

Vers un système d'extraction d'informations pour les textes de la presse arabophone en ligne

ArIExtract

ACHIT Abdelmadjid* and AZZOUNE Hamid**

* CRSTDLA/Division Informatique Linguistique, Alger, Algérie. Email: aachit@yahoo.com

** USTHB University/Département d'informatique, Alger, Algérie. Email: azzoune@yahoo.fr

Résumé— De nos jours, l'extraction d'informations suscite un intérêt grandissant auprès de la communauté des chercheurs activant dans le domaine de l'ingénierie des langues. En effet, elle est d'un intérêt capital pour de nombreux domaines connexes relevant du TALN. Cet article vise à donner un aperçu sur un travail qui est en cours de réalisation et qui concerne le développement d'un système d'extraction d'informations pour les textes de la presse arabophone en ligne. Généralement, toute solution d'extraction d'informations englobe les tâches suivantes: reconnaissance des entités nommées, reconnaissance des relations entre ces entités avec la résolution des co-références dans certains cas. Pour l'élaboration de notre système, nous nous sommes intéressés, dans une première étape, à l'identification des entités nommées arabes: noms propres (personnes, organisations, lieux), expressions temporelles et expressions de mesure, en s'inspirant de la catégorisation des entités nommées, issue des conférences MUC. Et ce, avec une certaine adaptation pour le cas de la langue arabe. La seconde étape, a concerné l'extraction des relations entretenues entre les actants/acteurs. Pour notre cas, nous nous intéressons aux relations liées à la notion de « Rencontre ».

Mots Clés—Extraction d'informations; Entités nommées; Exploration contextuelle; Traitement de la langue arabe;

I. INTRODUCTION

Le travail, exposé dans cet article, concerne le développement d'un système d'extractions d'informations à partir de textes en langue arabe, issus de la presse arabophone en ligne. Ce papier s'articule autour de deux principales parties. La première est dédiée à la présentation du domaine de l'extraction d'informations. La seconde est consacrée à l'étude conceptuelle et à la démarche suivie pour l'élaboration du système. En dernier, nous évoquons quelques problèmes rencontrés dans la conception et la réalisation de ce système. Dans cet article, vu le stade actuel de l'étude, les entités nommées seront beaucoup plus traitées que les relations entre elles.

II. PRESENTATION DE L'EXTRACTION D'INFORMATIONS

Il s'agit dans cette partie d'introduire le domaine de l'extraction d'informations.

A. Définitions

La revue de littérature du domaine TALN, en général et du domaine de l'extraction d'informations en particulier, nous a permis de sélectionner les définitions suivantes:

Selon [5]: « L'extraction d'information consiste à extraire de l'information précise du contenu d'un document et à la représenter sous forme structurée. Cette forme structurée peut ensuite être stockée dans une base de données ou être utilisée comme base à la génération automatique de résumés ». Pour leur part [10], ils définissent l'extraction d'information comme une activité qui consiste à remplir une source de données structurées (base de données) à partir d'une source de données non structurées (texte libre). Aussi, il y'a [8] qui définissent l'extraction d'informations comme étant la structuration et la combinaison sélective de données issues d'un ou plusieurs documents textuels.

Les conférences *MUC* (*Message Understanding Conferences*) définissent la tâche d'extraction d'informations comme la tâche consistant à extraire des informations spécifiques et bien définies à partir de textes écrits en langue naturelle dans des domaines restreints, avec l'objectif spécifique de remplir automatiquement des formulaires prédéfinis ou des bases de données [7]. Selon [21], un système d'extraction d'informations est un système qui produit une représentation de l'information textuelle pertinente dans un domaine particulier pour une application particulière.

A partir des précédentes définitions, nous énonçons une définition, synthétisant ces dernières :

Définition: L'extraction d'information est un processus automatique permettant d'extraire des informations pertinentes et précises à partir de documents non structurés ou semi structurés en langage naturel et permet leur sauvegarde sous une forme structurée du type formulaire ou base de données (voir Fig. 1). Les critères de pertinence étant fixés au préalable par des patrons d'extraction (règles d'extraction).

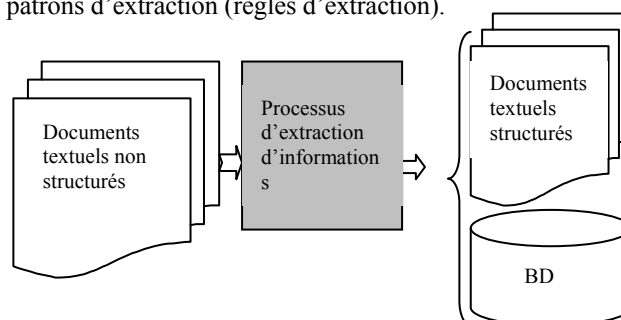


Figure 1. processus d'extraction d'informations

B. Historique de l'extraction d'informations

La réflexion sur les systèmes d'extraction d'informations fut entamée dès les années 1950, par les travaux de [12]. Par la suite, de nombreuses implantations ont été réalisées, nous pouvons citer celle réalisée à l'université de New York au travers du Linguistic String Project [20], dont le but était de remplir des formulaires à partir de textes médicaux (rapports de radiologie).

C'est à partir de la fin des années 1980, que la recherche en extraction d'information a connue une grande diffusion. Aux Etats-Unis, l'agence de la défense américaine DARPA s'est particulièrement intéressée à l'extraction d'information d'où l'organisation des *Message Understanding Conferences MUC*, une série de conférences organisée depuis 1987 afin d'évaluer les systèmes de compréhension de messages, cela a donné lieu à de nombreux projets de recherche, par exemple [11]. De son côté, l'Europe s'est intéressée au domaine et de nombreux systèmes ont été développés pour répondre aux besoins de l'industrie

Mais ce sont les conférences MUC (Message Understanding Conférences), qui ont considérablement réactivé ce courant de recherche. Durant l'évolution des conférences MUC, nous assistons à un passage de la compréhension de textes à la compréhension des messages puis à l'extraction d'informations. Le tableau ci-dessous, récapitule ces conférences en indiquant les tâches que s'est assignées chacune d'elles :

TABLE I. TABLEAU RECAPITULATIF DES CONFERENCES MUC

N°	MUC	Tâches
1	MUC 1 (1987)	Analyse des rapports d'opérations tactiques navales [5]
2	MUC 2 (1989)	Idem que MUC 1 [5]
3	MUC 3 (1991)	Analyse des textes journalistiques traitant du terrorisme en Amérique Latine, afin d'extraire des dépêches d'agence de presse le maximum d'informations sur des actes terroristes comme le nom de groupes terroristes impliqués, le nom des victimes, les types d'armes utilisées, les dates et les lieux... [15]
4	MUC 4 (1992)	Idem que MUC 3 [16]
5	MUC 5 (1993)	Traitement d'un corpus de nature économique (fusion, rachat, et création d'entreprises internationales et la fabrication de circuits électroniques). [17]
6	MUC 6 (1995)	C'est une suite de MUC 5: ont traité les changements de dirigeants à la tête des entreprises. [18]
7	MUC 7 (1998)	Analyse de textes journalistiques rapportant des crashes d'avion et de tirs de missiles. [19].

