Construction automatique des ontologies à partir des textes arabes

BENABDALLAH Ali
Université de Tlemcen

Résumé

Dans cet article nous proposons une approche de construction automatique d'ontologie à partir des textes arabes. Cette dernière se base sur trois étapes : d'abord on commence par l'application d'un ensemble de traitements préliminaire sur un corpus textuel, en suite on passe à l'étape d'extraction des termes qui se base sur l'hybridation de deux méthodes : une méthode statistique qui s'appelle « *les segments répétés* » et une autre méthode linguistique : *« méthode du dictionnaire »*. La dernière étape se focalise sur l'extraction des relations sémantiques à partir du texte en se basant sur un ensemble de patrons syntaxiques. En fin, on terminera par l'ajout des relations sémantiques à partir de l'ontologie lexicale WordNet Arabe.

1. Introduction

Parmi les sous-domaines de l'ingénierie des ontologies on distingue la construction d'ontologies à partir de documents textuels. Ces ontologies peuvent êtres utilisées dans plusieurs domaines de la recherche d'informations tels que l'annotation sémantique de ressources, l'indexation automatique des documents, les résumés automatiques des textes,…etc**.**

L'utilisation des textes dans le processus de construction des ontologies est justifiée par deux raisons : d'abord les textes sont souvent porteurs de connaissances stabilisées et partagées par des communautés de pratiques. En outre, même s'ils ne les

remplacent totalement, les textes sont plus facilement disponibles que les experts qui manquent de temps pour participer au processus de construction [Mon08]. Mais il convient de noter que le recours aux experts est nécessaire surtout dans l'étape de la validation des ontologies obtenues.

Une telle ontologie est constituée d'un ensemble de *concepts* à la fois organisés hiérarchiquement et structurés par des *relations* liant ces concepts. C'est-à-dire que chaque processus de construction d'ontologies doit passer par deux étapes importantes: une première concerne la conceptualisation et une deuxième concerne l'extraction des relations sémantiques.

L'objectif de notre approche est de construire une ontologie à partir des textes arabes d'une manière automatique. En se basant sur l'extraction des termes et des relations sémantiques à partir du texte et d'une base de données lexicale.

La suite de l'article est structurée comme suit : La première partie se focalise sur la préparation du corpus, en suite pour extraire les candidats-termes, on va adopter une méthode statistique *« les segments répétés »* suivie par une méthode linguistique *« méthode à base de dictionnaire »*. Et dans la dernière partie, on va utiliser deux méthodes d'extractions des relations, une méthode des patrons syntaxiques qui se base sur le texte et une autre méthode pour mettre à jour les relations de l'ontologie à partir de la ressource *WordNet Arabe.*

2. Etat de l'art

Dans le cadre de la construction des ontologies à partir des textes, il existe plusieurs travaux dans la langue anglaise, française, espagnole, …etc. Parmi ces travaux nous avons

privilégié des systèmes opérationnels, et disponibles sur la toile, qui sont cités dans [Mon08] :

Text2Onto [Cim05] est un outil conçu pour construire des ontologies à partir de textes de manière complètement automatique, il existe des versions anglaises, françaises et espagnoles. Text2Onto est codé en java et est composé de modules qui extraient à partir des textes des concepts, des relations entre ces concepts (relation d'équivalence, hiérarchiques, etc.) et des instances de concepts. Text2Onto utilise la ressource WordNet pour construire une hiérarchie de concepts.

OntoGen [For06], qui est codé en .net, implémente une approche semi-automatique pour la construction d'ontologies de thèmes (topic ontologies) à partir de collections de documents. C'est un outil interactif qui suggère à l'expert du domaine des concepts sous la forme de classes de documents, propose une dénotation et leur associe automatiquement des instances (les documents).

Terminae [Aus08] est une méthode supportée par un logiciel qui propose de guider l'ontologue dans la conception de l'ontologie. Terminae s'appuie sur les résultats des outils de traitement automatique des langues (extracteur de termes, concordancier, détecteur de synonymie, analyseur syntaxique) pour extraire des éléments qui peuvent aider l'ontologue à construire l'ontologie selon ses objectifs de modélisation.

