Chapitre 3 : État de l'art

Extraction des termes et des relations à partir de textes

1. Introduction

De nombreuses méthodes d'analyses linguistique ou statistique existent, issues des recherches de linguistes et d'informaticiens. Parmi ces méthodes, certaines sont susceptibles d'être appliquées à des textes, les processus utilisés alors, sont dans la majorité des cas itératifs, permettant un enrichissement de la connaissance. Ces approches peuvent être utilisées dans différentes applications. Ce qui nous intéresse dans ce chapitre c'est l'extraction des termes et des relations dont l'objectif est la construction d'ontologies à partir de textes. Nous allons présenter dans ce chapitre les différentes approches et outils utilisés dans l'extraction des termes et des relations**.**

2. La construction d'ontologies à partir de textes

2.1. Définitions

Avant d'aborder les étapes de création d'ontologies, donnons quelques définitions pour mieux comprendre les termes utilisés dans les différentes étapes.

2.1.1. Mot

Nous entendons par mot toute séquence de caractères délimitée par deux séparateurs (blanc ou autre marqueur de séparation, tel que la ponctuation) ou unité minimale de signification appartenant au lexique appelé lexème. Si une séquence de caractères de ce type se répète 2, 3, n fois, elle correspond à 2, 3 ou n mots, mais constitue un seul et même item (**Azé & Heitz, 2004**).

Exemple : La suite de caractère « **كتاب** » est un mot, il peut avoir plusieurs sens (livre, missive, etc).

2.1.2. Terme

Un terme est une expression possédant un sens unique pour un domaine particulier (**Azé & Heitz, 2004**). Dans le Coran considéré comme Corpus d'un domaine donné, le mot **« كتاب »** devient un terme par rapport au domaine, il a une signification unique, c'est une des écritures saintes révélées par l'Archange Gabriel aux différents prophètes (PSSE).

2.1.3. Concept

Un concept est « *l'idée générale et abstraite que se fait l'esprit humain d'un objet de pensée concret ou abstrait, et qui lui permet de rattacher à ce même objet les diverses perceptions qu'il en a, et d'en organiser les connaissances* » Larousse[30]. « *Faculté, manière de se représenter une chose concrète ou abstraite; résultat de ce travail; représentation* » CNRTL[31].

Ainsi si « **كتاب** » est un terme dans le domaine cité plus haut, le concept est l'idée, la perception que l'on se fait de ce terme.

2.1.4. Terminologie

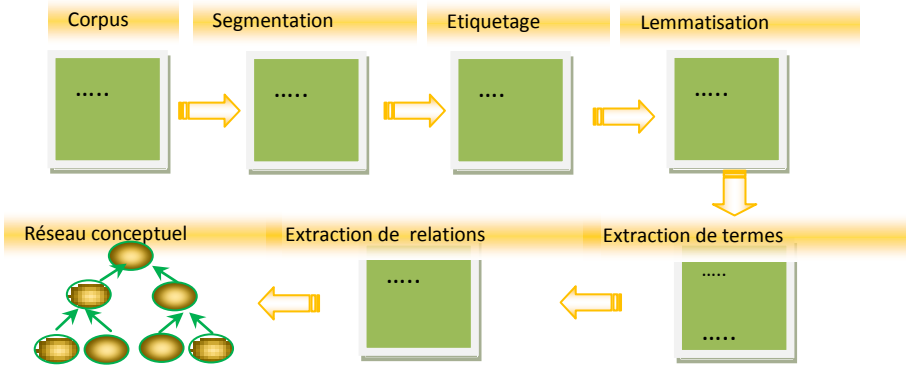
Une terminologie représente l'ensemble des termes particuliers à une science, un domaine ou un art (**Larousse, 1988**), à un groupe de personnes ou à un individu (**Office de la langue française, 2001**). La terminologie, considérée comme science, s'intéresse donc, au recensement des concepts d'un domaine et des termes qui le désignent pour faciliter l'échange de connaissances dans une langue et d'une langue à l'autre (**Baneyx, 2007**).