Il y'a lieu aussi d'indiquer l'existence d'autres manifestations scientifiques sous forme de conférences, d'ateliers et de campagnes d'évaluation et de benchmarking, consacrés à l'extraction d'informations ainsi qu'aux autres disciplines du traitement automatique du langage naturel TALN. Ces rencontres ont, également, contribué au développement et à l'évolution des techniques d'extractions d'informations. Le tableau ci-après, reprend quelque unes de ces rencontres:

TABLE II. TABLEAU DES DIVERS CONFERENCES SUR L'EI

Conférence	Détail
ACE	Automatic Content Extraction <i>Automatic Content Extraction (ACE)</i> http://www.itl.nist.gov/iad/mig/tests/ace/
NER	Language-Independent Named Entity Recognition at <i>Computational Natural Language Learning (CoNLL)</i> workshops http://www.cnts.ua.ac.be/conll/ http://www.cnts.ua.ac.be/conll2003/ner/
DUC	Document Understanding Conference http://duc.nist.gov/
TAC	Text Analysis Conference http://www.nist.gov/tac/
PASCAL Challenge	PASCAL Challenge for Evaluating Machine Learning for Information Extraction http://nlp.shef.ac.uk/pascal/
LREC	<i>International Conference on Language Resources and Evaluation</i> http://www.lrec-conf.org/
QA@TREC	Question Answering at The Text REtrieval Conference TREC http://trec.nist.gov/
QA@CLEF	Cross-lingual Question Answering at CLEF Cross Language Evaluation Forum http://www.clef-campaign.org/
MET	the <i>Multilingual Entity Task Conference (MET)</i> , TIPSTER Text project http://www-nlpir.nist.gov/related_projects/tipster/met.htm
IREX	<i>Information Retrieval and Extraction Exercise</i> http://nlp.cs.nyu.edu/irex/index-e.html
NTCIR	NTCIR (NII Test Collection for IR Systems) workshop http://research.nii.ac.jp/ntcir/

C. Domaines d'application

Dans la pratique, nous dénombrons une multitude de domaine d'applications, tels que :

1. Résumé automatique de documents relevant de domaines spécifiques ;
2. Traduction automatique, en améliorant la précision et la qualité de la traduction ;
3. Développement de systèmes Q/R (permet d'identifier le type de réponse attendu) ;
4. Amélioration de la précision des systèmes de IR (indexation et recherche) ;
5. Veille scientifique et technique, etc.
6. Extraction de terminologies à partir des textes et construction de thesaurus ;

D. Quelques solutions d'extraction d'informations

1) ÉCRAN

C'est un système d'extraction d'informations pour le français réalisé chez Thales (ex-Thomson-CSF) en 1998, sur la base de la plate-forme Gate à l'Université de Sheffield. Il comporte aussi des outils d'aide à l'acquisition de lexiques et de grammaires à partir de corpus qui ont été intégrés et testés. Il a été utilisé pour l'analyse de récits d'attentats terroristes et de comptes-rendus de films de cinéma.

2) EXIBUM

EXIBUM (EXtraction d'Information Bilingue de l'Université de Montréal) est un système bilingue (français, anglais) d'extraction d'informations qui réutilise

autant que possible des outils existants, en particulier, ceux développés au RALI, à l'Université de Montréal. Il traite des dépêches d'agences de presse sur les attentats en Algérie [13].

E. Processus général d'extraction d'informations

Généralement, un processus d'extraction d'informations englobe les tâches suivantes :

- Identification / reconnaissance des entités nommées ;
- Extraction des relations entre les entités nommées ;
- Formatage, sauvegarde et présentation des résultats sous une forme structurée.

1. IDENTIFICATION DES ENTITES NOMMEES

Dans la pratique, nous observons une multitude d'études et de travaux qui ont porté sur la reconnaissance des entités nommées dans des textes de presse (articles & dépêches), notamment ceux des Conférences MUC. Cette tâche importante est omniprésente dans tout processus d'extraction d'informations. L'identification des entités nommées inclut traditionnellement trois types d'expressions: les noms propres (ENAMEX) (noms de personnes, noms d'organisations, noms de lieux), les expressions temporelles (TIMEX) et les expressions numériques (NUMEX), ainsi, qu'un certain nombre d'entités qui ne sont pas toujours considérées comme noms propres : les noms collectifs (*les Algériens, les néandertaliens*, etc.), les maladies ou encore les noms de personnages mythiques ou fictifs (*Zorro, Tintin*, etc.).

Selon [9], la reconnaissance des entités nommées est une tâche complexe, qui nécessite le recours à de nombreuses ressources lexicales (liste de prénoms, d'entreprises, de régions, de fleuves, de groupes de musique, etc.). Les lexiques que nous exploitons sont le plus souvent incomplets. En effet, il est difficile de créer des lexiques exhaustifs : recenser l'ensemble des cours d'eau de la Planète serait une tâche presque impossible. D'autre part, la maintenance de ces lexiques est une activité très lourde, comme pour les noms d'entreprises, par exemple. Cette tâche a obtenu les meilleures performances. Les taux combinés de précision et de rappel sont comparables à ceux des humains, avec un taux de l'ordre de 0,90 de précision et de rappel sur des dépêches journalistiques en langue anglaise [19].

La catégorisation des EN issus des travaux des conférences MUC (voir Fig. 2) regroupe une grande partie des entités nommées présentes dans les textes journalistiques, mais, elle est limitée et inadaptée à la traduction par exemple, car elle reste insuffisamment exhaustive et peu fine. C'est pour cela qu'il faudrait selon le domaine d'application introduire une extension pour cette catégorisation qui répondra le mieux aux types d'informations à capturer. Il y'a lieu aussi de réfléchir à une solution permettant la constitution d'une catégorisation générale qui soit la plus complète possible afin de garantir son indépendance du domaine d'application.

Notons que la reconnaissance des noms propres peut se découper en deux tâches distinctes: identification des noms propres connus et la découverte de nouveaux noms propres et leurs catégorisations.