Ces systèmes de construction des ontologies ne supportent pas les documents textuels ***arabes***. Parmi les travaux sur la langue arabe, on trouve [Maz12] qui a proposé une approche de construction des ontologies à partir des textes arabes en utilisant deux techniques statistiques d'extraction des termes ; la première s'appelle *« repeated segments »* pour

identifier les termes pertinents qui dénotent les concepts associés au domaine et la deuxième technique s'appelle «co-occurrence» pour lier ces nouveaux concepts extraits à l'ontologie à travers des relations hiérarchiques ou non hiérarchiques.

3. Approche proposée

Pour construire notre ontologie à partir d'un corpus textuel arabe, on a adopté un processus d'extraction de *concepts* et de *relations* à partir des documents textuels en se basant sur l'ontologie lexicale WordNet Arabe. Ce processus se résume en trois grandes étapes ; la première est la formation du corpus de domaine. Cette étape est très importante, car la qualité de l'ontologie obtenue dépendra de la qualité du corpus traité et de la manière de préparer ce corpus qui doit couvrir entièrement le domaine traité. La deuxième étape est l'extraction de candidats-termes qui sont des future-concepts qui composent l'ontologie. Cette étape se déroule en trois phases : d'abord on va extraire tous les segments existants dans le texte, en suite on va appliquer le filtre de poids sur ces segments pour éliminer les termes non pertinents, puis on va éliminer les termes représentants des faux-concepts en se basant sur la ressource WordNet Arabe. Dans la troisième étape nous reliant les concepts obtenus à travers des relations sémantiques extraites à partir du texte en utilisant la méthode des patrons syntaxiques. On finira ce travail par la mise à jour de l'ontologie, en insérant de nouvelles relations sémantiques à partir de la base de donnés lexicale *WordNet arabe*.

3.1 Préparation du corpus

Dans un processus de construction d'ontologies à partir de textes, l'étape de constitution et de préparation du corpus est

à la fois primordiale et délicate. Puisque, le corpus est la source d'information essentielle pour tout le processus de construction [Bou03].

Les questions qui se posent dans la conception de tout corpus comprennent : le type de corpus (un corpus «spécialisé» est un corpus contenant des textes sur un sujet lié à un domaine de connaissance par exemple, dans notre cas, *la langue arabe*) , l'adéquation pour le projet visé, la possibilité de réutiliser ces corpus, la taille (nombre de mots), la représentativité (c'est-à-dire, la variété de textes, d'auteurs, de sources, etc.), l'utilisation de textes complets ou d'échantillons, …etc.[Mar03]

Après avoir construire le corpus brut, on doit le préparer pour être prés à traiter durant le reste du processus de construction. L'utilité de la préparation du corpus est d'enlever l'ambiguïté, réduire le nombre de transactions et d'adapter le corpus avec l'objectif de l'étape suivante «extraction de candidats-termes ». Cette étape de préparation se décompose en plusieurs sous-étapes qui sont :

- Normalisation : transforme le document dans un format standard plus facilement manipulable [Maz12]. Avant la lemmatisation, le document est normalisé comme suit:

  - Suppression des caractères spéciaux et des chiffres, par exemple : ٢,٣,٤,٥,(,[,+, »…
  - Suppression des mots latins: les caractères latins sont détectés par leurs graphiques.
  - Suppression des lettres isolées: les mots à une seule lettre en arabe et les abréviations. par exemple : *la numérotation* (paragraphe « B » الفقرة "ب"), *les*

abréviations de la date (هـ : تاريخ هجري , م:تاريخ ميلادي), *les voyelles* (حروف العلة: "و"), *...etc.*

- suppression des signes de voyelles, qui sont écrites sous la forme de signes diacritiques placés au-dessus ou au-dessous des lettres, par exemple : ً

Suppression des mots vides : consiste à éliminer tous les mots non significatifs, en comparant chaque mot reconnu avec un des éléments de la liste des mots vides :"stop-liste". Il s'agit d'une liste de tous les mots d'outils, liaison et d'articulation (pronoms(**الضمائر المنفصلة**) , prépositions (**حروف العطف**), conjonctions(**حروف الجر**) , etc.) [Maz12]. Exemple: إلى ، انه ، هذا ، هذه ، بين ، بعد ، مع، الذي ، عن ، التي ، في ، أن ، من ، على ... لم ، ما ، منذ

- Lemmatisation : C'est une tâche délicate du fait que l'arabe est une langue flexionnelle et fortement dérivable ; l'absence des diacritiques crée une ambiguïté et donc exige des règles morphologiques complexes, de plus la capitalisation n'est pas employée dans l'arabe ce qui rend difficile l'identification des *noms propres*, des *acronymes* et des *abréviations*. Pour résoudre l'ambiguïté, [Alj02] a montré que la lemmatisation légère (approche basée sur suppression de suffixe et de préfixe) surpasse celle basée sur détection de racine dans le domaine de la recherche d'information.