2.1.5. Ontologie

Une ontologie comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts (entités, attributs, processus), leurs définitions et leurs interrelations. On appelle cela une conceptualisation **(Baneyx, 2007**). Une définition plus détaillée de l'ontologie a été donnée au chapitre2.

2.2. Les étapes de construction d'une ontologie à partir de textes

La création d'une ontologie à partir de textes, nécessite une suite logicielle outillant une méthodologie globale de construction **(Zaidi & al, 2010b).**



[30] http://www.larousse.fr/dictionnaires/français
[31] Centre National de Ressources Textuelles et lexicales(http://www.cnrtl.fr)

Figure 9: Etapes pour la construction d'ontologies à partir de textes

La figure 9 donne une idée sur les différentes étapes nécessaires pour la construction automatique d'ontologie du domaine. Nous explicitons dans cette section chaque étape.

2.2.1. Le corpus

Un corpus est un ensemble de textes homogènes, présentés sous un format brut ou semi-structuré (**Azé & Heitz, 2004**). Le corpus doit être soigneusement choisi en fonction du domaine et de l'application visée. La taille des corpus et la masse d'informations contenues impliquent l'utilisation d'outils de terminologie textuelle spécifiques pour préparer les textes bruts à être utilisés par une quelconque application dont l'objectif serait par exemple de construire des ressources linguistiques telles que des ontologies ou des terminologies.

Le prétraitement de corpus est l'étape préliminaire pour identifier les données lexicales à partir des textes (**Harrathi, 2009**). Les prétraitements des données textuelles consistent à normaliser les diverses manières d'écrire un même mot, à corriger les fautes d'orthographe évidentes ou les incohérences typographiques et à expliciter certaines informations lexicales exprimées implicitement dans les textes. Les traitements consistent, par exemple, à remplacer (**ة, أ, ي**) qui sont écrites habituellement (**ه, ا, ى**), ou à extraire la structure superficielle des textes à partir d'indices comme une ligne vide pour délimiter les paragraphes (**Heitz, 2006**).

2.2.2. La segmentation

La segmentation est une étape quasiment obligatoire avant l'extraction d'information. Elle permet de découper le texte en unités linguistiques suffisamment élémentaires pour qu'elles soient traitées (**Dubois, 1994**). C'est une étape qui permet de découper un texte d'abord en section puis en phrase et enfin en mot.

Exemple : خَتَمَ/ اللَّهُ/ عَلَى/ قُلُوبِهِم

2.2.3. Étiqueteur : permet l'identification de la catégorie grammaticale (nom, verbe, adjectif, particule…) de chaque mot. Un texte étiqueté ressemblera grossièrement à ceci :

قُلُوبِهِم	**عَلَى**/	**اللّهُ**/	**خَتَمَ**/
/nom	préposition	/nom	*verbe*

Tableau 6: Etiquetage d'une phrase

2.2.4. Lemmatiseur

C'est un programme de traitement automatique du langage qui permet de passer d'un mot portant des marques de flexion (pluriel, forme conjuguée d'un verbe...) à sa forme de référence (lemme ou forme canonique).

Exemple :

قلوبهم ⟵ قلب

2.2.5. Concordancier

C'est un programme qui, pour un mot donné, recherche dans un texte toutes ses concordances, c'est-à-dire les phrases ou les groupes de mots dans lesquels il apparaît. Le concordancier n'apparait pas dans la figure 9, mais il est très utilisé dans la recherche d'information.

2.2.6. Extracteur de termes

L'extracteur terminologique analyse le contenu d'un document et recherche la terminologie disponible dans un corpus choisi.

2.2.7. Extracteur de relations sémantiques

Il s'agit de repérer des relations entre des termes préalablement extraits, telle la synonymie, l'hyponymie ou encore la méronymie.