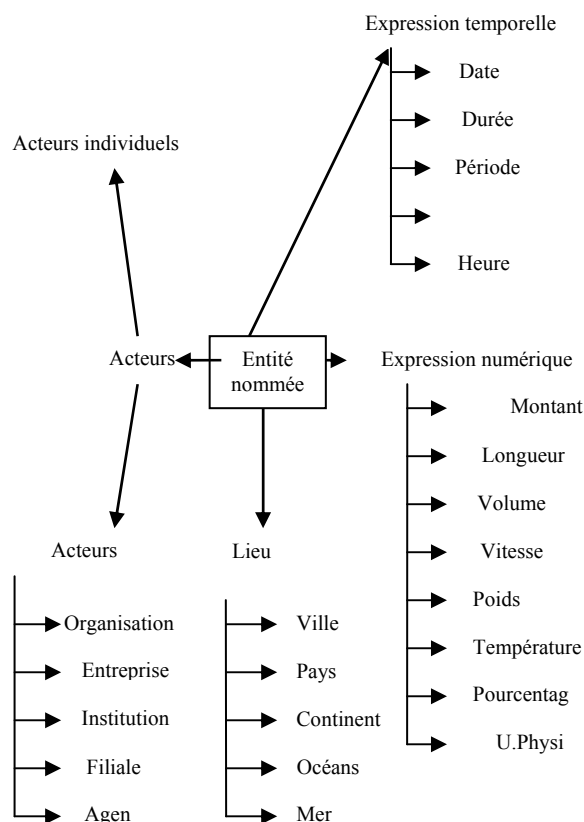


Figure 2. Catégorisation des EN – version MUC
6 et 7

2. EXTRACTION DES RELATIONS ENTRE LES ENTITES NOMMEES

Dans cette étape, nous nous intéressons à l'identification des relations existantes entre les entités nommées découvertes lors de la précédente étape. Le repérage de ces relations permettra, ainsi, la construction de modèles de représentation des connaissances. D'après la revue de littérature du domaine, les principes d'extraction de connaissances procèdent de deux manières :

- Etude de la distribution de contextes autour des entités (étude statistique);
- Détermination de patrons/ schémas/ formules linguistiques caractéristiques de relations lexicales (étude linguistique basée sur l'acquisition de marqueurs de relations liées à certaines notions/ concepts).

EXEMPLE :

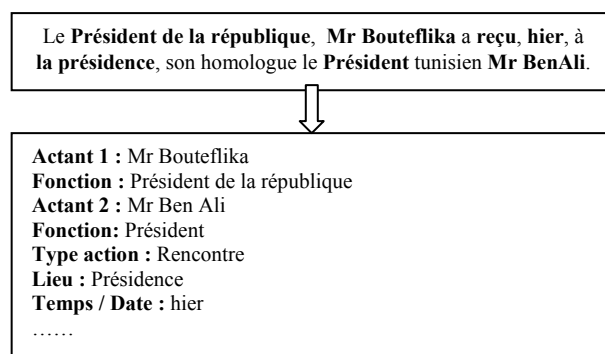


Figure 3. Cas d'EI d'une dépêche de presse

Le processus d'extraction d'informations implique la prise en charge du problème posé par les coréférences. Il s'agit de d'entités / de groupes nominaux qui se réfèrent au même objet. C'est le cas par exemple, des anaphores.

F. Approches et méthodes d'extraction

Les grandes approches suivies pour l'identification des entités nommées ainsi que des relations sont :

- Approche basée sur l'analyse linguistique et les techniques du TALN (dite manuelle) ;
- Approche Apprentissage automatique ;
- Approche hybride qui exploite les deux approches précédentes.

1. APPROCHE LINGUISTIQUE

L'approche linguistique est fondée sur la description syntaxique et lexicale des syntagmes recherchés. Des règles de grammaire utilisent des marqueurs lexicaux (ex. Mr pour Monsieur,..), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement pour repérer les mots inconnus) sont utilisés pour repérer et typer les syntagmes intéressants [1], [2], [11]. La méthode de l'exploration contextuelle [5], [6] en est un exemple. Citons aussi, le travail de [22] qui a porté sur la conception d'un système de reconnaissance des entités nommées arabes de type noms de personnes PERA basé sur l'utilisation de règles sous forme d'expressions régulières ainsi d'un lexique de noms de personnes.

2. APPROCHE APPRENTISSAGE AUTOMATIQUE

Il s'agit d'une approche basée sur les techniques d'apprentissage automatique et statistiques. Elle utilise un modèle de langage entraîné sur de larges corpus de textes pré-étiquetés. Parmi les travaux relevant de cette approche, nous citons [3] qui a porté sur l'utilisation des supports vecteurs machines SVM pour la reconnaissance des entités nommées arabes. Egalement, il y'a eu des travaux qui ont porté sur l'utilisation des Modèles de Markov Cachés MMC ainsi que de la méthode de l'entropie maximale [4].

3. APPROCHE HYBRIDE

Plus récemment sont apparues des approches hybrides tirant parti des avantages respectifs des méthodes linguistique et probabiliste. Dans les systèmes de ce type, un ensemble de règles est généralement appris automatiquement puis révisé par un expert [1].

G. Quelques difficultés rencontrées dans l'extraction d'informations

Parmi les problèmes et obstacles rencontrés dans la conception de systèmes d'EI, nous citons:

- La langue naturelle est flexible. Il y a toujours plusieurs façons d'exprimer la même idée ;

- La langue naturelle est ambiguë. Une phrase peut être interprétée de différentes manières ;

- La langue naturelle est dynamique. Elle évolue constamment ;

- Multilinguisme;

- Style de textes: textes journalistiques, textes d'un email (absence de règles et de style rédactionnel)

- L'information peut s'étendre sur plusieurs phrases;

- Complexité du processus d'EI du fait des différentes tâches :

- Identification des entités nommées ;
- Recherche des relations entre entités ;
- Résolution des coréférences ;

- Evaluation difficile;

- Données : quantité croissante, non standardisées et de types différents;

- Limites de l'état de l'art des systèmes d'EI. ;

- Difficulté de conception de systèmes d'EI. génériques ;

- Peu de systèmes d'E.I. commercialisés ;

- Trop peu d'interdisciplinarité (non informaticiens et informaticiens).

III. ETUDE CONCEPTUELLE

Notre étude vise à répondre à la problématique soulevée par la rareté des outils d'extraction d'informations pour les textes arabes et les besoins exprimés par une multitude de domaines comportant une analyse de textes arabes en vue d'extraire des informations précises. Ce genre d'applications serait donc, d'un apport considérable pour plusieurs cas pratiques.