Dans notre approche nous avons considéré la *lemmatisation légère* qui consiste à déceler si des préfixes ou suffixes ont été ajoutés au mot. Nous utilisons la liste de préfixes et de suffixes proposée par [Dar03] voir Tableau1. Cette liste regroupe les préfixes et les suffixes les plus utilisés dans la langue arabe tels que les conjonctions, préfixes verbaux, pronoms possessifs, pronoms compléments du nom ou suffixes verbaux exprimant le pluriel, etc.…

Préfixes							
والـ	بتـ	وتـ	بمـ	كمـ	للـ	فيـ	لا
فالـ	يتـ	ستـ	لمـ	فمـ	ليـ	وا	با
بالـ	متـ	نتـ	ومـ	الـ	ويـ	فا	
Suffixes							
ات	وه	ته	هم	ـه	ين	ا	نا
وا	ان	تم	هن	تك	يه	ـي	ها
ون	تي	كم	هما	تما	كما	كن	

Tableau 1 : Liste des préfixes et suffixes les plus fréquents en arabe

Exemple : Afin de montrer les démarches de notre approche sur un texte réel, on a choisi à partir de notre corpus l'échantillon de texte suivant :

٣- السيارة (la voiture) هي مركبة آلية تتكون من أجزاء ميكانيكية تعمل كل هذه الأجزاء بصورة متناسقة بحيث تؤدي إلى تحريك هذه المركبة، وتعتبر السيارة وسيلة النقل الأكثر انتشارا في عصرنا الحالي, بحيث أنها تساهم في دفع عجلة الاقتصاد إلى الأمام جنبا إلى جنب مع السكة الحديدية. و من بين أنواع السيارات: السيارة

Figure 1 : Echantillon de notre corpus textuel arabe

L'application de la normalisation sur cet échantillon supprime quelque caractère spéciaux comme : **و-, :** ,**...** ,٣, , **(** ,
et aussi les caractères latins comme « **la voiture** ».
Les mots vides qui seront supprimés sont : إلى ، من، أنها، مع ، في
هذه ، هي، بين ،بحيث،.
La lemmatisation consiste à comparer les préfixes (cités dans le tableau 1) avec les débuts de chaque mot du texte et les suffixes avec les fins de chaque mot et supprimer ces préfixes et ces suffixes.

Par exemple : السيارة → **سيارة**; المركبة → **مركبة**; عصرنا → **عصر** ; انتشارا → **انتشار** .

3.2 Extractions des candidats-termes

Après la préparation du corpus, nous passons à l'étape d'extraction des éléments de l'ontologie.

Parmi les méthodes utilisées pour l'extraction de termes, on trouve trois grandes familles : (cités dans [TUR01])

- Les méthodes linguistiques (basées sur des règles syntaxiques)

Exemples : méthode à base de dictionnaire, méthode des bornes, méthodes des schémas syntaxiques,…

- Les méthodes statistiques (basées sur les fréquences de séquences)

Exemples : méthode des segments répétés, méthode des cooccurrences, …

- Les méthodes hybrides (basées sur les deux méthodes précédentes)

Pour notre cas, nous avons choisi de travailler selon une méthode hybride.