Exemple : Une phrase telle que:

«ثم قست قلوبكم من بعد ذلك فهي كالحجارة »(**البقرة 74**)

(*Puis, et en dépit de tout cela, vos cœurs se sont endurcis; ils sont devenus comme des pierres*)[32] (**Chapitre2 : La vache 74**)

Génère une relation de similitude entre les *cœurs durs* et les *pierres*.

[32] http://www.mosquee-lyon.org/?cat=Coran

3. Approches et outils pour l'extraction de termes

L'extraction de Terminologie consiste à identifier des termes potentiels dans un texte spécifique ou un ensemble de textes (corpus) ainsi que les informations pertinentes liées à l'emploi de ces termes ou aux concepts auxquels ils renvoient (définition, contexte, etc.).

L'extraction de termes représente une étape importante dans la construction d'une ressource ontologique à partir de corpus. Les termes sont des mots ou des expressions ayant un sens précis dans un contexte donné et représentent les supports linguistiques des concepts.

Le problème de la constitution de ressources est au cœur de l'activité terminologique. Si la notion de "terme", qui fait appel à celle de concept et se fonde souvent sur un acte de référence particulier, semble peu se prêter à un traitement informatique, un certain nombre d'outils visant à extraire les termes d'un corpus ont vu le jour (**Piwowarski, 2003**).

La définition du terme donnée ci-dessus exerce des contraintes fortes sur la forme et le fonctionnement des unités terminologiques. Ces contraintes constituent les principes opérationnels des logiciels d'extraction de terminologie qui ont été développés ces dix dernières années. L'objectif de ces logiciels est de fournir automatiquement un lexique plus ou moins structuré du domaine. Les méthodes utilisées se distinguent par les caractéristiques du terme qu'elles choisissent d'exploiter. Généralement, on sépare les outils d'extraction de terminologie exploitant les caractéristiques linguistiques du terme et ceux basés sur une étude statistique du corpus, exploitant principalement le phénomène d'occurrence (**Bourigault & Jacquemin, 2000**).

On distingue deux grandes familles pour l'extraction automatique de termes: Les méthodes linguistiques et les méthodes statistiques. Ces deux types de méthodes ne s'excluent pas mutuellement et peuvent être combinés pour obtenir de meilleurs résultats, ce sont les méthodes mixtes ou hybrides.

3.1. Les approches linguistiques

Ces méthodes font généralement appel à des connaissances linguistiques *à priori* qui peuvent être syntaxiques, lexicales ou morphologiques.

Les méthodes linguistiques considèrent que la construction des unités terminologiques obéit à des règles de syntaxe plus ou moins stables, ce sont principalement des syntagmes formés de noms et d'adjectifs. Se basant sur ces connaissances, ces systèmes procèdent à l'extraction de candidats termes à l'aide de schémas syntaxiques (**Malaisé, 2005**).

On peut aussi utiliser des grammaires et un lexique acquis en cours d'analyse ou par le biais d'une collaboration avec des spécialistes pour générer l'ensemble des termes potentiels d'un domaine (**Drouin, 2002**).

Ces outils nécessitent donc un prétraitement du corpus par un analyseur syntaxique. La qualité des résultats dépend étroitement de la qualité de ces analyseurs. Ils ont l'inconvénient de dépendre directement de la langue des textes traités et nécessitent des ressources linguistiques (dictionnaires, liste de stop-word etc.). De plus ils ne sont efficaces que sur de petits corpus.

Les premiers travaux dans ce sens, sont ceux de David et Plante. Le premier outil spécifiquement dédié à la construction de bases terminologiques est *Termino* (**David & Plante, 1990**). Il a été élaboré dans le cadre d'une collaboration entre l'Office de la langue française du Québec et une équipe du Centre d'ATO de l'Université du Québec à Montréal. La première version de ce logiciel, qui se nommait donc *Termino*, a depuis été remplacée par un nouveau système nommé *Nomino* (**Perron 1996**).