Dans le cadre de notre étude, nous nous fixons comme objectif, le développement d'un système d'extraction d'informations pour des textes en arabe standard issus de sites web de la presse en ligne arabophone. Le choix des textes journalistiques n'est pas fortuit. Il est motivé par le fait que ces textes respectent un certain style rédactionnel et sont disponible en ligne. Comme thématique pour ces textes, nous avons sélectionner ceux relatant des événements de rencontre entre personnes et ceux décrivant des manifestations (scientifiques, économiques, culturelles, etc.).

A. Approche adoptée pour la conception du système

Pour le développement de notre système d'extraction d'informations, nous nous sommes appuyer sur une méthode linguistique dite méthode d'exploration contextuelle. Cette dernière a donné de bons résultats dans de nombreuses implémentations.

PRESENTATION DE LA METHODE D'EXPLORATION
CONTEXTUELLE [5]

D'après la présentation de l'Exploration Contextuelle (EC) faite dans [5], il s'agit d'une méthode issue des recherches effectuées par l'équipe LaLICC [6] pour le traitement automatique des textes en langue naturelle. De nombreuses applications informatiques, utilisant cette méthode, ont déjà été réalisées, notamment le résumé automatique [14], le filtrage d'informations selon différents points de vue, etc.

Elle est basée sur une analyse linguistique, permettant le repérage des entités nommées (acteurs, lieux, temps,...) ainsi que la mise en relation des acteurs avec leur environnement dans l'espace et le temps au moyen d'indices déclencheurs, d'indices complémentaires et de règles qui les combinent. En effet, Cette méthode a une portée sémantique et ne se base pas sur une représentation profonde du texte mais sur une identification automatique de certaines unités linguistiques (marqueurs) pertinentes pour une tâche donnée, appelées indices déclencheurs (indicateurs) et indices complémentaires. Les indices déclencheurs sont retenus en fonction d'objectifs précis (par exemple, déterminer une relation sémantique entre concepts et/ou la valeur sémantique contextualisée d'un marqueur grammatical ou lexical polysémique). Une analyse exploratoire du contexte permet d'identifier d'autres indices linguistiques, eux aussi jugés pertinents pour la tâche traitée (indices complémentaires). L'indice déclencheur et les indices complémentaires étant identifiés, ils permettent, au moyen de règles heuristiques, de prendre les décisions impliquées par l'objectif attendu dans un contexte bien défini. Ces règles se déclenchent pour attribuer à une unité lexicale (une phrase, un paragraphe, etc.) des étiquettes sémantiques, etc.

Afin de pouvoir utiliser cette méthode pour la découverte des relations recherchées, il est nécessaire au préalable de construire une base d'indices linguistiques (marqueurs) exprimant les relations entre actants et leurs environnements dans l'espace et le temps. Ces indices sont regroupés dans des classes sémantiques: celles des indicateurs et celles des indices complémentaires qui seront mises en relation par un ensemble de règles d'EC. L'action de l'ensemble des règles permet de construire progressivement des représentations sémantiques. Certaines règles permettent de créer des marqueurs d'EC complexes, d'autres d'attribuer une étiquette sémantique à une phrase.

Dans ce qui suit nous donnons une spécification littéraire d'une règle d'exploration contextuelle tel que rapporté dans [5].

1. Spécification de l'espace de recherche

E:= Créer espace(PhraseParent de l'indicateur principal) ;

2. Spécification des listes des indicateurs et des indices complémentaires

Li := liste de verbes / adjectif/ ...

3. Conditions

Concerne les contraintes d'agencements et d'ordonnement des marqueurs ainsi que des indices complémentaires dans l'espace de recherche considéré.

4. Actions

Attribuer une étiquette au segment textuel considéré (la phrase) ou déclencher une tâche.

B. Méthodologie de mise en œuvre de l'exploration contextuelle EC

1. Spécification de l'objectif de l'extraction : en fixant le domaine thématique ou plus précisément en spécifiant les notions et concepts liés aux relations à extraire ;

2. Constitution / collecte d'un corpus de textes du domaine objet de l'étude répondant au domaine sémantique sélectionnée;

3. Fouille systématique des textes du corpus en vue d'extraire et de réunir les données linguistiques nécessaires pour la mise en œuvre de cette méthode :

- les indicateurs (marqueurs) ;

- les indices complémentaires et informationnels ;

- les règles qui les combinent (élaboration des règles d'exploration contextuelle);

4. Informatisation de ces règles ;

5. Validation des règles d'extraction élaborées sur des textes pris en dehors de ceux du corpus initial.

C. Choix du corpus d'analyse

Du fait de l'absence de corpus de textes journalistiques en langue arabe, en libre téléchargement et non payant, nous avons décidé de nous constituer un petit corpus qui nous servira lors de la phase de développement de l'extracteur d'informations. Ainsi, pour les besoins de l'étude nous avons constitué un corpus de la taille de 100 documents. L'aspect thématique des documents concerne, en premier, les dépêches de presse concernant les rencontres (scientifiques, politiques, économiques, etc.).

Exemple : texte portant sur des rencontres politiques

التقى جورج ميتشيل مبعوث الرئيس الأميركي باراك أوباما إلى الشرق الأوسط بالقاهرة في مستهل جولة بالمنطقة تستغرق أسبوعاً الأمين العام للجامعة العربية عمرو موسى والمفوض الأعلى للأمن والسياسة الخارجية بالاتحاد الأوروبي خافيير سولانا.

D. Extraction des entités nommées et des relations**1) Reconnaissance des entités nommées EN arabes**

A l'instar des autres langues, la reconnaissance des entités nommées dans la langue arabe a suscité un très grand intérêt chez la communauté activant dans le domaine. Parmi les travaux qui ont concernés particulièrement la reconnaissance des entités nommées arabe, nous pouvons évoquer les travaux de [3] qui a eu recours à une méthode d'apprentissage automatique, en l'occurrence les supports vecteurs machines SVM et qui a donné un bon score (F-mesure=82,17). De son côté [22] a eu recours à une méthode linguistique manuelle dans l'élaboration du système PERA de reconnaissance des entités nommées arabes. Les règles utilisées ont été formalisées sous forme d'expressions régulières.