Notre approche se base sur l'hybridation d'une méthode statistique : « *méthodes des segments répétés* » avec une méthode linguistique qui s'appelle « *méthode du dictionnaire* ». Notre processus d'extraction de termes se déroule en trois phases :

Phase 01 : Extraction de tous les segments du texte

La méthode des « *segments répétés* » s'appuie sur la détection de chaînes constituées de morceaux existant plusieurs fois dans le même texte. Il s'agit d'une technique statistique pour l'extraction d'informations à partir des textes. La répétition des segments dans un texte indique que ceux-ci peuvent être utilisés pour désigner les concepts du domaine du

corpus. Un segment de texte peut contenir un ou plusieurs mots séparés par des espaces ou des signes de ponctuation (un terme peut être simple ou complexe, contenant un nombre fini de mot). La méthode identifie tous les segments répétés dans une fenêtre de trois mots dans la même phrase (le nombre de trois est choisi selon le principe qu'un terme désignant un concept contient un maximum de trois mots). A la fin de cette phase, les redondances sont éliminées en retirant les segments inclus dans les autres avec le même nombre d'occurrences. [Maz12]

Tous les segments obtenus dans cette phase seront ensuite filtrés, pour éliminer les segments non pertinents (moins fréquents), et ne garder que les termes représentant le domaine du corpus traité.

Phase 02 : Application du filtre de poids

La méthode des « *segments répétés* » est basée sur la préposition suivante : « *Un terme pertinent est utilisé plusieurs fois dans un texte spécialisé ».* Pour cette raison, nous utilisons le « *filtre de poids »* [Her06] pour sélectionner des termes avec suffisamment de poids. Le poids est mesuré par la fréquence totale d'un terme : c'est le nombre total d'occurrences de ce mot dans le corpus. Si cette fréquence est supérieure à un seuil : f_{min} (*un seuil indiquant la pertinence*), le terme sera gardé pour la prochaine étape du processus de construction, sinon ce terme sera ignoré.

Phase 03 : Filtrage des termes à base de dictionnaire

Malgré l'application du filtre de poids sur les segments détectés dans la première phase, il reste toujours des segments non désirés (c-à-d qui ne représentent pas des vrais concepts). La méthode du dictionnaire est une méthode d'extraction des

termes s'appuyant sur une *ressource externe* qui consigne les mots et expressions susceptibles d'être rencontrés dans un texte du domaine [TUR01]. Cette phase consiste à comparer successivement les segments qui restent de la phase précédente (segment d'un seul mot, de deux mots, ou de trois mots) avec tous les termes existants dans la base de données lexicale *WordNet Arabe*, pour valider les candidats termes et éliminer les segments représentant des faux concepts.

A la fin de cette phase, les termes validés à partir du *WordNet* seront considérés comme des nouveaux concepts de l'ontologie.

Résultats

Segments d'un seul mot		Segments de deux mots	Segments de trois mots
سيارة	سيار	سيارة مركبة	سيارة مركبة آلية
مركبة	سياحية	مركبة آلية	مركبة آلية تتكون
آلية	شاحنة	آلية تتكون	آلية تتكون أجزاء
تتكون	حافلة	تتكون مجموعة	تتكون أجزاء ميكانيكية
أجزاء	....	سيارة سياحية...	.....

En appliquant les trois phases précédentes sur l'échantillon de texte de la «*figure 1*» on obtiendra les résultats suivants:

- Dans la première phase, un grand nombre de segment seront extraits. Par exemple :

Tableau 2 : Liste des segments extraits à partir du texte de la figure 1

- Dans la phase de filtrage (filtre de poids), on calcule pour chaque segment sa fréquence d'apparition dans le corpus, en suite on supprime tous les segments ayant une fréquence

inférieure à la fréquence minimale f_{min} . (f_{min} est choisie d'une manière *empirique*, et dépend de la taille du corpus).

- La troisième phase consiste à diviser l'ensemble des segments restants en deux parties : les segments validés et les segments rejetés (ou non valides), en se basant sur la ressource *WordNet Arabe*. Les résultats de cette phase sur l'exemple précédant sont :

Exemples de segments validés (les concepts) : سيارة, سكة حديدية ..., وسيلة نقل, تحريك, مركبة

Exemples de segments rejetés : أنواع_سيار ...آلية_تتكون_أجزاء, انتشار_عصر ,

3.3 Extraction des relations sémantiques

La majorité des travaux existants dans le domaine de l'extraction des relations sémantiques à partir de texte se basent sur l'utilisation des patrons (ou marqueurs) lexico-syntaxiques. Ces travaux se base seulement sur le corpus textuel et une liste de patrons syntaxiques, mais il existe d'autres travaux qui utilisent des ressources externes pour extraire les relations, par exemple : [GIR02] a proposé une technique semi-automatique d'extraction des patrons syntaxiques en utilisant les relations existantes dans *WordNet Anglais*.