La majorité des outils, pour ne pas dire la totalité, pour le traitement de l'Arabe n'ont pas été construits à l'origine pour accomplir cette tache **(Zaidi & Laskri, 2009)**. Une grande partie de ces outils, a été développée pour le traitement de l'Anglais, du Français ou autre langue. Après des années de raffinement et d'amélioration l'idée ou le besoin s'est ressenti de l'adapter au traitement de la langue Arabe (**Zaidi & al., 2010d**).

La littérature regorge d'ouvrages présentant ou recensant les outils dédiés à l'extraction de termes. Les outils que nous allons cités dans les sections suivantes sont ceux qui supportent ou peuvent être modifiés pour supporter la langue arabe. L'objectif fondamental dans ce travail étant la construction d'ontologies en langue

arabe. Nous invitons aimablement le lecteur désireux de s'étaler sur les outils pour d'autres langues à consulter les ouvrages suivants : **(Assadi H.& Bourigault D., 1996), (Beguin & al, 1997), (Biebow & Szulman, 2000), (Condamines & Rebeyrolle1997), (Daille, 1994), (Daoust, 1992), (Garcia, 1998), (Hearst, (1992), (Le Priol, 1998), (Morin, 1999), (Smadja, 1993), (Gandon, 2002) (Voutilainen, 1993).**

3.1.1. GATE

GATE[33] (General Architecture for Text Engineering) est une boîte à outils logicielle écrite en Java à l'université de Sheffield, à partir de 1995. Il est très largement utilisé à travers le monde, par de nombreuses communautés (scientifiques, entreprises, enseignants, étudiants) pour le traitement du langage naturel dans différentes langues. La communauté de développeurs et de chercheurs autour de GATE est impliquée dans plusieurs projets de recherche européens comme TAO (Transitioning Applications to Ontologies) et SEKT (Semantically Enabled Knowledge Technology).

GATE offre une architecture, une interface de programmation d'applications(API) et un environnement de programmation graphique. Il comporte un système d'extraction d'information, ANNIE (A Nearly-New Information Extraction System, pour système quasi nouveau pour l'extraction d'information), lui-même formé de modules parmi lesquels un analyseur lexical, un gazetteer (lexique géographique), un segmenteur de phrases (avec désambiguïsation), un étiqueteur, un module d'extraction d'entités nommées et un module de détection de coréférences. Il existe de nombreux plugins d'apprentissage automatique (Weka, RASP, MAXENT, SVM light), d'autres pour la construction d'ontologies (WordNet), pour l'interrogation de moteurs de recherche comme Google et Yahoo et pour l'étiquetage (Brill, TreeTagger), etc.

GATE accepte en entrée divers formats de texte comme le texte brut, HTML, XML, Microsoft Word (Doc), PDF, ainsi que divers formats de bases de données comme Java Serial, PostgreSQL, Lucene, Oracle, grâce à RDBMS et JDBC.

[33] http://gate.ca.uk

GATE utilise également le langage JAPE (Java Annotation Patterns Engine) pour bâtir des règles d'annotation de documents. On trouve aussi un debugger et des outils de comparaison de corpus et d'annotations (**Cunningham & al, 2002**).