Dans notre étude, nous nous sommes inspiré de la catégorisation des EN de la conférence MUC 7, mais aussi, d'autres études. Au passage, nous observons, que la plupart s'accordent sur un minimum d'informations nécessaires à la description des événements occurrents et rapportés dans les textes journalistiques: acteurs ou actant (celui qui agit où celui qui subit l'action) en indiquant les coordonnées spatio-temporelles (lieu : où?, temps : quand?) associées. Ajouter à cela, l'information de mesure / monétaire nécessaire à décrire et à rendre compte des

événements économiques, par exemple. Les difficultés rencontrées sont pour la plupart liées : à l'absence de la voyéllisation dans la plupart des textes arabes disponibles en ligne (source d'ambiguïté) et à l'existence de plusieurs retranscriptions en caractères arabes de noms propres étrangers (absence de normes pour l'écriture de noms propres issus d'autres langues) etc.

Classification des entités nommées arabes adoptée

La catégorisation d'entités nommées retenue pour la conception de notre système est la suivante:

1. acteurs ou actants (agent de l'action ou cible de l'action) :

- particulier / individuel (personnes) ou
- collectifs (entreprise, organisme, institution, ...)

2. information de localisation (lieu géographique): villes, régions, pays, continents, océans, mers, fleuves, etc.

3. information temporelle : dates, durée, période, horaire, etc.

4. information numérique : mesure, monétaire ou pourcentage, etc.

L'idée centrale, c'est de capturer le maximum d'informations pertinentes rentrant dans la description d'événements ou de concepts scientifiques et technologiques.

TABLE III. APERÇU SUR LE JEU D'ETIQUETTES UTILISEES

N°	Type entité	Etiquette associée
1	Phrase	<Phrase>
2	Acteur particulier	<ActP>
3	Acteur collectif	<ActC>
4	Exp de localisation	<Lieu>
5	Exp de localisation : ville	<Ville>
6	Exp de localisation : région	<Région>
7	Exp de localisation : pays	<Pays>
8	Exp de localisation : local	<Loc>
9	Exp de localisation : lieu terrestre	<LieuT>
10	Exp de localisation : lieu maritime	<LieuM>
11	Exp temporelle	<Temps>
12	Exp temporelle de type durée	<Durée>
13	Exp temporelle de type horaire	<Horaire>
14	Exp temporelle de type age	<Age>
15	Exp temporelle de type période	<Période>
16	Exp temporelle de type date	<Date>
17	Exp numérique de monétaire	<ExpMon>
18	Exp numérique de longueur	<Long>
19	Exp numérique de poids	<Poids>
20	Exp numérique de volume	<Volume>
21	Exp numérique de vitesse	<Vitesse>
22	Exp numérique de température	<Température>
23	Exp numérique de pourcentage	<Pourcentage>
24	Nom propre	<NP>
25	Titre	<Titre>
26	Nom organisation	<Org>
27	Fonctions sociales	<FS>
28	Fonctions relationnelles	<FR>
29	Nationalité	<Nat>
30	Appartenance religieuse	<App-Rel>
31	Appartenance ethnique	<App-Eth>

RECONNAISSANCE DES ENTITES NOMMEES DE TYPE ACTEUR / ACTANT

Dans notre travail, la méthode suivie pour l'identification des entités nommées de type actant se base à la fois sur la structure interne de l'entité nommée ainsi que sur l'analyse du contexte

a) Reconnaissance des entités nommées de type actant particulier

Un acteur particulier est une personne qui est caractérisé par son nom propre (محمد، عبد الله) et sa fonction (... وزير، رئيس)، son titre (السيد، الدكتور، العاهل)، qui pourrait aussi avoir une classe d'appartenance : nationalité (... جزائري، مغربي، تونسي)، religion (... مسالم، مسيحي، يهودي). etc.

De ce fait, la reconnaissance et l'annotation des acteurs particuliers (personnes et ses attributs), nécessitent :

1. des ressources:

- un ensemble d'expressions régulières décrivant des entités selon leurs structures internes

- un lexique sous forme de dictionnaires et de classes d'indices: classe de fonctions sociales, classe d'appartenance (nationalité, religion, ...), etc.

2. un ensemble de règles lexico sémantiques: pour l'annotation finale de l'acteur. Ces règles sont indépendantes du domaine d'application.

Le travail de fond consiste donc à collecter et réunir les ressources nécessaires ainsi qu'à élaboration des règles lexico syntaxiques. Ces dernières sont en fait, des expressions régulières avec des règles dépendantes du contexte dites règles d'exploration contextuelles.

Exemple :



b) Figure 4. Cas de reconnaissance d'un actant particulier
actant collectif

D'une manière quasi identique au cas précédent, pour la reconnaissance des actants collectifs (noms d'organisation, d'entreprises, filiales, groupes, administration, institution, ...) nous exploitons des lexiques et nous faisons appel aux informations concernant la structure interne des entités en question ainsi qu'aux expressions régulières et aux règles dites d'exploration contextuelle.

Exemple :

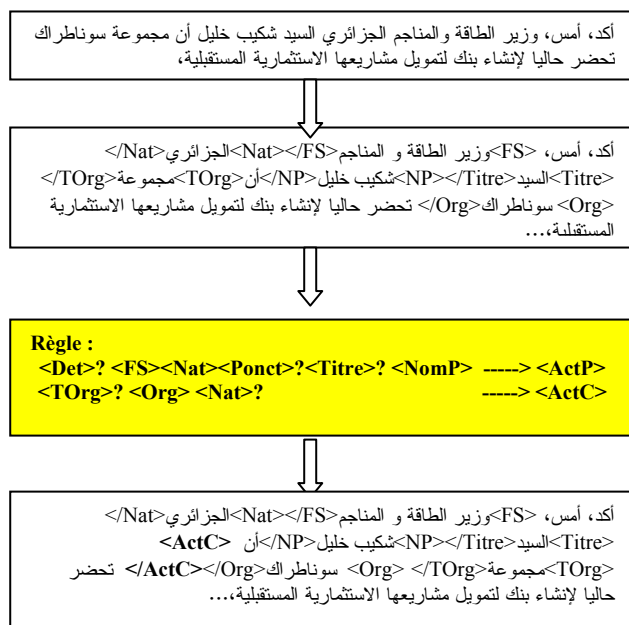


Figure 5. Cas de reconnaissance d'un actant collectif

c) Reconnaissance des noms de pays comme actants

Lors de l'opération de reconnaissance des entités nommées, nous avons le cas des noms de pays qui pose problème. S'agit il d'une localisation ou bien d'un actant collectif. Afin de lever ce problème nous avons eu recours à une règle d'exploration contextuelle :

Cas 1 : lieu géographique

غدا، بالجزائر، سينعقد اجتماع لوزراء الخارجية المغربية.