Dans notre approche, nous avons choisi de travailler avec la méthode des patrons syntaxiques dans la première phase et dans la deuxième phase on ajoutera des relations sémantiques (hiérarchiques) entre les concepts à partir du *WordNet Arabe*.

3.3.1 Extraction des relations à partir du texte (Patrons syntaxiques)

Dans les méthodes d'extraction de relations à partir d'un corpus textuel, le principe est de définir dans un premier temps, un ensemble de listes de patrons lexico-syntaxiques (une liste

pour chaque relation). En suite, ces patrons seront projetés sur le corpus de texte (non préparé) afin de repérer les instances des relations. La construction des patrons lexico-syntaxiques est alors une étape préliminaire afin de découvrir les relations dans un corpus. Par exemple :

Relations	Patrons syntaxiques
Hyperonyme et *Hyponyme (is-a)*	نوع من ،صنف من، هو، هي، هم، وغيره من ، .....
Méronyme (part-of)	جزء من ، تتكون من ، تنقسم الى ، تتألف من ، .....
Antonymie	ضد، عكس.....
....	

Tableau 3 : Liste de quelques patrons syntaxiques de langue arabe

A cause de la morphologie spécifique de la langue arabe telle que la vocalisation et l'agglutination, les listes des patrons syntaxiques doivent regrouper toutes les formes morphologiques susceptibles d'être rencontrées dans les textes arabes.

Le tableau suivant résume quelques relations sémantiques extraites à partir du texte précédant (*Figure 1*) par la méthode des patrons syntaxiques :

Les patrons syntaxiques détectés	Type de la relation	Les segments de textes	Concept1-Relation-Concept2
هي	Hyperonymie (is a)	«السيارة **هي** مركبة»	سيارة -is a– مركبة
تتكون من	Méronymie (part of)	«مركبة آلية **تتكون من** أجزاء ميكانيكية»	أجزاء ميكانيكية –part of- مركبة آلية
...	...	...	...

Tableau 4 : Liste de quelques relations extraites à partir

des patrons syntaxiques du texte de la figure 1.

4. L'ajout des relations sémantiques à partir de WordNet

Dans la phase précédente, on a relié les concepts de l'ontologie avec des relations sémantiques extraites à partir du texte. Cette phase consiste à enrichir l'ontologie obtenue, en ajoutant des relations sémantiques entre les concepts qui n'ont pas été déjà reliés dans la phase précédente afin de mettre à jour l'ontologie.

Notre processus de mise à jour de l'ontologie se déroule comme suit :

- Pour chaque terme T (de notre ontologie) relié avec des termes $T_{i\ (i=1,\ 2,\ 3\ldots)}$ par des relations extraites dans la première phase, on cherche dans la base de données lexicale *WordNet Arabe* la liste des termes $T_{j\ (j\ :\ 1,\ 2,3\ldots)}$ qui peuvent être reliés à ce terme T avec une des relations sémantiques $R_{j\ (j\ :\ 1,\ 2,3\ldots)}$ respectivement.

- On compare la liste des termes T_j avec la liste des termes T_i et on supprime les termes communs de la liste T_j

- Pour chaque terme qui reste dans la liste T_j on vérifie son existence dans notre ontologie :

  - Si ce terme existe déjà dans l'ontologie, alors on ajoutera sa relation R_j avec le terme T à l'ontologie.
  - Sinon, on n'ajoutera ni ce *terme* ni sa *relation* R_j à l'ontologie.

Résultats : Parmi les relations sémantiques qui peuvent être ajoutées à l'ontologie de l'échantillon précédant à partir de *WordNet Arabe* on trouve :

Concept 1	Type de la relation	Concept 2
سيارة	Hyperonyme de (Is-a)	وسيلة نقل
حافلة	Hyperonyme de (Is-a)	وسيلة نقل
شاحنة	Hyperonyme de (Is-a)	وسيلة نقل
…	…	…