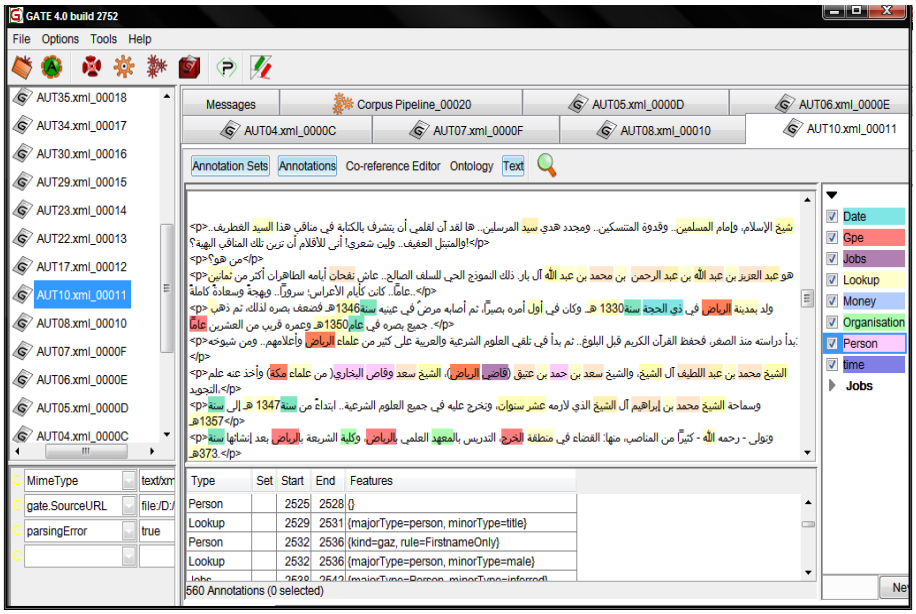


Figure 10: Extraction d'entités nommées arabes avec Gate **(Amari, 2009)**

GATE a été construit à l'origine pour l'extraction d'entités nommées **(Zaidi & al, 2010c)**, mais par la suite des améliorations ont permis de l'adapter pour de multiples fonctions dans plusieurs langues, nous présentons notre contribution dans le chapitre système proposé concernant l'extraction de termes et de relations à partir de textes arabes **(Zaidi & al, 2010e).**

3.1.2. NOOJ

NOOJ[34] est un système de traitement de corpus, reprenant et améliorant les fonctionnalités d'INTEX (**Koeva & al, 2007**), conçu pour l'enseignement des langues et de la linguistique. Le système intègre des outils TAL (analyse morphologique de mots simples et complexes, élaboration de transducteurs d'états finis, grammaires locales) qui permettent un prétraitement du corpus par l'enseignant, et des procédures de recherche et d'entraînement pour l'étudiant.

La nouvelle mouture du logiciel INTEX (appelée "NOOJ") a été réécrite à partir de zéro, en particulier pour répondre aux besoins des utilisations pédagogiques.

[34] http://www.nooj4nlp.net

Le système NOOJ présente des fonctionnalités de TAL qui paraissent prometteuses pour l'enseignement des langues et de la linguistique. Son principal avantage est sa simplicité d'utilisation : il permet à la fois à l'enseignant non spécialiste de TAL de constituer des ressources linguistiques (à l'aide d'interfaces simples) et de les paramétrer afin de constituer des projets pédagogiques destinés aux apprenants (**Silberztein & Tutin, 2004**).

Après construction d'un dictionnaire et d'une grammaire nous pouvons extraire de l'information à partir de textes arabes. La figure3 ci-dessous, donne un exemple d'extraction d'entités nommées Enamex.

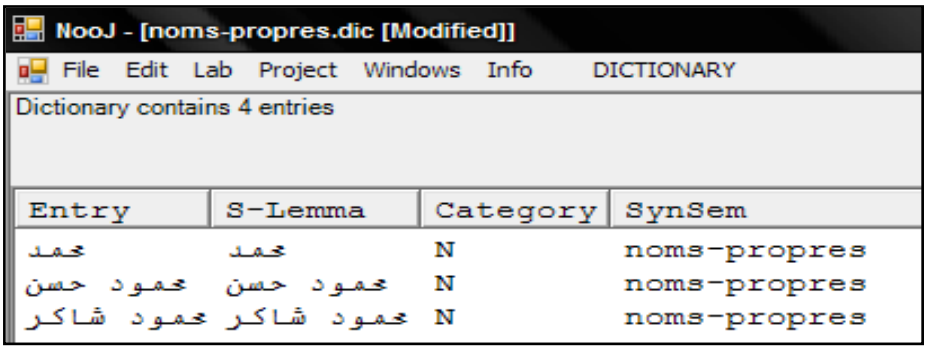


Figure 11: Extraction d'entités nommées arabes avec Nooj **(Lebhour, 2009)**

Il est aussi possible d'utiliser un concordancier sous NOOJ, ceci nous permettra avec l'aide d'une liste de termes simples, de chercher des termes composés, les sauvegarder dans une base de données par exemple et soumettre des requêtes sur des suites de mots respectant un certain marqueur.