البارحة، زار الأمين العام للأمم المتحدة، بان كي مون غزة لتفقد الخسائر الكبيرة التي خلفها العدوان الإسرائيلي.

Cas 2 : actant collectif

قدمت الجزائر طلب انضمامها إلى المنظمة الدولية للتجارة.

Dans notre cas, c'est une règle d'exploration contextuelle qui sera utilisé pour attribuer l'étiquette adéquate en se basant sur le contexte linguistique de l'unité en question. Cette règle sous la forme littéraire, est la suivante:

Règle :

Si l'entité nommée étiqueter par <pays> est précédée par une préposition (... نحو، إلى، ب) ou par un verbe du type (... وصل، غادر، انتقل، ذهب، زار، مكث، بقي، ...) alors attribuer l'étiquette <Lieu> sinon attribuer l'étiquette <ActantCollectif>

2) Reconnaissance d'entités de type temporelle

Dans cette tâche, nous nous intéressons à l'étiquetage des dates, des durées, des différentes expressions temporelles. Cela permettra ultérieurement d'associer une information temporelle à la relation extraite. Pour l'achèvement de cette tâche, nous faisons appel aux expressions régulières ainsi qu'à des règles d'explorations contextuelles.

Détection des dates

Elles peuvent apparaître sous une :

- une forme numériques (1990/01/15, 1990-01-15, ...) ;
- une forme mixte (15 جانفي 1990) ;
- seulement de mots (خمسة عشر جانفي ألف وتسعة مئة وتسعون) ;
- les dates non absolues ("في ماي", "5 مارس") ;
- les dates absolues ("05 جويلية 2009") ;

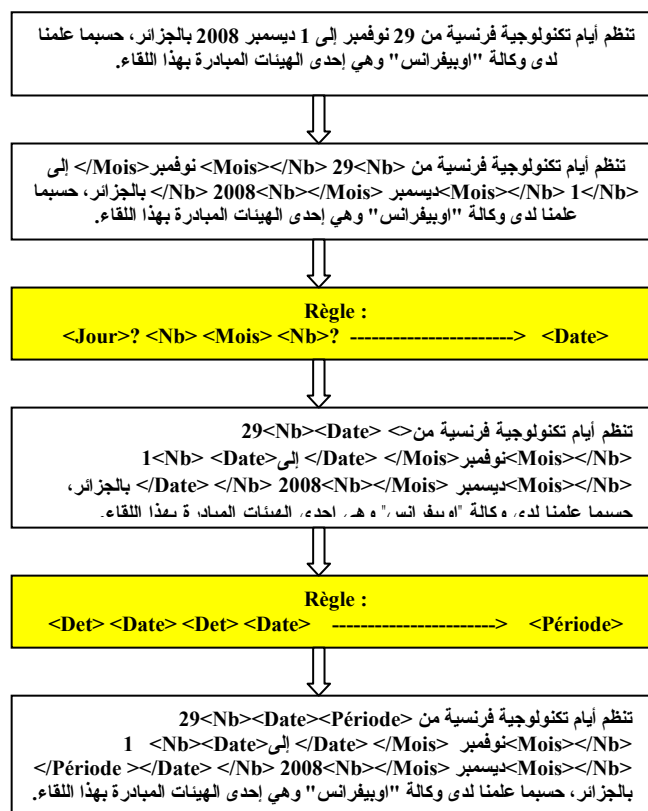
Détection des durées

- Des durées quelconques ("خلال 3 سنوات")
- Des intervalles temporels ("من 06 جوان إلى 15 جويلية")
- Des durées absolues ("انطلاقا من 05 جويلية")
- Des durées relatives au moment d'élocution ("منذ عام")

Détection des expressions temporelles

- Ce sont des expressions qui regroupent : par exemple :
- Des dates relatives, de forme particulière ("في القرن" (الماضي", "الأسبوع الفارط", "في بداية السنة") ;

Exemple :



3) Figure 6. Cas de reconnaissance d'expression temporelles

L'annotation de l'information spatiale, implique l'identification des noms de lieux géographiques : village, ville, pays, continent, mer, océan, fleuve, lac, montagnes, désert, plaines, etc. ainsi que toutes les unités linguistiques (noms de localisation, verbes de localisation, adjectif de localisation, classificateurs, adverbes de lieux, etc.) pouvant marquer et indiquer un nom de lieu ou contribuant à dénoter un nom de lieu. De la même manière, nous utiliserons des expressions régulières ainsi que des règles d'exploration contextuelles pour leur identification.

Exemple :

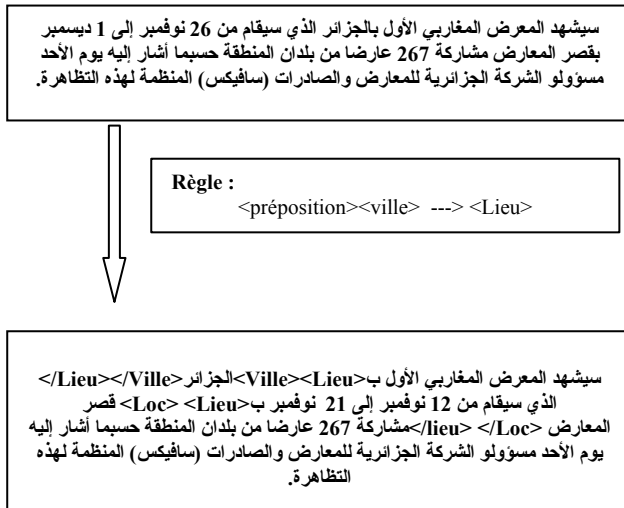


Figure 7. Cas de reconnaissance d'expression de localisation

4) Reconnaissance d'entités de type numérique (EN de mesure ou monétaire)

Il peut s'agir soit d'entités de mesure soit d'entités monétaires soit des pourcentages. Un nombre est soit :

- numérique simple : 10 أورو، 15 دولار، 100 مليون دينار،
- numérique avec virgule : %5,7
- numérique négatif : -6 %
- numérique composé : 7 آلاف دينار

Les classes utilisées pour l'identification et l'annotation des informations numériques sont :

- unités monétaires (أورو، دينار، دولار)
- unité de mesure (كغم، متر، لتر، كغ)
- de signes relatifs au pourcentage (%)

Exemple :