Tableau 5 : Liste de quelques relations extraites à partir de WordNet Arabe

5. Conclusion et perspectives

Dans cet article nous avons proposé une approche automatique de construction d'une ontologie à partir des textes arabes. Cette dernière se base sur l'extraction des concepts et des relations sémantiques à partir de deux ressources : un corpus de textes arabes et une ressource externe qui est la base de données lexicale *WordNet Arabe*. Notre processus de construction commence avec une étape préliminaire qui est la préparation du corpus passant par la normalisation, la suppression des mots vides et la lemmatisation. En suite pour extraire les candidats-termes, on a utilisé une méthode statistique *« les segments répétés »* suivie par l'application d'un filtre de poids et une méthode linguistique *« méthode à base de dictionnaire »*. Et pour relier les nouveaux concepts extraits on a adopté deux méthodes, une méthode des patrons syntaxiques qui se base sur le texte et une autre méthode pour mettre à jour les relations de l'ontologie à partir de la ressource *WordNet Arabe.*

De nombreuses perspectives sont offerts en se basant sur ce modeste travail ; d'abord, on propose de faire une désambigüisation des sens au niveau de l'étape d'extraction des relations sémantiques à partir de la base de donnés lexicale *WordNet Arabe*, et plus précisément dans le choix des synsets contenant les termes à relier avec des relations. Cette

désambigüisation permettra de choisir le synset qui correspond bien au sens de ces termes dans le texte.

Et comme deuxième perspective : dans l'étape de validation des candidats termes à partir de WordNet Arabe, on propose d'ajouter l'intervention d'un expert en langue arabe pour valider les termes qui ne figurent pas dans la base de données *WordNet arabe* et qui semblent corrects pour lui.

Références

[Alj02] ALJLAYL Mohammed and FRIEDER Ophir.(2002) *«On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach »,* In 11th International Conference on Information and Knowledge Management (CIKM), November 2002, pp. 340-347.

[Aus08] AUSSENAC-GILLES N., DESPRES, S. & SZULMANS. (2008). *«Bridging the Gap between Text and Knowledge: Selected Contributions to Ontology learning from Text »*, chapter The Terminae Method and Platform for Ontology Engineering from Texts, p.A paraitre. IOS Press.

[Bou03] BOURIGAULT Didier et AUSSENAC-GILLES, Nathalie . (2003). *«Construction d'ontologies à partir de textes».* Conférence sur le traitement automatique des langues (TALN), France, Juin 2003.

[Cim05] CIMIANO, P. & VOLKER, J. (2005). *«Text2onto - a framework for ontology learning and data-driven change discovery».* Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, p.227–238, Alicante, Spain :Springer.

[Dar03] DARWISH, K. : *«Probabilistic Methods for Searching OCR-Degraded Arabic Text»,* Doctoral dissertation, University of Maryland, 2003

[For06] FORTUNA, B., GROBELNIK, M. & MLADENIC, D. (2006). *«Semi-automatic data driven ontology construction system ».*

In Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia.
[GIR02] GIRJU et al. (2002). GIRJU, R., MOLDOVAN, D. *« Text mining for causal relations ».* In 15sup th international Florida Artificial Intelligence Research Society Conference, pp: 360-364.
[Her06] HERNANDEZ Nathalie *« Ontologies de domaine pour la modélisation du contexte en recherche d'information »* Thèse de Doctorat à l'Université Paul Sabatier France 2006.
[Mar03] MARSHMAN Elizabeth *«Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie»* Janvier 2003, "Observatoire de linguistique Sens-Texte" (OLST) de l'Université de Montréal.
[Maz12] MAZARI, A.C., ALIANE, H. and ALIMAZIGHI, Z. (2012). *«Automatic construction of ontology from Arabic texts»* ICWIT'2012 *« International Conference on Web and Information Technologies , Sidi Bel Abbes, Algeria »* ICWIT, volume 867 of CEUR Workshop Proceedings, page 193-202.
[Mon08] MONDARY,T., DESPRES, S., NAZARENKO, A. SZULMAN, S. *« Construction d'ontologies à partir de textes : la phase de conceptualisation »* IC2008 : « 19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008), Nancy : France » LIPN - UMR 7030 Université Paris 13 – CNRS.
[TUR01]TURENNE Nicolas *« Etat de l'art de la classification automatique pour l'acquisition de connaissances à partir de textes »,* UMR INRA-INAPG – Biométrie et Intelligence Artificielle (BIA), Technical Report, INRA, 2001.