de créer un lexique par la répétition des formes présentes dans un texte. La méthode des segments répétés se base sur la détection de chaînes constituées de fraction fréquentes dans le même texte. La méthode des bornes travaille avec des délimiteurs. [TUR01]

4.2. Outils d'extraction

Les méthodes n'agissent pas directement sur les corpus bruts (textes) mais utilisent un « *shallow text processing* » basé sur des études de traitement des textes peu profonde (TAL), et d'analyses syntaxiques ou tout autres traitement fournissant une sortie normalisée et exploitable par des algorithmes d'apprentissage automatiques. Ces outils empruntés au TAL, sont conçus avec plusieurs éléments chacun d'eux est dédié à une tâche bien précise :

- Tockenizer : Extrait toutes les unités lexicales d'une phrase ou d'un texte.
- Lemmatiseur : PoS tagger pour identifier la classe d'une unité : Nom, Verbe,…
- Name Entity : Reconnaisseur d'entité et décider si l'entité est une personne, un matériel, une date, un horaire, un nom de société, etc.

4.2.1. Méthodes statistiques

Une méthode très répandu dans la recherche d'information (IR) est le calcul de la fréquence d'occurrence d'un terme dans un corpus ou dans un texte. Mais très vite, d'autres techniques émergent et prouvent leurs efficacités, comme la méthode issue de la recherche d'information et basée sur la mesure Tfidf « *Frequency Term Inverted Document Frequency*». [MAE03] :

- *Term Frequency* Tf (t, d): fréquence d'occurrence du terme « t » dans le document « d » ∈ D (corpus, ensemble de document).

- *Documents frequency* df (***t***) : le nombre des documents dans le corpus D dans lesquels apparaît le terme.

- Inverse *Documents frequency* idf(***t***) : idf(t) = log (|D| / df(t)), où |D| : le nombre total de documents dans un corpus D. Un mot qui apparaît dans un peu de documents possède une grande valeur au calcul de la mesure idf(*t*), à l'inverse de celle qui a une valeur haute de tf*idf est reconnue comme un terme candidat et pertinent pour le document. Alors tfidf du terme ***t*** pour un document d est :

- tfidf (t,d) = tf(t,d) * log (|D| / df(t)).

- Corpus Frequency cf(***t***) : est le nombre d'occurrence du terme « t » dans tous les documents du corpus D. C'est clair que df(***t***) ≤ cf(***t***) et ∑ tf(***t,d***) = cf(***t***).

4.2.2. Méthodes à base de dictionnaires (notre axe de recherche)

Il existe des approches qui préfèrent des ressources issues des dictionnaires comme un outil d'amorce pour repérer les termes pertinents ou acquérir directement des termes contenus dans ces dictionnaires qui constituent une mine très riche d'information lexicale et sémantique (au cas où ils existent). Il offre une stabilité pour un bon amorçage du processus d'extraction.

Un souci majeur pour une exploitation facile se situe dans leur transformation en des représentations facilement exploitable par des machines. Kiez, dans [KIE00], a présenté des travaux pour la construction d'ontologie de domaine (assurance) ainsi que Maedche et Staab dans [MAE03] pour la télécommunication.

4.3. Extraction de relation

Plusieurs ressources lexicales sont utilisées pour relever les relations sémantiques entre les concepts, on cite alors : les dictionnaires, les ontologies (existantes), les patrons syntaxiques, la notion de collocations de termes ou bien la combinaison de toutes ces ressources.

A titre d'exemple, dans les partons lexico-syntaxiques (hérités du TAL), on trouve les relations sujet-verbes, verbes-objet, ou le groupement des termes selon leurs cooccurrences avec le verbe qui permettra d'acquérir par la suite des relations sémantiques.

4.4. Relations taxonomiques :

Deux grandes approches émergent dans l'apprentissage ou l'acquisition des taxonomies [MAE03] :

- Approches moyennant le *clustering* : Basé sur les hypothèses distributionnelles, ce sont des approches statistiques (groupement des termes et calcul de similarité,…).
- Approches utilisant les patrons lexico-syntaxiques : se sont des approches symboliques pour détecter les relations d'hyponymie proposé dans [HEA92].