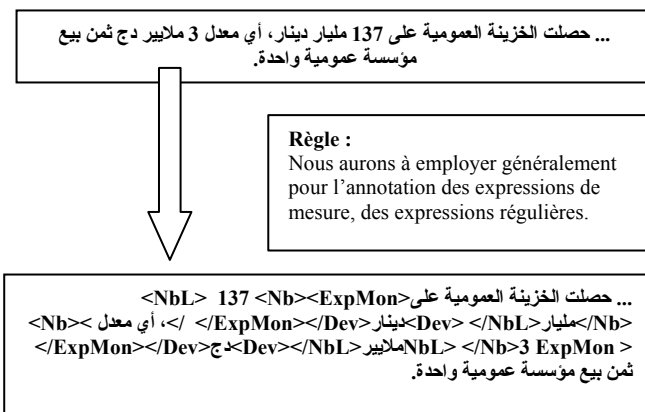


Figure 8. Cas de reconnaissance d'une expression de mesure

2) Extraction des relations entre les entités nommées

Ce deuxième module d'annotation prend en entrée un document ayant déjà subi une annotation de ces entités nommées : acteurs et les informations spatiales, temporelles. Cette étape va utiliser des ressources linguistiques plus importantes que celles de la précédente étape. En effet certaines règles de repérage de relations entre actants s'appuient sur des segments textuels déjà annotés (<actant>, <Temps>, <Lieu>...). Les règles d'annotation augmentent celles de la précédente étape de deux nouvelles formes: l'une qui prend en compte, dans ses prémisses, des segments textuels déjà annotés et l'autre prenant en charge les notions d'indicateur, d'indices complémentaires et d'espace de recherche selon la méthode d'exploration contextuelle.

Le repérage des relations entre les entités nommées, est réalisé grâce à des règles d'extractions, ou patrons d'extraction. Ces dernières sont exprimées sous forme d'expressions régulières et de règles d'exploration contextuelle.

Du fait, de la multitude de relations qui peuvent exister entre les entités nommées, nous nous sommes limités aux relations liées à la notion rencontre (scientifique, politique, économique, culturelle, religieuse, etc.). L'objectif étant d'essayer de repérer dans les dépêches de presse, les rencontres de personnes scientifiques, politiques, culturelles apparaissant dans les textes journalistiques et d'essayer d'extraire toutes les informations les décrivant (les personnes qui se sont rencontrées, date, lieu, ...). De ce fait, nous tâcherons de recenser les verbes véhiculant cette notion du genre (،... لقي، تحاور، عقد).

Exemple :

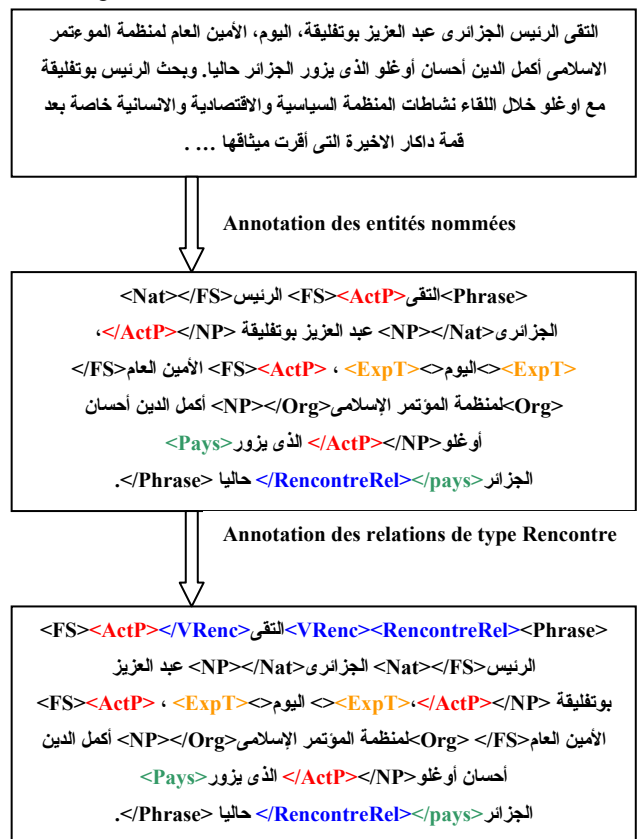


Figure 9. Cas de reconnaissance d'une relation de Rencontre

L'annotation des relations de type rencontre a été effectuée grâce à la règle suivante :

<VRenc><ActP1><Tps><ActP2><Lieu>---> <RencRel>

Concernant les contraintes imposées par cette règle:

La présence d'un verbe de type VRenc antéposé aux Acteurs ActP1 et ActP2, ainsi que des informations spatio-temporelles, nous pouvons annoter ce segment de texte par l'étiquette <RencontreRel>. Le repérage de ces relations se fait au travers d'une classe de marqueurs linguistiques (par exemple les verbes relatant l'organisation d'une rencontre). Parmi ces verbes (marqueurs déclencheurs) que nous avons retenus, nous citons:

، ضيف، تحاور، عقد، لقي، اجتمع،.....

Ces listes ont été constituées à partir des observations faites en analysant les dépêches et les articles journalistiques en langue arabe.

Dans la solution que nous proposons, le traitement des co-références n'est pas pris en charge, il sera l'objet des futurs travaux.

IV. PROBLEMES RENCONTRES DANS LE CAS DES TEXTES ARABES

1. Forme agglutinante des mots arabes : la langue arabe est une langue agglutinante. En effet, les mots arabes, peuvent être affixés, ce qui fait que des fois il y'a des particule qui colle à certaines entités ce qui ne facilite pas leur détection.

2. Absence de casse (indice naïf): absence de majuscules et de minuscules, dont la présence faciliter la reconnaissance des entités nommées du type noms propres par exemple.

3. Absence de normes d'écritures des noms propres : certains noms propres en langue latines sont retranscrits en langue arabe mais sous plusieurs formes, par exemple : Poutine est réécrit en arabe : بوتين، بوتن d'où la difficulté à réunir l'ensemble des formes possibles et d'où la nécessité de normaliser l'écriture des noms propres d'origine non arabe.

4. Non voyélisation des textes arabes est source d'ambiguïtés. En effet, le mot en arabe « مؤسسة » sans voyelles, peut s'interpréter selon deux sens distinct :

مؤسسة → entreprise

مؤسسة → fondatrice

5. Problèmes de délimitation des entités nommées pour deux raisons :

- mot inconnu : absence d'informations morphologiques (nécessite de disposer d'un analyseur morphologique)

- antonomase : passage du mot de la langue au nom propre

6. problème de la ponctuation qui n'est pas respectée dans la rédaction des textes arabes.