3.1.3. UNITEX

Unitex[35] est un ensemble de logiciels permettant de traiter des textes en langue naturelle, en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire.

UNITEX manipule des textes Unicode. Le prétraitement du texte consiste à lui appliquer les opérations suivantes : normalisation des séparateurs, découpage en unités lexicales, normalisation de formes non ambiguës, découpage en phrases et application des dictionnaires (**Paumier, 2009**).

[35] http://www-igm.univ-mlv.fr/~unitex

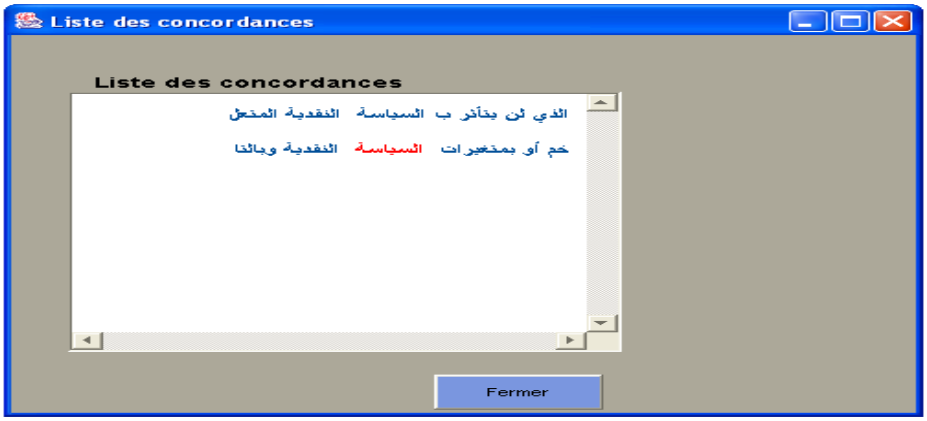


Figure 12: Extraction d'information arabe avec UNITEX **(Benmazou, 2009)**

En plus de l'extraction d'information, UNITEX dispose d'un concordancier permettant la localisation d'une entité dans le texte d'origine, en affichant ses contextes droit et gauche. L'outil peut etre facilement adapté à l'extraction de termes arabes de la même façon que l'outil NOOJ.

3.2. Les approches statistiques

L'approche statistique offre des avantages indéniables, puisqu'elle permet de s'attaquer à des ensembles de données d'une taille imposante qu'il serait tout à fait impensable de traiter manuellement (**Drouin, 2002**).

Les premiers travaux dans ce domaine, utilisant des données statistiques, datent des années 80, ils furent menés par Ludovic Lebart et André Salem sur les segments répétés (**Lebart &Salem, 1988**).

L'objectif des auteurs est de représenter un texte, qui est habituellement vu comme un enchaînement de formes simples, comme une succession de formes simples et de segments répétés (SR). Ces derniers sont définis comme des « *suites de formes graphiques non séparées par un caractère délimiteur de séquence, qui apparaissent plus d'une fois dans ce corpus de textes* » (**Drouin, 2002**).

Ces travaux exploitent des mesures de similarité. Il existe plusieurs méthodes statistiques appliquées à l'extraction de termes, la plupart sont fondées sur l'information mutuelle ou le coefficient de Dice (**Velardi, 2001**). Le principe est que l'association récurrente de deux mots ne peut être due qu'au fruit du hasard. Par conséquent, elle est forcément significative (**L'Homme, 2001**).