→ ***Clustering et les relations***

Dans la famille des méthodes de regroupement non supervisées, on distingue les méthodes agglomératives (plus proche voisin, distance maximum...) qui regroupent des clusters existants selon des mesures de similarité et des méthodes de divisions (bisection k-means).

[CIM04-b] expose un aperçu de plusieurs approches : Il commence avec les premiers travaux liés au *clustering,* citant tout d'abord les travaux de Hindle [HIN90], où les noms sont

regroupés selon leurs apparitions comme sujets ou objet de verbes similaires. Quand à Pereira [PER93], il présente une approche du « *Top-down clustering* » pour bâtir une taxonomie non étiquetée de noms (Les relations de la taxonomie non étiquetée). Par contre l'approche ittérative « bottom up of clustering » a été présenté dans [FAU98], privilégiant ainsi la fréquence des mots apparaissant dans un même contexte. Cette méthode nécessite un suivi manuelle (méthode supervisée), par conséquent elle est n'est pas privilégiée par rapport aux méthodes (semi) automatiques. Dans [BIS00], Bisson et al, fournit un outil complet assistant le concepteur dans le domaine de construction d'ontologie, en utilisant une comparaison des distances de similarités (distances sémantiques) afin d'arriver à un clustering « bottom up ». Des études assez récente dans [CIM04-a], Viz utilise une *FCA* (*Formal Concept Analysis*), analyse des concepts formelle pour grouper les concepts et d'en extraire une hiérarchie à partir des textes.

→ ***Patrons lexico-syntaxiques et les relations***

Les patrons lexico-syntaxiques fournissent une relation entre des concepts d'un domaine. Ces relations ne sont repérées que lorsque les concepts appartiennent à la même phrase. Deux axes supplémentaires se sont développés :

- Dans la littérature linguistique, des patrons relatifs aux relations hiérarchiques (hyperonymie, définition, méronymie – partie de –) ou de synonymie, ont été capitalisés avec l'espoir de pouvoir les réutiliser sur tout type de textes. L'état de l'art montre que ces patrons sont plus ou moins adéquats et doivent toujours être ajustés.

- Dans les recherches de l'extraction d'information, de nouveaux patrons sont redéfinis pour repérer des relations spécifiques au domaine étudié.

En 1992, Hearst a proposé une approche pour extraire des relations d'hyponymies à partir d'une encyclopédie scolaire « Grolier », cette méthode utilise des patrons lexico-syntaxiques manuellement capturés à partir d'un corpus. [CHA99] donne une approche pour apprendre la relation « Part of », mais ceux [VEL01] manipule des techniques heuristiques. [MOR98] développe Prométhée pour palier à la lourdeur de la méthode Hearst (confection manuelle des patrons). C'est un outil d'apprentissage automatique pour l'extraction des patrons lexico-syntaxiques relatifs à la spécification conceptuelle des relations.

Conclusion

Dans ce chapitre, nous avons fait un passage horizontal sur les différentes techniques, approches et outils de base utilisées dans la création d'une ontologie, en générale. Le point de rencontre commun à tous les systèmes étudiés est la réutilisabilité et le partage de l'ontologie.

L'extraction de connaissances ou communément parlant « apprentissage d'ontologies » a pour but la construction semi-automatique d'ontologie. Les méthodes de construction d'ontologies à partir des documents semi structuré favorisent souvent l'étude du texte, proprement dit, que ce soit selon une approche statistique, symbolique ou linguistique.

Le dernier chapitre va surtout mettre en lumière l'approche de la solution adoptée à la construction d'une ontologie lexicale en prenons l'ontologie WorNet comme modèle de travail, et en utilisant comme source d'entrée pour l'apprentissage, les données d'un dictionnaire arabe « Al ghannye ».