V. CONCLUSION

Dans cet article, nous avons rapporté une étude portant sur la conception et la réalisation d'un système d'extraction d'informations pour les textes de la presse arabophone en ligne. Ainsi, dans une première étape, nous nous sommes intéressés à l'identification des entités nommées arabes : noms propres (personnes, organisations, lieux), expressions temporelles et expressions de mesure,

en s'inspirant de la catégorisation issue des conférences MUC. Et ce, avec une certaine adaptation pour le cas de la langue arabe. La reconnaissance des différents types d'entités nommées a été possible grâce à l'utilisation de lexiques de noms de personnes, villes, ...mais, aussi grâce à l'utilisation d'une combinaison d'expressions régulières et de règles dites d'exploration contextuelle permettant l'annotation sémantique des entités nommées.

La seconde étape a concerné l'extraction des relations entretenues entre les actants concernant, pour notre cas, la notion de « Rencontre », et qui est marquée par des verbes déclencheurs du genre (لقي، اجتمع، تحاور، عقد، ضيف). Le repérage des relations est guidée par la présence dans l'espace de recherche (dans notre cas c'est la phrase) d'un mot marqueur caractérisant la relation recherchée de tel manière que les entités impliquées ainsi que ce dernier, respectent certaines contraintes d'agencement et d'ordres. Ces contraintes sont formulées sous forme d'une combinaison d'expressions régulières et de règles d'exploration contextuelles.

L'efficacité d'un extracteur d'informations dépend de la validité des règles d'extraction utilisées et de leur pouvoir d'identification des différentes entités nommées et à repérer les relations recherchées et ce malgré les nombreux obstacles linguistiques inhérents à la langue naturelle en général et à la langue arabe en particulier.

Notre système nécessite des améliorations certaines, tant sur le plan de la richesse du lexique nécessaire pour une reconnaissance efficace des ENs. Mais aussi, en ce qui concerne, la capacité des règles d'extraction à pouvoir repérer les relations cibles du domaine étudié. Il devra dans une étape ultérieure subir des testes pour calculer son de taux de précisions et de rappel pour pouvoir mesurer le degré de sa puissance en tant qu'outil d'extraction d'informations. Malgré les insuffisances existantes, ce travail, constitue une première ébauche pour développer un outil logiciel assurant cette tâche complexe et c'est un début pour évoluer vers des solutions plus robustes et ayant une large couverture en termes de capacités d'extractions et de captures d'informations pertinentes liées à un domaine spécifique.

Comme perspectives, nous envisageons d'enrichir nos ressources lexicales employées. Egalement, d'ajouter de nouvelles règles pour mieux appréhender l'ensemble des relations liées aux notions choisies dans notre cas d'étude. Mais, aussi, nous comptons nous intéresser à d'autres notions afin d'étendre les domaines couverts par notre application. Il y'a lieu aussi, de prendre en charge, le traitement des co-références dans les textes.

RÉFÉRENCES

- [1] J Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain. MITRE : Description of the alembic system used for MUC-6. In Proceedings of the Sixth Message Understanding
- [2] D. Appelt. An introduction to information extraction. Artificial Intelligence Communications, 12(3) :161–172, 1999.
- [3] Benajiba Yassine, Mona Diab and Paolo Rosso, Arabic Named Entity Recognition: an SVM-based Approach, Proceedings of 2008 Arab International Conference on Information Technology ACIT2008, Sfax, Tunisia December, 2008

- [4] Benajiba Yassine, Paolo Rosso and Jose Miguel Beneda, ANERsys: An Arabic Named Entity Recognition System based on Maximum Entropy, Proceedings of 2007 Conf. on Comp. Linguistics and Intelligent Text Processing CICLing2007, Mexico City, Mexico February
- [5] Bouhafs A., 2004. « Système d'extraction d'information dédié à la veille Qui est qui? Qui fait quoi? Où? Quand? Comment? » conférence Récital, FES, Le Maroc, 2004.
- [6] DESCLES J-P., Systèmes d'Exploration Contextuelle, dans Contexte et calcul du sens, Claude Guimier (éd).
- [7] N. Chinchor et E. Marsh, MUC-7 Information Extraction Task Definition, Proceedings of the Seventh Message Understanding Conference, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html, 1998.
- [8] Cowie J., Wilks Y. (2000), *Information Extraction*, in: *Handbook of Natural Language Processing*, R. Dale, H. Moisl, H. Somers (ed), Marcel Dekker inc.
- [9] Fourour et Morin, « APPORT DU WEB DANS LA RECONNAISSANCE DES ENTITÉS NOMMÉES », Revue québécoise de linguistique, vol. 32, n°1, 2003, p41-60.
- [10] Gaizauskas R. and Wilks Y., 1998. *Information Extraction: Beyond Document Retrieval*. Journal of Documentation, 54(1):70—105.
- [11] Grishman R., 1995. The NYU system for MUC-6 or where's the syntax?. *actes Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, San Francisco : 167-176.
- [12] Harris Z., Zellig S., 1951. *Structural Linguistics*. Chicago and London: The University of Chicago Press, 1960.
- [13] Kosseim L. et Lapalme G., 1998. EXIBUM : un système expérimental d'extraction bilingue. *Actes des rencontres internationales sur l'extraction, le filtrage et le résumé automatique (Rifra '98)*, Sfax : 129-140.
- [14] MINEL J-L., DESCLES J-P., CARTIER E., CCRISPINO G., BENHAZZEZ S., JACKIEWICZ A., "Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText ", revue TSI, 2000.
- [15] MUC-3, 1991, *Proceedings Third Message Understanding Conference (DARPA)*, Morgan Kaufmann Publisher, San Francisco.
- [16] MUC-4, 1992, *Proceedings Fourth Message Understanding Conference (DARPA)*, Morgan Kaufmann Publisher, San Francisco.
- [17] MUC-5, 1995, *Proceedings Fifth Message Understanding Conference (DARPA)*, Morgan Kaufmann Publisher, San Francisco.
- [18] MUC-6, 1996, *Proceedings of the sixth Message Understanding Conference (DARPA)*, Morgan Kaufmann Publisher, San Francisco, 1996.
- [19] MUC-7, 1998, *Proceedings of the seventh Message Understanding Conference*, <http://www.muc.saic.com>.
- [20] N. Sager, *Natural Language Information Processing: A Computer Grammar of English and Its Applications*, Addison-Wesley Publishing, Reading, Massachusetts, 1981.
- [21] Soderland S., Lehnert W., 1994. « Wrap-up: A trainable discourse module for information extraction ». In *Journal of Artificial Intelligence Research* (2), pp 131-158., 2007
- [22] Shaalan Khaled, Person Name Entity Recognition for Arabic, proceedings of the 5th workshop on Important Unresolved Matters, pages 17-24, Prague, Czech Republic, June 2007.